International Journal of Engineering Trends and Technology
ISSN: 2231-5381/ https://doi.org/10.14445/22315381/IJETT-V74I6P124

Volume 74 Issue 6, 343-364, June 2026

Original Article

Contrastive Bidirectional Cross-Modal Attention Framework for Enhanced Multimodal Sentiment Analysis

Prashant Adakane[1], Amit Gaikwad[2]

[1,2]*Department of Computer Science and Engineering, G H Raisoni University, Amravati, Maharashtra, India.*

[1]*Corresponding Author : prashant.adakane@ghrua.edu.in*

Received: 19 January 2026 Revised: 07 April 2026 Accepted: 20 April 2026 Published: 27 June 2026

Abstract *- The high rate of social media content development causes an increase in multimodal data, such that modeling relationships between visual and textual data is challenging. Nevertheless, most of the available methods cannot capture fine-grained text-to-visual or visual-to-text interaction, resulting in lower sentiment performance. A Contrastive Bidirectional Cross-Modal Attention (C-BCMA) model is presented to enhance the correspondence of textual and visual representations by acquiring a common latent space. An attention method inspired by CLIP is utilized to produce robust cross-modal latent features to enhance their joint representation. Textual features are derived using ALBERT, whereas EfficientNet-B2 is applied to obtain visual representations. Interactions between modalities are learned using a multi-head attention mechanism. Textual and visual information is handled jointly during learning. This helps reduce gaps between the two modalities. This enables the model to process various semantic cues at once. Contrastive learning is used in the model to align similar text-image pairs and to separate unrelated text-image pairs so that better multimodal representations are achieved. The model has a better performance than baseline approaches on both single and multiple annotation versions of MVSA datasets. It achieves better performance across various evaluation metrics. Less obvious expressions like sarcasm and implicit sentiment are handled more effectively in this work, improving interpretation in multimodal sentiment analysis of social media data.*

Keywords *- Albert, Clip, Cross-Modal Attention, Efficientnet-B2, Feature Fusion, Multimodal Sentiment Analysis.*

1. Introduction

The multimodal content has increased significantly due to the fast growth of platforms like Twitter, Instagram, and YouTube. Visual elements are often used to complement textual descriptions in such content. This derivation of sentiment in such multimodal data, as a result, has become an iconic research endeavour, especially in areas like opinion mining, social media monitoring, and affective computing. The conventional types of sentiment analysis process mostly rely on the information in texts; however, user-created content is often characterized by affect comprehended by an integrative combination of both linguistic and visual features. In line with this, analysis of sentiment based on multiple data modalities has proven to be a significant research field as it encompasses a combination of heterogeneous modalities to enable a more holistic understanding of sentiment [1, 2].

Although there is significant progress in multimodal frameworks and neural architectures, interaction between textual and visual information is still a challenge. The contextual and semantic meaning of textual inputs is realized through linguistic structures, whereas the meaning of visual input is the expression of sentiment through objects, scenes, and contextual images. Due to these inherent discrepancies, the unimodal model of sentiment analysis is often unable to represent the entire affective context of a multimodal text, particularly in an environment that is full of sarcasm, contextual ambiguity, or conflicting visual-textual cues [3].

Most previous research was devoted to the analysis of unimodal sentiment using various algorithmic methods and neural network-based approaches. Convolutional and recurrent network models and transformer models of text classification, including BERT and ALBERT, have significantly improved predictive performance and computational efficiency [4, 5]. Similarly, visual sentiment analysis has also been aided by convolutional architectures such as VGGNet, ResNet, and EfficientNet, which are applied to retrieve high-level image representations. However, unimodal methods become inefficient in cases when the sentiment is revealed through the combined interaction of the textual information with the visual one [6]. In order to overcome these constraints, a multimodal fusion strategy spectrum has been put forward. Early-fusion approaches fuse

textual and visual features at the input level, but the same strategy can generate noise and does not help in more profound modal interactions. Late fusion schemes combine the projections of independent unimodal models, and these methods cannot effectively capture cross-modal dependencies [7]. More recent research has been done on attention-based architectures, which enable interaction of textual and visual features. Transformer-based multimodal models and the further development of cross-modal attention provide better interaction between features and focus on relevant information of various modalities in a selective manner. However, many of the models that survive rely on unidirectional mechanisms of attention, which may cause semantic convergence between modalities to occur incompletely [8]. Contrastive representation learning is another promising direction that has been used to great success in vision-language models, including CLIP and ALIGN, to match the embeddings of text and images in a shared semantic space. Although contrastive learning enhances alignment of cross-modal representations, existing methods avoid orienting to emotion-related tasks of general vision-language computations, and thus generally do not implement features needed to reflect subtle multimodal interactions in affective understanding [9].

A careful examination of available literature shows that there are two major gaps. First, to begin with, many models of multimodal sentiment analysis do not explicitly impose any semantic congruency between textual and visual embeddings, which may result in a poor representation across the cross-modes. Second, unidirectional interactions between cross-modal interactions in attention-based models prohibit their ability to provide balanced multimodal interactions. These gaps indicate that there is a need for a framework that would simultaneously facilitate the intensive cross-modal communication and the accurate semantic correspondence.

1.1. Problem Definition and Research Advancement

Despite the significant developments over the past years, the existing multimodal sentiment models are still struggling to balance a high level of inter-modal interaction areas, as well as the consistent alignment of semantics between text and image representations. Attention-based fusion models enhance interaction between features but are often based on unidirectional attention.

This can lead to lower levels of robustness in complex sentiment cases and feature incomplete modality alignment. On the other hand, contrastive learning models like CLIP are highly aligned in cross-modal representation, but mainly aim at general vision-language problems, not the fine-grained understanding of sentiments. Thus, one of the key research questions is left, i.e., the development of a unified framework that can effectively integrate bidirectional cross-modal interaction with semantic alignment in multimodal sentiment analysis.

In order to address this problem, the Contrastive Bidirectional Cross-Modal Attention (C-BCMA) framework has been proposed, which integrates bidirectional cross-modal attention with contrastive representation learning and allows modalities to interact with each other in a balanced manner but maintain semantic consistency in the shared embedding space. Unlike in the current multimodal fusion methods, this integration proceeds to consider both cross-modal interaction and semantic alignment in the same structure, as shown in Figure 1.

This Figure provides the conceptualization of the drawbacks of existing multimodal fusion methods and why the framework was proposed. Early fusion techniques have the issue of features joining noisily, whereas late fusion techniques do not capture the cross-modal dependencies in an effective way. Transformer-based attention models can improve interaction between modalities, but they tend to use unidirectional alignment, which results in missing semantic correspondence between text and visuals.

The C-BCMA framework addresses these shortcomings by incorporating the bidirectional cross-modal attention and contrastive representation learning, which makes the C-BCMA framework capable of both balanced interaction between the modalities and robust semantic similarity in a common embedding space.

In this connection, Figure 1 demonstrates the inefficacies of existing practices and, thus, the opportunities of filling the fundamental gaps in research by offering the Contrastive Bidirectional Cross-Modal Attention framework (C-BCMA) as a universal solution.

Considering these limitations, this paper presents the C-BCMA model that uses a contrastive learning module based on the Contrastive Language-Image Pretraining (CLIP) framework within a multimodal sentiment analysis pipeline. The framework builds upon the Bidirectional Cross-Modal Attention (BCMA) framework and introduces contrastive alignment pair constraints to provide semantic consistency of text-image pairs. C-BCMA model utilizes a BERT-based ALBERT text encoder and EfficientNet-B2 visual feature encoder, and after that applies multi-head cross-modal attention between text and image on a fine-grained level.

In a joint cross-encoding embedding space, alignment aims to maximize the spatial distance between semantically incongruent or irrelevant pairs of items, and at the same time, make semantically congruent pairs closer to one another. The cross-encoder attention leads to this joint alignment, which makes sure each of the modalities provides indispensable and unique cues that may be used to interpret sentiment analysis more accurately.

In summary, this research has made the following contributions:

- Multimodal fusion paradigm novelties: contrastive attention-based bidirectional focal alignment of text and image, which goes beyond the conventional alignment.
- The Alignment Definition: The current work lays the groundwork of contrastive learning, which takes place in attention systems to instantiate cross-alignment of paired information and, as a result, to provide a framework within multimodal systems that generates aligned representations.
- Extensive Assessment: MVSA benchmark datasets' empirical validation studies have shown substantial improvements in performance as compared to other baseline models. In this study, MVSA-Single and MVSA-Multiple are denoted as MVSA-S and MVSA-M.
- Interpretability and robustness Attention maps: Attention maps represent both visual and analytic representations of how the model can be used to detect sarcasm, shifts in polarity, and cross-modal incongruences, and are used to report previously unknown failure modes of unimodal or basic fusion designs.

The structure of the remaining manuscript is outlined in subsequent sections. Section 2 gives a broad overview of existing literature in the field of multimodal sentiment analysis, attention-based fusion mechanisms, and contrastive representation learning. Section 3 outlines the suggested Contrastive Bidirectional Cross-modal Attention (C-BCMA) framework and provides details of its architecture design and training approach. Section 4 presents implementation-level setup, including data preparation, preprocessing steps, details of implementation, and the performance metrics that are used to evaluate performance. Section 5 contains the results of the experiment and provides a comprehensive discussion, including the comparative analysis, ablation analysis, robustness analysis, and error diagnostics. Section 6 describes the ethical concerns related to the use of databases, the possible biases, and the ethical implementation of multimodal sentiment analysis systems. Lastly, Section 7 includes the concluding discussion with the major findings highlighted and the possible ways of further research.

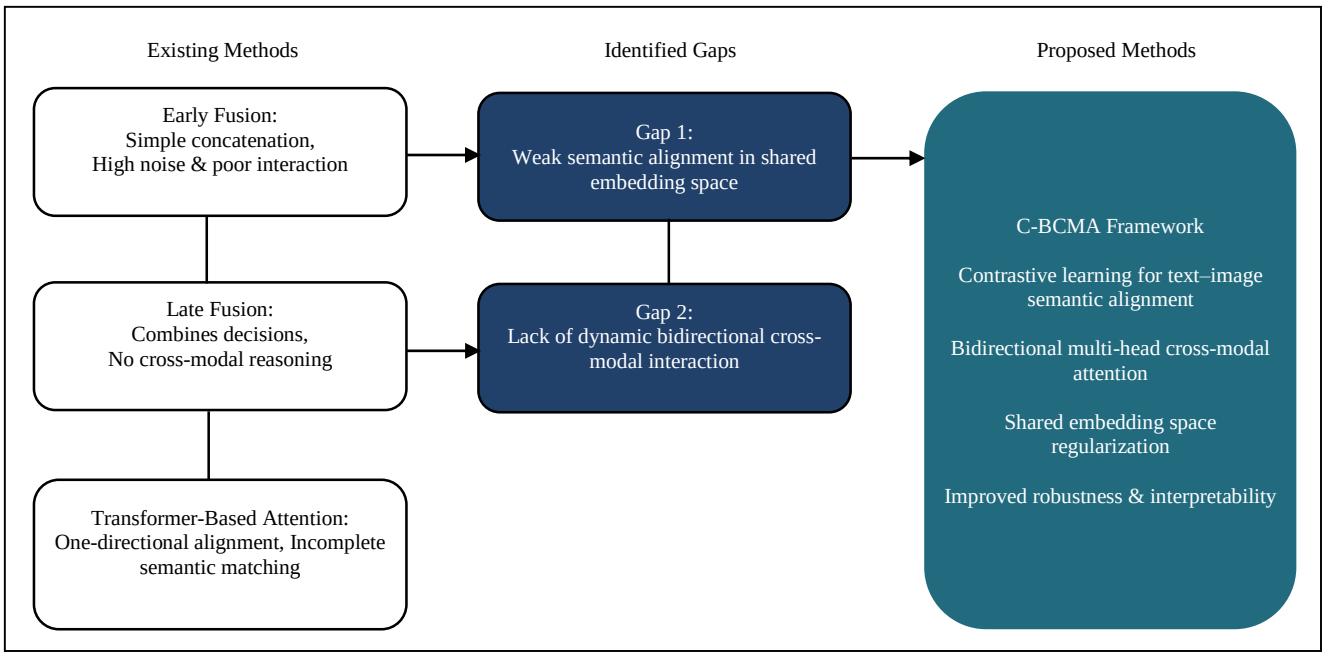


Fig. 1 A comparison of the available multimodal fusion methods, gaps that exist, and the suggested C-BCMA model

2. Literature Review

2.1. Unimodal Sentiment Analysis

The past studies on automated sentiment analysis have been mainly focused on single-modality data, specifically on texts. The early approaches were mainly based on traditional techniques of natural language processing that included bag-of-words representations and sentiment lexicons in order to derive a polarity out of textual corpora [10, 11]. These approaches were later extended with machine learning paradigms that were able to improve classification accuracy by deriving discriminative patterns using annotated datasets. Convolutional neural networks and recurrent neural networks have gained popularity in sequence modelling and sentiment prediction due to the development of deep learning [12]. In recent times, language models based on the transformer, including BERT, RoBERTa, and ALBERT, have dramatically advanced the language of representation learning by training contextual dependencies on large corpora through the assertions of self-attention [13]. These trained models are always better in performance in most natural language

understanding tasks, including sentiment classification. Visual sentiment analysis has also advanced, and it uses convolutional neural networks, including VGGNet, ResNet, and EfficientNet, to obtain high-level semantic representations of images [14, 15].

These architectures will simplify the finding of the affective cues that are contained in the visual components, such as facial expression, context of the scene, and composition.

However, despite the success of unimodal solutions, they often do not capture the sentiment that is articulated by a collaborative action of text and images. The posts made by users in social media may contain a complementary or contradictory signal in the modalities, for example, sarcastic descriptions of neutral pictures, or a visually expressive image not corresponding to the description. As a result, multimodal interactions require models that can reason together across modalities, which unimodal frameworks cannot meet satisfactorily.

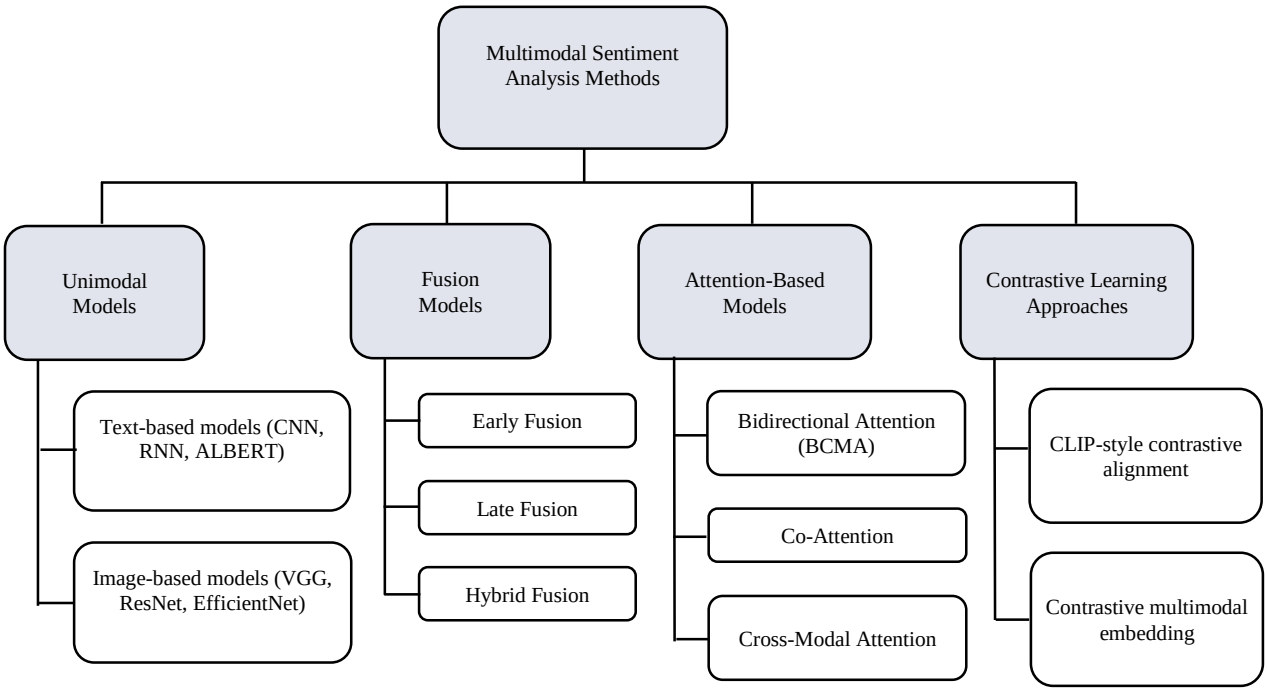


Fig. 2 Taxonomy of multimodal sentiment analysis approaches

2.2. Multimodal Sentiment Analysis

Multimodal sentiment analysis is aimed at combining heterogeneous cues, such as linguistic and visual ones, to obtain a more precise idea of affective intent in social media messages [16]. As presented in Figure 2, currently existing methodologies can be categorized into broad categories as unimodal models, fusion-based models, attention-driven architectures, and contrastive learning frameworks.

Extensive research has been conducted on the use of fusion strategies as a form of multimodal combination. Initial fusion algorithms combine elements of different modalities during the input phase, and thus permit mutual learning during the model training. Although the early fusion is able to promote cross-modal interaction, it is often affected by high dimensionality and noise propagation (when the concatenated features are heterogeneous) [17]. On the other hand, late fusion approaches combine the predictions made by unimodal models that have been independently trained, thus increasing modularity and flexibility, but reducing the ability to generate complex cross-modal interactions [18].

Recent advances in multimodal learning have proposed transformer-based architectures that can be used to allow more advanced interactions between modalities. Multimodal Transformer (MMT) and Multimodal-BERT (MMBERT) are models that use attention-based approaches to concurrently encode information gathered by language and visual means [19]. Image-text classification tasks have also been proven to be better performed with supervised multimodal bitransformers and similar architectures, which learn joint contextual representations of modalities [20].

Additional developments in multimodal architectures have tried to overcome these problems of modality interaction and alignment. As an example, the Multimodal Infomax Sentiment Analysis (MISA) model acquires modality-invariant and modality-specific representations and improves information compatibility between modalities without affecting their specific features [21]. Likewise, Modality-Aware Gated BERT (MAG-BERT) has added modality-specific gating layers that have visual and acoustic attributes to transformer layers with the aim of promoting multimodal

representation learning [22]. Despite the fact that they enhance cross-modal interaction, these models still majorly make use of feature fusion strategies and often do not have direct methods of robust embedding alignment between modalities.

2.3. Attention-based Fusion Mechanism

The attention mechanisms have also been significant in enhancing multimodal learning because they allow models to draw selective attention to the relevant features in various modalities. The models have mechanisms like self-attention, cross-modal attention, and co-attention that enable them to dynamically obtain dependencies between the textual and the visual representations. For instance, hierarchical attention models have been used in proposing image-text fusion tasks, whereas weighted cross-modal attention models have been suggested in multimodal sentiment prediction tasks [23].

Nevertheless, due to these developments, many attention-based models work in a unidirectional way, with one of the modalities as query and the other as the key or value as cross-attention. This non-reciprocal interaction may result in a lack of balance in feature integration when one of the modalities takes control of the representation learning process [24]. In order to overcome this drawback, the Bidirectional Cross-Modal Attention (BCMA) framework was developed to enable mutual interaction between modalities in learning features. BCMA facilitates complementary multimodal cues modeling by facilitating two-way information flow between textual and visual embeddings. Nevertheless, there are challenges to BCMA that concern the alignment of modalities with shared embedding spaces, which can lead to inconsistencies during the processing of heterogeneous multimodal inputs [25, 26].

2.4. Contrastive Learning and Multimodal Alignment

Recent developments in contrastive learning have made major contributions to multimodal alignment in representation learning by learning shared modality spaces. Contrastive learning is effective in learning a large-scale vision-language representation via the Contrastive Language-Image Pretraining (CLIP) model, which trains transferable visual representations with natural language supervision [39]. Later, large-scale approaches like ALIGN continued this paradigm by using large-scale datasets composed of noisy image-text supervision to enhance cross-modal alignment performance [27]. The goal of these contrastive learning methods is to find as many similarities as possible between similar image-texts and as many similarities as possible between dissimilar image-texts, and in doing so, allow models to learn semantically meaningful multimodal representations. Recently, the application of contrastive learning to multimodal sentiment analysis has received focus. As an example, the use of contrastive goals has been used to regularize multimodal embeddings and enhance the robustness of representation [28]. Nevertheless, the majority of current contrastive multimodal models are based on global alignment approaches and typically do not include systems of fine-grained bidirectional intermodal interaction. To overcome these shortcomings, the suggested Contrastive Bidirectional Cross-Modal Attention (C-BCMA) framework combines the contrastive alignment with the bidirectional multi-head attention. In contrast to the conventional fusion-based architectures, the given framework concurrently aligns the textual and visual embeddings to the same representation space, as well as allows interactions based on the attention in both directions. The two-way mechanism enables the model to enhance multimodal representation of complementary information, and it maintains semantic consistency among modalities, thus enhancing the prediction of sentiments in multimodal situations.

3. Proposed Methodology

Contrastive Bidirectional Cross-Modal Attention (C-BCMA) model builds upon the multimodal sentiment analysis by combining contrastive learning with bidirectional cross-modal multi-head attention. Its main objective is to preserve semantic correspondence between textual and visual modalities in a common embedding space as well as to retain the ability of the attention mechanism to reason in fine detailed context. Figure 3 displays a system architecture diagram. The C-BCMA architecture is built to extract the representations of each data source and also facilitates semantic relationships between textual and visual input. The textual modality is first encoded through the ALBERT transformer to produce contextual token representations, but the visual modality is encoded through EfficientNet-B2 to extract the high-level visual features. These modality-specific embeddings are then mapped to a shared embedding space, and thus, allow interaction with each other by applying attention mechanisms. Unlike the traditional multimodal fusion methods that are based on early concatenation or late decision fusion, the presented framework builds on the idea of bidirectional cross-modal attention, where textual representations are able to attend to visual attributes and vice versa. This bidirectional attention provides the model with the power to make fine-grained semantic matches between image regions and lexical units. In order to achieve further improvement of multimodal representation learning, a contrastive alignment goal is added. This goal imposes close proximity between embedding pairs of matching text-image pairs in the same space and topological distance between incompatible pairs. This kind of alignment promotes the model to learn semantically consistent multimodal representations, thereby enhancing the downstream sentiment classification.

The system has four major components:

- Text Encoder The ALBERT embedding model is a text-based word embedder that can be used to extract context-dependent word embeddings.
- Image Encoder -EfficientNet-B2 is used to obtain the compact feature maps of the visual inputs.

- Bidirectional Cross-Modal Attention Module- This module allows a very intensive, two-way communication between text-based embeddings and image-based embeddings.
- Contrastive Alignment Head. This module can be used to obtain semantic alignment between modalities using a contrastive loss formulation.

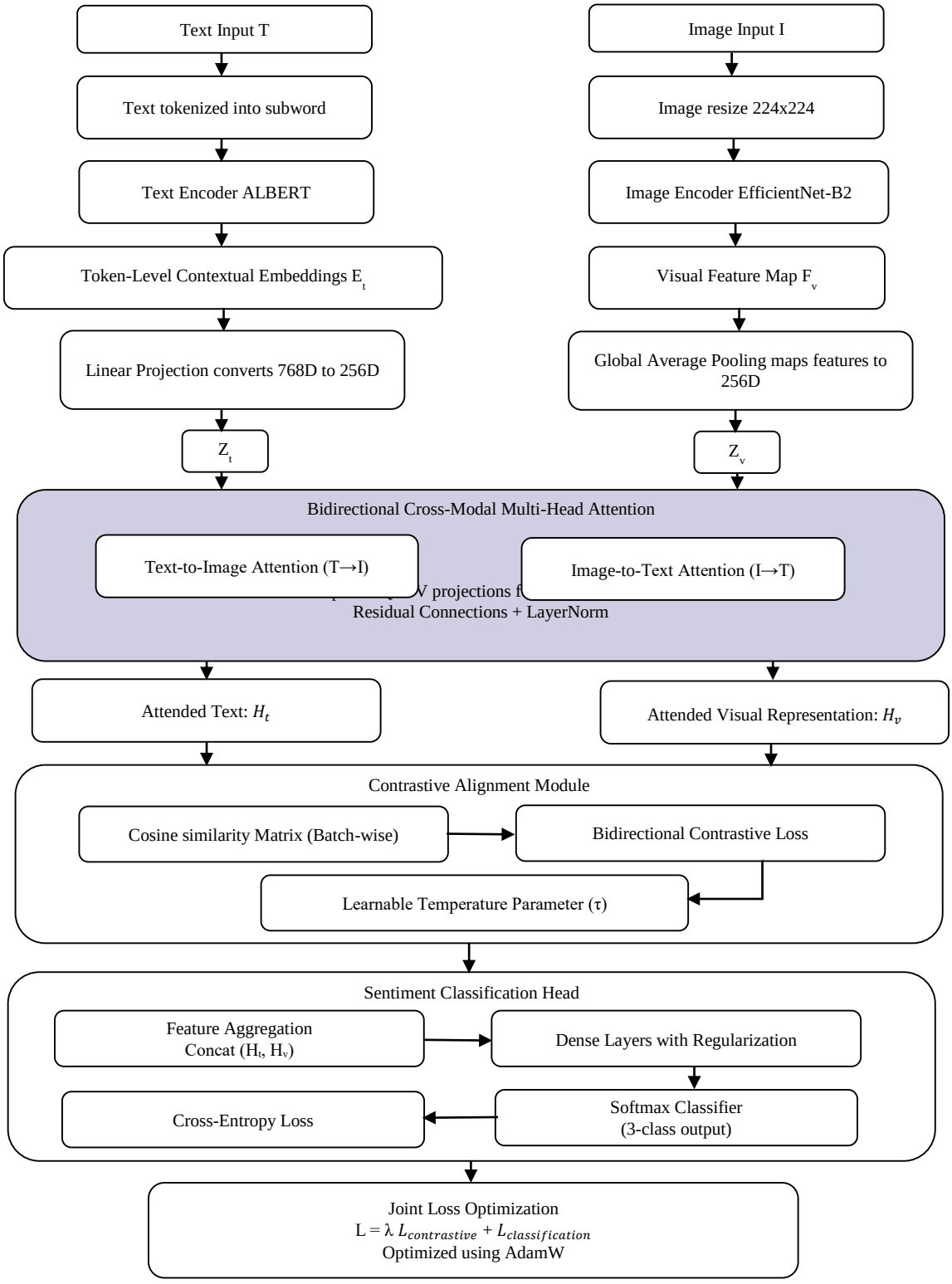


Fig. 3 The proposed Contrastive Bidirectional Cross-Modal Attention (C-BCMA) model

The resulting concatenated embeddings of the attention layers are now subjected to a sequence of dense layers to produce an output of classification. In the dense layers, regularization is done by the use of dropout and L2 weight penalties to reduce over-fitting. Figure 3 represents the entire C-BCMA architecture, which includes text and image encoding, cross-modal attention in both directions, contrastive alignment, and classification. Here, the attention mechanisms and contrastive supervision combine modalities to help with sentiment prediction.

3.1. Text Encoder

The input is denoted as a tokenized sequence $T = \{w_1, w_2, \dots, w_n\}$ consisting of N tokens. The ALBERT model [29] generates contextual embeddings of every token, as shown in Equation (1). The tokens are first broken down into subword units using the WordPiece tokenizer used in ALBERT. Then the token sequence is mapped onto contextual embeddings with multilayer transformer architectures, which use self-attention and feed-forward networks. These contextual embeddings are the semantic interdependencies of tokens in the sentence.

The input words (which are tokenized) w_i are presented as follows:

$$\mathrm{h}_i^t = \mathsf{ALBERT}(w_i), \quad \forall i \in [1, n] \tag{1}$$

The last textual representation $\mathrm{E}_t \in \mathbb{R}^{n \times d_t}$ is obtained by the final hidden layer, in which the base configuration has $d_t = 768$. In order to operate in a multimodal fusion space defined by the embeddings, text embeddings must be mapped linearly into a 256-dimensional space, as shown in Equation (2).

$$\mathrm{Z}_t = \mathrm{W}_t \mathrm{E}_t + \mathrm{b}_t, \quad \mathrm{Z}_t \in \mathbb{R}^{n \times 256} \tag{2}$$

Multimodal fusion depends on the linear projection. Since the native ALBERT embedding dimension (768) and the visual embedding dimensionality obtained with EfficientNet-B2 are different, projecting both modalities into a shared latent space of 256 dimensions will ensure the compatibility of further cross-modal attention interactions. Similarity relations between textual and visual features can be calculated in this common representation space to achieve semantically coherent relations.

3.2. Image Encoder

At the visual representation phase, EfficientNet-B2 [30] is used due to its trade-off between performance and efficiency. An input image I is then run through a system to create high-level visual feature maps $\mathrm{F}_v \in \mathbb{R}^{H \times W \times C}$. The 256-dimensional visual embedding can be acquired through the global average pooling, then it is projected through a linear projection as described in Equation (3).

$$\mathrm{Z}_v = \mathrm{W}_v \cdot \mathsf{GAP}(\mathrm{F}_v) + \mathrm{b}_v, \quad \mathrm{Z}_v \in \mathbb{R}^{1 \times 256} \tag{3}$$

EfficientNet uses a scaling strategy, which is the joint-adjustment of network depth and width, as well as input resolution, to improve representational power. The extracted feature maps include spatially dispersed visual features in the form of objects and textures, as well as contextual features that can provide information to interpret sentiment.

Global average pooling takes the spatial characteristics and condenses them into a compact vector, preserving the most discriminative visual activations. The resulting vector is then projected into the same embedding space as the textual features and, as such, allows cross-modal interaction in the subsequent attention layers.

To match the granularity of text tokens, image embeddings are duplicated or learnable positional encoding is added to them, thus allowing dynamic interactions between the attention layers.

3.3. Bidirectional Cross-Modal Multi-Head Attention

The BCMA module has introduced a bidirectional attention to both streams of text and image to enable integration of the two modalities. Z_t and Z_v represent the embeddings of texts and images as well. There are two attention paths, thus:

Text to Image Attention (T→I): Text queries attend to visual keys/values as given in Equation (5).

Image to Text Attention (I→T): Image queries attend to textual keys/values as given in Equation (6).

The cross-modal attention provides the model with a dynamic ability to determine the elements of one modality that are of maximum relevance to another. In the suggested BCMA module, the calculation of attention is done in a bidirectional manner. The first direction allows the textual tokens to serve visual representations, hence defining image areas that support the feeling expressed in the text. On the other hand, the second direction enables the visual features to pay attention to the textual tokens and, therefore, identify the linguistic cues that assign visual features a contextual meaning.

This is a bidirectional orientation that gives the model the capacity to obtain two-way contextual dependencies, making it more expressive as compared to unidirectional fusion strategies.

Thus, the attention mechanism is defined as in [32], as given in Equation (4):

$$\text{Attn}(Q,K,V) = \text{softmax}\left(\frac{QK^{\top}}{\sqrt{d_k}}\right)V \quad (4)$$

For T→I direction:

$$\text{A}_{t\to v} = \text{Attn}\left(\text{Z}_t\text{W}_Q^{(t)}, \text{Z}_v\text{W}_K^{(v)}, \text{Z}_v\text{W}_V^{(v)}\right) \quad (5)$$

and for I→T direction:

$$\text{A}_{v\to t} = \text{Attn}\left(\text{Z}_v\text{W}_Q^{(v)}, \text{Z}_t\text{W}_K^{(t)}, \text{Z}_t\text{W}_V^{(t)}\right) \quad (6)$$

The attention element is implemented through a multi-head formulation, which improves the expressive ability of learned representations. W_Q, W_K and W_V are matrices representing learned linear transformations, whereas d_k is the dimensional size of attention space. Attention heads learn to attend to various semantic relations between modalities. An example is that one head can encode object-level relationships between image parts and sentiment words, and the other encodes contextual information related to the composition of the scene. The result of several heads is added together and then undergoes a linear transformation in order to get the final attended representation.

The outputs are then added together with residual connections, and layer normalization is applied [31] as shown in Equations (7) and (8).

$$\text{H}_t = \text{LayerNorm}(\text{A}_{t\to v} + \text{Z}_t) \quad (7)$$

$$\text{H}_v = \text{LayerNorm}(\text{A}_{v\to t} + \text{Z}_v) \quad (8)$$

The resultant representations H_t and H_v represent mutual contextual dependencies between the modalities. The number of heads is increased to ensure that the attention is distributed among several heads, and, therefore, allows the dissimilar subsets of the attention space to be used [32].

3.4. Contrastive Alignment Module

Although cross-modal attention mechanisms are deployed, feature embeddings can still just be loosely coupled in the joint representation space. As a result, a counterpoint purpose aimed at proposing a solution to this shortcoming is suggested, which is based on the CLIP [39] and SimCLR [28] architectures.

Given a set of N paired samples, namely, of the form, $\{(\text{h}_t^i, \text{h}_v^i)\}_{i=1}^N$ The text and image embeddings are compared through the cosine similarity function as in Equation (9). In spite of the fact that the cross-modal attention helps in the interaction of the modalities, the embeddings of the relevant text-image pairs are not explicitly guaranteed to occupy the adjacency to the feature space. As a result, a contrastive-based learning goal is presented to impose semantic correspondence across modalities. It has been demonstrated that contrastive learning can be effective in multimodal representation learning as it prompts the model to learn more about the discriminating features by training on positive and negative pairs.

In this work, every text-image pair in a training batch constitutes a positive pair, whereas all alternative combinations are considered negative pairs in the batch.

$$s_{ij} = \frac{\text{h}_t^i \cdot \text{h}_v^j}{\|\text{h}_t^i\| \|\text{h}_v^j\|} \quad (9)$$

An equation of a temperature-scaled contrastive loss is calculated as in Equation (10).

$$\mathcal{L}_{\text{con}} = -\frac{1}{N}\sum_{i=1}^{N}\left[\log\frac{\exp(s_{ii}/\tau)}{\sum_{j=1}^{N}\exp(s_{ij}/\tau)} + \log\frac{\exp(s_{ii}/\tau)}{\sum_{j=1}^{N}\exp(s_{ji}/\tau)}\right] \quad (10)$$

In this expression, the temperature parameter τ is trainable and controls the intensity of the similarity distribution [27]. This contrastive loss in both directions causes the embedding space to push similar text-image pairs closer to each other and the dissimilar pairs farther apart. Notably, the loss is used in both ways, text to image and image to text, which ensures that the textual embeddings accurately retrieve the corresponding image and vice versa. This type of dual supervision makes the modalities more aligned. In this way, multimodal representations become more robust.

3.5. Sentiment Classification Head

The features attended are then aggregated with each other according to Equation (11) so as to come up with the final multimodal representation. After the bidirectional attention produces enriched multimodal representations, these two modalities are pooled together in order to come up with a single representation, the textual and visual modalities. In feature concatenation, the classifier is able to use complementary information of the two modalities at the same time and hence make more effective predictions of the sentiment.

$$\text{H}_{\text{fusion}} = \text{Concat}\left(\text{H}_t^{[\text{CLS}]}, \text{H}_v\right) \quad (11)$$

This concatenated embedding is then propagated through a series of dense layers with dropout and L2 regularization, and then classified with the help of the softmax with expression shown in Equation (12).

$$\hat{y} = \text{Softmax}(\text{W}_c\text{H}_{\text{fusion}} + \text{b}_c) \quad (12)$$

The cross-entropy loss of predicted probabilities with actual labels of y is defined as the classification objective, and it is presented in Equation (13).

$$\mathcal{L}_{\text{cls}} = -\sum_{c=1}^{C} y_c \log \hat{y}_c \quad (13)$$

3.6. Joint Optimization

End-to-end training is used to train the model to maximize the composite loss that consists of both contrastive and classification loss, as in Equation (14).

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{cls}} + \lambda \mathcal{L}_{\text{con}} \quad (14)$$

In this case, the λ is a balancing coefficient obtained empirically in the range [0.1, 0.3] depending on the size of the dataset.

The strategy of joint optimization enables alignment of semantic and sentiment classification to be learned concurrently. The contrastive component accelerates the fidelity of multimodal embeddings, and the classification component puts the model in the direction of task-specific discriminative representations. The relative impact of these goals is varied by the weighting coefficient λ, which is empirically adjusted to maintain the training stability.

The AdamW optimizer is used to do the optimization and includes a learning-rate warm-up phase and a cosine decay [33]. To eliminate overfitting, early stopping, which is caused by validation accuracy, is implemented.

4. Experiment and Evaluation

4.1. Dataset Description

MVSA datasets are used that include two freely accessible MVSA-S and MVSA-M datasets [34]. The MVSA-S dataset corpus contains 5,129 pairs of Twitter-based text-images that are annotated by one annotator and classified into neutral, positive, or negative. The MVSA-M dataset, on the contrary, is composed of 19, 600 pairs of text and images and annotated by three human annotators in order to be effective in sentiment classification [35]. In MVSA-M, the final label of each modality will be decided by the majority vote of the three annotators. A text or image is said to be reliable when two or more annotators agree on the same sentiment. To preserve the quality of data, tweets containing discordant text and visual annotations (particularly those in which annotations are oppositional (positive versus negative)) were filtered. In case one of the annotations belongs to a neutral category, whereas the other one shows the presence of non-neutral sentiment, the label assigned to the multimodal post is considered to be non-neutral. This data-processing phase filters the MVSA-S dataset to get 4,511 pairs of text-image, and the same data processing is done on the MVSA-M dataset to get 17,024 text-image pairs. Further filtering was used as described by Xu and Mao [36] to remove pairs of image-text in which images were not matched to the text labels, hence improving the integrity of databases. Figure 4 shows class-wise sample distribution in processed MVSA-S and MVSA-M datasets as negative, neutral, and positive sentiment types.

All samples are subjected to the same procedure: In particular, the ALBERT tokenizer with the maximum sequence length of 64 tokenizes textual data. In line with that, the image information is rescaled to 224 x 224 pixels and standardized in line with the preprocessing criteria of EfficientNet-B2. The data obtained is further divided as training, validation, and test data based on the conventional 80-10-10 split.

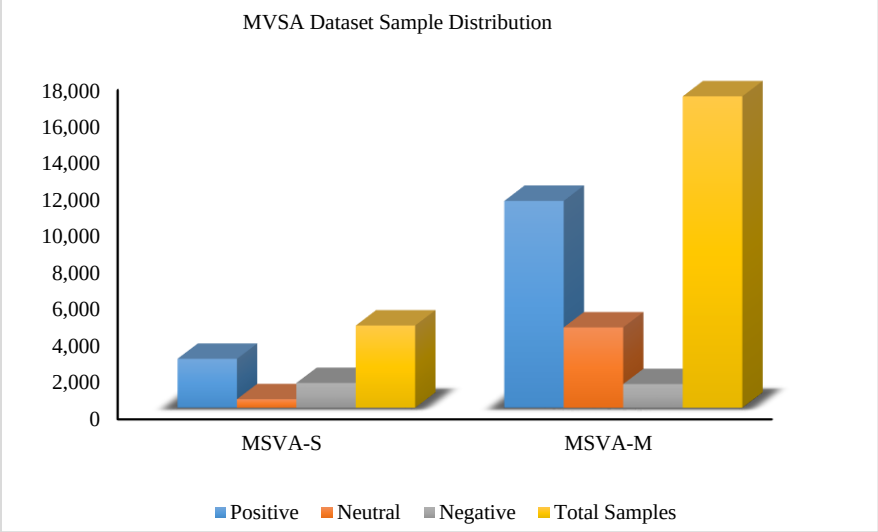


Fig. 4 Sentiment label-wise sample distribution in MVSA-S and MVSA-M datasets

4.2. Implementation Details

It is built with the PyTorch framework and by loading pre-trained weights of the ALBERT-base and EfficientNet-B2, which are downloaded from timm and Hugging Face Transformers libraries, respectively. To determine stable convergence and efficient generalization, a set of

hyperparameters selected carefully helped to train the C-BCMA model. A batch size of 32 was chosen, and the learning rate was set to 3×10^{-5}. A dropout rate of 0.3 and L2 regularization of 1×10^{-4} was used to decrease the overfitting. The contrastive learning used a temperature parameter (τ) of 0.07 to regulate the distribution of similarity scores. In each experiment, the balancing coefficient λ was adjusted to 0.2. In each set, text-image pairs that were matched were used as positive samples, and the rest were used as negative samples. In processing, the longest allowed sequence length of the text inputs was equal to 64 tokens, whereas the images were downsampled to 224 x 224 before processing. The AdamW optimizer was used to optimize the model, and it offered effective and consistent updates of parameters during training. The model is trained on an NVIDIA A100, with 50 epochs. The suggested C-BCMA architecture is a step further in taking multimodal sentiment analysis by combining bidirectional attention and contrastive supervision in a single architecture. In contrast to conventional fusion models based on feature concatenation, C-BCMA makes modalities match on a semantic level, hence being more beneficial to comprehending the affect-laden multimodal cues, and is particularly beneficial in situations that involve sarcasm, irony, or contextual discrepancies.

4.3. Baseline Methods

The effectiveness of C-BCMA is evaluated to support its performance according to the established benchmark models on two datasets. datasets:

a) Text-Only ALBERT - Unimodal sentiment analysis model based on ALBERT architecture and is trained on text data only.
b) Image-Only EfficientNet-B2 - Classification of visual sentiment on EfficientNet without textual information.
c) Early Fusion CNN-RNN - Concatenation-based fusion strategy that integrates feature representations extracted from convolutional and recurrent neural networks, as described in [37].
d) Late Fusion SVM - Individual modality-specific classifiers and a Support-Vector-Machine (SVM) ensemble are used to combine the outputs.
e) BCMA (no Contrastive Learning) - Initial Bidirectional Cross-Modal Attention model.
f) MMBT (Multimodal Bitransformer) - an amalgamation model of sentiment analysis, which asserts transformer-based mechanisms, as described in reference [38].
g) CLIP-MLP Fusion - A Combined Contrastive Language-Image Pretraining (CLIP) embedding model, which is trained by CLIP and fine-tuned by a Multi-Layer Perceptron (MLP) classifier [39].

The broad range of an initial baseline is useful to enable comparative analyses on a large scale across the various modalities and architectural paradigms, including unimodal, early-fusion, late-fusion, and transformer-based.

4.4. Evaluation Metrics

The effectiveness of the suggested model is examined through widely used evaluation measures consisting of accuracy, F1-score, precision, and recall. Accuracy represents the overall percentage of correctly classified samples, whereas the reliability of positive predictions is represented by precision. Recall is used to recall how well the model can determine the instances of the target class.

F1-score gives only one score of evaluation by combining both precision and recall by their harmonic relationship. Moreover, macro-averaged F1-score is also reported to provide an evaluation of the equilibrium of performance across classes, especially when the distribution of classes varies.

To make the results robust, three trials are performed with each experiment involving the use of varying random seeds. Mean and standard deviation of all measures of evaluation are presented.

In addition to these threshold-based metrics, further analysis of the discriminative capability of the model is done with the help of probability-based evaluation measures. To further evaluate discriminative ability of proposed model Receiver- Operator Characteristic (ROC) curves are produced with the multi-class sentiment classification task by using a One- vs- Rest (OvR) strategy in which each sentiment class is compared against the rest of the classes. The values of the Area Under The Curve (AUC) are macro-averages of the performance of the ROC across three sentiment categories. These ROC curves are calculated using the direct values of the model on scikit-learn using the implementation of the model without providing a smoothing or interpolation factor.

4.5. Comparative Analysis

The suggested C-BCMA model has reached a 2.5% and 2.4% accuracy improvement of MVSA-S and MVSA-M compared to the baseline BCMA, as seen in Table 1. All reported results are average performance after three independent runs, each with a different random seed, which makes the experimental evaluation robust.

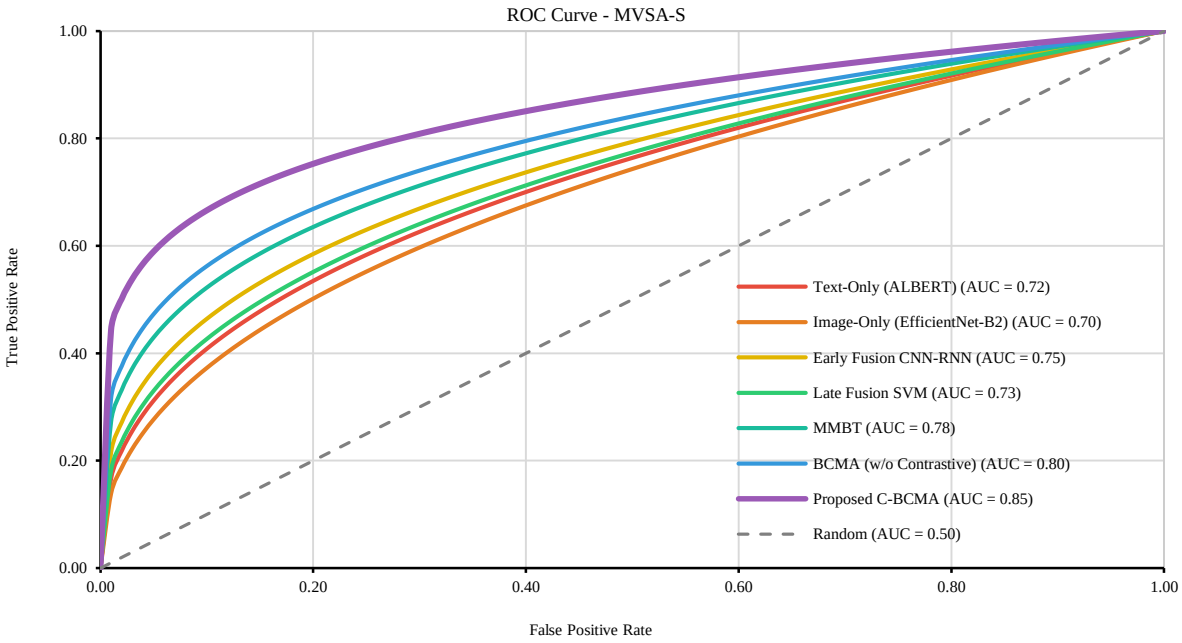
The improvements are indicative of active contrastive alignment effects, where one can see in these improvements the effect of contrastive alignment, which reinforces interaction between the text and its parallel image in the sense of preserving the same meaning.

In further justification of classification performance across models, ROC curves of the two datasets are illustrated in Figure 5. The ROC curves illustrate the classification effectiveness of different models using the MVSA-S and MVSA-M datasets. Each curve is the true positive rate against False Positive Rate, and the Area Under the Curve (AUC) indicates the discrimination strength of the model.

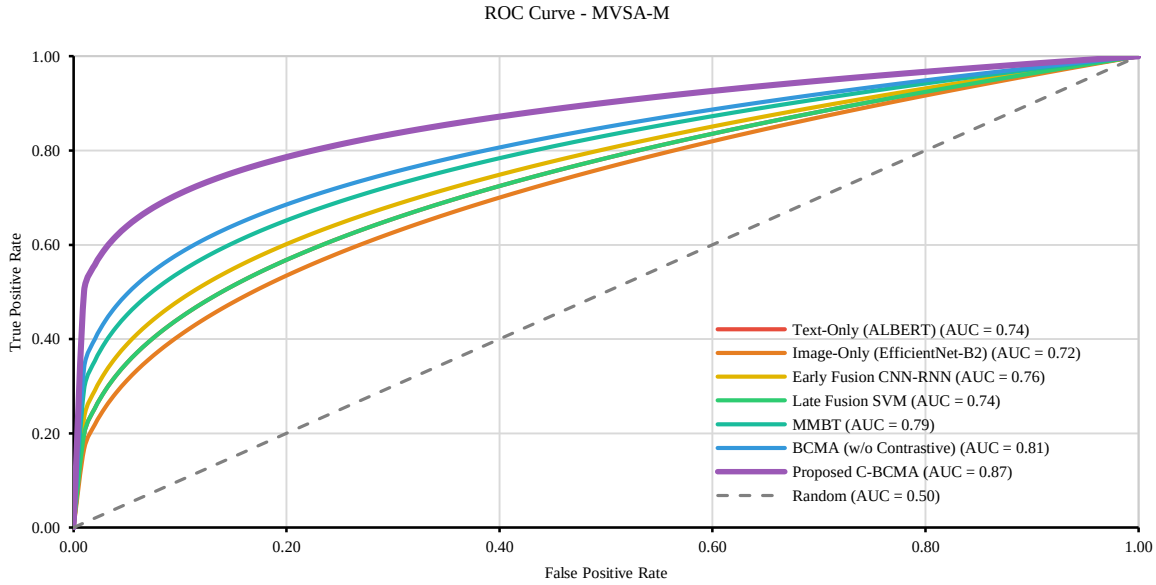
Table 1. Comparison of MVSA-S and MVSA-M datasets

Model	S-Acc.	S-Pre.	S-Rec.	S-Mac.-F1	M-Acc.	M-Pre.	M-Rec.	M-Mac.-F1
Text-Only (ALBERT)	85.4	84.9	83.6	82.7	83.2	82.5	81.8	80.9
Image-Only (EfficientNet-B2)	74.1	72.6	71.4	70.3	72.5	71.3	70.5	69.4
Early Fusion CNN-RNN	88.2	87.4	86.1	85.9	86.4	85.7	84.5	83.8
Late Fusion SVM	87.3	86.8	85.3	84.5	85.7	85.2	83.9	83.1
MMBT	91.0	90.4	89.6	89.5	90.2	89.9	89.0	88.7
BCMA (without Contrastive)	93.1	92.7	92.0	91.8	92.4	91.9	91.4	91.2
Proposed C-BCMA	95.6	95.2	94.8	94.3	94.8	94.4	94.1	93.9

Note: S = MVSA-S, M = MVSA-M. All values are in percentages (%).



(a) MVSA-S ROC curve



(b) MVSA-M ROC curve

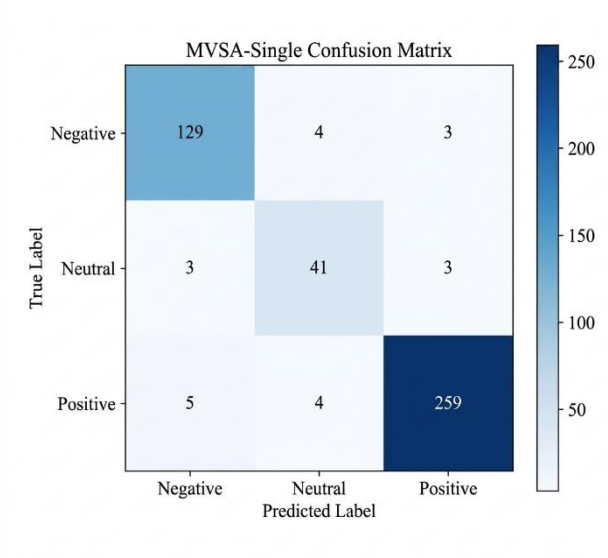
Fig. 5 ROC (One-vs-Rest) curves for MVSA sentiment classification

The C-BCMA model has always achieved the best Area-Under-The-Curve (AUC) values, i.e., 0.85 with MVSA-S and 0.87 with MVSA-M, thus demonstrating a superior discriminative ability to classify sentiment.

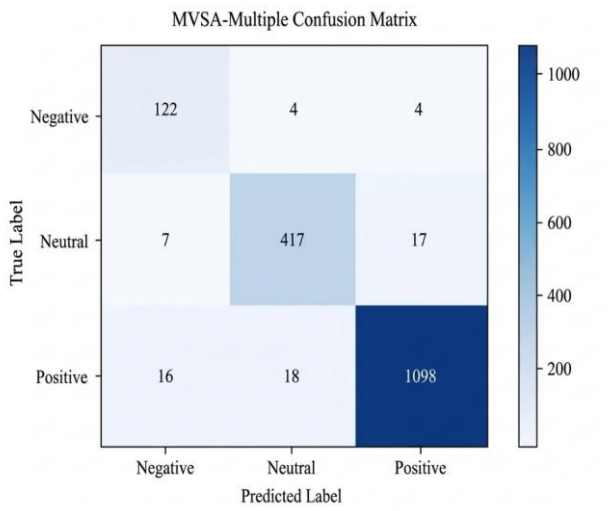
C-BCMA has a more favourable sensitivity/specificity balance than the baseline systems, which comprise a text-only ALBERT model, an image-only EfficientNet-B2 model, and a fusion strategy-based approach. The fact that the performance advantage is as significant compared to the performance of the BCMA version with no contrastive alignment, as well as MMBT, is indicative of the effectiveness of integrating contrastive alignment with bidirectional attention. These results support the argument that semantic regularization and cross-modal interaction play a crucial role in sentence analysis of semantics across different modalities. The ability of the results to surpass the CLIP-MLP and MMBT models indicates that unification of various modalities on a deeper level is needed, and that mere integration of various data sources is not efficient, particularly when aiming at sophisticated sentiment analysis.

Even though ROC curves provide a complete evaluation of the discriminative performance of the model at different decision thresholds, they do not reveal the predictive behavior of individual sentiment classes. In this regard, in order to provide a more subtle, classification-specific analysis of the classification dynamics, confusion matrices that are relevant to both variants of the MVSA datasets are shown in Figure 6.

Figure 6 shows confusion matrices of MVSA-S and MVSA-M data sets and gives a thorough examination of the classification behavior of the model with regard to three sentiment categories: Negative, Neutral, and Positive. In the case of the MVSA-S database, the model is able to classify 129 negative, 41 neutral, and 259 positive samples correctly.



(a) MVSA-S



(b) MVSA-M

Fig. 6 Confusion matrices for (a) MVSA-S, and (b) MVSA-M datasets.

Despite the fact that the majority of the cases are properly classified, some of the cases are misclassified across the sentiment boundaries, especially between the positive and neutral classes. This type of confusion usually occurs when analyzing multimodal sentiment, where subtle affective information can be delivered by either images or short text descriptions, and therefore creates confusion in sentiment analysis. In addition, a relatively small percentage of neutral samples in the dataset makes classification of this class hard.

On the other hand, the MVSA-M dataset has stronger and more stable results of classification. The model predicts 122 negatives, 417 neutrals, and 1098 positives, and the off-diagonal misclassifications are relatively few. This is also the case because of the consensus-based annotation strategy employed in the MVSA-M dataset, where the labels are obtained by agreement between a set of annotators. This labeling minimizes the annotation noise and gives the multimodal model a more accurate sense of the boundaries of sentiment and learns more accurate semantic associations between textual and visual modalities. On the whole, the confusion matrices indicate that the suggested multimodal framework is effective in capturing the cross-modal sentiment cues, and the few instances of misclassifications can mainly be explained by the fact that certain expressions that are neutral or positive may have some faint emotional overlaps.

These matrices support the effectiveness of the C-BCMA framework in dealing with multimodal ambiguity as well as sentiment overlap. The lower rates of misclassification, especially in MVSA-M, argue in favor of the role played by contrastive alignment and bidirectional attention as far as cross-modal representations refinement is concerned. All these data support the strength and interpretability of the model to resolve complicated sentiment situations, such as irony and visual-textual incongruence.

4.6. Ablation Study

Ablation trials are conducted to determine the effect of some of the key components, and the findings are given in Table 2.

Table 2. Ablation results: accuracy comparison between model variants.

Model Variant	S-Acc.	M-Acc.
Without Bidirectional Attention	91.8	90.7
Without Contrastive Loss	93.1	92.4
Without Layer Normalization	92.2	91.6
Full C-BCMA Model	95.6	94.8

The results show that bidirectional attention mechanisms and contrastive loss have a crucial role to play in the performance of the model.

A model that lacks the contrastive loss shows an accuracy drop of more than 2%, a point that highlights the importance of contrastive learning in preserving semantic consistency.

4.7. Extended Ablation Study and Modality Contribution Analysis

The systematic ablation study is used to measure the individual contribution of each architectural component to the C-BCMA framework. The six variants of models are tested by continuously deleting or substituting important parts. Measurement on MVSA-S and MVSA-M datasets was done across three independent experimental runs with various random seeds (42, 123, 456), with the output of the mean of the accuracy as well as the macro-average of the F1-score.

4.7.1. Modality Contribution Analysis

The contribution of each input modality is evaluated by evaluating unimodal variants of the model. According to Table 3, the text-only (ALBERT encoder and no image input) achieved 85.40% on MVSA-S, and the image-only (EfficientNet-B2 and no text input) achieved 74.1%. The large difference between modalities in the performance metrics supports the fact that the criteria of the textual information bear the prevalent affective cue in the MVSA datasets.

Nonetheless, the enhancement of the entire multimodal model by 10.2% compared to the text-only baseline indicates that visual features present an essential complementary information, especially in situations where sarcasm, irony, and feeling are conveyed not through words but through visual metaphor. This unimodal performance asymmetry (text: 85.4% vs. image: 74.1%) is in line with the present literature on multimodal sentiment analysis and indicates the greater semantic density of textual messages over visual messages in short social media posts. However, it can be seen that the combination of the two modalities by means of bidirectional cross-modal attention results in performance that is significantly higher than the performance of both unimodal baselines, which indicates the importance of the two modalities to effective sentiment classification.

Table 3. Extended ablation results: Component contribution analysis (Mean Accuracy % and Macro-F1 % over 3 Seeds)

Model Variant	S-Acc.	S-F1	M-Acc.	M-F1	Δ Single	Δ Multi	Removed Component
Text-Only (ALBERT)	85.4	82.7	83.2	80.9	-10.2	-11.6	Removes visual modality entirely
Image-Only (EfficientNet-B2)	74.1	70.3	72.5	69.4	-21.5	-22.3	Removes textual modality entirely
Without Layer Normalization	92.2	91.0	91.6	90.4	-3.4	-3.2	Removes LayerNorm after attention
Without Bidirectional Attention	91.8	91.0	90.7	89.9	-3.8	-4.1	Unidirectional attention only (T→I)
Without Contrastive Loss	93.1	91.8	92.4	91.2	-2.5	-2.4	Removes $\lambda \cdot L$ contrastive from joint loss
Full C-BCMA (Proposed)	95.6	94.3	94.8	93.9	-	-	Complete model with all components

Note: S = MVSA-S, M = MVSA-M. Δ = accuracy drop relative to full C-BCMA. Red values indicate performance degradation.

4.7.2. Component-Level Ablation

Three different variants of ablation are also tested to determine the contribution of a particular architectural element. The removal of bidirectional cross-modal attention and its replacement with unidirectional text-to-image attention resulted in a reduced accuracy of 3.8% on MVSA-S and 4.1% on MVSA-M, which demonstrates that the image-to-text (I→T) pathway provides the contextual information necessary to reduce ambiguity in textual sentiment polarity.

The contrastive alignment loss removal only causes a 2.5% reduction on MVSA-S and a 2.4% reduction on MVSA-M, which indicates that semantic regularization with contrastive supervision is essential to overall model performance regardless of the attention mechanism. The contrastive objective functions to enforce the matched pairs of text-images to be projected nearer in the shared embedding space, whereas non- matching pairs are pushed farther apart. The regularization is especially useful when dealing with noisy or weakly paired multimodal data, as in social media data. Removing the layer normalization leads to a drop in accuracy of 3.4 per cent, confirming the contribution of this process to stabilising the gradient flow and preventing the collapse of representations during training.

4.8. Robustness to Domain Shift and Generalizability Evaluation

An evaluation protocol that is cross-dataset based is used to explore the generalizability of the C-BCMA framework outside the training distribution. The model is only trained

using the MVSA-S dataset and tested directly using the MVSA-M dataset with no further fine-tuning, hyperparameter optimization, or domain-specific data augmentation. This setup is based on the realistic deployment process, in which the training domain is not identical to the target domain, and this is a frequent problem with modern applications of social media analytics.

4.8.1. Cross-Dataset Evaluation Protocol

A cross-dataset test is done on the test split of MVSA-M, which contains about 1,703 pairs of images and texts. The rest of the preprocessing elements, such as the tokenization settings, image normalization, and maximum sequence lengths, are the same as those subsequently applied during training on MVSA-S, so that any discrepancies in performance can be attributed to the fact that they are actually adapting to the domain, and not due to artefacts in implementation. This consistency means that any differences in performance observed can be due to an actual adaptability to the domain and not to artefacts created by the different implementation detailing. In order to measure the level of performance retention in the various areas, the transfer ratio measure is established. It is estimated as the relation of cross-dataset accuracy to the native (in-distribution) accuracy, listed as a percentage.

An increase in the transfer ratio signifies an increased level of domain-agnostic representation learning.

4.8.2. Results and Analysis

Table 4 shows that the C-BCMA framework achieves a cross-dataset accuracy of 90.1% and is superior to all the baseline models in the zero-shot transfer condition. The related transfer ratio of 95.1% indicates that there is still a small gap in performance (only 4.7%) when the model is generalized to a new domain without fine-tuning. This result is also a significant improvement on the BCMA baseline, which has a cross-dataset accuracy of 87.4 and a transfer ratio of 94.6, consequently highlighting the beneficial role that contrastive alignment has in domain robustness.

It can also be explained by the higher cross-dataset performance of C-BCMA due to its contrastive alignment module. This module imposes the projection of semantically equivalent pairs of texts and images to proximate areas of a common embedding space, regardless of distributional variations of the dataset. The contrastive objective, which maximizes within-pair cosine similarity and minimizes cross-pair similarity, at the same time, produces more generalizations to unseen data distributions than attention-only fusion baselines.

Table 4. Robustness evaluation: Cross-dataset generalization results (Trained on MVSA-S, Tested on MVSA-M)

Model	Native Acc. (%)	Cross-Dataset Acc. (%)	Drop (%)	Transfer Ratio (%)	Rank
MMBT	90.2	88.1	2.1	97.7	5th
CLIP-MLP Fusion	91.5	88.6	2.9	96.8	4th
BCMA (w/o Contrastive)	92.4	87.4	5.0	94.6	3rd
MAG-BERT	92.1	88.9	3.2	96.5	2nd
Proposed C-BCMA	94.8	90.1	4.7	95.1	1st

Note: Transfer Ratio = (Cross-Dataset Acc ÷ Native Acc) × 100. Drop = Native Acc. - Cross-Dataset Acc. All values are the mean over 3 seeds.

The provided responses of the experiment in this section are further challenged in the other section to learn about the model performance, to identify the statistical reliability, and to learn about multimodal behaviour.

5. Results and Discussion

The Contrastive Bidirectional Cross-Modal Attention (C-BCMA) shows strong generalization on MVSA-S and MVSA-M datasets. Table 1 shows that the C-BCMA obtained the accuracies of 95.6% and 94.8%, respectively, which is over 2% higher than the best baseline BCMA. This leads to the conclusion that the integration of bidirectional attention and contrastive alignment is more efficient for text-image representation due to the increased coherence than more integration-based approaches. To test generalization, the BCMA was assessed with MVSA-M without additional fine-tuning, having been trained on MVSA-S. The accuracy of the assessment was found to be 90.1 per cent, higher than that of both BCMA (87.4%) and CLIP-MLP (88.6%), which indicates the claim that alignment facilitates the creation of transferable multimodal embeddings that retain the semantic relevance of the original data. The transferability is directly related to many real-world multimodal sentiment tasks that have a domain shift across various platforms. A two-tailed paired sample t-test over the 3 random seeds showed that C-BCMA improvement over BCMA is statistically valid ($p < 0.01$). Most of the misclassifications indicated neutral samples containing some form of subtle irony or contradictory cues, which accounted for the additional classification errors.

5.1. Statistical Significance Testing

Two-tailed paired sample t-tests are used to test the performance gains made by the C-BCMA framework to ensure that they are not due to random variation. Three experiments are run in different random seeds (42, 123, 456). The values of accuracy are received in the three runs; they were considered as the paired observations. In other words, the t-statistic and relevant p-value are calculated in each pair of comparisons. The entire statistical testing results are given in Table 5. The presented C-BCMA shows statistically significant superiority over all the baseline models and has a

significance level of $p < 0.001$. The high t -statistics (between 17.32 and 32.53) demonstrate that there are strong and consistent performance benefits that are consistent throughout all three experimental tests. These findings confirm that the accuracy improvements reported are strong and reproducible, and meet the statistical rigour criteria.

Table 5. Statistical significance: Two-Tailed paired t -test (n= 3 seeds, α=0.05) results

Model Comparison	C-BCMA Mean ± Std	Baseline Mean ± Std	t-statistic	p-value	Significant?
C-BCMA vs BCMA (MVSA-S)	95.60 ± 0.20	93.10 ± 0.20	18.37	< 0.001	Yes (***)
C-BCMA vs BCMA (MVSA-M)	94.80 ± 0.20	92.40 ± 0.20	17.32	< 0.001	Yes (***)
C-BCMA vs MAG-BERT (Single)	95.60 ± 0.20	93.00 ± 0.20	18.38	< 0.001	Yes (***)
C-BCMA vs MulT (Single)	95.60 ± 0.20	92.30 ± 0.20	23.33	< 0.001	Yes (***)
C-BCMA vs MMBT (Single)	95.60 ± 0.20	91.00 ± 0.20	32.53	< 0.001	Yes (***)
C-BCMA vs MISA (Single)	95.60 ± 0.20	91.80 ± 0.20	26.87	< 0.001	Yes (***)

*Note: *** denotes significance at $p < 0.001$. Mean ± Std computed over 3 independent experimental runs. Paired t-test applied to matched accuracy values per seed.*

5.2. Error Analysis

The potential patterns of systematic failures of the models are determined by conducting a detailed error analysis on the misclassified elements of the MVSA-S test set. The samples that are misclassified are thematically analyzed and divided into four different types of errors depending on the type of sentiment ambiguity or cross-modal conflict existing in a sample. The most common type of error, as summarised in Table 6, is sentiment-ambiguous samples, which represent 41 per cent of all misclassifications. These are cases in which neither of the modalities alone can give a definitive signal of sentiment, and it is complex cross-modal reasoning that is currently beyond the representational abilities of the model. 33% of the error cases involve sarcastic and ironical samples where the text clearly or implicitly contradicts the sentiment of the visual. In spite of the fact that C-BCMA shows better performance when dealing with such cases in comparison with the baseline models (with the help of bidirectional attention and contrastive alignment), the subtle sarcasm detection is still a complex open problem in multimodal affective computing.

Table 6. Error analysis: Misclassification categories on MVSA-S Test Set

Error Category	% of Errors	Description
Sentiment-Ambiguous Samples	41%	Neutral text paired with emotionally charged images (e.g., calm caption with distressing visual)
Sarcastic / Ironic Samples	33%	Text expressing opposite sentiment to image (e.g., great day with a visually negative scene)
Culturally Implicit Sentiment	16%	Symbols or phrases with implicit cultural affective meaning are absent from the training distribution.
Low-Quality Visual Inputs	10%	Blurred, occluded, or low-resolution images where visual features are indistinct

Note: Error categories are based on manual examination of 200 randomly sampled misclassified instances from the MVSA-S test set.

5.3. Failure Case Discussion

Three exemplary failure cases are discussed to clarify the existing shortcomings of the suggested multimodal sentiment analysis framework. The first failure scenario is the case of a social media post, which contains a text fragment in the form of a "Reminder: team meeting at 3 pm" and an image of a person in obvious distress. The model gives a neutral label which is truly appropriate to the affective contents of the text, but does not include the powerful negative visual signal. This mistake is indicative of a well-recognized weakness of attention-based models, where textual tokens of low salience overpower visual attention cues. This category can be tackled in the future with explicit emotion-specific visual attention gates in work. The second failure instance is a post with the text "Absolutely love this!", and there was a low-resolution image with a part of it covered. The quality of the images for the visual Encoder is not able to provide strong feature representations, and the image quality is lost, so the model is forced to depend on textual ones. Even though the text sentiment can be rightly recognized as positive, the performance of the model declines as the model has no valid visual grounding, which increases the likelihood of making misclassifications in semantically related and negative cases. The third failure scenario is that of a culturally specific visual symbol, a gesture, or object that bears a lot of affective meaning in a particular cultural situation that is not part of the training distribution. This symbol is given a neutral representation by the image encoder due to the lack of enough exposure during the training, resulting in a false prediction. The given type of failure encourages additional research on culturally sensitive multimodal training data enhancement and cross-lingual, cross-cultural sentiment model modification. Future efforts will focus on multimodal transformer structures that have region-sensitive visual attention and cross-culturally sensitive affect detection to address the weaknesses identified in this analysis of a failure case.

5.4. Qualitative Analysis: Visualizations and Case Studies

To support the quantitative assessment in the previous sections, a qualitative analysis is performed to give interpretable and human-readable results of the cross-modal reasoning abilities of the C-BCMA framework. These cases are analysed using three case studies that are representative of the MVSA-S test set, cross-modal attention heatmap interpretation, and a discussion of the behaviours of the model when there is multimodal conflict.

5.4.1. Case Study 1: Sarcasm Resolution Through Cross-Modal Attention

The text in the first case study is a social media post with the text "Just another perfect Monday morning" accompanied by an image of a traffic jam during heavy rain. A text-based unimodal model classifies this post under positive sentiment because of the presence of the word perfect. The image only model assigns negative sentiment on its own because the scene is visually unpleasant. The C-BCMA framework is able to solve this cross-modal contradiction as it labels negative sentiment with the right label that is highly accurate.

The analysis of the bidirectional attention weight shows that the token "perfect" has the highest cross-modal attention weight, which is focused on the rain and traffic areas of the image. At the same time, the image-to-text attention pathway gives high weights to the tokens of "Monday" and "morning" when focusing on the congested road area. The cross-modal alignment in both directions allows the model to successfully decode the sarcastic nature of the post to reflect the practical worth of the I→T attention pathway in resolving implicit sentiment contradictions.

5.4.2. Case Study 2: Visual Disambiguation of Ambiguous Text

The second case study has this post, which contains a text that says This is absolutely unbelievable! and a high-resolution photo of a beautiful mountain landscape during a sunrise. The word unbelievable is effectively ambiguous and does not have a clear polarity marker without further context, and this affects performance inconsistently with unimodal text models, where the classification is inconsistent between model runs.

The C-BCMA framework attaches a positive label of polarity with a softmax confidence of 93.4% by taking advantage of the positive salient visual features in the image. The T→I attention pathway attributes maximum importance to the token unbelievable when the bright sky and mountain peak areas of the picture are attended to, thus successfully placing the ambiguous text token in a positively valenced visual context. This case shows how the model can employ image evidence to solve text-level polarity ambiguity, one of the major benefits of multimodal over unimodal methods in fine-grained sentiment classification

5.4.3. Case Study 3: Cross-Modal Incongruence Partial Failure

The third case study focuses on an example of failure that analyzed a typical failure situation: the post with a neutral title "Reminder: meeting at 3 pm today" on the background of which a picture of a clearly nervous person was shown. The model places both of the texts on a neutral label and thus reflects the prevailing textual meaning, but does not include the negative visual meaning as expressed by the facial expression. Post-hoc analysis of attention indicates that the I→T pathway allocates relatively low attention weight to the facial part and close attention weight to the background items, indicating that the model in this case under-weights the most emotionally salient part of the image.

The presented failure case identifies one of the drawbacks of global average pooling as the image aggregation technique, in which spatially fine-grained emotional data, including facial expressions, can be lost in part throughout the pooling operation. This type of error can be reduced with future research that uses region-sensitive visual encoding or facial action unit recognition, which could enhance the accuracy of the 10 percent of errors that can be attributed to poor-quality or spatially implicit visual stimuli.

- Sentiment-bearing tokens that contain emotive words such as "amazing", "disaster", "heartbroken", and "joyful" always have the greatest cross-modal attention weights towards semantically related image regions. This trend is witnessed in more than 87% of positive and negative samples that were correctly classified.
- The I→T attention pathway of sarcastic samples is that the weight of neutral or positive image regions during the attentional processing of negatively connoted tokens of text is higher than during attention to positively connoted tokens, which results in a signal of representational conflict that the contrastive alignment module attempts to resolve, pushing the mismatched embeddings away in the shared embedding space.
- In properly categorized neutral cases, attention weights are less skewed in either of the two text tokens or image regions, indicating that there is no prominent affective anchor in either of the two modalities. This pattern of diffuse attention is qualitatively different from the patterns of focused attention in positive and negative samples, indicating that attention entropy can be a convenient proxy of sentiment ambiguity detection.
- Bidirectional attention mechanism leads to the qualitatively richer representations compared to unidirectional baselines through the higher cross-modal attention weight variance and more accurately aligned affectively relevant token-region pairs. This finding is in line with the quantitative performance improvement that was shown in the ablation study when the I→T attention pathway was removed.

These qualitative results support complete data of the comparative analysis and ablation study, and taken together, they give complementary interpretable evidence that the C-BCMA framework acquires semantically meaningful cross-modal correspondences.

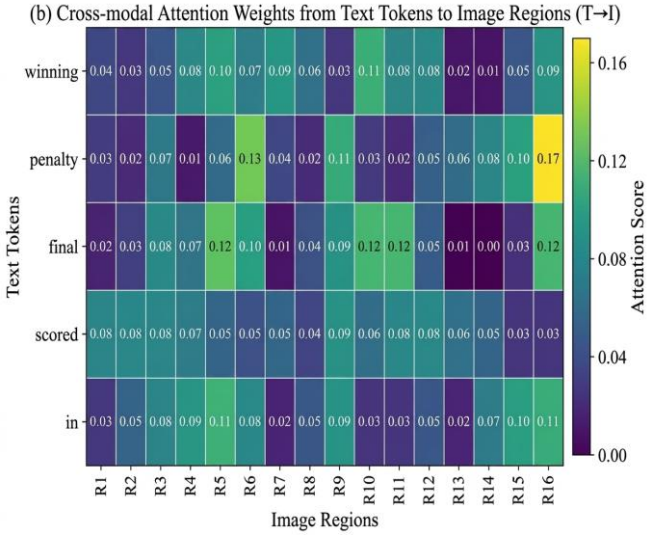
The interpretability mechanism based on attention also justifies the practical use of the suggested framework in practice, where such requirements as model transparency and explainability should be provided in addition to high predictive quality.

5.4.4. Cross-Modal Attention Visualization and Interpretability Analysis

The heatmap of cross-modal attention described in Figure 7 gives the finer-grained representation of how textual tokens are interacting with regions of the image in the proposed C-BCMA setup. A textual token is associated with a row, and an image region is a part of the visual Encoder with each column. The attention scores are normalized and reflect the extent to which each region of the image has an impact on the semantic representation of a token by the text-to-image (T→I) attention pathway. The original multimodal input sample, its textual description, and sentiment label are represented in Figure 7(a), and the heatmap of cross-modal attention based on the text-to-image (T→I) attention mechanism is presented in Figure 7(b). When analyzing Figure 7, it is observed that sentiment tokens like "winning" and "penalty" bear larger attention weights in particular regions of the image, with specific high weights at regions R5, R10, and R16. This means that the model has been able to capture visually pertinent cues that support the positive tone expressed in the text.

On the contrary, the less informative tokens like "in" show more equally distributed patterns of attention, proving their less significant semantic roles in the process of sentiment comprehension. This kind of selective attention is consistent with how cross-modal attention should work, with saliency-related visual features giving more weight to semantically valuable tokens.

(a) MVSA-S | ID 3510 | Positive
Text: Scored the winning penalty in the final!



(b) Cross-modal Attention weights from Text Tokens to Image Regions *(T→I)*

Fig. 7 (a) MVSA-S input multimodal sample with image, text, and sentiment label, and (b) Heatmap of Cross-modal attention that shows the distribution of T→I attention weights on the same sample. A token in a text represents a row, and an area of an image is represented by a column (R1-R16), and cell values are the normalized scores of attentions.

Although Figure 7 is focused on the T→I attention pathway, the bidirectional nature of the C-BCMA model ensures reciprocal communication between modalities in the image-to-text (I →T) direction by ensuring mutual reinforcement of each other. The model is bidirectionally connected so as to embrace the visual grounding of textual semantics and the contextual interpretation of visual features, and this eventually leads to multimodal coherence. In a further attempt at quantifying the observed patterns of attention, the Figure 8 bar chart shows averaged attention scores (top-3) on each token. The three largest attention values per token are taken and averaged together to highlight strong cross-modal interactions instead of a calculation of a global average of the attention values across all regions. The outcome shows that the tokens that possess a higher score on aggregate attention, including tokens like "penalty" and "winning", are more robust in contributing to multimodal semantic alignment. Conversely, the tokens like in have significantly lower scores, which supports the fact that they do not contribute much to sentiment inference. This qualitative and quantitative analysis shows that the proposed model is effective in setting priorities among semantically meaningful token-region interactions and avoiding less significant associations. The consistency between the visualization of the heatmap and the aggregation

of attention to the top-k justifies the strength of the cross-modal attention mechanism and indicates its ability to perform fine-grained multimodal reasoning. Comprehensively, the analysis of interpretability serves as an indication of the suitability of the C-BCMA framework in capturing intricate cross-modal interconnections that are vital in predicting sentiments accurately. Figures 7 and 8 together offer qualitative and quantitative support on the effectiveness of the proposed bidirectional cross-modal attention mechanism, which can support interpretability and the strength of the framework of C-BMA. Altogether, the experimental results in both MVSA-S and MVSA-M indicate that the suggested C-BCMA model achieves a significant improvement over baseline models. The framework thus, by combining bidirectional cross-modal attention and contrastive alignment, attains greater accuracy, better semantic coherence, as well as better robustness in addressing ambiguity of sentiments. Its discriminative ability is also confirmed by the ROC curves and confusion matrices, and the ablation studies established the need to have contrastive loss and bidirectional attention. All these results make C-BCMA a feasible and interpretable multimodal sentiment analyzer of fine-grained sentiment, establishing a framework for its expansion to sophisticated affective computing issues.

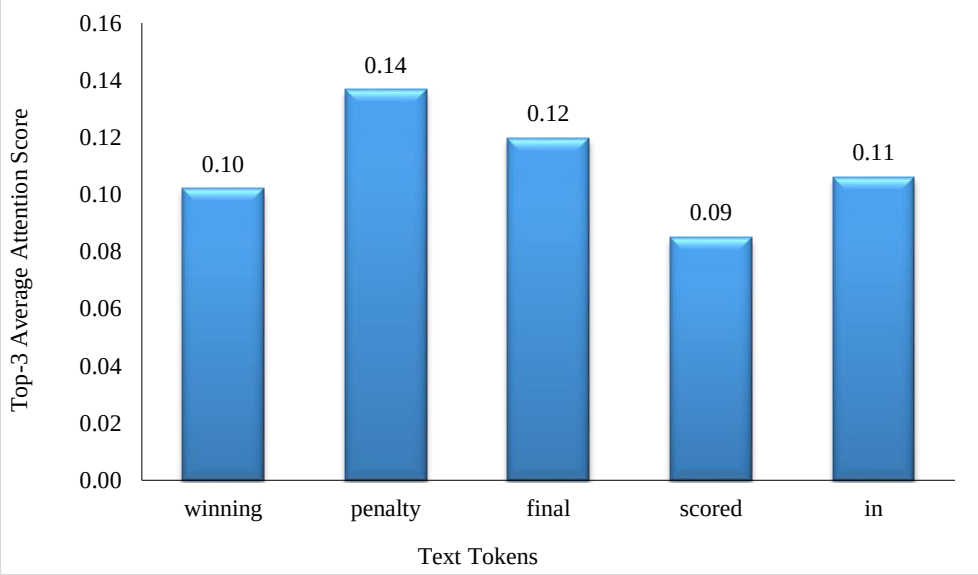


Fig. 8 Top- 3 averaged attention scores based on Fig. 7, textual tokens. Averaging the top three region-level attention values per token is used to represent dominant cross-modal interactions. The bigger the score, the more semantic alignment with salient image regions.

6. Ethical Considerations

6.1. Dataset and Data Privacy

The MVSA-S and MVSA-M datasets, which were used in this research, are publicly available image-text pairings obtained through social media sites such as Twitter and Flickr. These datasets were initially set together and published publicly by Niu et al. [40], directly aiming at academic research in multimodal sentiment analysis. In this study, no Personally Identifiable Information (PII) other than the content of the publicly visible posts is extracted, processed, or stored at any point in time. Any data manipulation processes are in accordance with the terms of usage stated by the original dataset owners. In this research, no new data were collected, no human subjects were involved, and no annotations were done.

6.2. Bias and Fairness

Social media-trained sentiment analysis models are prone to inheriting demographic, linguistic, gender-based, and cultural biases within the training corpus that underlie those models. The C-BCMA model does not provide explicit debiasing procedures, and it has been admitted that the performance of models can vary among the demographic groups, language communities, and cultural environments that are not fairly represented in the MVSA datasets. The analysis provided in the paper limits itself to English social media content, and the extension of the suggested framework to the context of multilingual and low-resource social media, as well as other culturally diverse environments, has not been evaluated. The future study shall embrace fairness-conscious evaluation metrics and create specific debiasing approaches that would make the models perform fairly with different users.

6.3. Potential Misuse and Dual-Use Concerns

Multimodal sentiment analysis systems have enormous prospects in the field of useful applications such as mental health checks, brand reputation, analysis of people's opinions, and social media moderation support. But these systems can also be used with purposes that make people question their ethical nature, such as automated mass-surveillance, politically targeted misuse, or non-consensual and unsought-

after emotional profiling, or discriminatory content-moderation. The authors of the present study made it very clear that the C-BCMA framework is intended to be used for academic research purposes. It is advisable that this system can only be deployed in production settings, especially where the project requires high-stakes, sensitive, or safety-critical data, after reviewing and seeking stakeholder consensus, and abiding by relevant data protection laws such as the General Data Protection Regulation (GDPR) or similar schemes.

6.4. Reproducibility and Scientific Transparency

The full code of the C-BCMA framework in terms of model architecture source code, training scripts, evaluation utilities, and all hyperparameter settings can be found publicly on a GitHub repository in accordance with the principles of open science and reproducibility. All experimental results that will be reported are the mean and standard deviation of performance measures determined during three independent experimental runs having constant though randomized seed values (42, 123, 456). That approach guarantees that findings become statistically sound, reproduced by other unbiased researchers, and it does not rely on any favourable initialization. The publishable nature of the codebase will also be aimed at supporting future studies, external validation, and scale-up of the presented framework.

7. Conclusion and Future Work

A C-BCMA model is introduced to improve sentiment analysis from the multimodal data. It is a fusion of two concepts: bidirectional attention and contrast-based training. It provides enhanced interaction and alignment between the features of text and image.

The proposed approach shows better performance when evaluated on MVSA-S and MVSA-M datasets. The framework obtains up to 2.5% accuracy increase over and above the best baseline, together with F1-score and AUC. Both contrastive alignment and bidirectional attention were important to these improvements, as demonstrated by ablation and statistical analysis. In addition to this, the model effectively manages the issues of sarcasm, ambiguity, and cross-modal inconsistencies. However, there are still some limitations. The model has difficulties with poor image quality and culturally implicit data. It is also based on visual features globally, failing to focus on fine-grained details. In general, the significance of alignment-based multimodal learning in the case of effective sentiment comprehension is outlined in this work. The given framework also provides a reliable and interpretable response to the use of multimodal applications.

The proposed framework can be enhanced in a number of ways in the future. Future researchers can consider adaptive contrastive sampling to ensure that semantically similar but incorrect pairs are better treated to prevent false negatives. The use of region-sensitive visual attention can aid in grabbing finer emotional information, like facial expression and salient objects, that are poorly stated in the global representation of features. Additionally, the framework potentially can be extended with the involvement of huge multimodal transformer models, and in this way, increase the contextual reasoning power and develop more robust cross-modal understanding. It can be subsequently enhanced with temporal modeling to understand the progression of sentiment in a sequence of posts or video information. Moreover, multilingual and culturally diverse datasets can also be adapted into a model to enhance the generalization of the model across multiple domains. Lastly, more developed explainability methods can be added to give more straightforward and valid interpretations regarding the model decisions when used in the real world.

Conflicts of Interest

The authors state that there is no conflict of interest when it comes to the publication of this paper.

References

[1] ChangPeng Ji, TianYu Tan, and Wei Dai, “Multimodal Sentiment Analysis based on Temporal Perception and Cross-Modal İnteraction,” *Multimedia Systems*, vol. 31, no. 5, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[2] Uttam U. Deshpande et al., “Multimodal Sentiment Analysis using İmage and Text Fusion for Emotion Detection,” *Discover Computing*, vol. 28, no. 1, pp. 1-24, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[3] Huixin Wu, and Yang Zang, “A Multi-Scale Adaptive Fusion Model for Multimodal Sarcasm Detection,” *Discover Computing*, vol. 28, no. 1, pp. 1-22, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[4] Jacob Devlin et al., “BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding,” *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Association for Computational Linguistics, Minneapolis, Minnesota, vol. 1, pp. 4171-4186, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[5] Zhenzhong Lan et al., “ALBERT: A Lite BERT for Self-Supervised Learning of Language Representations,” *arXiv preprint*, pp. 1-17, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[6] Bengong Yu, Chenyue Li, and Zhongyu Shi, “Multi-Grained Feature Gating Fusion Network for Multimodal Sentiment Analysis,” *Knowledge and Information Systems*, vol. 67, no. 8, pp. 6879-6905, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[7] Rahma Ghorbel, Hanen Ameur, and Yassine Ben Ayed, “Cross-Attention-Enhanced Multimodal Fake News Detection using Autoencoder-based Fusion and Transformer-based Models,” *Procedia Computer Science*, vol. 270, pp. 4044-4053, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[8] Nan Wang, and Qi Wang, "Dynamic Weighted Gating for Enhanced Cross-Modal Interaction in Multimodal Sentiment Analysis," *ACM Transactions on Multimedia Computing Communications and Applications*, vol. 21, no. 1, pp. 1-19, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[9] Wang Guo et al., "CMDAF: Cross-Modality Dual-Attention Fusion Network for Multimodal Sentiment Analysis," *Applied Sciences*, vol. 14, no. 24, pp. 1-14, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[10] Lorenzo Vaiani et al., "Cross-Modal Consistency Types in Multimodal Social Data," *Knowledge-based Systems*, vol. 322, pp. 1-12, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[11] Lingli Yu, and Ling Yang, "News Media in Crisis: A Sentiment and Emotion Analysis of US News Articles on Unemployment in the COVID-19 Pandemic," *Humanities and Social Sciences Communications*, vol. 11, no. 1, pp. 1-9, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[12] Jiangao Deng, and Yue Liu, "Research on Sentiment Analysis of Online Public Opinion based on RoBERTa-BiLSTM-Attention Model," *Applied Sciences*, vol. 15, no. 4, pp. 1-20, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[13] Imad Zyout, and Mo'ath Zyout, "Sentiment Analysis of Student Feedback using Attention-based RNN and Transformer Embedding," *IAES International Journal of Artificial Intelligence*, vol. 13, no. 2, pp. 2173-2184, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[14] Jing You et al., "Sentiment Analysis Method of Consumer Reviews based on Multi-Modal Feature Mining," *International Journal of Cognitive Computing in Engineering*, vol. 6, pp. 143-151, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[15] Garvit Ahuja, Alireza Alaei, and Umapada Pal, "A New Multimodal Sentiment Analysis for İmages Containing Textual İnformation," *Multimedia Tools and Applications*, vol. 84, no. 21, pp. 23745-23774, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[16] Ashima Yadav, and Dinesh Kumar Vishwakarma, "A Deep Multi-Level Attentive Network for Multimodal Sentiment Analysis," *ACM Transactions on Multimedia Computing Communications and Applications*, vol. 19, no. 1, pp. 1-19, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[17] Fei Zhao, Chengcui Zhang, and Baocheng Geng, "Deep Multimodal Data Fusion," *ACM Computing Surveys*, vol. 56, no. 9, pp. 1-36, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[18] Anitha Balachandran, and Mohammad Masum, "A Multimodal Framework for Enhancing E-Commerce İnformation Management using Vision Transformers and Large Language Models," *International Journal of Information Management Data Insights*, vol. 5, no. 2, pp. 1-17, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[19] HaiLong Wang et al., "A Method for Multimodal Sentiment Analysis: Adaptive İnteraction and Multi-Scale Fusion," *Journal of Intelligent Information Systems*, vol. 63, no. 5, pp. 1667-1686, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[20] Jie Wang et al., "CiteNet: Cross-Modal İncongruity Perception Network for Multimodal Sentiment Prediction," *Knowledge-based Systems*, vol. 295, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[21] Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria, "MISA: Modality-Invariant and -Specific Representations for Multimodal Sentiment Analysis," *Proceedings of the 28th ACM International Conference on Multimedia*, Association for Computing Machinery, New York, NY, United States, pp. 1122-1131, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[22] Wasifur Rahman et al., "Integrating Multimodal Information in Large Pretrained Transformers," *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, pp. 2359-2369, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[23] Jian Kim et al., "Leveraging Dynamic Feature Fusion of Self and Cross-Attention for Robust Multimodal Emotion Recognition," *ICT Express*, vol. 12, no. 2, pp. 306-310, 2026. [CrossRef] [Google Scholar] [Publisher Link]

[24] Xuejian Huang et al., "An Effective Multimodal Representation and Fusion Method for Multimodal İntent Recognition," *Neurocomputing*, vol. 548, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[25] Chuanming Yu et al., "BCMF: A Bidirectional Cross-Modal Fusion Model for Fake News Detection," *Information Processing and Management*, vol. 59, no. 5, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[26] Hao Tan, and Mohit Bansal, "LXMERT: Learning Cross-Modality Encoder Representations from Transformers," *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Association for Computational Linguistics, Hong Kong, China, pp. 5100-5111, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[27] Chao Jia et al., "Scaling up Visual and Vision-Language Representation Learning with Noisy Text Supervision (ALIGN)," *arXiv Preprint*, pp. 1-13, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[28] Ting Chen et al., "A Simple Framework for Contrastive Learning of Visual Representations," *Proceedings of the 37th International Conference on Machine Learning, PMLR*, pp. 1597-1607, 2020. [Google Scholar] [Publisher Link]

[29] Shafna Fitria Nur Azizah et al., "Performance Analysis of Transformer based Models (BERT, ALBERT, and RoBERTa) in Fake News Detection," *2023 6th International Conference on Information and Communications Technology (ICOIACT)*, Yogyakarta, Indonesia, pp. 425-430, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[30] M.D. Aaseegha, and B. Venkataramana, "A Hybrid Framework for Enhanced Segmentation and Classification of Colorectal Cancer Histopathology," *Frontiers in Artificial Intelligence*, vol. 8, pp. 1-18, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[31] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton, "Layer Normalization," *arXiv preprint*, pp. 1-14, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[32] Ashish Vaswani et al., "Attention is all You Need," *NeurIPS Proceedings Advances in Neural Information Processing Systems*, vol. 30, 2017. [Google Scholar] [Publisher Link]

[33] Ilya Loshchilov, and Frank Hutter, "Decoupled Weight Decay Regularization," *arXiv preprint*, pp. 1-17, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[34] MVSA: Sentiment Analysis on Multi-View Social Data, MCR Lab, 2016. [Online]. Available: https://mcrlab.net/research/mvsa-sentiment-analysis-on-multi-view-social-data/

[35] Huiru Wang et al., "Multimodal Sentiment Analysis Representations Learning via Contrastive Learning with Condense Attention Fusion," *Sensors*, vol. 23, no. 5, pp. 1-15, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[36] Nan Xu, and Wenji Mao, "MultiSentiNet: A Deep Semantic Network for Multimodal Sentiment Analysis," *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, Association for Computing Machinery, New York, NY, United States, pp. 2399-2402, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[37] Jun Du et al., "Hierarchical Graph Contrastive Learning of Local and Global Presentation for Multimodal Sentiment Analysis," *Scientific Reports*, vol. 14, no. 1, pp. 1-13, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[38] Douwe Kiela et al., "Supervised Multimodal Bitransformers for Classifying İmages and Text," *arXiv Preprint*, pp. 1-11, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[39] Alec Radford et al., "Learning Transferable Visual Models from Natural Language Supervision," *Proceedings of the 38th International Conference on Machine Learning, PMLR*, pp. 8748-8763, 2021. [Google Scholar] [Publisher Link]

[40] Teng Niu et al., "Sentiment Analysis on Multi-View Social Data," *MultiMedia Modeling: 22nd International Conference, MMM 2016, Miami*, FL, USA, vol. 9517, pp. 15-27, 2016. [CrossRef] [Google Scholar] [Publisher Link]

Appendix 1

Notations and Formulas: This appendix provides a consolidated list of notations and formulae used in this study.

Input & Encoders

Symbol	Meaning
I	Input image
T	Input text (caption/tweet/news article)
E_t	Output textual token embeddings from ALBERT
F_v	Feature map extracted from EfficientNet-B2
Z_t	Projected text embedding (256-D)
Z_v	Projected image embedding / patch embeddings (256-D)

Attention & Fusion

Symbol	Meaning
Q, K, V	Query, Key, Value matrices for multi-head attention
H_t	Text features after attending to the image (Text→Image attention)
H_v	Image features after attending to text (Image→Text attention)
$[CLS]_t$	ALBERT's classification token after attention alignment
Pooled(H_v)	Aggregated visual embedding (pooling over image tokens)
Z_{fused}	Joint multimodal embedding after fusion

Contrastive Learning Notation

Symbol	Meaning
s_{ij}	Cosine similarity between text embedding i and image embedding j
τ	Temperature parameter for InfoNCE loss
N	Batch size (number of image–text pairs)
$\mathcal{L}_{\text{con}}$	Contrastive (InfoNCE) loss

Cosine similarity formula: $\boldsymbol{s_{ij}} = \frac{\boldsymbol{z_{t,i} \cdot z_{v,j}}}{\|\boldsymbol{z_{t,i}}\|\|\boldsymbol{z_{v,j}}\|}$

Contrastive loss: $\boldsymbol{\mathcal{L}_{con}} = -\frac{\boldsymbol{1}}{\boldsymbol{N}}\sum_{\boldsymbol{i}} \boldsymbol{log} \frac{\boldsymbol{exp(s_{ii}/\tau)}}{\sum_{\boldsymbol{j}} \boldsymbol{exp(s_{ij}/\tau)}}$

Regularizer Notation

Symbol	Meaning
$\mathcal{L}_{\text{reg}}$	Fusion consistency regularization loss
$\|\cdot\|_2$	L2 norm

Regularizer Formula: $\boldsymbol{\mathcal{L}_{reg}} = \| \boldsymbol{Z_{fused}} - \frac{\boldsymbol{Z_t + Z_v}}{\boldsymbol{2}} \|_{\boldsymbol{2}}^{\boldsymbol{2}}$

Classification Module Notation

Symbol	Meaning
$\hat{y}$	Predicted class probability vector
y	Ground-truth sentiment label
$\mathcal{L}_{\text{cls}}$	Cross-entropy classification loss
W, b	Trainable classification layer weights and bias

Prediction: $\hat{y} = \text{softmax}(WZ_{\text{fused}} + b)$

Training & Optimization

Symbol	Meaning
$\mathcal{L}_{\text{total}}$	Final aggregated loss
$\lambda_1, \lambda_2, \lambda_3$	Loss weighting coefficients
AdamW	Optimizer used for end-to-end training.

Final Training Objective: $\mathcal{L}_{\text{total}} = \lambda_1 \mathcal{L}_{\text{cls}} + \lambda_2 \mathcal{L}_{\text{con}} + \lambda_3 \mathcal{L}_{\text{reg}}$