DOI Link: https://doi.org/10.61586/rdTmo
Vol.70, Issue.9, Part.1, Sep 2025, PP. 2-28

SENTIMENT ANALYSIS FROM MULTIMODAL DATA USING JOINT REPRESENTATIONS AND ADAPTIVE TRANSFORMER OPTIMIZATION

Vasanthi P [1] and Madhu Viswanatham V [2] *

[1,2] School of Computer Science and Engineering, Vellore Institute of Technology, Vellore-632014, Tamil Nadu, India.
*Corresponding author: E-mail: vmadhuviswanatham@vit.ac.in

Received Feb 2025 Accepted Aug 2025 Published Sep 2025

ABSTRACT

Multimodal Sentiment Analysis (MSA) represents an advancing research domain focused on identifying sentiment through Audio(A), Video (V), and Text (T) modalities. A significant challenge lies in capturing joint representations by associating information across various modalities with integration techniques. Numerous existing approaches rely on acquiring joint representation through the concatenation of input features without fully exploiting interactions to ensure consistency with complementarity among modalities. To address this problem, a novel framework named Joint Representation with Optimized Transformer (JRT) has been developed to overcome the challenges in multimodal sentiment analysis through hierarchical interactions among modalities. The process begins with a diverse dataset comprising text, audio, and visual data from the CMU-MOSI and CMU-MOSEI datasets. Preprocessing steps are performed with normalization, noise reduction, and feature scaling to ensure data quality. Feature extraction is used to isolate meaningful patterns from each modality for further analysis. The framework integrates these features into a unified representation through a Joint Representation Translator (JRT), which aligns heterogeneous data for compatibility using cyclic translation. This translation process captures joint representations of bimodality by translating one modality into another forward with backward passes using encoder-decoders to maintain consistency between modalities. To explore complementarity among modalities, a transformer-based prediction mechanism is optimized with the Adaptive Dragon Optimization Algorithm (ADOA), which strengthens unimodal features with common information extracted from bimodality, enhancing model accuracy and convergence. Extensive experiments conducted with CMU-MOSI and CMU-MOSEI datasets validate the effectiveness of this framework, showing superior performance compared to existing methods. The proposed method achieves 95.11% accuracy (%), 97.51% recall (%), 95.14% F-Score (%), 0.90 correlation coefficient(unitless), 0.051 false positive rate (%), 97% negative predictive value (%), and 0.0489 mean squared error (MSE). These results highlight substantial advancements in accuracy and robustness over traditional approaches.

Keywords: Multimodal data, Joint Representation Translator, Sentiment analysis, Adaptive dragonfly optimization, Transformer.

1. INTRODUCTION

Multimodal Sentiment Analysis (MSA) represents a rapidly evolving area within artificial intelligence, focusing on predicting human emotions by leveraging data from multiple sources such as audio, video, and text [1-3]. By integrating the complementary information from these modalities, MSA enhances human-computer interactions and enables the development of empathetic and context-aware systems [4]. For instance, emotion recognition facilitates personalized learning in online education, supports natural interactions in conversational AI, and enables adaptive gaming environments based on players' emotional states [5]. Despite its potential, MSA faces significant challenges in achieving accurate sentiment prediction from multimodal data [6-7]. Challenges include modality alignment due to varying levels of information provided by different data sources, inconsistencies in feature representations and data quality across modalities, and the requirement for advanced fusion techniques to capture synergistic effects among audio, video, and text [8].

Current models, including Recurrent Neural Networks (RNNs) [9-10], Transformers [11-12], and BERT-based models [13-14], have made progress in multimodal analysis, yet understanding the interaction and contribution of each modality remains a critical issue. Fusion mechanisms, such as tensor-based models, address some of these issues but struggle with modality gaps inherent in heterogeneous data. Addressing the heterogeneity among modalities, the Multimodal Transformer (MulT) was designed by Tsai et al. [15] employs a cross-modal attention mechanism to adapt information from one modality to another. However, limitations persist, such as the exclusion of audio-video interactions when adapting to textual modalities. Research emphasizes the importance of three-way interactions among modalities to achieve optimal multimodal fusion. While joint representation learning aims to overcome these challenges by correlating information across modalities, existing methods often fail to fully explore the complementarity required for effective fusion [16].

This research introduces a novel framework named Joint Representation with Optimized Transformer (JRT), designed to address the key challenges in multimodal sentiment analysis. The framework prioritizes learning robust joint representations and enhancing multimodal interactions. By translating data between source and target modalities while maintaining semantic similarity, the framework achieves both bimodal consistency and multimodal complementarity. The Adaptive Dragon Optimization Algorithm (ADOA) refines modality-reinforced representations through hierarchical interactions and multi-head attention thus ensuring seamless integration of modalities.

2. RELATED WORKS

To regulate the training of MSA models, Ying Zeng et al. [17] introduced a unique framework called M^3SA. This framework intends to investigate the best possible solution for unimodal networks and multimodal embedding. Several baselines have been recommended for reducing the harmful effects of

contradictory data along with achieving accurate integrating in several modes, including soft-filtering, hard-filtering, and unimodal embedding. Although the framework utilized fewer parameters and achieved efficient noise elimination, it lacked advanced strategies for aligning modalities dynamically. The absence of adaptive cross-modal synchronization limited its performance on datasets with highly variable modalities.

Mai et al. [18]introduced a unique framework called HyCon for cross-modal mixed learning with contrast structures to carry out analysis of emotions in multiple modes. HyCon was able to explore inter-class interactions while simultaneously learning about intra-/inter-modal dynamics. The proposed HyCon particularly featured three different learning modalities such as semi-contrastive, intra-modal, and inter-modal. Then a model successfully learned latent embeddings for sentiment prediction using the specified functionality of loss including a modification word. A visualisation results demonstrate that the proposed HyCon approach can learn an enhanced selective embedding area for various emotional groups. Despite its strengths in capturing latent embeddings for sentiment prediction, its fixed contrastive approach hindered adaptability to real-time or evolving datasets. Moreover, the framework did not address the challenges of dynamic feature representations across modalities.

Chu et al. [19] proposed a deep fused audio-text bimodal transformer structured upon a single-modal pre-trained models RoBERTa and HuBERT. To facilitate cross-modal learning as effectively as possible, a cross-modal prior training process that is phased in and a well-thought-out cross-modal merging technique were designed. Using a stage-wise training approach, the authors proposed initial pretraining, task-adaptive pretraining, and finetuning downstream tasks. While this method achieved state-of-the-art performance on benchmark datasets called CMU-MOSEI sentiment and IEMOCAP emotion datasets, the reliance on pre-trained models increased the computational cost. Additionally, its phased pretraining failed to fully resolve modality gaps, that particularly for unstructured or asynchronous inputs.

Hang Li et al. [20] presented a Cross-modal Transformer for Audio-and-Language (CTAL), which integrates audio and language through two proxy tasks. During the fine-tuning stage, the authors developed a novel cross-modality fusion technique that forces the pre-trained model to acquire composite characteristics from diverse perspectives. In a general way dataset of audio and language pairs, it was trained with two pretraining tasks. A wide variety of empirical evidence reveals that their CTAL excelled at activities related to language interpretation, such as voice identification, sentiment evaluation, and attitude categorization. This method excelled at voice identification and sentiment evaluation but struggled with handling heterogeneous data sources due to its dependence on predefined proxy tasks. Its fusion mechanism lacked flexibility in aligning multi-modal data dynamically.

In a recent article Mehdi Arjmand et al. [21] introduced a Transformer-Based Speech-Prefixed Language Model called TEASEL to get around these restrictions without training a whole Transformer model. To implement an efficient spatial encoding, the authors implemented a Lightweight Attentive Aggregation module

that generated speech prefixes for a pre-trained Transformer. In comparison to a conventional language model, TEASEL included voice for a dynamic prefix for texts. an authors conclude that this approach employs a conventional previously trained language framework as the foundation for the cross-modal Converter system. An CMU-MOSI dataset was used by the researchers to assess TEASEL for the Analysis of emotion in many modes. Although this approach was innovative, it focused primarily on text and audio, neglecting other critical modalities such as visual data. Its application was further constrained by limited exploration of modality gaps and feature-level fusion.

In the context of misaligned multiple temporal sequences, the authors note that emotional cues are limited and inconsistent across modalities. Cross-modal dynamic convolution has been presented by Wen et al. [22] for multi-modal emotion identification tasks. It was created to address the absence of emotional cues and their discordance in multi-modal temporal sequences. Based on this convolution and MulT, the authors construct a multi-modal emotion identification pipeline to put their theories to the test. The adoption of more effective CNN-style layers. The ability was provided via dynamic convergence across modes. They offered creative solutions for dealing with non-aligned multimodal human language sequences. However, the ability to handle data from different sources over time was limited before the introduction of the learnable temporal offset in deformable convolution. This approach provided improved convergence across modes, yet its reliance on convolutional structures reduced its ability to capture long-range temporal dependencies, it limits the effectiveness for complex multimodal data.

Regarding multivariate evaluation of feelings, Lianyang Ma et al. [23] presented the multi-scale cooperative multimodal transformer (MCMulT) architecture. In general, the "multi-scale" method was effective in utilizing the various levels of semantic data present in each modality that is utilized for precise cross-modal interactions. As a result of incorporating cross-modal interactions with multiple level characteristics from its source modality, each modality learns its feature hierarchies. In such a fashion, each pair of modalities cooperatively and gradually constructed their feature hierarchy. To optimize efficiency and lower computing costs, the approach is still investigating MCMulT's finer design, such as section count adjustment for every modality separately. While this approach effectively utilized semantic data at different levels, the fixed section count for each modality resulted in inefficiencies and scalability issues for larger datasets or real-time applications.

Lian et al. [24] introduced a conversational transformer network (CTNet), a multiple learning architecture for converting emotion recognition. The author specifically recommended utilising a transformer-oriented framework used to simulate intra-modal and cross-modal exchanges between multimedia properties. We were able to extract the utterance's temporal information from them using segment-level acoustic characteristics and word-level lexical data as inputs. To simulate context- and speaker-sensitive dependencies, in addition, they proposed integrating multi-head attention with speaker embeddings and the bi-directional GRU component. The productivity of the recommended modes was determined in

a test finding on the IEMOCAP and MELD datasets. Nevertheless, the system's classification accuracy needed improvement. Its ability to model speaker-sensitive dependencies made it suitable for conversational data; however, it required more sophisticated optimization to improve classification accuracy and better handle asynchronous modality alignment.

An MSA strategy that increased decision-making capacities was presented by Zhao et al. [30] in the year 2023. This approach made use of a variety of sources which provided information about sentiment. It was demonstrated by the results of the experiments that the CRNN-SVM achieved the greatest average accuracy rate for sentiment, which was 91%, and it also provided excellent stability. Due to the limited number of participants in the experiment, it is conceivable that the findings did not adhere to the standards of scientific rigour.

Ghosh et al. [31] proposed a multi-sensory assessment (MSA) method that would use a five-module approach to evaluate the aggression of a person who is experiencing pain. Text-based sentiment analysis was the focus of the first three modules, which classified patient pain into three categories: overtly aggressive, non-aggressive, and covertly hostile. Image-based sentiment analysis was the primary emphasis of the next two modules, which utilised deep learning for feature extraction and classification procedures. The results of the experiments showed that this system outperformed the most advanced methods now available, with the combination of image- and text-based ratings contributing to an improvement in total performance. A level of accuracy of 82.35% was achieved by the model. Performance was impeded by the fact that the system was dependent on the quality of the data that was input.

Ma et al. [32] adopted a BCD-MM strategy in 2024, which improved generalisations to OOD scenarios by minimising dependence on non-causal correlations and boosting cross-modal feature extraction. This made it possible for the researchers to improve generalisations. In order to maximise the efficiency of feature extraction, this method makes use of a gated cross-modal attention mechanism. Additionally, it incorporates attention score-based strategies in order to minimise redundancy. The BCD-MM algorithm displayed strong performance in both sentiment analysis and debiasing when it was evaluated on OOD MOSEI, OOD MOSI, MOSEI, MOSI, and SIMS. An accuracy of 80.05 was achieved by the model when applied to OOD MOSI. On the other hand, the physiological signals need to be incorporated into the application of BCD-MM in order to ensure that sentiment analysis is both robust and comprehensive.

In order to improve multi-hop interactions and modal position perception, Fan and Chen [33] proposed PPMFN, which combined a position-perceptive module with an aspect-guided multi-hop interactive attention fusion module. A multi-hop interactive attention network was utilised by the aspect-guided fusion module in order to merge multi-modal qualities and minimise inappropriate noise. The position-perceptive module was responsible for capturing positional data from both text and images. Extensive trials proved that PPMFN regularly outperformed powerful baseline models on two public datasets. This was demonstrated by the fact that such models were used. For this model to be able to

handle complicated multimodal data more effectively, it was necessary to focus on refining the fusion techniques and the architecture of the model.

Both Lu and Fu [34] were the ones who initially presented the SentDep model. In order to merge auditory, textual, and visual data, it was accomplished by integrating cutting-edge fusion techniques with sophisticated deep learning models. This allowed for the creation of the system. In addition to that, it provided a uniform representation of the multimodal representation. The CLIP paradigm was adjusted in order to fulfil the task of processing textual and visual modalities. This was accomplished through the use of the CLIP paradigm. The newly introduced model was found to perform extraordinarily well in comparison to the most advanced techniques in MSA when it was tested on well-known datasets such as CMU-MOSEI and CMU-MOSI. This was observed during the testing process. According to extensive evaluation tests and additional trials, the amount of pre-training data, the efficiency of the fusion algorithms, and the incorporation of temporal features were shown to be key factors in SentDep's improved performance. Additionally, the incorporation of temporal features was shown to be instrumental in the improvement. The performance of SentDep was superior to that of the other approaches, as it had a mean absolute error of 0.520 and an accuracy of 55.0%. The computing expenses associated with this model are higher than those associated with other models, and it requires more time for training.

3. PROBLEM DEFINITION WITH CONTRIBUTION

Sentiment analysis (SA) is a field that is changing quickly. Its goal is to use computers to figure out what people think about things like products, services, books, movies, and more. SA has mostly used text-based data in the past, which makes it good for things like tracking brands and figuring out how happy customers are. But with the rise of user-generated videos, pictures, and other types of content on sites like YouTube, Instagram, and TikTok, Multimodal sentiment Analysis (MSA) has become a more complete method. MSA is different from standard SA since it uses more than one type of data, like audio, video, and text, to better represent how people feel. Even if MSA has a lot of potential, there are still several problems that need to be solved:

- Existing approaches sometimes have trouble combining data from several modalities in a useful way.
- A lot of models don’t show how different modalities work together.
- Also, tweaking hyperparameters by hand or in a static way makes deep learning systems less effective in complicated multimodal scenarios.

To overcome these limitations, we propose an innovative multimodal sentiment analysis approach that combines joint representation learning with an enhanced cross-modal transformer architecture. Our principal contributions are as follows:

- We derive features from audio, video, and text modalities and project them into a common dimensional space utilising a 1D temporal convolutional layer.

- A Joint Representation Translator(JRT) module is implemented to acquire bimodal joint characteristics that encapsulate cross-modal interactions.
- We employ an optimised Cross-Modal Transformer(CMT) to improve multimodal fusion, with hyperparameters refined by the Adaptive Dragonfly Optimization(ADO) algorithm.
- Ultimately, modality-enhanced features are concatenated and transmitted through a fully connected layer for sentiment prediction.

4. OVERVIEW OF PROPOSED METHODOLOGY

The proposed framework effectively performs sentiment analysis by integrating three modalities such as face, audio, and text. The structure of multimodal sentiment analysis is illustrated in Figure 1. The framework comprises three stages, including feature extraction, a joint representation translator, and a transformer-based prediction mechanism optimized using the Adaptive Dragon Optimization Algorithm (ADOA).

The framework processes three modalities, such as Video (V), audio (A), and text (T), for sentiment prediction The input data consists of a training dataset$D^{d} = \left\{Y_{\{A,V,T\}}^{[i]}, o^{[i]}\right\}_{i=1}^{M}$, with M data samples and $Y_{\{A,V,T\}}$ represents the input data from audio (A), video (V), and text (T) modalities and o denotes the sentiment label. The aim involves designing a mapping function $f: Y_{\{A,V,T\}}$to ô, which predicts sentiment intensity from the extracted features.

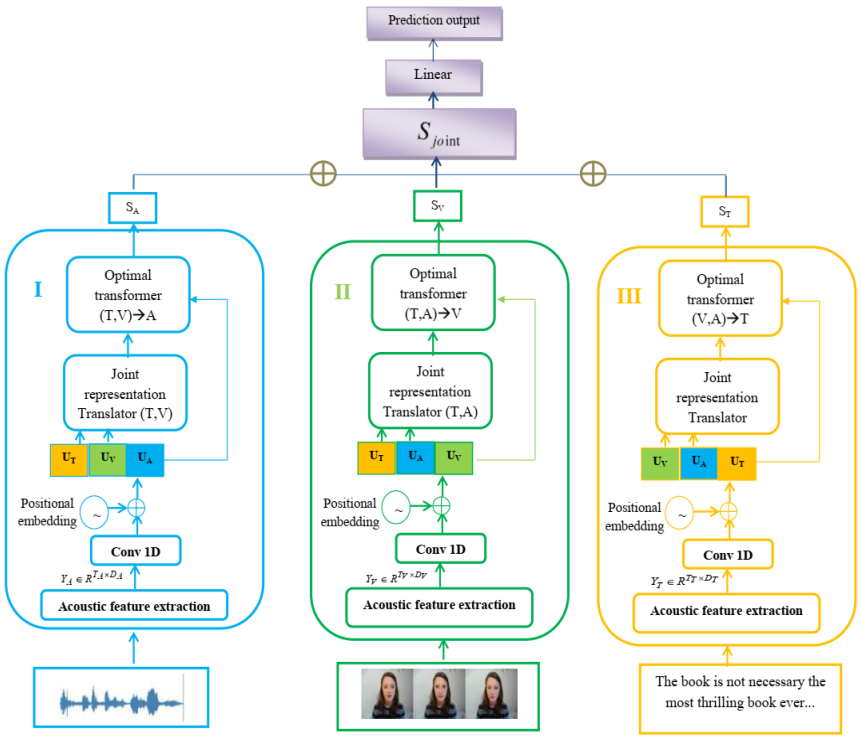


Fig. 1 Overall structure of multi-modal sentiment analysis

The methodology encompasses three stages such as feature extraction, a joint representation translator (JRT), and a transformer-based prediction mechanism optimized through the Adaptive Dragon Optimization Algorithm (ADOA). During feature extraction, modality-specific features are processed independently and aligned to a common feature space. The extracted modality feature sequences are represented as $Y_A \in \mathbb{R}^{L_A \times FD_A}, Y_V \in \mathbb{R}^{L_V \times FD_V}$ and $Y_T \in \mathbb{R}^{L_T \times FD_T}$, where $L(.)$ and $FD(.)$denote sequence length and feature dimension, respectively. Under a word-aligned setting, $L_A = L_V = L_T = L$.

To consolidate heterogeneous data, a 1D temporal convolutional layer is applied by ensuring uniform feature dimensionality across modalities. Positional embedding (PE) is used to encode temporal information into sequences by generating processed features denoted as U_A, U_V, and U_T. The processed representations pass through the Joint Representation Translator (JRT) to capture dependencies between modalities by using hierarchical fusion. The JRT aligns modalities by translating one modality into another in a cyclic manner to ensure consistency.

The outputs from the JRT are further processed by the transformer-based prediction mechanism, which integrates these features for final sentiment

prediction. The transformer is optimized using ADOA to enhance predictive accuracy and efficiency. Modality-reinforced representations S_A, S_V, and S_T are generated from the cross-modal transformer, which processes joint representations alongside individual modality sequences $U_{A,V,T}$. The outputs are concatenated into a joint vector S_{joint} for final sentiment label prediction through a fully connected layer.

4.1 FEATURE EXTRACTION

The feature extraction process independently processes audio, video, and text data to extract modality-specific features. Acoustic features are extracted using the Mel-frequency cepstral coefficients (MFCC) method, while video features are obtained through Facet, and text features are derived using the skip-gram model. The extracted modality-specific features are represented as $Y_A \in \mathbb{R}^{L_A \times FD_A}$, $Y_V \in \mathbb{R}^{L_V \times FD_V}$ and $Y_T \in \mathbb{R}^{L_T \times FD_T}$ respectively.

Audio feature extraction: For audio feature extraction, the Mel-frequency cepstral coefficients (MFCC) approach is employed. MFCC is selected due to its strong alignment with the auditory perception of humans. It extracts frequency components based on the Mel scale by capturing essential emotional signals such as pitch, tone, and intensity. This selection is crucial for identifying subtle emotional variations in audio, which enhances sentiment prediction accuracy. The MFCC method segments the speech signal into frames and computes cepstral features, which encapsulate emotional cues. The mathematical representation for MFCC features is given in equation (1).

$$MFCC = \log[\sum_{m=1}^{M} |Y(m)|^2 \times k(m)] \quad (1)$$

Where $Y(m)$ represents the Fourier-transformed signal, and $k(m)$ is the filter bank.

Video feature extraction: Video frames are pre-processed to eliminate noise, followed by feature extraction using Facet. Facet is effective in identifying micro-expressions and emotional cues from facial landmarks, including facial action units, gaze direction, and head movement. These features contribute significantly to the interpretation of non-verbal emotional signals. The video feature extraction method produces a 35-dimensional feature vector.

Text feature extraction: The skip-gram model is employed for text feature extraction. This model is proficient at capturing the semantic and syntactic relationships among words by learning the context of words within sentences. Skip-gram generates word vector representations that reflect both local and global linguistic patterns by enabling the capture of emotional nuances in text. The word vector y_j for a target word m_j is computed using equation (2-4).

$$w_l(m_j) = \phi\left(m^l w_l(m_{j-1}) + m^{sl} e(m_{j-1})\right) \quad (2)$$

$$w_r(m_j) = \phi\left(m^r w_r(m_{j+1}) + m^{sr} e(m_{j+1})\right) \quad (3)$$

$$y_j = [w_l(m_j); e(m_j); w_r(m_j)] \quad (4)$$

In this formulation, $w_l(m_j)$ and $w_r(m_j)$ represent the left and right context vectors, $e(m_j)$ is the embedding of m_j, and m^l, m^r, m^{sl}, m^{sr} are model parameters. ϕ represents a nonlinear activation function.
The integration of MFCC, Facet, and skip-gram ensures that emotional cues from all modalities are effectively represented. This comprehensive approach strengthens sentiment prediction capabilities across diverse data sources.

4.2 JOINT REPRESENTATION TRANSLATOR

The Joint Representation Translator (JRT) framework applies cyclic translation for pairing modalities, ensuring alignment and consistency across different input sources. The "joint representation" refers to the unification of features from multiple modalities to form a comprehensive representation. This method enables the model to utilize complementary information from diverse inputs, which results in improved prediction accuracy.

In the design of the JRT, a source modality is translated into a target modality. Following that, the target modality undergoes reverse translation into the source modality. The cyclic translation helps capture and retain common information shared across modalities, ensuring minimal discrepancies between the original and transformed representations. The architecture of the JRT framework is illustrated in Figure 2.

The necessity of the JRT arises from its capability to address alignment issues between modalities. This is achieved through the cyclic translation process, where forward translation maps the source modality to the target modality. The reverse translation reconstructs the original modality from the translated version. This bidirectional process allows the model to identify and preserve common latent features between paired modalities, thereby improving the accuracy of the shared representation.

Modality Reconstruction Loss (MRL) is incorporated to enhance performance. This loss function minimizes the difference between the original modality and the reconstructed modality. The loss function is computed using the mean squared error (MSE) to ensure the shared representation captures essential inter-modal features with minimal distortion. The result of modality translation is the refined shared representation, which improves tasks requiring multi-modal integration.

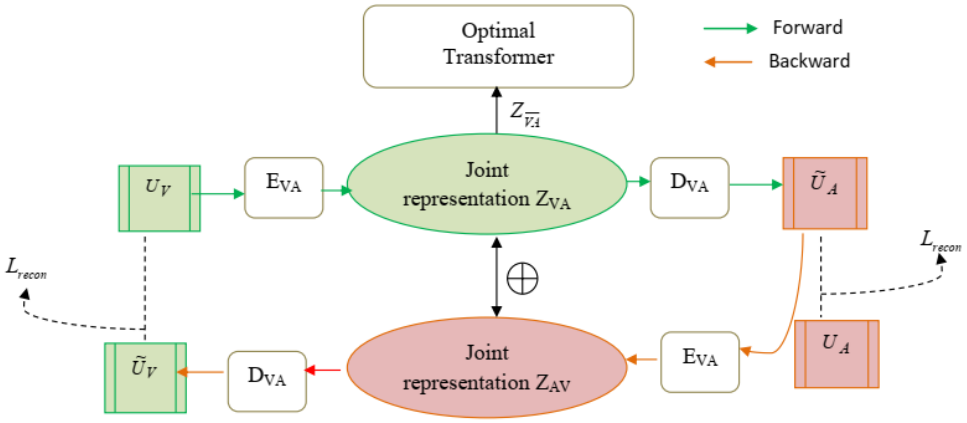


Fig. 6 The structure of JRT (V, A)

The JRT utilizes an encoder-decoder structure with an attention mechanism based on a seq2seq model. The forward translation process is defined by the following equation (5)

$$\tilde{\delta} = D(E(\eta)) \quad (5)$$

Here, the encoder E processes the input modality η and the decoder D transforms it into the target modality $\tilde{\delta}$. The reverse translation reconstructs the source modality η from the translated modality $\tilde{\delta}$ expressed using equation (6)

$$\tilde{\eta} = D(E(\tilde{\delta})) \quad (6)$$

The hidden representation Z_j is calculated using a GRU as $Z_j = GRU(Z_{j-1}, \delta_j), j = 1, \ldots T$. The decoder follows a probabilistic approach, where the output of the modality δ is computed in equation (7)

$$p(\delta) = \prod_{j=1}^{T} p(\delta_j | Z, \delta_1, \delta_2, \ldots., \delta_{j-1}) \quad (7)$$

To achieve optimal translation, the translated modality $\tilde{\delta}$ is defined by equation (8)

$$\tilde{\delta} = \arg \mathrm{Max}_{\delta}\, p(\delta\, |\eta) \quad (8)$$

For different modality pairs such as text-video, text-audio, and video-audio, the encoder-decoder processes are defined as follows:

For text-video modality pairs with U_T is the input text modality, E_{TV} is the encoder for text-video translation. A_{TV} is the forward representation of the text modality. D_{TV} is the decoder for text-video translation. $\tilde{U}_V$ is the translated video modality from equation (9-10)

$$A_{TV} = E_{TV}(U_T), \tilde{U}_V = D_{TV}(A_{TV}) \quad (9)$$

$$A_{VT} = E_{TV}(\tilde{U}_V), \tilde{U}_T = D_{TV}(A_{VT}) \quad (10)$$

For text-audio modality pairs using equation (11-12)

$$A_{TA} = E_{TA}(U_T),\ \widetilde{U}_A = D_{TA}(A_{TA}) \quad (11)$$

$$A_{AT} = E_{TA}(\widetilde{U}_A);\ \widetilde{U}_T = D_{TA}(A_{AT}) \quad (12)$$

For video-audio modality pairs using equation (13-14)

$$A_{VA} = E_{VA}(U_V);\ \widetilde{U}_A = D_{VA}(A_{VA}) \quad (13)$$

$$A_{AV} = E_{VA}(\widetilde{U}_A);\ \widetilde{U}_V = D_{VA}(A_{AV}) \quad (14)$$

The modality reconstruction loss (MRL) function is given by equation (15)

$$L_{\text{reconstruct}} = \sum_{m\in\{A,V,T\}} ||U_m - \widetilde{U_m}||_2^2 \quad (15)$$

Where $||.||_2^2$is the squared L_2-norm. This loss ensures that the reconstructed modality retains essential features by minimizing the semantic difference between the original and translated modalities. By concatenating modality representations such as A_{VA} and A_{AV}, these are projected into $\widetilde{A_{VA}}$ for capturing interactions between modalities. The final joint representations $\widetilde{A_{TV}}$, $\widetilde{A_{TA}}$ and $\widetilde{A_{VA}}$ are then integrated to enhance downstream task performance.

4.3 ADAPTIVE DRAGON OPTIMIZATION ALGORITHM (ADOA) OPTIMIZED TRANSFORMER

The Adaptive Dragon Optimization Algorithm (ADOA) optimized transformer leverages joint representations $\widetilde{A}_{\{TV,TA,VA\}}$ from bio-modalities to explore the complementarity between modalities. This method integrates common bio-modality information while retaining unimodal features, facilitating the transformation of combined representations into unimodal data. The ideal transformer, denoted as $B \in \mathbb{R}^{L_b \times FD_b}$ and $P \in \mathbb{R}^{L_p \times FD_p}$takes joint representations $\widetilde{A}_{\{TV,TA,VA\}}$ and processed unimodal sequences as inputs. Multi-Head Attention (MHA) with N_h Heads creates interactions across all modalities by enabling the model to focus on different parts of the input sequence simultaneously. The attention mechanism computes multiple attention distributions in parallel by enhancing the model's capability to capture complex relationships.

For the i^{th} head, the search term, price, and password are identified as $Q_i = PW_i^q, K_i = BW_i^k$and $V_i = BW_i^v$, respectively, where $W_i^q \in \mathbb{R}^{D_p \times D/N_h}$, $W_i^k \in \mathbb{R}^{D_b \times D/N_h}$ and, $W_i^v \in \mathbb{R}^{D_b \times D/N_h}$. D is the hidden dimension.

Each unimodal sequence $U_{\{A,V,T\}}$ updates as a query using the prior unimodal sequence, while joint representations B are transformed into key-value pairs for modality alignment. Message delivery (MD) transitions bimodal data into unimodal format through the equations (16-17)

$$\text{head}_i = \text{softmax}\left(\frac{PW_i^q\left(BW_i^k\right)^T}{\sqrt{D/N_h}}\right)BW_i^v \quad (16)$$

$$\text{MD}(B,P) = \text{Concat}(H_1,\ldots,H_{Nh})\,W_0 \quad (17)$$

Where,$W_0 \in \mathbb{R}^{D\times D}$,$H_i \in \mathbb{R}^{L_p\times D/Nh}$, $MD(B,P) \in \mathbb{R}^{L_p\times D}$.Concat represents the concatenation operation.

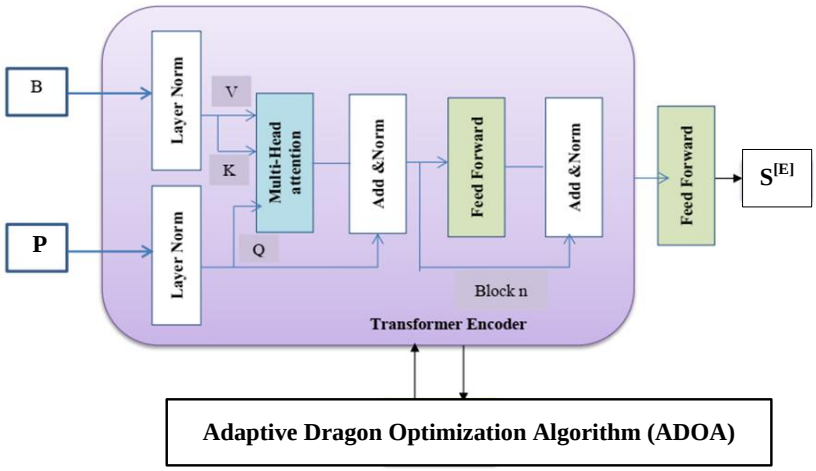


Figure 3. The structure of the Optimal transformer with Bi-modal

The structure of the optimal transformer with bio-modal data is shown in Figure 3. This transformer uses residual connections, position-wise feed-forward networks (FFN), and layer normalization (LN) for gradual transformation. Residual connections involve the addition of layer inputs to their outputs by addressing the vanishing gradient problem and facilitating the training of deep neural networks. This technique ensures the preservation of information flow across layers, which enhances convergence speed and improves predictive performance. Position-wise FFNs apply fully connected networks with non-linear activation functions independently at each position in the sequence by enabling the model to capture complex transformations and improve its expressiveness and adaptability. Layer normalization standardizes layer outputs by normalizing summed inputs by reducing internal covariate shifts and stabilizing the training process. After E stacked layers, the transformer's output $S^{[e]}$ is computed using the following equation (18-20)

$$S^{[0]} = P \tag{18}$$

$$\hat{S}^{[e]} = MD\left(LN(B), LN\left(S^{[e-1]}\right)\right) + LN\left(S^{[e-1]}\right) \tag{19}$$

$$S^{[e]} = FFN\left(LN\left(\hat{S}^{[e]}\right)\right) + LN\left(\hat{S}^{e}\right), e = 1, \ldots, E \tag{20}$$

where E is the number of layers of the Adaptive Dragon Optimization Algorithm (ADOA) Optimized transformer (OT).

The modality-reinforced representationsS_A, $S_V$and S_T combine data from specific modalities and shared modalities. These representations are calculated using equations (21-23)

$$S_A = OT(\,JRT(U_T, U_V), U_A)$$ (21)

$$S_V = OT(\,JRT(U_T, U_A), U_V)$$ (22)

$$S_T = OT(\,JRT(U_V, U_A), U_T)$$ (23)

The final joint vector S_{joint} is expressed using equation (24)

$$S_{joint} = S_A \oplus S_V \oplus S_T$$ (24)

Finally, using a joint vector S_{joint} as input, the sentiment score is estimated through a fully integrated network f using equation (25)

$$\hat{Y} = f(S_{joint})$$ (25)

4.3.1 PARAMETER OPTIMIZATION USING ADO ALGORITHM

The optimal transformer architecture is enhanced through the Adaptive Dragonfly Optimization (ADO) algorithm to fine-tune hyperparameters specific to multimodal sentiment analysis. The transformer's hyperparameters, including the number of encoder layers, attention head count, feed-forward layer dimension, dropout rate, and embedding size, are dynamically adjusted by the ADO algorithm to maximize performance.

During the optimization process, the dragonfly algorithm mimics the swarming behaviors of dragonflies, incorporating both static and dynamic swarm movements. The algorithm adjusts the position of each dragonfly by considering factors such as separation, alignment, cohesion, attraction to food sources, and avoidance of adversaries. These factors guide the algorithm in exploring and exploiting the search space for optimal hyperparameters.

Table 1: Hyperparameter range

Hyper-parameters	**Range**
Number of layers in encoder	[1-7]
Count of attention heads	1, 2, 4, 8, 16
Dimension of the feed-forward layer	128, 256, 512, 1024, 2048, 4096
Dropout	[0.1-0.5]
Embedding dimensions	256, 512, 1024

In the absence of neighboring solutions, a Levy flight operation is employed to introduce random walk behavior, allowing for further exploration of the search space. The location of each dragonfly is updated using the Levy flight step length, which facilitates exploration beyond local optima. The dragonfly's position is iteratively adjusted until a termination criterion is met, typically when the number of iterations reaches a predefined maximum or the best solution is found.

The steps involved in the adaptive dragonfly optimization algorithm are explained as follows:

Step 1: Initialization: The primary objective of this technique is to select the optimal hyperparameter. In this case, the dragonfly represents the hyperparameter level. Initially, the dragonfly is selected at random. The solutions are represented as follows using equation (26)

$$P_N = \{Y_1, Y_2, \dots, Y_N\} \quad (26)$$

Here, Y_N is the N^{th} solution or cat's position, and it is defined using (27).

$$Y_i = \{L, H, F, O, E\}_i \quad (27)$$

while L is the encoder's total number of levels, H is the total number of attention heads, F is the feed-back layer's scale, O is the dropout rate, and E is the embedded size.

Step 2: Fitness calculation: The fitness of each solution is determined after solution initialization. The fitness function is considered to have the highest accuracy. Equation (28) contains the fitness function.

$$\text{Fit=Max (Accuracy)} \quad (28)$$

In this equation,

$$\text{Accuracy} = \frac{TN+TP}{TP+FP+TN+FN} \quad (29)$$

Where TP depicts the True Positives; TN represents the True Negatives; FP depicts False Positives and FN depicts the False Negatives.

Step 3: Update the solution using ADO: In this case, the algorithm uses various principles to update the dragonfly's position. These principles are:

Separation: This avoids a collision between a swarm member and its neighbours. The separation is calculated using equation (30).

$$S_d = -\sum_{j=1}^{N}(Z - Z_j) \quad (30)$$

Alignment: This ensures that the direction of the swarm members is in alignment, calculated using equation (31).

$$A_d = \frac{\sum_{j=1}^{N} W_j}{N} \quad (31)$$

Cohesion: This ensures that the members of the swarm are drawn toward the centre of mass, calculated using equation (32)

$$C_d = \frac{1}{N} \times \sum_{j=1}^{N}(Z_j - Z) \quad (32)$$

The formula for calculating attraction to a food source is as follows:

$$F_d = Z^+ - Z \quad (33)$$

The following formula is used to calculate adversary distraction:

$$E_d = Z^- + Z \quad (34)$$

Where Z denotes the immediate location of the person, Z_j denotes the immediate location of the j^{th} individual, the number of neighbouring solutions is represented as N, W_j denotes the speed of the j^{th} neighbouring individual, $Z^+$and Z^-represent the locations of the food source and adversary, respectively.

A pair of vectors is taken into consideration, phase (ΔZ) and location (Z), which are used to duplicate the movements of synthetic dragonflies and modify their locations in search. Equation (35) indicates that a step vector, that also referred to as the quick, indicates the path of the dragonfly movements. The vector of position is modified by equation (36), after the initial matrix has been successfully computed.

$$\Delta Z_{t+1} = (sS_d + aA_d + cC_d + fF_d + eE_d) + \underline{\omega}\Delta Z_t \quad (35)$$

$$Z_{t+1} = Z_t + \Delta Z_{t+1} \quad (36)$$

where s, a, and c in equation (35) are constants related to separation, alignment, and cohesion, whereas f, e, w, and t, represent the food, adversary, and inertial factors, respectively. ω is the inertial weight, t represents the iteration.

Step 4: Levy flight operation: When there are no neighbouring solutions, a random walk (Levy Flight) is used to introduce randomness and promote exploration. The updated location is calculated as:

$$Z_{t+1} = Z_t + \text{Levy(i)} \times Z_t \quad (37)$$

Here, I represent the position vector's size. The following random walk distributions have been generated by the Levy Flight.

$$\text{levy(z)} = 0.01 \times \frac{r_1 \times \sigma}{|r_2|^{1/\beta}} \quad (38)$$

r_1 and r_2 denote a random number ranging within [0,1]. βis a constant value. σ represents step length scaling.

Levy flight step lengths s_l are as follows:

$$s_l = \left|\frac{v}{\mu}\right|^{\frac{1}{\beta}} \quad (39)$$

Where μ is a uniformly distributed random value,

$$\sigma = \left[\frac{1+x\times\sin(x/2)}{(((1+x/2))\times x2^{(-1)/2}}\right] \text{ and } \sigma_v = 1 \quad (40)$$

$$\tau(z)=(z-1)! \quad (41)$$

Step 5: Termination criteria

Steps 2 to 4 are repeated until the termination condition is met. The termination criterion is when the number of iterations t exceeds the maximum iterations, t_{max}, or the best solution is found.

5 RESULTS AND DISCUSSION

In this section, the performance of the proposed framework is assessed to determine its effectiveness in conducting efficient multimodal sentiment analysis. The framework consists of three primary components, including feature extraction, a joint representation translator, and a transformer-based prediction mechanism optimized by the Adaptive Dragon Optimization Algorithm (ADOA). For modality feature extraction, COVAREP [25] processes the acoustic signals, which yields a feature dimension of 74 for the acoustic modality. The video frames are processed using Facet [26] which produces 35-dimensional feature vectors. For the textual modality, Glove [27] is applied to generate 300-dimensional word embeddings. To ensure fairness in comparison, the experiments are repeated five times, and the average results are reported for both the proposed JRT and Optimized Transformer method and a selection of baseline models. All experiments are conducted on a single RTX-2080Ti GPU, and Python is the tool used for implementing and executing the model.

5.1 DATASET DESCRIPTION

For testing the proposed model, two public multimodal sentiment analysis datasets, CMU-MOSI and CMU-MOSEI, are adopted. A brief overview of the datasets are as follows,

CMU-MOSI: The CMU-MOSI dataset [28] serves as a popular benchmark for multimodal sentiment analysis. It contains 2,199 samples of short monologue video clips extracted from 93 YouTube movie review videos. Each sample is labeled with a continuous sentiment score ranging from [-3, 3], where -3 represents extremely negative emotions and 3 signifies extremely positive emotions. The dataset is divided into 1,284 training samples, 229 validation samples, and 686 test samples.

CMU-MOSEI: The CMU-MOSEI dataset [29] improves upon CMU-MOSI by incorporating a greater diversity of samples, speakers, and topics. It includes 23,453 annotated video clips from 5,000 videos, 1,000 different speakers, and 250 topics. Each sample is labeled with a sentiment score within the range of [-3, +3]. The dataset consists of 16,326 training samples, 1,871 validation samples, and 4,659 test samples.

Table 2: Performance metrics analysis for various input videos

Input video	Accuracy	Recall	F-Score	Correlation coefficient	False positive rate	Negative predictive value
	95.42	97.53	95.39	0.9094	0.0657	0.9757
	96.49	98.64	96.46	0.9308	0.0553	0.9867
	95.93	98.42	95.91	0.9200	0.0640	0.9845

Table 2 provides the analysis of performance metrics for different input videos. According to Table 1, the highest accuracy is achieved with 95.42% for Video 1, 96.49% for Video 2, 95.93% for Video 3, 94.55% for Video 4, and 93.20% for Video 5. The minimal error values are due to the effectiveness of the JRT and Optimized Transformer method.



Figure 4. Computation analysis based on execution time

The Execution Time vs. Dataset Size graph for JRT-ADOA illustrates the computational cost trend as the dataset size increases. As expected, execution time

grows with larger datasets, indicating the model's scalability. Initially, the execution time increases gradually, but as the dataset size reaches higher thresholds, the rise becomes more pronounced due to increased computational complexity in processing multimodal sentiment analysis. Despite this, JRT-ADOA maintains a reasonable execution time, suggesting its efficiency compared to traditional methods.

5.2 COMPARATIVE ANALYSIS

Table 2 presents the comparative analysis results for various methods, including LSTM, RNN, CNN, Bi-GRU, the method without optimization, and the proposed JRT and Optimized Transformer method. The evaluation includes metrics such as accuracy, recall, F1-score, correlation, false positive rate, negative predictive value, and mean squared error.

Table 3. Comparative analysis results

Measures	**LSTM**	**RNN**	**CNN**	**Bi-GRU**	**Without optimization**	**Proposed**
Accuracy (%)	87.06	75.20	87.74	81.23	83.33	95.11
Recall (%)	82.8	62.65	91.55	68.75	75.78	97.51
F1-Score (%)	87.15	75.40	87.55	81.11	83.01	95.14
Correlation coefficient	0.74	0.57	0.758	0.678	0.681	0.90
False Positive rate (%)	0.0813	0.0943	0.0756	0.089	0.078	0.051
Negative Predictive Value (%)	0.7256	0.6214	0.718	0.69	0.766	0.97
MSE	0.1294	0.248	0.1226	0.1877	0.1667	0.0489

Table 3 displays the comparative analysis results. The proposed JRT and Optimized Transformer method shows substantial improvement across various performance metrics compared to existing methods, including LSTM-based, RNN-based, CNN-based, Bi-GRU-based sentiment analysis, and the baseline JRT method without optimization. For accuracy, improvements of 9.25%, 26.48%, 8.40%, 17.09%, and 14.14% are achieved, thereby demonstrating the method's strong capability in accurately classifying instances across multiple sentiment

categories. This increase in accuracy highlights the model's effectiveness in minimizing classification errors across diverse sentiment analysis tasks. The recall metric experiences a significant enhancement, with increases of 17.77%, 55.64%, 6.51%, 41.83%, and 28.68%. These improvements illustrate the model's superior ability to identify true positive cases while minimizing false negatives. This is especially important in tasks where detecting positive instances is critical, such as in sentiment classification, where the improved recall ensures the model reliably identifies all relevant positive sentiment cases.

A notable F1-score improvement of 9.17%, 26.18%, 8.67%, 17.30%, and 14.61% further emphasizes the balanced performance between precision and recall in sentiment classification. A high F1-score demonstrates that the model effectively maintains equilibrium between minimizing both false positives and false negatives, which is crucial for applications where errors in sentiment classification significantly impact the analysis. The correlation coefficient shows increases of 21.62%, 57.89%, 18.73%, 32.74%, and 32.16%, thereby indicating a stronger alignment between predicted and actual sentiment values. This improvement is vital for sentiment analysis models where understanding the correlation between predictions and real-world outcomes is necessary, such as in sentiment classification tasks. The False Positive Rate (FPR) is reduced by 37.27%, 45.92%, 32.54%, 42.70%, and 34.62%, thereby signifying fewer incorrect positive predictions in sentiment classification. A lower false positive rate is crucial for sentiment analysis applications where avoiding misclassifications is essential, thus minimizing unnecessary reactions or decisions based on incorrect sentiment predictions.

Negative Predictive Value (NPV) shows an increase of 33.68%, 56.10%, 35.10%, 40.58%, and 26.63%, thereby highlighting the model's enhanced reliability in predicting negative sentiment cases. With a higher NPV, the accuracy of the model's negative sentiment predictions is improved, which enhances the trustworthiness of sentiment analysis results. Finally, the Mean Squared Error (MSE) shows a significant decrease of 62.21%, 80.28%, 60.11%, 73.95%, and 70.67% which reflects the model’s enhanced ability to make precise predictions with minimal deviation from actual values. This reduction in MSE highlights the model's improved accuracy and reliability in real-world applications where precise prediction is essential.

Table 4: Comparative analysis based on different optimization algorithms

Measures	**JRT+** grid search	**JRT+** Bayesian optimization	**JRT+** ADOA (**Proposed**)
Accuracy (%)	90.14	88.20	95.11
Recall (%)	91.45	86.65	97.51
F1-Score (%)	82.15	87.40	95.14
Correlation coefficient	0.82	0.58	0.90
False Positive rate (%)	0.0813	0.0943	0.051
Negative Predictive Value (%)	0.7256	0.6214	0.97
MSE	0.1294	0.248	0.0489

Table 4 presents a comparative analysis of different optimization algorithms, JRT+ Grid Search, JRT+ Bayesian Optimization, and the proposed JRT+ ADOA, based on key performance metrics. The proposed JRT+ ADOA outperforms the other methods across all measures, achieving the highest accuracy of 95.11%, recall of 97.51%, and F1-score of 95.14%, indicating its superior classification capability. Additionally, it demonstrates a higher correlation coefficient (0.90), suggesting a stronger relationship between predicted and actual values. The false positive rate (0.051%) is the lowest, reflecting its robustness in minimizing incorrect positive classifications. Furthermore, the negative predictive value of 0.97 is significantly higher than the other methods, showing improved reliability in identifying true negatives. Moreover, the MSE for JRT+ ADOA (0.0489) is the lowest, indicating minimal prediction errors compared to JRT+ Grid Search of 0.1294 and JRT+ Bayesian Optimization of 0.248. Overall, the results validate that the proposed ADOA-based approach significantly enhances classification performance, reducing error rates and improving predictive reliability.

5.3 COMPARISON WITH PUBLISHED WORK

We contrast our findings with previously published studies to demonstrate the effectiveness of the proposed method. In this, we compare our work with [17-24]. These works are mainly used for sentiment analysis, and they have been published recently. The detailed description of each work is already explained in section 2.

Table 5: Comparative analysis results

Modal	Accuracy	F1 score	Mean absolute error (MAE)	Correlation
Multimodal [17]	85.7	85.6	0.714	0.794
Multimodal [18]	85.2	85.1	0.713	0.790
Cross-modal model [19]	88.5	88.5	0.500	-
Cross-modal [20]	80.77	81.01	0.6027	0.6809
Cross-modal attention module [21]	89.33	89.31	0.644	0.842
Multi-modal [22]	80.5	80.2	0.608	0.667
Multimodal [23]	83.1	82.8	0.582	0.706
Multimodal [24]	83.6	83.8	-	-
Multimodal [proposed]	95.11	95.14	0.0171	0.90

Table 5 presents comparative analysis results. Here, we compared our work with already published research works. These works are mainly focused on multimodal-based sentiment analysis. When analyzing Table 3, the suggested work reached the maximum accuracy of 95.11% which is 83.6% for [24], 83.1% for [23], 80.5%% for [22], 89.33% for [21], 80.77% for [20], 88.5% for [19], 85.2% for [18] and 85.7% for [17]. Compared to existing work, our proposed method attained the maximum accuracy. This is due to the parameter optimization in the transformer encoder and JRT. Similarly, our work attained the minimum MSE of 0.0171. The results show that the recommended technique produced a better result than the other techniques.

5.4 ABLATION STUDY

An ablation study is a systematic analysis used to evaluate the impact of individual components or hyperparameters on a model's performance. By

selectively modifying specific parameters, such as dropout rate or embedding dimension, researchers can determine their optimal values for improving accuracy and generalization. The results provide insights into the trade-offs between different settings, guiding the selection of the best-performing model configuration.

Table 6: Performance analysis by varying dropout value

	Dropout				
Measures	0.1	0.2	**0.3**	0.4	0.5
Accuracy	91.5	93.32	**95.11**	92.45	91.63
Recall	93.4	95.34	**97.51**	94.54	93.32
F1-Score	91.32	92.71	**95.14**	93.22	92.63
MSE	0.0534	0.0342	**0.0171**	0.0234	0.0434

Table 6 presents an ablation study analyzing the impact of varying dropout values (0.1 to 0.5) on model performance using accuracy, recall, F1-score, and MSE as evaluation metrics. The results indicate that a dropout value of 0.3 yields the highest accuracy of 95.11%, recall of 97.51%, and F1-score of 95.14%, while also achieving the lowest MSE of 0.0171, suggesting an optimal balance between regularization and performance. Lower dropout values 0.1 and 0.2 lead to relatively lower accuracy and recall, implying potential overfitting, while higher dropout values 0.4 and 0.5 result in performance degradation, likely due to excessive neuron deactivation. Thus, a dropout value of 0.3 appears to be the most effective for enhancing model generalization while minimizing error.

Table 7: Performance analysis by varying embedding dimension

Measures	**Embedding dimension**		
	256	512	**1024**
Accuracy	91.8	**95.11**	**92.52**
Recall	93.7	**97.51**	**94.81**
F1-Score	91.50	**95.14**	**93.8**
MSE	0.0531	**0.0171**	**0.0421**

Table 7 presents an ablation study examining the effect of varying embedding dimensions (256, 512, and 1024) on model performance based on accuracy, recall, F1-score, and MSE. The results indicate that an embedding dimension of 512 achieves the highest accuracy of 95.11%, recall of 97.51%, and F1-score of 95.14%, while also obtaining the lowest MSE of 0.0171, suggesting optimal feature representation and model generalization. In contrast, lower 256 and higher 1024 embedding dimensions result in reduced performance, with increased MSE values of 0.0531 and 0.0421, respectively, possibly due to underfitting in the

former and overfitting in the latter. These findings highlight that an embedding dimension of 512 is the most effective in capturing meaningful features while maintaining model efficiency.

6. CONCLUSION

The proposed multimodal sentiment analysis framework, integrating advanced feature extraction, a Joint Representation Translator (JRT), and a transformer model optimized through the Adaptive Dragon Optimization Algorithm (ADOA) achieves significant advancements in sentiment prediction accuracy. The process begins with the extraction of video, audio, and text features to ensure a comprehensive representation of the multimodal input. This joint feature set is then fed into the JRT, which effectively translates and merges these features into a unified representation. The transformer-based architecture, further optimized by ADOA ensures that this representation is processed in the most efficient manner, leveraging attention mechanisms to enhance model performance. The ADOA optimization significantly boosts the model's predictive capabilities by refining the transformer's parameters, improving its ability to identify intricate patterns in sentiment-related data. The significant improvement in these metrics highlights the proposed method's superior ability to capture complex sentiment patterns by comprehensively integrating both textual and non-textual modalities for a more robust analysis. The comprehensive evaluation across multiple benchmarks further validates the approach's efficiency and effectiveness in real-world sentiment detection tasks. It is appropriate for a wide variety of global applications because of its generalisability and solid performance, which includes healthcare, education, entertainment, and social media analytics, amongst others.

Future work would focus on real-time data stream analysis and integration of edge computing for faster sentiment detection in dynamic environments. Expanding the framework to include additional multimodal data, such as wearable sensors or physiological signals, would enhance emotion recognition. Further advancements in model interpretability and domain-specific adaptations could improve performance and applicability across various industries.

ACKNOWLEDGMENT

The authors would like to thank the management of Vellore Institute of Technology, Vellore, India Author

CONTRIBUTIONS

Conceptualization, methodology, analysis, and paper writing by P.Vasanthi . Supervision, reviewing, and editing by V. Madhu Viswanatham. Both authors contributed to the article and approved the submitted version.

AVAILABILITY OF DATA

The data that support the findings of this study are openly available in kaggle at https://www.kaggle.com/datasets/mathurinache/cmu-mosi.

ETHICS APPROVAL
Not applicable.

CONFLICTS OF INTEREST
The author(s) declare(s) no conflicts of interest to report regarding the present study.

REFERENCES

1. Kaur, R., & Kautish, S. 2022. Multimodal sentiment analysis: A survey and comparison. Research anthology on implementing sentiment analysis across multiple disciplines:1846-1870.
2. Zhu, T., Li, L., Yang, J., Zhao, S., Liu, H., & Qian, J. 2022. Multimodal sentiment analysis with image-text interaction network. IEEE transactions on multimedia, 25: 3375-3385.
3. Vasanthi, P., & Viswanatham, V. M. (2024). Federated Learning for Multimodal Sentiment Analysis: Advancing Global Models with an Enhanced LinkNet Architecture. IEEE Access.
4. Mahammad, F. S., & Viswanatham, V. M. 2018. Performance analysis of data compression algorithms for heterogeneous architecture through parallel approach. Journal of Supercomputing, 1-14. DOI: 10.1007/s11227-018-2478-3.
5. Gilleade, K. M., Dix, A., & Allanson, J. 2005. Affective videogames and modes of affective gaming: assist me, challenge me, emote me. In Proceedings of DiGRA 2005 Conference: Changing Views: Worlds in Play.
6. Chandran, J., & Viswanatham, V. M. 2022. Dynamic node influence tracking based influence maximization on dynamic social networks. Microprocessors and Microsystems, 95, Article 104689. DOI: 10.1016/j.micpro.2022.104689.
7. P, Vasanthi & V, Madhu. 2024. Sentiment analysis of multi social media using machine and deep learning models: a review. Multimedia Tools and Applications. 1-19. 10.1007/s11042-024-18981-8.
8. P, Vasanthi & V, Madhu. 2024. Sentiment analysis of multi social media using machine and deep learning models: a review. Multimedia Tools and Applications. 1-19. 10.1007/s11042-024-18981-8.
9. Hochreiter, S., & Schmidhuber, J. 1997 . Long short-term memory. Neural computation: 9(8), 1735-1780.
10. Mai, S., Hu, H., Xu, J., & Xing, S. 2020. Multi-fusion residual memory network for multimodal human sentiment comprehension. IEEE Transactions on Affective Computing: 13(1), 320-334.
11. Tseng, H. Y., Lee, H. Y., Huang, J. B., & Yang, M. H. 2020. Cross-domain few-shot classification via learned feature-wise transformation. arXiv preprint arXiv:2001.08735.
12. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. 2017. Attention is all you need. Advances in neural information processing systems: 30.

13. Rahman, W., Hasan, M. K., Lee, S., Zadeh, A., Mao, C., Morency, L. P., & Hoque, E. 2020. Integrating multimodal information in large pretrained transformers. In Proceedings of the conference. Association for Computational Linguistics. Meeting (Vol. 2020, p. 2359). NIH Public Access.
14. Yang, K., Xu, H., & Gao, K. 2020. Cm-bert: Cross-modal bert for text-audio sentiment analysis. In Proceedings of the 28th ACM international conference on multimedia (pp. 521-528).
15. Tsai, Y. H. H., Bai, S., Liang, P. P., Kolter, J. Z., Morency, L. P., & Salakhutdinov, R. 2019. Multimodal transformer for unaligned multimodal language sequences. In Proceedings of the conference. Association for computational linguistics. Meeting (Vol. 2019, p. 6558). NIH Public Access.
16. Mai, S., Hu, H., Xu, J., & Xing, S. 2020. Multi-fusion residual memory network for multimodal human sentiment comprehension. IEEE Transactions on Affective Computing: 13(1), 320-334.
17. Zeng, Y., Mai, S., & Hu, H. 2021. Which is making the contribution: Modulating unimodal and cross-modal dynamics for multimodal sentiment analysis. arXiv preprint arXiv:2111.08451.
18. 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.
19. Chu, I. H., Chen, Z., Yu, X., Han, M., Xiao, J., & Chang, P. 2022. Self-supervised cross-modal pretraining for speech emotion recognition and sentiment analysis. In Findings of the Association for Computational Linguistics: EMNLP 2022 (pp. 5105-5114).
20. Li, H., Kang, Y., Liu, T., Ding, W., & Liu, Z. 2021. CTAL: Pre-training cross-modal transformer for audio-and-language representations. arXiv preprint arXiv:2109.00181.
21. Arjmand, M., Dousti, M. J., & Moradi, H. 2021. Teasel: a transformer-based speech-prefixed language model. arXiv preprint arXiv:2109.05522.
22. Wen, H., You, S., & Fu, Y. 2021. Cross-modal dynamic convolution for multi-modal emotion recognition. Journal of Visual Communication and Image Representation, 78, 103178.
23. Ma, L., Yao, Y., Liang, T., & Liu, T. 2022. Multi-scale cooperative multimodal transformers for multimodal sentiment analysis in videos. arXiv preprint arXiv:2206.07981.
24. Lian, Z., Liu, B., & Tao, J. 2021. CTNet: Conversational transformer network for emotion recognition. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 29, 985-1000.
25. Degottex, G., Kane, J., Drugman, T., Raitio, T. and Scherer, S., 2014, May. COVAREP—A collaborative voice analysis repository for speech technologies. In 2014 ieee international conference on acoustics, speech and signal processing (icassp) (pp. 960-964). IEEE.
26. Baltrušaitis, T., Robinson, P. and Morency, L.P., 2016, March. Openface: an open source facial behavior analysis toolkit. In 2016 IEEE winter conference on applications of computer vision (WACV) (pp. 1-10). IEEE.

27. Pennington, J., Socher, R. and Manning, C.D., 2014, October. Glove: Global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP) (pp. 1532-1543).
28. Zadeh, A., Zellers, R., Pincus, E. and Morency, L.P., 2016. Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. IEEE Intelligent Systems, 31(6), pp.82-88.
29. Zadeh, A.B., Liang, P.P., Poria, S., Cambria, E. and Morency, L.P., 2018, July. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 2236-2246).
30. Zhao, Y., Mamat, M., Aysa, A., & Ubul, K. (2023). Multimodal sentiment system and method based on CRNN-SVM. *Neural Computing and Applications*, *35*(35), 24713-24725.
31. Ghosh, A., Dhara, B. C., Pero, C., & Umer, S. (2023). A multimodal sentiment analysis system for recognizing person aggressiveness in pain based on textual and visual information. *Journal of Ambient Intelligence and Humanized Computing*, *14*(4), 4489-4501.
32. Ma, L., Li, J., Shao, D., Yan, J., Wang, J., & Yan, Y. (2024). BCD-MM: Multimodal sentiment analysis model with dual-bias-aware feature learning and attention mechanisms. *IEEE Access*, *12*, 74888-74902.
33. Fan, H., & Chen, J. (2024). Position perceptive multi-hop fusion network for multimodal aspect-based sentiment analysis. *IEEE Access*, *12*, 90586-90595.
34. Al-Tameemi, I. K. S., Feizi-Derakhshi, M. R., Pashazadeh, S., & Asadpour, M. (2023). Interpretable multimodal sentiment classification using deep multi-view attentive network of image and text data. *IEEE Access*, *11*, 91060-91081.