A New Advanced Multimodal Sentiment Classification Through Combined Attention Mechanisms: A-MSDA

Soukaina FATIMI, WAFAE SABBAR, Abdelkrim BEKKHOUCHA
LIM Laboratory-Faculty of Sciences and Techniques, University Hassan II of Casablanca, Mohammedia, Morocco

***Abstract*—Interest in multimodal sentiment analysis has grown significantly due to the widespread sharing of text and images on social platforms. Existing approaches often emphasize either sentiment features within textual–visual data or the correlation between modalities, leaving gaps in effectively capturing both aspects simultaneously. To address these limitations, we propose the Advanced Multimodal Sentiment Analysis with Dual Attention (A-MSDA) model, which integrates self-attention and cross-modal attention mechanisms in a unified dual-attention framework. This design enables robust multimodal fusion by extracting salient textual and visual features while modeling their image–text interaction comprehensively. Experimental evaluation on MVSA-Single and MVSA-Multiple datasets demonstrates that A-MSDA achieves notable improvements in accuracy and F1-score, outperforming existing techniques by up to 3.4% in F1-score on MVSA-Multiple, while maintaining competitive performance on MVSA-Single. These results highlight the potential of A-MSDA to advance research in deep multimodality and sentiment analysis.**

***Keywords*—Dual attention; self-attention; cross-modal attention; multimodal sentiment analysis; multimodal fusion; MVSA dataset; image–text interaction; deep multimodality**

I. Introduction

Multimodal analysis is an emerging and indispensable research field with the rise of the digital world. Nowadays, data is not only found in massive amounts, but the diversity of the content presents new analytical challenges. This emerging field has been empowered by the rapid evolution of digital platforms and the rise in the sharing of visual content and textual data. Combining textual data with other social, visual, or audio cues aims to improve knowledge extraction and reduce the uncertainties associated with single-modality analysis. Multimodal sentiment analysis leverages feature extraction and representation, combined with multimodal fusion models, to provide a more comprehensive understanding of sentiment [1], [2].

Texts and images complement each other and work jointly to convey nuanced emotions; in this case, the expressed feelings will not be limited to one type of data. This capability is critical for real-world applications such as social media monitoring, brand perception analysis, and marketing strategies, where user-generated content often combines text and visuals to express opinions about products, services, and events [3], [4].

Recent advances in attention mechanisms have significantly improved multimodal modeling by enabling models to focus on contextually relevant information [2]. However, most existing approaches employ either self-attention or cross-modal attention in isolation, which limits their ability to fully capture intra-modal and inter-modal relationships. This gap motivates the need for a unified framework that combines both attention types to enhance sentiment prediction [5]. The objectives of this work are:

- To design a dual-attention architecture that integrates self-attention for modality-specific feature extraction and cross-modal attention for interaction modeling.
- To evaluate the proposed model on benchmark datasets and demonstrate its effectiveness in improving accuracy and robustness.

This study is organized as follows: Section II reviews related work and highlights gaps in existing approaches. Section III details the proposed A-MSDA architecture, and Section IV presents the experimental setup, results, and discussion. Finally, Section V underlines limitations before concluding with future directions in Section VI.

II. Literature Review

Multimodal sentiment analysis has become an active field of study. It allows understanding emotional cues in heterogenous data from different types of sources, and using a variety of modalities, including text, images, audio, and video, to increase the efficiency of sentiment prediction over unimodal methods. [1], [6]. A multitude of approaches have been proposed to enhance this research field. Conventional approaches first concentrated on decision-level fusion, in which the categorization results were integrated after each modality was processed separately [7]. Subsequently, concatenated or weighted features from several modalities are used in feature-level fusion techniques prior to classification. [8]. Hybrid fusion strategies attempted to combine both early and late fusion techniques for better alignment [9].

Despite these decent developments, fusion-based models frequently fall short in real-world circumstances due to their inability to capture intricate cross-modal linkages. To address this, researchers introduced attention mechanisms, which allow models to focus on contextually relevant features. Bahdanau et al. [10] pioneered attention in sequence modeling, and subsequent works adapted it for multimodal tasks. For

instance, Liu et al. proposed shared-memory attention for modality interaction [11], while Yang et al. introduced multiview attentional networks integrating semantic features [12]. Huang et al. [13] developed deep multimodal attentive fusion models, and Zhang et al. explored hierarchical attention for image-caption relationships. More recent architectures, such as the ITIN model [14], significantly increase task accuracy for sentiment analysis. The CLMLF model [15] applies contrastive learning to more accurately represent image-text features. Yang et al. [16] propose dynamic primary modality selection to optimize multimodal sentiment analysis, improving robustness when one modality is noisy. Chen et al. [17] introduce an adapter-based architecture for efficient multimodal fusion, reducing computational cost while maintaining accuracy, and Wang et al. [18] present adaptive interaction and multi-scale fusion for better cross-modal alignment and sentiment prediction.

According to recent surveys, attention-based fusion has surpassed conventional concatenation-based techniques to become a prominent trend in MSA [19], [20]. Nevertheless, the majority of current models only use cross-modal or self-attention, which restricts their capacity to fully utilize intra-modal and inter-modal interactions. This gap motivates the development of dual-attention frameworks that integrate both mechanisms for richer feature representation and improved sentiment classification.

These developments have been aided by a number of benchmark datasets, such as MVSA-Single and MVSA-Multiple, which include image–text pairs from social media posts [21]. Multimodal sentiment analysis in video-based situations has made extensive use of other datasets, including CMU-MOSI, CMU-MOSEI, and CH-SIMS [22]. Large-scale multimodal samples for practical applications are provided by emerging resources such as RU-Senti and SM-MSD [23].

III. Advanced Multimodal Sentiment Analysis with Dual Attention

The proposed A-MSDA architecture is designed to address limitations in existing multimodal sentiment analysis models by combining intra-modal self-attention and inter-modal cross-attention within a unified framework. Unlike prior approaches such as ITIN and CLMLF [15], which focus primarily on cross-modal alignment or hierarchical attention, A-MSDA jointly captures modality-specific features and bidirectional interactions, enabling richer sentiment representation.

Our approach, presented in Fig. 1, contains two unimodal sub-models for sentiment analysis: textual self-attention-based sentiment analysis and visual self-attention-based sentiment analysis. We first handle texts and images individually. We extract attended characteristics from the text and image via different unimodal self-attention-based models. While the latter concentrates on phrases that have strong emotions, the former focuses on important areas/regions in images that have emotionally solid substances. The extracted features are used for predicting the sentiment polarity in both modalities separately.

In addition to the first layer of attended textual and attended visual feature extraction, we overlay a second level of attention as a bidirectional cross-modal attention-based model to explore how visual and textual data complement one another and interact. Textual cross-attention mechanisms emphasize emotionally relevant words that are linked with attended high-level visual regions. Moreover, the visual cross-attention mechanism focuses on image regions on the basis of attended textual features. This layer hence features text-to-image and image-to-text attention.

These dual-attended features are later used to extract the multimodal sentiment score of text-image data. The final sentiment classification is obtained by aggregating the predictions from the text model, visual model, and fusion output.

Before delving into the details, let us denote the list of images as V and the text descriptions as T. We will then proceed by explaining the various components of the model.

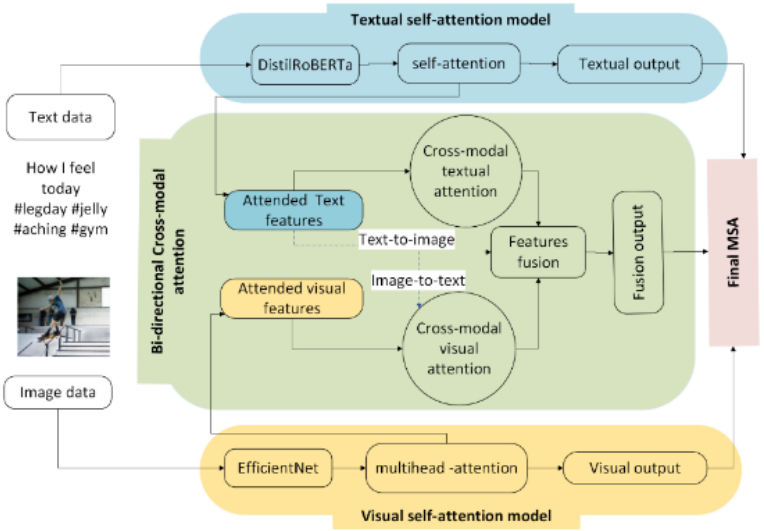

Fig. 1. Advanced multimodal sentiment analysis with a dual attention model architecture.

A. Unimodal Self-Attention Model for Text

In text sentiment analysis (see Fig. 2), the emotional content often hinges on specific words within the text. Our model begins by transforming the input text sequence $W = (W_1, W_2, \ldots, W_n)$, where each W_i is a token, into meaningful embeddings via DistilRoBERTa to capture these nuances, followed by self-attention mechanisms to focus on key features. DistilRoBERTa is a distilled version of RoBERTa model [24] using knowledge distillation [25], it is considered reduced in computational complexity, making it ideal for applications that require efficient processing without substantial loss of performance.

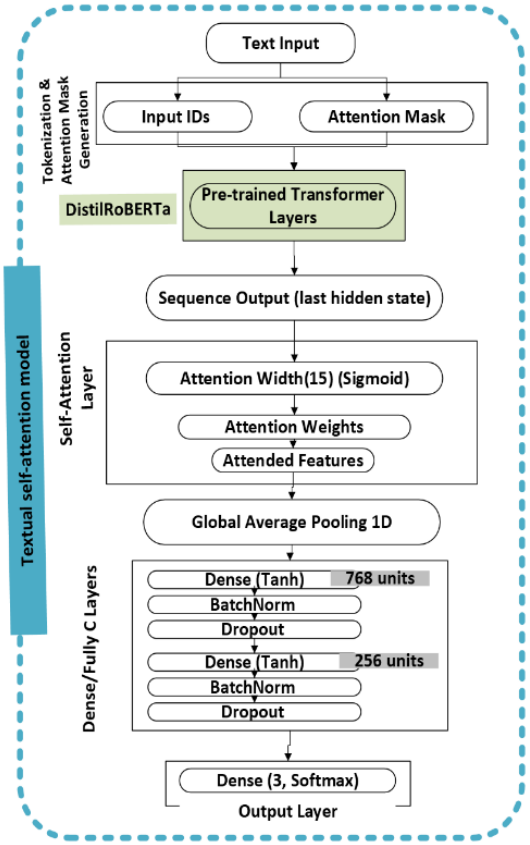

Fig. 2. Textual sentiment analysis model based on self-attention.

The input text sequence $W = (W_1, W_2, \ldots, W_n)$ consists of n tokens. Each token W_1 is transformed into an embedding vector. The model outputs the last hidden state E [Eq. (1)], where each vector e_1 in E is of size d and encapsulates semantic and syntactic information of the corresponding token.

$$\mathrm{E} = [e_1, e_2, \ldots, e_n] = \text{DistilRoBERTa (W)} \quad (1)$$

The embedding E is then passed through attention layers to refine the representation. The self-attention [see Eq. (2)] layer calculates the importance of each token represented by $q^{(i)}$ in relevance to the embedding E represented by K and V by generating attention scores α_1 for each token e_1. The attention score for token e_1 is computed via Eq. (3), where the score function measures the relevance of e_1 with respect to the entire sequence context E.

$$Self\_Attention(Q,K,V) = softmax(\frac{QK^T}{\sqrt{d_k}}) \text{ Where } q^{(i)} = W_q * x^i \text{ for i} \in [1,\mathrm{T}] \ , k^{(i)} = W_k * x^i \ for \ i \in [1,T] \ , v^{(i)} = W_v * x^i \ for \ \ i \in [1,T] \quad (2)$$

$$\alpha_i = \frac{exp(score(e_i,E))}{\sum_{j=1}^{n} exp(score(e_j,E))} \quad (3)$$

The result of this layer is a weighted sum [see Eq. (4)] of the input's embedding, where α_{ij} indicates the attention weight from token W_i to W_j.

$$\hat{e} = \sum_{j=1}^{n} \alpha_{ij} e_j \quad (4)$$

To condense the sequence-level information into a fixed-size representation, a global average pooling layer is employed. This operation computes the mean of the vectors along the sequence dimension [see Eq. (5)], resulting in a single vector $\hat{e}$ that summarizes the entire text, e $\in R^{\mathrm{d}}$ and serves as a comprehensive representation of the input text sequence.

$$e = \frac{1}{n}\sum_{i=1}^{n} \hat{e}_i \quad (5)$$

The pooled vector e is fed into a series of dense layers to capture higher-level abstractions. A nonlinear activation function is applied after a linear modification in each dense layer [see Eq. (6)]:

$$Dense_i(\hat{e}) = tanh(W_i \hat{e} + b_l) \quad (6)$$

where, the activation function used here is the hyperbolic tangent (Tanh). W_i is the weight matrix, and b_l is the bias for the i-th layer. Dropout layers are used to prevent overfitting.

The final layer generates a probability distribution across the sentiment classes [see Eq. (7)]. Softmax activation function is used for this output, where W_0 and b_0 are the output layer's weights and biases, respectively, and y represents the predicted probabilities for each class.

$$y = softmax(W_0 Dropout + b_0) \quad (7)$$

B. Unimodal Self-Attention Model for Images

Similar to how some words are more important than others in a text, some regions of an image are more relevant to sentiment.

Our visual attention model, Fig. 3, is devised to highlight these emotionally essential regions for a better representation of sentiment. The visual model for sentiment analysis is designed to extract and process features effectively from images via a combination of EfficientNetB0 [26], transformer encoders, self-attention mechanisms, and fully connected layers. This architecture seeks to improve the model's capacity to capture sentiment-related visual cues and integrate them with textual features for comprehensive sentiment analysis.

The visual self-attention model starts by extracting visual high-level feature maps with fine-tuned EfficientNetB0, excluding its top layers [see Eq. (8)]. Fine-tuning ensures pretrained weights adaptation to the specific task of visual sentiment analysis, where the input is the image tensor X_{image} and where F_{eff} represents the extracted feature maps.

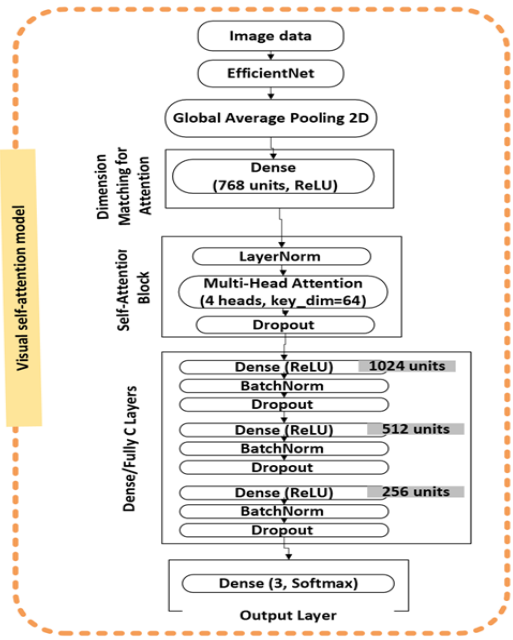

Fig. 3. Visual sentiment analysis model based on self-attention.

$$F_{eff} = EfficientNetB0(X_{image}) \quad (8)$$

To decrease the feature maps' spatial dimensions while keeping the depth information constant, global average pooling is used. This process calculates the mean value for each feature map, resulting in a single value per map [see Eq. (9)], where H and W are the height and width of the feature map, respectively.

$$f_{GAP} = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} F(i,j) \quad (9)$$

The pooled features are passed through a dense layer [Eq. (10)] with 768 units to align their dimensions with the text model's output. The features are then reshaped to prepare for self-attention processing [Eq. (11)], where W_d and b_d are the weights and biases of the dense layer.

$$f_{dense} = ReLU(W_d F_{se} + b_d) \quad (10)$$

$$F_{reshape} = Reshape((1,768))(f_{dense}) \quad (11)$$

The self-attention mechanism within the transformer encoder processes the reshaped feature vector, where Q, K and V are the query, key, and value matrices, respectively, derived from the input features.

$$A_{self} = MultiHeadAttention(Q,K,V) \quad (12)$$

To capture different contextual relationships in the input properties, several heads of attention are employed [see Eq. (12)]. A collection of attended features is produced by calculating and combining multiple attention scores through 4 attention heads with a key dimension of 8. The idea is to focus on various regions of the image contributing the most in understanding the visual sentiment. The residual connections combine the attention output with the original features [see Eq. (13)], followed by layers of normalization and dropout [see Eq. (14)]:

$$x = Add([A_{self}, Freshape]) \quad (13)$$

$$x = Dropout(0.4)(LayerNormalization(x)) \quad (14)$$

The model applies more refinement through several fully connected dense layers with 1024, 521, and 256 units, respectively [see Eq. (15)], with dropout and batch normalization [Eq. (16) and Eq. (17)] to overcome overfitting.

The output layer employs a Softmax activation function [Eq. (18)] to generate a probability distribution across the sentiment categories.

$$f_{fc1} = ReLU(W_{fc1} \cdot A_{self} + b_{fc1}) \quad (15)$$

$$f_{bn1} = BatchNormalization(f_{fc1}) \quad (16)$$

$$f_{drop1} = Dropout(0.4)(f_{bn1}) \quad (17)$$

$$y_{pred} = softmax(W_{out} \cdot f_{drop3} + b_{out}) \quad (18)$$

C. Cross-Modal Attention for Multimodal Sentiment Analysis

In the first unimodal processing layer, the text model processes input sequences to capture the underlying semantic and syntactic nuances and outputs of both sentiment prediction and an attention output, which is crucial for the subsequent fusion process. Simultaneously, the visual model processes images to extract features that reflect the sentiment conveyed visually and outputs both visual sentiment prediction and attention-enhanced features.

The second layer represents the core innovation of the model. At this level, we fuse the attended features from both text and visual modalities via cross-modal attention mechanisms for a dual attended features extraction, as shown in Fig. 4 and Algorithm 1.

Algorithm 1: The advanced multimodal sentiment analysis with dual attention

Inputs:
Textual data (T)
Visual data (V)
Outputs:
Predicted multimodal sentiment probabilities

Compute
For (every text T_i, image V_i) do
$Text_{sentiment}$, att_T ← self-att-based-model-text(T_i)
$Visual_{sentiment}$, att_V ← self-att-based-model-imaget(V_i)
dual-att-T-to-V← coss_attention(att_T, att_V)
dual-att-V-to-T← coss_attention(att_V, att_T)
concat-dual-att ← [dual-att-T-to-V,dual-att-V-to-T]
F_flattened ← flatten(concat-dual-att)
h_{fusion} ← dense_layers(F_flattened)
$Fusion_{sentiment}$ ← Softmax(W_{fusion} * h_{fusion} + b_{fusion})
Msa_{logits} ← (w_{text} * logits($Text_{sentiment}$,)) + (w_{visual} * logits($Visual_{sentiment}$,)) + (w_{fusion} * logits($Fusion_{sentiment}$))
$Multimodal_{sentiment}$ ← Softmax(Msa_{logits})
End

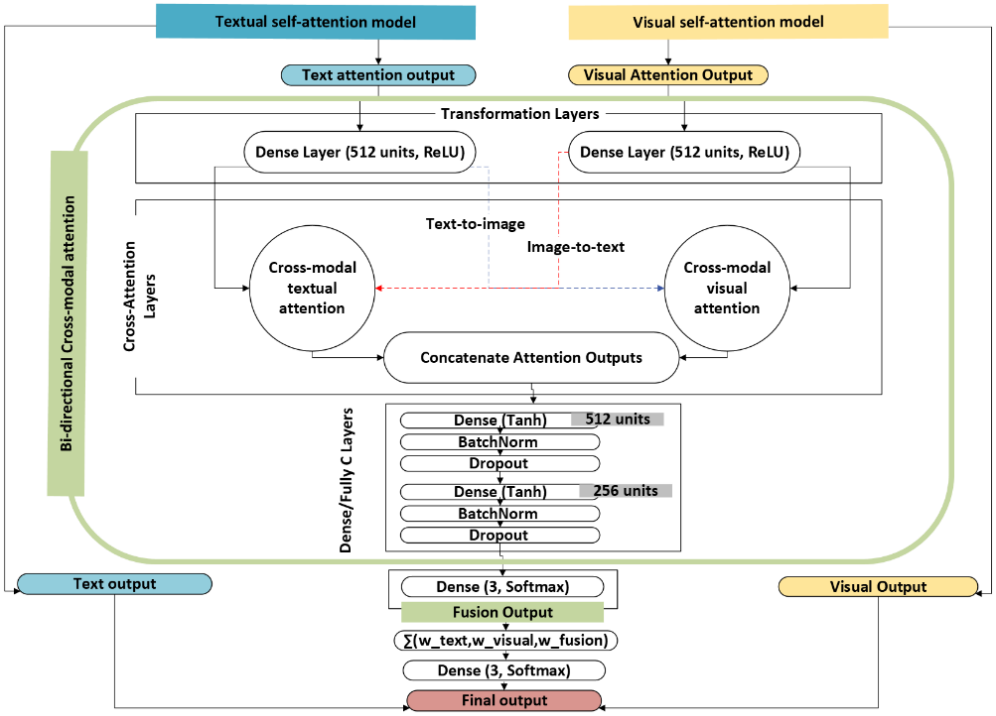

Fig. 4. Cross-modal attention-based fusion for textual–visual sentiment analysis architecture.

The attention mechanism computes the influence of text features on visual features. Given that, A_v is the visual attention output, and A_t is the text attention output [see Eq. (19)]. Similarly, the image-to-text attention computes the influence of visual features on the text features [Eq. (20)]. The outputs of both attentions are concatenated [Eq. (21)], flattened and passed through dense layers for further processing [Eq. (22)].

$$A_{text-image} = Attention(A_v, A_t) \quad (19)$$

$$A_{image-text} = Attention(A_t, A_v) \quad (20)$$

$$A_{concat} = [A_{text-image}, A_{image-text}] \quad (21)$$

$$h_1 = ReLU(W_1 * flatten(A_{concat}) + b_1) \quad (22)$$

A series of dense layers [see Eq. (23)] (512 and 256 units, respectively) followed by Softmax activation outputs [see Eq. (24)], the sentiment classification of the concatenated data as a result of the fusion process.

$$dense_{fusion} = Relu(W_i\ \hat{e} + b\_l\ ) \quad (23)$$

$$fusion\_outpu = softmax(W_0 dense_{fusion} + b_0) \quad (24)$$

The final sentiment prediction aggregates the outputs of the text model, visual model, and fusion output

Three trainable weights $w_{text}, w_{visual}, w_{fusion}$ determine the contribution of each modality to the final prediction. These weights are normalized to sum to 1 [Eq. (25)], and the outputs are combined as a weighted average [Eq. (26)]. Then, the combined logits are fed into a Softmax layer [Eq. (27)], which outputs the final probability distribution across the sentiment categories.

$$w_{normalized} = \sum(w_{text}, w_{visual}, w_{fusion}) \quad (25)$$

$$y_{logits} = w_{textnormalized} \cdot text_{output} + w_{visualnormalized} \cdot visual_{output} + w_{fusionnormalized} \cdot fusion_{output} \quad (26)$$

$$multimodal\_output = softmax(y_{logits}) \quad (27)$$

In summary, combining textual and visual data enables the multimodal model to detect a broader spectrum of sentiment signals, strengthening the sentiment analysis process. Self-attention mechanisms improve the model's focus on key elements of the input, while cross-attention contributes to the alignment between modalities, leading to precise and refined sentiment predictions. This all-encompassing approach guarantees that the outcome includes a thorough evaluation of the text and image data.

IV. Experiments

A. Datasets

Niu et al. [21] introduced two publicly available multimodal sentiment analysis datasets, MVSA-Single and MVSA-Multiple, sourced from Twitter. The MVSA-Single dataset consists of 5,129 image–text pairs, each classified into positive, neutral, or negative sentiment categories. In contrast, the MVSA-Multiple dataset contains 19,600 image–text pairs, with each pair independently annotated by three different annotators.

To ensure consistency in sentiment labels between the image and text, we preprocess these datasets for our experiments. In particular, pairs with different pictures and text labels are eliminated. The positive (or negative) sentiment label is applied to the pair when one label is positive (or negative) and the other is neutral. After this preprocessing, there are now

17,024 image–text pairings in the revised MVSA–Multiple dataset and 4,511 image–text pairs in the revised MVSA–Single dataset. We adhere to earlier preprocessing techniques to guarantee an equitable comparison.

TABLE I. MVSA DATASET STATISTICS

Dataset	Neutral	Positive	Negative	Total
MVSA-multiple	4408	11318	1298	17024
MVSA-single	470	2683	1358	4511

Table I presents the statistics for the two MVSA datasets. This finding demonstrates that the data distributions for MVSA-Single and MVSA-Multiple differ and that the various categories are extremely uneven. Without sampling, the dataset's smaller categories will be challenging to examine during training, and all of the data will be categorized into the same class. Because the MVSA datasets are tiny, the smallest category in each MVSA dataset is randomly oversampled to lessen the effect of data imbalance on the experiment.

B. Implementation Details

The proposed model is implemented in Python 3.7.13 using Keras on Google Colab with an NVIDIA T4 GPU (50 GB RAM).

1) Textual model: Text features are extracted using DistilRoBERTa, with sequences (max = 20) tokenized, padded, or truncated. The resulting 768-dimensional embeddings are refined by a self-attention layer and aggregated via global average pooling, followed by dense layers (512, 256) with 0.3 dropout. A Softmax layer classifies text into positive, neutral, or negative sentiment.

2) Visual model: Images (224×224×3) are processed using EfficientNetB0, followed by global pooling, a dense projection (768 units), and a multihead self-attention (4 heads, key = 8). Fully connected layers with batch normalization and dropout precede the Softmax classifier.

3) Cross-modal fusion: Textual and visual features (768 dims) are aligned and fused via cross-attention, where text acts as query and image as key/value. The fused representation passes through dense layers with batch normalization and dropout, and the final Softmax outputs the sentiment class.

C. Training Process

Our model is trained using the Adam optimizer with a learning rate of 0.00001, a batch size of 128 for MVSA-Multiple and 32 for MVSA-Single. Weighted focal loss is employed to handle class imbalance, combining class weights with focal loss parameters of α=0.27 and γ=1.9, ensuring greater focus on hard-to-classify samples.

To prevent overfitting, we employ early stopping with a patience of 5, while a learning rate scheduler reduces the rate by a factor of 0.1 if the validation loss plateaus for three epochs. Data is splinted using an 8:1:1 ratio for training, validation, and testing. The effectiveness and performance of A-MSDA are greatly influenced by the hyperparameter selection. For example, adding more attention heads increases the computational cost but enhances the capacity to record a variety of cross-modal connections. Overfitting is lessened by a larger dropout rate, but if it is set too high, convergence may be slowed. While the learning rate guarantees gradual convergence, dense layer sizes strike a balance between representation capacity and training stability. By highlighting difficult-to-classify samples, focal loss parameters aid in reducing class imbalance. Memory consumption and gradient stability are impacted by batch size; larger batches speed up training but demand more GPU power. Grid search and cross-validation were used to adjust these settings, which are presented in Table II, in order to maximize the trade-off between efficiency and accuracy.

TABLE II. HYPERPARAMETERS OF THE A-MSDA MODEL

Hyperparameter	MVSA-Multiple	MVSA-Single
Word vector dimensionality	768	768
Maximum input text length	20	20
Batch size	128	32
Optimizer	Adam	Adam
Learning rate	0.00001	0.00001
Dropout rate	0.4	0.4
Epochs	20	20
Loss function	Weighted Focal Loss	Weighted Focal Loss
Balancing factor	α=0.27	α=0.27
Focusing parameter	γ=1.9	γ=1.9

D. Main Results

To evaluate our work, we compare, in Table III, the proposed model against several benchmark models. All these models are evaluated and tested on the MVSA dataset, and the reported values are based on their original papers:

- The Co-memory model [27] enhances sentiment analysis by introducing a fusion module that strengthens the interactions within multimodal datasets.
- The CNN-Multi model [28] uses two separate CNNs to independently process text and image features and combine the extracted features to enhance emotion prediction.
- The MVAN model [29] uses a memory network within a multiview attention network to refine semantic image-text features, leading to increased accuracy.
- Xu et al. [30] proposed MultiSentiNet, a deep CNN technique with pretrained features that integrate text embedding with object and scene information encoded from photos. To capture the influence of visual characteristics on word embedding, they employed visual feature-guided attention.
- Braytee et al. [31], dual-pipeline based attention mechanism for multimodal social sentiment analysis.

For the MVSA-single dataset, the A-MSDA model achieves an accuracy of 73.01% and an F1 score of 71%. For the MVSA-multiple dataset, the model recorded an accuracy of

73.40% and an F1 score of 73.35%. A comparison with the state-of-the-art methods reveals that our A-MSDA approach outperforms the current state-of-the-art approaches.

TABLE III. PERFORMANCE ANALYSIS OF THE A-MSDA METHOD COMPARED WITH THE STATE-OF-THE-ART METHODS

Model	MVSA-Single		MVSA-Multiple	
	Accuracy	F1-score	Accuracy	F1-score
Unimodal Visual SA EfficientNetB0	62.31	62.28	57.46	56.10
Unimodal Textual SA DistilRoBERTa	65.56	65.12	66.56	66.00
Co-Memory [27]	70.51	70.01	68.92	68.83
CNN-Multi [28]	61.20	58.37	66.39	64.19
MVAN [29]	72.98	**72.98**	72.36	72.30
MultiSentiNet [30]	69.84	69.63	68.86	68.11
Dual-Pipeline Attentional architecture [31]	57.29	56.76	73.18	69.76
Proposed Model (A-MSDA)	**73.01**	71.10	**73.40**	**73.17**

Table IV shows that class imbalance has a significant effect on the performance indicators. The macro-average values for the precision, recall, and F1 score are significantly lower at 0.64, 0.57, and 0.60, respectively, despite the total accuracy of 0.73, which indicates a respectable degree of prediction accuracy. Severe class imbalance impacts the model's ability to consider and evaluate all classes equally, despite the use of weighted focus loss classes and sampling strategies; this discrepancy in the macro-average values highlights the challenge posed by class imbalance.

TABLE IV. CLASSIFICATION REPORT OF THE A-MSDA MODEL ON BOTH THE MVSA-SINGLE (A) AND MVSA-MULTIPLE (B) DATASETS

	Precision	Recall	F1-score	Support
Neutral	0.57	0.63	0.60	442
Positive	0.85	0.76	0.80	1114
Negative	0.51	0.80	0.63	137
Accuracy			0.73	1693
Macro avg	0.64	0.72	0.67	1693
Weighted avg	0.75	0.73	0.73	1693
	Precision	**Recall**	**F1-score**	**Support**
Neutral	0.52	0.35	0.42	43
Positive	0.77	0.88	0.82	290
Negative	0.63	0.49	0.55	118
Accuracy			0.73	451
Macro avg	0.64	0.57	0.60	451
Weighted avg	0.71	0.73	0.71	451

However, when the class distribution is considered, the weighted average values demonstrate improvements, with a precision of 0.71, a recall of 0.73, and an F1 score of 0.71. This implies that the model performs better when the class imbalance is taken into account, but the differences between the weighted and macro averages highlight how crucial it is to address and lessen the effects of class imbalance in the feature extraction process for a more thorough and impartial model evaluation.

E. Module Performance Validation and Case Study

We carried out several tests to confirm the performance improvements made by each module in multimodal sentiment analysis. The purpose of these studies is to compare unimodal and multimodal sentiment analysis with unimodal sentiment extraction for text and image features, as well as different feature fusion strategies, to assess the efficacy of the dual-attention MVSA model. The following summarizes the specifics of the model ablation experiments:

- Unimodal visual SA and unimodal textual SA: These experiments assess sentiment analysis using only the visual modality and only the textual modality, respectively.
- Textual sentiment analysis with cross attention: using visual data to assist in the prediction of textual sentiment.
- Visual sentiment analysis with cross attention: using text data to enhance the prediction of visual sentiments.
- Different self-attention models: Evaluation of various self-attention mechanisms.

Table V presents the results of these experiments. The comparison demonstrates that multimodal sentiment analysis surpasses unimodal sentiment analysis for both textual and visual data. The results also highlight the significant impact of cross-modal attention in leveraging one modality to focus on specific aspects of the other, as outlined in the proposed methodology. This mechanism enhances the models' ability to accurately understand and analyse sentiment, leading to improved performance.

TABLE V. CLASSIFICATION RESULTS OF DIFFERENT UNIMODAL AND MULTIMODAL ARCHITECTURES AGAINST THE PROPOSED A-MSDA

Ablation Experiment	MVSA-single		MVSA- multiple	
	Accuracy	F1-score	Accuracy	F1-score
Unimodal Visual SA EfficientNetB0	62.31	62.28	57.46	56.10
Unimodal Textual SA DistilRoBERTa	65.56	65.12	66.56	66.00
Visual SA w/cross attention text-to-image	64.51	63.81	61.64	60.98
Textual SA w/cross attention image-to-text	69.47	68.03	68.29	67.54
A-MSDA w/o self-attention	70.14	70.06	70.82	70.68
A-MSDA w/o cross attention	70.76	70.50	71.63	71.23
A-MSDA	**73.01**	**71.10**	**73.40**	**73.17**

For instance, unimodal textual sentiment analysis produced a moderate accuracy of 65.56% with an F1 score of 65.12%.

the textual emotion classification was enhanced by the addition of image-to-text cross-modal attention. This helps improve the accuracy by more than 3% for accuracy and F1 score. The advantage of integrating information across modalities is highlighted by this improvement.

Our suggested method, advanced multimodal sentiment analysis with dual attention, produced the best outcomes compared to unimodal analysis or single type of attention network integration. A-MSDA achieves a higher degree of interaction between modalities by combining self-attention and cross-modal mechanisms, which results in more reliable sentiment classification. These results demonstrate the effectiveness of our suggested A-MSDA method, which successfully blends self-attention and dual cross-modal mechanisms to produce challenging outcomes in multimodal sentiment analysis tasks.

To illustrate the practical impact of the A-MSDA model, we consider a sample from the MVSA dataset, the image is shown in Fig. 5, and the corresponding text is: "Q&A panel for #TheDressmaker. We laughed, we cried, we fell in love with Liam Hemsworth a LOT. #TIFF15 @TIFF_NET".

Fig. 5. Sample from MVSA-multiple Test Set (ID: 5001).

The self-attention-based text sentiment analysis highlights strong positive cues such as "laughed", "love" words, combined with emotional engagement markers like "we cried". Simultaneously visual self-attention model focuses on regions with vibrant colors and persons but conveys a more neutral sentiment. Then, with the cross-attention, we align these textual cues with visual features, amplifying the positive prediction.

A-MSDA combines both text and image cues; it is impacted by the unimodal results of each modality but also enhanced with the shared content, ensuring accurate sentiment classification. Here, A-MSDA Outperforms Unimodal Models because:

- Image-only models ignore nuanced textual enthusiasm and consider the image as neutral.
- Text-only model detected positive sentiment but is slightly influenced by the mixed emotional tone "we cried" as it is considered negative.

V. Limitations

Despite the promising results, this study has several limitations.

1) Dataset size and diversity: the used MVSA-Single and MVSA-Multiple were constructed from Twitter posts. They are relatively small and not well generalized. This may limit the generalizability of the model to other platforms or domains.

2) Class imbalance: Although weighted focal loss was applied, severe class imbalance in the datasets still affected macro-average performance, indicating challenges in handling minority classes.

3) Computational complexity: The dual-attention architecture increases computational cost compared to simpler fusion models, which may not be the optimum solution to resource-constrained environments.

Future work should address these limitations by incorporating larger and more diverse datasets, additional modalities, and optimization strategies for efficiency.

VI. Conclusion

In this work, we presented A-MSDA, an advanced multimodal sentiment analysis model where we focus on overcoming the interlinkage investigation between modalities, a major gap in traditional MSA strategies. To overcome this challenge, our approach integrates self- and cross-attention mechanisms to capture the complementary strengths of textual and visual data.

Our proposed model effectively focuses on sentiment-relevant features within each modality and aligns them through multiple layers of attention-based fusion. Each modality's use of self-attention mechanisms efficiently extracts high-quality characteristics, highlighting the most pertinent elements of both visual and textual material. Moreover, we emphasize the interaction between modalities by integrating the cross-attention fusion layer in order to enhance the accountability of the combined sentiment.

The experimental results demonstrate that A-MSDA improves overall sentiment classification performance, mainly compared to unimodal models. Beyond performance gains, the use of attention mechanisms provides a clearer and more interpretable view of how different modalities impact each other and contribute to the final decision.

Building on the current findings, and considering the existence of different limitations and improvement points, several directions can be explored: The multimodality doesn't reflect only text or image, we can explore extending the model to include different sources of information, such as audio-video, temporal information, likes and comments in social media platforms, and even the number of clicks. This could provide a richer representation of sentiment. Additionally, testing on larger multimodal datasets and investigating domain

adaptation techniques will potentially improve generalization across different social media platforms and languages. Furthermore, we think of integrating explainability tools to offer deeper insights into the reasoning behind model predictions. These directions aim to enhance the applicability, transparency, and robustness of multimodal sentiment analysis systems.

Overall, this study highlights the importance of attention-based fusion in bridging multimodal gaps and provides a strong foundation for more robust, transparent, and human-like sentiment understanding systems.

REFERENCES

[1] A. Gandhi, K. Adhvaryu, S. Poria, E. Cambria, and A. Hussain, "Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions," Information Fusion, vol. 91, pp. 424–444, 2023.

[2] H. Zhang, "A comprehensive survey on multimodal sentiment analysis: Techniques, models, and applications," Advances in Engineering Innovation, vol. 12, pp. 128–142, 2024.

[3] V. Vinitha and R. Jayanthi, "A Comprehensive Review of Multimodal Sentiment Analysis on Social Networks," in Proceedings of International Joint Conference on Advances in Computational Intelligence, Springer, 2024, pp. 655–664.

[4] M. S. Alam, M. S. Hossain, and M. A. Rahman, "Sentiment Analysis in Social Media: How Data Science Impacts Public Opinion," American Journal of Scholarly Research and Innovation, vol. 4, no. 1, 2025.

[5] X. Xu et al., "SentiMM: A Multimodal Multi-Agent Framework for Sentiment Analysis in Social Media," 2025.

[6] S. Lai, X. Hu, H. Xu, Z. Ren, and Z. Liu, "Multimodal Sentiment Analysis: A Survey," May 2023, [Online]. Available: https://arxiv.org/abs/2305.07611

[7] S. Fatimi, W. Sabbar, and A. Bekkhoucha, "Textual-Visual multimodal sentiment analysis : A Review of Approaches and Challenges," 2023 14th International Conference on Intelligent Systems: Theories and Applications (SITA), pp. 1–8, 2023, doi: 10.1109/SITA60746.2023.10373735.

[8] Q. Qi, L. Lin, and R. Zhang, "Feature Extraction Network with Attention Mechanism for Data Enhancement and Recombination Fusion for Multimodal Sentiment Analysis," Information 2021, Vol. 12, Page 342, vol. 12, no. 9, p. 342, Aug. 2021, doi: 10.3390/INFO12090342.

[9] X. Guo, M. Wushouer, and G. Tuerhong, "Survey of Sentiment Analysis Algorithms Based on Multimodal Fusion," Computer Engineering and Applications, vol. 60, no. 2, pp. 1–18, 2024.

[10] D. Bahdanau, K. Cho, and Y. Bengio, "Neural Machine Translation by Jointly Learning to Align and Translate," Sep. 2014, [Online]. Available: http://arxiv.org/abs/1409.0473

[11] Z. Zhao et al., "An image-text consistency driven multimodal sentiment analysis approach for social media," Inf Process Manag, vol. 56, no. 6, p. 102097, 2019, doi: 10.1016/j.ipm.2019.102097.

[12] X. Yang, S. Feng, D. Wang, and Y. Zhang, "Image-text multimodal emotion classification via multi-view attentional network," IEEE Trans Multimedia, vol. 23, no. c, pp. 4014–4026, 2021, doi: 10.1109/TMM.2020.3035277.

[13] K. Zhang, Y. Zhu, W. Zhang, and Y. Zhu, "Cross-modal image sentiment analysis via deep correlation of textual semantic," Knowl Based Syst, vol. 216, p. 106803, 2021, doi: 10.1016/j.knosys.2021.106803.

[14] T. Zhu, L. Li, J. Yang, S. Zhao, H. Liu, and J. Qian, "Multimodal Sentiment Analysis With Image-Text Interaction Network," IEEE Trans Multimedia, 2022, doi: 10.1109/TMM.2022.3160060.

[15] Z. Li, B. Xu, C. Zhu, and T. Zhao, "CLMLF:A Contrastive Learning and Multi-Layer Fusion Method for Multimodal Sentiment Detection," Findings of the Association for Computational Linguistics: NAACL 2022 - Findings, pp. 2282–2294, 2022, doi: 10.18653/v1/2022.findings-naacl.175.

[16] X. Yang and others, "Improving Multimodal Sentiment Analysis via Modality Optimization and Dynamic Primary Modality Selection (MODS),". arXiv preprint arXiv:2511.06328, 2025

[17] Y. Chen and others, "Mixture of Multimodal Adapters for Sentiment Analysis," in Proceedings of NAACL, 2025. doi: 10.18653/v1/2025.naacl-long.90.

[18] Z. Wang and others, "A method for multimodal sentiment analysis: adaptive interaction and multi-scale fusion (AIMS)," J Intell Inf Syst, 2025, doi: 10.1007/s10844-025-00957-1.

[19] M. A. Halim, N. Zainuddin, and K. A. F. Abu Samah, "Fusion in Multimodal Sentiment Analysis: A Review of Approaches and Challenges," International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems. Singapore: Springer Nature Singapore, 2025.

[20] P. Q. Dao, T. B. Nguyen-Tat, and M. Roantree, "Exploring Multimodal Sentiment Analysis Models: A Comprehensive Survey," 2024 International Conference on Multimedia Analysis and Pattern Recognition (MAPR). IEEE, 2024. p. 1-7

[21] T. Niu, S. A. Zhu, L. Pang and A. El Saddik, Sentiment Analysis on Multi-view Social Data, MultiMedia Modeling (MMM), pp: 15-27, Miami, 2016.

[22] W. Yu et al., "CH-SIMS: A Chinese Multimodal Sentiment Analysis Dataset with Fine-grained Annotation of Modality," pp. 3718–3727, 2020, doi: 10.18653/v1/2020.acl-main.343.

[23] J. Peng, Y. He, Y. Chang, Y. Lu, P. Zhang, and others, "A Social Media Dataset and H-GNN-Based Contrastive Learning Scheme for Multimodal Sentiment Analysis," Applied Sciences, vol. 15, no. 2, p. 636, 2025.

[24] Y. Liu et al., "RoBERTa: A Robustly Optimized BERT Pretraining Approach," no. 1, arXiv preprint arXiv:1907.11692, 2019.

[25] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT , a distilled version of BERT : smaller , faster , cheaper and lighter," pp. 2–6, 2019.

[26] M. Tan and Q. V Le, "EfficientNet : Rethinking Model Scaling for Convolutional Neural Networks,". International conference on machine learning. PMLR, 2019. p. 6105-6114.

[27] N. Xu, W. Mao, and G. Chen, "A co-memory network for multimodal sentiment analysis," 41st International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2018, pp. 929–932, 2018, doi: 10.1145/3209978.3210093.

[28] G. Cai and B. Xia, "Convolutional Neural Networks for Multimedia Sentiment Analysis," pp. 159–167, 2015, doi: 10.1007/978-3-319-25207-0.

[29] X. Yang, S. Feng, Y. Zhang, and D. Wang, "Multimodal sentiment detection based on multi-channel graph neural networks," ACL-IJCNLP 2021 - 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Proceedings of the Conference, pp. 328–339, 2021, doi: 10.18653/v1/2021.acl-long.28.

[30] N. Xu and W. Mao, "MultiSentiNet : A Deep Semantic Network for Multimodal Sentiment Analysis MultiSentiNet : A Deep Semantic Network for Multimodal Sentiment Analysis," no. November, 2017, doi: 10.1145/3132847.3133142.

[31] A. Braytee, A. S. Yang, and A. Anaissi, "A Novel Dual-Pipeline based Attention Mechanism for Multimodal Social Sentiment Analysis," pp. 1816–1822, 1816, doi: 10.1145/3589335.3651967.