electronics

Article

MFSC: A Multimodal Aspect-Level Sentiment Classification Framework with Multi-Image Gate and Fusion Networks

Lingling Zi [†], Xiangkai Pan [*,†] and Xin Cong

College of Computer and Information Science, Chongqing Normal University, Chongqing 401331, China; zll@cqnu.edu.cn (L.Z.); cx@cqnu.edu.cn (X.C.)

[*] Correspondence: 2022110516034@stu.cqnu.edu.cn

[†] These authors contributed equally to this work.

Abstract: Currently, there is a great deal of interest in multimodal aspect-level sentiment classification using both textual and visual information, which changes the traditional use of only single-modal to identify sentiment polarity. Considering that existing methods could be strengthened in terms of classification accuracy, we conducted a study on aspect-level multimodal sentiment classification with the aim of exploring the interaction between textual and visual features. Specifically, we construct a multimodal aspect-level sentiment classification framework with multi-image gate and fusion networks called MFSC. MFSC consists of four parts, i.e., text feature extraction, visual feature extraction, text feature enhancement, and multi-feature fusion. Firstly, a bidirectional long short-term memory network is adopted to extract the initial text feature. Based on this, a text feature enhancement strategy is designed, which uses text memory network and adaptive weights to extract the final text features. Meanwhile, a multi-image gate method is proposed for fusing features from multiple images and filtering out irrelevant noise. Finally, a text-visual feature fusion method based on an attention mechanism is proposed to better improve the classification performance by capturing the association between text and images. Experimental results show that MFSC has advantages in classification accuracy and macro-F1.

Keywords: sentiment analysis; multimodal sentiment classification; fusion network; multi-image gate

check for updates

Citation: Zi, L.; Pan, X.; Cong, X. MFSC: A Multimodal Aspect-Level Sentiment Classification Framework with Multi-Image Gate and Fusion Networks. *Electronics* **2024**, *13*, 2349. https://doi.org/10.3390/electronics13122349

Academic Editor: Ioannis Hatzilygeroudis

Received: 7 April 2024
Revised: 31 May 2024
Accepted: 12 June 2024
Published: 15 June 2024

1. Introduction

Multimodal aspect sentiment classification (MASC) is a subtask of multimodal aspect sentiment analysis that aims to classify the sentiment of a particular aspect, where an aspect refers to a word or phrase that describes any attribute related to an entity [1]. Nowadays, e-commerce platforms have many consumer reviews, which are very important for customers who are about to purchase goods. Categorizing consumer reviews through MASC enables customers to swiftly gain insights into various product aspects without the labor of manually evaluating each aspect on online retail sites. As shown in Figure 1, for text-image pairs that contain aspects of price–performance ratio, our inputs are text information, image information related to the text, and aspect information for a particular aspect. Our output is the result of classifying the emotional polarity of a particular aspect. We categorized the outputs into eight categories based on the level of satisfaction of the consumers and used eight different scores to indicate the different categories. A higher score means that consumers are more satisfied with an aspect of the product and a lower score means that consumers are less satisfied.

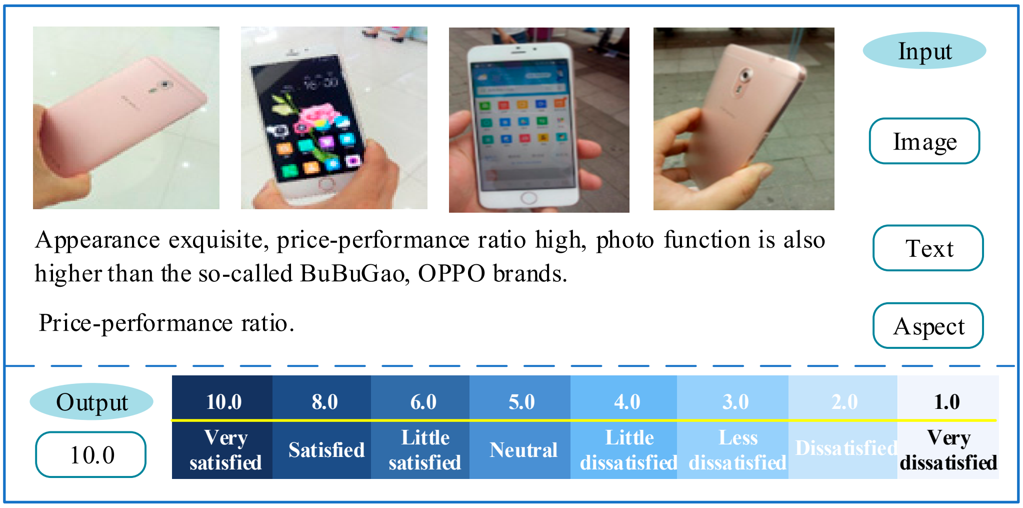


Figure 1. An example of MASC task.

In past research, most scholars have used traditional machine learning methods to solve sentiment classification problems. This type of method is based on manually labeling a portion of the data as a training set, then constructing a classification model by extracting features and learning from the data in the training set, and finally using the model to predict the unlabeled data to achieve classification automatically [2]. With the formal introduction of deep learning by Hinton et al. [3], deep learning has achieved impressive success in the field of natural language processing. Specifically, deep learning architectures, such as convolutional neural network, recurrent neural network, and their variants like long short-term memory networks, have demonstrated remarkable performance across a range of natural language processing tasks. Similarly, as a subtask of natural language processing tasks, it is also widely used in aspect sentiment classification tasks [4].

For aspect sentiment classification, image information is usually complemented and associated with text information in some way. On the one hand, for a short text message containing only one aspect, it is difficult to judge the affective polarity of a particular aspect by its text information due to the brevity and possibly informality of the text information. For example, for the aspect 'appearance and feeling', consumers only commented on the 'appearance and feeling really . . .'. At this point, our affective analyses of the aspect weigh heavily on their associated images. On the other hand, text and images are generally highly correlated with aspect sentiment. For example, when reviewing the 'screen' of a mobile phone, a customer might write a positive text message and add a picture of a high-definition mobile phone screen with no screen anomalies to his review to show that he is satisfied with the screen, or this consumer might express negative words and upload a picture of a mobile phone with a cracked or spotted screen to show his disappointment with the screen. In summary, aspect-level sentiment classification of multimodal data has a variety of complementarities and correlations. And there is a lot of noise from text and images in multimodal data compared to unimodal.

We aim to reduce noise and enhance the integration of text and visual information to improve sentiment classification of aspects in multimodal consumer reviews on e-commerce platforms. Our main contributions are as follows:

- A multimodal aspect-level sentiment classification framework with multi-image gate and fusion networks called MFSC is proposed, which uses a bidirectional long short-term memory network (LSTM) to extract text features and aspect features and uses a memory network to augment text features. Meanwhile, in image feature extraction, a more powerful ResNet-152 is used to deeply mine visual information.
- In the text feature enhancement stage, we propose a text feature extraction method. It utilizes the synergy of the text memory network and an adaptive weighting mechanism to extract text features. The method carefully extracts valuable information at each

hop of the text memory network to ensure that no relevant details are overlooked. In addition, the adaptive weighting strategy dynamically adjusts the importance of the extracted features at each hop. This results in a comprehensive and fine-grained extraction of the text.

- In the visual feature extraction stage, we propose a multi-image gate mechanism designed to integrate multiple images in a sample while filtering out irrelevant noise. This innovative approach first fuses visual features through two attention mechanisms to ensure that the relevant parts of the images are integrated. Subsequently, our method utilizes image gates to filter out irrelevant noise so that the most relevant image features are preserved.
- In the multi-feature fusion stage, we propose a feature fusion mechanism that combines text and images, i.e., using an attention mechanism to combine text features and image features. This fusion mechanism captures the association between text and images, which in turn improves the performance of the MFSC.

The remaining part of this paper is structured as follows: Section 2 gives a brief review on the related work. Section 3 describes the proposed MFSC framework in detail. Section 4 shows experimental simulations. Finally, Section 5 concludes this paper.

2. Related Work

Multimodal aspect sentiment analysis consists of two primary subtasks: one is multimodal aspect term extraction (MATE) and the other is multimodal aspect sentiment classification (MASC). For MATE, the input is text-image pairs, and the output is all the aspect terms contained in the text. For example, an adaptive co-attention network [5] was proposed to extract aspect terms using text and images from tweets. A multi-modal graph fusion method [6] is proposed, which does not directly utilize the entire image, but finds targeted visually guided regions in the image for visual information extraction. Zhai et al. [7] developed a model containing a suggested bi-axial attention module in order to address the core bottleneck of previous structured sentiment analysis methods. Considering that one of the subtasks of structured sentiment analysis is the extraction of level words, the model is perfectly suited to be combined with image features for aspect term extraction.

For MASC, it aims to classify the sentiment polarity of the specific target or aspect words in the current sentence, which is a fine-grained sentiment classification task [8–17]. The traditional method is text-based aspect sentiment classification, which aims to classify the sentiment of a specific aspect term. Initially, machine learning techniques are utilized to accomplish the sentiment classification task [18–21]. With the rapid development of deep learning in the fields of natural language processing (NLP) and computer vision (CV), it has been used to improve classification performance over machine learning techniques, such as LSTM [22] and its variant GRU [23], memory network [24,25], interactive attention network [26], target-sensitive memory network [27], multi-grained attention network [28], and so on. Specifically, LSTM can integrate aspect information and significantly improve the accuracy of sentiment classification. The memory network can explicitly capture the significance of every contextual word to infer the sentiment polarity of an aspect. In addition, it can also include relevant information about the neighboring aspects in the sentiment classification, modeling the relationship between the target aspect and neighboring aspect while filtering out irrelevant information. The interactive attention network uses two LSTM models to learn the initial representations of context and aspect, respectively, and then obtains the final text and aspect representations through the attention mechanism. The multi-grained attention network solves the problem of aspect sentiment not being solely dependent on contextual inferences.

The rise of multimodal content in e-commerce platforms, including text, images, and videos within reviews, has led to the recognition of the limitations of text-based sentiment analysis and the growing academic focus on MASC. Xu et al. [29] first proposed the new task of aspect-based multimodal sentiment classification and presented a new multi-interactive memory network model, which achieved good results on the publicly available dataset

Multi-ZOL. However, it adopts the same memory network model for text and image and interacts text and image information at each hop of the two-memory network, so it does not consider the problem of noise present in the image and therefore the performance of the model needs to be improved. Based on this, Yu et al. [30] proposed an aspect-sensitive attention and fusion network to accomplish this task. The model first generates aspect-sensitive text representations using an attention mechanism, then removes the noise information from images using a gate mechanism, and finally further fuses the text and image features to obtain cross-modal features. However, Yu et al. only used LSTM to extract text features, which is not enough to mine text features and also leads to suboptimal results. To solve the problem of users posting tweets that are too short to easily recognize their emotions, Khan et al. [31] proposed a transformer architecture for object detection, and used the generated model to transform the image into the input space of a pre-trained language model, and then constructed an auxiliary sentence to feed the translated image into the BERT (Bidirectional Encoder Representations from Transformers) [32] language model, so that the resulting features can be used for multimodal aspect-level sentiment classification. Ling et al. [33] manually annotated a dataset of image and evaluation object matching and proposed a new image and evaluation object matching model, which is mainly a network structure from coarse-grained to fine-grained matching. Ju et al. [34] first proposed to jointly perform two sub-tasks: multi-modal aspect terms extraction and multi-modal aspect sentiment classification. Their approach can not only model the cross-modal relation between text and image, determining how much visual information contributes to text, but also separately mine the visual information for two sub-tasks instead of collapsed tagging with the same visual feeding. Zhao et al. [35] drew inspiration from the concept of curriculum learning and proposed a multi-grained Multi-curriculum Denoising Framework (M2DF). It achieves denoising by adjusting the order of training data rather than by filtering image noise through threshold settings. Han et al. [36] proposed a model based on selective attention and natural contrastive learning, which uses a probe-based strategy to implement high attention weights for regions of higher importance. Moreover, the low memory consumption at runtime of this model is a prominent advantage compared to other methods. Lu et al. [37] proposed a language-guided reasoning network called LGR-NET. Instead of simply fusing textual and visual features, LGR-NET alternates the extracted textual features for text-guided cross-modal alignment and fusion. In addition, they designed a novel cross-modal loss to enhance cross-modal alignment between text and images.

In summary, our proposed framework takes into account the noise problem in visual features and extracts text features in more detail, and also exploits the interaction between textual and visual features in the final feature fusion stage.

3. Proposed Methodology

This section presents the proposed MFSC framework to accomplish multimodal aspect-level sentiment classification. First, Section 3.1 describes the task definition, followed by giving an overview of the MFSC framework in Section 3.2. Then, the details of the implementation are elaborated in Section 3.3.

3.1. Task Definition

Given a sample $B(T, I)$ of a multimodal dataset, where T denotes a text set containing N words W and I denotes an image set, $T = \{W_1, W_2, \ldots, W_\mathrm{N}\}$ ($|W| = N$) and $I = \{I_1, I_2, \ldots, I_M\}$ ($|I| = M$). Meanwhile, an aspect set $A = \{A_1, A_2, \ldots, A_\mathrm{P}\}$ ($|A| = P$) is also given. Our task is to predict the sentiment polarity Y with a given aspect phrase Ai. $Y \in \{C_1, C_2, \ldots, C_\mathrm{S}\}$ ($|C| = S$). Here, each C is described by a score, where a larger score indicates that the consumer's sentiment attitude is the more satisfied and a smaller score represents the consumer's more dissatisfied sentiment attitude.

3.2. Overview of MFSC

The framework of MFSC is shown in Figure 2 and it consists of four modules: text feature extraction, visual feature extraction, text feature enhancement, and multi-feature fusion. In the text feature extraction module, the text information and aspect information are converted into vectors through word embedding, and the obtained ones are then separately put into the Bi-LSTM network to obtain their corresponding hidden states. In the visual feature extraction module, the residual network [38] is used to extract image features and pass the obtained raw image information to multi-image gate. And for the multi-image gate, two novel attention mechanisms are used and final visual features are obtained. On this basis, the text feature can be enhanced in the text memory network, and the attention mechanism is adopted to fully learn the text context, thus highlighting the important text information. Finally, the text features V_t, visual feature V_{vis} and $V_{t\text{-}vis}$ are concatenated through softmax function, so as to achieve the task of sentiment prediction. For easy presentation, Table 1 summarizes the notations used in the proposed MFSC framework.

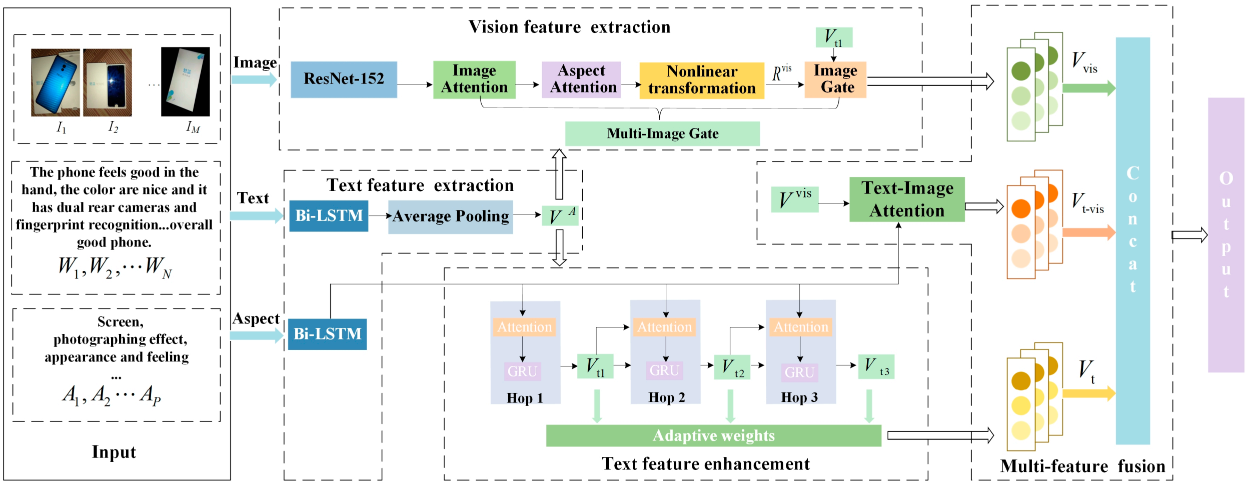


Figure 2. The proposed framework of MFSC.

Table 1. Notations.

Notation	Description
W_i	the vector representation of the *i*-th text word
h_i	the hidden representation of the *i*-th text word
H	the text hidden representation set
a_t	the vector representation of the *t*-th aspect word
S_t	the hidden representation of the *t*-th aspect word
S	the aspect hidden representation set
V^A	the final aspect feature vector
W_t	$t \in \{v, r\}$. W_v, W_r are learnable parameters in the tanh function
u^T	the learnable parameters in Equation (9)
b^i	$i \in \{\text{null}, img, vis, z, tex, t\text{-}vis\}$. b, b^{img}, b^{vis}, b^z, b^{tex}, $b^{t\text{-}vis}$ are learnable parameters in the tanh function
α	all weight of attention mechanism
c^T	the learnable parameters in Equation (12)
σ	the sigmoid function
λ_i	the *i*-th adaptive weight

3.3. Implementation of MFSC

3.3.1. Text Feature Extraction

In order to utilize the contextual information in both the front and back directions of the text sequence so that the framework can better understand the whole text sequence, and considering the limited computational resource, we use a Bi-LSTM for text feature extraction:

$$w_i = embed(W_i) \tag{1}$$

$$h_i = Bi - LSTM(w_i) \tag{2}$$

where *embed* denotes word embedding, *Bi-LSTM* denotes Bi-LSTM, $H = \{h_1, h_2, \ldots, h_\mathrm{N}\}$ denotes the text representation obtained after Bi-LSTM and $i \in [1, N]$. Similarly, the representation of aspect features is shown below:

$$a_t = embed(A_t) \tag{3}$$

$$s_t = Bi - LSTM(a_t) \tag{4}$$

We use the set $S = \{s_1, s_2, \ldots, s_P\}$ to denote the aspect representation obtained after Bi-LSTM, and $t \in [1, P]$. Then, we perform the same process as in [29], i.e., we take the average of all hidden representations as the final aspect feature vector:

$$v^A = \frac{1}{P}\sum_{t=1}^{P} s_t \tag{5}$$

3.3.2. Visual Feature Extraction

In aspect-level sentiment analysis, since many sample sentences are very short and contain information that is often insufficient for a single-text model to make a correct judgment, visual information is very important for analyzing sentiment polarity, which contains rich content that can effectively enhance the robustness of the framework. In order to solve the problem of difficulty in training neural networks with too much depth, the residual network [38] was first proposed and it has shown amazing performance in the image domain. Considering its excellent performance, here we use the pre-trained ResNet-152 for image extraction; its use of residual connectivity preserves the original features, making the learning of the network smoother and more stable, further improving the accuracy and generalization of the framework:

$$r_m = resnet(I_m), m \in [1, M] \tag{6}$$

In Equation (6), r_m is a 2048 × 7 × 7 feature vector, which indicates that for each image, it is divided into 49 visual blocks, and for each visual block its 2048-dimensional visual features are extracted. Here, M is 5, which indicates that for a sample, we only consider the features of its first five images. The detail view of the visual feature extraction is shown in Figure 3.

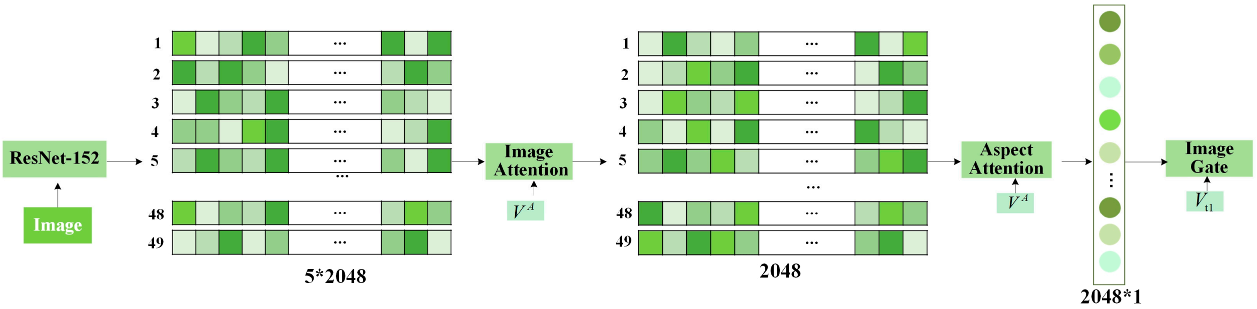


Figure 3. Detail view of the visual feature extraction.

Then, we apply multi-image gate to process the visual feature extracted by Equation (6), i.e., the visual feature will go through two attention modules and one image gate, and the first attention module is represented as follows:

$$q_m = u^{\top} \tanh(W_v v^A + W_r r_m + b) \tag{7}$$

$$\alpha_m = \exp(q_m) / \sum_{b=1}^{5} \exp(q_b) \tag{8}$$

$$r^{img} = \sum_{m=1}^{5} \alpha_m r_m \tag{9}$$

where α_m ($m \in [1, 5]$) is the weight of the first attention mechanism we obtained, and from this we obtain the first layer of visual features r^{img}, which is a 2048 × 7 × 7 feature vector. The second attention module is represented as follows:

$$g_i{}^{img} = c^{\top} \tanh(W_v{}^{img} v^A + W_R{}^{img} r_i{}^{img} + b^{img}) \tag{10}$$

$$\alpha_i{}^{img} = \exp(g_i{}^{img}) / \sum_{w=1}^{49} \exp(g_w{}^{img}) \tag{11}$$

$$r^{vis} = \sum_{i=1}^{49} \alpha_i{}^{img} r_i{}^{img} \tag{12}$$

$$R^{vis} = \tanh(W^{vis} r^{vis} + b^{vis}) \tag{13}$$

In which $\alpha_i{}^{img}$ ($i \in [1, 49]$) is the weight of the second attentional mechanism, and from this we obtain the second layer of visual features r^{vis}, which is a 2048-dimensional feature vector. Finally, we perform a nonlinear transformation in order to align the dimensions of the visual feature with the dimensions of the text features and thus obtain R^{vis}.

For image gate, the output is determined by the textual context and the visual context. The output value of the gate varies from 0 to 1. The size of the output value determines the size of the contribution of a certain part of the image. According to [30], the expression is shown below:

$$Z = \sigma(W_v{}^z V_{t1} + W_R{}^z r^{vis} + b^z) \tag{14}$$

Based on the output of the image gate, we can then obtain the final visual feature V_{vis}, where $\circ$ represents the Hadamard product:

$$V_{vis} = Z \circ R^{vis} \tag{15}$$

3.3.3. Text Feature Enhancement

The text memory network extracts important words for sentiment and aggregates text memory with the representation of the given aspect to account for the influence that the aspect brings to texts. According to [29], we adopt three memory hops to extract text contextual and aspect information. But unlike [29], we do not take the output of the last hop as the final text features; instead, we utilize the outputs of each hop and multiply the outputs of each hop by an adaptive weight to obtain the final text feature. We did not consider the interaction between textual and visual features; we believe this could provide rich visual information for textual features, but it may also introduce a large amount of redundant noise at each hop. The formula of the first hop is shown as follows:

$$l_i{}^1 = \tanh(W_{tex}{}^1 h_i + W_{tex}{}^1 v^A + b^{tex}) \tag{16}$$

$$\alpha_i{}^{tex} = \exp(l_i{}^1) / \sum_{i=1}^{N} \exp(l_i{}^1) \tag{17}$$

$$v^1{}_{tex} = \sum_{i=1}^{N} \alpha_i{}^{tex} h_i \tag{18}$$

$$V_{t^1} = gru(v^1{}_{tex}, v^A) \tag{19}$$

where $i \in [1, N]$, $V_t{}^1$ is the text feature of the first hop obtained after the GRU module. Similarly, the formula for the second and third hop is expressed as follows:

$$l_i{}^j = \tanh(W_{tex}{}^j h_i + W_{tex}{}^j V_{tj-1} + b_j{}^{tex}) \tag{20}$$

$$\alpha_i{}^{texj} = \exp(l_i{}^j) / \sum_{i=1}^{N} \exp(l_i{}^j) \tag{21}$$

$$v^j{}_{tex} = \sum_{i=1}^{N} \alpha_i{}^{texj} h_i \tag{22}$$

$$V_{tj} = gru(v^j{}_{tex}, v^A) \tag{23}$$

where $i \in [1, N]$, j = [2, 3], $V_t{}^2$ and $V_t{}^3$ are the text features obtained after the GRU module for the second and third hop, respectively.

The final text feature formula obtained from adaptive weighting is represented as follows:

$$\lambda_i = \exp(w_i) / \sum_{i=1}^{3} \exp(w_i) \tag{24}$$

$$V_t = \sum_{i=1}^{3} \lambda_i V_{ti} \tag{25}$$

where w_i and i (i = 1, 2, 3) is the initialized weight parameter, and V_t is the final text feature.

3.3.4. Multi-Feature Fusion

The final fusion layer consists of text feature V_t, feature $V_{t\text{-}vis}$, and visual feature V_{vis}, where $V_{t\text{-}vis}$ is obtained through an attention module using both textual feature H and visual feature V_{vis}. Considering that the image feature not only provides rich feature, but also helps to find out the important feature in the text, here we include an attention mechanism in the final fusion layer to help find out the important text information, i.e., $V_{t\text{-}vis}$ is calculated as follows:

$$e_i = \tanh(W_{t-vis} h_i + W_{t-vis} V_{vis} + b^{t-vis}) \tag{26}$$

$$\alpha_i{}^{t-vis} = \exp(e_i) / \sum_{i=1}^{N} \exp(e_i) \tag{27}$$

$$V_{t-vis} = \sum_{i=1}^{N} \alpha_i{}^{t-vis} h_i \tag{28}$$

We concatenate V_t, $V_{vis,}$ and $V_{t\text{-}vis}$ to generate the final multimodal representation, where $\oplus$ denotes concatenation:

$$V = V_t \oplus V_{vis} \oplus V_{t-vis} \tag{29}$$

Finally, we feed the multimodal representation V to the softmax function for classification of sentiment polarity:

$$pred = softmax(W^{\top}{}_{muti} V + b_{muti}) \tag{30}$$

4. Experiments and Discussion

In this section, the experimental results of MFSC are examined. MFSC is implemented on an Nvidia GeForce GTX3060 GPU with RAM 12 GB. And due to the high computational complexity of the framework, which can take up a lot of memory, the minimum memory for the device is 32 GB. Accuracy and Macro-F1 are analyzed compared to existing methods. Among them, accuracy is the most common metric used to measure the performance of classification models. It is simply the ratio of correctly predicted observations to the total number of observations. Macro-F1 is the average F1 score across all classes. It gives equal weight to each class, which is particularly useful in multi-class classification problems where some classes may have a small number of instances.

4.1. Datasets and Experimental Parameters

We conducted experiments on a publicly available multimodal sentiment classification dataset collected from (https://github.com/xunan0812/MIMN/tree/master/datasets/zolDataset, accessed on 15 April 2023), where each comment contains text content, a set of images, and at least one but no more than six aspects. The sentiment classification labels are represented as scores from 1.0 to 10.0. In this case, 1.0 means that consumers are extremely dissatisfied, and as the score increases, so does the level of consumer satisfaction, i.e., a score of 5.0 represents neutral consumer sentiment, and a score of 10.0 means that consumers are extremely satisfied. There are 22,743 training samples, 2843 validation samples, and 2843 test samples in the dataset.

For a fair comparison, we followed the experimental parameter settings in [29]. The number of the hops in the text memory network is 3, which is the optimal hop count in previous experiments. The dimension of the hidden representation of LSTM is 100. The maximal padding length of the text content is 320, and the maximal padding length of the aspect words is 4. For the images, we resized them to 224 × 224 and fed them into a pre-trained neural network ResNet152 [38] to extract 2048 × 7 × 7 dimensional embeddings. The maximum fill number of M is 5. The dropout is 0.2, the learning rate is 0.001, the batch size is 128, and the early stopping is 10; during the training process, it can monitor the performance on the validation set. Once performance begins to decline, training will stop to avoid overfitting. This method helps to find the optimal point for training time to achieve optimal generalization performance. In addition, we chose the cross-entropy loss function and updated the parameters using the Adam optimizer.

4.2. Comparative Methods

We compare our proposed framework with several comparative methods, including the representative text method and multimodal method, shown in Table 2.

Table 2. The descriptions of comparative methods.

Comparative Methods		**Description**
Text-based method	LSTM	A method to learn text features and perform sentiment classification using LSTM [39].
	AEAT-LSTM	A method to learn text features and perform sentiment classification using LSTM incorporating attention mechanisms [39].
	RAM	A method to learn text features and perform sentiment classification using Bi-LSTM [40].
	BERT	A method that enables understanding of contextual representations [32].
Multimodal method	Co-Memory + Aspect	Unlike Co-Memory, it adds the aspect feature on the inputs [41].
	MIMN	A method for utilizing text memory network and image memory network and interacting between them [29].
	TomBERT	A method for multimodal sentiment classification using pre-trained BERT [42].
	ESAFN	A method for introducing a gate mechanism to eliminate picture noise [30].
	EF-CapTrBERT	A method for adding text information using image captioning [31].
	MFSC-BERT	A method where we replace Bi-LSTM in MFSC with BERT for text feature extraction.

Text-based method: These models only use text information, include LSTM [39], which is a model based on a long- and short-term memory network to learn text features and perform sentiment classification. AEAT-LSTM [39] adds aspect information to text information to perform sentiment classification. RAM [40] is a memory-based model which builds memory on the hidden states of a Bi-LSTM and generates aspect representation also based on a Bi-LSTM. BERT [32], which is a pre-trained natural language processing model that uses the Transformer architecture for bidirectional training, is capable of understanding and generating contextual representations of human language.

Multimodal method: These models use information from both text and image modalities. Co-Memory + Aspect is a variant of Co-Memory [41] that adds aspect information to the original paper as input to both the textual and visual memory networks. MIMN [29] is the first model to begin investigating multimodal aspect-level sentiment classification, which primarily utilizes a multi-hop memory network to extract textual and visual representations. TomBERT [42] is the earliest method to utilize pre-trained BERT for multimodal aspect-level sentiment analysis. Because the Multi-ZOL dataset is a Chinese dataset, we utilize bert-base-Chinese rather than bert-base-uncased, which was used by the original authors. ESAFN [30] is an aspect-sensitive attention and fusion network, which incorporates a gate unit to remove noise from the image, but since we cannot divide the text from the Multi-ZOL dataset into left and right texts, we did not conduct the experiment by separating the text into left and right as performed in ESAFN. Instead, we processed it in a manner consistent with our method. EF-CapTrBERT [31], which uses translation in the input space to translate images into text, is followed by multimodal fusion using an auxiliary sentence input to the encoder of a language model. MFSC-BERT is a method to replace Bi-LSTM in MFSC with BERT for text feature extraction. We do this to verify that Bi-LSTM as a lighter weight text feature extraction method is more suitable for our framework than BERT.

4.3. Experimental Results

We show the results of all the compared methods in Figure 4. From this figure, we can see the following points. (1) Multimodal methods based on images and texts generally perform better overall than the text-based unimodal method, suggesting that visual information is useful information for consumer reviews and that it is an important complement to text features. (2) Our framework consistently outperforms text-based unimodal methods, which suggests that relying on text information alone is not sufficient for the model to make correct judgements. For multimodal methods, visual information can bring additional features to the model or bring features with different meanings than those expressed by text features, which can help the model to diversify its considerations and make correct judgments. (3) It can be seen that the MFSC outperforms the compared methods on accuracy and macro-F1. Among them, text feature enhancement can extract text features more fully, multi-image gates can filter image noise effectively, and the interaction of text features and image features can also increase useful information. This validates the effectiveness of our proposed framework in multimodal aspect-level sentiment classification.

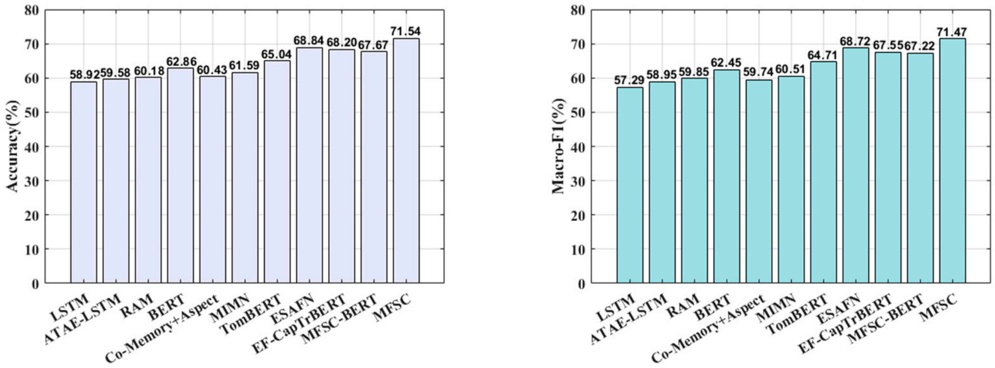
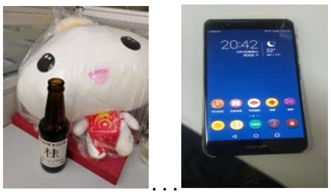


Figure 4. Performance comparison of different methods.

To further demonstrate the validity of MFSC, we show three test examples in Table 3. Where the first row represents the input image, the second row represents the input text, and the third row represents the input aspect information. MIMN, TomBERT, and ESAFN are all comparative multimodal methods. Here, the red score denotes the classification result is wrong and the blue score means the classification result is correct. In the first example, MIMN and TomBERT predicted a sentiment score of 6.0. For the second example, all but ESAFN predicted the correct result. For the third example, only MIMN predicted a sentiment score of 6.0. With all examples, our method predicts the correct category, which demonstrates the strength of our MFSC framework in predicting sentiment classification.

Table 3. Prediction of three test samples by different methods.

Image			
Text	The phone has a good design, light in the hand, thin and comfortable to touch.	The black version of the phone is particularly good-looking; taking pictures of the rear camera shoot effect is great; front camera has noise in selfies.	The exterior is one of Apple's most stylish mobile phones; red represents passion and youth; there is nothing wrong with it.
Aspect	Appearance and feeling	Photographing effect	Appearance and feeling
MIMN	6.0	8.0	6.0
TomBERT	6.0	8.0	8.0
ESAFN	8.0	10.0	8.0
MFSC	8.0	8.0	8.0

4.4. Ablation Study

To explore the impact of the modules of MFSC on the overall performance, different variants are tested, as shown in Table 4.

Table 4. Results of ablation experiments.

Module	Accuracy (%)	Macro-F1 (%)
ResNet-152-only	69.19	68.97
w/o Multi-Image Gate	69.96	69.83
w/o Text-Image Attention	70	69.97
w/o Adaptive Weights	71.02	70.90
MFSC	71.54	71.47

ResNet-152-only: Only visual features are extracted using ResNet-152; text features are extracted using LSTM, text features are processed using the text memory network, and textual and visual features are spliced together in the feature fusion stage.

w/o Multi-Image Gate: Using only ResNet-152 in the visual feature extraction stage, text features are extracted using LSTM and enhancing text features using the text memory network and adaptive weighting. The correlation is improved between text and images using the attention mechanism in the feature fusion stage. In this way, we investigate the impact of the Multi-Image Gate on MFSC.

w/o Text-Image Attention: In the visual feature extraction stage, visual features are extracted using ResNet-152 and noise is filtered using the Multi-Image Gate. Text features are extracted using LSTM and text features are enhanced using the text memory network and Adaptive Weight and finally text features and visual features are stitched together, in order to study the effect of Text-Image Attention on MFSC.

w/o Adaptive Weight: In the visual feature extraction stage, visual features are extracted using ResNet-152 and noise is filtered using the Multi-Image Gate. Text features are extracted using LSTM; text features are enhanced using the text memory network. And finally, features are obtained from text features, visual features, and Text-Image Attention, in order to study the effect of Adaptive Weight on MFSC.

As can be seen from Table 4, after greatly changing the structure of the framework, i.e., removing the Multi-Image Gate, the Text-Image Attention, and the Adaptive Weights, there is a significant decrease in the performance of the framework. This indicates that there is a significant deficiency in the features obtained from the network by relying solely on image feature extraction. Compared to our framework, the accuracy and macro-F1 value of the framework with the Multi-Image Gate removed decreased by 2.35% and 2.5%, respectively. This suggests that the Multi-Image Gate plays a role in fusing the feature information of multiple images and removing noise, which helps the model to capture the critical information of the images efficiently. The accuracy and macro-F1 value of the framework also decreased after the absence of the features obtained based on the Text-Image Attention, which reflects the importance of the interaction between text and image, and it illustrates that the interaction between text and image can improve the framework performance. In addition, there is a significant decrease in performance on the framework without Adaptive Weights, which indicates that the text features obtained at each hop of the text memory network are valuable.

4.5. Parameter Analysis

4.5.1. Impact of Different Hops

MFSC uses a memory network-based method for extracting text features. Therefore, the number of memory hops is a major hyperparameter that affects the performance. Referring to the MIMN [29], we show the experimental results of MFSC with 1 to 5 memory hops and compare them; the results are shown in Figure 5. When the number of the memory hop is 1, this leads to its worst results due to its inability to adequately learn the interaction of text and aspect features. The performance of MFSC becomes better as the number of memory hops increases, and it reaches its best performance when the memory hop is 3. However, when the memory hop is more than 3, the performance of MFSC does not become better; this is because it has higher complexity and lower generalization with the increase in the number of memory hops.

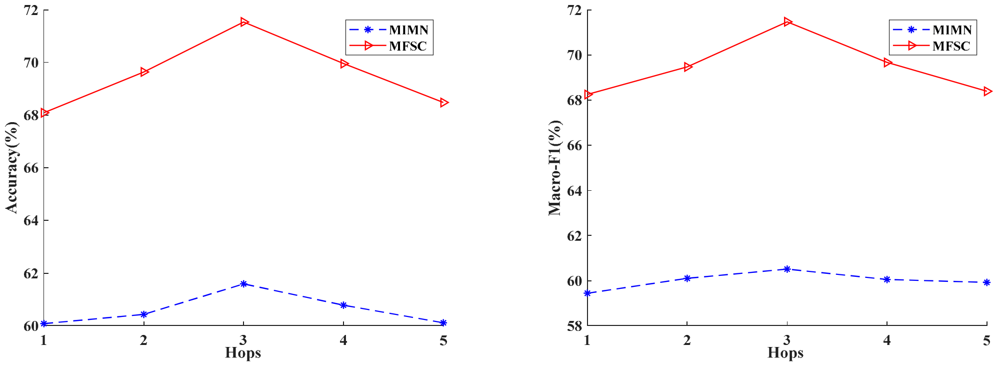


Figure 5. Results of different memory hops. Left figure shows the accuracy results and right figure shows the macro-F1 results.

4.5.2. Impact of Different Batch Sizes

In this section, the effect of batch sizes on the experimental results is tested, where batch sizes denote the number of data samples captured in a training session. The size of the batch sizes affects the model training; larger batch sizes yield more precise gradient estimates but can cause memory issues, slow convergence, and reduced generalization. And smaller batch sizes can lead to a better generalization error, but it will bring noise and even lead to non-convergence. Therefore, the choice of batch sizes is also an important issue. The results of the variation of accuracy and macro-F1 values under different sizes of batch sizes are shown in Figure 6, and we can see that MFAC works better when the batch size is 128.

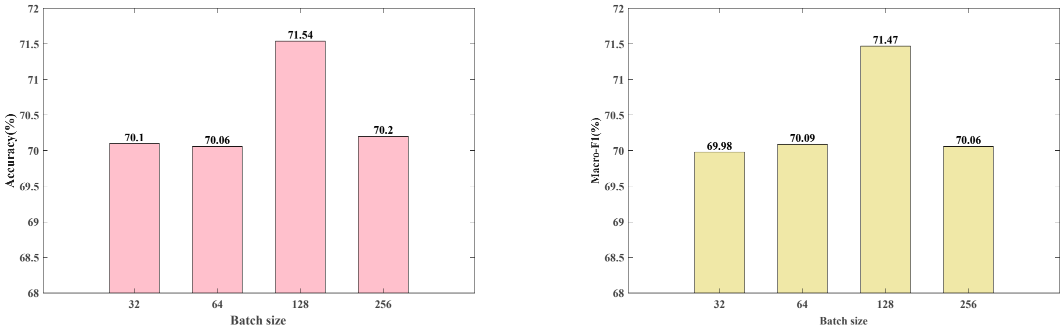


Figure 6. Results of different batch sizes.

5. Conclusions

In this paper, we propose a novel MFSC framework for aspect-level multimodal sentiment classification. MFSC utilizes multi-image gates to efficiently filter noise from multiple images, highlighting important visual information. Meanwhile, it enhances text features using a text memory network and adaptive weights. And the association between textual and visual information is enhanced using text-image attention in the multi-feature fusion stage. The effectiveness of MFSC is verified through comparison experiments with other methods. In practical terms, the MFSC framework can be applied to e-commerce platforms. For example, online retailers can use MFSC to analyze customer reviews that include text and images to gain insights into the sentiment expressed towards different aspects of products, such as quality, design, and functionality. This information can help in making decisions about marketing strategies and product improvements. We believe that by replacing pre-trained word-embedding vectors in other languages, this model can also be applied to other languages. The limitation of our approach is that the framework is somewhat complex, which makes the computation slower and leads to overfitting. Our subsequent work is how to streamline the structure of the framework so that the parameters and computation time of the framework can be reduced without affecting the performance of the framework. We believe that small and high-performance models are a promising direction for future research.

Author Contributions: Conceptualization, L.Z. and X.P.; methodology, L.Z. and X.P.; validation, X.P.; formal analysis, X.P.; writing—original draft preparation, L.Z. and X.P.; writing—review and editing, X.P. and X.C.; visualization, X.P.; supervision, X.C.; funding acquisition, L.Z. and X.C. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Key Program of Chongqing Education Science Planning Project (No. K22YE205098) and the Doctoral Research Foundation of Chongqing Normal University (No. 21XLB030, No. 21XLB029).

Data Availability Statement: Publicly available datasets were analyzed in this study. The data can be found at: https://github.com/xunan0812/MIMN/tree/master/datasets/zolDataset, accessed on 15 April 2023.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Pang, B.; Lee, L. Opinion mining and sentiment analysis. *Found. Trends Int. Ret* **2008**, *2*, 1–135. [CrossRef]
2. Kiritchenko, S.; Zhu, X.; Cherry, C.; Mohammad, S. Nrc-Canada-2014: Detecting aspects and sentiment in customer reviews. In Proceedings of the 8th International Workshop on Semantic Evaluation, Dublin, Ireland, 23–24 August 2014; pp. 437–442.
3. Hinton, G.; Salakhutdinov, R. Reducing the dimensionality of data with neural networks. *Science* **2006**, *313*, 504–507. [CrossRef] [PubMed]
4. Yu, K.; Jia, L.; Chen, Y.; Xu, W. Deep learning: Yesterday, today, and tomorrow. *J. Comput. Res. Dev.* **2013**, *20*, 1349.
5. Zhang, Q.; Fu, J.; Liu, X.; Huang, X. Adaptive co-attention network for named entity recognition in Tweets. In Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018; pp. 5674–5681.
6. Zhang, D.; Wei, S.; Li, S.; Wu, H.; Zhu, X.; Zhou, G. Multi-modal graph fusion for named entity recognition with targeted visual guidance. In Proceedings of the AAAI Conference on Artificial Intelligence, Vancouver, BC, Canada, 2–9 February 2021; pp. 14347–14355.
7. Zhai, Z.; Chen, H.; Li, R.; Wang, X. USSA: A Unified Table Filling Scheme for Structured Sentiment Analysis. In Proceedings of the Annual Meeting of the Association for Computational Linguistics, Toronto, ON, Canada, 9–14 July 2023; pp. 14340–14353.
8. Nguyen, H.-T.; Nguyen, L.M. ILWAANet: An interactive lexicon-aware word-aspect attention network for aspect-level sentiment classification on social networking. *Expert Syst. Appl.* **2020**, *146*, 113065. [CrossRef]
9. Park, H.J.; Song, M.; Shin, K.S. Deep learning models and datasets for aspect term sentiment classification: Implementing holistic recurrent attention on target-dependent memories. *Knowl.-Based Syst.* **2020**, *187*, 104825. [CrossRef]
10. Nazir, A.; Rao, Y.; Wu, L.; Sun, L. Issues and challenges of aspect-based sentiment analysis: A comprehensive survey. *IEEE Trans. Affect. Comput.* **2020**, *13*, 845–863. [CrossRef]
11. Wang, J.; Li, J.; Li, S.; Kang, Y.; Zhang, M.; Si, L. Aspect sentiment classification with both word-level and clause-level attention networks. In Proceedings of the 27th International Joint Conference on Artificial Intelligence, Stockholm, Sweden, 13–19 July 2018; pp. 4439–4445.
12. Tian, Y.; Yue, R.; Wang, D.; Liu, J.; Liang, X. Part-of-speech- and syntactic-aware graph convolutional network for aspect-level sentiment classification. *Multimed. Tools Appl.* **2023**, *83*, 28793–28806. [CrossRef]
13. Wu, D.; Wang, Z.; Zhao, W. XLNet-CNN-GRU dual-channel aspect-level review text sentiment classification method. *Multimed. Tools Appl.* **2023**, *83*, 5871–5892. [CrossRef]
14. Wang, Q.; Qian, Q. Malicious code classification based on opcode sequences and textCNN network. *J. Inform. Secur. Appl.* **2022**, *67*, 103151. [CrossRef]
15. Venugopalan, M.; Gupta, D. An enhanced guided LDA model augmented with BERT based semantic strength for aspect term extraction in sentiment analysis. *Knowl.-Based Syst.* **2022**, *246*, 108668. [CrossRef]
16. Liang, B.; Su, H.; Gui, L.; Cambria, E.; Xu, R. Aspect-based sentiment analysis via affective knowledge enhanced graph convolutional networks. *Knowl.-Based Syst.* **2022**, *235*, 107643. [CrossRef]
17. Yan, C.; Liu, J.; Liu, W.; Liu, X. Research on public opinion sentiment classification based on attention parallel dual-channel deep learning hybrid model. *Eng. Appl. Artif. Intel.* **2022**, *116*, 105448. [CrossRef]
18. Vo, D.-T.; Zhang, Y. Target-dependent twitter sentiment classification with rich automatic features. In Proceedings of the 24th International Joint Conference on Artificial Intelligence, Buenos Aires, Argentina, 25–31 July 2015; pp. 1347–1353.
19. Pang, B.; Lee, L.; Vaithyanathan, S. Thumbs up? Sentiment classification using machine learning techniques. In Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing, Philadelphia, PA, USA, 6 July 2002; pp. 79–86.
20. Zhang, Z.; Lan, M. ECNU: Extracting effective features from multiple sequential sentences for target-dependent sentiment analysis in reviews. In Proceedings of the 9th International Workshop on Semantic Evaluation, Denver, CO, USA, 4–5 June 2015; pp. 736–741.

21. Naz, S.; Sharan, A.; Malik, N. Sentiment classification on Twitter data using support vector machine. In Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence, Santiago, Chile, 3–6 December 2018; pp. 676–679.
22. Tang, D.; Qin, B.; Feng, X.; Liu, T. Effective lstms for target-dependent sentiment classification. In Proceedings of the COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, Osaka, Japan, 11–16 December 2016; pp. 3298–3307.
23. Xu, L.; Bing, L.; Lu, W.; Huang, F. Aspect Sentiment Classification with Aspect-Specific Opinion Spans. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Online, 16–20 November 2020; pp. 3561–3567.
24. Tang, D.; Qin, B.; Liu, T. Aspect level sentiment classification with deep memory network. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Austin, TX, USA, 1–4 November 2016; pp. 214–224.
25. Majumder, N.; Poria, S.; Gelbukh, A.; Akhtar, M.S.; Cambria, E.; Ekbal, A. Iarm: Inter-aspect relation modeling with memory networks in aspect-based sentiment analysis. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, 31 October 2018; pp. 3402–3411.
26. Ma, D.; Li, S.; Zhang, X.; Wang, H. Interactive attention networks for aspect-level sentiment classification. In Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence Main track, Vienna, Austria, 19–25 August 2017; pp. 4068–4074.
27. Wang, S.; Mazumder, S.; Liu, B.; Zhou, M.; Chang, Y. Target-sensitive memory networks for aspect sentiment classification. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, Melbourne, Australia, 15–20 July 2018; pp. 957–967.
28. Fan, F.; Feng, Y.; Zhao, D. Multi-grained attention network for aspect-level sentiment classification. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, 31 October 2018; pp. 3433–3442.
29. Xu, N.; Mao, W.; Chen, G. Multi-Interactive memory network for aspect based multimodal sentiment analysis. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January 2019; pp. 371–378.
30. Yu, J.; Jiang, J.; Xia, R. Entity-Sensitive attention and fusion network for entity-level multimodal sentiment classification. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2019**, *28*, 429–439. [CrossRef]
31. Khan, Z.; Fu, Y. Exploiting BERT for multimodal target sentiment classification through input space translation. In Proceedings of the MM '21: 29th ACM International Conference on Multimedia, Virtual Event, Chengdu, China, 20–24 October 2021; pp. 3034–3042.
32. Devlin, J.; Chang, M.; Lee, K.; Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, MN, USA, 2–7 June 2019; pp. 4171–4186.
33. Ling, Y.; Yu, J.; Xia, R. Vision-Language Pre-Training for Multimodal Aspect-Based Sentiment Analysis. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, Dublin, Ireland, 22–27 May 2022; pp. 2149–2159.
34. Ju, X.; Zhang, D.; Xiao, R.; Li, J.; Li, S.; Zhang, M.; Zhou, G. Joint Multi-modal Aspect-Sentiment Analysis with Auxiliary Cross-modal Relation Detection. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Online and Punta Cana, Dominican Republic, 7–11 November 2021; pp. 4395–4405.
35. Zhao, F.; Li, C.; Wu, Z.; Quyang, Y.; Zhang, J.; Dai, X. M2DF: Multi-grained Multi-curriculum Denoising Framework for Multimodal Aspect-based Sentiment Analysis. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Singapore, 6–10 December 2023; pp. 9057–9070.
36. Han, W.; Chen, H.; Hai, Z.; Poria, S.; Bing, L. SANCL: Multimodal Review Helpfulness Prediction with Selective Attention and Natural Contrastive Learning. In Proceedings of the 29th International Conference on Computational Linguistics, Gyeongju, Republic of Korea, 12–17 October 2022; pp. 5666–5677.
37. Lu, M.; Li, R.; Feng, F.; Ma, Z.; Wang, X. LGR-NET: Language Guided Reasoning Network for Referring Expression Comprehension. *IEEE Trans. Circuits Syst. Video Technol.* **2024**. [CrossRef]
38. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
39. Wang, Y.; Huang, M.; Zhu, X.; Zhao, L. Attention-based LSTM for aspect-level sentiment classification. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Austin, TX, USA, 1–4 November 2016; pp. 606–615.
40. Chen, P.; Sun, Z.; Bing, L.; Yang, W. Recurrent Attention Network on Memory for Aspect Sentiment Analysis. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Copenhagen, Denmark, 7–11 September 2017; pp. 452–461.
41. Xu, N.; Mao, W.; Chen, G. A co-memory network for multimodal sentiment analysis. In Proceedings of the SIGIR '18: The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, Ann Arbor, MI, USA, 8–12 July 2018; pp. 929–932.
42. Yu, J.; Jiang, J. Adapting BERT for target-oriented multimodal sentiment classification. In Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, Macao, China, 10–16 August 2019; pp. 5408–5414.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.