

# AoM: Detecting Aspect-oriented Information for Multimodal Aspect-Based Sentiment Analysis

Ru Zhou<sup>1</sup> Wenyu Guo<sup>1\*</sup> Xumeng Liu<sup>1</sup> Shenglong Yu<sup>1</sup>  
Ying Zhang<sup>1</sup> Xiaojie Yuan<sup>1</sup>

<sup>1</sup> College of Computer Science, TKLNDST, Nankai University, Tianjin, China  
{zhouru, guowenya, liuxumeng, yushenglong, zhangying}@dbis.nankai.edu.cn  
yuanxj@nankai.edu.cn

## Abstract


Multimodal aspect-based sentiment analysis (MABSA) aims to extract aspects from text-image pairs and recognize their sentiments. Existing methods make great efforts to align the whole image to corresponding aspects. However, different regions of the image may relate to different aspects in the same sentence, and coarsely establishing image-aspect alignment will introduce noise to aspect-based sentiment analysis (*i.e.*, visual noise). Besides, the sentiment of a specific aspect can also be interfered by descriptions of other aspects (*i.e.*, textual noise). Considering the aforementioned noises, this paper proposes an Aspect-oriented Method (AoM) to detect aspect-relevant semantic and sentiment information. Specifically, an aspect-aware attention module is designed to simultaneously select textual tokens and image blocks that are semantically related to the aspects. To accurately aggregate sentiment information, we explicitly introduce sentiment embedding into AoM, and use a graph convolutional network to model the vision-text and text-text interaction. Extensive experiments demonstrate the superiority of AoM to existing methods. The source code is publicly released at <https://github.com/SilyRab/AoM>.

## 1 Introduction

As an important and promising task in the field of sentiment analysis, Multimodal Aspect-Based Sentiment Analysis (MABSA) has attracted increasing attention (Lv et al., 2021; Ju et al., 2021). Given an image and corresponding text, MABSA is defined as jointly extracting all aspect terms from image-text pairs and predicting their sentiment polarities (Ju et al., 2021).

In this scenario of fine-grained sentiment recognition for multimodal information, the input image-text pairs are always complex. (1) The semantics of sentence is complex which adds sentiment confusion among different aspects. Take Figure 1 as an

\*Corresponding author.



Aspect	Kyoto	mayor	Mayor Kadokawa
Sentiment	Negative	Neutral	Positive

Figure 1: An example of MABSA task, including the aspects, their corresponding descriptions, and sentiments.

example, there are 3 aspects in the sentence with 3 different sentiments. The sentiment of “mayor” can be easily confused by the keyword, “Interesting”, which is of positive sentiment. (2) The images contain a large amount of detailed information, and the visual contents are usually related to only one or several of the aspects. For example, as shown in Figure 1, the objects in red boxes are more helpful in analyzing the sentiment of “Mayor Kadokawa” than the other aspects. The complex input greatly challenges the recognition of aspect-based sentiment.

Considering the multimodal input, existing methods are typically dedicated to associated visual and textual contents (Ju et al., 2021; Ling et al., 2022; Yang et al., 2022). Ju et al. (2021) uses image-text relation to evaluate the contribution of visual contents to aspect sentiment, based on which to determine whether the image is involved in sentiment analysis. Ling et al. (2022) and Yang et al. (2022) align visual representations of objects and their attributes with corresponding textual contents. To summarize, the whole image is directly associated with textual content in these methods. Intuitively, without aligning image blocks to corresponding aspects, the coarse whole-image-text association can introduce aspect-irrelevant visual noise, which further hinders aspect sentiment analysis. In addition, the performance can be further impacted by the

textual noise from the confusion among different aspects.

In this paper, we propose an Aspect-oriented Method (AoM) to mitigate aforementioned noises from both image and text. AoM can detect aspect-relevant information from perspectives of both semantics and sentiment. There are two key modules in AoM: Aspect-Aware Attention Module ( $A^3M$ ) for semantically fine-grained image-text alignment and Aspect-Guided Graph Convolutional Network (AG-GCN) for sentiment information aggregation. In  $A^3M$ , we first extract aspect features associated with each visual and textual token. And then aspect-relevant token representations are computed based on their relevance to the corresponding aspect features. In AG-GCN, we first explicitly add sentiment embeddings to the obtained representations of visual and textual tokens. A multimodal weighted-association matrix is constructed containing aspect-to-image-block similarity and word-to-word dependency. Then we use a graph convolutional network to aggregate sentiment information according to the constructed multimodal matrix.

The contributions can be summarized as follows:

(1) We propose an aspect-oriented network to mitigate the visual and textual noises from the complex image-text interactions.

(2) We design an aspect-aware attention module and an aspect-guided graph convolutional network to effectively detect aspect-relevant multimodal contents from the perspectives of semantic and sentiment, respectively.

(3) Experiments on two benchmark datasets, including Twitter2015 and Twitter2017, show that our approach generally outperforms the state-of-the-art methods.

## 2 Related Work

In this section, we review the existing methods for both ABSA and MABSA.

### 2.1 Aspect-based Sentiment Analysis

In the past few years, Aspect-Based Sentiment Analysis (ABSA) in the textual fields has attracted much attention and gained mature research (Chen and Qian, 2020; Oh et al., 2021; Xu et al., 2020). On the one hand, most recent works are based on the pre-trained language model BERT because of its remarkable performance in many NLP tasks (Liang et al., 2022a). On the other hand, some recent efforts focus on modeling the dependency

relationship between aspects and their corresponding descriptions, in which graph convolutional networks (GCNs) (Chen et al., 2022; Liang et al., 2022b, 2020; Li et al., 2021a; Pang et al., 2021) or graph attention networks (GATs) (Yuan et al., 2020) over dependency with the syntactic structure of a sentence are fully exploited.

### 2.2 Multimodal Aspect-based Sentiment Analysis

With the enrichment of multimodal users' posts in social media, researchers find that images offer great supplementary information in aspect term extraction (Wu et al., 2020a; Zhang et al., 2018; Asgari-Chenaghlu et al., 2021) and sentiment analysis (Wu et al., 2022; Li et al., 2021b; Hazarika et al., 2020; Cai et al., 2019). Thus, Multimodal Aspect-based Sentiment Analysis (MABSA) begins to be widely studied. MABSA task can be divided into two independent sub-tasks, i.e., Multimodal Aspect Term Extraction (MATE) and Multimodal Aspect-oriented Sentiment Classification (MASC). The former extracts all aspect terms in the sentence at the prompt of the image, and the latter predicts the sentiment polarities for the aspects.

Ju et al. (2021) first realizes MABSA in a unified framework and designs an auxiliary cross-modal relation detection to control whether the visual information will be used in prediction. For capturing cross-modal alignment, Ling et al. (2022) constructs a generative multimodal architecture based on BART for both vision-language pre-training and the downstream MABSA tasks. Yang et al. (2022) dynamically controls the contributions of the visual information to different aspects via the trick that the lower confidence of the results predicted by purely textual is, the more contributions from images will be considered.

On the one hand, the above methods ignore the alignment of fine-grained visual blocks and the corresponding aspects, which introduce irrelevant visual noise. On the other hand, modeling of syntax dependency and sentiment information for aspect descriptions is absent in these methods, which is proved important in sentiment analysis (Liang et al., 2022a; Kalaivani et al., 2022; Xu et al., 2022).

To tackle the aforementioned issues, we propose an aspect-oriented model consisting of Aspect-Aware Attention Module and Aspect-Guided Graph Convolutional Network which respectively work to capture semantic information by fine-grained

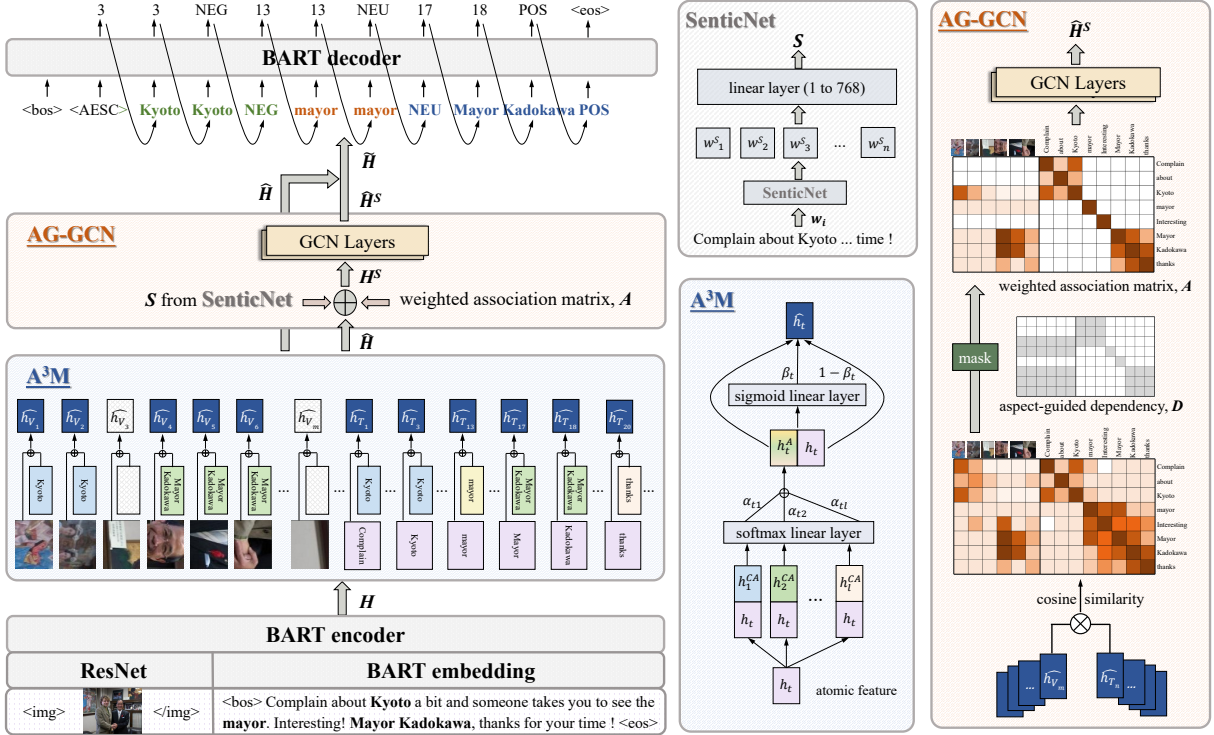


Figure 2: The overview of our proposed aspect-oriented model AoM.

image-text alignment and effectively aggregate aspect-relevant sentiment information.

### 3 Methodology

#### 3.1 Overview

**Task Definition.** Formally, given a tweet that contains an image  $V$  and a sentence with  $n$  words  $S = (w_1, w_2, \dots, w_n)$ , our goal is to acquire the sequence  $Y$  representing all aspects and their associated sentiment polarities. We formulate the output of MABSA as  $Y = [a_1^s, a_1^e, s_1, \dots, a_i^s, a_i^e, s_i, \dots, a_k^s, a_k^e, s_k]$ , where  $a_i^s, a_i^e$  and  $s_i$  depict the start index, end index of the  $i$ -th aspect and its sentiment polarity in the tweet, and  $k$  is the number of aspects.

**Model preview.** Figure 2 shows the overview of our model architecture, which builds on an encoder-decoder architecture based on BART (Lewis et al., 2019). Between the encoder and the decoder of BART, we creatively implement the Aspect-Aware Attention Module ( $A^3M$ ) and Aspect-Guided Graph Convolutional Network (AG-GCN) to align the textual aspect to its associated visual blocks and textual description, simultaneously mitigate interference both from semantics and sentiment among different aspects. In the following subsections, we will illustrate the details of

the proposed model.

**Feature Extractor.** The initial word embeddings are obtained from the pre-trained BART due to its excellent ability of textual representation. The embeddings of visual blocks are obtained by pre-processing via ResNet (Chen et al., 2014) following (Yu et al., 2019). We consider every feature of a visual block or word token as an atomic feature. We add  $\langle \text{img} \rangle$  and  $\langle / \text{img} \rangle$  before and after the visual features,  $\langle \text{bos} \rangle$  and  $\langle \text{eos} \rangle$  for the textual features. Then, we concatenate the multimodal features as  $X$  which is the input of BART encoder.

We can get the multimodal hidden state  $H = \{h_0^V, \dots, h_i^V, \dots, h_m^V, h_0^T, \dots, h_j^T, \dots, h_n^T\}$  with  $m$  visual blocks and  $n$  words, where  $h_i^V$  and  $h_j^T$  refer to features of the  $i$ -th visual block and the  $j$ -th word in the sentence.

#### 3.2 Aspect-Aware Attention Module ( $A^3M$ )

Since aspects are not specially modeled by BART encoder, we creatively design the Aspect-Aware Attention Module ( $A^3M$ ) aiming to capture aspect-relevant semantic information. For this purpose, we align the multimodal information of target objects and filter out the semantic noise from images.

First, as aspects are usually noun phrases from the sentences, we extract those phrases as the

candidate aspects (CA) with the NLP tool Spacy<sup>1</sup>. And from the hidden state  $H$  of the BART encoder, we obtain the features of all candidate aspects denoted as  $H^{CA} = \{h_1^{CA}, \dots, h_i^{CA}, \dots, h_l^{CA}\}$ , where  $l$  is the number of noun phrases in the sentence. To get the relationship between candidate aspects and atomic features, we implement an attention-based mechanism guided by the candidate aspects. Given the  $t$ -th hidden feature  $h_t$ , its attention distribution  $\alpha_t$  over  $k$  candidate aspects is obtained by:

$$Z_t = \tanh((W_{CA}H^{CA} + b_{CA}) \oplus (W_H h_t + b_H)), \quad (1)$$

$$\alpha_t = \text{softmax}(W_\alpha Z_t + b_\alpha), \quad (2)$$

where  $Z_t \in \mathbb{R}^{2d \times k}$  is the comprehensive feature extracted from both the candidate aspects and the hidden states.  $H^{CA} \in \mathbb{R}^{d \times k}$  denotes the features of candidate aspects.  $W_{CA} \in \mathbb{R}^{d \times d}$ ,  $W_H \in \mathbb{R}^{d \times d}$ ,  $W_\alpha \in \mathbb{R}^{1 \times 2d}$ ,  $b_{CA}$ ,  $b_H$  and  $b_\alpha$  are the learned parameters.  $\oplus$  is an operator between a matrix and a vector, where the vector is repeated into the appropriate size to concatenate with the matrix. We then get the aspect-related hidden feature  $h_t^A$  by calculating the weighted sum of all candidate aspects following the below equation:

$$h_t^A = \sum_i^k \alpha_{t,i} h_i^{CA}. \quad (3)$$

For example, if a visual block is strongly associated with the  $j$ -th aspect, the corresponding  $\alpha_{t,j}$  is approximately 1.  $h_t^A$  would be equal to the aspect semantically. And if the visual block is not related to any specific candidate aspects, both  $\alpha_t$  and  $h_t^A$  would be zero-like vectors of no information.

Considering that not every visual block can be used for prediction,  $\beta_t$  is learned to add up the atomic feature  $h_t$  and its aspect-related hidden feature  $h_t^A$ . Details are as follows:

$$\beta_t = \text{sigmoid}(W_\beta [W_1 h_t; W_2 h_t^A] + b_\beta), \quad (4)$$

$$\hat{h}_t = \beta_t h_t + (1 - \beta_t) h_t^A, \quad (5)$$

where  $W_\beta$ ,  $W_1$ ,  $W_2$ ,  $b_\beta$  are parameters, and  $[\cdot]$  denotes the concatenation operator for vectors.  $\hat{h}_t \in \hat{H}$  is the final output of A<sup>3</sup>M after the semantic alignment and the noise reduction procedure. Thus we get the noiseless and aligned information for every atomic feature.

**Pre-training** To align the two modalities and reduce noise, we conduct a pre-training task in A<sup>3</sup>M.

<sup>1</sup>Spacy: <https://spacy.io/>

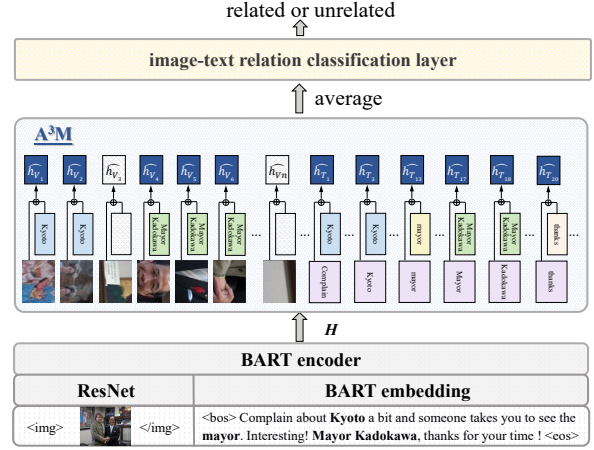


Figure 3: The framework of the pre-training task.

Specifically, we detect the image-text relationship on the datasets TRC (Vempala and Preoțiu-Pietro, 2019) as illustrated by Figure 3. We first obtain the average feature of image blocks from the output of A<sup>3</sup>M and then pass it to a fully connected softmax layer, which outputs a probability distribution over whether the image is related to the text. Finally, we use cross entropy loss to train our model.

### 3.3 Aspect-Guided Graph Convolutional Network (AG-GCN)

The aspect-focused interaction between visual modality and textual modality in A<sup>3</sup>M concentrates on the context semantics, and that is not adequate for MABSA. Sentiment interference among different aspects still exists and influences sentiment prediction. Thus, we design the Aspect-Guided Graph Convolutional Network (AG-GCN) module to introduce external sentiment information and mitigate emotional confusion among different aspects to a certain extent.

Specifically, for word  $w_i$  in the sentence, we gain its affective score  $w_i^S$  from SenticNet (Ma et al., 2018) and project it to the space with the same dimension as  $h_t^A$ , with  $s_i$  obtained. Then we add the sentiment feature  $s_i$  to the output of A<sup>3</sup>M:

$$w_i^S = \text{SenticNet}(w_i), \quad (6)$$

$$s_i = W_S w_i^S + b_S, \quad (7)$$

$$h_i^S = \hat{h}_i + s_i, \quad (8)$$

where  $W_S$ ,  $b_S$  are the learned parameters.  $h_i^S$  is the feature with affective knowledge.

Next, we build a boolean dependency matrix  $D$  among visual blocks and words. First, for the word-to-word part, submatrix  $D_{TT}$  representing



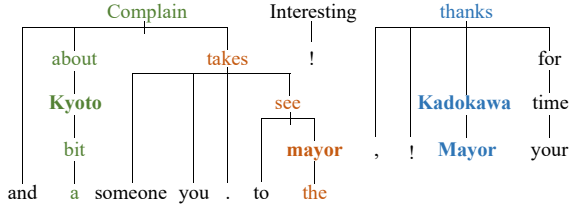


Figure 4: The dependency tree of the example mentioned in the introduction.

the dependency tree<sup>2</sup> of the input sentence like Figure 4. If two words can be associated within two generations, the element of  $D_{TT}$  would be set to 1, otherwise 0 instead. For example, “Kyoto” is associated with “bit” (child), “a” (grandchild), “about” (father) and “Complain” (grandfather). Second, the visual dependency submatrix  $D_{VV}$  is initialized as a diagonal matrix. And as for the word-image-block dependency, denoted as  $D_{TV}$  and equaled to  $D_{VT}^T$ , we set all the elements in the  $i$ -th line of  $D_{TV}$  to 1 if the  $i$ -th word is an aspect, otherwise 0. And the matrix  $D$  is defined as:

$$D = \begin{bmatrix} D_{VV} & D_{VT} \\ D_{TV} & D_{TT} \end{bmatrix}, \quad (9)$$

Considering the different importance of different dependencies, we attach weights onto  $D$  with cosine similarity among  $\hat{h}_i$  as follows:

$$A_{ij} = D_{ij} F_{\text{cosine\_similarity}}(\hat{h}_i, \hat{h}_j), \quad (10)$$

where both  $D, A \in \mathbb{R}^{(m+n) \times (m+n)}$ , and  $A$  is the weighted association matrix.

AG-GCN takes  $H^S$  from Eq.8 as initial node representations in the graph. For the  $i$ -th node at the  $l$ -th layer, the hidden state  $h_{i,l}^S$  is updated by the following equation:

$$h_{i,l}^S = \text{ReLU}\left(\sum_{j=1}^n A_{ij} W_l h_{j,l-1}^S + b_l\right), \quad (11)$$

where  $W_l, b_l$  are learned parameters and we use ReLU as activation function. Significantly,  $h_{i,0}^S$  is equal to  $h_i^S$ . Accordingly, we get the final output  $\hat{H}^S$  from the last GCN layer which is rich in sentiment information. Every underlying aspect aggregates its relevant information from both the image-text pair. Moreover, sentiment confusion of different aspects is weakened because the association matrix makes little interference among different aspects.

<sup>2</sup>We use spaCy toolkit to construct the dependency tree referring from <https://spacy.io>

	Twitter2015	Twitter2017
#sentence	3,502	2,910
#with one aspect	2,159 (61.65%)	976 (33.54%)
#with multiple aspects	1,343 (38.35%)	1,934 (66.46%)
#with multiple sentiments	1,257	1,690

Table 1: Statistics of the two benchmark datasets. Line 1 is the number of sentences. #X in the last 3 lines denotes the number of sentences with such characteristics X.

### 3.4 Prediction and Loss Function

The BART decoder takes the combination of  $\hat{H}$ ,  $\hat{H}^S$ , and the previous decoder output  $Y_{<t}$  as inputs, and predicts the token probability distribution as follows:

$$\tilde{H} = \lambda_1 \hat{H} + \lambda_2 \hat{H}^S, \quad (12)$$

$$h_t^d = \text{Decoder}(\tilde{H}; Y_{<t}) \quad (13)$$

$$\bar{H}_T = (W + \tilde{H}_T)/2 \quad (14)$$

$$P(y_t) = \text{softmax}([\bar{H}_T; C^d] h_t^d) \quad (15)$$

where  $\lambda_1, \lambda_2$  are the hyper-parameters to control the contribution from the two modules.  $\tilde{H}_T$  is the textual part of  $\tilde{H}$ .  $W$  denotes the embeddings of input tokens.  $C^d$  means the embeddings of the [positive, neutral, negative, <eos>]. The loss function is as follows:

$$\mathcal{L} = -\mathbb{E}_{X \sim D} \sum_{t=1}^O \log P(y_t | Y_{<t}, X), \quad (16)$$

where  $O = 2M + 2N + 2$  is the length of  $Y$ , and  $X$  denotes the multimodal input.

## 4 Experiment

### 4.1 Experimental settings

**Datasets.** Our two benchmark datasets are Twitter2015 and Twitter2017 (Yu and Jiang (2019)). As shown in the statistics of Table 1, sentences with multiple aspects take up a considerable part of the two datasets.

**Implementation Details.** Our model is based on BART (Lewis et al., 2019), and the pre-training task is trained for 40 epochs with batch size 64, and for 35 epochs with batch size 16 on MABSA. The learning rates are both  $7e-5$  and hidden sizes are 768. Hyper-parameters  $\lambda_1$  and  $\lambda_2$  are 1 and 0.5 respectively. Besides, we pre-train A<sup>3</sup>M on TRC dataset (Vempala and Preojuic-Pietro, 2019), which is divided into two groups according to whether the text is represented.

Methods	Twitter2015			Twitter2017			
	P	R	F1	P	R	F1	
Text-based	SPAN* (Hu et al., 2019)	53.7	53.9	53.8	59.6	61.7	60.6
	D-GCN* (Chen et al., 2020)	58.3	58.8	59.4	64.2	64.1	64.1
	BART* (Yan et al., 2021)	62.9	65.0	63.9	65.2	65.6	65.4
Multimodal	UMT+TomBERT* (Yu et al., 2020; Yu and Jiang, 2019)	58.4	61.3	59.8	62.3	62.4	62.4
	OSCGA+TomBERT* (Wu et al., 2020c; Yu and Jiang, 2019)	61.7	63.4	62.5	63.4	64.0	63.7
	OSCGA-collapse* (Wu et al., 2020c)	63.1	63.7	63.2	63.5	63.5	63.5
	RpBERT-collapse* (Sun et al., 2021)	49.3	46.9	48.0	57.0	55.4	56.2
	UMT-collapse (Yu et al., 2020)	61.0	60.4	61.6	60.8	60.0	61.7
	JML (Ju et al., 2021)	65.0	63.2	64.1	66.5	65.5	66.0
	VLP-MABSA* (Ling et al., 2022)	<u>65.1</u>	68.3	<u>66.6</u>	66.9	69.2	68.0
	CMMT (Yang et al., 2022)	64.6	<u>68.7</u>	66.5	<u>67.6</u>	<u>69.4</u>	<u>68.5</u>
	AoM (ours)	<b>67.9</b>	<b>69.3</b>	<b>68.6</b>	<b>68.4</b>	<b>71.0</b>	<b>69.7</b>

Table 2: Results of different methods for MABSA on the two Twitter datasets. \* denotes the results from Ling et al. (2022). The best results are bold-typed and the second best ones are underlined.

**Evaluation Metrics.** We evaluate the performance of our model on MABSA task and MATE task by Micro-F1 score (F1), Precision (P) and Recall (R), while on MASC task we use Accuracy (Acc) and F1 following previous studies.

## 4.2 Baselines

We compare our proposed model with four types of methods listed below.

**Approaches for textual ABSA.** 1) **SPAN** (Hu et al., 2019) detects opinion targets with their sentiments. 2) **D-GCN** (Chen et al., 2020) models dependency relations among words via dependency tree. 3) **BART** (Yan et al., 2021) solves seven ABSA subtasks in an end-to-end framework.

**Approaches for MATE.** 1) **RAN** (Wu et al., 2020b) focus on alignment of text and object regions. 2) **UMT** (Yu et al., 2020) takes text-based entity span detection as an auxiliary task. 3) **OSCGA** (Wu et al., 2020c) focus on alignments of visual objects and entities.

**Approaches for MASC.** 1) **ESAFN** (Yu et al., 2019) is an entity-level sentiment analysis method based on LSTM. 2) **TomBERT** (Yu and Jiang, 2019) applies BERT to obtain aspect-sensitive textual representations. 3) **CapTrBERT** (Khan and Fu, 2021) translates images into text and construct an auxiliary sentence for fusion.

**Approaches for MABSA.** 1) **UMT-collapse** (Yu et al., 2020), **OSCGA-collapse** (Wu et al., 2020c) and **RpBERT-collapse** (Sun et al., 2021) are adapted from models for MATE by using collapsed labels to represent aspect and sentiment pairs. 2) **UMT+TomBERT**, **OSCGA+TomBERT** are two pipeline approaches by combining UMT (Yu et al., 2020) or OSCGA (Wu et al., 2020c) with

**TomBERT** (Yu and Jiang, 2019). 3) **JML** (Ju et al., 2021) is the first joint model for MABSA with auxiliary cross-modal relation detection module. 4) **CMMT** (Yang et al., 2022) implements a gate to control the multimodal information contributions during inter-modal interactions. 5) **VLP-MABSA** (Ling et al., 2022) performs five task-specific pre-training tasks to model aspects, opinions and alignments.

## 4.3 Main Results

In this section, we show the excellent performance of AoM on the two datasets for the three tasks compared with SOTAs.

**Performance on MABSA:** The results for MABSA are shown in Table 2. **First**, our AoM far exceeds all text-based models, which means detection of richer visual information and textual information in our model is helpful. **Second**, multimodal pipeline methods and adaptive methods are generally unsatisfactory, because they ignore the interaction between the semantic information and sentiment for the two sub-tasks. **Last**, AoM outperforms all multimodal methods in every metric. Especially, AoM achieves the improvement of 2% and 1.2% with respect to F1 in contrast with the second best models on two datasets (*VLP-MABSA* for Twitter2015 and *CMMT* for Twitter2017), which demonstrates the effectiveness of learning aspect-relevant visual blocks and textual words compared to focusing on all visual and textual inputs.

**Performance on MATE:** As shown in Table 3, AoM is ahead of most of the current models and performs the best in Twitter 2015 by 0.3% higher than the second best *CMMT* on F1. The performance of *CMMT* in Twitter2017 is 0.8% higher

Methods	Twitter2015			Twitter2017		
	P	R	F1	P	R	F1
RAN*	80.5	81.5	81.0	90.7	90.7	90.0
UMT*	77.8	81.7	79.7	86.7	86.8	86.7
OSCGA*	81.7	82.1	81.9	90.2	90.7	90.4
JML*	83.6	81.2	82.4	92.0	90.7	91.4
VLP-MABSA*	83.6	87.9	85.7	90.8	92.6	91.7
CMMT	83.9	88.1	85.9	92.2	93.9	93.1
AoM (ours)	84.6	87.9	86.2	91.8	92.8	92.3

Table 3: Results of different methods for MATE. \* denotes the results from Ling et al. (2022).

Methods	Twitter2015		Twitter2017	
	ACC	F1	ACC	F1
ESAFN	73.4	67.4	67.8	64.2
TomBERT	77.2	71.8	70.5	68.0
CapTrBERT	78.0	73.2	72.3	70.2
JML	78.7	-	72.7	-
VLP-MABSA	78.6	73.8	73.8	71.8
CMMT	77.9	-	73.8	-
AoM (ours)	80.2	75.9	76.4	75.0

Table 4: Results of different methods for MASC.

than ours, probably due to our model wrongly predicting some noun phrases as aspects. But considering the improvement in MASC and MABSA, it is still worthy treating all noun phrases in the sentence as candidate aspects when acquiring aspect-relevant visual information.

**Performance on MASC:** Table 4 shows the performance of MASC. It is exciting that our model outperforms the second-best results by 1.5% and 2.6% in accuracy, 2.1% and 3.2% points in F1 score on Twitter2015 and Twitter2017. It demonstrates that AoM has the ability to detect aspect-related sentiment information from both images and text, even disturbed by other noisy aspects.

#### 4.4 Ablation Study

In this section, we research the effectiveness of each component in AoM, the results are shown in Table 5.

**W/o A<sup>3</sup>M&AG-GCN** shows that after removing the two specially designed modules, the per-

Methods	Twitter2015			Twitter2017		
	P	R	F1	P	R	F1
Full	67.9	69.3	68.6	68.4	71.0	69.7
w/o A <sup>3</sup> M&AG-GCN	65.7	67.3	66.5	66.5	69.0	67.8
w/o A <sup>3</sup> M&TRC	62.1	61.0	61.6	63.7	64.1	63.9
w/o TRC	66.8	68.4	67.6	67.8	69.8	68.8
w/o AG-GCN	67.0	69.4	68.2	67.8	69.7	68.8
w/o SenticNet	65.7	70.5	68.0	68.1	69.4	68.7
w/o TRC&AG-GCN	66.7	69.2	68.0	67.8	69.5	68.6

Table 5: The performance comparison of our full model and its ablated approaches.

Text	(a) NBA Western Conference Finals: Golden State Warriors shock Oklahoma City Thunder,...	(b) This subtle difference between Daniel Radcliffe and Elijah Wood is pretty unsettling.
	Image	
VLP-MABSA	(NBA, NEU) (✓, ✓) -- (Oklahoma, NEU) (✗, ✗)	(Daniel Radcliffe, NEU) (✓, ✗) (Elijah Wood, NEU) (✗, ✗)
BART+A <sup>3</sup> M	(NBA, NEU) (✓, ✓) (Golden State Warriors, POS) (✓, ✓) (Oklahoma City Thunder, NEG) (✓, ✓)	(Daniel Radcliffe, NEU) (✓, ✗) (Elijah Wood, NEG) (✓, ✓)
AoM	(NBA, NEU) (✓, ✓) (Golden State Warriors, POS) (✓, ✓) (Oklahoma City Thunder, NEG) (✓, ✓)	(Daniel Radcliffe, NEG) (✓, ✓) (Elijah Wood, NEG) (✓, ✓)

Figure 5: Two cases with predictions by VLP-MABSA (Ling et al., 2022), BART+A<sup>3</sup>M, and our model.

formance declines by 2.1% on Twitter2015 and 1.9% on Twitter2017. It fully demonstrates their contributions to learning effective information.

**W/o A<sup>3</sup>M&TRC** performs worse after removing A<sup>3</sup>M including the pre-training on TRC. It proves the necessity of modeling semantic alignment between visual blocks and aspects in A<sup>3</sup>M. With the alignment, AG-GCN can obtain appropriate aspect-image-block and text-text association.

**W/o TRC pre-training** shows a slight drop after we remove the TRC pre-training on A<sup>3</sup>M, which implies relevant pre-training task is useful for the model to learn better parameters.

**W/o AG-GCN** displays the performance without AG-GCN, declining by 0.42% on Twitter2015 and 0.9% on Twitter2017. It means that AG-GCN does make the prediction focus on specific aspects related to blocks and words with syntax dependencies. In other words, the multimodal interference from other aspects can be mitigated.

**W/o SenticNet** is the model without sentiment information in AG-GCN. Its performance shows adding external affective knowledge can enhance the sentiment comprehension of the model.

**W/o TRC&AG-GCN** is the BART model only with our A<sup>3</sup>M module. We can see from Table 5 that *w/o TRC&AG-GCN* improves *w/o A<sup>3</sup>M&AG-GCN* by 1.5% and 0.8%. So it is effective to align the fine-grained visual block to related aspect and reduce irrelevant information.

#### 4.5 Case Study

To better analyze how the Aspect-Aware Attention Module and Aspect-Guided Graph Convolutional Network work, we present the case study as follows.

Figure 5 displays two examples with predictions from VLP-MABSA (Ling et al., 2022),

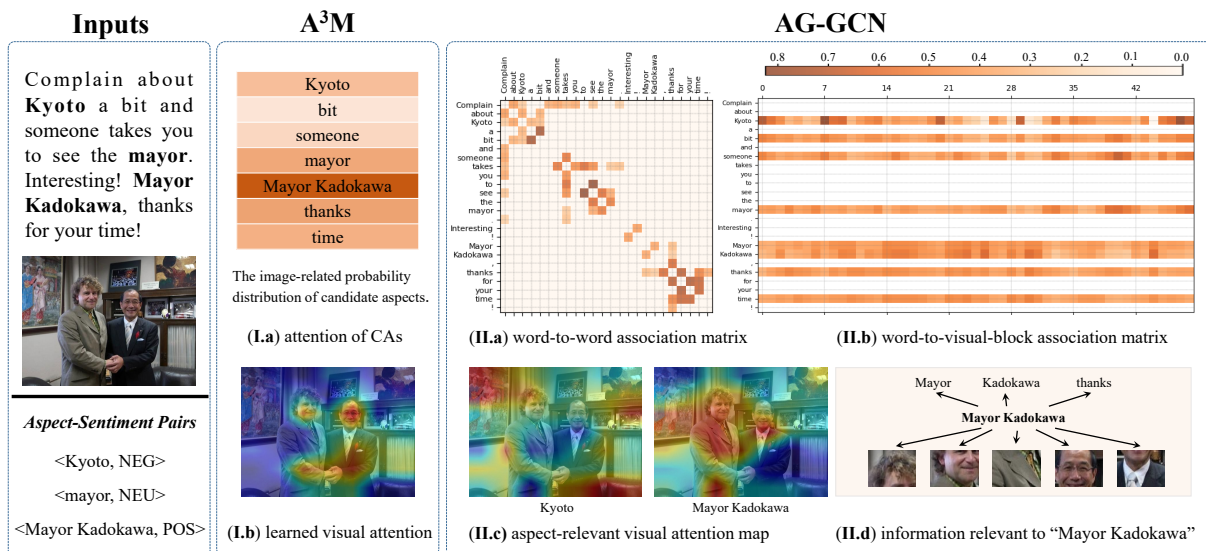


Figure 6: Visualization of the attention maps in A<sup>3</sup>M and the sub-parts of the weighted-association matrix AG-GCN.

BART+A<sup>3</sup>M and our AoM. In example (a), VLP-MABSA misses the aspect “Golden State Warriors”, gets an incomplete aspect “Oklahoma City Thunder” and wrongly predicts the sentiment. It may be caused by the interference from the visual region which represents pride expression of a person. However, BART+A<sup>3</sup>M gets all right predictions due to the ability of aspect-oriented attention. In example (b), compared with our whole model, BART+A<sup>3</sup>M wrongly predicts the sentiment of “Daniel Radcliffe” which should be negative. We attribute the wrong prediction to lacking syntax association which benefits sentiment transmission. In other words, AG-GCN contributes to the correctness.

#### 4.6 Attention Visualization

To investigate the effectiveness of detecting aspect-relevant information, we visualize the attention process as shown in Figure 6.

**For A<sup>3</sup>M:** (i) Figure 6-(I.a) shows the attention weights of candidate aspects computed according to the images. We can see that “Mayor Kadokawa” is the most relevant aspect. (ii) Figure 6-(I.b) shows the proportions of the visual information reserved at the last step in A<sup>3</sup>M, where we weighted add up the representations of visual blocks and the corresponding aspects. The heat map shows that the visual information associated with “Mayor Kadokawa” is reserved to a great extent, while the helpless information from other blocks is disregarded as noise. It demonstrates that attention in A<sup>3</sup>M is able to detect aspect-relevant information.

**For AG-GCN:** (i) Figure 6-(II.a) shows the word-to-word part of the weighted association matrix. The matrix effectively excludes sentiment interference from other aspects by adding syntax dependency information. For example, the sentiment of “mayor” cannot be influenced by irrelevant keywords, such as “Complain” and “thanks”. (ii) Figure 6-(II.b) shows the dependencies between visual blocks and words. (iii) Specifically, we visualize the visual attention of aspects “Kyoto” (see Figure 6-(II.c) left) and “Mayor Kadokawa” (see Figure 6-(II.c) right). We can see that “Kyoto” pays more attention to the pictures hanging on the wall which are full of Japanese elements related to the place, while “Mayor Kadokawa” focus more on the joyful expressions of the two people. (iv) Figure 6-(II.d) shows the words and image blocks “Mayor Kadokawa” focused on in sentiment transmission. It’s obvious that these attentions are helpful for the prediction.

## 5 Conclusion

In this paper, we proposed an aspect-oriented model (AoM) for the task of multimodal aspect-based sentiment analysis. We use two specially designed modules to detect aspect-relevant information from the semantic and sentiment perspectives. On the one hand, to learn aspect-relevant semantic information especially from the image, we construct the Aspect-Aware Attention Module to align the visual information and descriptions to the corresponding aspect. On the other hand, to detect the aspect-relevant sentiment information,



we explicitly add sentiment embedding into AoM. Then, a graph convolutional network is used to aggregate the semantic and sentiment embedding under the guidance of both image-text similarity and syntax dependency in sentences. The experimental results on two widely used datasets demonstrate the effectiveness of our method.

## Limitations

Though our proposed method outperforms current state-of-the-art methods, there are still many challenges we should overcome in future research. First, for colloquial expression which confuses current dependency tree parser, we should come up with new solutions. Second, emotional prediction of tweet posts describing current issues needs external knowledge, which is absent in existing research.

## Acknowledgments

We thank anonymous reviewers for their valuable comments. This work was supported by the Natural Science Foundation of Tianjin, China (No.22JCQJC00150, 22JCQNJC01580), the National Natural Science Foundation of China (No.62272250), Tianjin Research Innovation Project for Postgraduate Students (No.2022SKYZ232), and the Fundamental Research Funds for the Central Universities (No. 63231149).

## References

- Meysam Asgari-Chenaghlu, M. Reza Feizi-Derakhshi, Leili Farzinvash, M. A. Balafar, and Cina Motamed. 2021. [CWI: A multimodal deep learning approach for named entity recognition from social media using character, word and image features](#). *Neural Computing and Applications*, 34(3):1905–1922.
- Yitao Cai, Huiyu Cai, and Xiaojun Wan. 2019. [Multi-Modal Sarcasm Detection in Twitter with Hierarchical Fusion Model](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2506–2515, Florence, Italy. Association for Computational Linguistics.
- Guimin Chen, Yuanhe Tian, and Yan Song. 2020. Joint aspect extraction and sentiment analysis with directional graph convolutional networks. In *Proceedings of the 28th international conference on computational linguistics*, pages 272–279.
- Hao Chen, Zepeng Zhai, Fangxiang Feng, Ruifan Li, and Xiaojie Wang. 2022. [Enhanced Multi-Channel Graph Convolutional Network for Aspect Sentiment Triplet Extraction](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2974–2985, Dublin, Ireland. Association for Computational Linguistics.
- Tao Chen, Damian Borth, Trevor Darrell, and Shih-Fu Chang. 2014. [Deepsentibank: Visual sentiment concept classification with deep convolutional neural networks](#).
- Zhuang Chen and Tiejun Qian. 2020. [Relation-Aware Collaborative Learning for Unified Aspect-Based Sentiment Analysis](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3685–3694, Online. Association for Computational Linguistics.
- Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. 2020. [Misa: Modality-invariant and -specific representations for multimodal sentiment analysis](#). In *Proceedings of the 28th ACM International Conference on Multimedia, MM '20*, page 1122–1131, New York, NY, USA. Association for Computing Machinery.
- Minghao Hu, Yuxing Peng, Zhen Huang, Dongsheng Li, and Yiwei Lv. 2019. Open-domain targeted sentiment analysis via span-based extraction and classification. [arXiv preprint arXiv:1906.03820](#).
- Xincheng Ju, Dong Zhang, Rong Xiao, Junhui Li, Shoushan Li, Min Zhang, and Guodong Zhou. 2021. [Joint Multi-modal Aspect-Sentiment Analysis with Auxiliary Cross-modal Relation Detection](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4395–4405, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- KS Kalaivani, M Rakshana, K Mounika, and D Sindhu. 2022. Senticnet-based feature weighting scheme for sentiment classification. In *Mobile Computing and Sustainable Informatics*, pages 839–848. Springer.
- Zaid Khan and Yun Fu. 2021. Exploiting bert for multimodal target sentiment classification through input space translation. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 3034–3042.
- Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. [arXiv preprint arXiv:1910.13461](#).
- Ruifan Li, Hao Chen, Fangxiang Feng, Zhanyu Ma, Xiaojie Wang, and Eduard Hovy. 2021a. [Dual Graph Convolutional Networks for Aspect-based Sentiment Analysis](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*

- and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 6319–6329, Online. Association for Computational Linguistics.
- Yuanqing Li, Ke Zhang, Jingyu Wang, and Xinbo Gao. 2021b. [A cognitive brain model for multimodal sentiment analysis based on attention neural networks](#). *Neurocomputing*, 430:159–173.
- Bin Liang, Hang Su, Lin Gui, Erik Cambria, and Ruifeng Xu. 2022a. [Aspect-based sentiment analysis via affective knowledge enhanced graph convolutional networks](#). *Knowledge-Based Systems*, 235:107643.
- Bin Liang, Rongdi Yin, Lin Gui, Jiachen Du, and Ruifeng Xu. 2020. [Jointly Learning Aspect-Focused and Inter-Aspect Relations with Graph Convolutional Networks for Aspect Sentiment Analysis](#). In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 150–161, Barcelona, Spain (Online). International Committee on Computational Linguistics.
- Shuo Liang, Wei Wei, Xian-Ling Mao, Fei Wang, and Zhiyong He. 2022b. [BiSyn-GAT+: Bi-Syntax Aware Graph Attention Network for Aspect-based Sentiment Analysis](#). In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1835–1848, Dublin, Ireland. Association for Computational Linguistics.
- Yan Ling, Jianfei Yu, and Rui Xia. 2022. [Vision-Language Pre-Training for Multimodal Aspect-Based Sentiment Analysis](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2149–2159, Dublin, Ireland. Association for Computational Linguistics.
- Yanxia Lv, Fangna Wei, Lihong Cao, Sancheng Peng, Jianwei Niu, Shui Yu, and Cuirong Wang. 2021. [Aspect-level sentiment analysis using context and aspect memory network](#). *Neurocomputing*, 428:195–205.
- Yukun Ma, Haiyun Peng, and Erik Cambria. 2018. [Targeted aspect-based sentiment analysis via embedding commonsense knowledge into an attentive lstm](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1).
- Shinhyeok Oh, Dongyub Lee, Taesun Whang, IINam Park, Seo Gaeun, EungGyun Kim, and Harksoo Kim. 2021. [Deep Context- and Relation-Aware Learning for Aspect-based Sentiment Analysis](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 495–503, Online. Association for Computational Linguistics.
- Shiguan Pang, Yun Xue, Zehao Yan, Weihao Huang, and Jinhui Feng. 2021. [Dynamic and Multi-Channel Graph Convolutional Networks for Aspect-Based Sentiment Analysis](#). In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2627–2636, Online. Association for Computational Linguistics.
- Lin Sun, Jiquan Wang, Kai Zhang, Yindu Su, and Fangsheng Weng. 2021. [Rpbert: A text-image relation propagation-based bert model for multimodal ner](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(15):13860–13868.
- Alakananda Vempala and Daniel Preoțiuc-Pietro. 2019. [Categorizing and inferring the relationship between the text and image of twitter posts](#). In *Proceedings of the 57th annual meeting of the Association for Computational Linguistics*, pages 2830–2840.
- Hanqian Wu, Siliang Cheng, Jingjing Wang, Shoushan Li, and Lian Chi. 2020a. [Multimodal aspect extraction with region-aware alignment network](#). In *Natural Language Processing and Chinese Computing*, pages 145–156, Cham. Springer International Publishing.
- Hanqian Wu, Siliang Cheng, Jingjing Wang, Shoushan Li, and Lian Chi. 2020b. [Multimodal aspect extraction with region-aware alignment network](#). In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 145–156. Springer.
- Yang Wu, Yanyan Zhao, Hao Yang, Song Chen, Bing Qin, Xiaohuan Cao, and Wenting Zhao. 2022. [Sentiment Word Aware Multimodal Refinement for Multimodal Sentiment Analysis with ASR Errors](#). In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1397–1406, Dublin, Ireland. Association for Computational Linguistics.
- Zhiwei Wu, Changmeng Zheng, Yi Cai, Junying Chen, Ho-fung Leung, and Qing Li. 2020c. [Multimodal representation with embedded visual guiding objects for named entity recognition in social media posts](#). In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1038–1046.
- Junjie Xu, Shuwen Yang, Luwei Xiao, Zhichao Fu, Xingjiao Wu, Tianlong Ma, and Liang He. 2022. [Graph convolution over the semantic-syntactic hybrid graph enhanced by affective knowledge for aspect-level sentiment classification](#). In *2022 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.
- Lu Xu, Hao Li, Wei Lu, and Lidong Bing. 2020. [Position-Aware Tagging for Aspect Sentiment Triplet Extraction](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2339–2349, Online. Association for Computational Linguistics.

- Hang Yan, Junqi Dai, Xipeng Qiu, Zheng Zhang, et al. 2021. A unified generative framework for aspect-based sentiment analysis. [arXiv preprint arXiv:2106.04300](#).
- Li Yang, Jin-Cheon Na, and Jianfei Yu. 2022. [Cross-Modal Multitask Transformer for End-to-End Multimodal Aspect-Based Sentiment Analysis](#). *Information Processing & Management*, 59(5):103038.
- Jianfei Yu and Jing Jiang. 2019. [Adapting bert for target-oriented multimodal sentiment classification](#). In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 5408–5414. International Joint Conferences on Artificial Intelligence Organization.
- Jianfei Yu, Jing Jiang, and Rui Xia. 2019. Entity-sensitive attention and fusion network for entity-level multimodal sentiment classification. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:429–439.
- Jianfei Yu, Jing Jiang, Li Yang, and Rui Xia. 2020. [Improving multimodal named entity recognition via entity span detection with unified multimodal transformer](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3342–3352, Online. Association for Computational Linguistics.
- Li Yuan, Jin Wang, Liang-Chih Yu, and Xuejie Zhang. 2020. [Graph attention network with memory fusion for aspect-level sentiment analysis](#). In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 27–36, Suzhou, China. Association for Computational Linguistics.
- Qi Zhang, Jinlan Fu, Xiaoyu Liu, and Xuanjing Huang. 2018. Adaptive co-attention network for named entity recognition in tweets. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

## ACL 2023 Responsible NLP Checklist

---

### A For every submission:

- A1. Did you describe the limitations of your work?  
*We discuss it in section Limitations.*
- A2. Did you discuss any potential risks of your work?  
*Our research is foundational research and not tied to particular applications.*
- A3. Do the abstract and introduction summarize the paper's main claims?  
*In the abstract and section 1 Introduction.*
- A4. Have you used AI writing assistants when working on this paper?  
*Left blank.*

### B Did you use or create scientific artifacts?

*In section 3 Methodology and 4 Experiment.*

- B1. Did you cite the creators of artifacts you used?  
*In section 3.1 Overview and 4 Experiment.*
- B2. Did you discuss the license or terms for use and / or distribution of any artifacts?  
*The datasets used in section 4 and pre-trained models in section 3 are in public domain and licensed for research purposes.*
- B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?  
*In section 4 Experiment.*
- B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?  
*Not applicable. The data is in public domain and licensed for research purposes.*
- B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?  
*In section 4 Experiment.*
- B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.  
*In section 4 Experiment.*

### C Did you run computational experiments?

*In section 4 Experiment.*

- C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?  
*We only use the most commonly used pre-trained models and the parameters or GPU hours are not focus of our research.*

*The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.*



- C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?

*In section 4 Experiment.*

- C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?

*In section 4 Experiment.*

- C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?

*In section 4 Experiment.*

**D  Did you use human annotators (e.g., crowdworkers) or research with human participants?**

*Left blank.*

- D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?

*No response.*

- D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?

*No response.*

- D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?

*No response.*

- D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?

*No response.*

- D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?

*No response.*