[1]Vani Golagana

[2] Prof. S. Vizianandа Row

[3] Prof. P. Srinivasa Rao

# Adaptive Multimodal Sentiment Analysis: Improving Fusion Accuracy with Dynamic Attention for Missing Modality

Journal of Electrical Systems

***Abstract:*** **-** Multimodal sentiment analysis combines data from different sources, like text and images, to improve the accuracy of sentiment predictions. Our research tackles two critical challenges in MSA: missing data and effective fusion through attention mechanisms. The challenge of missing data can be a significant obstacle in real-time applications. We propose a method to handle cases where either text or images are missing by utilizing the knowledge of available data to fill in the gaps. Our approach begins with feature extraction. For text, we utilize advanced natural language processing models to obtain rich, context-aware representations. For images, we employ deep convolutional neural networks to capture detailed visual features. After extracting these features, we calculate sentiment scores for both modalities to identify the most relevant modality. These sentiment scores play a crucial role in determining attention weights, allowing the model to focus dynamically on the most significant features from each modality. We then concatenate the text and image features according to these attention weights, ensuring a robust and accurate fusion of information. These fused features are fed into a classification algorithm to predict the overall sentiment. Our method outperforms previous approaches, demonstrating the effectiveness of using attention-based fusion networks for multimodal sentiment analysis. This framework also underscores the importance of effectively handling missing data to maintain robust performance in real-time scenarios. It shows the potential for improving sentiment analysis in practical applications by intelligently combining multimodal data using attention-weighted fusion

***Keywords:*** Multimodal Sentiment Analysis, Natural Language Processing, Missing Data Handling, Attention Mechanisms, Feature Fusion, Deep Learning, Sentiment Classification

## I. INTRODUCTION

Multimodal sentiment analysis (MSA) is an emerging area that combines natural language processing and computer vision to understand and interpret emotions conveyed through various forms of data, such as text and images. Sentiment analysis aims to uncover the emotional context in textual or visual content, which is crucial for tasks like analyzing customer feedback and monitoring sentiments on social media. MSA seeks to enhance the accuracy and depth of sentiment predictions by integrating information from multiple sources. Text provides semantic context and clear sentiment indicators, while images offer visual cues that convey implicit emotional signals. These visual cues can either reinforce or conflict with the sentiment expressed in text, providing a more comprehensive understanding of the overall sentiment conveyed by multimedia content. By combining these modalities, MSA seeks to provide richer insights into human emotions expressed through diverse forms of communication. MSA faces challenges like data heterogeneity, where text and images reside in different feature spaces, and missing modalities, where one may be absent or incomplete. In real-world applications, the absence of one or more modalities, such as text and image data, is a common challenge that can significantly impact the performance of multimodal sentiment analysis (MSA) systems. Another challenge is directly concatenating data without attention mechanisms can lead to the loss of modality-specific information and reduced performance in sentiment classification [1]. Many multimodal learning approaches naively concatenate the two modalities without deeply investigating the intricate correlations among them, which results in suboptimal performance in sentiment classification. Direct

[1,2,3] Department of CS&SE, AUCE, Andhra University, Visakhapatnam, AP, India.
[1]Email: vani.srr22@gmail.com

concatenation of data from more than one modality without using an attention mechanism may lead to the loss of important information on the modality level. In such scenarios, it may fail to capture and exploit the specific features of all of the modalities. Direct concatenation does not make the interaction and dependency among them as all the features being treated are considered equally and not judged for relevance [2][3].

Attention-based networks leveraging deep learning techniques have gained widespread adoption for modeling the semantic relationships across image and textual data [4], [5]. These networks typically utilize convolutional neural networks (CNNs), recurrent neural networks (RNNs), or a combination thereof. More recently, graph neural networks (GNNs) and transformers have also surged in popularity for attention-based approaches in multimodal tasks involving image and text. These methodologies incorporate attention mechanisms that enhance model performance, showcasing a diverse range of effective strategies in deep learning. Recent advancements in deep learning have significantly elevated MSA, enabling the extraction of rich representations from both textual and visual data. Techniques based on deep neural networks, attention mechanisms [6], and strategies for integrating multimodal information are instrumental in synthesizing diverse knowledge to predict sentiment accurately. These methods have been crucial in our research, where combining textual and visual cues ensures a nuanced comprehension of emotional content across different media formats [7].

Advanced deep learning techniques, including attention mechanisms and multimodal fusion strategies, play pivotal roles in extracting rich representations and integrating information effectively across modalities [8].

Our main contributions are as follows:

- Handling Missing Data: Developed a method to effectively handle cases where either text or images are missing in multimodal sentiment analysis applications.
- Feature Extraction: Utilized advanced techniques for feature extraction from both textual and visual data.
- Attention Mechanism: Introduced attention mechanisms to dynamically weight the contributions of text and image features based on their relevance.
- Fusion Strategy: Developed an attention-weighted fusion strategy that concatenates text and image features based on their respective attention weights.
- Highlighted the practical implications of the research by demonstrating how intelligently combining multimodal data using attention-weighted fusion, along with handling missing modalities, can significantly enhance sentiment analysis in real-world applications.

This paper is organized as follows: Section 3 outlines the related work on multimodal sentiment analysis, providing a comprehensive review of existing approaches and methodologies. Section 4 presents the detailed methodology for building a multimodal sentiment analysis system, including preprocessing steps, feature extraction techniques, sentiment score calculation, and fusion strategies. Section 5 discusses the results and provides an analysis of the system's performance. Finally, Section 6 concludes the paper with a summary of findings and suggestions for future research directions.

## II. RELATED WORK

Yang et al. [9] introduce the Multi-View Attentional Network (MVAN), aimed at multimodal emotion classification by integrating image and text modalities. The network uses attention mechanisms to highlight important features within each modality and capture the interactions between them. This multi-view approach enables MVAN to leverage complementary information from both image and text data, thereby enhancing the accuracy of emotion classification tasks. Jin et al. (2020) [10] focus on Deep Multi-View Attentive Network (DMVAN), designed for interpretable multimodal sentiment classification, utilizing both image and text data. The network employs attention mechanisms to highlight relevant features within each modality. This enhances interpretability by identifying the features that most significantly contribute to sentiment classification. In Israa K. Salman Al-Tameemi et al. [11] multimodal data (combining visual and textual content) can convey user emotions more effectively than unimodal content. However, existing approaches often treat modalities independently or simply combine features, neglecting semantic details and the relationship between visual and textual content. Peng et al. (2021) [12] discuss a Cross-Modal Complementary Network (CMCN), designed for multimodal sentiment classification. The network utilizes

hierarchical fusion mechanisms to integrate both intra-modal and cross-modal complementary information effectively. This approach enhances sentiment understanding by synthesizing insights from diverse modalities at multiple levels of abstraction. In [13] Keith April Araño et al., discuss a new manifold of hyperbolic space for multimodal sentiment and emotion recognition and initiated its application for the first time in this area. This will enrich semantic modeling by dealing more comprehensively with complex relationships and hierarchies inherent in data modalities. This approach is supposed to enhance the performance and interpretability of sentiment analysis systems against traditional methods based on Euclidean space. Chen et al. (2022) [14] proposed a weighted cross-modal attention mechanism, WCAM, combined with a senti- ition auxiliary task for multimodal sentiment analysis. WCAM enhances modality fusion by adjusting attention weights accordng to the sentiment carried by each modality. Tang et al. [15] discuss a bi-directional attention mechanism for fusing information from multiple modalities in the context of sentiment analysis. It utilizes attention mechanisms to capture relevant information from both textual and visual modalities. The bi-directional attention enables the model to attend not only to the input modalities but also to the interactions between them. By dynamically attending to salient features in both directions, the model effectively integrates information from different modalities for sentiment analysis. Qingfu Qi et al. [16] proposed a framework for multimodal sentiment analysis that utilizes a Multimodal Encoding-Decoding Transformer (MEDT) network. MEDT leverages transformer architecture for encoding and decoding both textual and visual modalities in a multimodal context. It employs attention mechanisms to capture inter-modal interactions and enable effective fusion of information from multiple modalities for sentiment analysis. In Xiaojun Xue et al. [6] utilizes attention mechanisms at multiple levels to capture informative features from both textual and visual modalities. It generates attention maps to highlight important regions in each modality and integrates them for sentiment analysis, enabling effective fusion of multimodal information. Tong Zhu et al.,[17] highlight a Multi-Level Semantic Reasoning Network (MLSRN) for multimodal emotion classification. MLSRN utilizes semantic reasoning mechanisms at multiple levels to capture nuanced emotional cues from different modalities. It leverages semantic embeddings to represent textual and visual information and integrates them through reasoning modules for effective emotion classification. Yanan Wang et al. [18] propose a method for video-level sentiment analysis that utilizes a Variational Autoencoder (VAE)-based adversarial multimodal domain transfer approach. This method aims to transfer sentiment knowledge across different domains by learning a shared latent space representation.

It employs VAE to model the joint distribution of features from multiple modalities and adversarial training to align the latent distributions between source and target domains. Junling Zhang et al. [19] introduced a novel adaptive modality-specific weight fusion network for multimodal sentiment analysis, integrating modality-specific layers and fusion mechanisms. It proposes weight-based feature fusion and a weight-mapping network to optimize multimodal integration, leveraging tailored unimodal feature generators for enhanced sentiment prediction. Licai Sun et al. [20] aim to propose an efficient multimodal transformer (EMT) architecture for robust multimodal sentiment analysis. It employs dual-level feature restoration (DFR) to enhance the fusion of modalities and mitigate the data imbalance issue. It combines both transformer-based models and feature restoration techniques for effective sentiment analysis. Qiupu Chen et al. [14] presents a multimodal sentiment analysis approach that combines an Attentional Temporal Convolutional Network (ATCN) with multi-layer feature fusion. ATCN is used to capture temporal dependencies in sequential data, while multi-layer feature fusion integrates features from multiple modalities at different abstraction levels for sentiment analysis. In [21] Alireza Ghorbanali et al. focus current methodologies, challenges, and future prospects in leveraging deep learning for multimodal sentiment analysis. Yanping Fu et al. [22] propose a hybrid cross-modal interaction learning approach for multimodal sentiment analysis. The proposed method leverages interactions between different modalities (e.g., text and image) to enhance sentiment analysis. It utilizes cross-modal attention mechanisms to capture informative interactions between modalities and fusion techniques to integrate multimodal features effectively. In [23] Peicheng Wang et al., discuss a novel method for multimodal sentiment analysis that explores cross-correlation in dual-attention mechanisms. This approach may involve leveraging dual-attention mechanisms to capture cross-modal dependencies and correlations between different modalities (such as text, images, and audio) for sentiment analysis tasks. The method may utilize deep learning architectures and attention mechanisms to learn informative representations and perform sentiment classification across modalities. The paper Bo Yang et al. [24] introduced the two-phase multi-task sentiment analysis (TPMSA) framework, enhancing pre-trained models with a two-phase training strategy and a novel multi-task learning approach. Chakraborty et al. (2024) [25] present a multimodal sentiment analysis (MSA)

approach for determining the sentiment of an image-text tweet. The approach utilizes multimodal decision-level fusion by integrating features from individual modalities and considering inter-modal semantic relationships to derive the final sentiment classification. In [26] focus on advancing visual sentiment analysis using data-augmented deep transfer learning techniques. By leveraging transfer learning, these models are adapted to enhance their feature extraction capabilities without relying on their original classification layers. Augmentation methods are employed to enrich the dataset, aiming to enhance model robustness and performance. Jiang, D., et al. (2021) [27] propose a deep learning framework that simultaneously addresses sarcasm detection, sentiment analysis, and emotion analysis in a multi-modal conversational context. They introduce two attention mechanisms: Inter-segment Inter-modal Attention (Ie-Attention) and Intra-segment Inter-modal Attention (Ia-Attention). These attentions are combined to improve sarcasm detection performance by leveraging two secondary tasks: emotion and sentiment analysis.

## III. METHODOLOGY

In this section, we outline the methodology used in our research, which involves several key stages: preprocessing, feature extraction, modality sentiment score, fusion strategy, and classificaiton. We proposed an Adaptive Attention-Based Fusion for Multimodal Sentiment Analysis (AAFMSA), the work of which is shown in Figure 1, and elaborated in Algorithm 1.

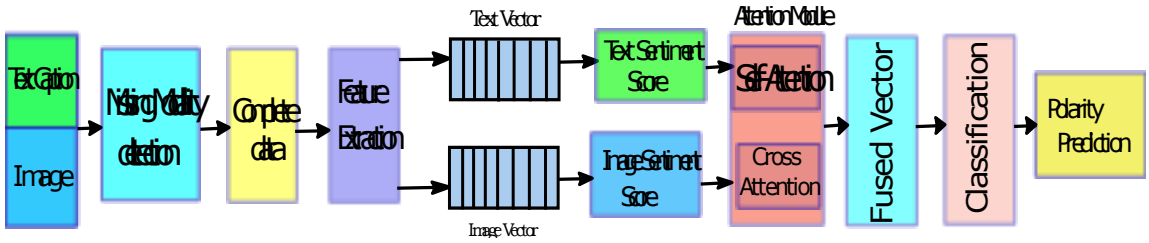


Figure 1: Proposed Workflow of Adaptive Attention-Based Fusion for Multimodal Sentiment Analysis (AAFMSA)

### *A. Preprocessing*

The preprocessing phase is essential for preparing the data for effective analysis and involves several critical tasks. One of the key aspects of preprocessing is handling missing modalities [28], which can significantly impact the performance of multimodal sentiment analysis".

1. Determine missing modality: We use specific tags to identify whether the text or image modality is missing in our dataset. If a modality is absent, a tag is assigned (Text_Missing or Image_Missing), guiding the subsequent generation process.

2. Handling Missing Modality: We handle missing modalities by adapting the knowledge from the existing modality. This approach involves employing various techniques to address the missing text caption and missing image data and reconstruct a complete dataset, as illustrated in Figure 2.

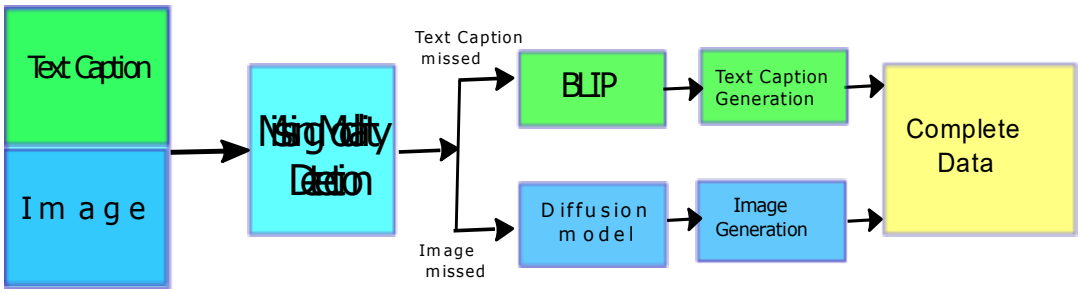


Figure. 2: Workflow of Handling Missing Modality

**Algorithm 1**

**Input:** Dataset with text and image modalities, each associated with sentiment labels. **Output:** Predicted sentiment labels for each text-image pair.

**1. Preprocessing Load Dataset:** Import the dataset containing text and image modalities with sentiment labels. **Handle Missing Modalities:** Identify and handle missing modalities: **a.** Impute missing text/image data where applicable. **b.** Exclude incomplete pairs from analysis if imputation is not feasible.

**2. Feature Extraction Extract Text Features:** Use BERT to obtain text embeddings from the text data. **Extract Image Features:** Use ResNet50 to obtain image embeddings from the image data.

**3. Sentiment Score Calculation Calculate Sentiment Scores for Text:** Apply a sentiment analysis model to the text embeddings. Let $S_t$ be the sentiment score for text, computed as:

$$S_t = f_{\text{text}}(E_t)$$

where $E_t$ represents the text embeddings and $f_{\text{text}}$ is the sentiment analysis model.

**Calculate Sentiment Scores for Images:** Apply a sentiment analysis model to the image embeddings. Let $S_i$ be the sentiment score for images, computed as:

$$S_i = f_{\text{image}}(E_i)$$

where $E_i$ represents the image embeddings and $f_{\text{image}}$ is the sentiment analysis model.

**4. Fusion of Features Align Features:** Align text and image features based on their corresponding sentiment scores. **Apply Fusion Strategy:** Combine the aligned features using a fusion strategy. For concatenation, the fused feature vector $F$ is:

$$F = [E_t; E_i]$$

For attention-based fusion, calculate attention weights $\alpha_t$ and $\alpha_i$ as:

$$\alpha_t = \frac{\exp(S_t)}{\exp(S_t) + \exp(S_i)}$$

$$\alpha_i = \frac{\exp(S_i)}{\exp(S_t) + \exp(S_i)}$$

**a.** Calculate attention weights based on sentiment scores to emphasize relevant features. The weighted feature vector $F_w$ is:

$$F_w = \alpha_t \cdot E_t + \alpha_i \cdot E_i$$

**5. Classification Prepare Fused Features:** Use the fused features $F$ or weighted features $F_w$ as input to a classification algorithm. **Train Classifier:** Train a classifier (e.g., fully connected neural network) on the dataset with the fused features and sentiment labels.

**6. Prediction Predict Sentiment Labels:** Use the trained classifier to predict sentiment labels for new text-image pairs.

**Handling missing Text Caption**: When dealing with a missing text caption, the goal is to generate a meaningful description based on the associated available image. To do this effectively, we need a model that can understand both visual content and language. The Bootstrapping Language Image Pretraining (BLIP) model, which was highlighted by Song et al. in 2024 [29], is a technology that generates captions from images with a high level of efficiency. It uses convolutional neural networks (CNNs) to analyze and extract features from images and transformer networks to understand and generate relevant caption. The process starts by preparing the image: it's resized and its pixel values are normalized to make sure the model can work with it properly. The image is then converted into a format that the model can process. Once this preparation is done, the image is fed into the BLIP model. The CNN part of the model dives into the image to pull out important visual details, and then the transformer network takes over to craft a text caption that describes what's in the image.

**Handling Missing Images**: When an image is missing, Stable Diffusion XL (SDXL) [30] can generate it based on the existing text caption through a sophisticated diffusion process. This process starts with a random noise tensor, which the model iteratively refines into a coherent image. The model leverages a reverse diffusion process, where it begins with a noisy image and applies successive denoising steps, progressively enhancing the image quality. The denoising process is governed by the following equation:

$$\mathbf{x}_{t-1} = \mathbf{x}_t - \alpha_t \nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t|\mathbf{x}_0) \quad (1)$$

where $\mathbf{x}_t$ is the image tensor at step $t$, $\alpha_t$ is the step size, and $\nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t|\mathbf{x}_0)$ is the gradient of the log-probability of the image at step $t$ given the final image $\mathbf{x}_0$. The model continues this process iteratively, applying each denoising step until a realistic and clear image emerges that aligns with the text description provided

*B.* ***Feature Extraction*** It is critical step in processing both textual and visual data for multimodal sentiment analysis In this section we describe the methods used for extracting features from text and images..

**Text Feature Extraction**: For text feature extraction, the process begins with tokenization, where the text input is split into tokens that BERT (Bidirectional Encoder Representations from Transformers) can handle.

This includes the addition of special tokens such as [CLS] at the beginning and [SEP] at the end of the sentence. Following tokenization, the tokens are converted into embeddings using BERT's embedding layer, which encompasses position embeddings, segment embeddings, and token embeddings. This transformation allows the model to capture the semantic and syntactic information necessary for further analysis.

**Image Feature Extraction**: For image feature extraction, we utilize ResNet50, a deep convolutional neural network known for its effectiveness in handling very deep networks through its residual learning framework. ResNet50 is widely used in image processing due to its ability to capture detailed visual information while remaining both robust and efficient. The features extracted by ResNet50 are high-dimensional vectors that represent various visual aspects of the image such as shapes, textures, and patterns.

*C.* ***Modality Sentiment Score*****:** In this section, we outline the method for calculating sentiment scores from both text and image data using Convolutional Neural Networks (CNNs), followed by normalization to obtain a unified sentiment score.

CNNs excel at capturing local patterns within text and image data, which is crucial for accurate sentiment analysis. By combining text and image modalities, multimodal sentiment analysis offers a more comprehensive perspective on sentiment. To emphasize the most informative modality, we dynamically adjust the contribution of each modality based on its sentiment score.

**Sentiment Score using CNN:** The features extracted from text using BERT and from images using ResNet50 are then fed into convolutional layers for further processing. The convolutional layers are designed to capture additional patterns and enhance the overall feature representation. The convolution operation for text is described as:

$$\mathbf{c}_i = \text{ReLU}(\mathbf{W}_c * \mathbf{E}_{i:i+h-1} + \mathbf{b}_c) \quad (2)$$

where, $c_i$ represents the output feature map at position $i$. $\mathbf{W}_c$ is the filter matrix, $\mathbf{E}_{i:i+h-1}$ represents the pre-extracted text features from the $i$-th to the $i+h-1$-th positions, $*$ denotes the convolution operation, and $\mathbf{b}_c$ is the bias term.

Similarly, the convolution operation for image is described as:

$$\mathbf{f}_{i,j} = \text{ReLU}\left(\mathbf{W}_r * \mathbf{I}_{i:i+h-1,j:j+w-1} + \mathbf{b}_r\right) \quad (3)$$

where $\mathbf{W}_r$ is the filter matrix in the residual block, $\mathbf{I}_{i:i+h-1,j:j+w-1}$ represents the image patch, and $\mathbf{b}_r$ is the bias term.

Following convolution, pooling layers (e.g., max pooling or average pooling) are used to reduce the dimensionality of the feature maps. This step helps to retain essential features while decreasing computational complexity. The output feature maps from the convolutional layers are concatenated and flattened into a one-dimensional vector. This vector is then passed through a fully connected layer to integrate the features and produce sentiment scores. The final sentiment score is computed using:

$$\text{Sentiment Score} = \sigma\left(\mathbf{W}_f \mathbf{c} + \mathbf{b}_f\right) \quad (4)$$

where $\mathbf{W}_f$ and $\mathbf{b}_f$ are the weights and biases of the fully connected layer, and $\sigma$ is the sigmoid activation function.

**Sentiment score Normalization:** The sentiment scores are normalized to provide a unified sentiment measure. This is achieved using following formula:

$$\text{Normalized Sentiment Score} = \frac{\text{positive}-\text{negative}}{\text{positive}+\text{neutral}+\text{negative}} \quad (5)$$

These sentiment scores play a key role in subsequent fusion processes within multimodal sentiment analysis.

### *D. Attention based on sentiment scores*

After calculating the sentiment scores for both text and image modalities, the scores are used to assess the relevance of each modality. Based on this relevance, self-attention is applied to the more influential modality and cross-attention between the two modalities. This method allows for effective combination of the text and image vectors, prioritizing the modality with the stronger sentiment signal [19,31,32,33].

If the text sentiment score is higher than the image sentiment score, self-attention is applied to the text data, and cross-attention is used between the text and image modalities. Conversely, if the image sentiment score surpasses the text sentiment score, self-attention is applied to the image data, with cross-attention between the two modalities. The outcomes of these processes are then merged using attention weights [34].

**Self-Attention on Text:** The sentiment score is greater, and self-attention is applied to the text features to capture the most relevant information as $\mathbf{A}_{\text{text}}$.

$$\mathbf{A}_{\text{text}} = \text{softmax}\left(\frac{\mathbf{Q}_{\text{text}}\mathbf{K}_{\text{text}}^{T}}{\sqrt{d_k}}\right)\mathbf{V}_{\text{text}} \quad (6)$$

where $\mathbf{Q}_{\text{text}}$, $\mathbf{K}_{\text{text}}$, and $\mathbf{V}_{\text{text}}$ are the query, key, and value matrices for the text features, and $d_k$ is the dimensionality of the key vectors.

**Self-Attention on Image:** If the image sentiment score is greater, self-attention [36] is applied to the image features to capture the most relevant information as $\mathbf{A}_{\text{image}}$.

$$\mathbf{A}_{\text{image}} = \text{softmax}\left(\frac{\mathbf{Q}_{\text{image}}\mathbf{K}_{\text{image}}^{T}}{\sqrt{d_k}}\right)\mathbf{V}_{\text{image}} \quad (7)$$

where $\mathbf{Q}_{\text{image}}$, $\mathbf{K}_{\text{image}}$, and $\mathbf{V}_{\text{image}}$ are the query, key, and value matrices for the image features, and $d_k$ is the dimensionality of the key vectors.

**Cross-Attention Between Text and Image:** If the image sentiment score is greater, self-attention [36] is applied to the image features to capture the most relevant information as $\mathbf{A}_{\text{image}}$.

$$\mathbf{A}_{\text{cross}} = \text{softmax}\left(\frac{\mathbf{Q}_{\text{main}}\mathbf{K}_{\text{other}}^{T}}{\sqrt{d_k}}\right)\mathbf{V}_{\text{other}} \quad (8)$$

where $\mathbf{Q}_{\text{main}}$ is the query matrix for the modality with the higher sentiment score, and $\mathbf{K}_{\text{other}}$ and $\mathbf{V}_{\text{other}}$ are the key and value matrices for the other modality.

### *D. Fusion based on attention weights*

The outputs from self-attention and cross-attention mechanisms are integrated using learned attention weights, as described in [37, 38]. This integration results in a combined feature representation that leverages the strengths of both attention mechanisms for more effective multimodal sentiment analysis as shown in Figure 3. The combined feature representation, denoted as $\mathbf{F}_{\text{combined}}$, is calculated as follows:

$$\mathbf{F}_{\text{combined}} = \alpha\mathbf{A}_{\text{main}} + \beta\mathbf{A}_{\text{cross}} \quad (9)$$

where $A_{\text{main}}$ represents the output from the self-attention applied to the dominant modality, i.e., either $\mathbf{K}_{\text{text}}$ or $\mathbf{K}_{\text{image}}$.

while $\mathbf{A}_{\text{cross}}$represents the output from the cross-attention between modalities.The terms $\alpha$ and $\beta$ are the respective attention weights that determine the contribution of each attention mechanism to the final representation.

This combined feature representation $F_{\text{combined}}$ is then processed through a fusion layer to predict the final sentiment score:

$$\text{Final Sentiment Score} = \sigma(\mathbf{W}_{\text{fusion}}\mathbf{F}_{\text{combined}} + \mathbf{b}_{\text{fusion}}) \quad (10)$$

where $\mathbf{W}_{\text{fusion}}$ and $\mathbf{b}_{\text{fusion}}$ are the weights and bias of the fusion layer, and $\sigma$ is the sigmoid activation function.

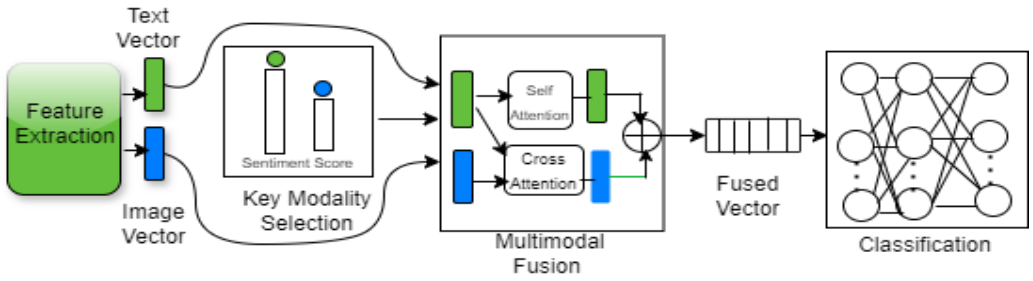


Figure 3: Multimodal Fusion Based on Attention

### *E. Classification*

To complete the sentiment analysis, we use a convolutional neural network (CNN) for classification. The CNN takes the fused features from the attention mechanisms followed by ReLu activation to introduce non-linearity. Pooling layers then reduce the spatial dimensions, retaining key information while lowering computational costs. The output is flattened and passed through fully connected layers, where the final softmax activation generates probability scores for classification. This process ensures that our final sentiment prediction is accurate and meaningful by effectively leveraging the strengths of both modalities [39].

## IV. RESULTS

### *A. Dataset*

We use the MVSA-S(Single)datasets to evaluate our model. The MVSA-S image-text pairs and collected from Twitter, with each pair annotated separately for sentiment by a single annotator. To ensure a fair comparison and align with previous research, we concentrate on pairs where the text and image share the same sentiment label. By applying the filtering and processing methods described in [41,42], we refine the datasets to 4511, that have consistent sentiment labels. The MVSA-S data set consists of 2683 positive, 470 neutral and 1358 negative samples.

### *B. Baseline methods*

The sentiment analysis models span text, image, and multimodal approaches. For text, CNN-T uses CNNs [42] to capture local sentiment patterns, BiLSTM-T [43] employs bidirectional LSTMs to understand contextual information, and BiACNN-T [43] combines CNN, BiLSTM, and an attention mechanism for enhanced sentiment extraction. In image sentiment analysis, Incep-V [44] utilizes the Inception V3 architecture for feature extraction, while OSDA-V applies double attention mechanisms to focus on both object and scene features. For multimodal sentiment analysis, SF-M uses simple fusion techniques to combine text and image data, DSN-M and MultiSentiNet-M [45] incorporate advanced fusion strategies, HSAN-M [46] employs a hierarchical semantic attentional network for nuanced understanding, and Co-Memory-M [47] iteratively models interactions between visual and textual data, refining the sentiment representation through a co-memory mechanism. MVAN-M [9] is a multimodal sentiment analysis model that uses attention mechanisms to capture important interactions between image and text data for accurate sentiment prediction.

### *C. Evaluation and comparison*

In our experiments, we thoroughly evaluated the performance of our model under various conditions. First, we assessed the impact of missing modalities by comparing results obtained with and without missing modality. We then conducted a comparative analysis between unimodal (single modality) and multimodal data, examining how the integration of both text and image modalities influences the overall model performance. Additionally, we explored different encoding techniques to identify the most effective methods for text and image processing. The results are evaluated on the MVSA-S dataset in comparison to several baseline models. To understand the impact of missing modalities, we tested the data accuracy in both cases: with missing data and without missing data on two sample datasets. The results, shown in Table 1, indicate that when data is missing, the accuracy drops to 63.5% and 68.9% for Sample 1 and Sample 2, respectively. However, when there is no missing data, the accuracy significantly improves to 81% for Sample 1 and 88%

for Sample 2 respectively. This highlights the importance of complete data in achieving higher accuracy in prediction tasks. For instances with missing text or image data, we applied imputation techniques to generate synthetic data that closely matches the existing modality. Specifically, when text data was missing, we generated text captions from the associated images. Conversely, when image data was missing, we created images from the available text captions.

TABLE 2: Comparison of actual and generated captions for the given images

| Image | Actual Caption | Generated Caption |
|---|---|---|
|  | RT @McDo_PH: Crushed graham, banana flavor, creamy vanilla soft serve Â Minion-approved! #MinionsAtMcDo | a close up of a cup of food with bananas and a banana |
|  | Valley of the Giants - where @WestAustralia makes you feel very small: #thisiswa | giraffe bridge over a forest with trees and a skywalk |
|   | RT @Ruptly: #RightSector set up checkpoints outside Kiev | a group of people standing around a pile of sandbags on the side of a road |
|   | Flight to Sicily for International Computer Vision Summer School ICVSS 2015 | giraffe airlines plane parked on the tarmac at an airport |
|  | Running clothes, gum in mouth, guac, bananas, sweet potato fries, and sparkling water. My life in one photo. | giraffe woman standing in front of a cash register in a store |

TABLE 1: Accuracy Comparison for Samples with Missing vs. Complete Data

| | Accuracy for Sample 1 | Accuracy for Sample 2 |
|---|---|---|
| **Missing Data** | 0.63 | 0.68 |
| **No missing Data** | 0.81 | 0.88 |

To address the issue of missing modalities in our dataset, we employed generation techniques to estimate the missing data and then compared the generated data with the actual data. Examples of these generated captions and images are provided in Table 2 and Table 3respectively. It demonstrates the effectiveness of our approach in handling missing modalities. This preprocessing step ensured dataset completeness by imputing any missing values. Consequently, our sentiment score fusion and subsequent classification steps were performed on a fully populated dataset.

TABLE 3: Comparison of actual and generated images for the given captions

| Text Caption | Actual Image | Generated Image |
|---|---|---|
| RT @rhettandlink: leftover america cake? |  |  |
| Found this, discarded and weather-worn, while taking a walk yesterday. Thinking of making it to @soundingline. |  |  |
| RT @LoveStoneArts: Evil Eye Earrings Lampwork Hearts in Plum Pink- Hamsa E107 http://t.co/gRWzr9i0ns #jewelryonetsy #GoldFilled |  |  |

| Text Caption | Actual Image | Generated Image |
|---|---|---|
| These prairie fields look absolutely amazing during sunset |  |  |
| A beautiful landscape with mountains and a lake where children are playing |  |  |

For the MVSA dataset, we specifically used BERT for text and compared ResNet50, VGG16, and Xception for image encoding. The performance differences are detailed in Table 5, showing that ResNet50 provided the best results for image features. Similarly, for image input, we used ResNet50 while testing different text encoders (BERT, GloVe, and Universal Sentence Encoder), and found that BERT consistently outperformed the others. This confirms the results from earlier experiments on the Flickr8k dataset, where BERT proved superior for text and ResNet50 was the optimal choice for image feature extraction. The integration of different modalities in sentiment analysis is crucial for improving classification accuracy. By combining textual and visual data, we can capture a richer understanding of sentiment than either modality can provide alone. This is evident in our findings, where single modality inputs (Text or Image) achieved an accuracy of only 55%. In contrast, the combination of Text and Image resulted in a significantly higher accuracy of 81% as shown in Table 4.

Following preprocessing, features were extracted from each modality using appropriate encoding techniques. Sentiment scores were then computed for both text and image modalities. Given that one modality often proved more relevant than the other for sentiment analysis, we applied self-attention mechanisms to the more relevant modality. Additionally, we employed cross-attention mechanisms to integrate information from both modalities, facilitating a comprehensive understanding of sentiment through their interaction.

TABLE 4: Accuracy scores for single and multimodal input

| **Input data** | **Accuracy** |
|---|---|
| Text (BERT) | 0.55 |
| Image (ResNet50) | 0.55 |
| **Text+Image(BERT+ResNet50)** | **0.81** |

Table 5 presents the accuracy scores for Sample 1 using the BERT encoding technique for textual data combined with various image encoding techniques, including ResNet50, VGG16, and Xception. The combination of BERT with ResNet50 achieved a notable accuracy score of 81%. Furthermore, when an autoencoding technique is applied to the ResNet50 model (referred to as ResNet50+AE), the accuracy further improves to 88%, demonstrating the effectiveness of integrating autoencoders in enhancing the performance of image data processing. When evaluating the performance of different text embedding techniques combined with the ResNet50 image model, we observed notable differences in accuracy.

TABLE 5: Performance of BERT combined with different image models for Sample 1

| EncodingTechnique | sample1 |
|---|---|
| **BERT+(ResNet50+AE)** | **0.88** |
| **BERT+ResNet50** | **0.81** |
| BERT+VGG16 | 0.78 |
| BERT+Xception | 0.79 |

As shown in Table 6, three prominent text embedding methods—GLOVE, USE, and BERT—were tested alongside the ResNet50 image model. Among these combinations, the ResNet50+BERT pairing demonstrated

the best performance, achieving an impressive accuracy of 89%. This was followed by the ResNet50+GLOVE and ResNet50+USE combinations, which achieved accuracy scores of 86% and 84%, respectively.

TABLE 6: Performance of ResNet50 combined with different text embeddings for Sample1

| Encoding Technique | sample1 |
|---|---|
| **ResNet50+BERT** | **0.89** |
| ResNet50+GLOVE | 0.84 |
| ResNet50+USE | 0.86 |

After feature extraction, attention was applied based on sentiment scores, and fusion was performed by dynamically updating the weights. The fused features were then passed through a CNN for final classification, using ReLU activation in intermediate layers and Softmax in the output layer to generate the classification results.

In our research, we explored various configurations for the sentiment analysis model, focusing on optimizing hyperparameters using K-Fold Cross-Validation and dropout regularization. We experimented with different values of k (3, 5, and 10) to evaluate the model's performance across multiple data subsets, ensuring a comprehensive assessment of its robustness and minimizing the risk of overfitting. Additionally, dropout layers were incorporated into the neural network architecture, with dropout rates ranging from 0.2 to 0.9 to enhance generalization. When k=3, dataset was split into 3 equal folds. Each fold serves as a validation set once, while the remaining two-thirds of the data are used for training. This process is repeated 3 times, ensuring that every part of the data is used for both training and validation. Our findings revealed that the optimal configuration was achieved with k=3 and a dropout rate of 0.2, which provided the most reliable results by effectively balancing model complexity and generalization. This combination significantly reduced variance between training and test sets, resulting in improved predictive accuracy on both training and unseen test data.

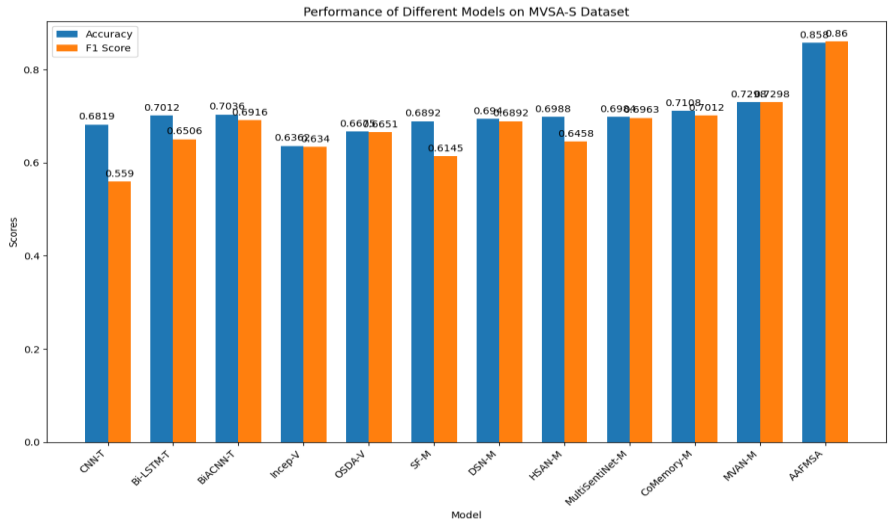


Figure. 4: Performance of different models on MVSA-S dataset

Table VII presents the performance comparison of different models on the MVSA-S dataset. Our model, AAFMSA, achieved an accuracy and F1-score of 0.858and 0.86 respectively, outperforming all the baseline models, including CoMemory-M and MVAN-M, which attained an accuracy of 0.7108 and 0.7298, respectively. This superior performance is also visually depicted in Figure 4, where our model stands out with the highest accuracy and F1-score, demonstrating the effectiveness of our attention-based fusion mechanisms in handling multimodal data compared to existing approaches.

In summary, our results validate the effectiveness of our preprocessing and fusion network strategies for sentiment classification, particularly in managing missing modalities. The substantial improvements in accuracy and F1-scores highlight the robustness of our methods in multimodal sentiment analysis. Our

comprehensive preprocessing and sentiment score fusion approach significantly improved the accuracy of sentiment prediction. The fusion network effectively combined information from both text and image modalities, resulting in more accurate sentiment predictions.

| **Model** | **MVSA-S (Accuracy)** | **MVSA-S (F1)** |
|---|---|---|
| CNN-T | 0.6819 | 0.5590 |
| Bi-LSTM-T | 0.7012 | 0.6506 |
| BiACNN-T | 0.7036 | 0.6916 |
| Incep-V | 0.6362 | 0.6340 |
| OSDA-V | 0.6675 | 0.6651 |
| SF-M | 0.6892 | 0.6145 |
| DSN-M | 0.6940 | 0.6892 |
| HSAN-M | 0.6988 | 0.6458 |
| MultiSentiNet-M | 0.6984 | 0.6963 |
| CoMemory-M | 0.7108 | 0.7012 |
| MVAN-M | 0.7298 | 0.7298 |
| **AAFMSA** | **0.858** | **0.86** |

TABLE VII: Performance of different models on MVSA-S dataset

## V. CONCLUSION

This paper introduces an advanced approach to Multimodal Sentiment Analysis (MSA) that enhances sentiment prediction accuracy by effectively integrating both text and image data. Advanced imputation techniques are employed to address missing modalities, ensuring seamless processing of incomplete data. Attention-based fusion strategies are used to dynamically combine the most relevant features from each modality, leading to improved classification performance. Experimental results validate the effectiveness of our approach, demonstrating how multimodal data fusion produces richer and more accurate sentiment predictions. This work contributes to the field of multimodal machine learning, with practical applications in areas such as social media monitoring, customer feedback analysis, healthcare sentiment assessment, and human-computer interaction. Future research will focus on further refining fusion techniques and exploring additional modalities, such as audio and video, to enhance the capabilities and applicability of multimodal sentiment analysis.

## REFERENCES

[1] Meng Xu, Feifei Liang, Xiangyi Su, and Cheng Fang. Cmjrt: Cross-modal joint representation transformer for multimodal sentiment analysis. IEEE Access, 10:131671–131679, 2022.

[2] Mahesh G Huddar, Sanjeev S Sannakki, and Vijay S Rajpurohit. Attention-based multimodal contextual fusion for sentiment and emotion classification using bidirectional lstm. Multimedia Tools and Applications, 80(9):13059–13076, 2021.

[3] Faliang Huang, Xuelong Li, Changan Yuan, Shichao Zhang, Jilian Zhang, and Shaojie Qiao. Attention-emotion-enhanced convolutional lstm for sentiment analysis. IEEE transactions on neural networks and learning systems, 33(9):4332–4345, 2021.

[4] Dichao Hu. An introductory survey on attention mechanisms in nlp problems. In Intelligent Systems and Applications: Proceedings of the 2019 Intelligent Systems Conference (IntelliSys) Volume 2, pages 432–448. Springer, 2020.

[5] Yunji Liang, Huihui Li, Bin Guo, Zhiwen Yu, Xiaolong Zheng, Sagar Samtani, and Daniel D Zeng. Fusion of heterogeneous attention mechanisms in multi-view convolutional neural network for text classification. Information Sciences, 548:295–312, 2021.

[6] Xiaojun Xue, Chunxia Zhang, Zhendong Niu, and Xindong Wu. Multi-level attention map network for multimodal sentiment analysis. IEEE Transactions on Knowledge and Data Engineering, 35(5):5105–5118, 2022.

[7] Ramandeep Kaur and Sandeep Kautish. Multimodal sentiment analysis: A survey and comparison. Research anthology on implementing sentiment analysis across multiple disciplines, pages 1846–1870,2022.

[8] Benyamin Ghojogh and Ali Ghodsi. Attention mechanism, transformers, bert, and gpt: tutorial and survey. OSF Preprints, 2020.

[9] Xiaocui Yang, Shi Feng, Daling Wang, and Yifei Zhang. Image-text multimodal emotion classification via multi-view attentional network. IEEE Transactions on Multimedia, 23:4014–4026, 2020.

[10] Ning Jin, Jiaxian Wu, Xiang Ma, Ke Yan, and Yuchang Mo. Multi-task learning model based on multi-scale cnn and lstm for sentiment classification. IEEE Access, 8:77060–77072, 2020.

[11] Israa K Salman Al-Tameemi, Mohammad-Reza Feizi-Derakhshi, Saeed Pashazadeh, and Mohammad Asadpour. Interpretable multimodal sentiment classification using deep multi-view attentive network of image and text data. IEEE Access, 2023.

[12] Cheng Peng, Chunxia Zhang, Xiaojun Xue, Jiameng Gao, Hongjian Liang, and Zhengdong Niu. Cross-modal complementary network with hierarchical fusion for multimodal sentiment classification. Tsinghua Science and Technology, 27(4):664–679, 2021.

[13] Keith April Ara˜no, Carlotta Orsenigo, Mauricio Soto, and Carlo Vercellis. Multimodal sentiment and emotion recognition in hyperbolic space. Expert Systems with Applications, 184:115507, 2021.

[14] Qiupu Chen, Guimin Huang, and Yabing Wang. The weighted cross-modal attention mechanism with sentiment prediction auxiliary task for multimodal sentiment analysis. IEEE/ACM Transactions on Audio,Speech, and Language Processing, 30:2689–2695, 2022.

[15] Jiajia Tang, Dongjun Liu, Xuanyu Jin, Yong Peng, Qibin Zhao, Yu Ding, and Wanzeng Kong. Bafn:Bi-direction attention based fusion network for multimodal sentiment analysis. IEEE Transactions on Circuits and Systems for Video Technology, 33(4):1966–1978, 2022.

[16] Qingfu Qi, Liyuan Lin, Rui Zhang, and Chengrong Xue. Medt: Using multimodal encoding-decoding network as in transformer for multimodal sentiment analysis. IEEE Access, 10:28750–28759, 2022.

[17] Tong Zhu, Leida Li, Jufeng Yang, Sicheng Zhao, and Xiao Xiao. Multimodal emotion classification with multi-level semantic reasoning network. IEEE Transactions on Multimedia, 25:6868–6880, 2022.

[18] Yanan Wang, Jianming Wu, Kazuaki Furumai, Shinya Wada, and Satoshi Kurihara. Vae-based adversarial multimodal domain transfer for video-level sentiment analysis. IEEE Access, 10:51315–51324, 2022.

[19] Junling Zhang, Xuemei Wu, and Changqin Huang. Adamow: Multimodal sentiment analysis based on adaptive modality-specific weight fusion network. IEEE Access, 11:48410–48420, 2023.

[20] Licai Sun, Zheng Lian, Bin Liu, and Jianhua Tao. Efficient multimodal transformer with dual-level feature restoration for robust multimodal sentiment analysis. IEEE Transactions on Affective Computing, 15(1):309–325, 2023.

[21] Alireza Ghorbanali and Mohammad Karim Sohrabi. A comprehensive survey on deep learning-based approaches for multimodal sentiment analysis. Artificial Intelligence Review, 56(Suppl 1):1479–1512, 2023. International Journal of Cyber Security and Information Management (IJCIM) Vol. xx, No. yy, PP. x-y, year

[22] Yanping Fu, Zhiyuan Zhang, Ruidi Yang, and Cuiyou Yao. Hybrid cross-modal interaction learning for multimodal sentiment analysis. Neurocomputing, 571:127201, 2024.

[23] Peicheng Wang, Shuxian Liu, and Jinyan Chen. Ccda: A novel method to explore the cross-correlation in dual-attention for multimodal sentiment analysis. Applied Sciences, 14(5):1934, 2024.

[24] Bo Yang, Lijun Wu, Jinhua Zhu, Bo Shao, Xiaola Lin, and Tie-Yan Liu. Multimodal sentiment analysis with two-phase multi-task learning. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 30:2015–2024, 2022.

[25] Debatosh Chakraborty, Dwijen Rudrapal, and Baby Bhattacharya. A multimodal sentiment analysis approach for tweets by comprehending co-relations between information modalities. Multimedia Tools and Applications, 83(17):50061–50085, 2024.

[26] Zhiguo Jiang, Waneeza Zaheer, Aamir Wali, and SAM Gilani. Visual sentiment analysis using data augmented deep transfer learning techniques. Multimedia Tools and Applications, 83(6):17233–17249,2024.

[27] Dazhi Jiang, Runguo Wei, Hao Liu, Jintao Wen, Geng Tu, Lin Zheng, and Erik Cambria. A multitask learning framework for multimodal sentiment analysis. In 2021 International conference on data mining workshops (ICDMW), pages 151–157. IEEE, 2021.

[28] Zhizhong Liu, Bin Zhou, Dianhui Chu, Yuhang Sun, and Lingqiang Meng. Modality translation-based multimodal sentiment analysis under uncertain missing modalities. Information Fusion, 101:101973,2024.

[29] Haisheng Song and Yingdong Song. Target research based on blip model. Academic Journal of Science and Technology, 9(1):80–86, 2024.

[30] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas M¨uller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. ArXiv preprint arXiv:2307.01952, 2023.

[31] Ignazio Gallo, Alessandro Calefati, and Shah Nawaz. Multimodal classification fusion in real-world scenarios. In 2017 14th IAPR international conference on document analysis and recognition (ICDAR), volume 5, pages 36–41. IEEE, 2017.

[32] Chuanbo Zhu, Min Chen, Sheng Zhang, Chao Sun, Han Liang, Yifan Liu, and Jincai Chen. Skeafn:sentiment knowledge enhanced attention fusion network for multimodal sentiment analysis. Information Fusion, 100:101958, 2023.

[33] Ringki Das and Thoudam Doren Singh. Image–text multimodal sentiment analysis framework of assamese news articles using late fusion. ACM Transactions on Asian and Low-Resource Language Information Processing, 22(6):1–30, 2023.

[34] Yifeng Wang, Jiahao He, Di Wang, Quan Wang, Bo Wan, and Xuemei Luo. Multimodal transformer withadaptive modality weighting for multimodal sentiment analysis. Neurocomputing, 572:127181, 2024.

[35] Jian Yang and Juan Yang. Aspect based sentiment analysis with self-attention and gated convolutional networks. In 2020 IEEE 11th International Conference on Software Engineering and Service Science (ICSESS), pages 146–149. IEEE, 2020.

[36] Abdul Mueed Hafiz, Shabir Ahmad Parah, and Rouf Ul Alam Bhat. Attention mechanisms and deep learning for machine vision: A survey of the state of the art. arXiv preprint arXiv:2106.07550, 2021.

[37] Hongju Cheng, Zizhen Yang, Xiaoqi Zhang, and Yang Yang. Multimodal sentiment analysis based on attentional temporal convolutional network and multi-layer feature fusion. IEEE Transactions on Affective Computing, 14(4):3149–3163, 2023.

[38] Feiran Huang, Xiaoming Zhang, Zhonghua Zhao, Jie Xu, and Zhoujun Li. Image–text sentiment analysis via deep multimodal attentive fusion. Knowledge-Based Systems, 167:26–37, 2019.

[39] IK Salman Al-Tameemi, Mohammad-Reza Feizi-Derakhshi, Saeed Pashazadeh, and Mohammad Asadpour. Multi-model fusion framework using deep learning for visual-textual sentiment classification. Computers, Materials & Continua, 76(2):2145–2177, 2023.

[40] Nan Xu and Wenji Mao. Multisentinet: A deep semantic network for multimodal sentiment analysis.In Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, pages 2399–2402, 2017.

[41] Nan Xu, Wenji Mao, and Guandan Chen. A co-memory network for multimodal sentiment analysis. In The 41st international ACM SIGIR conference on research & development in information retrieval, pages 929–932, 2018.

[42] A Rakhlin. Convolutional neural networks for sentence classification. GitHub, 6:25, 2016.

[43] Siwei Lai, Liheng Xu, Kang Liu, and Jun Zhao. Recurrent convolutional neural networks for text classification. In Proceedings of the AAAI conference on artificial intelligence, volume 29, 2015.

[44] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 2818–2826, 2016.

[45] Nan Xu and Wenji Mao. Multisentinet: A deep semantic network for multimodal sentiment analysis. In Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, pages 2399–2402, 2017.

[46] Nan Xu. Analyzing multimodal public sentiment based on hierarchical semantic attentional network. In 2017 IEEE international conference on intelligence and security informatics (ISI), pages 152–154. IEEE, 2017.

[47] Nan Xu, Wenji Mao, and Guandan Chen. A co-memory network for multimodal sentiment analysis. In The 41st international ACM SIGIR conference on research & development in information retrieval, pages 929–932, 2018.

[48] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. Advances in neural information processing systems, 30, 2017.