World Wide Web (2021) 24:1957–1974
https://doi.org/10.1007/s11280-021-00955-7

# ModalNet: an aspect-level sentiment classification model by exploring multimodal data with fusion discriminant attentional network

Zhe Zhang[1] · Zhu Wang[1] · Xiaona Li[1] · Nannan Liu[1] · Bin Guo[1] · Zhiwen Yu[1]

Received: 19 January 2021 / Revised: 14 June 2021 / Accepted: 13 September 2021 /
Published online: 20 September 2021

**Abstract**
Aspect-level sentiment classification aims to identify sentiment polarity over each aspect of a sentence. In the past, such analysis tasks mainly relied on text data. Nowadays, due to the popularization of smart devices and Internet services, people are generating more abundant data, including text, image, video, et al. Multimodal data from the same post (e.g., a tweet) usually has certain correlation. For example, image data might has an auxiliary effect on the text data, and reasonable processing of such multimodal data can help obtain much richer information for sentiment analysis. To this end, we propose an aspect-level sentiment classification model by exploring multimodal data with fusion discriminant attentional network. Specifically, we first leverage two memory networks for mining the intra-modality information of text and image, and then design a discriminant matrix to supervise the fusion of inter-modality information. Experimental results demonstrate the effectiveness of the proposed model.

**Keywords** Multimodal data · Aspect-level sentiment classification · Discriminant attention network · Feature fusion

## 1 Introduction

Sentiment analysis is an important task in natural language processing, which has attracted much attention in recent years due to its usefulness in business intelligence, social media and public management. For example, companies usually need to know people's preferences for personalized recommendations.

Compared with traditional sentiment analysis [1–5], aspect level sentiment classification [6] is a more fine-grained task [7] that aims to determine the sentiment polarity of each aspect in a sentence. Traditional sentiment analysis approaches are not able to accurately classify the sentiment polarity of a sentence at the aspect level [8]. For example, given a

✉ Zhu Wang
wangzhu@nwpu.edu.cn

[1] Northwestern Polytechnical University, Xi'an, China

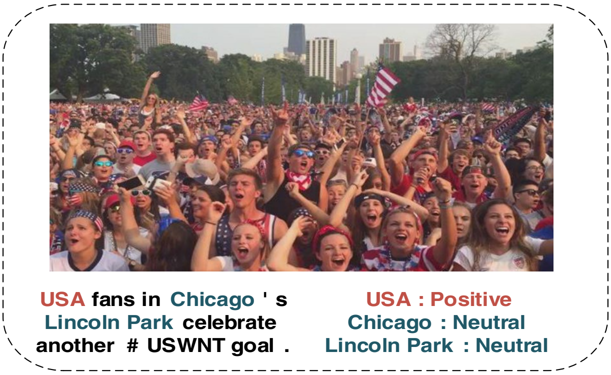

**Fig. 1** Example of aspect-level multimodal sentiment classification

sentence “the monitor is clear, but the keyboard is uncomfortable”, where the sentiment polarity of "monitor" and "keyboard" are positive and negative, respectively.

To solve the problem of aspect-level sentiment classification, various supervised learning techniques [9, 10] and attention based deep neural networks [11, 12] have been proposed to model the interaction among aspects and context. However, one major limitation of existing approaches is that these methods mainly rely on text data, while people are generating more abundant data, including text, image, video, et al. For example, online social media posts usually contain both text and image (e.g., 42% of tweets contain images[1]). Similarly, 40% of the user comments in online shopping sites contain both text and picture [13]. Moreover, compared with plain text, such multimodal contents (e.g., posts and comments) attract much more "replies" and "likes" (3 and 4 times). Meanwhile, multimodal contents in the same package usually are correlated with each other. For example, as shown in Fig. 1, the positive sentiment expressed in the image plays a distinct and auxiliary role, compared with that expressed in the text. Thereby, to facilitate aspect-level sentiment classification, it is necessary to explore multimodal contents (e.g., text and image) rather than only the text.

Specifically, for fine-grained aspect-level sentiment classification, it might be difficult to detect the expected sentiment if only a signal modality of data is explored. In other words, text and image can complement each other and enhance fine-grained sentiment analysis. Moreover, different images or different parts of the image may be related to different aspects, which have an enhanced indicative effect on sentiment analysis.

To explore multimodal contents for aspect-level sentiment classification, a key issue is how to accurately extract and fuse the complementary information from different modalities. To deal with this issue, Yu et al. [14] uses 4 BERT modules to capture intra-modality dynamics and inter-modality dynamics. It models the alignment of the target image through a stacked BERT architecture. However, while BERT may have advantages in the field of coarse-grained sentiment analysis, it has certain defects for fine-grained sentiment analysis. For example, the dependency information of adjacent words lost by BERT is fatal to fine-grained sentiment analysis. Thereby, besides the extraction and fusion of complementary

[1] https://buffer.com/resources/twitter-data-1-million-tweets/

multimodal information, another key issue is how to characterize and utilize the location information contained in the multimodal contents.

In this paper, to address the above issues, we achieve aspect-level multimodal sentiment classification by making the following contributions:

First, to fully explore multimodal contents especially the location information, we design a multimodal interactive network to model the textual and visual information of each aspect. Two memory network-based modules are used to capture intra-modality features, including text to aspect alignment and image to aspect alignment.

Second, a fusion discriminant matrix is designed to learn the interaction among 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.

Finally, extensive experiments have been conducted based on two Twitter datasets, which demonstrate that the proposed model outperforms the baseline method.

The rest of this paper is organized as follows. We present the related work in Sect. 2, followed by the detailed methodology in Sect. 3. Experimental results are described in Sect. 4, and the work is concluded in Sect. 5.

## 2 Related work

### 2.1 Aspect-level sentiment analysis

Aspect-level sentiment classification (ASC) is an important subtask in the field of sentiment analysis [15, 16]), which takes into account the factors that affect the sentiment of the sentence.

To enable aspect-level sentiment classification, classical approaches usually leverage a series of manually designed rules and combine them with external resources for feature engineering [8, 9, 17]), and then solve the problem as a text classification issue using statistical learning methods [10, 18]. While such approaches greatly improve the performance of classification, they are labor-intensive and require careful extraction of rich features.

In recent years, neural networks have been widely used for sentiment classification, which can extract rich semantic information from sentences by automatically encoding original data as feature vectors. Particularly, Dong et al. [19] first introduced recursive neural networks into the field of aspect-level sentiment classification, and classified sentiments based on syntactic relationships. Since then RNN starts to play a more and more important role in the ASC task.

Tang et al. [20] introduced LSTM for aspect-level sentiment classification, which models context words on the left and right side of the aspect word separately and connects the last hidden vector on both sides to capture the semantic information of context words more flexibly. However, most neural network models failed to consider the relationship between specific aspects and their context words.

With the successful application to tasks on NLP [21], attention mechanisms were also introduced into aspect-level sentiment classification, which enable the model to focus on important parts of context words relative to aspect words. For example, Wang et al. [22] proposed a single-hop attention-based LSTM model, which uses the hidden state of LSTM for attention calculation to capture the important part of the sentence for a given aspect. Ma et al. [23] developed an interactive network that uses attention-based LSTM and pooling operations to capture important parts of aspects and their context words, which interact with each other to influence the generation of two-part representations. Fan et al. [24] proposed a model that combines fine-grained and coarse-grained attention mechanisms, which explores interactions at the word and sentence levels separately to reduce information loss. Zeng et al. [25] developed an LSTM model based on location attentions, which uses position-aware vectors to represent locations of aspect words and context words.

Furthermore, to address the shortcoming of lacking of labeled data, Ma et al. [26] and He et al. [27] introduced knowledge from common sense and document-level sentiment datasets into the field of aspect-level sentiment classification through pre-training and multitask learning, which can improve the model's ability to filter irrelative information.

Different from the above methods, in this paper we explore multimodal contents for sentiment analysis, which increase the robustness of the model without using external knowledge.

### 2.2 Multimodal sentiment analysis

Nowadays, more and more people prefer to express their emotions and opinions in online social media through multimodal information. Multimodal data provides a new perspective and opportunity for sentiment analysis. Existing studies on multimodal sentiment analysis can be roughly divided into two categories.

The first line of research mainly focuses on building feature extraction models. For example, Wang et al. [28] used a unified cross-media bag of words model to represent textual and visual features, which proves that multimodal sentiment analysis outperforms single-modal sentiment analysis. Poria et al. [29] adopted a variety of ways to extract single-modal sentiment information and merge multimodal information through feature-level and decision-level fusion.

The other category of multimodal sentiment analysis studies is based on deep learning technology. Yu et al. [30] utilized CNN and DNN to extract textual features and visual features respectively, and then adopted logistic regression to fuse the features for sentiment analysis. Zadeh et al. [31] used a tensor fusion network (TFN) to calculate the correlation between different modalities, which uses LSTM to encode text and uses DNN to encode audio and video. In the phase of fusion, the outer product is performed on the output vectors of the three modalities. Similarly, to facilitate the fusion of multimodal information, Zadeh et al. [32] proposed memory fusion network, which uses delta-memory attention and multi-view gated memory to discover both cross-view and temporal interactions across different dimensions of memories. Xu et al. [13] developed a multi-interactive attention mechanism, which fuses textual and visual information through interactive attentions.

While the above studies mainly focus on coarse-grained (e.g., sentence-level) sentiment analysis, less work has been done to facilitate fine-grained (e.g., aspect-level) sentiment analysis using multimodal data. To this end, we develop a fusion discriminant attentional network to capture fine-grained interactions among different modalities, and then adopt multimodal learning to achieve efficient fusion of textual and visual information.

## 3 Methodology

In this section, we first give the formal definition of aspect-level multimodal sentiment classification task, followed by the overview of the proposed framework. Finally, we present the technical details of the proposed approach.

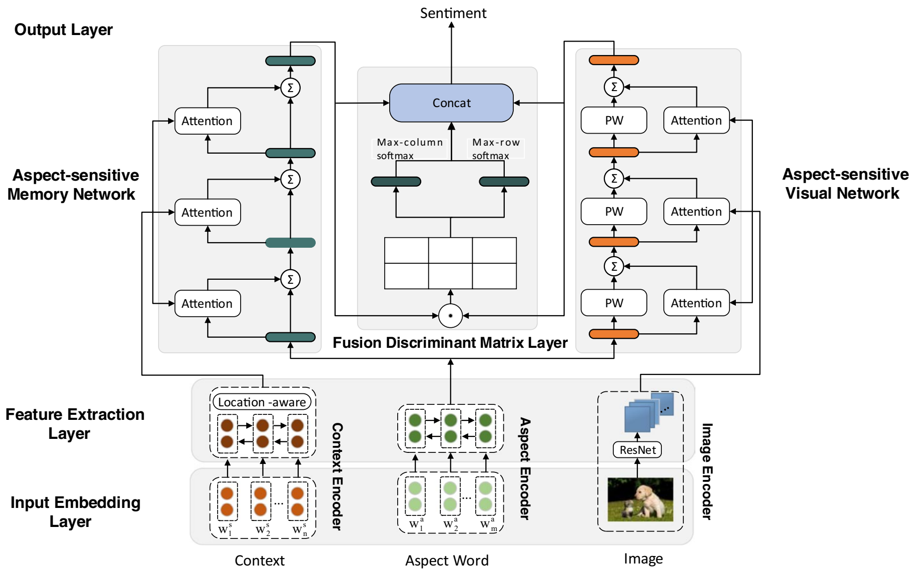

**Fig. 2** Overview of the proposed framework

## 3.1 Task definition

Given a set of multimodal dataset C, each sample $c \in C$ includes a context sentence $L = \left\{w_1^s, w_2^s, \ldots, w_n^s\right\}$ and an associated image I. Meanwhile, there is also an aspect sequence $A = \left\{w_1^a, w_2^a, \ldots, w_m^a\right\}$, where A is a sub-sequence of L. The task is to achieve aspect-level sentiment classification by exploring both text and image data, i.e., identify the sentiment label y for each aspect.

## 3.2 Overview of the proposed framework

In this section, we present the overall architecture of the proposed Multimodal fusion discriminant attentional Network (ModalNet), which consists of an input embedding layer, a feature extraction layer, an aspect-sensitive memory network, a fusion discriminant matrix layer, and an output layer, as shown in Fig. 2.

The input layer is responsible for mapping words into high dimensional vectors, while the feature extraction layer extracts eigenvectors of both text and image. The aspect-sensitive memory network and fusion discriminant matrix layer fuse multimodal information from the aspect level. Finally, the sentiment classification result is given by the output layer.

## 3.3 Input embedding layer

The input embedding layer includes word embedding and image embedding. The input embedding layer maps text data to a high dimensional vector space. The image embedding converts the image to the appropriate size.

### 3.3.1 Word embedding

Specifically, we mainly adopt two approaches for word embedding, which are Glove embedding and BERT embedding.

Glove [33] embedding uses a large pre-trained embedding matrix to obtain the fixed word embedding of each word. Let $\mathrm{M} \in \mathrm{R}^{\mathrm{d_v} \times |\mathrm{V}|}$ be the embedding matrix, where $\mathrm{d_v}$ is the dimension of word vectors and $|\mathrm{V}|$ is the vocabulary size. After embedding, each word corresponds to a row of the matrix.

BERT [34] is a language model based on bidirectional transformer proposed by Google. The pre-trained BERT can generate word vectors of the sequence, which severe as high-quality feature input for downstream processing. Specifically, we transform the sentence and aspect word to "[CLS] + sentence + [SEP]" and "[CLS] + aspect word + [SEP]" respectively to represent the entire input. Specifically, a vector is obtained for each token, and the length of the vector is 768.

### 3.3.2 Image embedding

The image size needs to be adapted to the neural network input requirements, so we need to preprocess the image. We first preprocess each image by filling it into a square. In order to find the image area most related to aspect words, we divide the image into k equal size areas. The division lays the foundation for obtaining the aspect-sensitive visual information. Afterwards, the image is scaled to a given size range, and then resized to 224 × 224 pixels.

## 3.4 Feature extraction layer

### 3.4.1 Contextual text feature extraction

To enable aspect-level sentiment classification, considering the strong dependence among words in a sentence, we employ a Bi-LSTM network on top of the embedding layer to capture the contextual semantic information.

Specifically, given a sentence $\mathrm{L} = \left\{\mathrm{w}_1^{\mathrm{s}}, \mathrm{w}_2^{\mathrm{s}}, \ldots, \mathrm{w}_{\mathrm{n}}^{\mathrm{s}}\right\}$, each word $\mathrm{w}_{\mathrm{i}}^{\mathrm{s}}$ is embedded to a vector $\mathrm{W}_{\mathrm{i}}^{\mathrm{s}} \in \mathrm{R}^{\mathrm{d_v}}$. At time step $t$, the LSTM unit converts the input word embedding $\mathrm{W}_{\mathrm{i}}^{\mathrm{s}}$ into the output hidden state $\mathrm{h}_{\mathrm{i}}^{\mathrm{s}} \in \mathrm{R}^{\mathrm{d_h}}$. The update process of the Bi-LSTM network can be formalized as follows:

$$W_i^s = embed\left(w_i^s\right), i \in [1, n] \tag{1}$$

$$\overrightarrow{h_i^s} = \overrightarrow{LSTM}\left(W_i^s\right), i \in [1, n] \tag{2}$$

$$\overleftarrow{h_i^s} = \overleftarrow{LSTM}\left(W_i^s\right), i \in [n, 1] \tag{3}$$

$$h_i^s = \left[\overrightarrow{h_i^s}, \overleftarrow{h_i^s}\right], i \in [1, n] \tag{4}$$

With the Bi-LSTM model, we can obtain concatenated output $\mathrm{h_i^s} \in \mathrm{R}^{2*\mathrm{d_h}}$. Given the embedding of a context sentence L, the sentence contextual output is $\mathrm{H^s} \in \mathrm{R}^{2*\mathrm{d_h}\times \mathrm{n}}$. The above operations are also used for the embedding of aspect words as follows:

$$W_i^a = embed\left(w_i^a\right), i \in [1, m] \tag{5}$$

$$\overrightarrow{h_i^a} = \overrightarrow{LSTM}\left(W_i^a\right), i \in [1, m] \tag{6}$$

$$\overleftarrow{h_i^a} = \overleftarrow{LSTM}\left(W_i^a\right), i \in [m, 1] \tag{7}$$

$$h_i^a = \left[\overrightarrow{h_i^a}, \overleftarrow{h_i^a}\right], i \in [1, m] \tag{8}$$

Similarly, we adopt the Bi-LSTM model to get the hidden state of the aspect word $\mathrm{h_i^a} \in \mathrm{R}^{2*\mathrm{d_h}}$, and get the aspect contextual output as $\mathrm{H^a} \in \mathrm{R}^{2*\mathrm{d_h}\times \mathrm{m}}$.

The importance of different words in a sentence is not equal, especially for aspect-level sentiment classification. Some words are more valuable for sentiment identification, such as context words that are closer to the aspect word. Therefore, we consider the location information of context words in the contextual feature extraction layer, and assign a weight to each word according to its importance in the sentence.

We use Normal distribution to model the relative position between the aspect word and its context words, and utilize location encoding to simulate the expression habit. The weight of the context word $\mathrm{w_i^s}$ is defined as follows:

$$dis = pos\left(w_i^s\right) - pos\left(w_i^a\right) \tag{9}$$

$$p\left(w_i^s\right) = \exp\left(\frac{-dis^2}{2\sigma^2}\right) \tag{10}$$

where *dis* means that there is a *dis* word-level distance between the context word and the aspect word, $\mathrm{p}\left(\mathrm{w_i^s}\right)$ represents the weight of the context word, and σ denotes the propagation scope. Then, the final context output is $H^s = \left[h_1^s \times p\left(w_1^s\right), h_i^s \times p\left(w_2^s\right), \dots, h_n^s \times p\left(w_n^s\right)\right]$ .

### 3.4.2 Visual feature extraction

We adopt the ResNet-50 model [35] to extract feature from the pre-processed images, and discard the top fully connected layer to obtain the output.

$$I = \left\{I_1, I_2, \dots, I_k\right\} \tag{11}$$

$$m_v = \mathrm{Re}sNet(I) \tag{12}$$

where $m_v \in R^{d_{img} \times k}$ is a visual feature vector extracted by the pre-trained model. We use linear transformation to project visual feature vectors to the textual feature space:

$$V = W_v m_v \tag{13}$$

where $W_v \in R^{2*d_h \times d_{img}}$ is the learnable parameter, and $V = \{v_1, v_2, \ldots, v_k\} \in R^{2*d_h \times k}$ is the visual feature information mapped to the text features space, which contains all the visual information and serves as the basis for interactions between the modalities.

## 3.5 Aspect-sensitive memory network

We did not focus on the aspect extraction. More specifically, we conduct experiments based on the labeled aspects already marked in the existing datasets.

### 3.5.1 Aspect-sensitive textual features

To obtain context sensitive information for the aspect words and eliminate possible noises, we adopt a memory network to adaptively identify the interaction between aspect words and context words.

Currently, BERT is used to model the interaction between aspect words and context words [14]. The core of BERT is deep bidirectional transformers, and its training process usually takes a long time for convergence and may cause the loss of inter-character dependency information due to the huge number of masks. Another shortcoming of BERT is that, even though a position encoding mechanism is added, it still cannot fully capture the position relationship between words in a sequence, which is very important for aspect-level sentiment classification. Moreover, for a sentence with multiple aspects, the bidirectional encoder may blur the sentiment of a given aspect. Meanwhile, for short text such as tweets, the importance of long-term dependence is very low.

Therefore, we adopt sequence modeling to obtain aspect-sensitive textual features. If an aspect term consists of one word, the word vector of that aspect is the expression of the aspect term. If an aspect term is composed of multiple words, we take the average value of the vector of multiple words to get the expression $h^a$.

Specifically, in each layer, we use the attention mechanism to supervise the generation of textual vector with aspect information. Taking the embeddings of context words and aspect words H and A extracted in Sect. 3.4 as input, the model outputs vector $ct_i \in R^{2*d_h \times n}$ by calculating the weighted sum of each unit as follows:

$$ct_i = \sum_{j=1}^{n} \alpha_i h_i^s \tag{14}$$

where $n$ is the number of context feature vectors, and $\alpha_{ij} \in [0, 1]$ is the weight of $h_j$. For each $h_i$, a feedforward neural network is used to calculate its correlation to the corresponding aspect as:

$$g_i = \tanh\left(W_{att}\left[h_i^s; h^a\right] + b_{att}\right) \tag{15}$$

where $W_{att} \in R^{n \times 4*d_h}$ and $b_{att} \in R^{1 \times 1}$ are learnable parameters. Then, the attention weight is normalized to obtain the final importance score:

$$\alpha_i = \frac{\exp(g_i)}{\sum_i g_i} \tag{16}$$

Each computing layer takes a context feature vector information $h_j$ as well as the output of the previous layer as inputs, and the last memory vector is $m^s \in R^{2*d_h \times n}$.

### 3.5.2 Aspect-sensitive visual features

To identify the interaction between images and aspect words, we design another memory network. Specifically, different from the calculation of aspect-sensitive textual features, we first conduct point-wise convolution to transform the aspect feature vector from the previous layer as follows:

$$pw\left(a_i\right) = conv(\sigma(conv(h^a, I), I)) \tag{17}$$

where conv() represents the convolution operation, and $pw\left(a_i\right) \in R^{2*d_h}$ is the obtained feature information after transformation. Point-wise convolution plays the role of intra-group information dissemination, which enables layers to integrate information in depth.

Afterwards, we adopt the attention mechanism to calculate interactions between the visual feature vector and the aspect feature vector as follows:

$$cv_i = \sum_{i=1}^{n} \alpha_i v_i \tag{18}$$

where $cv_i \in R^{2*d_h \times k}$.

Finally, output of the memory network is a vector $m^v \in R^{2*d_h \times k}$.

## 3.6 Fusion discriminant matrix layer

As we known, multimodal contents usually contain complementary information, and better sentiment classification performance would be achieved if such complementary information can be fused efficiently. In the previous sections, we have designed two memory networks to extract aspect-sensitive features of different modalities. In this section, we will explore how to fuse the obtained features.

A reasonable assumption is that distributions of modality invariant features should be of high similarity. To quantify the similarity between two different modalities, we introduce a fusion discriminant matrix as follows:

$$D_{ij} = W_{fd}([m^{s_i};m^{v_j};m^{s_i} * m^{v_j}]) \tag{19}$$

where $D_{ij} \in R^{n \times k}$ denotes the similarity between context word feature vector $m^{s_i}$ and visual feature vector $m^{v_j}$, $W_{fd} \in R^{1 \times 6*d_h}$ is a learnable parameter matrix, and $*$ represents element-wise multiplication. Then we use the fusion discriminant matrix to calculate the attention vectors between modalities.

On one hand, to measure the textual information most relevant to visual features, and let $\beta_i^{vh}$ be the attention weight of sentence contextual feature vector $h_i$, the definition is as follows:

$$\gamma_i^{sv} = \max\left(D_i, :\right) \tag{20}$$

Using $\gamma_i^{vh} \in R^n$ to obtain the maximum similarity in the entire row, and uses the following method to obtain the sentence contextual attention vector $m^{vh} \in R^{2*d_h}$:

$$\beta_i^{sv} = \frac{\exp\left(\gamma_z^{sv}\right)}{\sum_{z=1}^{n} \exp\left(\gamma_z^{sv}\right)} \tag{21}$$

$$m^{sv} = \sum_{j=1}^{n} \beta_i^{sv} \times h_j \tag{22}$$

On the other hand, we use the similarity matrix to further measure the visual area most related to the sentence, and obtain the visual attention vector $m^{vs} \in R^{2*d_h}$.

In such a way, we can quantify the interaction among different modalities.

### 3.7 Output layer

By concatenating the obtained textual and visual feature vectors, we get the final feature $m = \left[m^h;m^v;\, m^{vh};m^{hv}\right]$ which is fed to a full connected layer for sentiment classification. Specifically, the final feature is as follows:

$$m = [m^s;m^v;m^{sv};m^{vs}] \tag{23}$$

$$y = soft\max\left(W_p * m + b_p\right) \tag{24}$$

where $y \in R^C$ is the probability distribution for polarity of aspect-level sentiment, $W_p \in R^{1\times C}$ and $b_p \in R^C$ denote the learnable weight matrix and bias.

### 3.8 Model training

Considering that the identification of neutral sentiment is a challenge issue due to its ambiguity, we introduce a label smoothing regular (LSR) term to the loss function [36], aiming to improve the model's generalization ability.

As there might be training samples that are wrongly labelled, LSR chooses to replace the one hot label with a smooth value, which gives the label a certain degree of fault tolerance and avoids excessive trust to training samples.

Assuming that y is the ground-truth label distribution of a sample, the label smoothing formula is defined as:

$$y' = (1 - \varepsilon) \times y + \varepsilon \times u \tag{25}$$

where $y'$ is the label after smoothing, $\varepsilon$ is the smoothing factor, and $u$ denotes the label's prior distribution. Based on empirical results, we set $\varepsilon$ as 0.2 and u as $1/C$.

We adopt the Adam optimization algorithm [37] to train the model by minimizing the cross-entropy loss, which is defined as:

$$H\left(y',p\right) = -\sum_{i=1}^{C} y_i' \log\left(p_i\right) = (1-\varepsilon)H(y,p) + \varepsilon H(u,p) \tag{26}$$

where p is the predicted probability, and H(u, p) is defined from the perspective of relative entropy as:

**Table 1** Statistics of the multimodal Twitter datasets

| | Positive | Neutral | Negative | Total | Avg. targets | Avg. lengths | Max lengths |
|---|---|---|---|---|---|---|---|
| Twitter-15 | | | | | | | |
| Train | 928 | 1883 | 368 | 3179 | 1.348 | 16.06 | 35 |
| Dev | 303 | 670 | 149 | 1122 | 1.336 | 16.06 | 40 |
| Test | 317 | 607 | 113 | 1037 | 1.354 | 16.36 | 36 |
| Twitter-17 | | | | | | | |
| Train | 1508 | 1638 | 416 | 3562 | 1.410 | 15.56 | 39 |
| Dev | 515 | 517 | 144 | 1176 | 1.439 | 15.75 | 31 |
| Test | 493 | 573 | 168 | 1234 | 1.450 | 15.79 | 38 |

**Table 2** Statistics of the multi-ZOL dataset

| Attribute | Statistic |
|---|---|
| #Review | 5228 |
| #Label | 10 |
| #Aspect-review pair | 28,469 |
| Avg. of #aspect/review | 5.45 |
| Avg. text length/review | 315.11 |
| Max text length/review | 8511 |
| Min text length/review | 5 |
| Avg. of #image/review | 4.5 |
| Max of #image/review | 111 |
| Min of #image/review | 1 |

$$\mathcal{L}_{lsr} = H(u,p) = -D_{KL}(u||p) \tag{27}$$

where u is a uniform distribution. The final objective function to be optimized is the cross-entropy loss function with LSR and L2 regularization, which is defined as:

$$H(y',p) = -\sum_{i=1}^{C} y_i \log\left(p_i\right) + \mathcal{L}_{lsr} + \lambda \sum_{\theta\in\Theta} \theta^2 \tag{28}$$

where p is the estimated probability given by the output layer, and λ is the coefficient for L2 regularization. Meanwhile, to avoid overfitting, we also introduce a dropout layer to the model [38].

# 4 Experiments

## 4.1 Dataset and experiment setup

To evaluate the performance of the proposed model, we used two public multimodal English datasets TWITTER-15 and TWITTER-17, and one Chinese dataset Multi-ZOL. TWITTER-15 and TWITTER-17 include tweets published within 2014–2015 and 2016–2017, respectively. The datasets provide aspects of each tweet together with three kinds of sentiment polarity labels. Statistics of the datasets are summarized in Table 1.

Multi-ZOL contains 5,288 multimodal reviews, each of which contains a textual content, an image set, and at least one but no more than six aspects. Statistics of this dataset are summarized in Table 2.

In the learning process, if we use BERT for word embedding, the pre-training parameters will be fine-tuned. Specifically, the embedding dimension $d_v$ is set to 300, and the embedding dimension of BERT pre-training word vector is 768.

For text, we set the maximum fill length to 40. For image, we set K to 9 which means that each image is cut into 9 equal sized regions.

The dimension of hidden layer state $d_h$ is set to 300, and the dropout rate is set to 0.1. The weights of the model were initialized by a uniform distribution [39].

Finally, we use accuracy and Macro-F1 to evaluate the performance of the model.

### 4.2 Baselines

To evaluate the performance of the proposed model, we chose a set of state-of-the-art models as baselines and also designed several ablations of the ModalNet.

(1) Res-Aspect-ATT, which concatenates the extracted visual feature and the embedding of aspect words with an attention layer.
(2) Res-Aspect-TFN, which fuses the extracted visual feature and the embedding of aspect words using a tensor fusion network [31].
(3) TD-LSTM [20], which uses two LSTM models to learn representations from the left and right contexts of the aspect. Only the last hidden vector is connected, which depends on the aspect.
(4) MemNet [40], a memory model that uses aspects as queries. Based on word embedding and position embedding, a multi-hop attention mechanism is applied to update the stored memory to achieve deep storage.
(5) IAN [23], which is based on interactive attention, where the aspect and the whole context are considered interactively. Two attention-based LSTMs are used to capture the important words of aspect terms and their context interactively.
(6) RAM [41], which adopts a dynamic attention structure based on Bi-LSTM to obtain the representation based on the memory model. GRU is used to construct a nonlinear neural system and enhance the expression ability of the global memory.
(7) MGAN [24], which construct an attention network from both coarse-grained and fine-grained perspectives to extract textual features.
(8) BERTABSA-ATT[42], which uses a progressive self-supervised attention learning approach to automatically and incrementally mine attention supervision information.
(9) Multimodal-TFN, which fuses the extracted visual feature, the embedding of aspect words and the embedding of context words using a tensor fusion network.
(10) Res-IAN, which uses ResNet as the image encoder and IAN as the text encoder. The extracted textual features and visual features are stitched together for classification.
(11) Res-IAN-TFN, which uses TFN to fuse the extracted visual features and textual features.
(12) MIMN[13], which uses two interactive memory networks to supervise and fuse the textual and visual information with the given aspect.
(13) TomBERT, which consists of three BERT models. The first one is used to capture the interaction between the aspect and the image, the second one is designed to extract textual features, and the third one is for the fusion of multimodal information.

**Table 3** Performance of ModalNet and baselines

| Modality | Models | TWITTER-15 | | TWITTER-17 | |
|---|---|---|---|---|---|
| | | Acc | Mac-F1 | Acc | Mac-F1 |
| Visual | Res-Aspect-ATT | 0.5972 | 0.4730 | 0.5861 | 0.5303 |
| | Res-Aspect-TFN | 0.5758 | 0.4432 | 0.5894 | 0.5426 |
| Text | TD-LSTM | 0.6830 | 0.6143 | 0.6067 | 0.5697 |
| | MemNet | 0.7011 | 0.6176 | 0.6418 | 0.6090 |
| | IAN | 0.7090 | 0.6332 | 0.6461 | 0.6120 |
| | RAM | 0.7068 | 0.6305 | 0.6442 | 0.6101 |
| | MGAN | 0.7117 | 0.6421 | 0.6475 | 0.6146 |
| | BERTABSA-ATT | 0.7566 | 0.6739 | 0.6805 | 0.6647 |
| Text + Visual | Multimodal-TFN | 0.6786 | 0.6094 | 0.6179 | 0.5702 |
| | Res-IAN | 0.7185 | 0.6384 | 0.6653 | 0.6335 |
| | Res-IAN-TFN | 0.7102 | 0.6397 | 0.6568 | 0.6283 |
| | MIMN | 0.7353 | 0.6649 | 0.6722 | 0.6385 |
| | TomBert | 0.7715 | 0.7175 | 0.7034 | 0.6803 |
| | ModalNet w/o FD | 0.7283 | 0.6512 | 0.6734 | 0.6499 |
| | ModalNet w/o LA | 0.7331 | 0.6753 | 0.6788 | 0.6533 |
| | ModalNet w/o LSR | 0.7310 | 0.6642 | 0.6803 | 0.6692 |
| | ModalNet | 0.7515 | 0.6995 | 0.6951 | 0.6536 |
| | ModalNet-Bert | **0.7903** | **0.7250** | **0.7236** | **0.6919** |

(14) ModalNet ablations:
(15) ModalNet w/o FDM, which ablates the fusion discriminant module.
(16) ModalNet w/o LA, which ablates the position-aware module.
(17) ModalNet w/o LSR, which ablates the label smoothing regularization module.

## 4.3 Experimental results

### 4.3.1 Performance of different models

The performance of the proposed ModalNet as well as the baseline models are demonstrated in Table 3. Accordingly, we can find that ModalNet-Bert achieves the best performance, which adopts BERT to pre-train word vectors and fully captures the information within the modalities and the interactive information between different modalities. We summarize the observations as follows.

The performance of Res-Aspect-ATT is limited, and an accuracy of about 60% indicates the importance of textual information. Res-IAN outperforms IAN which uses only text data, and is superior to all the models that have not explored multimodal data. Therefore, multimodal information is useful for fine-grained sentiment classification. The addition of a TFN module (e.g., Res-IAN to Res-IAN-TFN) leads to the model's performance reduction on both datasets, indicating that TFN-based fusion is not suitable for such kind of multimodal fine-grained sentiment classification.

Among all the models that only use text data, the performance of TD-LSTM is the worst. The reason is that it has not effectively process the context information of aspect

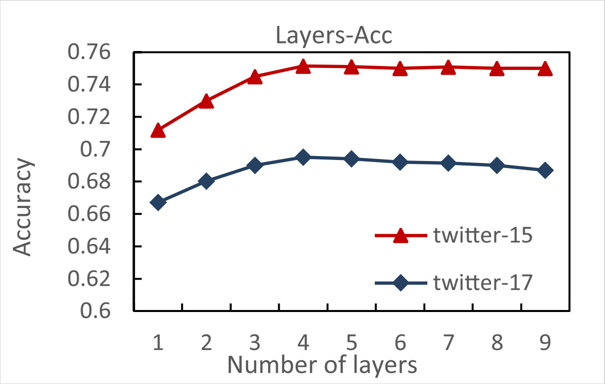

**Fig. 3** Performance of different computing layers

words, and important sentiment related words might be ignored. On the contrary, BERT-ABSA-ATT performs the best, which is based on BERT and has leveraged attention mechanisms to avoid information loss.

By fusing text and image modalities for sentiment classification, ModalNet outperforms all the models that only use text data. Meanwhile, the overall performance of ModalNet-Bert is also better than TomBert, indicating the effectiveness of customizing the downstream network for fine-grained sentiment classification. In other words, simply using BERT to extract the interaction of different modalities for fine-grained sentiment classification is not the best choice.

Finally, ModalNet outperforms all of its ablations, which proves that all the modules are indispensable to the model. Specifically, by comparing the performance of different ablations, we find that the most important module is the fusion discriminant matrix (i.e., the performance of ModalNet w/o FD is the worst), indicating that the proposed model can extract complementary information from data of different modalities.

### 4.3.2 Effects of multiple computing layers

In ModalNet, we extract the internal characteristics of each modality using a memory network with multiple computing layers, where the number of layers is an important hyperparameter that affects the performance. We validate the model by setting the number of layers to the interval of [13, 19], and the result is shown in Fig. 3.

Generally, the model's performance improves with the increase of the number of computing layers, especially when the number of layers is relatively small (i.e., smaller than 4). Meanwhile, as the number of layers continues to increase, the performance tends to be stable. Thereby, to balance the model's performance and efficiency, we set the number of layers to 4.

In order to better understand the advantages of ModalNet, we further calculate the accuracy of the sentences with single aspect and those with multiple aspects. The results are shown in Table 4, where "Aspect = 1" means that there is only one aspect word in a sentence, and "aspect ≥ 2" means that there are two or more aspect words in a sentence.

**Table 4** Classification accuracy of sentences with different number of words

| Models | Twitter-15 | | Twitter-17 | |
|---|---|---|---|---|
| | Aspect = 1 | Aspect ≥ 2 | Aspect = 1 | Aspect ≥ 2 |
| ModalNet-GloVe | 75.54 | 74.68 | 70.68 | 68.47 |
| ModalNet-Bert | 79.64 | 78.30 | 71.96 | 74.61 |

**Table 5** Performance of MIMN and ModalNet

| Models | Acc | Mac-F1 |
|---|---|---|
| MIMN | 0.6159 | 0.6051 |
| ModalNet | 0.6271 | 0.6094 |

**Table 6** Complexity of different models

| Models | Params $\times 10^6$ |
|---|---|
| Res-Aspect-ATT | 2.67 |
| TD-LSTM | 1.45 |
| MemNet | 0.36 |
| IAN | 2.17 |
| RAM | 2.25 |
| MGAN | 3.62 |
| TomBert | 369.26 |
| ModalNet | 8.09 |
| ModalNet-Bert | 117.96 |

When there are multiple aspect words in a sentence, the proposed model can still produce satisfactory performance. For the Twitter-17 dataset, when there are two or more aspect words in a sentence, the accuracy of using ModalNet-Bert is higher than that of using only one aspect word in a sentence. Such a result proved that ModalNet is suitable for aspect-level sentiment classification, which is consistent with the motivation of this paper.

We also compare our model with MIMN[13] using the Multi-ZOL dataset, where each multimodal review contains at least one but no more than six aspects. We randomly divide the Multi-ZOL dataset into training set (80%), development set (10%) and test set (10%). The experimental results are shown in the Table 5. Accordingly, we can find that the proposed model slightly outperforms MIMN.

### 4.3.3 Complexity analysis

The computational complexity is another key issue of machine learning models. Thereby, we analyze the complexity of ModalNet and all the baseline models, and present the statistical results in Table 6.

Accordingly, we can see that although ModalNet requires more parameters than all the single-modality models, it has much fewer parameters than TomBert. Meanwhile, the size of ModalNet-Bert is also smaller than TomBert, even though word vectors are embedded in the same way. BERT uses deep bidirectional transformers, and the model size is

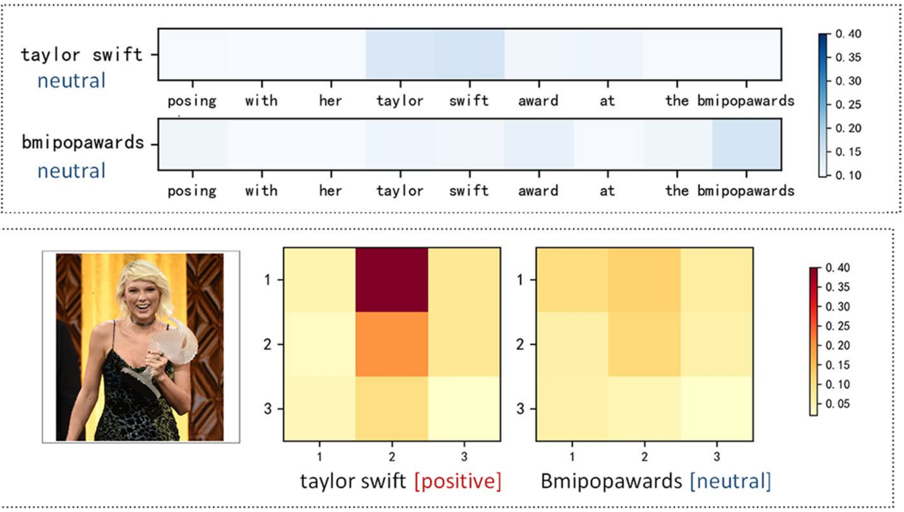

**Fig. 4** Examples of attention weights for a sentence and an image

larger and the complexity is higher. Using LSTM to model the hidden states of sequences requires fewer parameters. Moreover, the Memory Network used by our model does not calculate the hidden state of the embedding vector, so the model size is smaller.

### 4.3.4 Case study

In order to better reflect the interaction between text and image, we select a sentence which contains two aspect words for visual analysis, as shown in Fig. 4. The color depth indicates the importance of a word or an image area, the daker the more important.

It can be observed that while there is no definite emotional words in the sentence "posting with her Taylor Swift award at the bmippawards", the correlated image provides useful information for sentiment analysis. According to the results of attention weights visualization, larger weights are assigned to the smiling face and the whole person, which indicates the advantage of joint sentiment classification.

## 5 Conclusion

In this paper, we studied the problem of aspect-level sentiment classification by exploring multimodal data, i.e., text and images. In particular, we proposed an aspect-level sentiment classification model, which adopts a pair of memory networks to effectively capture intra-modality information and a fusion discriminant attentional network to extract interactive information between different modalities. Experiments on two multimodal datasets validated the effectiveness of the proposed model in the field of aspect-level sentiment classification.

**Acknowledgements** This work is partially supported by the National Natural Science Foundation of China (No. 61960206008, 62072375), and the Fundamental Research Funds for the Central Universities (No. 3102019AX10).

## References

1. Hsu, W.Y., Hsu, H.H., Tseng, V.S.: Discovering negative comments by sentiment analysis on web forum. World Wide Web **22**, 1297–1311 (2019)
2. Chauhan, U.A., Afzal, M.T., Shahid, A., Abdar, M., Basiri, M.E., Zhou, X.: A comprehensive analysis of adverb types for mining user sentiments on amazon product reviews. World Wide Web **23**, 1811–1829 (2020)
3. Ouyang, Y., Guo, B., Zhang, J., Yu, Z., Zhou, X.: SentiStory: multi-grained sentiment analysis and event summarization with crowdsourced social media data. Pers. Ubiquit. Comput. **21**(1), 97–111 (2017)
4. Yu, Z., Wang, Z., Chen, L., Guo, B., Li, W.: Featuring, detecting, and visualizing human sentiment in chinese micro-blog. ACM Trans. Knowl. Discov. Data **10**(4), 1–23 (2016)
5. D. Yang, D. Zhang, Z. Yu, and Z. Wang. A sentiment-enhanced personalized location recommendation system. Proceedings of the 24th ACM conference on hypertext and social media, 119–128, 2013.
6. M. Pontiki, D. Galanis, J. Pavlopoulos, H. Papageorgiou, I. Androutsopoulos, and S. Manandhar, Semeval-2014 task 4: Aspect based sentiment analysis. In: Proc. 8th Int. Workshop Semantic Eval. (SemEval), 2014, pp. 27–35.
7. Lai, Y., Zhang, L., Han, D., Wang, G.: Fine-grained emotion classification of Chinese microblogs based on graph convolution networks. World Wide Web **23**, 2771–2787 (2020)
8. D-T Vo and Y Zhang (2015) Target-dependent twitter sentiment classification with rich automatic features. In: IJCAI. pp. 1347–1353.
9. Jiang, L., Yu, M., Zhou, M., Liu, X., Zhao, T.: Target-dependent twitter sentiment classification. ACL **1**, 151–160 (2011)
10. SM Mohammad, S Kiritchenko, and X Zhu. Nrc-canada: building the state-of-the-art in sentiment analysis of tweets. arXiv preprint https://arxiv.org/abs/1308.6242.
11. T. Luong, H. Pham, and C. D. Manning: Effective approaches to attention-based neural machine translation. In Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP). Lisbon, Portugal, Sep. 2015, pp. 1412–1421.
12. Feng, S., Wang, Y., Liu, L., Wang, D., Yu, G.: Attention based hierarchical LSTM network for context-aware microblog sentiment classification. World Wide Web **22**, 59–81 (2019)
13. N. Xu, W. Mao, and G. Chen. Multi-interactive memory network for aspect based multimodal sentiment analysis. In: Proceedings of the AAAI Conference on Artificial Intelligence. 33, 01 (2019), 371-378
14. J. Yu and J. Jiang. Adapting BERT for target-oriented multimodal sentiment classification. In: Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence Main track. pp. 5408–5414.
15. Pang, Bo., Lee, L.: Opinion mining and sentiment analysis. Found. Trends R Inf. Retr. **2**(1–2), 1–135 (2008)
16. Liu, B.: Sentiment analysis and opinion mining. Synth. Lect. Hum. Lang. Technol. **5**(1), 1–167 (2012)
17. V. Perez-Rosas, C. Banea, and R. Mihalcea. Learning sentiment lexicons in spanish. In: LREC. pp. 3077–3081, 2012.
18. S. Kiritchenko, X. Zhu, C. Cherry, and S. Mohammad. Nrc-canada-2014: detecting aspects and sentiment in customer reviews. In: Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014). pp. 437–442.
19. L. Dong, F. Wei, C. Tan, D. Tang, M. Zhou, and K. Xu. Adaptive recursive neural network for target dependent twitter sentiment classification. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. pp. 49–54.
20. D. Tang, B. Qin, X. Feng, and T. Liu: Effective LSTMs for targetdependent sentiment classification. In: Proc. COLING 26th Int. Conf. Comput. Linguistics, Tech. Papers, 2016, pp. 3298–3307.
21. Y. Kim, C. Denton, L. Hoang, and A. M. Rush, ``Structured attention networks,'' Feb. 2017, Available https://arxiv.org/abs/1702.00887.
22. Y. Wang, M. Huang, and L. Zhao: Attention-based lstm for aspectlevel sentiment classification. In: Proc. Conf. Empirical Methods Natural Lang. Process., 2016, pp. 606–615.
23. D. Ma, S. Li, X. Zhang, and H. Wang: Interactive attention networks for aspect-level sentiment classification. In: Proc. IJCAI, 2017. pp. 4068–4074.
24. F. Fan, Y. Feng, and D. Zhao: Multi-grained attention network for aspect-level sentiment classification. In: Proc. Conf. Empirical Methods Natural Lang. Process., 2018, pp. 3433–3442.
25. Zeng, J., Ma, X., Zhou, K.: ‘Enhancing attention-based LSTM with position context for aspect-level sentiment classification.’ IEEE Access **7**, 20462–20471 (2019)

26. Y. Ma, H. Peng, and E. Cambria: Targeted aspect-based sentiment analysis via embedding commonsense knowledge into an attentive LSTM. In: Proc. AAAI, 2018, pp. 5876–5883.
27. R. He, W. S. Lee, H. T. Ng, and D. Dahlmeier. Exploiting document knowledge for aspect-level sentiment classification. In: Proc. 56th Annu. Meeting Assoc. Comput.Linguistics (ACL), Melbourne, VIC, Australia, vol. 2, Jul. 2018, pp. 579–585.
28. Min Wang, Donglin Cao, Lingxiao Li, Shaozi Li, and Rongrong Ji. Microblog sentiment analysis based on cross-media bag-of-words model. In: Proceedings of International Conference on Internet Multimedia Computing and Service (ICIMCS'14). Association for Computing Machinery, New York, NY, USA, 76–80.
29. Poria, S., Cambria, E., Howard, N., Huang, G.-B., Hussain, A.: Fusion audio, visual and teatual clues for sentiment analysis from multimodal content. Neurocomputing **2016**(174), 5059 (2016). https://doi.org/10.1016/j.neucom.2015.01.095
30. Yu, Y., Lin, H., Meng, J., Zhao, Z.: Visual and textual sentiment analysis of a microblog using deep convolutional neural networks. Algorithms **9**, 41 (2016)
31. Zadeh A, Chen Minghai, Poria S, E. Cambria, and L.P. Morency. Tensor fusion network for multimodal sentiment analysis. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP), Copenhagen, Denmark, 2017.
32. Zadeh A, Liang P, Mazumder N, Poria S, Cambria E, and Morency P. Memory fusion network for multi-view sequential learning. In: Proceedings of the AAAI Conference on Artificial Intelligence. 2018.
33. J. Pennington, R. Socher, and C. Manning (2014) Glove: global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). pp. 1532–1543.
34. J. Devlin, M. Chang, K. Lee, and K. Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. https://arxiv.org/abs/1810.04805, 2018.
35. K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In: CVPR. pp. 770–778, 2016.
36. C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna (2016) Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2818–2826
37. Kingma, D. P., and Ba, J. (2014) Adam: a method for stochastic optimization. arXiv preprint https://arxiv.org/abs/1412.6980.
38. Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. R. (2012) Improving neural networks by preventing co-adaptation of feature detectors. arXiv preprint https://arxiv.org/abs/1207.0580.
39. X. Glorot and Y. Bengio (2010) Understanding the difficulty of training deep feedforward neural networks. In: Proceedings of the thirteenth international conference on artificial intelligence and statistics, pages 249–256.
40. D. Tang, B. Qin, and T. Liu: Aspect level sentiment classification with deep memory network. In: Proc. Conf. Empirical Methods Natural Lang. Process., 2016, pp. 214–224.
41. P. Chen, Z. Sun, L. Bing, and W. Yang: Recurrent attention network on memory for aspect sentiment analysis. In: Proc. Conf. Empirical Methods Natural Lang. Process., 2017, pp. 452–461.
42. Su, J., Tang, J., Jiang, H., Lu, Z., Ge, Y., Song, L., Xiong, D., Sun, L., Luo, J.: Enhanced aspect-based sentiment analysis models with progressive self-supervised attention learning. Artif. Intell. **296**, 103477 (2021)