



## OPEN ACCESS

## EDITED BY

Yu Liu,  
Hefei University of Technology, China

## REVIEWED BY

Yunchun Zhang,  
Yunnan University, China  
Zhongqing Wang,  
Soochow University, China

## \*CORRESPONDENCE

Junjun Guo,  
✉ guojjgb@163.com

RECEIVED 16 March 2023

ACCEPTED 12 April 2023

PUBLISHED 10 May 2023

## CITATION

Xiang Y, Cai Y and Guo J (2023), MSFNet: modality smoothing fusion network for multimodal aspect-based sentiment analysis. *Front. Phys.* 11:1187503. doi: 10.3389/fphy.2023.1187503

## COPYRIGHT

© 2023 Xiang, Cai and Guo. This is an open-access article distributed under the terms of the [Creative Commons Attribution License \(CC BY\)](https://creativecommons.org/licenses/by/4.0/). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

# MSFNet: modality smoothing fusion network for multimodal aspect-based sentiment analysis

Yan Xiang<sup>1,2</sup>, Yunjia Cai<sup>1,2</sup> and Junjun Guo<sup>1,2\*</sup>

<sup>1</sup>Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming, China, <sup>2</sup>Yunnan Key Laboratory of Artificial Intelligence, Kunming University of Science and Technology, Kunming, China

Multimodal aspect-based sentiment classification (MABSC) aims to determine the sentiment polarity of a given aspect in a sentence by combining text and image information. Although the text and the corresponding image in a sample are associated with aspect information, their features are represented in distinct semantic spaces, creating a substantial semantic gap. Previous research focused primarily on identifying and fusing aspect-level sentiment expressions of different modalities while ignoring their semantic gap. To this end, we propose a novel aspect-based sentiment analysis model named modality smoothing fusion network (MSFNet). In this model, we process the unimodal aspect-aware features via the feature smoothing strategy to partially bridge modality gap. Then we fuse the smoothed features deeply using the multi-channel attention mechanism, to obtain aspect-level sentiment representation with comprehensive representing capability, thereby improving the performance of sentiment classification. Experiments on two benchmark datasets, Twitter2015 and Twitter2017, demonstrate that our model outperforms the second-best model by 1.96% and 0.19% in terms of Macro-F1, respectively. Additionally, ablation studies provide evidence supporting the efficacy of each of our proposed modules. We release the code at: <https://github.com/YunjiaCai/MSFNet>.

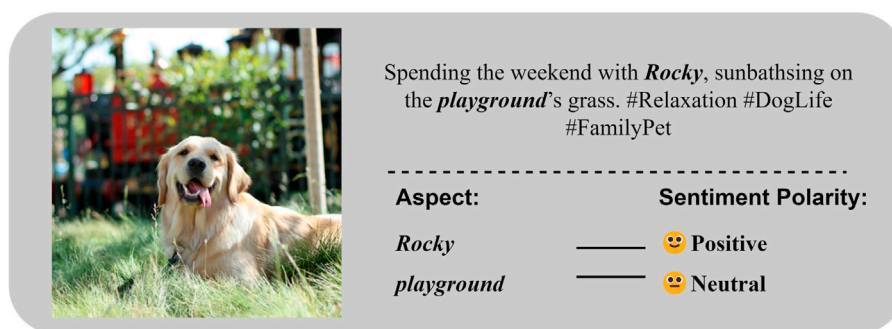
## KEYWORDS

multimodal sentiment analysis, aspect-based sentiment analysis, multimodal fusion, feature smoothing, semantic gap

## 1 Introduction

In recent years, there has been a significant increase in the amount of multimodal data from various social, shopping, and news platforms. These data consist primarily of a piece of text and an associated image, and are often accompanied by a personal sentiment tendency. Analyzing the sentiment towards specific aspects in this type of data can provide valuable insights into people's personalized preferences or predict public opinion trends. Therefore, multimodal aspect-based sentiment classification (MABSC) has received extensive attention. The objective of this task is to combine a piece of text, its associated image and a given aspect from the text to determine the sentiment polarity of the given aspect. As shown in [Figure 1](#), the sentiment polarity of the aspect {Rocky} could be determined as {Neutral}, according to the text alone. However, by combining image information, it can be determined that the aspect term has a {Positive} sentiment polarity. Therefore, the key to this task lies in effectively extracting and combining the sentiment features from both images and texts.

From the feature learning perspective, images and texts are commonly represented in distinct feature spaces, which creating a semantic gap between the two modalities and posing substantial



**FIGURE 1**  
Example of MABSC tasks.

challenges for subsequent inter-modal interactions [1, 2]. As a result, the major difficulty of MABSC is to bridge the gap between modalities and model the deep interactions of them. Early MABSC research primarily relied on directly modeling the interaction between modalities to achieve multimodal fusion. Xu et al. [3] proposed a memory-based model which extracted text and image features using pre-trained Bert and ResNet models respectively, and stacked interactive attention mechanism with several memory hops to learn the deep abstraction of multimodal data. Similarly, Zhang et al. [4] sent features of two different modalities into a fusion discriminant matrix to learn the interaction of different modalities and a similarity matrix is used to capture modal invariant features, based on which the consistency and redundancy of different modalities can be identified. However, the deficiency of these methods was that they did not consider the semantic gaps on subsequent interactions. Khan et al. [1] recognized the influence of semantic gaps on multimodal fusion and used a cross-modal Transformer to map image content to the text space. They then utilized a pre-trained Bert structure to model the interactions between image, text, and aspect. However, the performance is limited due to the lack of in-depth exploration of inter-modal interactions.

To tackle the problem of insufficient fusion, we propose a novel MABSC model called “modality smoothing fusion network (MSFNet)”. The main contribution can be summarized as follows.

- Unlike existing works of MABSC that mainly study extracting and fusing aspect-level sentiment expressions, we focus on the problem that modality discrepancy influence their subsequent fusions.
- The proposed MSFNet adopts the feature smoothing strategy and the multi-channel attention to effectively bridge the semantic gap and achieve better fusion of text-image modalities.
- Experimental results on two benchmark datasets verify that MSFNet achieve effective interaction of multimodalities and obtains state-of-the-art performance in MABSC.

## 2 Related work

Aspect-based sentiment classification (ABSC) was first proposed on text datasets. With the increase of multimodal data, multimodal

sentiment analysis (MSA) gained great attention, and MABSC is the research combining ABSC and MSA.

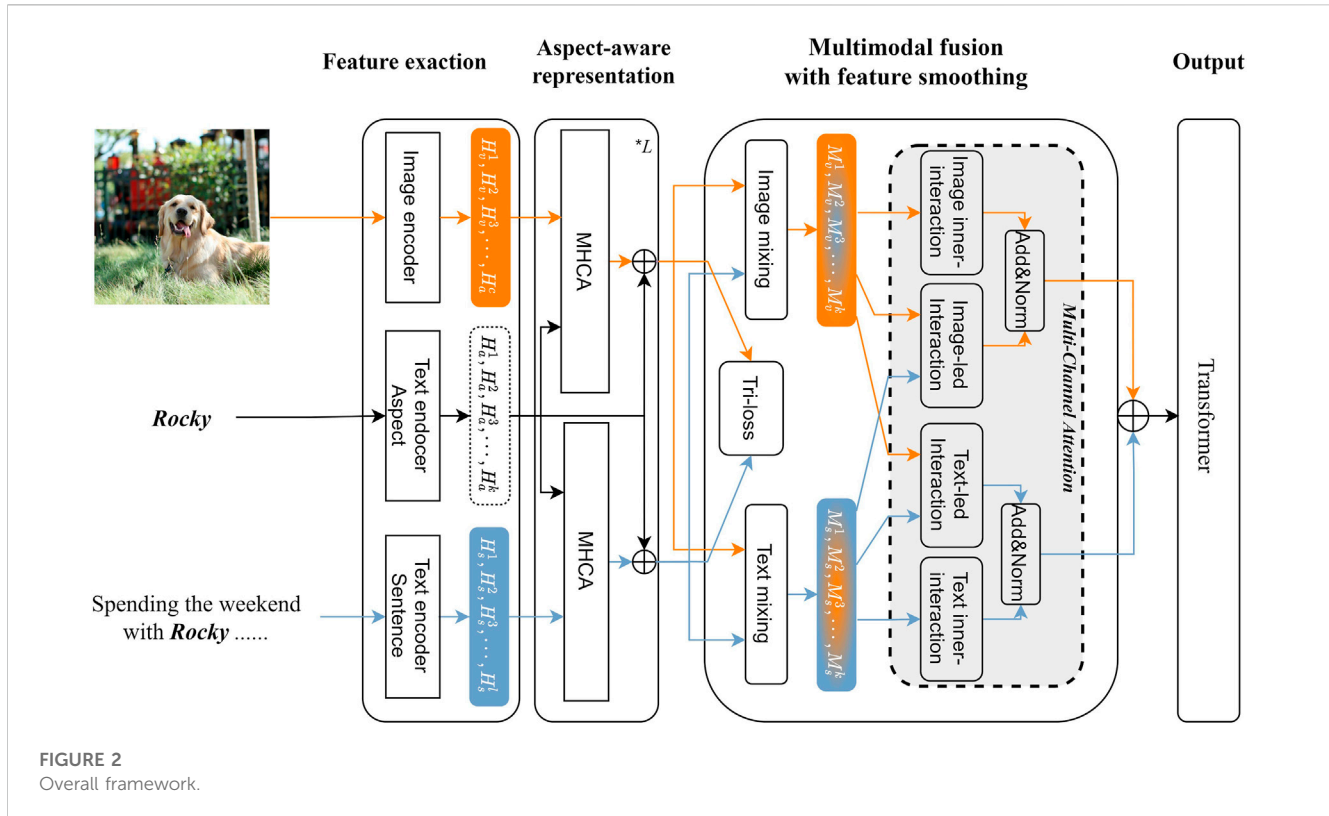
### 2.1 Aspect-based sentiment classification (ABSC)

Aspect-Based Sentiment Classification (ABSC) is a task that involved predicting the sentiment polarity of a target entity within a given sentence. Traditional methods for ABSC relied on manually annotated features, such as language rules [5] and feature engineering [6]. In recent years, neural networks have shown great promise in this area and have led to significant performance improvements. Early neural network approaches typically used Long Short-Term Memory to model the interaction between the aspect and its context [7]. More recent works have incorporated attention mechanisms to select aspect-related sentiment features [8], with some studies introducing more complex interactive attention methods to learn aspect-specific representations [9, 10]. These methods demonstrate the significance of contextual information in the task of aspect sentiment analysis. Pre-trained language models, such as BERT [11], have also been utilized to improve the ABSC performance [12].

### 2.2 Multimodal sentiment Analysis (MSA)

MSA aims to combine multimodal information such as text, visual, and audio to understand human emotions [13]. Previous researchers have primarily focused on unimodal representation learning and multimodal fusion.

Unimodal representation learning: Wang et al. [14] constructed a recurrent variational embedding network that projects text representations into a common space by calculating offset vectors between linguistic and non-linguistic information. Hazarika et al. [15] proposed modality-invariant and modality-specific representations to learn complementary information between modalities, reducing redundancy and merging a set of diverse information. Yu et al. [16] designed a label generation module based on a self-supervised learning strategy to capture consistency and differences between three modalities by jointly



learning unimodal and multimodal tasks. Effective unimodal representations can mitigate the impact of the semantic gap.

**Multimodal fusion:** For multimodal fusion, Zadeh et al. [17] proposed a tensor fusion network that obtains multimodal fusion representation by calculating outer products between all unimodal representations. Liu et al. [18] proposed an improved low-rank multimodal fusion network based on tensor fusion network, which uses low-rank tensors to reduce the computational complexity of tensor-based methods and achieve better performance. Zadeh et al. [19] proposed a memory fusion network that first models unimodal representations using LSTM, and then models intermodal interactions using Delta-memory Attention Network and Multi-view Gated Memory. Transformer structures are widely used to model interactions between modalities due to the success of Transformer-based models. Tsai et al. [20] used Directional Cross-Modal Attention modules to extend the standard Transformer network [21] for modeling unaligned multimodal language sequences. Wang et al. [22] used forward and backward translation from one modality to another and back to better fuse multimodal features. Modeling efficient interactions between different modalities can fully utilize information between modalities for multimodal emotional expression.

### 2.3 Multimodal aspect-based sentiment Classification (MABSC)

MABSC is the research combining ABSC and MSA. Similar to text aspect-based sentiment classification, different parts of the sentence and image play different roles in specific aspects, and

attention mechanisms are widely used to obtain aspect-specific representations. Xu et al. [3] first used interactive attention to obtain aspect-specific unimodal representations, and then stacked several interactive attention mechanisms and memory hops to learn deep abstractions for multimodal data. Zhang et al. [4] used an aspect-sensitive memory network to capture intra-modal features, then designed a fusion discriminative matrix to learn interactions between different modalities. Inspired by the success of BERT-based models, Yu et al. [23] proposed a target-oriented multimodal BERT (TomBERT), which constructs a BERT-based structure to match the target text and target image and capture dynamics within and between modalities. Khan et al. [1] used a pre-trained transformer-based image captioning model to convert images into textual image captions, then fused information from both modalities by constructing sentence pairs and inputting the image caption, aspect, and original sentence into a BERT language model. Yu et al. [2] modeled pairwise interactions between inputs using an interactive transformer, and bridged the semantic gap between the two modalities by calculating the loss between the representations of the two modalities and the original context. Additionally, Huang et al. [24] constructed sequential cross-modal semantic graphs to fully extract the information contained in the image, and used an encoder-decoder model with a target prompt template to achieve MABSC task.

The importance of integrating image information into text information has been repeatedly proved in the research of MABSC. However, this integration invariably encounters the issue of semantic gaps between two modalities. Therefore, we focus on easing the semantic gap before integration.

### 3 Methodology

In this section, we first give the definition of multimodal aspect-based sentiment analysis task, and introduce the overall framework of the proposed model. Then, we present the details of each module of the proposed model.

#### 3.1 Task definition

Given a set of multimodal dataset  $D$ , each sample  $d \in D$  includes a context sentence  $t$ , an associated image  $i$ , a given aspect  $a$ , and a golden label  $y$ . Specifically, the sentence  $t = (w_1, w_2, w_3, \dots, w_m)$ , where  $m$  is the length of the sentence. The given aspect is a sub-sequence of sentence  $t$  and is represented as  $a = (w_x, w_{x+1}, \dots, w_{x+n})$ , where  $n$  is the length of the given aspect. As shown in Figure 2, this task is to take  $t$ ,  $i$  and  $a$  as inputs to determine the sentiment polarity  $y \in \{Positive, Neutral, Negative\}$  associated with the given aspect  $a$ .

#### 3.2 Overview of the proposed model

The overall architecture of the model is shown in Figure 2, which consists of a feature extraction layer, an aspect-aware representation layer, a multimodal fusion layer with feature smoothing, and an output layer. We extract separate representations of the image, text and aspect in the feature extraction layer. In the aspect-aware representation layer, we mine the aspect-related representations of each modality with the guidance of the aspect. In the multimodal fusion layer, we use feature smoothing strategy and multi-channel attention to model the deep interaction between the two modalities. Finally, we obtain the sentiment polarities in the output layer.

#### 3.3 Feature extraction layer

We utilize two different unimodal feature encoders to extract original representations of the text and image inputs.

##### 3.3.1 Text encoder

The pre-training language model BERT [11] can capture advanced text representations. To distinguish sentence and aspect representations, we fine-tune two different pre-trained BERTs to encode sentence and aspect respectively. Specifically, for the input sentence, we add a special token [CLS] in front of the original sentence and a special token [SEP] in the back to form new tokens  $\mathbf{I}_s \in \mathbb{R}^l$ , and then input  $\mathbf{I}_s$  to a pre-trained BERT to obtain the encoded sentence representation  $\mathbf{h}_s$ , as follows:

$$\mathbf{h}_s = BERT(\mathbf{I}_s) \tag{1}$$

where  $\mathbf{h}_s \in \mathbb{R}^{l \times d_t}$  is the obtained sentence representation,  $d_t$  is the hidden dimension.

Similarly, for a given aspect, we add the special tokens [CLS] and [SEP] to form tokens  $\mathbf{I}_a \in \mathbb{R}^k$ , and then input  $\mathbf{I}_a$  into another pre-trained BERT to obtain the encoded aspect representation  $\mathbf{h}_a$ , as follows:

$$\mathbf{h}_a = BERT(\mathbf{I}_a) \tag{2}$$

where  $\mathbf{h}_a \in \mathbb{R}^{k \times d_t}$  is the obtained aspect representation.

After obtaining the sentence and aspect representation, we use the linear layer to map their hidden dimension to the same dimension  $d_h$  for the subsequent interaction:

$$\mathbf{H}_s = \mathbf{W}_1 \mathbf{h}_s + \mathbf{b}_1 \tag{3}$$

$$\mathbf{H}_a = \mathbf{W}_2 \mathbf{h}_a + \mathbf{b}_2 \tag{4}$$

where  $\mathbf{H}_s \in \mathbb{R}^{l \times d_h}$  and  $\mathbf{H}_a \in \mathbb{R}^{k \times d_h}$ .

##### 3.3.2 Image encoder

Different from coarse grained sentiment analysis tasks, MABSC should focus on aspect-related information to determine the sentiment polarity. We use the object detection model Faster R-CNN [25] to extract aspect-level features of images. Specifically, we input the image  $i$  into a pre-trained Faster R-CNN model to obtain the candidate regions in the image, and retain the features with the highest confidence as image features:

$$\mathbf{h}_i = \text{FasterR-CNN}(i) \tag{5}$$

where  $\mathbf{h}_i \in \mathbb{R}^{c \times d_v}$  is the obtained image representation,  $c$  denotes the number of image regions retained, and  $d_v$  is the hidden dimension of Faster R-CNN.

Then we use a linear layer to map the hidden dimension of image representation to  $d_h$ :

$$\mathbf{H}_i = \mathbf{W}_3 \mathbf{h}_i + \mathbf{b}_3 \tag{6}$$

where  $\mathbf{H}_i \in \mathbb{R}^{c \times d_h}$ .

We obtain the final image representation by a multi-head self attention (MHSA) [21] to pay more attention to the important image regions:

$$\mathbf{H}_v = \text{MHSA}(\mathbf{H}_i) \tag{7}$$

where  $\mathbf{H}_v \in \mathbb{R}^{c \times d_h}$ .

#### 3.4 Aspect-aware representation layer

After obtaining the initial sentence representation and image representation, we need to further interact them with the aspect representation to focus on aspect-related information. We adopt an interactive attention mechanism to enable interaction between the aspect representation and unimodal representation, and retain more aspect representations through residual connections. Specifically, we use the aspect representation as the query, and the sentence representation as the key-value in the multi-head cross attention (MHCA) [21], to generate the aspect-sentence representation, as follows:

$$\mathbf{R}_s = \text{MHCA}(\mathbf{H}_a, \mathbf{H}_s) \tag{8}$$

where  $\mathbf{R}_s \in \mathbb{R}^{k \times d_h}$ .

Then we add the aspect-sentence representation and the aspect representation, and perform one layer normalization (LN) to obtain the one-layer aspect-aware text representation:

$$\mathbf{A}_s = \text{LN}(\mathbf{R}_s + \mathbf{H}_a) \tag{9}$$

where  $\mathbf{A}_s \in \mathbb{R}^{k \times d_h}$ .

Finally, we stack  $l$  layers of the aspect-aware layer to learn the deep interaction of aspect and text, as follows:

TABLE 1 Dataset statistics.

	Twitter2015			Twitter2017		
	Train	Dev	Test	Train	Dev	Test
Positive	928	303	317	1508	515	493
Neutral	1883	670	607	1638	417	573
Negative	368	149	113	416	144	168
Total Samples	3179	1122	1037	3562	1176	1234
Avg Aspect	1.348	1.336	1.354	1.410	1.439	1.450
Avg Length	16.72	16.74	17.05	16.21	16.37	16.38
Max Length	35	40	36	39	31	38
Total Sentence	2101	727	674	1746	577	587

$$\mathbf{A}_{s(l)} = LN(MHCA(\mathbf{H}_a, \mathbf{A}_{s(l-1)}) + \mathbf{H}_a) \quad (10)$$

where  $\mathbf{A}_{s(l)} \in \mathbb{R}^{k \times d_h}$  is the final aspect-aware text representation, and  $l$  is the number of stacked layers.

For the image representation, we input it together with the aspect representation into the similar aspect-aware layer to obtain the aspect-aware image representation, as follows:

$$\mathbf{A}_{v(l)} = LN(MHCA(\mathbf{H}_a, \mathbf{A}_{v(l-1)}) + \mathbf{H}_a) \quad (11)$$

where  $\mathbf{A}_{v(l)} \in \mathbb{R}^{k \times d_h}$ .

### 3.5 Multimodal fusion layer with feature smoothing

After obtaining the aspect-aware representations of two modalities, we propose a multimodal fusion layer with feature smoothing to combine information from different modalities. Firstly, to relieve the semantic gap between the two modalities, a feature smoothing strategy is used to smooth the aspect-aware representations of the two modalities. Then we use a multi-channel attention interaction network to achieve deep interaction between the two modal representations.

#### 3.5.1 Feature smoothing

We integrate the partial representation of one modality into the representation of another modality via a feature-level mixing approach, and obtain two smoothed unimodal representations, as follows:

$$\mathbf{M}_s = W_{mix} * \mathbf{A}_{s(l)} + (1 - W_{mix}) * \mathbf{A}_{v(l)} \quad (12)$$

$$\mathbf{M}_v = W_{mix} * \mathbf{A}_{v(l)} + (1 - W_{mix}) * \mathbf{A}_{s(l)} \quad (13)$$

where  $W_{mix}$  is a hyperparameter. The obtained  $\mathbf{M}_s$  and  $\mathbf{M}_v$  are the smoothed text and image representations, respectively. We will use these smoothed representations for further interaction.

In addition, we use the average representation of the two modalities as an anchor, to bridge the semantic gap between the two modalities via the constraint of the mean square error Tri-loss:

$$\mathbf{A}_l = MEAN(\mathbf{A}_{s(l)}, \mathbf{A}_{v(l)}) \quad (14)$$

TABLE 2 The hyperparameter Setting.

	Twitter2015	Twitter2017
Learning rate	2e-5	4e-5
Warm up step	37	35
$l$	2	1
$W_{mix}$	0.85	0.85
$(\alpha_1, \alpha_2, \alpha_3)$	(1,1,0.5)	(1,1,0.5)
$\lambda$	4e-3	4e-3
Batch size	32	32
Attention heads	8	8
Attention dimension	512	512

$$L^{tri} = \alpha_1 * MSE(\mathbf{A}_{s(l)}, \mathbf{A}_l) + \alpha_2 * MSE(\mathbf{A}_{v(l)}, \mathbf{A}_l) + \alpha_3 * MSE(\mathbf{A}_{s(l)}, \mathbf{A}_{v(l)}) \quad (15)$$

where the *MEAN* operator refers to averaging values of each dimension in the two tensors. *MSE* is mean square error loss, and  $(\alpha_1, \alpha_2, \alpha_3)$  are hyperparameters. The above loss will be added to the main loss to guide the training of the model parameters.

#### 3.5.2 Multi-channel attention-based interaction

In order to effectively utilize the complementary information between modalities to enhance the expression of sentiment, we propose a multi-channel attention interaction network (MCA) including four channels, named text self-attention, text-led multimodal attention, image self-attention and image-led multimodal attention channels respectively.

In the text self-attention channel, we use a multi-head self attention to process the smoothed text representation acquired in the preceding stage and obtain the text inner-interaction representation, denoted as  $\mathbf{CS}_s \in \mathbb{R}^{k \times d_h}$ :

$$\mathbf{CS}_s = MHSA(\mathbf{M}_s) \quad (16)$$

In the text-led multimodal attention channel, we take the smoothed text representation as the query and the smoothed image representation as the key-value, and sent them to a multi-head interactive attention network, to obtain the text-led inter-interaction representation, denoted as  $\mathbf{CC}_s \in \mathbb{R}^{k \times d_h}$ :

$$\mathbf{CC}_s = MHCA(\mathbf{M}_s, \mathbf{M}_v) \quad (17)$$

Final, we add up the representations of the two channels and normalize it to obtain the text-led multimodal representation  $\mathbf{F}_s \in \mathbb{R}^{k \times d_h}$ :

$$\mathbf{F}_s = LN(\mathbf{CS}_s + \mathbf{CC}_s) \quad (18)$$

Similarly, following the same procedure as the two text channels above, we feed the smoothed image representation into the two image channels to obtain the image inner-interaction representation and the image-led inter-interaction representation. We then add and normalize them to obtain the image-led multimodal representation  $\mathbf{F}_v \in \mathbb{R}^{k \times d_h}$ .



TABLE 3 Comparison of our method and baseline Macro-F1.

Modality	Method	Twitter2015	Twitter2017
Visual	Res-Aspect	46.58	54.01
	FasterRCNN-Aspect	37.71	54.71
Text	IAN [9]	63.32	63.32
	MGAN [10]	64.21	61.46
	BERT [11]	70.01	66.15
	Res-BERT	71.46	66.89
Text + Visual	Faster R-CNN-BERT	70.85	66.21
	TomBERT (ResNet) [23]	71.75	68.04
	TomBERT (FasterR-CNN) [2]	72.95	68.49
	ModalNet [4]	72.50	69.19
	IFNRA [27]	71.79	69.48
	MSFNet (Ours)	74.46	69.67

After obtaining the two representations  $F_s$  and  $F_v$ , we concatenate them and send it to a transformer and a average pooling, to get the final multimodal sentiment representation  $H_m \in \mathbb{R}^{d_h}$ :

$$F_m = [F_s; F_v] \tag{19}$$

$$H_m = \text{averagepooling}(\text{Transformer}(F_m)) \tag{20}$$

### 3.6 Output layer

We send the multimodal sentiment representation  $H_m$  to a fully connected layer and a softmax layer to obtain the classification result:

$$p(y|H_m) = \text{softmax}(W_c H_m + b_c) \tag{21}$$

where  $W_c \in \mathbb{R}^{r \times d_h}$  and  $b_c \in \mathbb{R}^r$  are learnable parameters,  $y \in \mathbb{R}^r$  is the probability distribution of sentiment polarity,  $r$  is number of classes.

The loss function of the model is as follows:

$$L = -\frac{1}{N} \sum_i \left( \sum_j g_{ij} \log p(y_{ij}|H_m) - \lambda L_i^{tri} \right) \tag{22}$$

where  $g_{ij}$  is the golden label,  $\lambda$  is a hyperparameter.

## 4 Experimental

In this section, we conducted comprehensive experiments on the proposed overall model and its individual modules.

### 4.1 Experiment setting

Datasets: We adopt two standard datasets Twitter15 and Twitter17 to evaluate the performance of our model. Twitter15 and Twitter17 datasets contain multimodal tweets

TABLE 4 Ablation study of feature-level mixing (Macro-F1).

Method	Twitter2015	Twitter2017
MSFNet (Ours)	74.46	69.67
w/o feature mixing	72.83	68.23
w/o Text mixing	73.88	68.91
w/o Image mixing	73.86	68.92

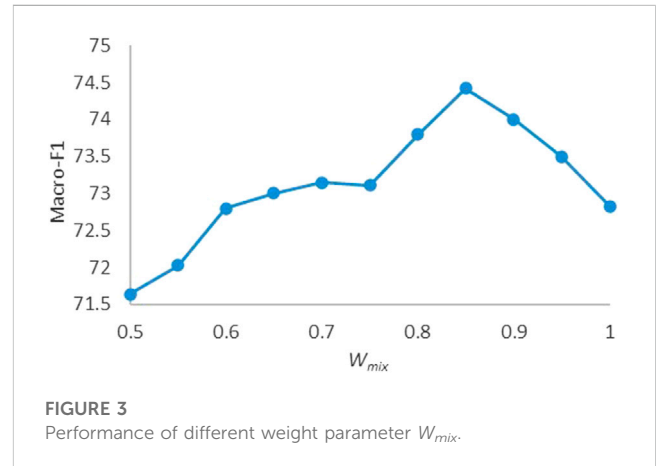


FIGURE 3 Performance of different weight parameter  $W_{mix}$ .

published on Twitter between 2014–2015 and 2016–2017 respectively. These datasets were originally annotated the given aspect by Zhang et al. [26] for the Multimodal Named Entity Recognition (MNER) task, and then Yu et al. [23] annotated the sentiment polarity of each given aspect for the MABSA task. The datasets provide tweet text, tweet image, aspect and the sentiment polarity of the given aspect. The specific data statistics are shown in Table 1.

Evaluation Metrics: To measure the performance of different approaches, we use Macro-F1 as evaluation metrics, as follows:

$$Macro - F1 = \frac{1}{r} \sum_{i=1}^r F1_i \tag{23}$$

$$F1_i = \frac{2 * P_i * R_i}{P_i + R_i} \tag{24}$$

where  $F1_i$  is the *f1-score* of class  $i$ ,  $P_i$  and  $R_i$  are the precision and recall of class  $i$ , and  $r$  is the number of classes.

Implement Details: For text input, we leverage the pre-trained BERT [11] model to encode the text. For image input, we utilized the Faster R-CNN structure proposed by Anderson et al. [25] and used a pre-trained Faster R-CNN model to extract region features of the image. We fix all the hyper-parameters after tuning them on the development set. The specific hyperparameter settings are shown in Table 2. We implemented all models in the PyTorch framework and ran experiments on RTX3090 GPU.

### 4.2 Baseline

In this section, we use the following methods as baselines to compare with our model.

TABLE 5 Ablation study of Tri-loss (Macro-F1).

Method	Twitter2015	Twitter2017
MSFNet (Ours)	74.46	69.67
Rep anchor of Tri-loss	73.71	69.02
w/o Tri-loss	73.08	68.36

TABLE 6 Ablation study of MCA (Macro-F1).

Method	Twitter2015	Twitter2017
MSFNet (Ours)	74.46	69.67
w/o MCA	71.21	67.47
Rep MCA to CMT	70.73	67.24

- Res-Aspect: ResNet and BERT are used to extract image and aspect features respectively, and an attention layer is used to obtain multimodal representation.
- Faster R-CNN-Aspect: Another baseline is similar to Res-Aspect, but image features are extracted by Faster R-CNN.
- IAN [9]: Capturing the interaction between aspect and context with bidirectional interactive attention.
- MGAN [10]: Based on IAN, a fine-grained attention is further proposed for interaction.
- BERT [11]: Sentence pairs constructed by context and aspect are fed into pre-trained BERT for sentiment classification.
- Res-BERT: The context and aspect are input as sentence pairs into a pre-trained BERT model to obtain text features. The image features are extracted by ResNet. And then modeling multimodal interaction using attention.
- FasterR-CNN-BERT: Another baseline is similar to Res-BERT, but image features are extracted by Faster R-CNN.
- TomBERT (ResNet) [23]: A target-oriented multimodal BERT architecture that utilizes ResNet for image representation, and leverages multiple BERT structures for text feature extraction, image aspect interaction, and multimodal interaction.
- TomBERT (Faster R-CNN) [2]: Same structure as TomBERT (ResNet), but the image representation is obtained by Faster R-CNN.
- ModalNet [4]: Use aspect-sensitive memory network to perform aspect-sensitive fusion of two modalities, and construct a fusion discriminant matrix to obtain multimodal sentiment representation.
- IFNRA [27]: Use GRU to achieve image denoising and multimodal fusion. And a decoder with recurrent attention is designed to gradually learn aspect-specific sentiment features.

### 4.3 Main result

Table 3 shows the performance of different methods on the twitter2015 and twitter2017 datasets. The following observations can be drawn: (1) Our model has achieved the best performance on

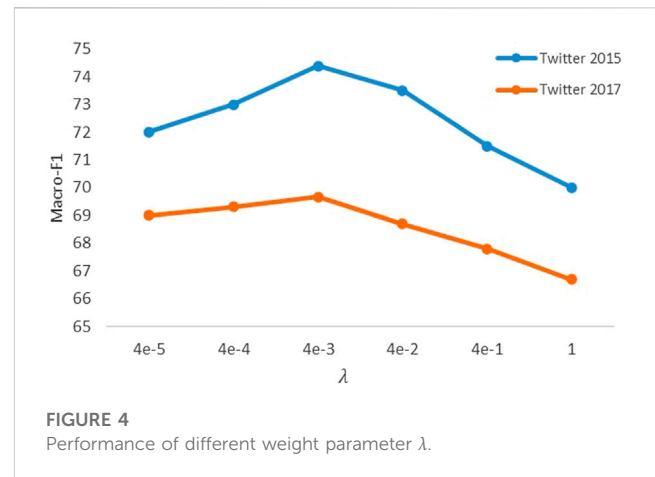


FIGURE 4 Performance of different weight parameter  $\lambda$ .

the two datasets, which are respectively improved by 1.96% and 0.19% compared with the second best model. This illustrates that our proposed multimodal fusion method is effective and has obvious advantages. (2) Sufficient multimodal fusion can effectively improve classification performance. For example, both TomBERT (Faster R-CNN) and Faster R-CNN-BERT use Faster R-CNN to extract regional features, but the latter performs much worse than the former because it only performs simple multimodal fusion. Similarly, for the models that use ResNet to extract image features, TomBERT (ResNet) shows better performance than Res-BERT, but it is still not as good as ModalNet. Our proposed method has significant advantages when compared to ModalNet. The latter focuses on multimodal fusion without considering the semantic gap of multimodal features. Our proposed model performs feature smoothing before multimodal fusion, which enables deeper interactions and achieves better performance. (3) Using the regional features extracted by FasterR-CNN can help the model focus on the object-level information in images. However, if the model cannot obtain information enabling to expressing sentiments from the image representation via a good image-text interaction method, using FasterR-CNN may result in performance degradation. This conclusion can be drawn from comparing Res-BERT and Faster R-CNN-BERT, as well as Res-Aspect and Faster R-CNN-Aspect. (4) The performance of image-based methods is much lower than that of text-based methods among the unimodal-based methods. This is mainly because the given aspect is a subsequence in the initial sentence. If image information is considered alone, it may introduce some noises that have nothing to do with the given aspect, resulting in wrong classification.



### 4.4 Ablation study of multimodal fusion layer

In this section, we conduct ablation studies to verify the effectiveness of multimodal fusion layer with feature smoothing.

#### 4.4.1 Feature-level mixing

To test the effect of feature-level mixing, we feed the unprocessed aspect-aware representations into the multi-channel attention interaction network instead of smoothed representations. The results are shown in Table 4.

**TABLE 7 Comparison between predicted results and golden labels for several representative samples on Bert, Faster R-CNN-BERT and MSFNet (Ours), respectively.**

Image		
Text	(a) <i>Charlie</i> is decidedly not excited about @ <i>ussoccer_ynt</i> at 4 am. #U20WC	(b) The final chapter of the fairytale— <i>Leicester</i> gear up for historic <i>Premier League</i> title
Golden Label	(Charlie, Negative)	(Leicester, Positive)
	(ussoccer_ynt, Neutral)	(Premier League, Neutral)
Bert	(Charlie, Neutral) ✗	(Leicester, Neutral) ✗
	(ussoccer_ynt, Neutral) ✓	(Premier League, Neutral) ✓
FasterR-CNN-BERT	(Charlie, Negative) ✓	(Leicester, Neutral) ✗
	(ussoccer_ynt, Neutral) ✓	(Premier League, Neutral) ✓
MSFNet (Ours)	(Charlie, Negative) ✓	(Leicester, Positive) ✓
	(ussoccer_ynt, Neutral) ✓	(Premier League, Neutral) ✓

It can be seen that if the unprocessed aspect-aware features of one modality are used to interact with the smooth features of another modality, the Macro-F1 of the twitter15 and twitter17 datasets drop by about 0.5% and 0.7%, respectively, compared to the full model. If all four interaction channels use unmixed features, the Macro-F1 drops by more than 1.44% on the two datasets. The above results further shows that feature smoothing before image-text interaction can better achieve multimodal fusion and improve classification performance.

In feature-level mixing, we set a hyperparameter to control the smoothing weight. Figure 3 shows the impact of different weight on the model performance of the twitter15 dataset. Setting the hyperparameter to 1 means that the two modalities do not perform feature smoothing, while setting to 0.8–0.95 means that we take one modality as the dominant information and incorporate a little information from another modality. It can be seen that when the hyperparameter is set to 0.8–0.95, the model can obtain better results than 1 or less than 0.75. This may be because, when feature smoothing is not performed, the semantic gap between modalities will make subsequent interactions insufficient. In addition, if we incorporate too much information from another modality, the dominant modal will lose its own representational ability. The best performance is achieved when the dominant modal feature introduces around 15% of the other modal feature.

#### 4.4.2 Tri-loss

In Table 5 we report the ablation study of the Tri-loss. It can be seen that the performance drops sharply after the removal of Tri-loss, which illustrates the effectiveness of reducing the semantic

distance between the two modalities via the constraint of Tri-loss. What’s more, if we use the initial aspect representation instead of the average representation of the two modals as the anchor in the Tri-loss, the performance decreases too. The reason may be that the model would learn from the lower-level aspect representation if using the initial aspect representation as the anchor after aspect-aware fusion, which is ineffective.

We adjusted the weight parameter  $\lambda$  of Tri-loss in the total loss to observe its effect. It can be seen from Figure 4 that the model achieves the best performance when  $\lambda$  is  $4e-3$ , while assigning too large or small weight leads to a decrease in the final performance. This illustrates that using appropriate constraints of Tri-loss can benefit the model.

#### 4.4.3 Multi-channel attention

We verified the effectiveness of the multi-channel attention-based interaction (MCA) by deleting it or replacing it with the Cross-Modal Transformer (CMT) [20]. As can be seen in Table 6, the performance decreases by 3.25% and 2.23% on the two datasets respectively after removing the module, which illustrates the necessity of performing deep image-text fusion. Furthermore, the performance decreases by 3.73% and 2.43% on the two datasets after replacing MCA with CMT, which fully illustrates the effectiveness of our proposed MCA module.

### 4.5 Case study

In this section, we choose two representative samples to compare the prediction results of our model with the two



baselines. Firstly, in Table 7, BERT predicted the sentiment polarity of the aspect {Charlie} incorrectly, which could be due to BERT only predicts based on text content and cannot recognize the negative sentiment expressed by the corresponding aspect in the image. In addition, the model Faster R-CNN-BERT, which also uses Faster R-CNN to capture image object-level features, made wrong predictions for the aspect Leicester in Table 7, while our model made correct predictions. It may be due to our excellent fusion network that enables our model to accurately capture the positive emotions expressed by waving flag in the image.

## 5 Conclusion

In this paper, we propose a MABSC model based on a multimodal feature smoothing fusion network. We extract aspect-aware representations of text and image modals at first. Then, we introduce a feature smoothing strategy to get smoothed representations, which are sent to the proposed multi-channel attention-based network for image-text information interaction. By this process, the comprehensive aspect-level sentiment representation is obtained for better classification. Experiments demonstrate that the model achieves better performance than the other baselines on the two datasets. The ablation experiments further demonstrate the effectiveness of the various modules of the model. In the future work, we will further consider how to align aspect-related information in image and text content, given that MABSC task requires to focus on fine-grained information in image and text.

## Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors upon request, without undue reservation.

## References

- Khan Z, Fu Y. Exploiting bert for multimodal target sentiment classification through input space translation. In: Proceedings of the 29th ACM International Conference on Multimedia; New York, NY, USA. New York, NY: Association for Computing Machinery (2021). p. 3034–42. MM '21. doi:10.1145/3474085.3475692
- Yu J, Chen K, Xia R, Wang Y, Feng K, Wan T, et al. Comprehensive comparisons of ocular biometry: A network-based big data analysis. *IEEE Trans Affective Comput* (2022) 10:1. doi:10.1186/s40662-022-00320-3
- Xu N, Mao W, Chen G. Multi-interactive memory network for aspect based multimodal sentiment analysis. In: Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence; July 2019. Palo Alto, CA: AAAI Press (2019). AAAI'19/IAAI'19/EAAI'19. doi:10.1609/aaai.v33i01.3301371
- Zhang Z, Wang Z, Li X, Liu N, Guo B, Yu Z. Modalnet: An aspect-level sentiment classification model by exploring multimodal data with fusion discriminant attentional network. *World Wide Web* (2021) 24:1957–74. doi:10.1007/s11280-021-00955-7
- Nasukawa T, Yi J. Sentiment analysis: Capturing favorability using natural language processing. In: Proceedings of the 2nd International Conference on Knowledge Capture; October 2003; New York, NY, USA. New York, NY: Association for Computing Machinery (2003). p. 70–7. K-CAP '03. doi:10.1145/945645.945658
- 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 (SemEval 2014); January 2014; Dublin, Ireland. Dublin, Ireland: Association for Computational Linguistics (2014). p. 437–42. doi:10.3115/v1/S14-2076
- Tang D, Qin B, Feng X, Liu T. Effective LSTMs for target-dependent sentiment classification. In: Proceedings of COLING 2016, the 26th International Conference on

## Author contributions

YX: Conceptualization, Methodology, Validation, Formal analysis, Writing—Review and Editing. YC: Software, Investigation, Writing—Original Draft. JG: Conceptualization, Methodology, Software, Writing—Review and Editing. All authors contributed to manuscript revision, read, and approved the submitted version.

## Funding

This work is supported by the National Natural Science Foundation of China (Grant Nos 62162037, 62266027, U21B2027, 62266028), General projects of basic research in Yunnan Province (Grant Nos 202001AT070047, 202001AT070046, 202301AT070444), Kunming University of Science and Technology “double first-class” joint project (Grant No. 202201BE070001-021).

## Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Computational Linguistics: Technical Papers; Osaka, Japan. The COLING 2016 Organizing Committee (2016). p. 3298–307.

8. Nguyen HT, Le Nguyen M. Effective attention networks for aspect-level sentiment classification. In: Proceedings of the 2018 10th International Conference on Knowledge and Systems Engineering (KSE); November 2018; Ho Chi Minh City, Vietnam. IEEE (2018). p. 25–30. doi:10.1109/KSE.2018.8573324

9. 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 IJCAI-17; August 2017. Palo Alto, CA: Association for Computing Machinery (2017). p. 4068–74. doi:10.24963/ijcai.2017/568

10. 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. Brussels, Belgium: Association for Computational Linguistics (2018). p. 3433–42. doi:10.18653/v1/D18-1380

11. Devlin J, Chang MW, 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, Volume 1 (Long and Short Papers); Minneapolis, Minnesota. Minneapolis, Minnesota: Association for Computational Linguistics (2019). p. 4171–86. doi:10.18653/v1/N19-1423

12. Sun C, Huang L, Qiu X. Utilizing BERT for aspect-based sentiment analysis via constructing auxiliary sentence. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers); June 2019; Minneapolis, Minnesota. Minneapolis, Minnesota: Association for Computational Linguistics (2019). p. 380–5. doi:10.18653/v1/N19-1035

13. Morency LP, Mihalcea R, Doshi P. Towards multimodal sentiment analysis: Harvesting opinions from the web. In: Proceedings of the 13th International Conference on Multimodal Interfaces; November 2011; New York, NY, USA. New York, NY: Association for Computing Machinery (2011). p. 169–76. ICMI '11. doi:10.1145/2070481.2070509
14. Wang Y, Shen Y, Liu Z, Liang PP, Zadeh A, Morency LP. Words can shift: Dynamically adjusting word representations using nonverbal behaviors. In: Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence; July 2019. Palo Alto, CA: AAAI Press (2019). AAAI'19/IAAI'19/EAAI'19. doi:10.1609/aaai.v33i01.33017216
15. Hazarika D, Zimmermann R, Poria S, Misa: Modality-invariant and -specific representations for multimodal sentiment analysis. In: Proceedings of the 28th ACM International Conference on Multimedia; New York, NY, USA. New York, NY: Association for Computing Machinery (2020). p. 1122–31. MM '20. doi:10.1145/3394171.3413678
16. Yu W, Xu H, Yuan Z, Wu J. Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. *Proc AAAI Conf Artif Intelligence* (2021) 35:10790–7. doi:10.1609/aaai.v35i12.17289
17. Zadeh A, Chen M, Poria S, Cambria E, Morency LP. Tensor fusion network for multimodal sentiment analysis. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing; September 2017; Copenhagen, Denmark. Copenhagen, Denmark: Association for Computational Linguistics (2017). p. 1103–14. doi:10.18653/v1/D17-1115
18. Liu Z, Shen Y, Lakshminarasimhan VB, Liang PP, Bagher Zadeh A, Morency LP. Efficient low-rank multimodal fusion with modality-specific factors. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); July 2018. Melbourne, Australia: Association for Computational Linguistics (2018). p. 2247–56. doi:10.18653/v1/P18-1209
19. Zadeh A, Liang PP, Mazumder N, Poria S, Cambria E, Morency LP. Memory fusion network for multi-view sequential learning. *Proc AAAI Conf Artif Intelligence* (2018) 32. doi:10.1609/aaai.v32i1.12021
20. Tsai YHH, Bai S, Liang PP, Kolter JZ, Morency LP, Salakhutdinov R. Multimodal transformer for unaligned multimodal language sequences. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics; July 2019; Florence, Italy. Palo Alto, CA: Association for Computational Linguistics (2019). p. 6558–69. doi:10.18653/v1/P19-1656
21. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. *Adv Neural Inf Process Syst* (2017) 30.
22. Wang Z, Wan Z, Wan X. Transmodality: An end2end fusion method with transformer for multimodal sentiment analysis. In: Proceedings of The Web Conference 2020; September 2020; New York, NY, USA. New York, NY: Association for Computing Machinery (2020). WWW '20, 2514–2520. doi:10.1145/3366423.3380000
23. Yu J, Jiang J. Adapting bert for target-oriented multimodal sentiment classification. In: Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19. San Mateo, CA: International Joint Conferences on Artificial Intelligence Organization (2019). p. 5408–14. doi:10.24963/ijcai.2019/751
24. Huang Y, Chen Z, Zhang W, Chen J, Pan JZ, Yao Z, et al. *Aspect-based sentiment classification with sequential cross-modal semantic graph* (2022). ArXiv abs/2208.09417.
25. Anderson P, He X, Buehler C, Teney D, Johnson M, Gould S, et al. Bottom-up and top-down attention for image captioning and visual question answering. In: Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE (2018). p. 6077–86. doi:10.1109/CVPR.2018.00636
26. Zhang Q, Fu J, Liu X, Huang X. Adaptive co-attention network for named entity recognition in tweets. In: Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence. Palo Alto, CA: AAAI Press (2018). AAAI'18/IAAI'18/EAAI'18.
27. Wang J, Wang Q, Wen Z, Liang X, Xu R. Interactive fusion network with recurrent attention for multimodal aspect-based sentiment analysis Artificial Intelligence: Second CAAI International Conference, CICA 2022, Beijing, China, August 27–28, 2022, Revised Selected Papers, Part III (Berlin, Heidelberg: Springer-Verlag) (2022), 298–309. doi:10.1007/978-3-031-20503-3\_24