Journal of Information Systems Engineering and Management
2025, 10(62s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Attention-Enhanced Deep Feature Fusion: A Comprehensive Framework for Multimodal Sentiment Analysis

Soumya Sharma*^, Dr. Srishti^, Dr. Deepak Gupta#
^*The NorthCap University, Gurugram, India*
#*MAIT, Rohini, Delhi, India*
**Corresponding Author: Soumya Sharma*
soumya21csd006@ncuindia.edu

ARTICLE INFO

Received: 29 Dec 2024

Revised: 12 Feb 2025

Accepted: 27 Feb 2025

ABSTRACT

Multimodal Sentiment Analysis (MSA) is an evolving area absorbed on capturing multifaceted human feelings by integrating diverse data types, inclusive of text, audio, and visuals. Leveraging multiple modalities enhances sentiment prediction by enriching the representational capacity of model architectures. Addressing the challenges in this domain, we put forward a novel approach that unites advanced feature extraction techniques—BERT for textual data, Wave2Vec2 for acoustic data, and Vision Transformer for visuals—with BiLSTM networks augmented by self-attention and multi-head attention mechanisms. The proposed architecture effectively extracts and fuses modality-specific features to construct a robust multimodal representation. Evaluated on the benchmark CMU-MOSI dataset, the projected model achieves 82.45% accuracy with the self-attention mechanism and 86.04% with the multi-head attention mechanism, surpassing several transformer-based states of the art approaches. This superiority stems from the handcrafted feature extraction, effective fusion strategy, and the ability of BiLSTM with multi-head attention to seize diverse relationships without overfitting. The findings highlight that hybrid architectures that integrate the advantages of transformers with attention-augmented recurrent models can outperform pure transformer-based designs for multimodal sentiment analysis.

Keywords: Multimodal Sentiment Analysis, Feature Extraction, Text Features, Audio Features, Visual Features, Self-attention Mechanism

1. INTRODUCTION-

MSA is an innovative and important area of SA study. It utilizes various modalities besides textual modality to understand the feelings or opinion. The breakthrough in ML and DL, expressed by the incorporation of transformers and generative AI, have simplified the evolution of various models and methods intended to address the growing complication of SA tasks. When paralleled to SA, MSA delivers better efficiency and an all-inclusive data, thus substituting SA in the recent years. Multimodal information comprises of multiple modalities that exist in world, like text, audio, visual and many other physical signs. These input modalities can be combined in any number of combinations. Amongst these combinations, the text-audio-visual combination are frequently utilized for SA. These combinations are predominantly preferred as they are particularly simple for both scientists and common masses to comprehend and document [1-3]. Fusing or combining multiple modalities is beneficial for SA. SA corpora on social networking websites need fusion techniques for combining the modalities [4]. More and more users on social media platforms like Facebook, YouTube, and TikTok favour to communicate their viewpoint, feelings, and opinions through visual and acoustic content. Intensified use of smartphones and low-cost internet service has enhanced multimodal data revealing. Individuals communicate their feedback related to latest released films, or merchandises, or travel places or any additional topic in the method of audio-visual content. This provides a fantastic chance for various companies to multiply their income and to give excellent customer knowledge by mining and testing customer sentiment, grievances, and references from the multimodal feedbacks. The understanding mined from multimodal data advances the level of living of a client by making precise choices as to which product to purchase or which destination to go to or which film to see or whose facility to use etc [5-8].

Journal of Information Systems Engineering and Management
2025, 10(62s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

The benefit of analysing various modalities for sentiment assignment over only text or audio or visual input is the existence of visual and audio cues, and facial aspects in video. The verbal pitch, physique movements and facial gestures in the audio-visual content, with recorded text data, helps to recognize the sentiment efficiently. Hence, an amalgamation of these various modalities aids in making a stronger and more efficient sentiment identification model. Multimodal information has data in 3 various modalities like text (or transcribed audio), acoustic and visual. Basic MSA methods excerpt the features from the various input sources or modalities, implement different feature fusion methods, either feature/early-level fusion or decision/late level fusion to combine useful information from various modalities [9-11]. An utterance is a section of the video (may or may not be a full sentence) and these review videos comprise of an arrangement of such numerous utterances. In utterance level sentiment analysis, every individual utterance is analysed individually and assigned a label.

2. RELATED WORKS

As a huge quantity of multimodal data is exchanged on social networking platforms, SA has become quite prominent among in the scientific society. The data gathered from these social media platforms are studied for guessing the end-user's sentiment for an event, entity, product or topic [12]. Fundamentally, SA is performed on text data [13] but lately audio-visual data is being examined for the same [14]. Recent advancements in this field are being performed in the English language, but lately linguistics diversity matters are also being addressed [15-16]. SA methods are planned utilizing either extracted features [17] or ontologies [18-19] or sentiment lexicons [20] or sentiment networks [21].

From the outside, SA appears as an easy and direct classification method, but essentially it is a demanding and difficult area of research as it is needed to study many assignments such as word polarity [22], topic recognition [23], subjectivity identification and its study [24], text encapsulation [25], and aspect deliberation [26]. The world of commerce is utilizing SA to undertake many outstanding problems like election results prediction [27], financial analysis [28], standard of commodity and service [29], travel and tourism boosting [30], healthcare facilities [28], movie performance forecast [31], etc.

Initially, SA was being performed only on 1 type of modality, naming it as the unimodal SA model. although easy to implement, it was observed that the unimodal models were not best at the task of SA. Researchers then progressed to combining 2 different types, terming it as the bimodal SA model. it was observed that the bimodal models have better performance than the unimodal models. it was combining 2 different types of input features contributed to more and better learning of the model. further advancements let to the researchers more than 2 modalities, hence, terming the model as multimodal sentiment analysis (MSA). The time taken to combine these modalities and to train the model was relatively high, but the results achieved from these multimodal models were the best. To combine the features, various feature fusion methods have been used. Over the period of time, various ML algorithms have been used for the task of SA in the unimodal, bimodal and trimodal or multimodal setup. Moving on from the ML algorithms, the DL algorithms were also used for the same task. It was observed and hence concluded that the DL algorithms perform better than the ML algorithms in all experimental setups- unimodal, bimodal and trimodal or multimodal.

Many researchers over time have projected various approaches for sentiment prediction in the bimodal and the multimodal experimental setup, as expressed in table 1. One of the models were proposed by Mai et al. [33]. They have created a model that works with the redundant embeddings that are created during the multimodal fusion process and the distinctive unimodal embeddings that ignored during the fusion process. This can often influence the overall accuracy of the model. they proposed their model MIB, multimodal information bottleneck, whose main aim was to acquire a controlling and adequate multimodal representation which is not influenced by repetitiveness and can screen noisy data in unimodal depiction. Their proposed model regulates multimodal and unimodal depictions, equally, which is an all-inclusive and accommodating framework that is well-suited with all fusion procedured. Their proposed model had 3 different variants- early fusion MIB, late fusion MIB and complete MIB. Their model, implemented on CMU-MOSI dataset gave an accuracy of 79.8% with early fusion variant. To further understand and analyze the problem of MSA, Hazarika et al. [34] have proposed their novel framework, Modality-Invariant and -Specific Representations for Multimodal Sentiment Analysis (MISA). Their proposed framework

Journal of Information Systems Engineering and Management
2025, 10(62s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

primarily deals with 2 underlying problems of MSA- modality invariation and modality specification. Their projected approach shows an accuracy of 81.8% on the CMU-MOSI dataset. Yuan et al. [35] have proposed their novel framework (NIAT)that works with data imperfections. In their model, they we articulate the inadequacy along with missing modality feature at the time of training the model and suggested an adversarial training framework based on novel noise intimation to enhance the efficiency in contrast to many possible inadequacies at the conclusion time. Their projected model utilizes temporal feature erasing followed by adversarial training approach with semantic rebuilding. Their model shows an accuracy of 68.02% with random missing, with temporal missing shows and accuracy of 67.95% and with structural temporal missing shows an accuracy of 61.99%. a novel approach, SentDep has been proposed by Lu et al. [36] to understand the complex human emotion by integrating various data types. In their, they have proposed a novel feature fusion technique. They have used the CLIP model to process the text and the visual features. and then have combined the acoustic features with the derivations of the CLIP model in their proposed feature fusion model. their proposed approach delivers a result of 45.5%. Sun et al. [37] in their novel approach have dealt with the issue of adding the false associations between multimodal features and sentiment labels. They have proposed a universal debiasing task, which targets to improve the Out-Of-Distribution (OOD) generalization capability by decreasing their dependence on spurious correlations and have developed a new robust feature fusion method. Their model shows an accuracy of 81.30%. Huddar et al. [38] have proposed another fusion technique to enhance MSA. They have described a new attention-based multimodal contextual fusion strategy, which abstracts the contextual material amongst the utterances prior to fusion. Their proposed model of BiLSTM with attention mechanism gives an accuracy of 80.18%. Xiao et al. [39] have proposed a model where they have incorporated sentimental facts into inter-modality knowledge. Their model consists of 2 main sections- cross-modality interactive learning and sentimental feature fusion. Their models, Multi-channel Attentive Graph Convolutional Network (MAGCN), delivers an accuracy of 80.6%. Another approach, SWAFN: Sentimental Words Aware Fusion Network for Multimodal Sentiment Analysis, proposed by Chen and Li [40], present a more advanced feature fusion approach where they integrate sentimental word facts into the fusion process to steer the gathering of combined depiction of multimodal features. Their approach contains shallow fusion and aggregation and shows an accuracy of 80.2%.

Table 1 Previous work on CMU-MOSI dataset

S.no	Reference	Modalities	Feature Extraction	Fusion Method	Model used	Accuracy
1	Mai et al. [33]	Text, Audio, Video	FACET Transformer for visual, BERT for text features, COVAREP for audio features	Early fusion, late fusion, complete fusion	Early-fusion MIB, late-fusion MIB, and complete MIB	Accuracy (with the early fusion)- 79.8%
2	Hazarika et al. [34]	Text, Audio, Video	BERT based-uncased pre-trained model for text, Facet for visual, COVAREP for audio	Self-attention supported by a succession of all the 6 converted modality vector	MISA	Accuracy-81.8%,

Journal of Information Systems Engineering and Management
2025, 10(62s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

S.no	Reference	Modalities	Feature Extraction	Fusion Method	Model used	Accuracy
3	Yuan et al. [35]	Text, Audio, Video	BERT toeknizer for text features, audio and video from CMU-Multimodal SDK	Late-Fusion method	NIAT	Accuracy with temporal missing-67.95% with random missing-68.02% with structural temporal missing-61.99%
4	Lu et al. [36]	Text, Audio, Video	Text features from transcription of the utterances, acoustic features like pitch, intensity and tempo, visual features from body gestures and facial gestures.	Early Fusion, Late Fusion, and Hybrid Fusion, Tensor fusion module	SentDep	Accuracy - 45.5%
5	Sun et al. [37]	Text, Audio, Video	Pre-trained BERT model for text features, acoustic and visual features	Robust Feature Fusion	GEAR	Accuracy - 81.30%
6	Huddar et al. [38]	Text, Audio, Video	CNN for textual features, OPENSmile for acoustic features, 3D-CNN for visual features	Attention-based multimodal contextual fusion	BiLSTM with attention mechanism	Accuracy - 80.18%
7	Xiao et al. [39]	Text, Audio, Video		Multi-head Self-attention	MAGCN	Accuracy - 80.6%

From the survey of existing works in multimodal sentiment analysis, it is evident that most frameworks either rely solely on unimodal or bimodal inputs, or they primarily leverage transformer-based architectures without deeply integrating handcrafted feature extraction. While these methods have shown promising results, they often disappoint in capturing the full complexity of utterance-level sentiment, especially when modalities are noisy or incomplete.

Our proposed approach addresses these limitations through several key advantages and novel contributions:

i. **Novelty in Feature Extraction and Fusion**: Unlike many existing methods, we have exclusively **handcrafted feature extraction pipelines** tailored for each modality: Enhanced BERT + LSTM + attention for text, Wave2Vec2 for audio, and Vision Transformer (ViT) for video. These features were integrated using **feature-level (early) fusion**, enabling the model to combine raw contextual information from all modalities before sequence modeling.

ii. **BiLSTM with Dual Attention Strategies**: We extend the BiLSTM framework with two attention mechanisms: **self-attention** and **multi-head attention**. While self-attention provides a lightweight mechanism to capture dependencies, **multi-head attention captures diverse relationships within data simultaneously**, enhancing robustness and contextual understanding.

Journal of Information Systems Engineering and Management
2025, 10(62s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

iii. **Utterance-Level Processing**: Many prior studies focus on video-level or segment-level sentiment, but our framework emphasizes **utterance-level sentiment analysis**, which provides more fine-grained predictions and better captures sentiment shifts within conversations.

iv. **Generalization and Robustness**: By implementing **5-fold cross-validation**, our framework avoids overfitting and learns from the entire dataset, unlike many prior works that rely on fixed train-test splits. This ensures that the performance gains are consistent and not dataset-specific.

3. PROPOSED APPROACH

In this segment, we elaborate on the proposed architecture of utterance-level MSA with 2 different attention mechanism- self-attention mechanism and multi-head attention mechanism using the BiLSTM architecture. The architecture is being used in the unimodal, bimodal and multimodal experimental setup. The summary of the proposed approach is as below-

a. First, we extract utterance level text, audio and visual features.

b. Second, features are fused in bimodal manner (text+audio, audio+video, video+text) and multimodal manner (text+audio+video)

c. Third, we implement our BiLSTM model with 2 different attention mechanism- self-attention mechanism and multi-head attention mechanism and all the features are inputted to the different models in unimodal (text, audio, video), bimodal (text+audio, audio+video, video+text) and multimodal manner (text+audio+video) and the accuracy for each experimental setup is calculated.

3.1 Dataset Used

For this research project, we have used the openly available CMU-MOSI dataset [40]. The CMU-MOSI dataset comprises of 93 videos in which there are 89 separate spokesperson reviewing various themes/products in English language. The videos clips are broken into 2199 opinion small videos, as shown in table 2, with a typical utterance segment being of 4.2s and 12 words each utterance. Every utterance is physically annotated by 5 different annotators, labelling each video with a count between −3 and +3, where the score of −3 shows strongly negative and +3 shows strongly positive. A typical of these 5 annotations is chosen to calculate the overall sentiment polarity. For our research work, we have worked with only positive and negative classes.

	Positive	**Negative**
Train	709	738
Test	467	285

Table 2 Train-Test distribution of CMU-MOSI dataset

Figure 1 Video Snapshot from the dataset [41]

3.2 Feature Extraction

This segment discusses the techniques used for extracting features from text, audio and visual inputs.

a. Text feature extraction

To extract features from text input, the inputs are converted to Bidirectional Encoder Representations from Transformers (BERT) [42] embeddings. BERT converts text data into numerical vectors or embeddings which help in capturing the semantic meaning of the words and also summarizes the contextual dependency amongst the words in a sentence. BERT is a language model that produces contextualized embeddings for words, meaning the

Journal of Information Systems Engineering and Management
2025, 10(62s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

representation of a word changes based on the adjacent words. For our research purpose, we have custom an EnhancedBERT model where the model has a BERT base model along with an LSTM layer for sequence processing. To make the model more efficient, an attention mechanism is also added to it. If f_{ti} represents text feature vector of i^{th} utterance, the text feature vector f_t is outlined as

$$f_t = \langle f_{t1}, f_{t2}, f_{t3}, \dots, f_{tn} \rangle \quad (1)$$

where, n= number of utterances.

b. Audio feature extraction

To extract the audio features, Wave2Vec2 [43] was used. Wave2Vec2 is a self-supervised DL model that is primarily used for automatic speech recognition jobs like sentiment analysis and emotion recognition. Developed by MetaAI, it takes the raw audio waveforms as the input without needing transcripts for the pre-training phase. As it uses the transformer model, they have an in-built capacity for contextual understanding. For our research purpose, we have resampled the audio waveforms are 16kHz and have extracted contextual embeddings from the audio waveforms. If f_{ai} represents audio feature vector of i^{th} utterance, the audio feature vector f_a is outlined as

$$f_a = \langle f_{a1}, f_{a2}, f_{a3}, \dots, f_{an} \rangle \quad (2)$$

where, n= number of utterances.

c. Visual Feature Extraction

To extract the visual features, we have created a VideoClassifier using the Vision Transformer (ViT) [44]. The ViT is a visual model which is built on the design of a transformer, which was initially designed for text-based undertakings. It shows an input image as a sequence of image patches, and straightforwardly calculates class tags for the image. For our research work, we have analysed the input video in a frame-wise manner. Let f_{vi} represents visual feature vector of i^{th} utterance, the visual feature vector f_v is outlined as

$$f_v = \langle f_{v1}, f_{v2}, f_{v3}, \dots, f_{vn} \rangle \quad (3)$$

where, n= number of utterances.

3.3 **Feature Fusion**- in our research work, we have performed SA not only on unimodal data but also on bimodal and trimodal data. For the bimodal experimental setup, we have combined features extracted from the inputs in pairs of 2- text+audio, audio+video and video+text. Similarly, in the trimodal experimental setup, we have combined the extracted features of all the 3 inputs, i.e., features from text, audio and video are combined together. For our research work, we have implemented feature level or early fusion [45]. In feature level fusion, all the features are combined or concatenated into a single feature vector. The resultant length of the feature vector is the sum of the number features of all the modalities combined. This level of feature fusion is also called early level fusion as the features are first combined and then are inputted to the algorithm or the model. however, the features concatenated from different modalities have different nature and different range. One main advantage of feature-level fusion is the capability to deal with noisy or inadequate data. If one modality is incomplete, others can reimburse for it and complete the information gap as all contributing modalities add knowledge to the model. For example- in audio-visual speech recognition, background noise might damage audio attributes, but facial features data from video can fulfil the shortcomings in understanding spoken words.

In our experimental procedure, for all the 3 input modalities, the features extracted are numerical are nature. Hence, all features have the same nature but they all belong to different range and also vary from each other in great deal. in order to ensure that they all belong to the same range, features are normalized before they are fused. Feature normalization, also called as feature scaling, is a data pre-processing method that transforms the values of the numerical features in a dataset to a general scale. Doing this ensures that features with greater numerical variations do not unreasonably influence the learning algorithm as contrasted to features with smaller variations. For our research purpose, we have used z-score normalization. It centres data around a mean of 0 and standard deviation of 1, which makes it an appropriate algorithm that assume a normal distribution. It improves the model's performance by dealing with the outliers and scaling them to a common range.

Journal of Information Systems Engineering and Management
2025, 10(62s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

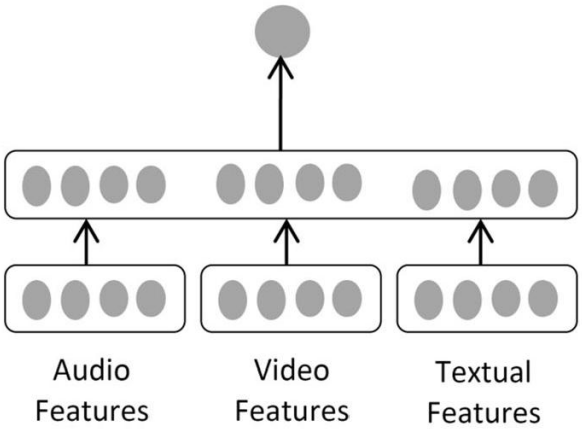

Figure 2 Feature/early Fusion

However, the duration and count of utterances in every video is dissimilar; thus, smaller videos are amplified with null vectors to make every video of identical length. After combining the extracted features using the feature-level fusion or early fusion technique in a bimodal manner, we get the following feature vector

$$f_{at} = \langle f_{t1}, f_{t2}, f_{t3}, \dots, f_{tn}, f_{a1}, f_{a2}, f_{a3}, \dots, f_{an} \rangle \quad (4)$$

$$f_{av} = \langle f_{a1}, f_{a2}, f_{a3}, \dots, f_{an}, f_{v1}, f_{v2}, f_{v3}, \dots, f_{vn} \rangle \quad (5)$$

$$f_{tv} = \langle f_{t1}, f_{t2}, f_{t3}, \dots, f_{tn}, f_{v1}, f_{v2}, f_{v3}, \dots, f_{vn} \rangle \quad (6)$$

where, n= number of utterances

f_{at}= feature vector of fused text and audio features

f_{av} =feature vector of fused audio and video features

f_{tv} = feature vector of fused video and text features

Similarly, after combining the extracted features using the feature-level fusion or early fusion technique in a trimodal manner, we get the following feature vector

$$f_{atv} = \langle f_{t1}, f_{t2}, f_{t3}, \dots, f_{tn}, f_{a1}, f_{a2}, f_{a3}, \dots, f_{an}, f_{v1}, f_{v2}, f_{v3}, \dots, f_{vn} \rangle \quad (7)$$

where, n= number of utterances

f_{atv}= feature vector of fused text, audio and video features

3.4 Bidirectional LSTM with Self-Attention model

The vanishing gradient problems affects the back propagation-based problems and the gradient-based problems. The long short-term memory (LSTM) based recurrent neural network (RNN) works on the vanishing gradient issue. Many various versions of LSTM were recommended. For our research work, we have adopted an alternative of the LSTM network, bidirectional LSTM model, BiLSTM. Much unlike the traditional LSTM, BiLSTM network read the input in both forward and backward direction. Due to this, the BiLSTM network is capable of a better contextual understanding and hence better accuracy. The BiLSTM has 2 sub-networks- a forward network and a backward network for sequence processing.

Journal of Information Systems Engineering and Management
2025, 10(62s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

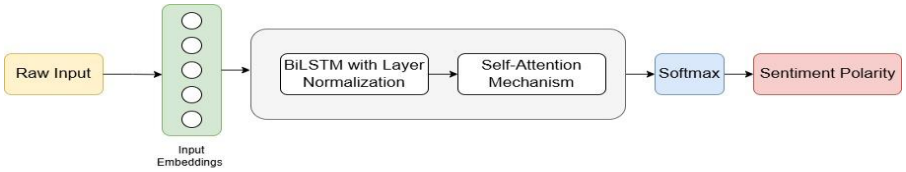

Figure 3 Proposed Unimodal Approach

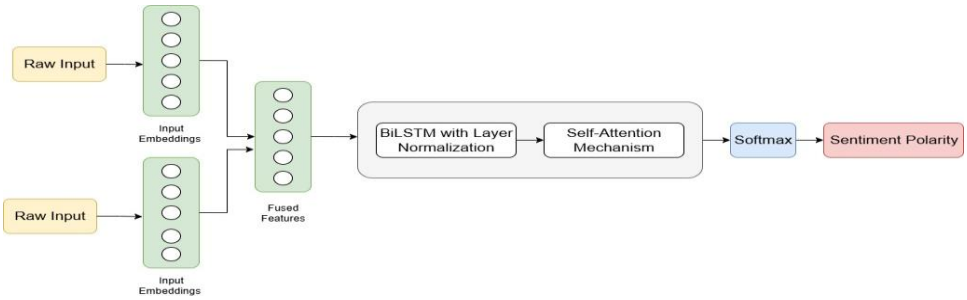

Figure 4 Proposed bimodal approach

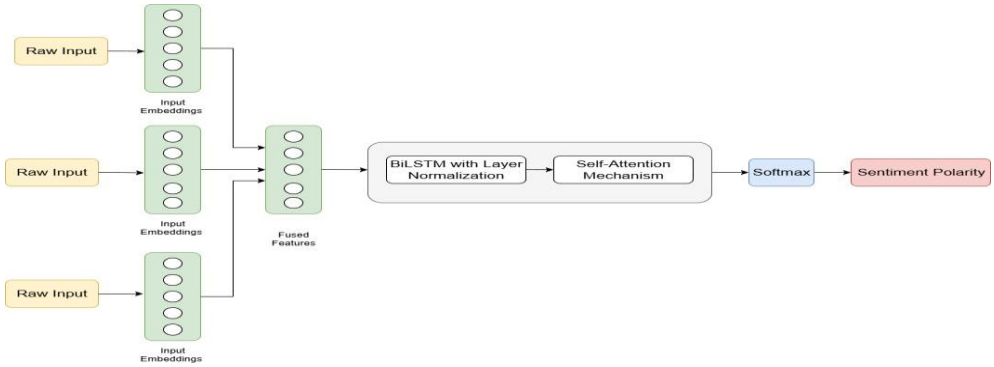

Figure 5 Proposed trimodal approach

In this research work, we have proposed 2 different approaches- in 1st method, we implement BiLSTM network with self-attention mechanism with added layer normalization and dropout regularization and in the 2nd method, we implement BiLSTM network with multi-head attention mechanism with added layer normalization and dropout regularization. Although adding the attention mechanism to the BiLSTM network improves the contextual understanding of the model and captures long range dependencies, there is, however, a great difference between the 2 approaches. The self-attention mechanism works with 1 set of attention weight whereas the multi-head attention mechanism makes multiple "heads", where each head has its own set of attention weights, thus, allowing the model to concentrate on various aspects of the input. Self-attention is capable of capturing only 1 relationship/pattern in the sequence whereas with the multi-head mechanism, each head is capable of learning of different relationships. In other words, self-attention calculates a single set of attention weights for each position, whereas, the multi-head attention extends this by making manifold sets of weights in parallel, letting the model to capture different relationships and perspectives within the input. Due to this although the self-attention mechanism is computationally less expensive, it might not be able to capture the complete relationship within data. To make the model more accurate and efficient, layer normalization is also done.

In traditional neural networks (NN), activation functions of every layer can differ extremely which can lead to problems like that exploding gradient or vanishing gradients, which can reduce the training process speed. Adding layer normalization to the network deal with this by normalizing the output of every layer, thereby, safeguarding the model so that the activation functions stay within a stable range. To ensure that the model does not overfit, we have implemented a dropout mechanism. In DL, dropout is a regularization mechanism where neurons are randomly

Journal of Information Systems Engineering and Management
2025, 10(62s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

deactivated during the model training phase to prevent the model from overfitting and to enhance accuracy. For our research work, we have implemented a dropout rate of 0.30 and have used Adam optimizer.

Hyperparameter	Self-Attention Parameters	Multi-Head Attention Parameters
Dropout rate	0.30	0.30
Optimizer	Adam Optimizer	Adam Optimizer
Epoch	20	50
Batch Size	32	32
Hidden Layers	256	256
Self-Attention Head count	8	16
Cross-Validation Rate	5	5
Initial Learning Rate	1.00E-03	1.00E-05

Table 3 Implementation Details

In this research work, we are trying to compute and compare the efficiency of the proposed model on the unimodal, bimodal and trimodal experimental. For the features extracted and the model implemented, we define the following equations-

$$B_t = biLSTM\langle f_t \rangle \quad (8)$$

$$B_a = biLSTM\langle f_a \rangle \quad (9)$$

$$B_v = biLSTM\langle f_v \rangle \quad (10)$$

$$B_{at} = biLSTM\langle f_{at} \rangle \quad (11)$$

$$B_{av} = biLSTM\langle f_{av} \rangle \quad (12)$$

$$B_{tv} = biLSTM\langle f_{tv} \rangle \quad (13)$$

$$B_{atv} = biLSTM\langle f_{atv} \rangle \quad (14)$$

Where, B_t= proposed biLSTM model implemented on extracted text features

B_a= proposed biLSTM model implemented on extracted audio features

B_v= proposed biLSTM model implemented on extracted video features

B_{at}= proposed biLSTM model implemented on fused text-audio features

B_{av}= proposed biLSTM model implemented on fused audio-video features

B_{tv}= proposed biLSTM model implemented on fused video-text features

B_{atv}= proposed biLSTM model implemented on fused text-audio-video features

4. RESULT ANALYSIS

To ensure more efficiency and better results of our proposed mode, detailed in table 3, we have implemented 5-fold cross validation. Implementing cross validation ensures that the model trains and tests on the whole dataset and does not overfit. We have compared our work with some of the previously proposed benchmark works.

Journal of Information Systems Engineering and Management
2025, 10(62s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Proposed Model (with 5-fold cross validation)	Accuracy (with self-attention mechanism	Accuracy (with Multi-head attention mechanism
B_t	70.11%	72.17%
B_a	71.42%	72.66%
B_v	70.95%	71.19%
B_{at}	72.47%	73.39%
B_{av}	71.67%	74.01%
B_{av}	72.01%	77.89%
$\boldsymbol{B_{atv}}$	**82.45%**	**86.04%**

Table 4 Results from the proposed approach

As observed from table-4, we conclude that the multi-head attention mechanism outperforms the existing state-of-the-art model, from table-5 and even out performs our self-attention approach. This is due to the fact that with multi-head attention, we can capture long-term contextual dependencies between the inputs. We also conclude that the bimodal experimental setup, under both the approaches, outperforms the unimodal approaches, thus, confirming to our literature.

Model	Accuracy
MISA [34]	81.80%
NIAT [35]	81.82%
SentDep [36]	45.50%
GEAR [37]	81.30%
BiLSTM with Attention Mechanism [38]	80.18%
MAGCN [39]	80.60%
SWAFN [40]	80.20%
CMJRT [46]	82.40%
HyCON [47]	85.20%
MEDT [48]	84.91%
Proposed Approach (with self-attention)	**82.45%**
Proposed Approach (with multi-head attention)	**86.04%**

Table 5 Proposed Approach VS State-of-the-Art Models

Upon contrasting our projected approach, detailed in table-3, with the advanced model, like that of [43] or [45] which have implemented transformer architecture for the task of MSA, we can confirm that our proposed approach outperforms all the advanced models. The primary reason behind this is feature extraction methods that are used. The features extracted deeply influence the architecture of the model that we use and also have deep impact on hyper-parameter tuning of the model. Another reason for better accuracy is that in our approach, we have we have performed 5-fold cross-validation. Cross validating the model ensure that the model is trained and tested on the whole of the dataset, rather that just the same train-test split. This also ensure that the model learns the insights from the whole dataset and does not overfit by learning from the same data again and again. In our proposed model, we

Journal of Information Systems Engineering and Management
2025, 10(62s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

have added an additional layer of self-attention. Adding self-attention mechanism helps in capturing long term dependencies and also makes the model more robust to noisy data. This robustness improves the overall accuracy of the model.

5. CONCLUSION AND FUTURE WORK

Multimodal Sentiment Analysis (MSA) is an ever-developing field, with new models and architectures being proposed regularly. In this work, we have presented a novel utterance-level MSA framework that integrates a BiLSTM architecture with two different attention mechanisms—self-attention and multi-head attention. Unlike existing approaches, we have exclusively handcrafted features for all three modalities—text, audio, and video—ensuring robust contextual representation and deep semantic understanding. Textual embeddings were extracted using an Enhanced BERT model with an additional LSTM layer and attention mechanism, audio embeddings were derived using Wave2Vec2 with resampling at 16kHz, and visual features were captured through Vision Transformer (ViT) for frame-wise analysis. To unify multimodal information, feature-level (early) fusion was applied in unimodal, bimodal (text+audio, audio+video, text+video), and trimodal (text+audio+video) experimental setups, followed by training BiLSTM with self-attention and multi-head attention.

Our experiments on the CMU-MOSI dataset demonstrated that the trimodal setup consistently outperformed unimodal and bimodal setups, achieving the highest accuracy of 86.04% with the multi-head attention mechanism and 82.45% with self-attention. This performance is superior not only to our own unimodal and bimodal models but also to several benchmark and transformer-based state-of-the-art architectures. The results highlight that multimodal models, which leverage diverse sources of information, create a more comprehensive and precise interpretation of reality compared to unimodal or bimodal models. Furthermore, the use of 5-fold cross-validation ensured robust generalization and minimized overfitting.

The superiority of our proposed approach lies in three key aspects:

1. Robust handcrafted feature extraction from all modalities.
2. BiLSTM's bidirectional sequence modeling, capturing contextual dependencies from both past and future utterances.
3. Multi-head attention, which captures diverse relationships within the input data and improves model robustness without overfitting.

Despite these achievements, future work can focus on enhancing cross-modal interactions through advanced fusion techniques, larger multilingual datasets, and temporal modeling of inter-utterance dependencies. Additionally, exploring multimodal LLMs and optimizing efficiency with lightweight architectures or model compression can further improve scalability and real-time applicability.

Conflict of Interest

Competing Interests- Not Applicable
Funding Information- Not Applicable
Author contribution- Not Applicable
Data Availability Statement- Not Applicable
Research Involving Human and /or Animals- Not Applicable
Informed Consent- Not Applicable

REFERENCES

[1] Jim, J. R., Talukder, M. A. R., Malakar, P., Kabir, M. M., Nur, K., & Mridha, M. F. (2024). Recent advancements and challenges of NLP-based sentiment analysis: A state-of-the-art review. *Natural Language Processing Journal*, 100059.

[2] Das, R., & Singh, T. D. (2023). Multimodal sentiment analysis: a survey of methods, trends, and challenges. *ACM Computing Surveys*, *55*(13s), 1-38.

[3] Gandhi, A., Adhvaryu, K., Poria, S., Cambria, E., & Hussain, A. (2023). Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions. *Information Fusion*, *91*, 424-444.

[4] Zhu, L., Zhu, Z., Zhang, C., Xu, Y., & Kong, X. (2023). Multimodal sentiment analysis based on fusion methods: A survey. *Information Fusion*, *95*, 306-325.
[5] César, I., Pereira, I., Rodrigues, F., Miguéis, V., Nicola, S., Madureira, A., ... & De Oliveira, D. A. (2024). A Systematic Review on Responsible Multimodal Sentiment Analysis in Marketing Applications. *IEEE Access*.
[6] Zhu, L., Zhu, Z., Zhang, C., Xu, Y., & Kong, X. (2023). Multimodal sentiment analysis based on fusion methods: A survey. *Information Fusion*, *95*, 306-325.
[7] Sharma, S., Sharma, S., & Gupta, D. (2023, November). Multimodal Sentiment Analysis and Multimodal Emotion Analysis: A Review. In *International Conference on Computing and Communication Networks* (pp. 371-382). Singapore: Springer Nature Singapore.
[8] Sharma, S., Sharma, S., & Gupta, D. (2023, November). Unveiling Emotions in the Digital Arena: Sentiment Analysis of YouTube Comments on FIFA World Cup 2022. In *2023 3rd International Conference on Technological Advancements in Computational Sciences (ICTACS)* (pp. 994-997). IEEE.
[9] Murthy, J. S., & Siddesh, G. M. (2024). Multimedia video analytics using deep hybrid fusion algorithm. *Multimedia Tools and Applications*, 1-19.
[10] Paul, A., & Nayyar, A. (2024). A context-sensitive multi-tier deep learning framework for multimodal sentiment analysis. *Multimedia Tools and Applications*, *83*(18), 54249-54278.
[11] Zhang, R.; Xue, C.; Qi, Q.; Lin, L.; Zhang, J.; Zhang, L. Bimodal Fusion Network with Multi-Head Attention for Multimodal Sentiment Analysis. Appl. Sci. **2023**, 13, 1915. https://doi.org/10.3390/app13031915
[12] Cambria E (2016) Affective computing and sentiment analysis. IEEE Intell Syst 31(2):102–107
[13] Lo SL, Cambria E, Chiong R, Cornforth D (2017) Multilingual sentiment analysis: from formal to informal and scarce resource languages. Artif Intell Rev 48(4):499–527.
[14] Huddar MG, Sannakki SS, Rajpurohit VS (2019) A survey of computational approaches and challenges in multimodal sentiment analysis. Int J Comput Sci Eng 7(1):876–883.
[15] Gandhi, A., Adhvaryu, K., Poria, S., Cambria, E., & Hussain, A. (2023). Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions. *Information Fusion*, *91*, 424-444.
[16] Penga H, Ma Y, Lib Y, Cambria E (2018) Learning multi-grained aspect target sequence for Chinese sentiment analysis. Knowl-Based Syst 148:167–176.
[17] Kiritchenko S, Zhu X, Mohammad SM (2014) Sentiment analysis of short informal texts. J Artif Intell Res 50:723–762.
[18] Thakora P, Sasi DS (2015) Ontology-based sentiment analysis process for social media content. Procedia Comput Sci 53:199–207.
[19] Sharma, S., Saraswat, M., & Dubey, A. K. (2022, November). Multi-aspect sentiment analysis using domain ontologies. In *Iberoamerican Knowledge Graphs and Semantic Web Conference* (pp. 263-276). Cham: Springer International Publishing.
[20] Mohammad SM, Kiritchenko S, Zhu X (2013) NRC-Canada: Building the state-of-the-art in sentiment analysis of tweets. In: Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013), Atlanta, pp 321–327.
[21] Nalisnick ET, Baird HS (2013) Extracting sentiment networks from Shakespeare's plays in 12th International Conference on Document Analysis and Recognition, Washington, DC, USA.
[22] Lyu K, Kim H (2016) Sentiment analysis using word polarity of social media. Wirel Pers Commun 89(3): 941–958.
[23] Peng B, Li J, Chen J, Han X, Xu R, Wong K-F (2015) Trending sentiment-topic detection on twitter. In: International Conference on Intelligent Text Processing and Computational Linguistics. Springer, Cham, pp 66–77.
[24] Korayem M, Crandall D, Abdul-Mageed M (2012) Subjectivity and sentiment analysis of arabic: A survey. In: International conference on advanced machine learning technologies and applications. Springer, Berlin, Heidelberg, pp 128–139.
[25] Gupta P, Tiwari R, Robert N (2016) Sentiment analysis and text summarization of online reviews: a survey, " in International Conference on Communication and Signal Processing (ICCSP), Melmaruvathur, India.

[26] de Kok S, Punt L, van den Puttelaar R, Ranta K, Schouten K, Frasincar F (2018) Review-aggregated aspect based sentiment analysis with ontology features. Prog Artif Intell 7(4):295–306.
[27] Ramteke J, Shah S, Godhia D, Shaikh A (2016) Election result prediction using twitter sentiment analysis. International Conference on Inventive Computation Technologies (ICICT), Coimbatore.
[28] Li X, Xie H, iChenb L, Wang J, Deng X (2014) News impact on stock price return via sentiment analysis. Knowledge-Based Syst 69:14–23.
[29] Mars A, GouiderMS (2017) Big data analysis to features opinions extraction of customer. Procedia Comput Sci 112:906–916.
[30] Kiritchenko S, Zhu X, Mohammad SM (2014) Sentiment analysis of short informal texts. J Artif Intell Res 50:723–762.
[31] Gohil S, Vuik S, Darzi A (2018) Sentiment analysis of health care tweets: review of the methods used, JMIR Public Health Surveill 4(2).
[32] Nagamma P, Pruthvi HR, Nisha KK, Shwetha NH (2015) An improved sentiment analysis of online movie reviews based on clustering for box-office prediction, in International Conference on Computing. Communication & Automation, Noida
[33] Mai, Sijie, Ying Zeng, and Haifeng Hu. "Multimodal information bottleneck: Learning minimal sufficient unimodal and multimodal representations." *IEEE Transactions on Multimedia* 25 (2022): 4121-4134
[34] Hazarika, D., Zimmermann, R., & Poria, S. (2020, October). Misa: Modality-invariant and-specific representations for multimodal sentiment analysis. In *Proceedings of the 28th ACM international conference on multimedia* (pp. 1122-1131).
[35] Yuan, Z., Liu, Y., Xu, H., & Gao, K. (2023). Noise imitation based adversarial training for robust multimodal sentiment analysis. *IEEE Transactions on Multimedia*, *26*, 529-539.
[36] Lu, C., & Fu, X. (2024). SentDep: Pioneering Fusion-Centric Multimodal Sentiment Analysis for Unprecedented Performance and Insights. *IEEE Access*, *12*, 21277-21286.
[37] Sun, T., Ni, J., Wang, W., Jing, L., Wei, Y., & Nie, L. (2023, October). General debiasing for multimodal sentiment analysis. In *Proceedings of the 31st ACM International Conference on Multimedia* (pp. 5861-5869).
[38] Huddar, M. G., Sannakki, S. S., & Rajpurohit, V. S. (2021). Attention-based multimodal contextual fusion for sentiment and emotion classification using bidirectional LSTM. *Multimedia Tools and Applications*, *80*(9), 13059-13076.
[39] Xiao, L., Wu, X., Wu, W., Yang, J., & He, L. (2022, May). Multi-channel attentive graph convolutional network with sentiment fusion for multimodal sentiment analysis. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 4578-4582). IEEE.
[40] Chen, M., & Li, X. (2020, December). Swafn: Sentimental words aware fusion network for multimodal sentiment analysis. In *Proceedings of the 28th international conference on computational linguistics* (pp. 1067-1077).
[41] Zadeh, A., Zellers, R., Pincus, E., & Morency, L. P. (2016). Mosi: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. *arXiv preprint arXiv:1606.06259*.
[42] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019, June). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)* (pp. 4171-4186).
[43] Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, *33*, 12449-12460.
[44] Khan, S., Naseer, M., Hayat, M., Zamir, S. W., Khan, F. S., & Shah, M. (2022). Transformers in vision: A survey. *ACM computing surveys (CSUR)*, *54*(10s), 1-41.
[45] Ross, A. (2009). Fusion, Feature-Level. In: Li, S.Z., Jain, A. (eds) Encyclopedia of Biometrics. Springer, Boston, MA. https://doi.org/10.1007/978-0-387-73003-5_157
[46] Xu, M., Liang, F., Su, X., & Fang, C. (2022). CMJRT: Cross-modal joint representation transformer for multimodal sentiment analysis. IEEE Access, 10, 131671-131679.
[47] Mai, S., Zeng, Y., Zheng, S., & Hu, H. (2022). Hybrid contrastive learning of tri-modal representation for multimodal sentiment analysis. *IEEE Transactions on Affective Computing*, *14*(3), 2276-2289.

Journal of Information Systems Engineering and Management
2025, 10(62s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

[48] Qi, Q., Lin, L., Zhang, R., & Xue, C. (2022). MEDT: Using multimodal encoding-decoding network as in transformer for multimodal sentiment analysis. *IEEE Access*, *10*, 28750-28759.