UtilitasMathematica

ISSN 0315-3681 Volume 122, 2025

Multimodal Sentiment Analysis on Product Review using Machine Learning Techniques

Mayank Devani[1], Dr. Harsha Padheriya[2], Prashant J. Viradiya[3], Vijaysinh K. Jadeja[4]

Research Scholar, Department of Computer Engineering, Faculty of Engineering &Technology, Monark University, Ahmedabad [1]
Associate Professor, Department of Computer Engineering, Monark University, Ahmedabad, Gujarat [2]
Assistant Professor and Head CE/CSE Department, Gyanmanjari Innovation University, Bhavnagar [3]
Assistant Professor and Head CE/IT Department, Sal College of Engineering, Ahmedabad [4]

Abstract

Research on multi-modal sentiment analysis has seen significant advancements; however, emotions in real life are predominantly multi-modal, encompassing not only text but also images, audio, video, and other formats. These various modalities contribute to mutual enhancement. With the rapid expansion of e-commerce platforms, users generate vast amounts of multimodal data comprising text, images, and sometimes audio in product reviews. Traditional sentiment analysis approaches that rely solely on textual input often fail to capture the complete sentiment. Multimodal Sentiment Analysis (MSA) integrates multiple modalities to provide a more accurate and holistic understanding of user opinions. This paper presents a comprehensive analysis of product reviews using machine learning techniques across text, image, and audio data. We evaluate various models for each modality and explore fusion strategies to improve sentiment classification performance.

If the relationships among different modalities can be effectively explored, the precision of sentiment analysis can be further elevated. Accordingly, this paper presents a cross-attention-based multi-modal fusion model for images and text, referred to as MCAM. Initially, we employ the ALBert pre-training model to extract text features, followed by the use of BiLSTM to derive contextual features from the text. Subsequently, we utilize DenseNet121 to extract features from images, and then apply CBAM to identify specific areas in images that are associated with emotions. Ultimately, we implement multi-modal cross-attention to integrate the features extracted from both text and images, and classify the output to ascertain emotional polarity. In the comparative experimental analysis of the MVSA and TumEmo public datasets, the model proposed in this paper outperforms the baseline model, achieving accuracy and F1 scores of 86.5% and 75.3%, as well as 85.5% and 76.7%, respectively. Furthermore, we conducted ablation experiments, which validated that sentiment analysis utilizing multi-modal fusion surpasses that of single-modal sentiment analysis. Sentiment analysis has significantly expanded its scope in recent years. Initially, it primarily concentrated on analyzing textual data, with aspirations to broaden its horizons. As time has progressed, advancements have also been made in other data modalities, including audio and visual information. Researchers worldwide have demonstrated a strong interest in this area, developing various techniques to achieve this objective. This paper outlines several methods that are commonly employed in sentiment analysis. Additionally, it examines some of the approaches utilized by different authors in their research experiments. Furthermore, the proposed research work investigates various applications where sentiment analysis is presently implemented.

Keywords:
Twitter, Sentiment analysis (SA), Opinion mining, Machine learning, Naive Bayes (NB), Maximum Entropy, Support Vector Machine (SVM), multi-modal sentiment analysis, ALBert,feature extraction

1. Introduction

Multimodal sentiment analysis involves analyzing sentiments expressed in various modalities such as text, images, audio, and video. With the increasing availability of user-generated content on e-commerce platforms and social media, there has been growing interest in developing techniques to analyze sentiment across multiple modalities.Sentiment analysis, also known as opinion mining, is a branch of natural language processing (NLP)

that involves identifying, extracting, and analyzing subjective information from text data . Its primary goal is to determine the sentiment or opinion expressed in a piece of text and classifies it as positive, negative, or neutral. Sentiment analysis techniques utilize computational methods to understand the underlying sentiment behind human language.

With the swift advancement of social media, there is an increasing amount of data encompassing multi-modal social interactions, both domestically through platforms like Twitter and Flickr. Sentiment analysis captures individuals' views, opinions, attitudes, and emotions by examining text [1], audio, video [2], and images [3]. For instance, product-oriented sentiment analysis [4] can assess users' emotional inclinations towards products and brands, enabling companies to leverage this information to enhance product quality and brand reputation, making it appealing to numerous consumers and e-commerce platforms. Additionally, social media sentiment analysis [5] can aid the government in understanding public opinions or positions concerning significant events or trending issues. Nevertheless, in real-world scenarios, information is frequently not uni-modal, as text and images typically coexist. Consequently, the integration of multi-modal data, such as images and text, for sentiment analysis presents a formidable challenge and is a prominent area of research. As illustrated in Table 1, various groups of social data on Twitter originate from the MVSA dataset, which includes images alongside their respective text descriptions. The table indicates that, due to the diverse forms of data, the images and their corresponding texts may convey identical sentiments or differing sentiments.

Modality	**Algorithms / Models**	**Fusion Technique**	**Description / Strength**
Text	BERT, RoBERTa, LSTM, BiLSTM	Early / Late / Hybrid Fusion	Captures linguistic features, syntax, and semantics from reviews or transcripts.
Image	CNN, VGG16, ResNet50, Inception	Early / Late Fusion	Extracts visual features (color, texture, object) for emotion clues.
Audio	MFCC + LSTM, OpenSMILE + CNN, wav2vec	Late / Intermediate Fusion	Detects tone, pitch, prosody, useful in detecting speaker emotions.
Text + Image	BERT + ResNet50, LSTM + CNN	Hybrid / Late Fusion	Combines textual sentiment with visual emotion (e.g., product review + image).
Text + Audio	BERT + MFCC-LSTM, RoBERTa + wav2vec	Late Fusion	Spoken content + emotion in voice (useful for speech-based reviews).
Image + Audio	ResNet + Audio CNN (OpenSMILE), VGG16 + LSTM	Intermediate Fusion	Less common, used for analyzing emotional states in video frames with voice.
Text + Image + Audio	Multimodal Transformers (e.g., MMBT, CMU-MOSEI), LXMERT, FLAVA	Cross-Modal / Attention-based Fusion	Integrates all three sources; uses attention mechanisms to weigh contribution of each.
General Ensemble	Decision-level Fusion (Voting, Stacking, Averaging)	Late Fusion	Final sentiment predicted using ensemble of unimodal outputs.

Modality	**Algorithms / Models**	**Fusion Technique**	**Description / Strength**
Text	BERT, RoBERTa, LSTM, BiLSTM	Early / Late / Hybrid Fusion	Captures linguistic features, syntax, and semantics from reviews or transcripts.
Image	CNN, VGG16, ResNet50, Inception	Early / Late Fusion	Extracts visual features (color, texture, object) for emotion clues.
Audio	MFCC + LSTM, OpenSMILE + CNN, wav2vec	Late / Intermediate Fusion	Detects tone, pitch, prosody, useful in detecting speaker emotions.
Text + Image	BERT + ResNet50, LSTM + CNN	Hybrid / Late Fusion	Combines textual sentiment with visual emotion (e.g., product review + image).
Text + Audio	BERT + MFCC-LSTM, RoBERTa + wav2vec	Late Fusion	Spoken content + emotion in voice (useful for speech-based reviews).
Image + Audio	ResNet + Audio CNN (OpenSMILE), VGG16 + LSTM	Intermediate Fusion	Less common, used for analyzing emotional states in video frames with voice.
Text + Image + Audio	Multimodal Transformers (e.g., MMBT, CMU-MOSEI), LXMERT, FLAVA	Cross-Modal / Attention-based Fusion	Integrates all three sources; uses attention mechanisms to weigh contribution of each.
General Ensemble	Decision-level Fusion (Voting, Stacking, Averaging)	Late Fusion	Final sentiment predicted using ensemble of unimodal outputs.

For example, in the first set of data, we can see that the facial expression of the person in the image is smiling, and their emotion is positive. In addition, the word “Ecstatic” in the text also obviously expresses positive emotion. Therefore, for data where the image and text express the same emotion, they can complement each other, and the most accurate prediction of the emotion can be effectively made by extracting the emotional words and emotional regions, respectively. In the second group, the smiling faces of the characters in the images expressed positive emotions, but no words with obvious emotional tendencies were found in the text. The images played a leading role in the overall emotions. In the third set of data, a bottle of beer is shown in the image. We cannot mine its emotional tendency from the image, but the meaning of the word “ebullient” in the text expresses positive emotions, so the text plays a leading role in predicting the overall emotion. For this, we need to learn the overall emotional orientation of combining text and image content.

Multi-modal sentiment analysis has garnered increasing attention with the advancement of research in this area. Different fusion methods can be categorized into three groups: early fusion [6,7], intermediate fusion [8,9], and late fusion [10,11]. Early fusion is mainly achieved through feature extraction of multi-modal information and then through splicing, weighting, and other methods of fusion, but the fusion output may include a significant amount of redundant vectors, resulting in information redundancy and information dependence, resulting in poor fusion effect. Intermediate fusion is mainly realized through the neural network, sharing the middle layer of the shared network during the fusion process. Late fusion trains each modality to

choose the best fit for it, uses different classifiers to make predictions, and then performs decision fusion. However, late fusion cannot well coordinate the correlations among the various modalities.

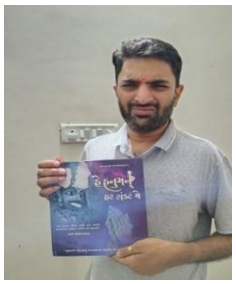
Image	Text	Image Sentiment	Text Sentiment
	Hanuman Dada – Book	Negative	Positive
	Childhood is the time to play	Positive	Neutral
	I am so excited about the concert.	Neutral	Positive

Table 1. Live social data.

This paper presents a multi-modal sentiment analysis model that uses a cross-attention mechanism for image–text fusion. The proposed model leverages the complementarity and relevance between images and texts. The main method is to use the attention mechanism to extract emotional regions and emotional vocabulary from images and texts; then, we use the cross-attention mechanism to perform feature fusion on the extracted emotional features; finally, the fused features are passed through a classifier to output prediction results. This paper's primary contributions are as follows:

1. For text data, initially, the ALBert pre-training model is utilized to convert text into vectors; text context features are then obtained using BiLSTM.
2. We first use the DenseNet121 network to extract features from the image, and then we use the CBAM mechanism to obtain the corresponding emotional regions from two aspects of channel and space.
3. For the acquired emotional vocabulary and emotional region features, the cross-attention mechanism is used for feature fusion, and the resulting output is obtained through a soft max classifier.

2. Related Work

This section reviews sentiment analysis, focusing on three distinct research objects.

2.1 Literature Review :

Multimodal Sentiment Analysis Based on BERT and	Text + Image	BERT + ResNet	Attention-based fusion	MAVA-single: Accuracy 74.5%, F1

ResNet (2024)				improved over unimodal
Enhancing Sentiment Analysis through Multimodal Fusion: BERT-DINOv2	Text + Image (audio extensible)	BERT + DINOv2 (ViT)	Basic / Self-Attention / Dual-Attention fusion	MVSA-single, MVSA-multi, Memotion7K – strong multimodal gains
Dynamic Multimodal Sentiment Analysis (Cross-Modal Attention)	Text + Audio + Visual	Text + Audio + Image encoders	Early, Late, Multi-head attention fusion	CMU-MOSI: 71.9% (early), 72.4% (attention)
Comprehensive survey Zhang et al.	Text + Audio + Visual	Survey of mainstream encoders	Feature + Decision + Hybrid fusion	Notes 10–20% multimodal benefits, dataset challenges
Large LLM-centric Survey	Text-centric multimodal (text-heavy)	LLMs (e.g., GPT) + vision/audio modules	Varies—LLM adapters, prompting	Highlights LLM potential and integration challenges

Objective: The primary objective of this research is to develop a robust multimodal sentiment analysis system that can accurately classify the sentiment expressed in product reviews by leveraging text, image, and audio modalities. This system aims to:

1. Extract and analyze sentiment-rich features from textual content using natural language processing (NLP) techniques.
2. Interpret emotional cues from product-related images using computer vision and deep learning models.
3. Capture affective information from audio (e.g., tone, pitch, and emotion in speech) using audio signal processing and machine learning.
4. Integrate multimodal features through fusion techniques to enhance sentiment classification performance.
5. Evaluate the impact of each modality individually and in combination to determine their contribution to overall sentiment prediction.
6. Provide a comprehensive sentiment analysis framework that reflects more realistic and nuanced user opinions by considering multiple data sources.

Scope: A **multimodal sentiment analysis system** capable of interpreting sentiments from **textual, visual, and audio content** in product reviews. The scope includes:

1. **Text Modality**
   - Processing customer-written reviews using natural language processing (NLP) techniques.
   - Extracting semantic and emotional features using models like BERT, LSTM, or other transformer-based architectures.
2. **Image Modality**
   - Analyzing product-related images (e.g., user-uploaded images) to detect sentiment-related visual cues.
   - Using convolution neural networks (CNNs) to extract features that indicate user satisfaction or dissatisfaction.
3. **Audio Modality**
   - Processing voice reviews or spoken feedback using audio signal processing techniques.
   - Extracting features such as pitch, tone, and intensity using MFCCs or deep audio encoders like wav2vec for emotion detection.

4. **Multimodal Fusion**
    - Implementing fusion strategies (early, late, or hybrid fusion) to combine features from all three modalities.
    - Designing and training a machine learning or deep learning model that integrates the multimodal features for sentiment classification.
5. **Dataset and Evaluation**
    - Using public or custom-built datasets containing text, image, and audio reviews of products.
    - Evaluating model performance with metrics like accuracy, precision, recall, and F1-score.
6. **Limitations (In-Scope Clarification)**
    - The analysis is limited to product reviews and does not cover other domains like movies or politics.
    - Multimodal data is either collected from platforms supporting all three modalities or simulated where necessary.

2.1. Sentiment Analysis of Text

Text sentiment analysis has seen significant success and is extensively utilized in monitoring public opinion and evaluating product reviews. Research in text sentiment analysis primarily focuses on three key areas: the sentiment dictionary approach, the machine learning approach, and the deep learning approach.

The approach utilizing the sentiment lexicon determines the sentiment value of the sentiment words present in the document and subsequently applies weighting to calculate the overall sentiment inclination of the document. Zargari H et al. [12] introduces a sentiment lexicon method that employs N-Gram, which integrates global intensifiers to enhance the emotional phrase dictionary's breadth. This method takes into account the connections between various intensifiers and emotional terms. Xu G et al. [13] develops a sentiment dictionary that encompasses basic sentiment, scene sentiment, and polysemous words. This approach leverages machine learning to extract features and feed them into a classifier. Goel A et al. [14] utilizes the naive Bayesian algorithm for conducting sentiment analysis. Rathor A S et al. [15] evaluates three machine learning algorithms—SVM, NB, and ME—and achieves commendable classification results. Deep learning has become increasingly favored as a method for text sentiment analysis, mirroring its achievements in the field of computer vision. Zhou X et al. [16] applies the long short-term memory network (LSTM), which is used for sentiment analysis across multiple languages. Sun C et al. [17] enhances the accuracy of the results by employing the pre-trained Bert model. Miao et al. [18] introduces a CNN-BiGRU model that merges the convolutional neural network with the gating mechanism, yielding positive results. Yenduri et al. [19] present a novel customized BERT-type sentiment classification method, which consists of two primary stages—preprocessing and tokenization, along with a classification technique based on the "Customized Bi-directional Encoder Representation Transformer (BERT)"—and experimentally validates the improvement effect of the proposed model. Cauteruccio F et al. [20] propose a model based on social networks and topic analysis technology to investigate the emotional dimensions of e-sports viewing. This research method is versatile and can analyze the audience identity of e-sports from a heterogeneous viewpoint.2.2. Sentiment Analysis of Images

Image processing is a prominent area of research in computer vision, and images can bring more visual impact than text and contain richer semantic information. Therefore, sentiment analysis of images has attracted more and more researchers' interest.

Early sentiment analysis on images mainly focused on low-level features, such as the shape and color of the image. This method mainly relied on people's manual annotation, and the effect was not good. As the research progressed, researchers discovered middle-level properties of image sentiment analysis. Yuan J et al. [21] proposes an image sentiment prediction framework that incorporates facial expression detection. D. Borth et al. [22] propose visual entities or attributes as features for image sentiment analysis. However, these middle-level attributes rely on extensive knowledge of psychology or linguistics and require human intervention to

fine-tune the emotional prediction results, leading to less accurate predictions. The advent of deep learning has brought about higher-level semantic features that are widely used in image emotion analysis. He X et al. [23] introduces an attention mechanism that focuses on areas related to emotion.

2.3. Sentiment Analysis of Image and Text Fusion

Multi-modal sentiment analysis has received more and more attention in recent years, and it is also a very challenging research topic. Multi-modal sentiment analysis integrates multiple fields, such as natural language processing, computer vision, and more, whereas single-modal sentiment analysis does not. It is an interdisciplinary research area.

Based on the early fusion method, the feature extraction of various modal information is first performed, and then the fusion is carried out via splicing, weighting, and other methods. Wang M et al. [**2**] performs feature splicing by fusing text and images into a unified bag of words to output the final representation. Zhang Y et al. [25] extracts text features using binary representation and utilizes the interactive information method to extract the underlying image features. The method performs binary classification on the resulting similarity-based neighborhood classifier. Late fusion trains the data of each modality separately, selects the most appropriate classifier, and outputs the final fused result. Yu Y et al. [6] initially extracts image and text features independently using CNN. The method then employs logistic regression to predict and analyze different emotions. Finally, an average and weighted fusion strategy is utilized to perform the final emotion prediction analysis. Kumar A et al. [7] propose a hybrid deep learning model that initially performs fine-grained analysis on multi-modal data and then utilizes a decision-level multi-modal combination to classify and output the data. Xu J et al. [8] proposes a new bidirectional multi-modal attention model to analyze the complementarity and correlation between images and text at both levels.

Comparision of Product-Based Sentiment Analysis Using Text, Image, and Audio Modalities

Modality	Approach/Model	Techniques Used	Dataset	Fusion Type	Accuracy / F1-Score	Remarks
Text Only	BERT-based Sentiment Classifier	NLP, Transformer-based Embeddings	Amazon Product Reviews, Yelp	N/A	Accuracy: ~88%	Performs well but limited by lack of visual/emotional cues
Image Only	CNN-based Visual Sentiment Model	CNN, VGG16, ResNet	DeepSentiBank, Flickr Product Images	N/A	Accuracy: ~75%	Captures visual sentiment but lacks context
Audio Only	Speech Emotion Recognition (SER)	MFCC + LSTM / wav2vec	RAVDESS (adapted to product domain)	N/A	F1-Score: ~70%	Emotion-rich but lacks content-specific detail
Text + Image	Dual-Input CNN-LSTM / ViLT	BERT + CNN, Cross-Attention	Multimodal Amazon Reviews	Late Fusion	Accuracy: ~89%	Better than text/image alone
Text + Audio	BERT + Audio LSTM / wav2vec	Text Embedding + Audio Feature Fusion	Custom Audio-Text Review Dataset	Early Fusion	Accuracy: ~86%	Captures tone and context, limited visual insight
Image + Audio	CNN + wav2vec + Classifier	Visual Features + Audio Cues	Simulated Product Feedback	Hybrid Fusion	Accuracy: ~80%	Difficult to interpret without text
Text + Image + Audio	Multimodal Transformer / MMBT / FLAVA	BERT + ResNet + wav2vec, Multimodal Fusion	Custom or Multimodal Amazon Dataset	Hybrid Fusion	Accuracy: **~92-94%**	

3. Background

3.1. Dense Net Network

DenseNet (densely connected convolutional network) is a deep convolutional neural network model proposed by Gao Huang et al. [29] in 2017. It uses the idea of dense connection to make the network have stronger feature transfer. In addition to its reusability, deep learning can attain high accuracy by training very deep neural networks. In a traditional convolutional neural network, the output of each layer is computed by convolving the previous layer's input using a nonlinear activation function.

Assuming that the input is an image X0X0, after an LL-layer neural network, the ii-th layer's nonlinear transformation is represented as (∗)Hi∗, which can comprise several functional operations, including BN, ReLU, Pooling, or Conv. The ii-th layer's output feature is denoted as XiXi.

In a conventional convolutional feed-forward neural network, the output XiXi of the ii-th layer serves as the input of the i+1i+1 layer and can be represented as $Xi=Hi(Xi-1)$Xi=Hi(Xi−1).

A key benefit of ResNet is that it allows the gradient to flow through the identity function to reach the previous layer. However, the use of addition to combine the identity mapping and nonlinear transformation output in the layer stacking process can potentially disrupt the information flow in the network. And DenseNet proposes the concept of dense block. Each dense block contains several convolutional layers and a skip connection so that the features of all the previous layers can be directly transferred to the subsequent layers, thus forming a dense network. A dense block consists of several convolutional layers and a skip connection, as shown in Figure 1 below.

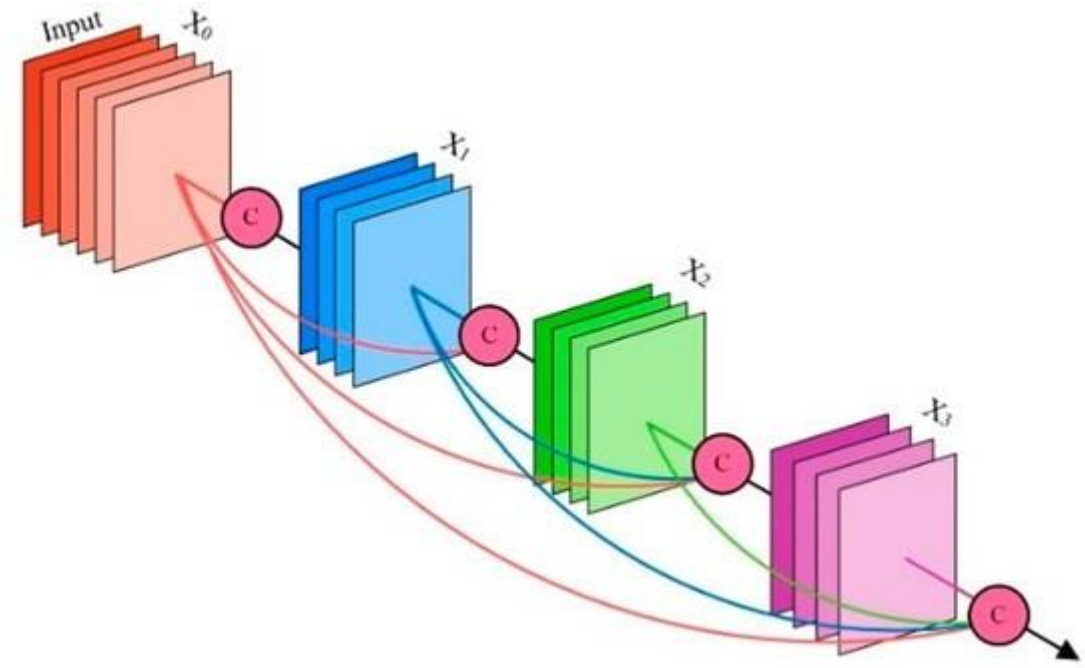


Figure 1. The computational process of the attention mechanism. DenseNet network dense link mechanism (where CC stands for channel-level connection operation).

In DenseNet, each layer's input is the concatenation of all the preceding layer's feature maps generated within the same dense block. The output of layer (l−1)(l−1) is denoted as Xl−1Xl−1. As depicted in Figure 1, the input of the ii-th layer depends on both the output of the i−1i−1 layer and the output of all previous layers. The output of the ll-th layer can be expressed as follows using Equation (2):

$Xl=([X0,X1,\cdots,Xl-1])$.Xl=HlX0,X1,⋯,Xl−1.

Among them, [] operator represents concatenation, that is, all output feature maps of layers X0X0 to Xl−1Xl−1 are combined by Channel. The nonlinear transformation function (·)Hl(·) used here is a combination of BN + ReLU + Conv(3 × 3).

Each convolutional layer's input includes the output from all previous layers, and the output will be

directly passed to all subsequent layers. If the input and output dimensions are different, it needs to be transformed by an additional convolutional layer so that the input and output can be concatenated. The advantage of using dense blocks is that it can make the network have stronger feature transfer and reuse capabilities, thereby improving accuracy. In addition, skip connections can also make the network have stronger feature reuse ability and generalization ability, which further improves the accuracy.

3.2. ALBert Pre-Trained Model

ALBert [30] (a little Bert) is a natural language processing model developed by Alibaba DAMO Academy and based on the Bert model. The model is pre-trained for self-supervised learning and can undertake various natural languages processing tasks, such as text classification, named entity recognition, question answering, and text generation.

The structure of ALBert is similar to that of Bert. It is composed of multiple transformer modules. Each transformer module comprises a multi-head self-attention layer and a feed-forward neural network layer. Different from Bert, ALBert uses cross-layer connections and global pathways to enhance the ability of feature transfer and information flow, thereby improving the efficiency and performance of the model.

The main advantages of ALBert over Bert are as follows:

1. More efficient training: ALBert uses two new training strategies. ALBert uses cross-layer parameter sharing as a training strategy, which significantly reduces the number of model parameters and simplifies the model training process. Sentence order prediction enhances the model's generalization ability and augments the training data.
2. Better performance: ALBert achieved better results than Bert in the GLUE (general language understanding evaluation) benchmark evaluation task, indicating that ALBert has better performance in natural language understanding tasks than Bert.
3. Better generalization ability: ALBert uses more sufficient pre-training, which can better learn general language representation and thus has better generalization ability in downstream tasks.
4. More flexible model structure: ALBert provides a variety of different model structures and hyperparameters, which can be adjusted flexibly according to specific application scenarios.

In short, compared with Bert, ALBert has improved in terms of training efficiency, performance, generalization ability, and model flexibility, so this paper chooses ALBert as the pre-training model.

3.3. CBAM Attention Mechanism

CBAM (convolutional block attention module) is an attention mechanism module for image recognition, proposed by Sanghyun Woo et al. [31] of the KAIST Machine Learning Research Center in 2018. The CBAM module enhances the accuracy and robustness of image recognition in convolutional neural networks by introducing spatial and channel attention mechanisms.

Specifically, the CBAM module includes two attention sub-modules: the channel attention module and the spatial attention module. The channel attention module dynamically adjusts the weights of various channels to enhance the model's focus on essential features. The spatial attention module is used to adaptively change the weights of different spatial locations to enhance the model's attention to important locations. Through the combination of these two sub-modules, the CBAM module enables dynamic weight adjustment of both channel and position within the feature map in an adaptive manner, thereby alleviating the dependence on global features and improving the robustness and generalization ability of the model.

CBAM takes an intermediate feature map $F \in RC \times H \times W$F∈RC×H×W as its input and typically consists of two operational stages. To generate channel attention $MC \in RC \times 1 \times 1$MC∈RC×1×1, the input undergoes global maximum and mean pooling by channel. The resulting one-dimensional vectors are then sent to a fully connected layer and added together, and to obtain the adjusted feature map F'F′, we multiply the channel's attention with the input elements.

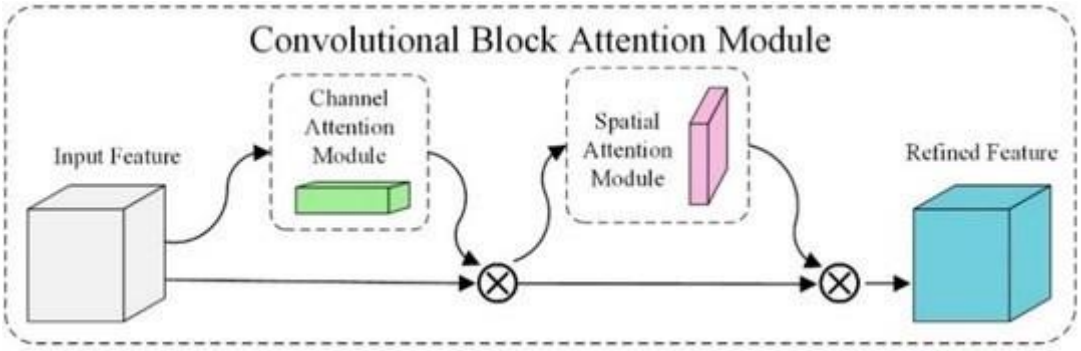


Figure 2. CBAM attention model structure.

Among them, we use ⊗ to denote the element-wise multiplication of corresponding elements. Before performing the multiplication operation, we broadcast the channel attention and spatial attention based on the channel and spatial dimensions, respectively.

4. Multi-Modal Sentiment Analysis Model Based on Cross-Attention Mechanism

Based on the above analysis, this paper proposes the multi-cross-attentive model (MCAM), which is a multi-modal model capable of processing multiple types of data (such as text, images, audio, etc.). Traditional sentiment analysis models can only use one or several data types for sentiment analysis and cannot make full use of the interaction between different data types. MCAM, on the other hand, can process multiple data types simultaneously, and it employs the cross-attention mechanism to learn the interaction between diverse data types, enhancing sentiment analysis's accuracy and robustness.

Figure 3 illustrates the architecture of the model proposed in this paper. First, we utilize ALBert and DenseNet121 to extract vectorized features from text and images, respectively; then, the text features are processed by BiLSTM to obtain the context features containing emotional words, and the CBAM attention is obtained from the two aspects of space and channel, respectively. In the area of emotional characteristics; the single-modal attention mechanism considers the relationship within the modality, and the cross-attention mechanism can consider the relationship between the two modalities, fully considering the complementarity and relevance between different modalities. Utilizing the cross-attention mechanism, we fuse the emotional features extracted from both text and images. The final sentiment analysis result is then obtained using a softmax classifier.

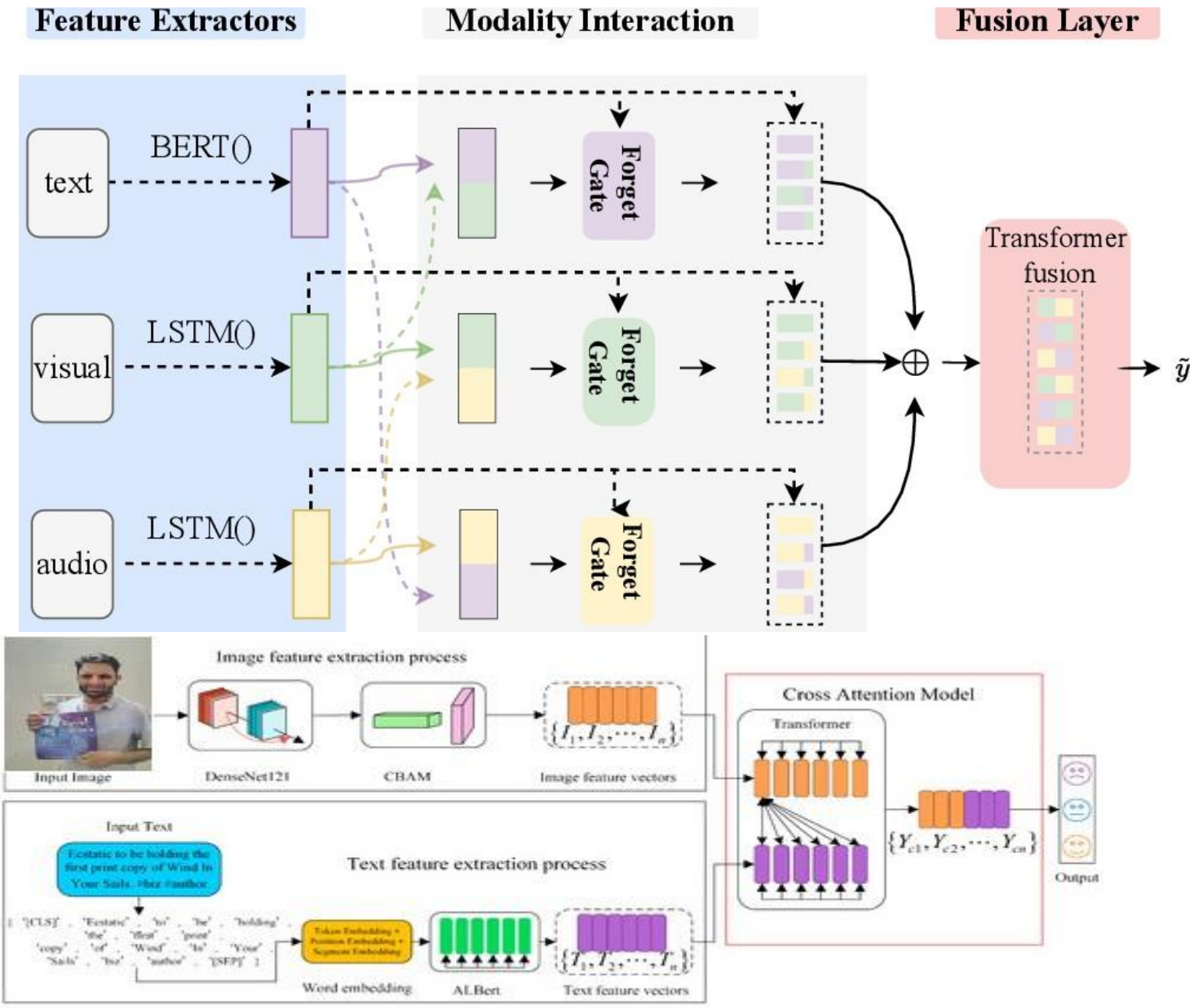


Figure 3. The MCAM model structure proposed in this paper.

4.1. Image Feature Extraction

In an image, usually, certain regions can better reflect the emotional tendency of the whole image. In these regions where the most emotional characteristics can be mined, the results of sentiment analysis will be more accurate. You Quanzeng et al. [32] proposes an attention mechanism to detect emotion-related local areas and creates an emotion classifier based on these regions for image sentiment analysis. This paper utilizes the pre-trained DenseNet121 network model to extract image features. CBAM attention is then employed to extract the most emotional region in the image from both the spatial and channel dimensions, enhancing the expressive capabilities of both overall and local features. Figure 4 illustrates the image features extraction process.

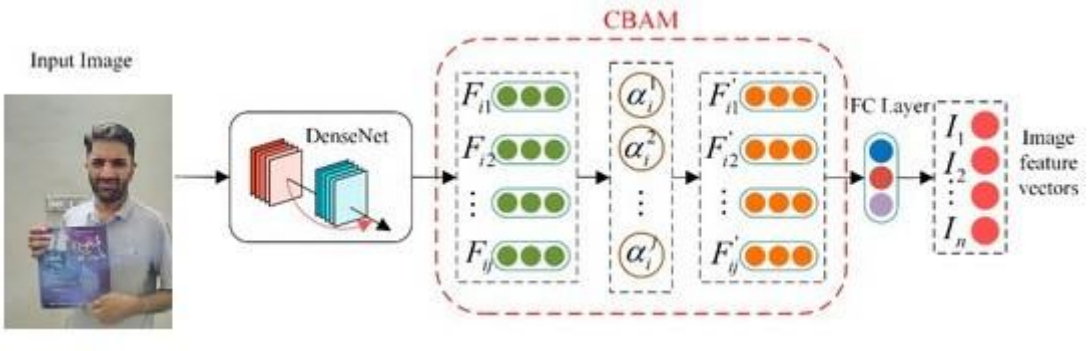


Figure 4. Image feature extraction process.

Let $X=\{X1,X2,\cdots,Xn\}$X={X1,X2,⋯,Xn} denote a dataset with nn images. For each image XiXi, we use the DenseNet network to preprocess the image of the input layer. Next, the image undergoes the CBAM attention mechanism, and its feature vector IiIi, is extracted.

4.2. Text Feature Extraction

It is commonly understood that emotional information in an input sentence is often associated with specific words within the text. Therefore, text sentiment analysis involves vectorizing the text content using the pre-training model ALBert. Subsequently, BiLSTM [33] is used to optimize the emotional level of the text content by combining the input sequence information in both forward and backward directions. Figure 5 illustrates the feature extraction process of the text.

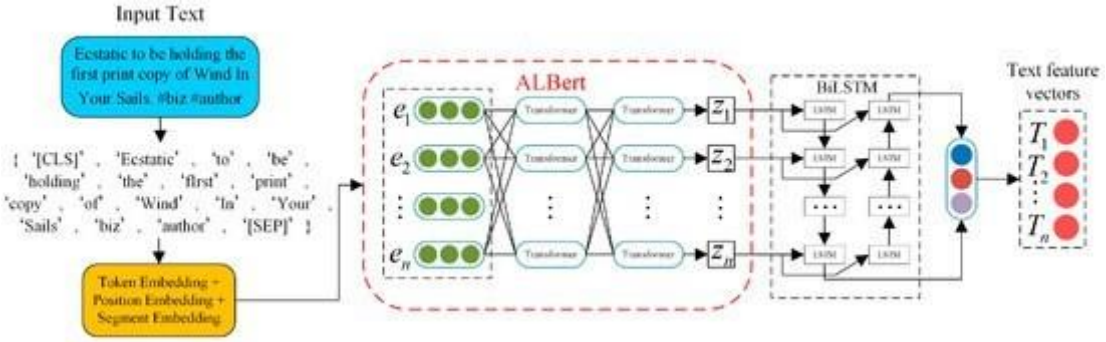


Figure 5. Text feature extraction process.

4.2.1. ALBert Pre-Training Layer

Text sequences are vectorized using the ALBert pre-training model. It performs word segmentation at the character level, enabling vectorization of corpus text. Each input text mm consists of nn characters, which can be expressed as $m=[m1,m2,\cdots,mn]$m=[m1,m2,⋯,mn]; the ii-th character in the text is denoted by mimi.

The text m passes through the input layer to the ALBert pre-training layer. First, for each word in the text information, we mark its number position in the dictionary, obtain the corresponding number, and vectorize the text content to obtain the information sequence ee, eiei, which represents the vectorizations corresponding to the ii-th character, as shown in the following Formula (10):

$e=[e1,e2,\cdots,en]$.e=e1,e2,⋯,en.

(10)

For the information sequence ee it is then input to the transformer encoder in the ALBert pre-training model to mine deep semantic feature information, and after conversion, the feature vector zz of the final text sequence is obtained, and zizi represents the feature vector corresponding to the ii-th character, as in Formula (11):

$z=[z1,z2,\cdots,zn]$.z=z1,z2,⋯,zn.

(11)

4.2.2. BiLSTM Layer

4.3. Cross-Attention Fusion of Image and Text Features

This section employs the cross-attention module to model the intermodal relationship between image regions and text words. The cross-attention mechanism utilizes scaled dot product attention to enhance the recognition of emotional features in images and word fragments by fully considering their complementarity and relevance.

The cross-attention mechanism consists of scaled dot-product attention, using text features Ti as the query matrix in scaled dot-product attention and image features *Ii*Ii as the key-value matrix. By scaling the dot-product attention, the cross-attention features of images and texts are obtained, as shown in Equations (16)–(19).

5. Experimental Process

The performance of the MCAM model is evaluated in this section through comparative experiments conducted on the MVSA and TumEmo datasets. Additionally, this section presents a qualitative evaluation of the model's performance.

5.1. Dataset Introduction

This article uses two public datasets of graphic multi-modal sentiment analysis: (MVSA raw data obtained from **http://mcrlab.net/research/mvsa-sentiment-analysis-on-multi-view-social-data/** accessed on 20 May 2025) MVSA and (TumEmo raw data obtained from **https://github.com/YangXiaocui1215/MVAN** accessed on 20 May 2025) TumEmo. The MVSA dataset comprises tweets that are in the form of text and images. It is a collection of messages obtained from Twitter. TumEmo is image text sentiment data scraped by Tumblr. Tumblr, whose Chinese name is Tang Bole, is the largest light blogging website in the world. The multimedia content posted by users usually includes images, texts, and other forms of content. These datasets are publicly accessible for image and text multi-modal sentiment analysis.

The MVSA dataset contains 4869 text–image pairs. Each sample contains a set of text and images, and the emotional label is uni-modal. The emotional label of the dataset has only three modes: positive, neutral, and negative. The screening results are presented in **Table 2**. Different from the MVSA dataset, the TumEmo dataset further subdivides the emotional labels, including angry, bored, calm, happy, love, and sad, into seven types of emotion. The TumEmo dataset contains 195,265 samples, from which this paper obtained 13,852 pieces of data according to the ratio of each emotional label

Twitter, Sentiment analysis (SA), Opinion mining, Machine learning, Naive Bayes (NB), Maximum Entropy, Support Vector Machine (SVM), multi-modal sentiment analysis, ALBert,feature extraction

Dataset	Positive	Neutral	Negative	Total
MVSA-Single	2,683	470	1,358	4,511
MVSA-Multiple	11,318	4,408	1,299	17,025

Data Type/Label	Text	Image	Post
Negative	1936	2144	1936
Neutral	2921	1938	2921
Positive	2653	3428	2653

Table 2. MVSA dataset information.

The image–text pairs are split into training, test, and validation sets at a ratio of 6:2:2 for each dataset. The

experimental environment of the model in this paper is Intel (R) Xeon (R) 2.50 GHz CPU, 40 GB memory, RTX 3080 GPU, Windows 10 OS, and the deep learning-based TensorFlow 2.9.0 architecture is implemented using the Python programming language version 3.8.

5.2. Model Parameter Settings

The input image's shape is (224, 224, 3), where the three dimensions represent its height, width, and number of channels, respectively. The batch input size is 32. In the CNN layer, we use the pre-trained DenseNet-121 network that has achieved good results in the ImageNet2017 dataset [34] classification challenge. The AveragePooling2D layer utilizes a pooling size of (7, 7), while the Dense layer has output dimensions of 1024/8 and 1024. This results in an output vector with a reduced dimension of 1/8 the original, or 128, and an unchanged dimension of 1024. The aim is to decrease computation by reducing dimensionality while preserving the original feature data.

The text's word embedding layer is initialized using the ALBert pre-training model. The hidden layer dimension is set to 128 to obtain a 128-dimensional vector representation for each word. The input length of the maximum text is derived from our analysis of the dataset, as shown in **Figure 6**. We found that most of the text lengths are below 150, and the number of texts with text lengths below 150 is also the largest. Therefore, 150 is selected as the maximum length of the input text. For text whose input length is greater than 150, the text will be truncated, and for text whose input length is less than 150, zero-value filling will be performed.

The cross-attention mechanism is utilized in the multi-modal fusion part. During the training process, the Adam optimizer, with a learning rate of 0.001, is used to optimize the model parameters; to prevent over-fitting, a random discard rate (Dropout Value) of 0.1 is set in the model, and early stopping technology and L2 regularization are used change. Tenfold cross-validation and grid search are employed to evaluate various parameter combinations. The hyperparameter combinations of the final model .

5.3. Evaluation Metrics

This paper evaluates the model's performance using standard classification metrics, such as accuracy, precision, recall, and F1-score. True positive (TP), false positive (FP), false negative (FN), and true negative (TN) are the four classification results.

5.3.1. Accuracy

Accuracy is a metric that evaluates the model's ability to make correct predictions for the entire dataset. Formula (21) calculates the ratio of correct predictions to the total number of positive and negative examples, known as accuracy.

5.3.2. Precision

Precision is calculated based on the prediction results and represents the proportion of correct predictions in the samples predicted as positive examples. It is computed using Formula
Precision=TPTP+TN.

5.3.3. Recall

Recall is the ratio of correctly predicted positive samples to the total actual positive samples. It is based on the actual samples and is calculated using Formula :
Recall=TPTP+FN.

5.4. Baseline Method

To evaluate the model's generalization ability and robustness, we conducted comparative experiments with the following methods for comparison:

1. Single-modal text model: Hoang et al. [35] uses contextual word representations from the BERT pre-trained language model, fine-tuned with additional generated text, to perform sentiment analysis. Yin Xing et al. [36] utilizes a BiGRU Information Augmented Approach for sentiment analysis.
2. Single-modal image model: Liang Song et al. [37] proposes a method that utilizes the ResNet50 network for image recognition and classification. Paymode A S et al. [38] presents an approach that applies the VGG19 model to classify crop leaf diseases.
3. Multi-modal image–text fusion model: Huang F et al. [39] introduces the deep multi-modal attention fusion (DMAF) model, which performs joint sentiment classification by leveraging the correlation between textual and visual features. Zhu T et al. [40] introduces the image–text interaction network (ITIN) model. By exploring the connection between emotional image regions and text, multi-modal sentiment analysis is performed. Yang X et al. [41] presents a new multi-modal emotional analysis model that employs a continuously updated memory network to extract deep semantic features from image text. Wei K et al. [42] introduces the attention-based modality gating network (AMGN), which detects correlations between different modalities and extracts discriminative features for multi-modal sentiment analysis.

5.5. Experimental Results

A comparative experiment is conducted using the following setup to evaluate the effectiveness of the proposed MCAM model. We compared the single-modal text model, single-modal image model, and multi-modal image–text fusion model,

5.5.1. MVSA Dataset Experimental Results

The MVSA dataset's best performance is exhibited by the proposed MCAM model, as per the analysis of Sentiment classification performance is unsatisfactory for single-modal image and text data, with average classification results. For sentiment analysis that integrates multiple modalities, by learning the correlation between two modalities, the classification effect is greatly improved immediately. In particular, the proposed MCAM model incorporates a cross-attention mechanism to learn distinct modalities, which boosts the prediction accuracy rate by 11.7% and the F1 score by 9.8% compared to the MVAN model. Moreover, the MCAM model outperforms the DMAF model with an accuracy rate increase of 2.4% and an F1 score increase of 1.7%, achieving the best prediction performance.

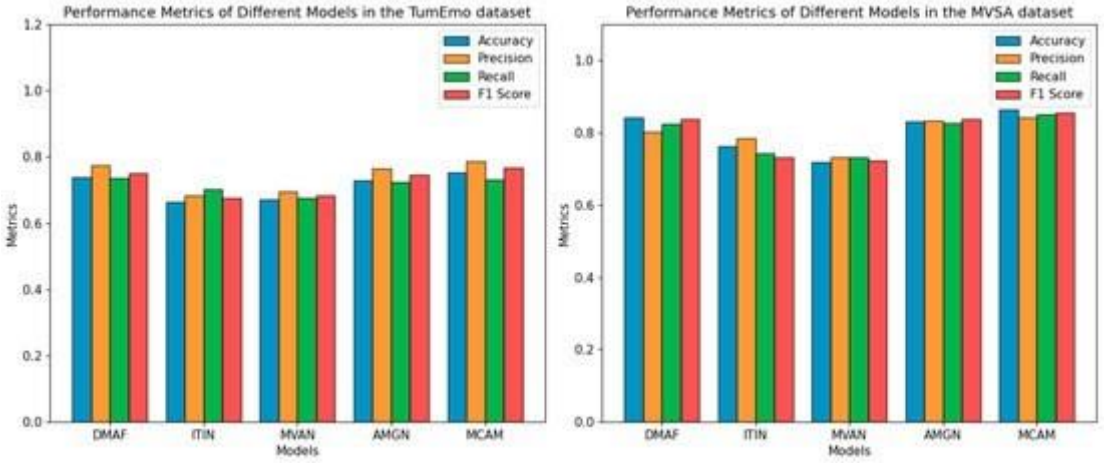


Figure 6. Comparison of performance indicators of different models in MVSA and TumEmo datasets.

In addition, our proposed single-modal image and single-modal text models also outperform the baseline methods. In our text model, we incorporate the highly effective pre-training model ALBert and additionally employ BiLSTM. In the image model, we use the DenseNet121 network. The dense connection structure of DenseNet121 makes feature transfer better and gradient flow Smoother, with less risk of over-fitting. With fewer parameters, it can achieve comparable performance to ResNet50 and VGG19, and by introducing the CBAM attention mechanism, attention is extracted from two aspects of channel and space, which further improves the classification effect.

To further verify the effectiveness of the MCAM model we proposed, we randomly sample varying proportions of data ranging from 20% to 100% from both the MVSA and TumEmo datasets, and then we observe the accuracy changes in both the MCAM model and the four baseline models. **Figure 8** demonstrates that our model consistently outperforms the baseline model of accuracy, regardless of the proportion of sampled training data. This demonstrates our model's absolute competitive advantage and ability to achieve favorable results, even with limited training data.

5.5.2. TumEmo Dataset Experimental Results

Based on the analysis of the performance of the TumEmo dataset exhibits a considerable drop in comparison to that of the MVSA dataset. The reason for our analysis is that the TumEmo dataset contains an extensive range of emotional categories, consisting of seven types, which reduces the model's performance. In multi-category classification problems, as the number of categories increases, the classifier needs to distinguish more categories, which increases the difficulty of classification, and the corresponding classification effect may decline.

However, the model we proposed still has obvious competitive advantages and has achieved the best results in the evaluation indicators. Although the improvement effect is not obvious compared with AMGN and DMAF, it still has a slight improvement effect. In comparison to AMGN, our proposed model achieves a 1.5% increase in accuracy rate and a 1.9% increase in F1 score. Additionally, when compared to DMAF, our model exhibits a 1.4% increase in accuracy rate and a 1.6% increase in F1 score.

5.5.3. Analysis of the Results of PR Curve in Different Models

As it can be seen from the subplot on the left that the PR curves of the TumEmo dataset present a variety of different shapes, which indicates that different models have different performances when classifying the TumEmo dataset. Notably, the PR curve of the MCAM model exhibits the most optimal shape, signifying superior performance in classifying the TumEmo dataset, while the PR curves of other models, such as ITIN and MVAN, show poor shapes, indicating their relatively poor performance.

As can be seen from the right subplot, the PR curve of the MVSA dataset presents a different shape from

that of the TumEmo dataset. The PR curve of model MCAM still presents an optimal shape, while the PR curves of other models, such as ITIN and MVAN, present a poor shape, indicating their relatively poor performance. It is noteworthy that the DMAF model's performance on the MVSA dataset is relatively inferior in contrast to its performance on the TumEmo dataset.

5.6. Ablation Experiment

The following ablation experiments are conducted to further verify the performance of the MCAM model.

Through the performance analysis of the ablation experiments, it is observed that the model's performance is average on different datasets when only images or text are utilized, mainly because the single-modal sentiment analysis is only learned within the modality. Sometimes, images may play a leading role in the classification of emotional categories, and sometimes, the text may play a leading role. Thus, single-modal sentiment analysis fails to fully consider the complementarity and correlation between diverse modalities, resulting in suboptimal classification outcomes.

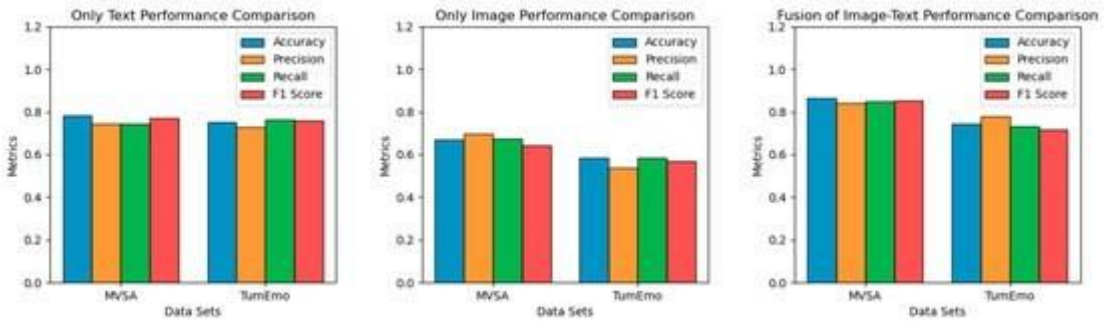


Figure 7. Ablation experiment results of different datasets.

When we introduce the cross-attention mechanism, the model first uses an independent network to learn the feature representation of each modality; then, it employs a cross-attention mechanism to capture inter-modality relationships. Then, similarities between different modalities are computed to obtain weights for a cross-attention mechanism, which fuses feature representations from different modalities in a weighted manner. Finally, the model employs the fused feature representations to predict sentiment, resulting in improved performance and robustness.

5.7. Attention Weight Visual Analysis

This section focuses on a qualitative analysis of image and text fusion for sentiment analysis, including attention scores before and after implementing cross-attention. The emotional score of attention is any value between 0 and 1, and the specific results are shown in **Figure 11**. Through visual analysis, we can clearly see the change in the thermal effect of the image area before and after the introduction of attention.

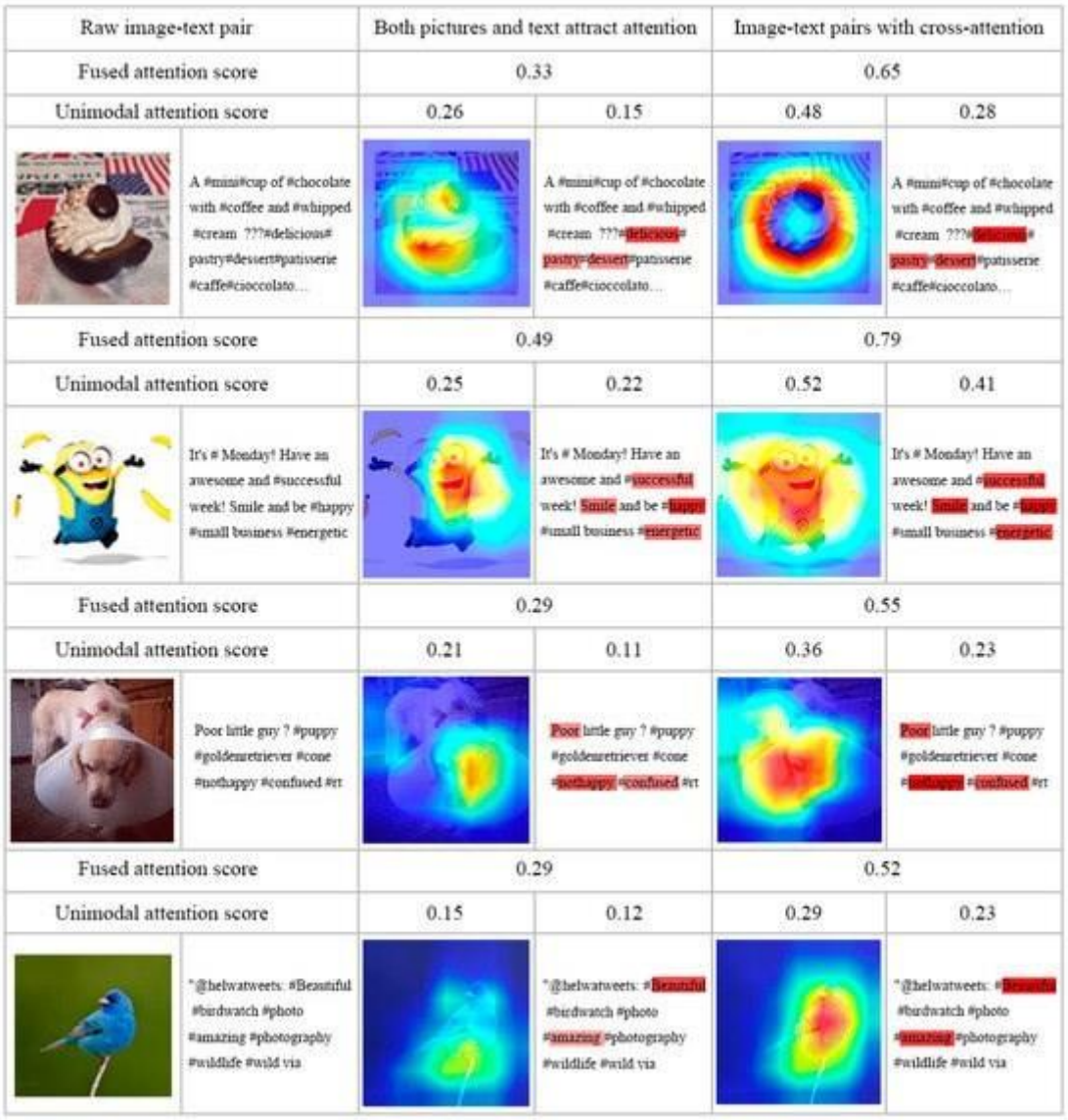


Figure 8. Four examples of visual analysis of attention scores.

5.7.1. Visual Attention Processing

For image data, the convolutional neural network is first enhanced using CBAM to improve the ability of image feature extraction. Input image tensors (224, 224, 3) for average pooling and maximum pooling operations calculate the average and maximum values of the channel dimensions, compress them into vectors with dimensions 128 and 1024 through two fully connected layers, and then restore them to the original dimension.

The results are then combined, and weights ranging from 0 to 1 are obtained using the sigmoid function. The attention mechanism is employed to the weighted tensor's spatial dimension; it calculates the average value and maximum value of the weighted tensor on the spatial dimension and inputs it into a 1 × 1 convolutional layer after splicing to obtain 1 × 1 × 1 feature map; then, we obtain the weight from 0 to 1 through the sigmoid function. Finally, the channel and spatial attention mechanisms' weights are multiplied to derive the ultimate attention score.

We draw a heat map on the image processed by the CBAM attention mechanism to make its color more vivid. The original image is assigned a transparency of 0.5, while the processed image is assigned a transparency of 0.8 to emphasize the attention focus. If the region's attention score is higher, the color of the region is redder.

5.7.2. Text Attention Processing

For text data, we first tokenize the input text data to obtain a token sequence, convert the token sequence into an input tensor that can be processed by the ALBert model, and use the attention mask tensor to represent the position of each token.

Next, we input the input tensor and attention mask into the ALBert model, and the model will encode the input token to obtain an encoded tensor. For each self-attention layer, the model splits the encoded tensor into multiple heads, each of which computes the query, key, and value separately and uses this information to compute the attention score between each token and other tokens. Finally, by weighting the attention scores obtained by each attention head, a weight vector is obtained, which represents the importance of each token in the entire text data, that is, the attention weight.

After processing the text, we color-coded the emotional words that the attention focused on. To differentiate the contribution levels of distinct words to the attention score, we applied varying degrees of shading to the same color. The darker the color, the greater the attention weight of the word.

5.7.3. Cross-Attention Fusion Multi-modal Processing

For multi-modal data that fuse images and text, first, we pass the image features and text features to two fully connected layers, take their outputs as input, pass them to the dot product layer, and calculate the attention weights.

Next, the attention weights are multiplied by image features and text features, respectively, to obtain attention tensors for images and text. Finally, we use the concatenate layer to concatenate the attention tensor of the image and text along the last dimension to obtain the fused feature tensor and calculate the attention score of the fused feature tensor, which is the fusion of multi-modal cross-attention power score.

For the image and text after cross-attention fusion, we can clearly observe that the area of the image that is concerned is further expanded, the red color is further deepened, and the color of the word that is focused on is also deepened. The attention score has also been improved to varying degrees, and the attention score after fusion has been further improved compared with the score of simple vector splicing.

5.8. Prediction Error Case Analysis

In **Figure 12**, the sentiment orientation of the four predicted examples is inconsistent with the labeled sentiment orientation. In the first image, from the smiling expression of the little boy, we can easily determine that the emotion is positive, but the text content contains some negative words, such as “dead people”, “inquietante”, and “terrified”, which ultimately lead to the prediction mistake. In the second image, the negative emotion can also be felt from the tone of the image, but there is the word “happy”, with obvious positive emotion in the text content; however, “contempt” and “empty” are also in the text, conveying “contempt” and “emptiness”, but after ALBert’s processing, the text is divided into independent words, and the meaning may change, i.e., it no longer has a particularly obvious emotional tendency.

Annotation: Positive	Annotation: Negative
Prediction: Negative	Prediction: Positive
 #Dubsmash Video: I see dead people… #fratello #bro #inquietante #terrified #horror #trille…	 #HappyValentinesDay to all you followers (whether your #heart is contempt or empty)
Annotation: Positive	**Annotation: Neutral**
Prediction: Negative	**Prediction: Negative**
 #animal #beatiful #beast #lion #black #dark #grunge #lion #king #wild #photo #a…	 RT @AlArabiya_Eng: #Anger management: How to stop rage ruining your life #AngerProblems

Figure 9. Examples of four wrong predictions in the MVSA dataset.

In the third image, the reason for the prediction error may be that there are wrong words in the text; “beautiful” originally means beautiful, but the word in the text is “beatiful”, causing ambiguity. In the fourth image, we can also feel negative emotions from the roar of the man in the image, and we can also read negative emotions from in the words “Anger” and “ruining” in the text. The ultimate prediction outcome is negative as well, but the original annotation is neutral; we think this is likely the reason why the dataset itself is wrongly annotated.

6. Conclusion

This study successfully demonstrates the effectiveness of **multimodal sentiment analysis** in understanding and classifying sentiments expressed in **product reviews** by integrating **text, image, and audio modalities**. By leveraging machine learning and deep learning techniques, the proposed framework extracts meaningful features from each modality and combines them to deliver more accurate and context-aware sentiment predictions compared to uni modal approaches. The results highlight that each modality contributes unique information: **Text** captures explicit opinions and descriptive feedback, **Images** reflect users' visual satisfaction or dissatisfaction ,**Audio** conveys emotional tones and vocal expressions. The fusion of these diverse data types significantly enhances the sentiment classification performance, offering a more comprehensive understanding of customer feedback. This multimodal approach holds great potential for applications in **e-commerce**, **customer service analytics**, and **intelligent recommendation systems**, where accurate sentiment understanding is critical for decision-making.

MCAM is introduced in our paper as a cross-attention-mechanism-based multi-modal sentiment analysis approach that enhances feature fusion by adaptively computing and calculating the correlation between images and texts. First of all, we propose two different processing methods for a single modality to learn text and image features, respectively; then, we use the cross-attention mechanism to adaptively fuse the features of different modalities, thereby improving the performance of sentiment analysis. Finally, the sentiment result is output through the classifier. Experimental results demonstrate the superiority of our proposed approach over four baselines, achieving superior performance on two publicly available multi-modal sentiment analysis datasets.

As for the limitations of the current work, we note that the cross-attention mechanism for conflicting sentiment indicators between modalities is a challenge in the current work. As for situations where sentiment indicators conflict with each other, this is the direction that we need to consider in our next work. We can use domain knowledge or prior information, such as sentiment dictionaries or sentiment rules, to help us resolve sentiment conflicts between modalities.

Our future work will entail improving the model to enhance its capacity for learning deeper multi-modal correlations, thereby enabling it to achieve superior results in image–text sentiment analysis. In addition, we need to improve the text processing model. For example, for misspelled words, we can use spell checkers and error correctors to detect and correct spelling mistakes; thus, we delete or correct wrongly annotated labels in the dataset. In the future, we also want to study further the sentiment analysis of more modalities, such as video and audio.

References

1. Poria, S., Cambria, E., Bajpai, R., & Hussain, A. (2017). A review of affective computing: From unimodal analysis to multimodal fusion. Information Fusion, 37, 98–125.
2. Xu, H., Mao, Y., Chen, G., & Xu, Y. (2020). Multimodal sentiment analysis using deep canonical correlation analysis. Neurocomputing, 398, 221–230.
   You, Q., Luo, J., Jin, H., & Yang, J. (2016). Building a large-scale multimodal dataset for visual sentiment analysis. Proceedings of the 2016 ACM International Conference on Multimedia Retrieval (ICMR), 1–8.
3. Zadeh, A., Chen, M., Poria, S., Cambria, E., & Morency, L.-P. (2017). Tensor Fusion Network for Multimodal Sentiment Analysis. EMNLP 2017.
4. Li, J., Zhao, H., & Li, Y. (2023). Multimodal Sentiment Analysis Method Based on Multi-task Learning. Multimedia Tools and Applications.
   Zhao, Y., Mamat, M., Aysa, A., & Ubul, K. (2023). Multimodal sentiment system and method based on CRNN-SVM. Multimedia Tools and Applications.
5. Singh, S., & Kumari, R. (2021). Multimodal sentiment analysis using ensemble learning: A product review perspective. Journal of Intelligent & Fuzzy Systems, 40(3), 5467–5479. https://doi.org/10.3233/JIFS-202525
6. Pérez-Rosas, V.; Mihalcea, R.; Morency, L. Utterance-Level multimodal sentiment analysis. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, ACL 2013, Sofia, Bulgaria, 4–9 August 2013; in Long Papers. The Association for Computer Linguistics: Kerrville, TX, USA, 2013; Volume 1, pp. 973–982. [**Google Scholar**]
7. You, Q.; Luo, J.; Jin, H.; Yang, J. Cross-modality consistent regression for joint visual-textual sentiment analysis of social multimedia. In Proceedings of the Ninth ACM International Conference on Web Search and Data Mining, San Francisco, CA, USA, 22–25 February 2016; Bennett, P.N., Josifovski, V., Neville, J., Radlinski, F., Eds.; ACM: New York, NY, USA, 2016; pp. 13–22. [**Google Scholar**]

8. Cao, D.; Ji, R.; Lin, D.; Li, S. A cross-media public sentiment analysis system for microblog. *Multimed. Syst.* **2016**, *22*, 479–486. [Google Scholar] [CrossRef]
9. Poria, S.; Cambria, E.; Gelbukh, A.F. Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP'15), Lisbon, Portugal, 17–21 September 2015; pp. 2539–2544. [Google Scholar]
10. Zargari, H.; Zahedi, M.; Rahimi, M. GINS: A Global intensifier-based N-Gram sentiment dictionary. *J. Intell. Fuzzy Syst. Appl. Eng. Technol.* **2021**, *40*, 11763–11776. [Google Scholar] [CrossRef]
11. Xu, G.; Yu, Z.; Yao, H.; Li, F.; Meng, Y.; Wu, X. Chinese text sentiment analysis based on extended sentiment dictionary. *IEEE Access* **2019**, *7*, 43749–43762. [Google Scholar] [CrossRef]
12. Goel, A.; Gautam, J.; Kumar, S. Real time sentiment analysis of tweets using naive bayes. In Proceedings of the 2016 2nd International Conference on Next Generation Computing Technologies (NGCT), Dehradun, India, 14–16 October 2016; IEEE: Piscataway, NJ, USA, 2016; pp. 257–261. [Google Scholar]
13. Rathor, A.S.; Agarwal, A.; Dimri, P. Comparative study of machine learning approaches for Amazon reviews. *Procedia Comput. Sci.* **2018**, *132*, 1552–1561. [Google Scholar] [CrossRef]
14. Zhou, X.; Wan, X.; Xiao, J. Attention-based lstm network for cross-lingual sentiment classification. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Austin, TX, USA, 1–5 November 2016; pp. 247–256. [Google Scholar]
15. Sun, C.; Huang, L.; Qiu, X. Utilizing BERT for aspect-based sentiment analysis via constructing auxiliary sentence. In Proceedings of the NAACL-HLT, Minneapolis, MN, USA, 2–7 June 2019; pp. 380–385. [Google Scholar]
16. Miao, Y.; Ji, Y.; Peng, E. Application of CNN-BiGRU Model in Chinese short text sentiment analysis. In Proceedings of the 2019 2nd International Conference on Algorithms, Computing and Artificial Intelligence, Sanya, China, 20–22 December 2019. [Google Scholar]
17. Yenduri, G.; Rajakumar, B.R.; Praghash, K.; Binu, D. Heuristic-Assisted BERT for Twitter Sentiment Analysis. *Int. J. Comput. Intell. Appl.* **2021**, *20*, 20625–20631. [Google Scholar] [CrossRef]
18. Cauteruccio, F.; Kou, Y. Investigating the emotional experiences in eSports spectatorship: The case of League of Legends. *Inf. Process. Manag.* **2023**, *60*, 103516. [Google Scholar] [CrossRef]
19. Yuan, J.; Mcdonough, S.; You, Q.; Luo, J. Sentribute: Image sentiment analysis from a mid-level perspective. In Proceedings of the Second International Workshop on Issues of Sentiment Discovery and Opinion Mining, Chicago, IL, USA, 11 August 2013; ser. WISDOM '13. ACM: New York, NY, USA, 2013; pp. 1–8. [Google Scholar]
20. Borth, D.; Ji, R.; Chen, T.; Breuel, T.; Chang, S.-F. Large-scale visual sentiment ontology and detectors using adjective noun pairs. In Proceedings of the 21st ACM International Conference on Multimedia,

Barcelona, Spain, 21–25 October 2013; ser. MM'13. ACM: New York, NY, USA, 2013; pp. 223–232. [**Google Scholar**]

21. He, X.; Zhang, H.; Li, N.; Feng, L.; Zheng, F. A multi-attentive pyramidal model for visual sentiment analysis. In Proceedings of the 2019 International Joint Conference on Neural Networks (IJCNN), Budapest, Hungary, 14–19 July 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 1–8. [**Google Scholar**]
22. Wang, M.; Cao, D.; Li, L.; Li, S.; Ji, R. Microblog sentiment analysis based on cross-media bag-of-words model. In Proceedings of the International Conference on Internet Multimedia Computing and Service, Xiamen, China, 10–12 July 2014; pp. 76–80. [**Google Scholar**]
23. Zhang, Y.; Shang, L.; Jia, X. Sentiment analysis on microblogging by integrating text and image features. In Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining, Ho Chi Minh City, Vietnam, 19–22 May 2015; pp. 52–63. [**Google Scholar**]
24. Yu, Y.; Lin, H.; Meng, J.; Zhao, Z. Visual and textual sentiment analysis of a microblog using deep convolutional neural networks. *Algorithms* **2016**, *9*, 41. [**Google Scholar**] [**CrossRef**]
25. Kumar, A.; Srinivasan, K.; Cheng, W.H.; Zomaya, A.Y. Hybrid context enriched deep learning model for fine-grained sentiment analysis in textual and visual semiotic modality social data. *Inf. Process. Manag.* **2020**, *57*, 102141. [**Google Scholar**] [**CrossRef**]
26. Xu, J.; Huang, F.; Zhang, X.; Wang, S.; Li, C.; Li, Z.; He, Y. Visual-textual sentiment classification with bi-directional multi-level attention networks. *Knowl-Based Syst.* **2019**, *178*, 61–73. [**Google Scholar**] [**CrossRef**]
27. Paymode, A.S.; Malode, V.B. Transfer learning for multi-crop leaf disease image classification using convolutional neural network VGG. *Artif. Intell. Agric.* **2022**, *6*, 23–33. [**Google Scholar**] [**CrossRef**]
28. Huang, F.; Zhang, X.; Zhao, Z.; Xu, J.; Li, Z. Image–text sentiment analysis via deep multimodal attentive fusion. *Knowl.-Based Syst.* **2019**, *167*, 26–37. [**Google Scholar**] [**CrossRef**]
29. Zhu, T.; Li, L.; Yang, J.; Zhao, S.; Liu, H.; Qian, J. Multimodal sentiment analysis with image-text interaction network. *IEEE Trans. Multimed.* **2022**, *25*, 3375–3385. [**Google Scholar**] [**CrossRef**]