


Aspect-Aware Affective Focus Network For Joint Multimodal Aspect Sentiment Analysis

Xiangbo Ji[†], Haoyu Shi[†], Wei Wu , Na Li[#], and Jinyang Wang[#]

Department of Computer Science, Inner Mongolia University, Hohhot, China
{jixiangbo, shihaoyu, leena, wangjinyang}@mail.imu.edu.cn,
cswuwei@imu.edu.cn

Abstract. Joint Multimodal Aspect Sentiment Analysis (JMASA) aims to jointly extract aspect terms and their associated sentiments from given text-image pairs. Existing methods focus on associating the entire image with the corresponding text. However, redundant information in the image introduces noise, hindering the highlighting of crucial affective visual regions. Additionally, simply utilizing attention mechanisms to adaptively search for the associated sentiment between aspects in a sentence may ignore that sentiment judgments can be easily interfered with by other irrelevant words. To address these challenges, we propose a novel Aspect-Aware Affective Focus Network (AAFN) for multimodal sentiment analysis. Specifically, our model contains an aspect-aware enhancement module that is sensitive to aspect-related semantic information based on syntactic structure and part-of-speech information. Furthermore, we introduce a candidate affective visual focus module that precisely identifies candidate visual sentiment regions under linguistic guidance. To effectively eliminate the semantic gap, we introduce a language-guided fusion module to achieve fine-grained interaction between visual focus and aspect-related information, thereby enhancing the relevance between image-text pairs. Extensive experiments conducted on two benchmark datasets, Twitter-2015 and Twitter-2017, demonstrate the effectiveness of our proposed method.

Keywords: candidate affective visual focus · aspect-aware enhancement · multimodal sentiment analysis.

1 INTRODUCTION

With the rapid development of social media, multimedia data featuring multiple modalities, such as text, and images have experienced explosive growth on platforms like Facebook and Twitter. Its wide applications include helping businesses evaluate user reactions to their products and aiding governments in understanding public attitudes on controversial issues [11]. In our paper, we focus on Joint Multimodal Aspect-Sentiment Analysis, defined as identifying the aspect terms and their associated sentiment polarities in text-image pair.


[†]Equal contribution.  Corresponding author. [#]Co-third author.



Fig. 1: An example of the JMASA task.

In this scenario of fine-grained JMASA task, existing methods measure the relation score between the entire image and text, and then assign lower contribution weights to images with lower similarity for multimodal fusion [6]. However, an image typically contains some sentiment-irrelevant information. Simple overall multimodal fusion may introduce redundant visual information, which hinders the effective highlighting of key affective visual regions. Following this, recent works have focused on extracting salient visual objects [4, 8, 17, 19] by leveraging object detectors to analyze the entire image. Although this approach eliminates some redundant patches, it restricts performance to the quality of object detectors and inadvertently filters out valuable private information within the visual content. As shown in Fig.1, it is sufficient to judge the sentiment of **Kevin Love** based on facial expressions alone, excluding background, clothing, and other information. In addition, for aspect-related information awareness, most of the aforementioned studies utilize attention mechanisms to adaptively search for the associated dependencies between words in a sentence [6, 7, 9, 13, 21]. Others employ a textual aspect-sentiment extraction module to leverage an auxiliary end-to-end textual ABSA task to learn the aspect-oriented text representation [15, 16]. However, these methods ignore that phrases should be considered as a whole and sentiment judgments can be easily interfered with by other irrelevant words. As shown in Fig.1, the aspect words **GameOfThrone** can be easily mispredicted as having a neutral sentiment if the dependency relationships corresponding to the context of **beating** in the sentence are not considered.

To address the above challenges, we propose an Aspect-Aware Affective Focus Network (AAFN), which consists of three components: the Aspect-Aware Enhancement module (AAM), the Candidate Affective Visual Focus module (CVM), and the Language-Guided Fusion module (LFM). Specifically, for each text-image pair, we utilize the AAM to make the model more sensitive to aspect-related information by leveraging syntactic structure and the part-of-speech information. Moreover, to reduce noise from redundant image patches and better highlight affective visual regions, we utilize the CVM to enhance the model’s capability to capture crucial emotional visual focus. Finally, we introduce the LFM to reduce the semantic gap across the two modalities and utilize cross-modal attention to guide the fusion of fine-grained visual features based on text features. The contributions can be summarized as follows:

- We propose an aspect-aware visual focus interaction network(AAFN), which achieves precise correspondence between aspect-related information and visual focus for the joint multimodal aspect-sentiment analysis.

Aspect-Aware Affective Focus Network

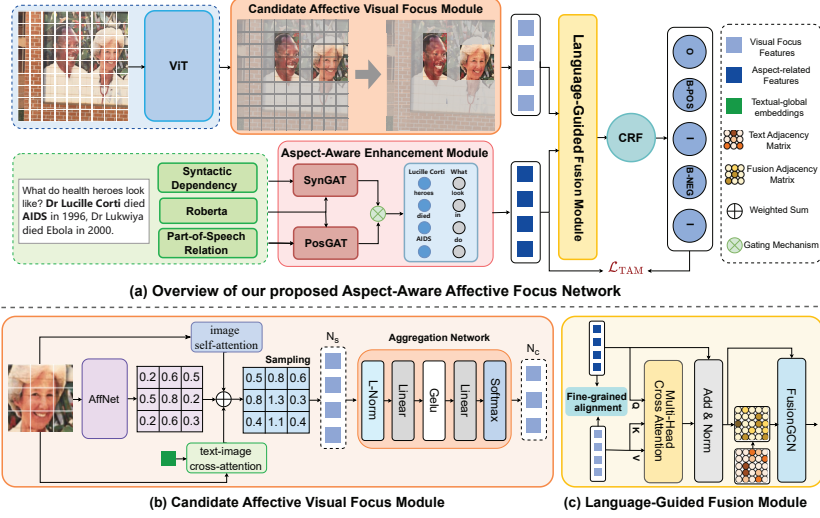


Fig. 2: The framework of our proposed AAFN.

- We introduce an aspect-aware enhancement module, which combines the syntactic information learned from the dependency tree with the part-of-speech information learned from constructed lexical relations to enhance the model’s sensitivity to aspect-related information.
- We design a novel candidate affective visual focus module, which identifies visual patches under linguistic supervision to efficiently address the resultant issue of visual patch redundancy and patch ambiguity.
- We present a language-guided fusion module, which leverages linguistic information to guide the extraction of visual information and performs fine-grained fusion while reducing the semantic gap between different modalities.

2 Related Work

As an emerging research area, joint multimodal aspect-based sentiment analysis is gaining significant practical importance and attracting increasing attention. In recent years, many advanced methods of joint multimodal analysis has emerged. [6] is the first proposed JMASA task and introduces an auxiliary cross-modal relation detection mechanism to control the use of visual information in predictions. [16] extracts image features and employs these features to predict noun-adjective pairs, while [15] employs a novel Aesthetic Assessment Module to discern the aesthetic attributes of images, thereby enhancing the quality of the image features. [13] uses a contrastive learning approach for JMASA, which is enhanced by an auxiliary momentum strategy to align representations of multimodal data before fusion. The recent advancement in this field, [21] innovatively utilizes a graph neural network to model the correlation between images and texts. [9] apply mutual information to reduce intra-modal feature noise and inter-modal semantic gap.

3 Methodology

3.1 Task Definition

Given an input sample X containing a sentence S and an image I , the JMASA task aims to predict a label sequence $\mathbf{y} = (y_1, y_2, \dots, y_m)$, where m is the length of S . Besides, $y_i \in \{\text{B-POS}, \text{B-NEU}, \text{B-NEG}, \text{I}, \text{O}\}$, where 'B' represents the beginning of the aspect word, 'I' indicates the end of the aspect word, and 'POS', 'NEU', and 'NEG' represent the sentiment towards the aspects.

3.2 Aspect-Aware Enhancement Module

The contextualized representations from current pre-trained encoders lack sensitivity to specific aspects. Additionally, the sentiment of an aspect can be interfered with by irrelevant words within a sentence. To address this, we design an Aspect-Aware Enhancement Module, which is intended to leverage syntactic and part-of-speech information for capturing aspect-related semantic information.

Part-of-speech and Syntactic Graph Construction. To characterize aspect and opinion term formation patterns, we construct a part-of-speech graph $G^{\text{Pos}} = (V, R^{\text{Pos}})$ using the spaCy library. Each edge $R_{i,j}^{\text{Pos}}$ between words i and j is based on their part-of-speech tags, with representation $r_{i,j}^p \in \mathbb{R}^{d_p}$, where d_p is the embedding dimension. Additionally, we build a syntactic dependency graph $G^{\text{Syn}} = (V, R^{\text{Syn}})$ via dependency parsing tree. The syntactic relation $R_{i,j}^{\text{Syn}}$ between words w_i and w_j is represented as $r_{i,j}^s \in \mathbb{R}^{d_s}$, where d_s is the dimension of syntactic relation embeddings.

High-order Feature Fusion. For sentence S , we employ a pre-trained Roberta [10] as the text extractor to obtain the text embeddings $\mathbf{h} \in \mathbb{R}^{m \times d}$. To capture diverse linguistic features and high-order word interactions, we employ two graph-based attentional modules: SynGAT for syntactic dependency graphs and PosGAT for part-of-speech graphs. These modules effectively distinguish syntactic and part-of-speech relationships when computing attention weights. Specifically, for the i -th node, the update process is as follows:

$$h_i^{\text{syn}}(l) = \parallel_{z=1}^Z \text{Sigmoid} \left(\sum_{j \in N(i)} \hat{\alpha}_{i,j}^{l,z} \left(W_s^{l,z} h_j^{\text{syn}}(l-1) + W_{s2}^{l,z} r_{i,j}^s \right) \right) \quad (1)$$

$$h_i^{\text{pos}}(l) = \parallel_{z=1}^Z \text{Sigmoid} \left(\sum_{j \in N(i)} \hat{\beta}_{i,j}^{l,z} \left(W_p^{l,z} h_j^{\text{pos}}(l-1) + W_{p2}^{l,z} r_{i,j}^p \right) \right) \quad (2)$$

where $W_{s2}^l \in \mathbb{R}^{\frac{d}{z} \times d}$ and $W_{p2}^l \in \mathbb{R}^{\frac{d}{z} \times d}$ are parameter matrices. z denotes the number of attention heads. N_i is the set of immediate neighbors of node i . $\hat{\alpha}_{i,j}^{(l,z)}$, $\hat{\beta}_{i,j}^{(l,z)}$ are the normalized attention coefficients for the z -th head at the l -th layer.

To fuse syntactic dependency and part-of-speech relation features, we introduce a gating mechanism that merges these two perspectives to obtain the final aspect-related representation $\mathbf{h}^a = (h_1^a, h_2^a, \dots, h_m^a) \in \mathbb{R}^{m \times d}$ as follows:

$$\mathbf{h}^a = g \circ \mathbf{h}^{\text{syn}} + (1 - g) \circ \mathbf{h}^{\text{pos}} \quad (3)$$

where the gating vector $g = \text{sigmoid}(W_g [h^{\text{syn}} : h^{\text{pos}}] + b_g)$, \circ denotes the element-wise product, $[\cdot]$ represents concatenation, and W_g , b_g are trainable parameters.

The final aspect-related representation \mathbf{h}^a is then passed through a softmax layer to predict the auxiliary label for the i -th token. After obtaining predicted probability distribution \mathbf{p}^t , we utilize the standard cross-entropy loss to optimize the aspect-aware enhancement task:

$$\mathcal{L}_{\text{AAM}} = \frac{1}{K} \sum_{j=1}^m \sum_{i=1}^{n_j} \log(p_i^t, r_i) \quad (4)$$

where K and n_j denote the number of samples, tokens in the j -th sample. r_i indicates the auxiliary label for the i -th token.

3.3 Candidate Affective Visual Focus Module

Since previous research associates the entire image with corresponding aspects, introducing redundant information hinders the highlighting of affective visual regions. To address this issue, we introduce a Candidate Affective Visual Focus (CVM) module to effectively capture candidate affective visual regions and aggregate them into an optimal semantic representation.

Affective Patch Selection. For an image I , we can obtain a set of visual patch features $\mathbf{v} = (v_1, v_2, \dots, v_n) \in \mathbb{R}^{n \times d}$ from the ViT [2] encoder, where n denotes the number of non-overlapping patches. Subsequently, we incorporate spatial information from images into patch features and use an Affective Score Prediction Network (AffNet) to obtain the affective scores $\mathbf{a}^p = (a_1^p, a_2^p, \dots, a_n^p)$.

$$a_i^p = \text{Sigmoid}(\text{FFN}(v_i)) \quad (5)$$

where $a_i^p \in [0, 1]$ is the visual affective score corresponding to the i -th patch v_i .

Relying solely on AffNet for visual focus is inaccurate without linguistic supervision. Therefore, we compute cross-attention between visual patches and aspect representations to derive aspect-relevant scores \mathbf{a}^r , and apply self-attention within visual patches to obtain image affective scores \mathbf{a}^s .

$$a_i^r = \text{Norm}(v_i^\top \cdot h_{\text{glo}}/d), \quad a_i^s = \text{Norm}(v_i^\top \cdot v_{\text{glo}}/d) \quad (6)$$

where Norm represents the normalization of attentive scores into a 0-1 range, aligning them consistently with the affective scores a_i^p . And $v_{\text{glo}}, h_{\text{glo}}$ represent the global visual/textual embeddings, obtained by applying average pooling to the patch/word features. We combine the above scores to calculate the final significant affective visual score $\mathbf{s} \in \mathbb{R}^N$, using γ as a weighting parameter.

$$s_i = (1 - \gamma)a_i^p + \gamma(a_i^s + a_i^r) \quad (7)$$

Then, we employ the Gumbel-Softmax technique [12] to provide a smooth and differentiable sampling process, and apply arg-max operation to derive a differentiable decision matrix $\mathbf{D} = \text{Sampling}(\mathbf{G})_{*,1} \in \{0, 1\}$, where \mathbf{G} denotes the Gumbel-Softmax matrix, '1' indicates an aspect-relevant patch, and '0' signifies a redundant patch.

Affective Patch Aggregation. The selected affective patches are denoted as $\mathbf{v}^s = \{v_1^s, \dots, v_{N_s}^s\} \in \mathbb{R}^{N_s \times d}$, where N_s is the number of candidate affective

patches. We then utilize an aggregation network to transform these patches into optimal affective semantic patches.

$$h_j = \sum_{i=1}^{N_s} (\mathbf{W})_{ij} \cdot v_i \quad (8)$$

where $\mathbf{W} = \text{Softmax}(\text{MLP}(\mathbf{v}^s)) \in \mathbb{R}^{N_s \times N_c}$ is the elements of the normalized weight matrix, $j \in [1, N_c]$ and N_c represents the number of aggregated affective patches. Although redundant patches could be directly discarded, they may contain overall sentiment tendency, hence we fuse them into one patch p_r .

$$h_r = \sum_{i \in Z} \hat{a}_i \cdot v_i, \quad \hat{a}_i = \frac{\exp(s_i) D_i}{\sum_{j=1}^n \exp(s_j) D_j} \quad (9)$$

where Z denotes the index set for redundant patches, \hat{a}_i signifies the normalized weights based on the aspect-relevant affective score s_i . Eventually, we obtain the final candidate visual focus patches $\mathbf{h}^p = \{h_1^p, h_2^p, \dots, h_{N_c}^p, h_r^p\}$.

3.4 Language-Guided Fusion Module

Since the candidate visual focus features and the aspect token embeddings reside in their own modal spaces. To eliminate the above semantic gap and enhance the relevance between image-text pairs, we design a language-guided fusion Module (LFM), which infuses visual private affective information into aspect-related information after achieving fine-grained alignment between the candidate’s visual focus and the aspect-related information.

Aspect-Patch Fine-grained Alignment. The similarity score $S(\mathbf{h}^p, \mathbf{h}^a)$ measures the semantic similarity between the aspect terms \mathbf{h}^a and the visual focus patches \mathbf{h}^p . It is determined by calculating the maximum similarity between each visual focus embedding and all aspect-related embeddings in bi-directions.

$$S(\mathbf{h}^p, \mathbf{h}^a) = \sum_{i=1}^N \max_{j=1}^M \langle h_i^p, h_j^a \rangle + \sum_{j=1}^M \max_{i=1}^N \langle h_i^p, h_j^a \rangle \quad (10)$$

where M, N denote the numbers of visual and text embeddings in the i -th sample pair. To enhance the relevance between image-text pairs, we utilize a hinge-based triplet loss [3] to achieve fine-grained alignment between the candidate visual focus feature \mathbf{h}^p and the aspect-related feature \mathbf{h}^a .

$$\mathcal{L}_{FGA} = \frac{1}{B} \sum_{(h^p, h^a) \in B} \left[S(h^p, \hat{h}^a) - S(h^p, h^a) + w \right]_+ + \left[S(\hat{h}^p, h^a) - S(h^p, h^a) + w \right]_+ \quad (11)$$

where w is a margin factor and $[x]_+ = \max(x, 0)$. \hat{h}^p and \hat{h}^a are the hardest negative samples in the given batch corresponding to (a, p) .

Aspect-Guided Fusion. To utilize aligned candidate visual focus information to improve the ability to classify the sentiment of aspects, we feed the aspect-related feature \mathbf{h}^a and the candidate visual focus feature \mathbf{h}^p into a cross-modal attention model [14], thus obtaining the initial multimodal representation $\mathbf{h}^f = \{h_1^f, h_2^f, \dots, h_N^f\}$. The formula is as follows:

$$\text{Cross-Attn}(Q_a, K_p, V_p) = \text{softmax} \left(\frac{Q_a K_p^T}{\sqrt{d_c}} \right) V_p \quad (12)$$

$$\mathbf{h}^f = \text{Norm}(\text{Cross-Attn}(Q_a, K_p, V_p) W_O) \quad (13)$$

where Q_a, K_p, V_p represents the query, key, and value obtained by projecting the aspect-related features \mathbf{h}^a and visual focus \mathbf{h}^p . Norm denotes L2 normalization. The parameter matrix W_O is a projection weight matrix.

To achieve the precise text-image interaction, we build a fusion graph on the fusion features \mathbf{h} and aggregate related features through the edges in the fusion graph. We calculate the similarity \mathbf{S}^f between elements of the fused features:

$$s_{i,j}^f = \cos(h_i^f, h_j^f) \quad (14)$$

where $s_{i,j}^f$ represents the similarity between the i -th node and the j -th node in the fusion features. Then, we compute the Hadamard product of the text graph's adjacency matrix \mathbf{A}^t with the similarity matrix \mathbf{S}^f and perform L_2 normalization to obtain the adjacency matrix $\mathbf{A}^f = \|\mathbf{A}^t \circ \mathbf{S}^f\|_2 \in \mathbb{R}^{N \times N}$ of the fusion graph.

In FusionGCN, we use the fused features \mathbf{h}^f as the initial nodes and allow relevant node information to interact through a graph convolution operation:

$$h_i^{(k)} = \sigma \left(\sum_{h_j \in \mathcal{N}_i} A_{ij}^f W_f^{(k)} h_j^{(k-1)} + b^{(k)} \right) \quad (15)$$

where $h_i^{(k)}$ is the hidden representation of the i -th node in the k -th layer of the fusion graph. After K^f layer graph convolution, we obtain the final graph-based fusion representation $\mathbf{H}^{f'} = \{h_1^{f'}, h_2^{f'}, \dots, h_N^{f'}\}$.

3.5 Prediction and Loss Function

Since Conditional Random Fields (CRF) evaluate correlations between labels in neighbourhoods and score the entire sequence, effectively capturing aspect-label dependencies, we feed the graph-based fusion representation $\mathbf{H}^{f'}$ into a CRF layer to predict the label sequence \mathbf{y} as follows:

$$P(\mathbf{y}) = \frac{\exp(\text{score}(\mathbf{H}^{f'}, \mathbf{y}))}{\sum_{\mathbf{y}' \in \mathcal{Y}_H} \exp(\text{score}(\mathbf{H}^{f'}, \mathbf{y}'))} \quad (16)$$

$$\text{score}(\mathbf{H}^{f'}, \mathbf{y}) = \sum_{i=0}^n T_{y_i, y_{i+1}} + \sum_{i=1}^n \mathbf{W}^{y_i} \cdot \mathbf{H}_i^{f'} \quad (17)$$

where $T_{i,j}$ is the transition score from label i to j , $\text{score}(\mathbf{H}^{f'}, \mathbf{y})$ indicate the score function of the label sequence \mathbf{y} under the features $\mathbf{H}^{f'}$, \mathbf{W}_{y_i} is the weight vector for y_i , and \mathcal{Y}_H represents all possible label sequences for input tokens.

The main training loss function is to minimize the negative log-probability of the correct label sequence as follows:

$$\mathcal{L}_{\text{JMASA}} = -\frac{1}{K} \sum_{j=1}^N \left(\text{score}(\mathbf{H}^{f'}, y_j) \right) - \log \sum_{y'_j \in \mathcal{Y}_{H_j}} \exp \left(\text{score}(\mathbf{H}^{f'}, y'_j) \right) \quad (18)$$

where K is the number of samples, y_j is the j -th sample's correct label sequence, and \mathcal{Y}_{H_j} represents all possible label sequences for input tokens.

Table 1: Results of different methods for JMASA on the two Twitter datasets.

	Methods	Twitter 2015			Twitter 2017		
		P	R	F1	P	R	F1
Text Only	SPAN [5]	53.7	53.9	53.8	59.6	61.7	60.6
	D-GCN [1]	58.3	58.8	59.4	64.2	64.1	64.1
	Roberta [10]	61.8	65.3	63.5	65.5	66.9	66.2
Multimodal	JML [6]	65.0	63.2	64.1	66.5	65.5	66.0
	CMMT [16]	64.6	68.7	66.5	67.6	69.4	68.5
	VLP [8]	65.1	68.3	66.6	66.9	69.2	68.0
	MOCOLNet [13]	66.3	67.9	67.1	67.3	68.7	68.0
	M2DF [20]	67.0	68.3	67.6	67.9	68.8	68.3
	AoM [21]	<u>67.9</u>	69.3	<u>68.6</u>	68.4	<u>71.0</u>	69.7
	Atlantis [15]	65.6	69.2	67.3	68.6	70.3	69.4
	RNG [9]	67.8	<u>69.5</u>	<u>68.6</u>	<u>69.5</u>	<u>71.0</u>	<u>70.2</u>
	Our	68.1	70.0	69.0	70.3	71.5	70.9

Table 2: Comparison of different module ablations on both datasets.

Components			Twitter-2015			Twitter-2017		
AAM	CVM	LFM	P	R	F1	P	R	F1
✓			62.8	65.9	64.3	65.9	67.5	66.7
	✓		65.4	67.7	66.5	66.8	68.2	67.5
		✓	64.6	67.9	66.2	66.1	67.5	66.8
✓	✓		68.1	69.7	68.9	69.5	70.2	69.9
✓		✓	66.5	68.3	67.4	67.9	68.8	68.4
	✓	✓	67.4	68.9	68.1	68.3	69.7	69.0
✓	✓	✓	68.1	70.0	69.0	70.3	71.5	70.9

To optimize our model’s parameters, we integrate the loss functions from our main task and two auxiliary tasks:

$$\mathcal{L} = \mathcal{L}_{\text{JMASA}} + \alpha \cdot \mathcal{L}_{\text{AAM}} + \beta \cdot \mathcal{L}_{\text{FGA}} \quad (19)$$

where α and β are trade-off hyperparameters.

4 EXPERIMENTS

4.1 Datasets and Evaluation Metrics

We validate the effectiveness of our proposed method on Twitter-2015 and Twitter-2017 datasets [18]. In the two multimodal datasets, Twitter-2015 consists of 2101 training, 727 development, and 674 test samples, while Twitter-2017 contains 1746 training, 577 development, and 587 test samples. Each aspect is labelled with three sentiment categories: positive, neutral, and negative in both datasets. And we use Precision (P), Recall (R), and F1-score (F1) as evaluation metrics.

4.2 Implementation Details

The learning rate for both datasets is configured at $3e-5$. We have configured the hyper-parameters α at 0.4, β at 0.6, and γ at 0.5. Other hyper-parameters, such as text embedding dimension d , the number of attention heads, and temperature τ , have values of 768, 12, and 0.07, respectively. The selection/aggregation ratio is set to 0.5/0.4. Furthermore, the maximum sentence length is fixed at 128, and the training is 150 epochs in both datasets. All experiments are conducted on a Tesla V100 GPU, by using the PyTorch-1.9.0 library.

4.3 Performance Comparison

In this subsection, we compare our proposed method with two groups of competitive models, as shown in Table 1. These benchmark methods can be broadly divided into two categories: The first category is the Text-only method, including SPAN, D-GCN, and Roberta. The second category is the Multimodal method, including JML [6], CMMT [16], VLP [8], MOCOLNet [13], M2DF [20], AoM [21], Atlantis [15], and RNG [9]. Our method AAFN significantly outperforms all text-based models, indicating that the integration of richer visual and textual information in our model is highly effective. Besides, methods like VLP and MOCOLNet, which utilize significant regions in images as visual information, outperform others like JML and CMMT which use the entire image. This implies that noise in images can impede sentiment analysis, and leveraging fine-grained information can help mitigate this issue. For multimodal methods, our model outperforms the previous best model (RNG) by improving the F1-score by 0.4% and 0.7% on Twitter-2015 and Twitter-2017, respectively. This is due to our method’s focus on affective visual regions and its emphasis on the fine-grained interaction between the visual focus and aspect term.

4.4 Ablation Study

We conduct a thorough ablation study to verify the effectiveness of the proposed components in our methods, as shown in Table 2. For the first group, only a single component is used for enhancing the sentiment analysis process. We can observe that the proposed CVM module shows significant improvements, highlighting the crucial role of capturing the candidate’s visual focus. For the second group, we pair different components to investigate their interrelated effects. We can observe that the most significant improvement is achieved in the integration of the CVM and AAM modules. This may be attributed to the fine-grained candidate visual focus and aspect-related information, which strengthens the relevance of image-text pairs and improves sentiment analysis performance.

5 Conclusion

In this paper, we present an Aspect-Aware Affective Focus Network for the JMASA task. We design a candidate affective visual focus module to efficiently address the resultant issues of visual patch redundancy and patch ambiguity that hinder highlighting the affective visual region. Additionally, we combine the syntactic information learned from the dependency tree with the part-of-speech information learned from constructed lexical relations to enhance the model’s sensitivity to aspect-related information. Comprehensive experiments show that our model outperforms state-of-the-art methods. In the future, we will further explore how to mine sentiment-related information from different modalities.

Acknowledgments. This work is supported by the Inner Mongolia Natural Science Foundation Project No.2024MS06007.

References

1. Chen, G., Tian, Y., Song, Y.: Joint aspect extraction and sentiment analysis with directional graph convolutional networks. In: COLING. pp. 272–279 (2020)
2. Dosovitskiy, A., Beyer, L., Kolesnikov, A.: An image is worth 16x16 words: Transformers for image recognition at scale. In: ICLR (2020)
3. Faghri, F., Fleet, D.J., Kiros, J.R., Fidler, S.: Vse++: Improving visual-semantic embeddings with hard negatives. In: British Machine Vision Conference (2018)
4. Fu, Z., Zhang, L., Xia, H.: Linguistic-aware patch slimming framework for fine-grained cross-modal alignment. In: CVPR. pp. 26307–26316 (2024)
5. Hu, M., Peng, Y., Huang, Z., Li: Open-domain targeted sentiment analysis via span-based extraction and classification. In: ACL. pp. 537–546 (2019)
6. Ju, X., Zhang, D., Xiao, R., Li, J., Li: Joint multi-modal aspect-sentiment analysis with auxiliary cross-modal relation detection. In: EMNLP. pp. 4395–4405 (2021)
7. Li, P., Li, P., Zhang, K.: Dual-channel span for aspect sentiment triplet extraction. In: Bouamor, H., Pino, J., Bali, K. (eds.) EMNLP. pp. 248–261 (2023). <https://doi.org/10.18653/v1/2023.emnlp-main.17>
8. Ling, Y., Yu, J., Xia, R.: Vision-language pre-training for multimodal aspect-based sentiment analysis. In: ACL. pp. 2149–2159 (2022)
9. Liu, Y., Zhou, Y., Li, Z., Zhang, J.: Rng: Reducing multi-level noise and multi-grained semantic gap for joint multimodal aspect-sentiment analysis. ICME (2024)
10. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer: Roberta: A robustly optimized bert pretraining approach. ICLR (2020)
11. Liu, Z., Zhou, B., Chu: Modality translation-based multimodal sentiment analysis under uncertain missing modalities. Information Fusion **101**, 101973 (2024)
12. Maddison, C.J., Mnih, A., Teh, Y.W.: The concrete distribution: A continuous relaxation of discrete random variables. In: ICLR (2022)
13. Mu, J., Nie, F., Wang, W.: Mocolnet: A momentum contrastive learning network for multimodal aspect-level sentiment analysis. TKDE pp. 1–14 (2023)
14. Tsai, Y.H.H., Bai, S., Liang, P.P., Kolter, J.Z., Morency: Multimodal transformer for unaligned multimodal language sequences. In: ACL. vol. 2019, p. 6558 (2019)
15. Xiao, L., Wu, X., Xu, J.: Atlantis: Aesthetic-oriented multiple granularities fusion network for joint multimodal aspect-based sentiment analysis. Information Fusion **106**, 102304 (2024)
16. Yang, L., Na, J.C., Yu, J.: Cross-modal multitask transformer for end-to-end multimodal aspect-based sentiment analysis. IPM **59**(5), 103038 (2022)
17. Yu, J., Chen, K., Xia, R.: Hierarchical interactive multimodal transformer for aspect-based multimodal sentiment analysis. IEEE Transactions on Affective Computing **14**(3), 1966–1978 (2022)
18. Yu, J., Jiang, J.: Adapting bert for target-oriented multimodal sentiment classification. In: IJCAI-19. pp. 5408–5414 (7 2019)
19. Yu, J., Wang, J., Xia, R., Li, J.: Targeted multimodal sentiment classification based on coarse-to-fine grained image-target matching. In: IJCAI. pp. 4482–4488 (2022)
20. Zhao, F., Li, C., Wu, Z., Ouyang, Y., Zhang, J., Dai, X.: M2df: Multi-grained multi-curriculum denoising framework for multimodal aspect-based sentiment analysis. In: EMNLP 2023. pp. 9057–9070 (2023)
21. Zhou, R., Guo, W., Liu, X.: Aom: Detecting aspect-oriented information for multimodal aspect-based sentiment analysis. In: ACL. pp. 8184–8196 (2023)