SSRG International Journal of Electronics and Communication Engineering
ISSN: 2348-8549/ https://doi.org/10.14445/23488549/IJECE-V12I5P130

Volume 12 Issue 5, 363-378, May 2025

Original Article

Enhancing Multimodal Sentiment Prediction with Cross-Modal Attention and Adaptive Feature Weighting

Prashant Adakane[1], Amit Gaikwad[2]

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

[1]*Corresponding Author : prashantadakane2020@gmail.com*

Received: 20 March 2025 Revised: 21 April 2025 Accepted: 22 May 2025 Published: 27 May 2025

Abstract *- This study introduces a novel framework, Contextual Adaptive Cross-Modal Attention Fusion (CA-CMAF), designed to solve multimodal sentiment analysis's difficulties. The framework leverages dynamic modality fusion and cross-modal attention mechanisms to effectually integrate textual and visual data, enabling a more nuanced understanding of sentiment in heterogeneous datasets. By focusing on the interplay between modalities, CA-CMAF aims to increase the accuracy and interpretability of the sentiment prediction system. The proposed approach combines textual features extracted through BERT, which has been both pre-trained and subsequently fine-tuned, with visual features derived from VGG-16. Through a cross-modal attention technique, these modalities are fused and aligned, capturing fine-grained interactions between text and images. The attention mechanism computes attention scores to prioritize the most relevant aspects of each modality, depending on the context provided by others. Additionally, an adaptive learning mechanism dynamically adjusts the contribution of each modality, ensuring optimal fusion for sentiment classification. The model is optimized using a blended loss approach. One part of this approach is based on cross-entropy principles for sentiment classification. Another component includes a regularization term to ensure balanced modality contributions. Experimental findings on MVSA-Single and MVSA-Multiple benchmark datasets indicate that CA-CMAF surpasses current baseline methods in state-of-the-art performance. The framework shows significant boosts in performance for metrics like accuracy, F1-score, precision, recall, with readings of 91%, 89%, 90%, and 90%, respectively, particularly in scenarios where one modality is more informative than the other.*

Keywords *- Cross-Modal Attention, Adaptive learning, BERT, VGG-16, Dynamic modality integration.*

1. Introduction

Sentiment analysis has become an essential technique for understanding content posted by users across many platforms, including social media, e-commerce, and review sites. [1, 2] By analyzing textual and visual data, sentiment analysis enables businesses and researchers may learn more about user opinions, preferences, and emotions. [3] However, the increasing prevalence of multimodal data, where information is conveyed through a combination of text and images, poses significant challenges. [4] While text provides explicit sentiment cues, images often convey implicit emotional context, making it essential to effectively integrate both modalities for accurate sentiment analysis. [5] Despite the potential of multimodal data, analyzing it remains a complex task. Traditional methods often treat text and Images independently, missing complex modalities' interconnections. [6] Moreover, data types heterogeneity and the varying relevance of each modality to the sentiment task further complicate the analysis. These challenges highlight the need for advanced techniques that can seamlessly fuse and balance contributions from different modalities to improve sentiment analysis performance. [6-8]

Existing methods for multimodal sentiment analysis often fall short when it comes to effectively combining and balancing the contributions of text and images. Most approaches use static fusion techniques, which assume that each modality contributes equally, regardless of the input. [5, 9] This one-size-fits-all approach leads to suboptimal performance, especially when one modality (like text or visuals) is more informative than the other. [10]

The research gap, therefore, lies in the absence of a flexible and context-aware framework that can (i) capture fine-grained interactions between modalities, and (ii) adaptively balance their contributions to sentiment prediction. Moreover, existing models often overlook the need for interpretability and the importance of aligning features across modalities in a context-sensitive manner. [11]

To tackle these challenges, the proposed Contextual Adaptive Cross-modal Attention Fusion (CA-CMAF), a novel framework that dynamically fuses text and image features utilizing cross-modal attention. Novelty of this approach derives from its integrated use of cross-modal attention and

adaptive learning mechanisms, allowing the model to not only align multimodal features contextually but also dynamically adjust the contribution of each modality. Framework employs BERT, which is pre-trained for textual and VGG-16 for images-to extract high-quality features from each modality. [12]

A cross-modal attention mechanism aligns and fuses these features, allowing the model to concentrate on the most pertinent aspects of each modality based on the context provided by the other. An adaptive learning mechanism further ensures that the contributions of text and visuals are dynamically adjusted, optimizing the fusion process for sentiment prediction.

The primary contributions of this work are,

- Cross-Modal Attention Mechanism - Proposed mechanism aligns text and visual features and enhances the model's capability to understand complex data from diverse modalities by capturing fine-grained interactions.
- Adaptive Learning for Modality Fusion - The Adaptive learning approach dynamically balances the influence of textual and visual according to their significance, ensuring robust performance across diverse datasets.
- Comprehensive Evaluation - Comprehensive evaluations conducted on standard datasets show that the developed model achieves superior performance compared to current approaches, highlighting its proficiency in dealing with heterogeneous data and its applicability for practical use.

The structure of the manuscript is - Review of recent studies given in Section 2, their methodologies, and associated challenges. Section 3 introduces the methodology for an efficient sentiment prediction model. Section 4 presents results obtained from applying the proposed model, and Section 5 presents the conclusion of the manuscript.

2. Literature Review

The growing interest in multimodal sentiment analysis is driven by the availability of content in both textual and visual formats. Early research in this domain employed independent feature extraction for each modality, followed by simple fusion techniques such as concatenation, averaging, or summation. Zadeh et al. [13] presented Tensor Fusion Networks (TFN) to interplay between modalities through tensor products. However foundational, outlined approaches often failed to capture the deep semantic interplay between modalities, especially in emotionally nuanced content.

In this study, T. Zhu et al. [14] introduced a unique method for fine-grained image-text multimodal emotion categorization, MULSER - Multi-Level Semantic Reasoning network focuses on semantic links between words and objects. It employs graph attention modules for both image and text modalities to enhance feature extraction and interdependencies. The module for cross-modal attention fusion integrates these enhanced features for accurate emotion classification. Experimental outcomes reveal that MULSER outperforms cutting-edge methods, confirming its usefulness in the analysis of multimodal sentiment.

Z. Liu et al. [15] suggested CMAFusion, a cross-modal attention-driven framework for fusing infrared and visible images to integrate pictures with multi-layered text features and object properties. A cross-modal feature aggregation mechanism is introduced to properly integrate infrared and visible pictures' deep complementary features. Conversely, a composite loss function was developed to regulate similarity, texture, and structural properties. CMAFusion surpasses cutting-edge techniques in comprehensive comparison and generalization testing.

In another paper, T. Zhu et al. [16] suggested the utilization of cross-modal attention through its innovative ITIN framework, which includes mechanisms for aligning and integrating features from both text and visual. This method aims to improve the accuracy of multimodal sentiment analysis by fully leveraging interactions between different modalities.

In this study, X. Luo et al. [17] introduced attribute Word-to-Face (W2F) hybridization, utilizing attribute-word patterns valuable in meaningful data as input. New Generative Adversarial Network, Cross-Modal Attention Fusion (CMAFGAN), was developed to create faces using descriptive facial trait words. CMAFGAN is built on two key mechanisms: transformation of Word-Level Features (WFT) and fusion through Cross-Modal Attention (CMAF). These components analyze the relationship between visual elements and corresponding textual attributes. Testing was performed using CelebA and LFW datasets. The results confirmed that CMAFGAN delivers notable enhancements in generating realistic synthesized faces.

Y. Wang et al. [18] developed a method for multi-label image classification that merges cross-modal fusion techniques with generalized convolutional neural networks, incorporating attention-based modules to effectively capture both local and global label relationships. Their approach consists of three primary components.

The first is an attention-driven method for extracting relevant features. The second utilizes a Graph Convolutional Network (GCN) to capture relationships between labels through co-occurrence patterns. The third integrates information across modalities using a specialized fusion framework that employs factorized bilinear transformation for effective feature combination. Evaluations on MS-COCO and VOC2007 datasets demonstrate CFMIC's improved efficacy

and superior classification performance over existing advanced methods.

In this work, M. Xu et al. [19] presented a novel transformer-based framework, referred to as CMJRT, designed to learn joint representations for effective multimodal sentiment prediction. CMJRT leverages multilevel relationships among modalities to transfer combined projections from bimodal to unimodal frameworks. Cyclic transformation is utilized to generate bimodal combined projections, where one modality is transformed into another and back using encoders and decoders, ensuring alignment across modalities.

A cross-modal transformer further enhances unimodal representations by integrating insights from bimodality, exploring how modalities interact. Extensive testing on CMU-MOSI and CMU-MOSEI datasets confirms CMJRT's superior performance over existing methods.

H. Yu et al. [20] introduced a novel approach, Cross Attention for Cross-Modal Retrieval Method (CACRM). This novel approach is designed to achieve local alignment in image retrieval. CACRM utilizes a Cross Attention Model for capturing relational patterns. It focuses on linking different image regions with corresponding textual information. It constructs a similarity matrix, which is then pooled to derive global similarity. Evaluations on public datasets such as the dam inspection log, MS-COCO, and Flicker30k demonstrate CACRM's notable advancements over prior techniques.

T. Zhou et al. [21] introduced a multi-level model for cross-modality interactions, focusing on correlation and consistency between modalities for visual-textual sentiment predictions. The system utilizes a multi-level attention design. It captures semantic interactions and filters out noise, combined with a multimodal convolutional neural network. It improves joint representation. By incorporating transfer learning, the framework effectively handles noise in social data. Extensive experiments highlight its superior performance over existing methods. It underscores the significance of phrase-level text fragments in interacting with image regions.

H. Wen, S. You, and Y. Fu [22] introduced a novel approach for emotion recognition that leverages multiple modalities through a technique known as cross-modal dynamic convolution. The method models the temporal dimension of emotion-related information, capturing useful interactions while minimizing unrelated information. It addresses challenges related to limited data density and mismatches in emotion-associated signals, making it easier to identify and utilize cross-modal interactions. The method is stackable, achieving competitive performance compared to existing approaches.

S. Thuseethan et al. [23] introduced a deep learning-based approach for predicting sentiment across modalities, integrating key visual and high-attention textual cues to overcome limitations of indiscriminate modality fusion. The framework uses dual unimodal deep feature extractors to gather relevant features from images and text, coupled with a late fusion mechanism for sentiment prediction. It outperforms existing unimodal and basic multimodal methods, effectively utilizing interrelationships in multimodal web data to achieve precise prediction of sentiment.

F. Huang et al. [24] proposed a technique called DMAF, which focuses on sentiment prediction by effectively modeling the underlying interactions between visual content and textual data through attentive fusion mechanisms. DMAF utilizes distinct attention mechanisms for images and text, followed by an intermediate-level fusion architecture with multimodal attention. A late fusion strategy integrates these models to improve sentiment prediction. Extensive experiments confirm DMAF's effectiveness on partially labeled and expertly annotated corpus. It is showcasing superior performance in multimodal sentiment prediction.

Z. Wang et al. [25] proposed method called CAMP, which dynamically manages how data is shared and integrated across various input types to improve cross-modal comprehension. CAMP addresses negative pairings and irrelevant data using an adaptive gating mechanism, alongside detailed and fine-grained cross-modal attention. It introduces a strongest negative twofold cross-entropy loss for model fitting. It infers matching scores based on fused features instead of traditional joint embedding techniques. Results on COCO and Flickr30k datasets highlight CAMP's effectiveness, significantly outperforming existing methods.

In this study [26], the author proposed using a multiplicative approach in cross-modal feature modeling. This method enables the system to learn both high-level abstract patterns and specific contextual features together. Effective representation learning for multimodal data is essential for cross-modal extraction efficiency. Extensive experimentations validate that the introduced framework efficiently matches images and text with complex content, achieving cutting-edge cross-modal extraction outcomes on the MSCOCO corpus.

2.1. Motivation of Research

Despite advancements in attention mechanisms and fusion techniques, the field still faces challenges in dynamically balancing and contextually aligning multimodal inputs. Many existing models either:

- Employ static fusion, assuming equal modality relevance,
- Rely on single-level or coarse-grained alignment, ignoring nuanced cross-modal relationships, or
- Lack of mechanisms to adaptively adjust modality weights based on contextual informativeness.

These gaps reveal a need for a context-aware, adaptive framework that can:

- Capture intricate interactions between text and image data,
- Adjust fusion based on input-specific dynamics, and
- Offer interpretability alongside performance.

2.2. Problem Statement

Existing literature lacks a unified framework that can simultaneously:

- Capture fine-grained semantic alignment across text and image modalities,
- Dynamically adapt to varying levels of modality informativeness, and
- Maintain interpretability and generalizability across diverse datasets.

To address these issues, the proposed Contextual Adaptive Cross-Modal Attention Fusion (CA-CMAF) is a framework that integrates BERT and VGG-16 extracted features using cross-modal attention, complemented by an adaptive learning mechanism and a blended loss function. This design ensures modality-aware fusion and optimized sentiment prediction, surpassing the limitations of static or non-contextual models.

3. Proposed Methodology

This research is conducted with the intention to provide a remarkable improvement in the Multimodal Sentiment Predictions. Figure 1 shows the entire flow of suggested work. Proposed framework, Contextual Adaptive Cross-Modal Attention Fusion (CA-CMAF), is a deep learning model for multi-modal inputs designed to effectively integrate textual and visual information. It utilizes the BERT model for contextualized text feature extraction and the VGG-16 model for visual feature extraction from images. These are pre-trained models. Both modalities are projected into a common dimensional space to ensure compatibility. The cross-modal attention technique is employed. This technique extracts interactions from textual and image features. It enables the model to emphasis pertinent parts of each modality. The framework further refines these features using an Adaptive Learning layer, which incorporates an Attention-Based Fusion Layer to dynamically compute attention weights and adaptively combine the attended features with the original visual features. This fusion process ensures that the final representation optimally balances the contributions of both modalities. Finally, class probabilities are produced by sending through combined features through a dense layer by means of softmax activation. Training of the model is performed utilizing sparse categorical cross-entropy loss and the Adam optimizer. This makes it highly effective for multi-modal sentiment analysis tasks requiring joint understanding of text and images.

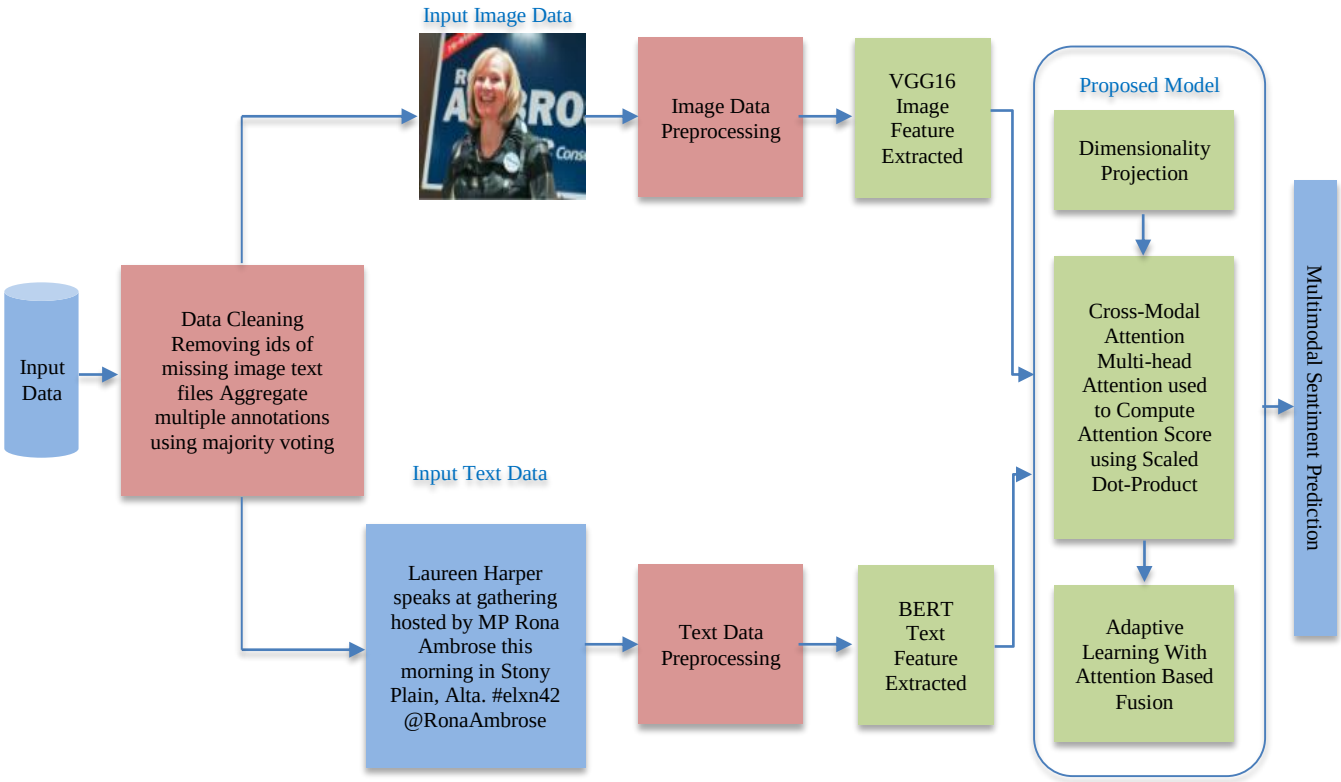


Fig. 1 The proposed system model

The problem of multimodal sentiment prediction is formulated as a categorization activity where the input consists of a pair of text and image data, and the output is a sentiment label. Specifically, given a textual input T (e.g., a sentence or paragraph) and an associated image I, the aim is to predict the sentiment label y (positive, negative, neutral). Introduced a framework designed to implement a deep learning model. This model efficiently merges textual and pictorial features. Its objective is to predict sentiment labels with high accuracy. This involves extracting meaningful features from both modalities, modeling their interactions, and fusing them into a unified representation for sentiment classification.

3.1. Textual Feature Extraction using BERT

Linguistic characteristic retrieval is performed using BERT, the transformer-based framework developed for natural language processing tasks. Figure 2 depicts BERT's workflow for processing textual input. BERT captures contextual embeddings by considering the bidirectional sentence structure and word dependencies. Sentiment prediction greatly benefits from this capability. [27]

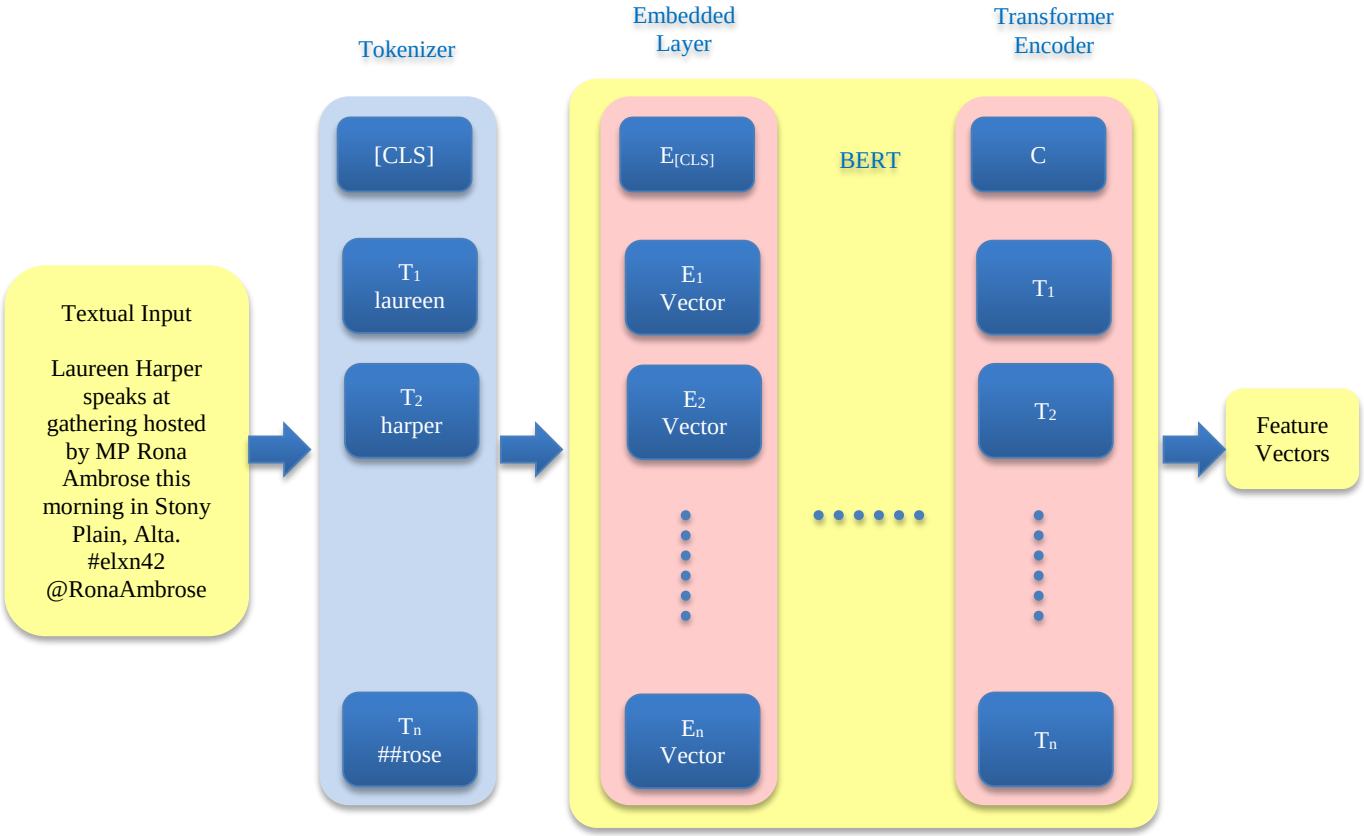


Fig. 2 BERT's workflow for processing textual input

Input text, such as "Laureen Harper speaks at gathering hosted by MP Rona Ambrose this morning in Stony Plain, Alta. #elxn42 @RonaAmbrose" is first tokenized into words or subwords. Using BERT tokenizer, the input text is tokenized. Tokens such as [CLS] and [SEP] are included. These tokens mark the beginning and end of sequences. These tokens are then converted into embeddings in the Embedded Layer, where every token is encoded as a vector capturing its semantic interpretation. Between the Embedded Layer and the Transformer Encoder, BERT adds Positional Encoding and Segment Embeddings to incorporate positional and segment information, ensuring the model understands word order and sentence boundaries. The Transformer Encoder is BERT's core component. Self-attention layers process embeddings first. Feed-forward layers then follow this step. The model uses self-attention to weigh the importance of each token relative to others. This creates contextualized feature vectors. These vectors capture bidirectional context. This process enables BERT to obtain contextual embeddings. For sentiment analysis, embedding associated with the CLS token is typically utilized as a textual feature representation. [28] Textual features are extracted using the Bert-base-uncased model, pre-trained and obtained via the HuggingFace Transformers library. The BERT model generates a pooled feature vector of size 768, which is further projected to 512 dimensions using a dense layer with ReLU activation in all our multimodal fusion models. The input to BERT consists of tokenized text sequences, which include input IDs and attention masks, created using the BertTokenizer.

3.1.1. BERT Configuration

In the proposed framework, BERT is implemented via the TFBertModel from the Hugging Face Transformers library. A custom wrapper, referred to as BertLayer, is employed to abstract the internal components and return the pooled output, which corresponds to the contextual embedding of a special classification (CLS) token.

Model accepts two primary inputs: tokenized input IDs and corresponding attention masks, each defined with a flexible shape to accommodate variable-length sequences. These inputs are tokenized using the pretrained BERT tokenizer (e.g., bert-base-uncased). The output from BERT is a fixed-length dense vector representing the semantic content of the input text. The resulting vector is transformed using a dense ReLU-activated layer, projecting it into a 512-dimensional latent space to align with the image feature dimensions.

3.2. Visual Feature Extraction using VGG-16

Extraction of visual features using VGG-16 is carried out, an architecture built upon CNN. This framework was produced by Visual Geometry Group, which is located at the University of Oxford. This is pre-trained on the ImageNet dataset. Convolutional layers are 13, and fully connected layers are 3 in VGG-16. [29] As shown in Figure 3, the model extracts high-level visual features from images. It removes the final fully connected layers. It utilizes output from the last convolutional layer. Input images are resized to 224x224 pixels, with normalization applied to the range [0, 1]. A series of convolutional layers is applied to the network. ReLU activation functions follow each of these layers. This process extracts features like edges, textures, and complex patterns from the image.

The network's depth is expanded by accumulating these convolutional layers. This gives the network the ability to understand better sophisticated features. To refine feature maps, max-pooling layers are used following multiple convolutional layers. They help in adjusting spatial dimensions, ensuring effective feature representation. This helps manage computational complexity and prevents overfitting. The derived features are reshaped into a one-dimensional vector for further processing. Subsequently, the data is processed using dense neural units to learn higher-level representations.

This forms a high-level representation of the image. Finally, the output layer produces a vector. We utilized the VGG16 model from Keras applications, omitting the top fully connected layers by setting include_top=False, and leveraging pre-trained ImageNet weights for feature extraction. Final convolutional outputs of the model are passed through a GlobalAveragePooling2D layer to generate a fixed-length feature vector. This vector is subsequently projected to 512 dimensions using a dense layer with ReLU activation.

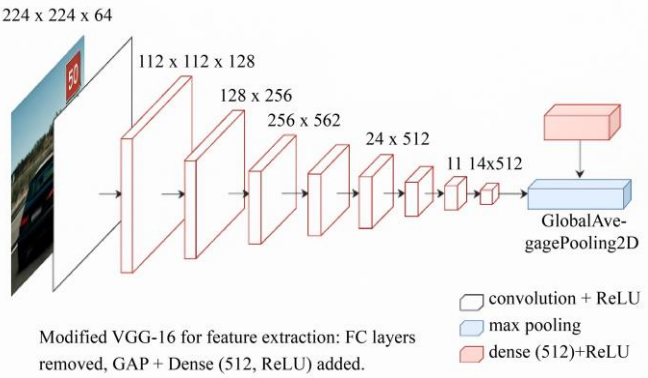


Fig. 3 Architecture of the modified VGG-16 model used for visual feature extraction

3.2.1. VGG-16 Configuration

For processing visual data, the model employs VGG-16 architecture, initialized with weights learned from a large-scale image classification benchmark and implemented using TensorFlow's Keras library. The network is initialized without its top classification layers, focusing solely on feature extraction. Input images are resized to a uniform resolution of 128×128 pixels with three color channels and normalized to fall within the [0, 1] range to standardize the input.

The convolutional feature maps produced by VGG-16 are subjected to global average pooling to obtain a compact, fixed-length feature vector. Following the approach used for textual features, the resulting vector undergoes transformation through a fully connected layer comprising 512 neurons and a ReLU function, ensuring alignment with text feature representation. This transformation enables seamless fusion of the two modalities.

3.3. Dimensionality Projection

Text features, extracted using BERT, are high-dimensional contextualized embeddings, which are 768 dimensions, while image features, extracted using VGG-16, are typically lower-dimensional, which are 4096 dimensions. To align these features, both are passed through separate dense layers (fully connected layers) that project them into a shared D-dimensional space where D = 512. This projection resolves the dimensionality mismatch and scales the features to a common representation, making them suitable for combination. Non-linearity is introduced by applying the ReLU activation function following the projection step, enhancing the representational capacity of features. The batch normalization technique is used to stabilize training and ensure consistent feature scaling. The output of this step is a pair of text and image features with the same dimensionality (D), which are now ready for further processing. This alignment is essential for the subsequent Cross-Modal Attention mechanism, which relies on the features being in the same space to compute meaningful interactions, and for the Adaptive Learning layer, which dynamically balances the contributions of both modalities.

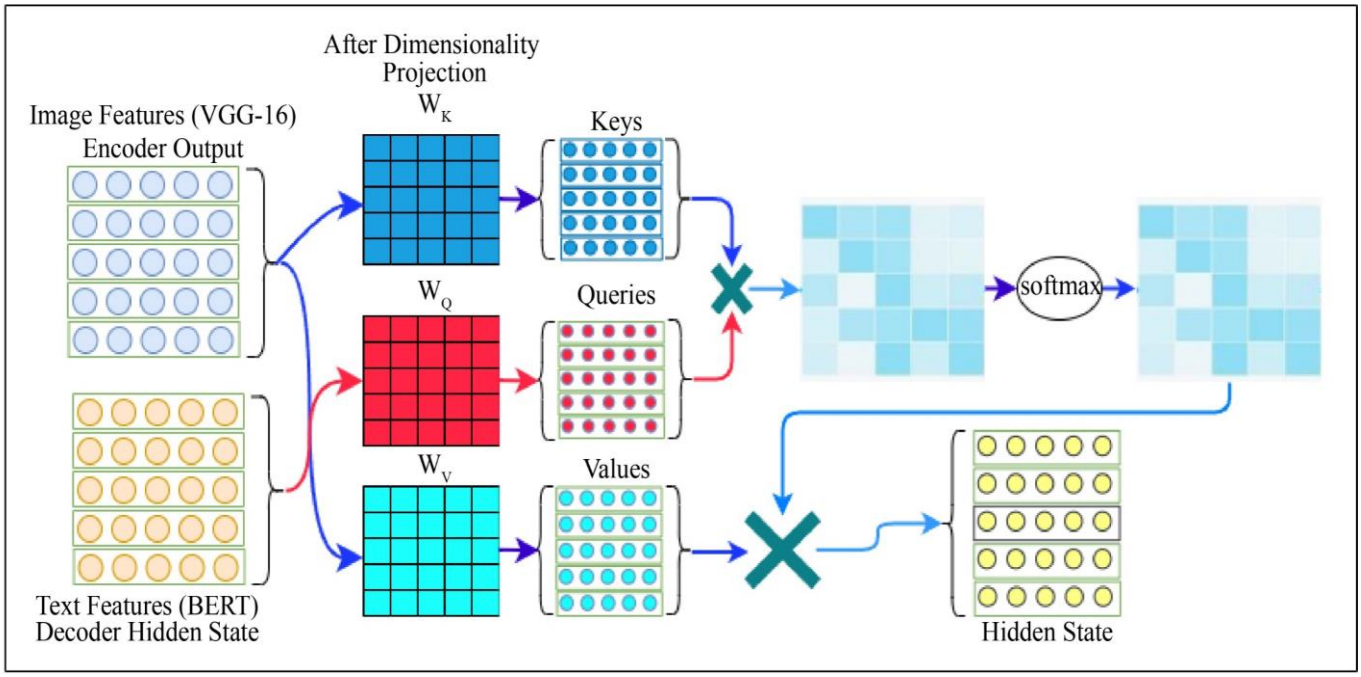


Fig. 4 Cross-modal attention mechanism

3.4. Cross-Modal Attention

Cross-modal attention supports the model to dynamically emphasis on most relevant measures of text and image, improving the representation of multi-modal data. As depicted in Figure 4, this mechanism models interactions between text and image features by taking two inputs, the Encoder Output, which is image features from VGG-16, and the Decoder Hidden State, which is text features from BERT.

Encoder Output is used as Keys (WK) and Values (WV), while the Decoder Hidden State is used as Queries (WQ). The mechanism employs a multi-head attention layer, where the text features (projected as Queries) are used to "ask" which parts of the image features (projected as Keys and Values) are relevant. As shown in Equation 1, attention scores are quantified by means of the scaled dot-product attention formula:

$$\text{Attention}(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right) V \qquad (1)$$

Here, Q represents the projected text features. These features are called Queries. K and V represent projected image features. K stands for Keys and V stands for Values. d_k is the dimensionality of key vectors. The dot product is scaled by $\sqrt{d_k}$ to stabilize gradients. This process is enhanced by a multihead attention layer, which applies attention mechanisms multiple times in parallel to capture diverse interactions between modalities. All attention head outputs are joined together by concatenation. They are then fed into a dense layer. This process produces the final attention output. Attended features are then weighted and combined to create a cross-modal representation, which captures the most relevant interactions between text and image. This representation is passed to the Adaptive Learning Layer for further refinement, enabling the model to dynamically balance and integrate multi-modal information. By focusing on semantically related parts of text and image, the Cross-Modal Attention Mechanism improves the model's capacity to analyze and manage multi-modal data in a meaningful way.

3.5. Adaptive Learning with Attention-Based Fusion Layer

This is an important component designed to refine and dynamically balance the contributions of text and image features. After the Cross-Modal Attention Mechanism generates a unified representation by capturing interactions between modalities, this layer further enhances the feature representation. It employs an Attention-Based Fusion approach, where attention weights are dynamically computed to decide the significance of each modality (text and image) for sentiment classification.

These weights are used to adaptively combine the attended cross-modal features with the original visual features from VGG-16, ensuring that the final representation optimally balances the contributions of both modalities. The fusion process is guided by the model's learning objective, allowing it to prioritize the most relevant features. The refined features are then passed to the next module for further processing, such as sentiment classification.

3.6. Sentiment Classification

The sentiment classification module in the CA-CMAF model takes the refined features from the Adaptive Learning

with Attention-Based Fusion Layer and maps them to sentiment classes. This module consists of one or more fully connected layers that transform the combined features into a format suitable for classification. Softmax activation function in the final layer generates a probability distribution across sentiment classes: positive, negative, and neutral.

The model is trained by means of categorical cross-entropy loss, which quantifies the difference between predicted and actual sentiment labels. Adam optimizer is used for gradient descent, ensuring efficient and stable training. To prevent overfitting, techniques such as dropout and L2 regularization are applied, enhancing the model's generalization ability. This module ensures that the CA-CMAF model can effectively perform sentiment classification based on the integrated multi-modal features.

3.7. Training Parameters

The training parameters for the model are as follows: The image input size is set to 128 × 128 × 3, ensuring a manageable resolution for processing. The text sequence length is variable, with sequences dynamically padded, and no fixed maximum length is enforced. The batch size is 32, providing balance between computational efficiency and model performance. Training occurs over 50 epochs, allowing sufficient time for the model to learn.

Adam optimization algorithm is employed with a 1e-4 learning rate is chosen for effective weight updates. Model utilizes sparse categorical cross-entropy as a loss function, making it appropriate for addressing multi-class classification problems. To mitigate overfitting, the model applies a 0.3 dropout rate, along with an L2 regularization factor of 1e-4.

ReLU is utilized as an activation function within hidden layers, while the final layer uses softmax activation to produce probabilities for classification. The output dimension of the dense layer is set to 512, enabling the model to capture complex features.

4. Results and Discussion

4.1. Experimental Arrangement

The model is developed using a Python script on a Windows 11 system with 8 GB RAM.

4.2. Dataset Used

MVSA datasets are utilized, which comprise two databases: MVSA-Single and MVSA-Multiple. [30] These are freely available. MVSA-Single includes 5,129 text-image pairings extracted from Twitter, each annotated by only one annotator who assigns any of the three sentiments: neutral, positive, or negative. Conversely, MVSA-Multiple includes 19,600 text-image pairings, assessed by three separate annotators to guarantee variation in sentiment classification. [32, 33]

For MVSA-Multiple, the actual label for individual modality is determined using majority vote among 3 annotators, ensuring the text or image becomes credible if minimum 2 annotators share the same sentiment. To maintain data of excellent quality, excluded tweets where text and visual annotations are inconsistent, particularly when one.

The annotation is positive, and the other is negative. If one annotation falls into the neutral category, while following one expresses either a positive or negative sentiment, the overall emotion polarity of the multimodal tweeter post is determined by the positive (or negative). This preprocessing step helps refine the MVSA-Single dataset, resulting in 4,511 text-image pairs.

Similarly, the MVSA-Multiple input dataset is refined to contain 17,024 text-image pairs. Preprocessing steps, as described by Xu and Mao [31], removed image-text pairs where the images are entirely unrelated to the text labels, further enhancing dataset quality. Table 1 summarizes the statistical details of preprocessed MVSA-Single and MVSA-Multiple datasets.

Table 1. Distribution of datasets based on sentiment categories

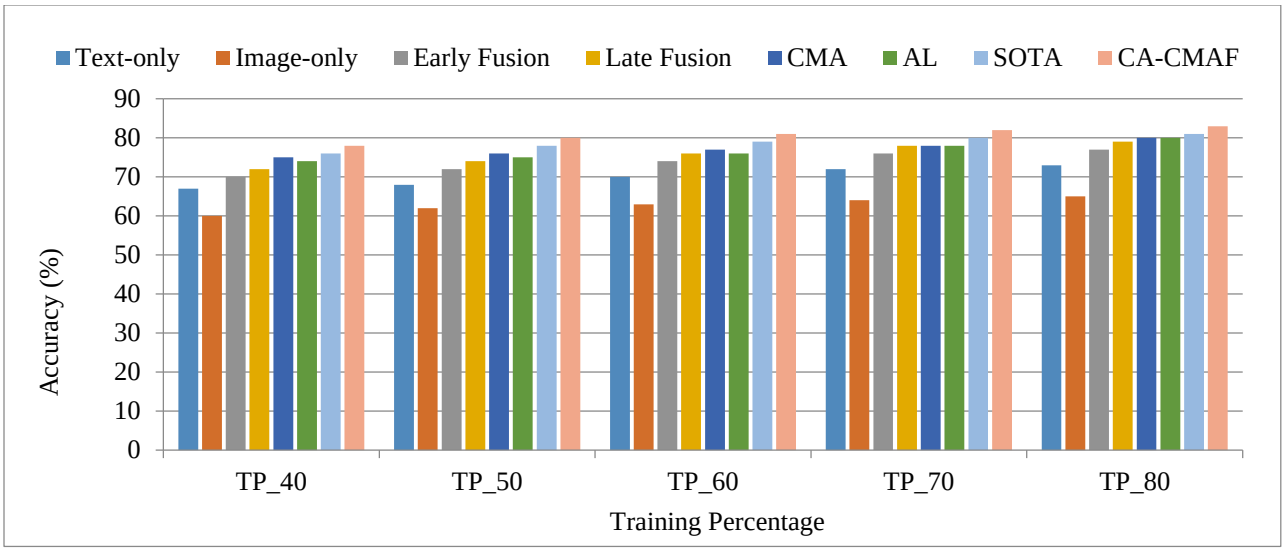
Dataset Type	Positive	Neutral	Negative	Total Samples
MSVA-Single	2,683	470	1,358	4,511
MSVA-Multiple	11,318	4,408	1,298	17,024

4.3. Comparative Analysis

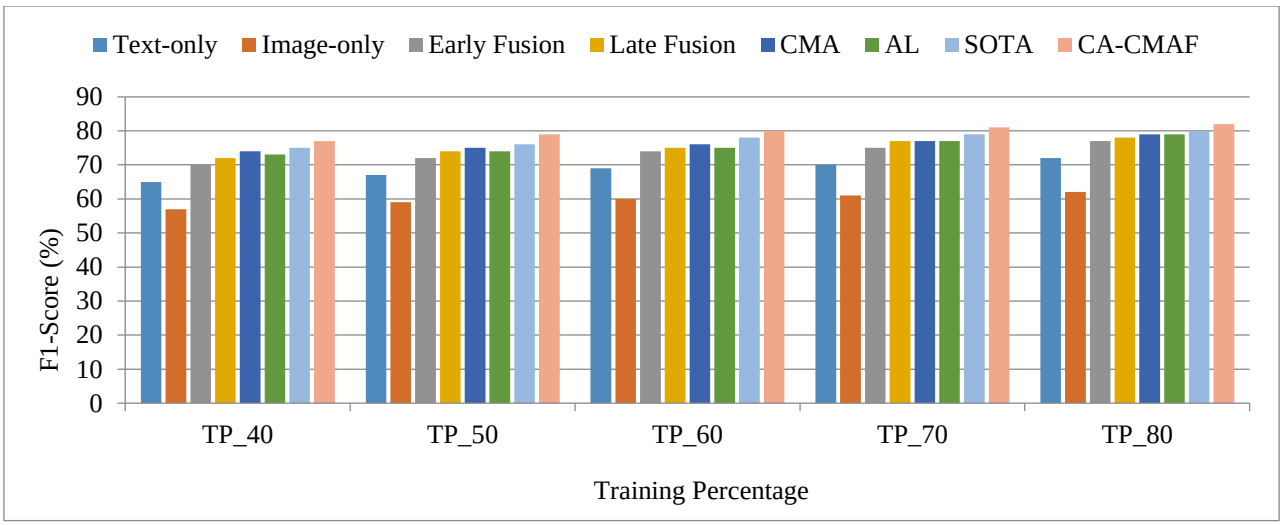
In comparative evaluation, the effectiveness of the CA-CMAF model is assessed in comparison to other models, including Text-only (BERT), Image-only (VGG-16), Early Fusion (Concatenation), Late Fusion (Averaging), Cross-Modal Attention (CMA), State-Of-The-ART (SOTA), and Adaptive Learning (AL) model.

Figure 5 displays the comparative analysis of the CA-CMAF model on the MVSA-Single dataset across different Training Percentages (TP). When comparing the proposed CA-CMAF framework with other SOTA models across all TP values 40, 50, 60, 70, and 80, it is observed that CA-CMAF consistently surpasses all competing methods.

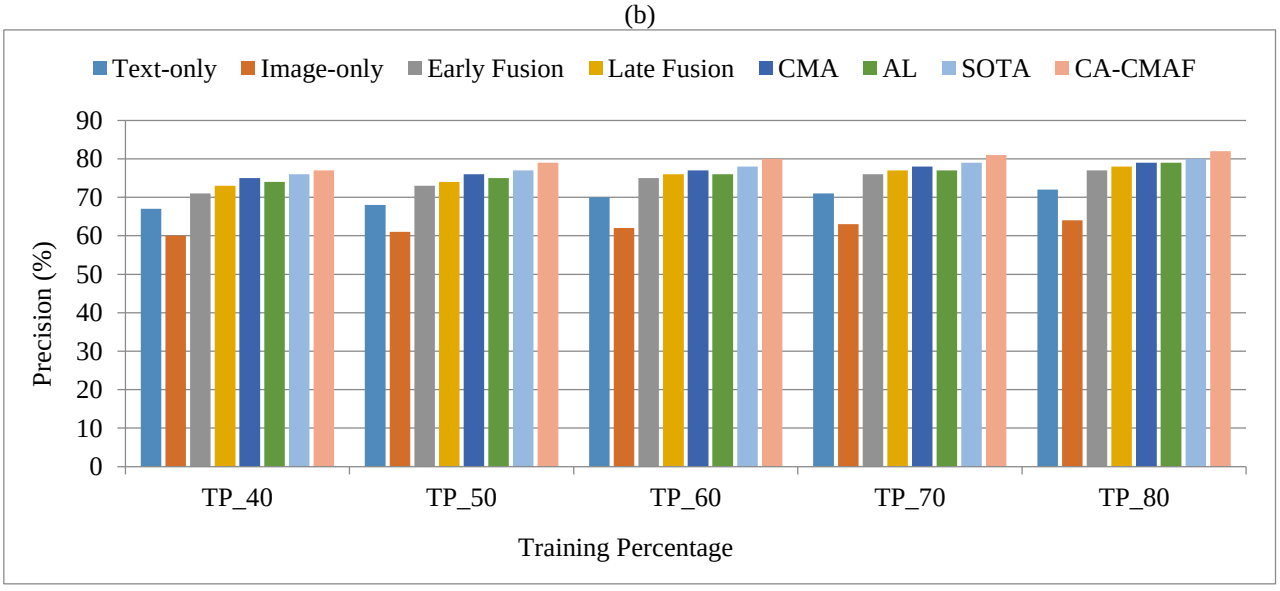
As shown in Figure 5(a), at TP of 40, CA-CMAF achieves an accuracy of 78%, surpassing SOTA by 2%. CA-CMAF reaches an accuracy of 83% at TP of 80, outperforming SOTA by 2% and CMA by 3%. Compared to Text-only (BERT) and Image-only (VGG-16), CA-CMAF demonstrates a substantial improvement, achieving 10% and 18% higher accuracy, respectively.


Text-only
Image-only
Early Fusion
Late Fusion
CMA
AL
SOTA
CA-CMAF
Accuracy (%)
90
80
70
60
50
40
30
20
10
0
TP_40
TP_50
TP_60
TP_70
TP_80
Training Percentage

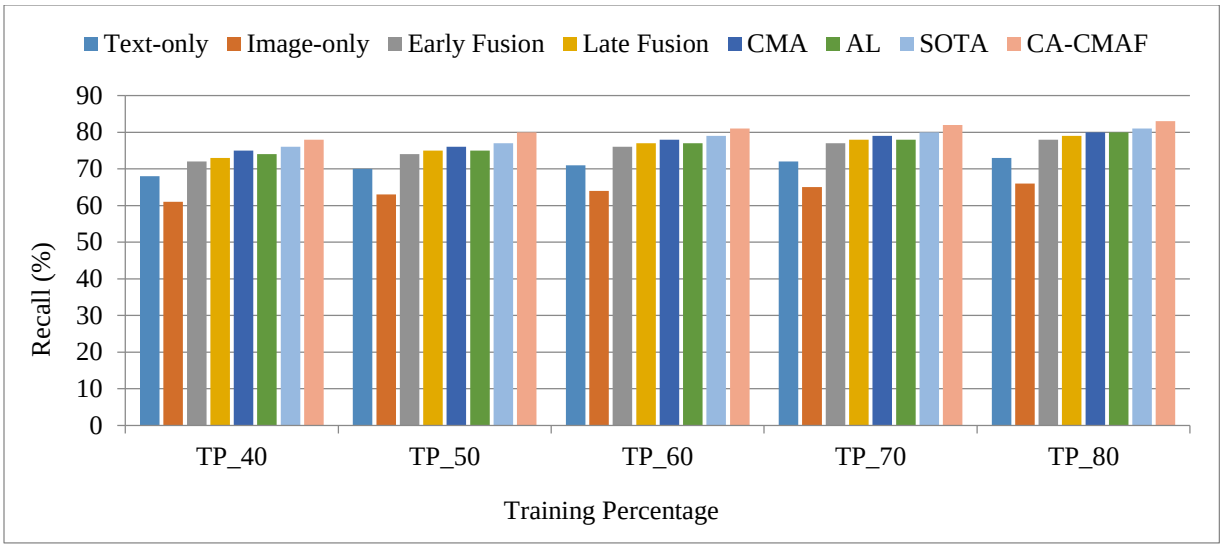
(a)


Text-only
Image-only
Early Fusion
Late Fusion
CMA
AL
SOTA
CA-CMAF
F1-Score (%)
90
80
70
60
50
40
30
20
10
0
TP_40
TP_50
TP_60
TP_70
TP_80
Training Percentage

(b)


Text-only
Image-only
Early Fusion
Late Fusion
CMA
AL
SOTA
CA-CMAF
Precision (%)
90
80
70
60
50
40
30
20
10
0
TP_40
TP_50
TP_60
TP_70
TP_80
Training Percentage

(c)



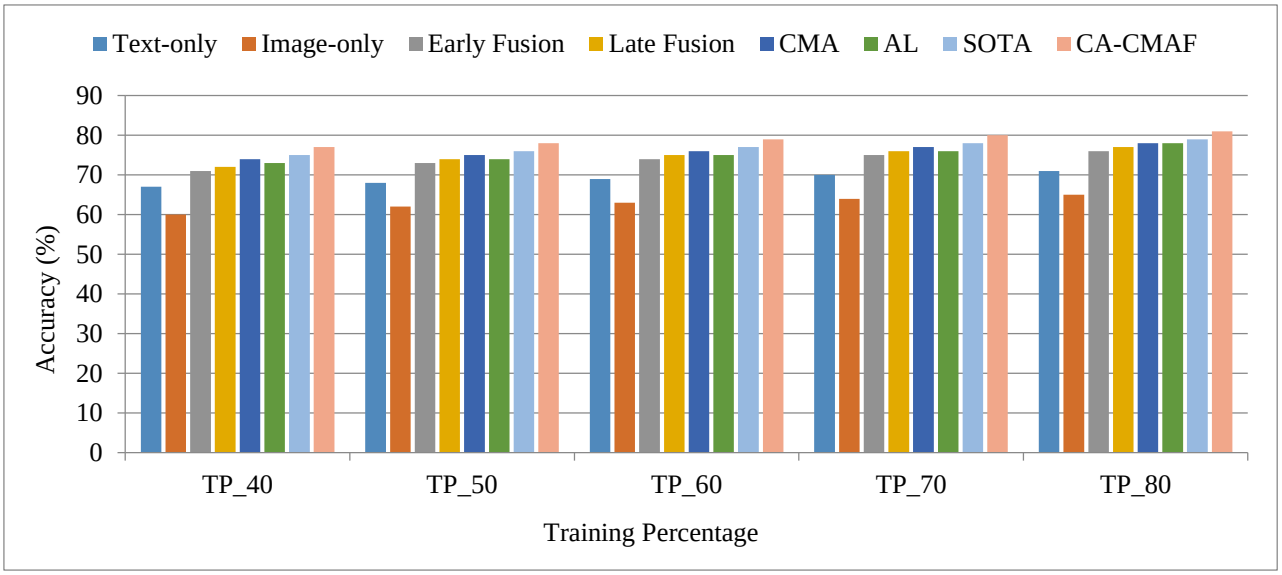
(d)

Fig. 5 Comparative analysis on MVSA-single dataset across various TP values, illustrating a) Accuracy, b) F1-score, c) Precision, and d) Recall.

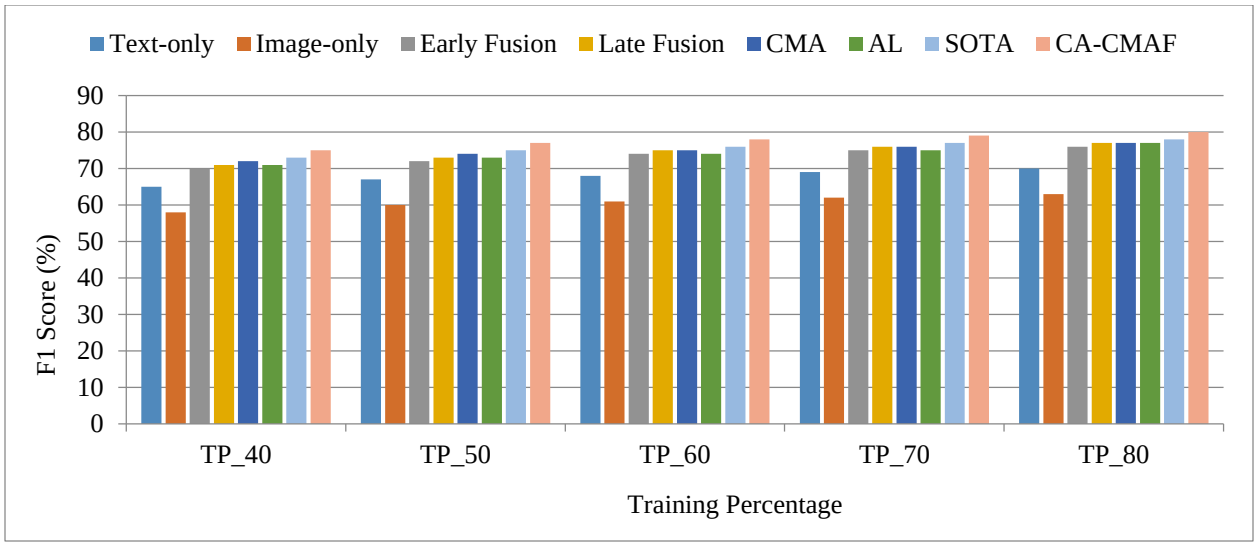
This indicates that CA-CMAF is the most effective in integrating multimodal data for accurate predictions. As shown in Figure 5(b), at TP of 40, CA-CMAF achieves an F1-Score of 76%, surpassing SOTA by 3%. F1-Score of 81% at TP of 80, outperforming SOTA by 3% and CMA by 3%. Compared to Text-only (BERT) and Image-only (VGG-16), CA-CMAF demonstrates a significant improvement, achieving 10% and 18% higher F1-Score, respectively.

When compared to Early Fusion (Concatenation) and Late Fusion (Averaging), CA-CMAF outperforms them by 7% and 5%, respectively. As shown in Figure 5(c), at TP of 80, CA-CMAF reaches a precision of 82%, outperforming SOTA by 2% and Cross-Modal Attention (CMA) by 3%. Compared to Text-only (BERT) and Image-only (VGG-16), CA-CMAF demonstrates improvement, achieving 10% and 18% higher precision, respectively.

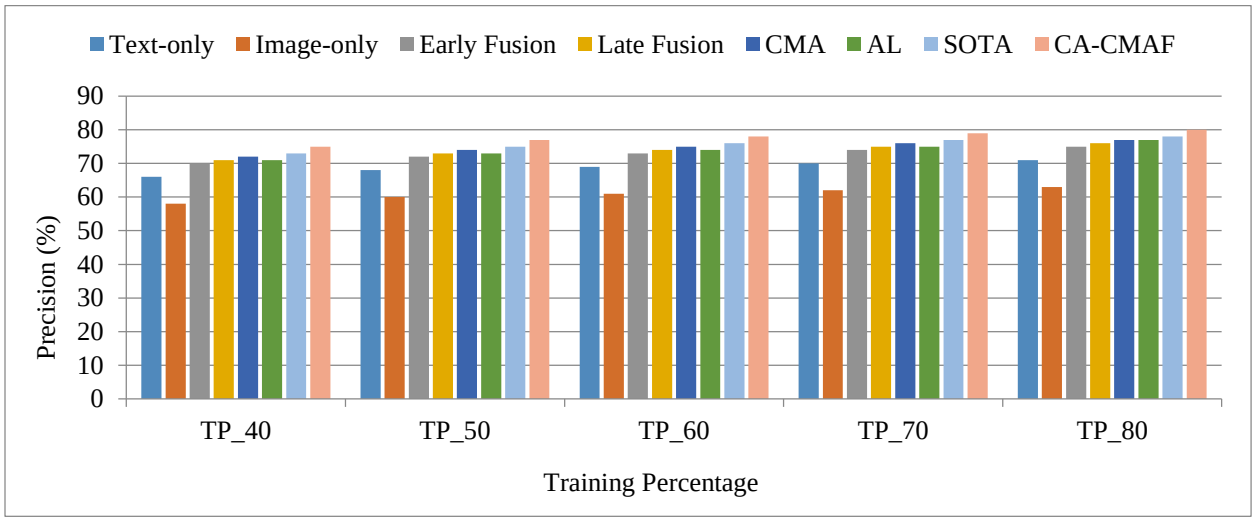
This indicates that CA-CMAF is the most effective in minimizing false positives and maximizing true positives. When compared to fusion-based methods like Early Fusion (Concatenation) and Late Fusion (Averaging), CA-CMAF outperforms them by 7% and 5%, respectively. As shown in Figure 5(d), at TP of 80, outperforming SOTA by 3% and Cross-Modal Attention (CMA) by 4%. CA-CMAF demonstrates a substantial improvement, achieving 9% and 16% higher recall, compared to Text-only (BERT) and Image-only (VGG-16), respectively. When compared to Early Fusion (Concatenation) and Late Fusion (Averaging), CA-CMAF outperforms them by 6% and 5%, respectively.



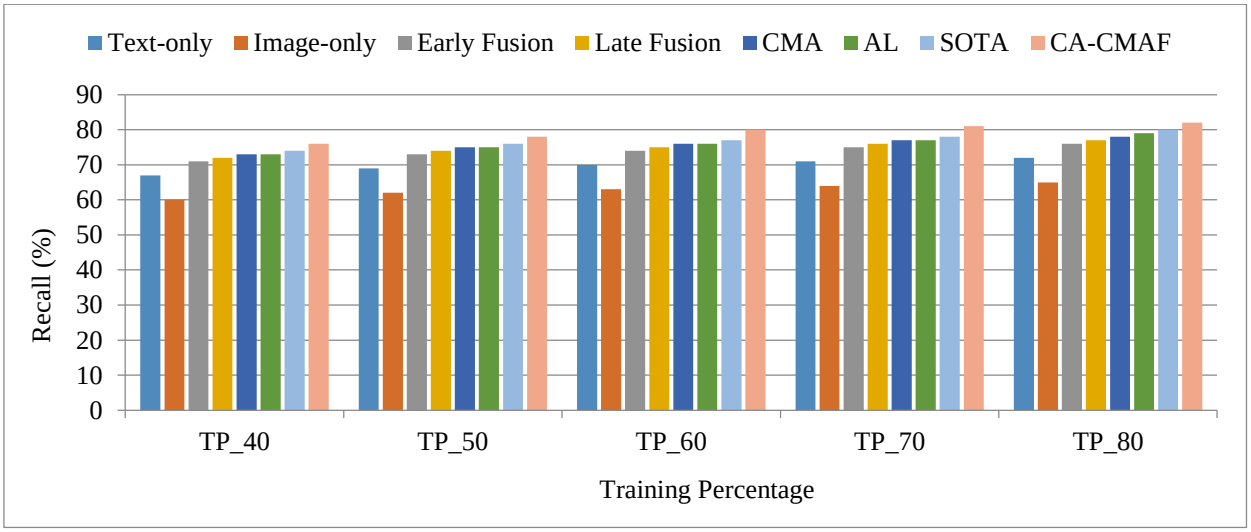
(a)



(b)



(c)



(d)

Fig. 6 Comparative analysis on MVSA-multiple dataset across various TP values, illustrating a) Accuracy, b) F1-score, c) Precision, and d) Recall.

Figure 6 displays a systematic comparison of the presented CA-CMAF model on the MVSA-Multiple dataset across different TP. After comparing the proposed CA-CMAF framework with other SOTA models across all TP values 40, 50, 60, 70, and 80, it is observed that CA-CMAF consistently surpasses all competing methods. As shown in Figure 6(a), at TP of 80, CA-CMAF achieves an accuracy of 81%, surpassing SOTA by 2% and CMA by 3%. Compared to Text-only (BERT) and Image-only (VGG-16), CA-CMAF achieves 9% and 18% higher accuracy, respectively. This indicates that CA-CMAF is the most effective in integrating multimodal data for accurate predictions.

As shown in Figure 6(b), F1-Score of CA-CMAF reaches 79% at TP of 80, outperforming SOTA by 3% and CMA by 3%. CA-CMAF achieves 9% and 18% higher F1Score, compared to Text-only (BERT) and Image-only (VGG-16), respectively. When compared to Early Fusion (Concatenation) and Late Fusion (Averaging), CA-CMAF outperforms them by 7% and 5%, respectively. As shown in Figure 6(c), CA-CMAF reaches a precision of 80% at TP of 80, outperforming SOTA by 2% and CMA by 3%. Compared to Text-only (BERT) and Image-only (VGG-16), CA-CMAF achieves 10% and 18% higher precision, respectively. When Early Fusion (Concatenation) and Late Fusion (Averaging), CA-CMAF outperforms them by 7% and 5%, respectively.

Table 2. Comparative analysis of various methods at TP of 90

Method	MVSA-Single				MVSA-Multiple			
	Acc.	F1-Score	Precision	Recall	Acc.	F1-Score	Precision	Recall
Text-only (BERT)	78%	76%	77%	78%	76%	74%	75%	76%
Image-only (VGG-16)	70%	68%	69%	70%	70%	66%	68%	68%
Early Fusion (Concatenation)	81%	79%	80%	81%	79%	77%	78%	79%
Late Fusion (Averaging)	83%	81%	82%	83%	81%	78%	80%	81%
Cross-Modal Attention (CMA)	85%	83%	84%	85%	84%	82%	83%	83%
Adaptive Learning (AL)	84%	82%	83%	84%	82%	80%	81%	82%
State-of-the-Art (SOTA)	86%	84%	85%	86%	84%	82%	83%	84%
CA-CMAF	**92%**	**90%**	**91%**	**91%**	**91%**	**89%**	**90%**	**90%**

As shown in Figure 6(d), CA-CMAF reaches a recall of 81% at TP of 80, outperforming SOTA by 2% and CMA by 3%. CA-CMAF achieves 9% and 18% higher recall, compared to Text-only (BERT) and Image-only (VGG-16), respectively. This indicates that CA-CMAF is the most effective in capturing true positives and minimizing false negatives. When compared to fusion-based methods like Early Fusion (Concatenation) and Late Fusion (Averaging), CA-CMAF outperforms them by 7% and 5%, respectively.

As shown in Table 2, a comparison of the performance metric for various models on two different datasets, i.e., MVSA-Single and MVSA-Multiple, for a training percentage of 90. Across all metrics - Accuracy, F1-Score, Precision, and Recall- the CA-CMAF model consistently outperforms other SOTA methods, demonstrating its superiority in integrating and leveraging multimodal data. This makes CA-CMAF the most effective and reliable model for tasks requiring high performance in multimodal data integration.

The confusion matrix presented in Figure 7(a) shows the classification performance of the proposed CA-CMAF model on the MVSA-Single dataset across three sentiment categories: Positive, Neutral, and Negative. The model demonstrates strong predictive performance for the Positive and Negative classes, with 2,450 and 1,218 correct predictions, respectively. The Neutral class also shows reasonable accuracy, with 420 correct classifications. Misclassifications are relatively low, indicating the model's effectiveness in discerning sentiment polarity in multimodal data. The matrix highlights the model's capability to handle class imbalances while maintaining high precision and recall across sentiment categories.

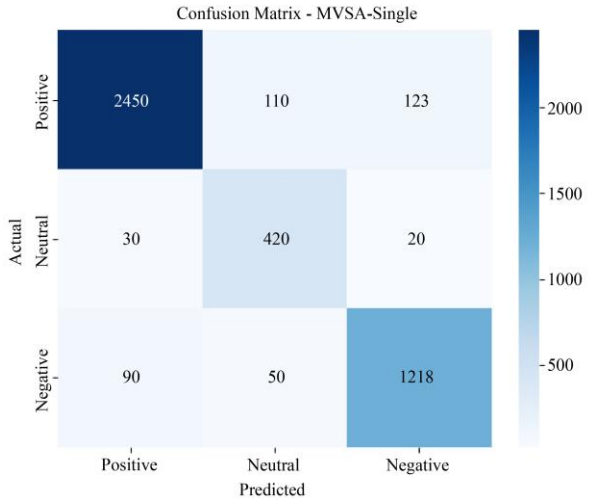


Fig. 7(a) Confusion matrix for MVSA-single dataset

The confusion matrix depicted in Figure 7(b) presents the sentiment classification performance of the CA-CMAF model on the MVSA-Multiple dataset. The model accurately identifies a majority of Positive instances (10,400), followed by Neutral (3,800) and Negative (1,108) sentiments. Misclassifications are relatively minimal, with some overlap between Neutral and Positive classes, indicating the inherent ambiguity in sentiment interpretation for certain samples. Despite the larger and more diverse MVSA-Multiple dataset, the proposed CA-CMAF model maintains strong generalization across all sentiment categories, reflecting its robustness and adaptability to complex multimodal data.

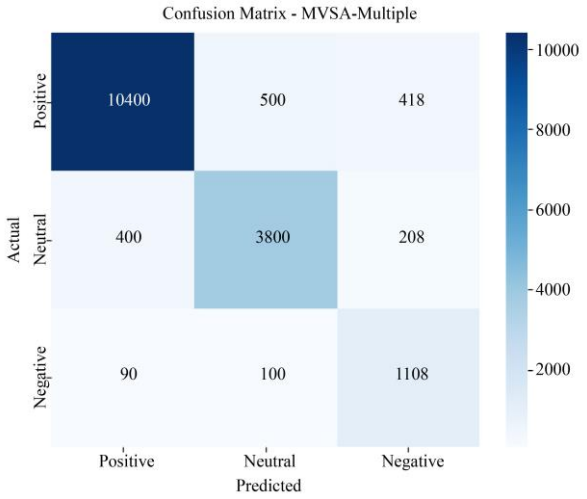


Fig. 7(b) Confusion matrix for MVSA-multiple dataset

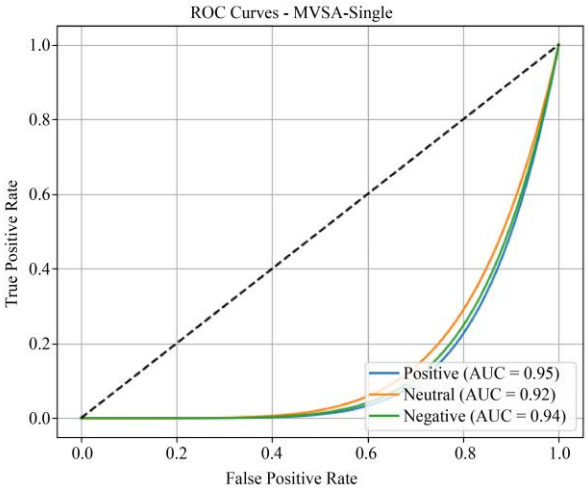


Fig. 8(a) ROC curves for MVSA-single dataset

The ROC curves plotted in Figure 8(a) for the MVSA-Single dataset illustrate the model's classification effectiveness across the three sentiment categories. Each sentiment class, positive, neutral, and negative, exhibits strong discriminative performance, achieving AUC scores of 0.95, 0.92, and 0.94, respectively. These scores indicate that the model performs with high accuracy, particularly for the positive sentiment category. The curves ascend steeply toward the ideal region of the plot, indicating effective sentiment prediction with minimal classification errors and strong detection accuracy.

ROC analysis depicted for MVSA-Multiple dataset in Figure 8(b) reflects strong classification capabilities across all three sentiment categories. The model achieves Area Under the Curve (AUC) scores of 0.94 for positive, 0.91 for neutral, and 0.93 for negative sentiments. These values indicate that the model maintains reliable discrimination power for each class, with slightly higher performance for positive and negative categories. The close proximity of the curves to the top-left corner demonstrates a favorable balance between sensitivity and specificity, reinforcing the model's effectiveness in handling multi-sentiment visual-textual inputs.

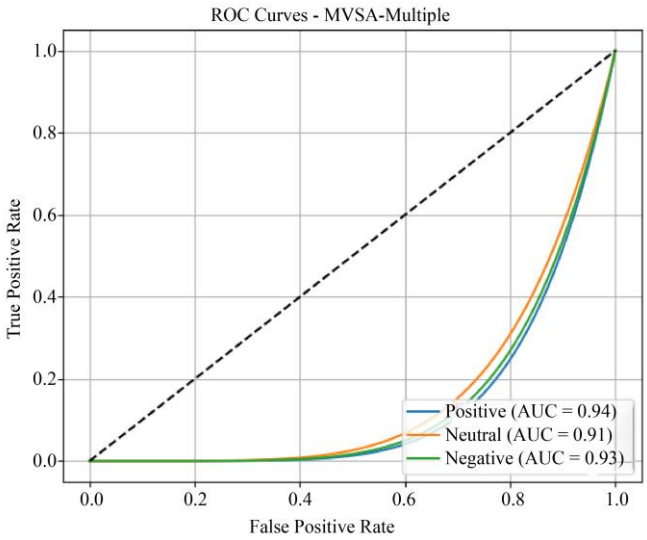


Fig. 8(b) ROC curves for MVSA-multiple dataset

4.4. Discussion

Existing sentiment analysis models face significant challenges, particularly when dealing with multimodal information such as text and images. These challenges include effective integration of diverse data types, handling noisy datasets, addressing context-dependent sentiments, and managing the complexities of sarcasm and ambiguity. Traditional approaches, such as Early Fusion and Late Fusion, often struggle to seamlessly combine textual and visual modalities, leading to suboptimal performance in capturing sentiment cues. Models like Cross-Modal Attention (CMA) and Adaptive Learning (AL) may encounter scalability issues and high computational costs, limiting their practical applicability in real-world scenarios. Content-based methods and conventional neural networks frequently fail to capture the nuanced interactions between textual and visual features, especially when sentiments are expressed implicitly or sarcastically.

The proposed Contextual Adaptive Cross-Modal Attention Fusion (CA-CMAF) framework addresses these challenges by leveraging a new cross-modal attention technique that dynamically adapts to the contextual relevance of textual and visual features. This approach ensures a seamless integration of multimodal data, enhancing the model's capability to capture complex sentiment cues that are often expressed across different modalities. By incorporating adaptive learning techniques, CA-CMAF optimizes the fusion process, improving both exploration and exploitation capabilities. The framework's ability to handle noisy and ambiguous datasets is further enhanced through its robust feature extraction and fusion mechanisms, which mitigate the impact of irrelevant or conflicting information from different modalities. Additionally, CA-CMAF addresses the challenge of context-dependent sentiments by effectively utilizing both textual and visual features, even when sentiments are expressed implicitly or sarcastically.

5. Conclusion

The proposed Contextual Adaptive Cross-Modal Attention Fusion (CA-CMAF) framework represents a significant advancement in multimodal sentiment predictions, addressing critical challenges such as seamless integration of textual and visual data, handling noisy and ambiguous datasets, and capturing context-dependent sentiments, including sarcasm and implicit expressions. By leveraging a novel cross-modal attention mechanism and adaptive learning techniques, CA-CMAF ensures robust and accurate sentiment classification, even in complex and noisy scenarios. The framework's ability to dynamically adapt to the contextual relevance of multimodal features enhances its performance, making it scalable and computationally efficient.

As demonstrated by outcomes on MVSA-Single and MVSA-Multiple datasets, CA-CMAF outperforms all competing methods across all metrics. On MVSA-Single, CA-CMAF achieves an accuracy of 92%, surpassing SOTA by significant margins. Similarly, on MVSA-Multiple, proposed CA-CMAF achieves an accuracy of 91%, consistently outperforming SOTA and other baseline methods like CMA and Late Fusion. These results highlight the framework's superior capability to integrate and leverage multimodal data for sentiment analysis. CA-CMAF focuses on context-aware sentiment analysis, leading to more precise sentiment predictions. It establishes a strong reference point for handling multimodal data in sentiment-related tasks. By integrating multimodal fusion, adaptive learning, and attention mechanisms, CA-CMAF offers a robust and versatile solution. This makes it highly effective for various datasets and practical applications. It contributes to the advancement of sentiment analysis models, ensuring greater accuracy and dependability moving forward.

Building upon the promising outcomes of the CA-CMAF framework, several avenues for future exploration are evident. First, extending the model to support additional modalities such as audio or video could further enrich sentiment interpretation in social media and real-world scenarios. Moreover, adapting the CA-CMAF framework for low-resource languages or multilingual datasets could broaden its applicability and inclusiveness. Incorporating explainable AI techniques may also enhance interpretability, helping users understand the contribution of each modality in decision-making. Additionally, exploring lightweight or real-time versions of CA-CMAF would be valuable for deployment in mobile or edge computing environments. Further research may also focus on improving the model's ability to handle evolving language trends, such as slang, emojis, and memes, which frequently appear in social media content. Integrating reinforcement learning could enhance adaptive feature selection based on feedback loops or user preferences. Another promising direction involves incorporating emotional intensity detection to better capture subtle sentiment variations. Finally, collaborative learning approaches, such as federated learning, could help build more privacy-preserving and decentralized sentiment models without sharing raw data.

References

[1] Danlei Chen et al., "Joint Multimodal Sentiment Analysis Based on Information Relevance," *Information Processing & Management*, vol. 60, no. 2, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[2] 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*, vol. 22, no. 6, pp. 1-30, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[3] Ankita Gandhi et al., "Multimodal Sentiment Analysis: A Systematic Review of History, Datasets, Multimodal Fusion Methods, Applications, Challenges and Future Directions," *Information Fusion*, vol. 91, pp. 424-444, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[4] Israa Khalaf Salman Al-Tameemi et al., "A Comprehensive Review of Visual-Textual Sentiment Analysis from Social Media Networks," *arXiv Preprint*, pp. 1-44, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[5] Lianting Gong, Xingzhou He, and Jianzhong Yang, "An Image-Text Sentiment Analysis Method Using Multi-Channel Multi-Modal Joint Learning," *Applied Artificial Intelligence*, vol. 38, no. 1, pp. 1-20, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[6] Songning Lai et al., "Multimodal Sentiment Analysis: A Survey," *Displays*, vol. 80, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[7] P. Ganesh Kumar, "A Context-Sensitive Multi-Tier Deep Learning Framework for Multimodal Sentiment Analysis," *Multimedia Tools and Applications*, vol. 83, pp. 54249-54278, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[8] Hongchan Li, Yantong Lu, and Haodong Zhu, "Multi-Modal Sentiment Analysis Based on Image and Text Fusion Based on Cross-Attention Mechanism," *Electronics*, vol. 13, no. 11, pp. 1-22, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[9] Ringki Das, and Thoudam Doren Singh, "Multimodal Sentiment Analysis: A Survey of Methods, Trends, and Challenges," *ACM Computing Surveys*, vol. 55, no. 135, pp. 1-38, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[10] Zhibang Quan et al., "Multimodal Sentiment Analysis Based on Cross-Modal Attention and Gated Cyclic Hierarchical Fusion Networks," *Computational Intelligence and Neuroscience*, vol. 2022, no. 1, pp. 1-12, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[11] Weijing Chen, Linli Yao, and Qin Jin, "Rethinking Benchmarks for Cross-Modal Image-Text Retrieval," *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, Taipei, Taiwan, pp. 1241-1251, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[12] Priyanka Meel, and Dinesh Kumar Vishwakarma, "Multi-Modal Fusion Using Fine-Tuned Self-Attention and Transfer Learning for Veracity Analysis of Web Information," *Expert Systems with Applications*, vol. 229, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[13] Amir Zadeh et al., "Tensor Fusion Network for Multimodal Sentiment Analysis," *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Copenhagen, Denmark, pp. 1103-1114, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[14] Tong Zhu et al., "Multimodal Emotion Classification with Multi-Level Semantic Reasoning Network," *IEEE Transactions on Multimedia*, vol. 25, pp. 6868-6880, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[15] Zhichao Liu et al., "CMAFusion: Cross Modal Attention Based End-to-End Infrared and Visible Image Fusion Network," *2023 7th CAA International Conference on Vehicular Control and Intelligence*, Changsha, China, pp. 1-6, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[16] Tong Zhu et al., "Multimodal Sentiment Analysis with Image-Text Interaction Network," *IEEE Transactions on Multimedia*, vol. 25, pp. 3375-3385, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[17] Xiaodong Luo et al., "CMAFGAN: A Cross-Modal Attention Fusion Based Generative Adversarial Network for Attribute Word-to-Face Synthesis," *Knowledge-Based Systems*, vol. 255, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[18] Yangtao Wang et al., "Cross-Modal Fusion for Multi-Label Image Classification with Attention Mechanism," *Computers and Electrical Engineering*, vol. 101, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[19] Meng Xu et al., "CMJRT: Cross-Modal Joint Representation Transformer for Multimodal Sentiment Analysis," *IEEE Access*, vol. 10, pp. 131671-131679, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[20] Huimin Yu et al., "CACRM: Cross-Attention Based Image-Text Cross Modal Retrieval," *2022 IEEE Eighth International Conference on Big Data Computing Service and Applications (BigDataService)*, Newark, CA, USA, pp. 137-142, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[21] Tao Zhou et al., "Visual-Textual Sentiment Analysis Enhanced by Hierarchical Cross-Modality Interaction," *IEEE Systems Journal*, vol. 15, no. 3, pp. 4303-4314, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[22] Huanglu Wen, Shaodi You, and Ying Fu, "Cross-Modal Dynamic Convolution for Multi-Modal Emotion Recognition," *Journal of Visual Communication and Image Representation*, vol. 78, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[23] Selvarajah Thuseethan et al., "Multimodal Deep Learning Framework for Sentiment Analysis from Text-Image Web Data," *2020 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*, Melbourne, Australia, pp. 267-274, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[24] Feiran Huang et al., "Image–Text Sentiment Analysis via Deep Multimodal Attentive Fusion," *Knowledge-Based Systems*, vol. 167, pp. 26-37, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[25] Zihao Wang et al., "CAMP: Cross-Modal Adaptive Message Passing for Text-Image Retrieval," *2019 IEEE/CVF International Conference on Computer Vision*, Seoul, Korea (South), pp. 5763-5772, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[26] Jiuxiang Gu et al., "Look, Imagine and Match: Improving Textual-Visual Cross-Modal Retrieval with Generative Models," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, pp. 7181-7189, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[27] Javad Hassannataj Joloudari et al., "BERT-Deep CNN: State of the Art for Sentiment Analysis of COVID-19 Tweets," *Social Network Analysis and Mining*, vol. 13, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[28] Mahsa Hadikhah Mozhdehi, and AmirMasoud Eftekhari Moghadam, "Textual Emotion Detection Utilizing a Transfer Learning Approach," *The Journal of Supercomputing*, vol. 79, pp. 13075-13089, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[29] Raheel Siddiqi, "Effectiveness of Transfer Learning and Fine Tuning in Automated Fruit Image Classification," *Proceedings of the 2019 3rd International Conference on Deep Learning Technologies*, Xiamen China, pp. 91-100, 2019. [CrossRef] [Google Scholar] [Publisher Link]

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

[31] 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*, Singapore, pp. 2399-2402, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[32] Ranjan Satapathy et al., "Phonetic-Based Microtext Normalization for Twitter Sentiment Analysis," *2017 IEEE International Conference on Data Mining Workshops*, New Orleans, LA, USA, pp. 407-413, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[33] 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]