Learning from Adjective-Noun Pairs: A Knowledge-enhanced Framework for Target-Oriented Multimodal Sentiment Classification

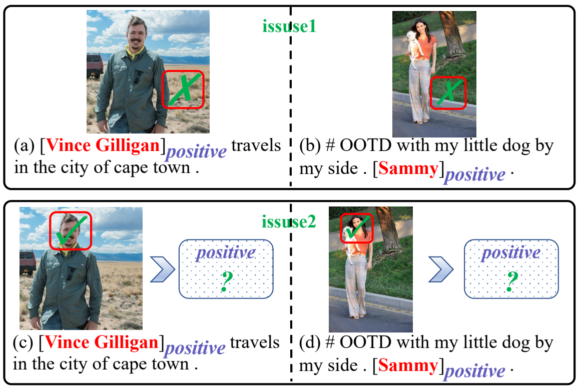
Fei Zhao Zhen Wu[*] Siyu Long Xinyu Dai Shujian Huang Jiajun Chen
National Key Laboratory for Novel Software Technology, Nanjing University, China
Collaborative Innovation Center of Novel Software Technology and Industrialization, China
{zhaof, longsy}@smail.nju.edu.cn
{wuz, daixinyu, huangsj, chenjj}@nju.edu.cn

Abstract

Target-oriented multimodal sentiment classification (TMSC) is a new subtask of aspect-based sentiment analysis, which aims to determine the sentiment polarity of the opinion target mentioned in a (sentence, image) pair. Recently, dominant works employ the attention mechanism to capture the corresponding visual representations of the opinion target, and then aggregate them as evidence to make sentiment predictions. However, they still suffer from two problems: (1) The granularity of the opinion target in two modalities is inconsistent, which causes visual attention sometimes fail to capture the corresponding visual representations of the target; (2) Even though it is captured, there are still significant differences between the visual representations expressing the same mood, which brings great difficulty to sentiment prediction. To this end, we propose a novel Knowledge-enhanced Framework (*KEF*) in this paper, which can successfully exploit adjective-noun pairs extracted from the image to improve the visual attention capability and sentiment prediction capability of the TMSC task. Extensive experimental results show that our framework consistently outperforms state-of-the-art works on two public datasets.

1 Introduction

Target-oriented multimodal sentiment classification (TMSC) is a new sub-task of aspect-based sentiment analysis (Pang et al., 2008; Liu, 2012; Pontiki et al., 2014), which aims to predict the sentiment polarity of the opinion target mentioned in a pair of sentence and image. The assumption behind this task is that the image information can help the text content identify the sentiment of the opinion target. Fig. 1(a) and Fig. 1(b) show two representative examples. It is hard to detect the sentiment of the opinion target (i.e., "*Vince Gilligan*" or "*Sammy*") only depending on informal sentences, but the visual content (i.e., smiling face) associated with the target can clearly reflect its sentiment polarity.



(a) [Vince Gilligan]$_{positive}$ travels in the city of cape town .

(b) # OOTD with my little dog by my side . [Sammy]$_{positive}$.

(c) [Vince Gilligan]$_{positive}$ travels in the city of cape town .

(d) # OOTD with my little dog by my side . [Sammy]$_{positive}$.

Figure 1: Two examples of Target-Oriented Multimodal Sentiment Classification (TMSC). Opinion targets and their corresponding sentiment polarities are highlighted in the sentence. The red bounding box denotes the visual clues that the opinion target focuses on.

From the above examples, we can conclude that aligning the opinion target of two modalities and capturing helpful visual sentiment features play a critical role in the TMSC task. Given its importance, dominant works employ the attention mechanism (Bahdanau et al., 2015) to automatically learn alignment of text and image, and then aggregate the captured the visual representations of the opinion target as auxiliary evidence to make sentiment predictions (Xu et al., 2019b; Yu et al., 2019; Yu and Jiang, 2019; Zhou et al., 2021; Zhang et al., 2021; Wang et al., 2021).

Despite achieving some improvements, the aforementioned methods still suffer from two key problems: (i). These methods easily fail to align two modalities because of the granularity gap of opinion target across text and image. Specifically, the opinion target appearing in the image often refers to a coarse-grained object concept (e.g., the `man` in Fig. 1(a)), while corresponding opinion target in the sentence are usually a fine-grained entity (e.g.,

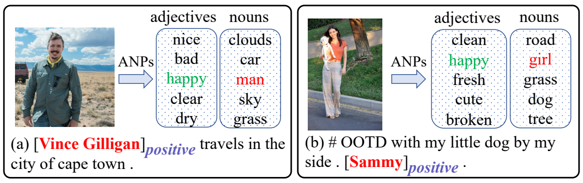
[*] Corresponding author.

Proceedings of the 29th International Conference on Computational Linguistics, pages 6784–6794
October 12–17, 2022.

the man's name [1] "*Vince Gilligan*"). The inconsistency of target granularity causes visual attention sometimes fail to capture the corresponding visual representations. (ii). Even though it is captured, diversified visual representations expressing the same mood also bring challenges for sentiment prediction. Take Fig.1(c) and Fig.1(d) as an example, the opinion target "*Vince Gilligan*" and "*Sammy*" separately focus on the coarse-grained object concepts man and girl in the image, and from their facial expressions we can tell that they are smiling, but the angle and magnitude of the smile are quite different. The variety of visual representations inevitably leads to its sparsity, which makes it hard to learn the accurate mapping function between visual representations and sentiment labels.

In this work, we provide a new idea to tackle the above problems, i.e., exploiting adjective-noun pairs (ANPs) (Borth et al., 2013) extracted from images (e.g., "*nice clouds*", "*bad car*", "*happy man*", "*clear sky*" and "*dry grass*" in Fig. 2(a)). For the first issue, we observed that the nouns of ANPs are also coarse-grained concepts, so an intuitive idea is to map a fine-grained opinion target (e.g. "*Vince Gilligan*") to a coarse-grained noun (e.g. "*man*"[2]) in ANPs. In this manner, it is easier to bridge the granularity gap of two modalities and align text and image. For the second issue, we observed that ANPs can usually extract the same adjectives from different visual content expressing the same mood, so an intuitive idea is to map diversified visual representations (e.g., smiling faces) to the same adjective (e.g., "*happy*"[3]). Apparently, it is easier to learn the mapping function between these same adjectives and sentiment labels.

To facilitate the TMSC task with ANPs, we propose a novel Knowledge-enhanced Framework (*KEF*), which consists of two components: Visual Attention Enhancer and Sentiment Prediction Enhancer. Specifically, the former first finds the noun most related to the opinion target from ANPs with our designed mapping method, and then uses it to improve the effectiveness of visual attention. The latter aims to build the connection between the adjective and target-relevant visual representations, and then utilizes it as the complementary information of visual representations to reduce the difficulty of predicting sentiment labels.



(a) [**Vince Gilligan**]$_{positive}$ travels in the city of cape town .

(b) # OOTD with my little dog by my side . [**Sammy**]$_{positive}$.

Figure 2: Extract Top-5 adjective-noun pairs (ANPs) from each image in our Twitter datasets.

Our contributions are summarized as follows:

- To the best of our knowledge, we are the first to propose leveraging adjective-noun pairs (ANPs) extracted from the image to help align text and image in the TMSC task.
- We propose a novel Knowledge-enhanced Framework (*KEF*), which contains a Visual Attention Enhancer to improve the effectiveness of visual attention, and a Sentiment Prediction Enhancer to reduce the difficulty of sentiment prediction.
- The *KEF* framework has good compatibility and is easily extended to existing attention-based models. In this work, we apply it to two latest TMSC models: *SaliencyBERT* (Wang et al., 2021) and *TomBERT* (Yu and Jiang, 2019). Experimental results on two public datasets prove the validity of our framework.

2 Notations and Preliminaries

In this section, we first present the task formalization, and then give brief introductions to adjective-noun pairs (Borth et al., 2013).

2.1 Task Formalization

We are given a set of multimodal samples $\mathcal{D}$. For each sample $c \in \mathcal{D}$, it contains a review sentence S with n words $(w_1, w_2, \cdots, w_n)$, an associated image I, as well as an opinion target T (refers to a span in sentence S). Our goal is to predict the sentiment label y of each opinion target mentioned in a pair of sentence and image, where y can be either *positive*, *negative*, or *neutral*.

2.2 Adjective-Noun Pairs

We extract the adjective-noun pairs (ANPs) from each image in our Twitter datasets to serve as an external knowledge base, where nouns denote coarse-grained object concepts in the image, and adjectives

[1] There are respectively 38.4% and 48.5% of opinion targets are different names of people in TWITTER-15 and TWITTER-17 datasets.

[2] Fig. 2(a) extracts the noun "*man*" from the image.

[3] Fig. 2(a) and Fig. 2(b) both extract same adjective "*happy*" from different smiling faces, this phenomenon is common.

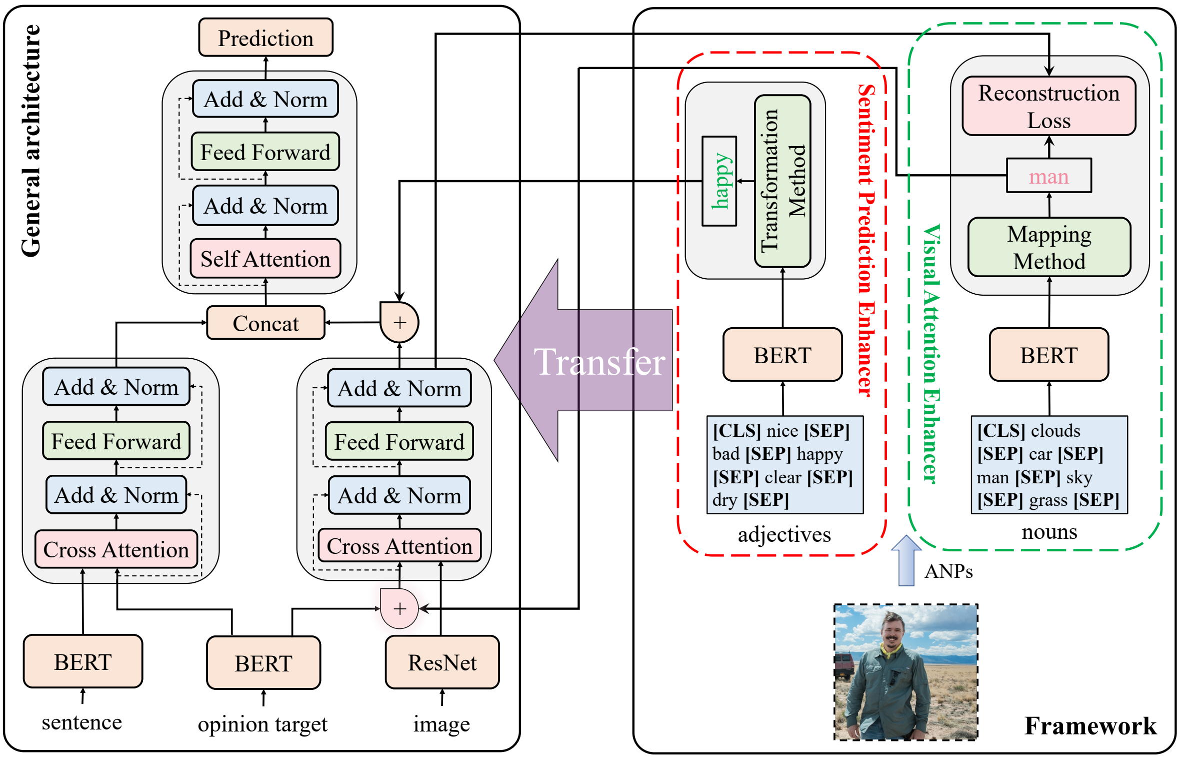


Figure 3: The overview of our KEF framework.

are modifiers of nouns. As shown in Fig. 2, we use SentiBank toolkit[4] to extract 1200 ANPs and select the Top-5[5] ANPs as the visual semantic information of each image, in which each pair contains an adjective word A_i and a noun word N_i:

$$\text{ANP} = (A_i, N_i), \tag{1}$$

Considering that adjectives and nouns in ANPs contain different semantics, i.e., the nouns mainly involve information about the opinion target, while the adjectives usually contain sentiment information. Thus, it is more appropriate and reasonable to encode them separately. For example, the BERT input for adjectives and nouns is given in Figure 3. We feed them to a pre-trained model BERT (Devlin et al., 2019) to obtain the hidden representations:

$$H_A = \text{BERT}(A), \tag{2}$$

$$H_N = \text{BERT}(N), \tag{3}$$

where $H_A \in \mathbb{R}^{(2l+1)\times d}$ and $H_N \in \mathbb{R}^{(2l+1)\times d}$ separately denote the adjective representations and noun representations, d indicates the hidden dimension, and l means the length of ANPs.

3 Knowledge-enhanced Framework

In this section, we will describe how to integrate Knowledge-enhanced Framework (*KEF*) into *TomBERT* (Yu and Jiang, 2019) and *Saliency-BERT* (Wang et al., 2021), which both achieve satisfying performance and thus are chosen as the foundation of our work.

3.1 Overview

Figure 3 shows the overall architecture of *KEF*, which contains two components: Visual Attention Enhancer and Sentiment Prediction Enhancer. Concretely, we first abstract a general attention architecture based on the well-designed *TomBERT* and *SaliencyBERT* models. Then, with the help of ANPs, we successively present a Visual Attention Enhancer and a Sentiment Prediction Enhancer. The former aims to improve the effectiveness of visual attention through a mapping method and a reconstruction loss, and the latter introduces a simple yet effective transformation approach to reduce the difficulty of predicting sentiment labels.

3.2 General Attention Architecture

Given an input sentence S, we first split S into two sub-sentences: the opinion target T and

[4] ee.columbia.edu/ln/dvmm/vso/download/sentibank.html

[5] we will give the reason why the Top-5 ANPs are extracted from each image in the Section 5.3.

the remaining context[6] C, and then separately feed them to pre-trained BERT to obtain the hidden representations:$H_C = \text{BERT}(C), H_T = \text{BERT}(T)$, where $H_C \in \mathbb{R}^{n \times d}$ and $H_T \in \mathbb{R}^{t \times d}$ denote the text representations and target representations, d is the hidden dimension, n and t are the length of C and the target T.

Similarly, for the associated image I, we adopt one of the state-of-the-art image recognition models ResNet-152 (res5c) (He et al., 2016) to obtain the output of the last convolutional layer:

$$\text{ResNet}(I) = \{r_j | r_j \in \mathbb{R}^{2048}, j = 1, 2, ..., 49\}, \tag{4}$$

which splits the original image into 7 × 7 = 49 regions and each region is represented by a 2048-dimensional vector r_j. Next, we use a linear function to project the visual features to the same space of textual features: $H_V = W_v \text{ResNet}(I)$, where $W_v \in \mathbb{R}^{d \times 2048}$ is the learnable parameter.

After that, we employ a cross-attention block to capture target-aware visual representation $H_{T \to V}$ and target-aware text representation $H_{T \to C}$:

$$H_{T \to V} = \text{Cross-ATT}(H_T, H_V), \tag{5}$$

$$H_{T \to C} = \text{Cross-ATT}(H_T, H_C), \tag{6}$$

where Cross-ATT$(\cdot)$ denotes the cross-modal multi-head attention as (Tsai et al., 2019). Then, we concatenate $H_{T \to C}$ and $H_{T \to V}$ together and further stack the attention block on top to obtain the multimodal output representation H.

Finally, we feed the first token H^0 of the multimodal representation to a softmax layer for the sentiment classification:

$$p(y|H^0) = \text{softmax}(W_M^\top H^0), \tag{7}$$

where $W_M \in \mathbb{R}^{d \times 3}$ is the weight matrix.

To optimize all the parameters, the objective is to minimize the standard cross-entropy loss function:

$$\mathcal{L}_t = -\frac{1}{|\mathcal{D}|} \sum_{i=1}^{|\mathcal{D}|} \text{log} p(y^i | H^0). \tag{8}$$

3.3 Visual Attention Enhancer

As mentioned before, the target appearing in the image is a *coarse-grained concept*, while the target mentioned in the sentence is a *fine-grained concept*, the inconsistency of target granularity causes visual attention in Eq. 5 sometimes fail to capture the corresponding visual representations of the target.

Basic Intuition. Apparently, the nouns extracted from the image are also *coarse-grained concepts*, so an intuitive idea is to map a fine-grained opinion target to a coarse-grained noun, and then use it as a bridge to capture the coarse-grained visual representations. However, most of the nouns extracted from the image are target-independent, so we cannot use them directly.

Mapping Method. To tackle the above challenge, we first measure the strength of target-noun relevance by calculating the semantic similarity between noun representation and target representations in the embedding space:

$$\alpha^i = \cos(H_T, H_N^i), \tag{9}$$

where H_N^i denotes single noun representation of H_N in Eq. 3, cos$(\cdot)$ is a cosine function and α^i means the similarity score.

Based on the largest similarity score, we can find the most relevant noun to the opinion target:

$$\alpha^m = \max_{i=1}^{l}(\alpha^i), \tag{10}$$

where l denotes the length of ANPs and H_N^m indicates the noun representation corresponding to the highest similarity score α^m.

Next, we aggregate them together as complementary information for the opinion target to capture the corresponding visual representations $H_{T \to V}$. Formally, we update H_T in Eq. 5 by:

$$\widetilde{H}_N = \alpha^m H_N^m, \tag{11}$$

$$H_T = H_T + \lambda_N \widetilde{H}_N, \tag{12}$$

where λ_N is a hyperparameter that controls the importance of $\widetilde{H}_N$ and can be adjusted.

Reconstruction Loss. To ensure that visual attention can capture the visual features associated with the opinion target more accurately, we also devise a reconstruction loss to minimize the divergence between target-relevant noun representations and target-aware visual representations. Formally,

$$\mathcal{L}_a = -\frac{1}{|\mathcal{D}|} \sum_{i=1}^{|\mathcal{D}|} (\widetilde{H}_N - H_{T \to V})^2, \tag{13}$$

In the Visual Attention Enhancer, the final loss is $\mathcal{L} = \mathcal{L}_t + \lambda \mathcal{L}_a$, where λ measures the importance of reconstruction loss $\mathcal{L}_a$ and can be adjusted.

[6] we replace the opinion target with \$T\$ in the context.

	TWITTER-15							TWITTER-17						
	Pos	Neg	Neu	Total	AT	Words	AL	Pos	Neu	Neg	Total	AT	Words	AL
Train	928	368	1883	3179	1.348	9023	16.72	1508	416	1638	3562	1.410	6027	16.21
Dev	303	149	670	1122	1.336	4238	16.74	515	144	517	1176	1.439	2922	16.37
Test	317	113	607	1037	1.354	3919	17.05	493	168	573	1234	1.450	3013	16.38

Table 1: The basic statistics of our two multimodal Twitter datasets. Pos: Positive, Neg: Negative, Neu: Neutral, AT: Avg. Targets, AL: Avg. Length

3.4 Sentiment Prediction Enhancer

Even if visual features are captured, there are still significant differences between the visual representations expressing the same mood, which brings challenges to learn the mapping function between visual representations and sentiment labels.

Basic Intuition. Considering that ANPs can usually extract the same adjectives from different visual representations expressing the same mood, so an intuitive idea is to map dversified visual representations to the same adjective. However, the adjective most relevant to visual representations is unknown, we need to find it explicitly.

Transformation Method. Actually, in the mapping method, we have found that the noun representation H_N^m is most relevant to target-aware visual representations $H_{T\to V}$. Since an adjective is a modifier of a noun, the adjective corresponding to this noun is also most relevant to target-aware visual representations. Finally, we use it as the complementary information of visual representations to reduce the difficulty of sentiment prediction. Formally, we rewrite $H_{T\to V}$ in Eq. 5 by:

$$H_{T\to V} = H_{T\to V} + \lambda_A H_A^m. \tag{14}$$

where H_A^m denotes the adjective representation corresponding to the noun representation H_N^m, λ_A is a hyperparameter and can be adjusted.

4 Experiments

4.1 Experimental Settings

Datasets. To evaluate the effect of *KEF* Framework, we carry out experiments on two public multimodal datasets `TWITTER-15` and `TWITTER-17` from (Yu and Jiang, 2019), which include user tweets posted during 2014-2015 and 2016- 2017, respectively. General information for both datasets is presented in Table 1.

Implementation Details. We build our *KEF* on top of the pre-trained uncased BERT-based model released by (Devlin et al., 2019), and tune the hyperparameters on the development set of each dataset. Specifically, we set λ_N, λ, λ_A to be {0.5, 0.2, 0.6} on `TWITTER-15` and {0.2, 0.3, 0.2} on `TWITTER-17`, the learning rate as 2e-5 and the dropout (Hinton et al., 2012) rate as 0.9. In addition, the mini-batch size is set to 16, the maximum length of the sentence input and the target input are respectively set as 64 and 32, the hidden dimension and the number of attention heads set as 768 and 12. All the models are implemented by the Tensorflow framework with an NVIDIA Tesla V100 GPU.

Evaluation Metrics and Significance Test. Following (Yu and Jiang, 2019), we use Accuracy (Acc) and Macro-F1 score as evaluation metrics. Besides, the paired t-test is conducted to test the significance of different methods. Finally, we report the average performance and standard deviation over 5 runs with random initialization. Our code and datasets are available at https://github.com/1429904852/KEF.

4.2 Compared Methods

We choose three kinds of baselines. **The first** is a frequently-used visual-based model *ResNet-Target*. **The second** is some classical text-based models, including *AE-LSTM* (Wang et al., 2016), *MemNet* (Tang et al., 2016b), *RAM* (Chen et al., 2017), *MGAN* (Fan et al., 2018), *BERT* (Devlin et al., 2019). **The third** is the recent multi-modal models, including *Res-MGAN*, *MIMN* (Xu et al., 2019b), *ESAFN* (Yu et al., 2019), *MMAP* (Zhou et al., 2021), *mPBERT* (Yu and Jiang, 2019), *ModalNet-BERT*(Zhang et al., 2021), EF-CapTrBERT (Khan and Fu, 2021), *TomBERT* (Yu and Jiang, 2019) and *Saliencybert* (Wang et al., 2021).

The *KEF* framework contains two plug-and-play components which can be easily combined or extended to existing attention-based methods. To better verify the effectiveness of *KEF*, we chose

Model	TWITTER-15		TWITTER-17	
	Acc	Macro-F1	Acc	Macro-F1
Visual				
Res-Target	59.88	46.48	58.59	53.98
Text				
AE-LSTM	70.30	63.43	61.67	57.97
MemNet	70.11	61.76	64.18	60.90
RAM	70.68	63.05	64.42	61.01
MGAN	71.17	64.21	64.75	61.46
BERT	74.15	68.86	68.15	65.23
Text + Visual				
Res-MGAN	71.65	63.88	66.37	63.04
MIMN	71.84	65.69	65.88	62.99
ESAFN	73.38	67.37	67.83	64.22
MMAP♣	73.50	66.53	67.31	64.34
mPBERT	75.79	71.07	69.61	67.12
ModalNet-Bert♣	76.71	70.93	69.55	67.28
EF-CapTrBERT★	77.01	71.79	69.00	66.71
Our Framework				
SaliencyBERT	77.03	72.36	69.69	67.19
KEF-SaliencyBERT	**78.15**†±0.33	**73.54**†±0.55	**71.88**†±0.21	**68.96**†±0.14
Δ	+1.12	+1.18	+2.19	+1.77
TomBERT	77.15	71.75	70.50	68.04
KEF-TomBERT	**78.68**†±0.30	**73.75**†±0.27	**72.12**†±0.15	**69.96**†±0.25
Δ	+1.53	+2.00	+1.62	+1.92

Table 2: Test accuracy on the TWITTER-15 and TWITTER-17 datasets (%). For the baseline model, the results with ♣ are produced with our implementation, the results with ★ are generated by running the code from (Khan and Fu, 2021), and the other results without symbols are retrieved from the original papers. For a fair comparison, we do not give the result of EF-CapTrBERT-DE from (Khan and Fu, 2021) since it use a domain-specific pre-trained encoder BERTweet from (Nguyen et al., 2020) instead of BERT-base. Δ denotes the difference between the performance of SaliencyBERT and KEF-SaliencyBERT, as well as TomBERT and KEF-TomBERT. We report the average performance and standard deviation over 5 runs. Best results are in bold. The marker † refers to significant test p-value $<$ 0.05 when comparing with other multi-modal models.

two latest BERT-based multimodal models as the foundations of our work, i.e., *TomBERT* and *Saliencybert*. In other words, we integrate *KEF* into *TomBERT* and *Saliencybert* to obtain the final model *KEF-TomBERT* and *KEF-Saliencybert*.

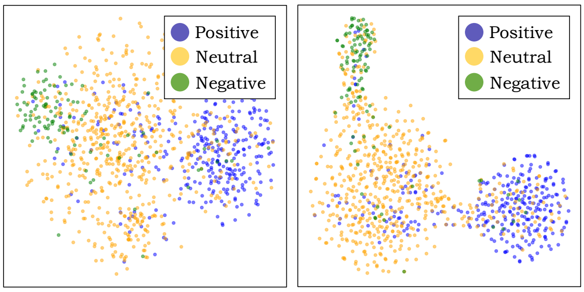
5 Results and Discussion

5.1 Main Results

The main experiment results are shown in Table 2. Based on the results, we can make a couple of observations: (1) We can see that pure visual-based methods perform very poorly, which implies that the associated images only play a supporting role to the text, and cannot be treated independently for target-oriented sentiment prediction; (2) For text-based methods, it is clear that *BERT* consistently outperforms all the baselines, which demonstrates the effectiveness of a strong pre-trained model. We attribute this to the fact that pre-trained models can provide rich semantic features; (3) *Res-MGAN* generally perform better than *MGAN*, which implies that the image information complements the textual information, and thus improves the performance for sentiment classification; (4) *TomBERT* and *Saliencybert* perform better than most multimodal models. A possible reason is that they employ self-modal and cross-modal multi-head attention to learn more robust representation; (5) *KEF-Saliencybert* and *KEF-TomBERT* both achieve competitive results on the TWITTER-15 and TWITTER-17 datasets. Specifically, compared to base version *TomBERT*, *KEF-TomBERT* obtains about 2.0% and 1.5% improvements in Macro-F1 and Accuracy. In contrast, *KEF-Saliencybert* outperforms *Saliencybert* by 1.5% and 1.7% on average. These results reveal that our framework have good compatibility; (6) *KEF-TomBERT* performs better than *KEF-Saliencybert* in most setting, which indicates that our framework is more effective for *TomBERT*.

Model	TWITTER-15		TWITTER-17	
	Acc	Macro-F1	Acc	Macro-F1
TomBERT	77.15	71.75	70.50	68.04
TomBERT+VAE	78.06±0.30	72.82±0.45	71.79±0.07	69.55±0.16
TomBERT+SPE	77.86±0.21	72.42±0.32	71.55±0.29	69.16±0.37
KEF-TomBERT	**78.68**±0.30	**73.75**±0.27	**72.12**±0.15	**69.96**±0.25
Δ (SPE)	+0.62	+0.93	+0.33	+0.41

Table 3: Ablation study of two main components. Δ represents the difference between the performance of *KEF-TomBERT* and *TomBERT+VAE*.



(a) TomBERT+VAE (b) KEF-TomBERT

Figure 4: Visualization of multimodal output representations for *TomBERT+VAE* and *KEF-TomBERT*.

5.2 Ablation Study

Without loss of generality, we choose *KEF-TomBERT* model for the ablation study to investigate the effects of different modules in *KEF*.

Model	TWITTER-15		TWITTER-17	
	Acc	Macro-F1	Acc	Macro-F1
TomBERT	77.15	71.75	70.50	68.04
TomBERT+MA	77.72±0.41	72.37±0.21	71.23±0.24	69.09±0.21
TomBERT+VAE	78.06±0.30	72.82±0.45	71.79±0.07	69.55±0.16
Δ (RL)	+0.34	+0.45	+0.56	+0.46

Table 4: Detailed ablation test over Visual Attention Enhancer. Δ represents the difference between the performance of *TomBERT+VAE* and *TomBERT+MA*.

Effects of Knowledge-enhanced Framework. We study the two main components of *KEF*: Visual Attention Enhancer (VAE) and Sentiment Prediction Enhancer (SPE). Based on the results reported in Table 3, we can observe the following: (1) In comparison with the base model *TomBERT*, *TomBERT+VAE* achieves competitive performance on both datasets, which validates the rationality of exploiting adjective-noun pairs to improve the visual attention capability; (2) After integrating *SPE* into *TomBERT+VAE*, *KEF-TomBERT* achieves the state-of-the-art performance, which demonstrates that *SPE* can improve the sentiment prediction capability through adjective-noun pairs; (3) *VAE* is more effective than *SPE*. This is explainable since the effectiveness of the attention mechanism is the core factor of sentiment prediction. Hence, it contributes more to our framework; (4) As depicted in Figure 4, we can see that multimodal output representations[7] learned by *KEF-TomBERT* are obviously more separable than those by *TomBERT+VAE*. This suggests that *SPE* can indeed reduce the difficulty of sentiment prediction.

Analysis over components of Visual Attention Enhancer. We further disassemble the Visual Attention Enhancer to see the contributions of the two sub-components: Mapping Method (MA) and Reconstruction Loss (RL). From the results in Table 4, we can observe that: (1) Compared to the base model *TomBERT*, *TomBERT+MA* achieves better performance, which indicates that the mapping method can help the opinion target capture corresponding visual representations; (2) After integrating *RL* into *TomBERT+MA*, *TomBERT+VAE* achieves further improvements, which demonstrates that the reconstruction loss can indeed improve the effectiveness of visual attention. This is consistent with our motivation.

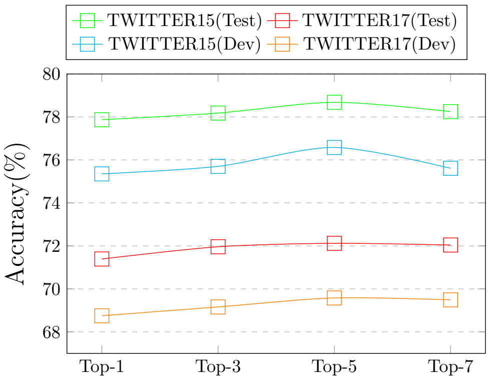


Figure 5: The results of *KEF-TomBERT* under different numbers of ANPs. Dev is short for development set.

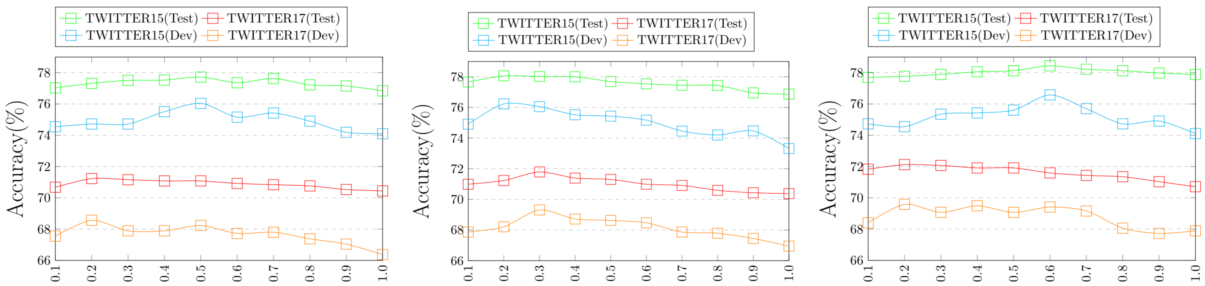
5.3 Parameter Analysis

In this subsection, we explore the effect of hyperparameters on our model. Specifically, we tune the hyperparameters on the development set, and then evaluate the performance on the test set.

Effect of the number of ANPs. To verify the impact of ANPs on *KEF-TomBERT* model, we extract the top 1, 3, 5, and 7 ANPs from each image. The results are shown in Figure 5. Obviously, as the number of ANPs increases, the performance of *KEF-TomBERT* gets better. And when the number of ANPs is equal to 5, *KEF-TomBERT* achieves the best results. However, once the number of ANPs is greater than 5, the performance does not continue to increase and even begins to fall. The reason behind this may be that: each sentence contains at most 5 opinion targets, so it will bring some noise when the number of ANPs is greater than the maximum number of opinion targets.

Effects of λ_N and λ. To investigate the effect of hyperparameters λ_N and λ on the Visual Attention Enhancer (VAE), we conduct experiments for values set at 0.1 intervals in the range (0, 1). Figure 6(a) and Figure 6(b) show the performance of *TomBERT+MA* and *TomBERT+VAE* with different λ_N and λ on both datasets, respectively. Actually, as λ_N and λ increase, the performance of *TomBERT+MA* and *TomBERT+VAE* has an initial upward trend, and then flattens out or begins to fall. Initially, the nouns in ANPs help the opinion target capture the corresponding visual representations more accurately, thus improving the performance. However, once the weight λ_N or λ exceeds a certain value, the nouns begin to dominate the attention process and perform poorly. It makes sense because we inevitably extract the wrong ANPs, so it

[7] visualized by t-SNE (Van der Maaten and Hinton, 2008).



(a) Effect of λ_N (b) Effect of λ (c) Effect of λ_A

Figure 6: The effect of hyper-parameters λ_N, λ and λ_A on the development set and test set.

tends to have a negative impact when the adjectives account for a large proportion. This is also why we only use it as complementary information to the opinion target instead of directly replacing the opinion target. Finally, we set λ_N, λ to be $\{0.5, 0.2\}$ on TWITTER-15 and $\{0.2, 0.3\}$ on TWITTER-17.

Effects of λ_A. To analyze the effect of different λ_A on Sentiment Prediction Enhancer (SPE), we adjust λ_A in (0, 1) to conduct experiments and the step is 0.1. Figure 6(c) shows the results of *KEF-TomBERT* with varying λ_A on two datasets. According to the trend of the curve, we set λ_A to be 0.6 on TWITTER-15 and 0.2 on TWITTER-17, the reason behind this is similar to λ_N and λ.

Jointly observing Figure 5 and Figure 6, we found that the best results of the development set and test set are basically consistent, which indicates that our framework has good robustness.

5.4 Case Study

To better understand the advantage of Visual Attention Enhancer (*VAE*) and Sentiment Prediction Enhancer (*SPE*), we randomly select some samples from the Twitter dataset for a case study.

Effects of Visual Attention Enhancer. As shown in Figure 7(a), the base model *TomBERT* incorrectly predicts the sentiment of the opinion target "*Korkie*". It is reasonable since we found that *TomBERT* focuses on visual clues (highlighted by the yellow bounding boxes) that are not related to the opinion target. After integrating *VAE* into *TomBERT*, *TomBERT+VAE* maps fine-grained opinion target "*Korkie*" to the coarse-grained noun "*man*" in ANPs. With the aid of the noun "*man*", *TomBERT+VAE* successfully captures the target-relevant visual clues (highlighted by the red bounding boxes), thus giving the right predictions.

Effects of Sentiment Prediction Enhancer. As shown in Figure 7(b) and Figure 7(c), although *TomBERT+VAE* accurately captures the corresponding visual representations (i.e., smiling faces) of the opinion target, the diversification of smile expressions increases the difficulty of sentiment prediction, thus *TomBERT+VAE* incorrectly predict the sentiment over "*Sammy*" in Figure 7(c). After integrating *SPE* into *TomBERT+VAE*, *KEF-TomBERT* maps different smiling faces to the same adjectives "*happy*". Apparently, it is easier for the *KEF-TomBERT* to learn the mapping function between these "*happy*" and sentiment label "*positive*", thus making the right prediction.

5.5 Error Analysis

Although our model improves the overall performance of TMSC, *KEF-TomBERT* and *KEF-SaliencyBERT* make some wrong predictions due to extracting some noise ANPs. According to the statistics, for *KEF-TomBERT* model, about 5.50% and 7.05% samples of the dataset TWITTER-15 and TWITTER-17 are predicted successfully by the model *TomBERT* but incorrectly by *KEF-TomBERT*. In contrast, for *KEF-SaliencyBERT* model, almost 5.79% and 7.29% samples of the dataset TWITTER-15 and TWITTER-17 are predicted successfully by the model *SaliencyBERT* but incorrectly by *KEF-SaliencyBERT*. Additionally, the mapping method is unable to consistently find the correct noun for a given opinion target. There ought to be more advanced natural language processing techniques to address them.

6 Related Work

Target-oriented Sentiment Classification. As an important task in aspect-based sentiment analysis, Target-oriented Sentiment Classification (TSC) has been extensively studied in recent years. With

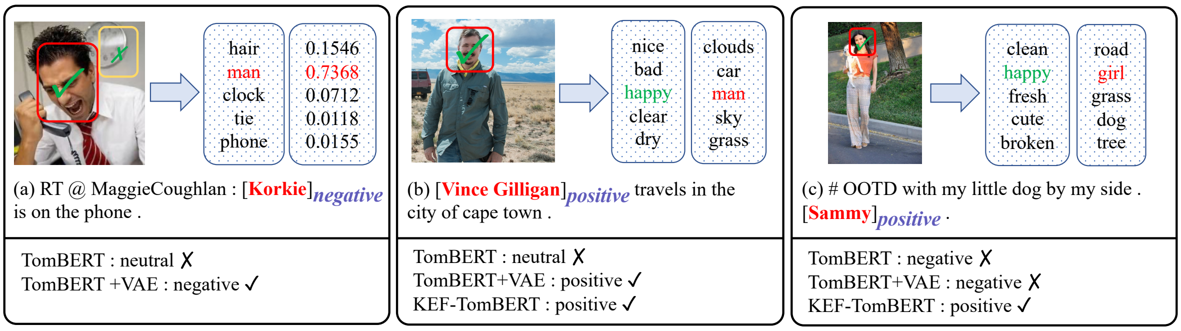


Figure 7: Predictions of *TomBERT*, *TomBERT+VAE* and *KEF-TomBERT* on three test samples. ✗and ✔denote incorrect and correct predictions. Opinion targets and their sentiment polarities are highlighted in the sentence. The yellow bounding box and red bounding box denote the visual clues that the opinion target focuses on under different methods. The numbers in the first sample indicate the similarity scores between the target and each noun in ANPs.

the development of deep learning, various neural networks have been designed for this task and have obtained promising results (Tang et al., 2016a; Li et al., 2018; Xue and Li, 2018). Recently, many studies have designed attention-based methods (Tang et al., 2016b; Wang et al., 2016; Ma et al., 2017; Chen et al., 2017; Fan et al., 2018; Zhao et al., 2020; He et al., 2018; Xu et al., 2019a; Hu et al., 2019; Xu et al., 2020; Wang et al., 2020) and graph-based methods (Zhang et al., 2019; Zhang and Qian, 2020; Huang and Carley, 2019; Sun et al., 2019; Tang et al., 2020; Chen et al., 2020) to model the interactions between the target and the context. However, none of the above works take visual modality into consideration, which can complement each other with these text-based methods.

Target-oriented Multimodal Sentiment Classification. With the growth of multimodal data (e.g., image) on the web, researchers proposed a new subtask of aspect-based sentiment analysis, namely Target-oriented Multimodal Sentiment Classification (TMSC), which has been explored in a few studies(Xu et al., 2019b; Yu et al., 2019; Yu and Jiang, 2019; Zhou et al., 2021; Zhang et al., 2021; Khan and Fu, 2021; Wang et al., 2021; Ling et al., 2022). Among them, based on the LSTM architecture, Xu et al. (2019b), Yu et al. (2019) and Zhou et al. (2021) proposed the MIMN, ESAFN and MMAP network to effectively model the *target-text* and *target-image* interactions. In contrast, Yu and Jiang (2019), Wang et al. (2021), Zhang et al. (2021) and Khan and Fu (2021) aim to explore the usefulness of the BERT architecture for TMSC and propose the TomBERT, Saliency-BERT, ModalNet-BERT and EF-CapTrBERT. Different from previous studies, this paper leverages the adjective-noun pairs (ANPs) to align text and image in the TMSC task.

7 Conclusion and Future Work

In this paper, we propose a novel knowledge-enhanced Framework (*KEF*) for the TMSC task. Specifically, with the aid of ANPs, we design two novel knowledge enhancers, Visual Attention Enhancer and Sentiment Prediction Enhancer, to improve the visual attention capability and sentiment prediction capability of the TMSC task. Results from numerous experiments indicate that our model achieves better performance than other state-of-the-art methods. Further analysis also validates the superiority of our framework.

In the future, we would like to apply our idea to other multimodal tasks since the adjective-noun pairs (ANPs) extracted from the image are easy to extend to other multimodal tasks, e.g., multi-modal entity linking, multi-modal machine comprehension and multi-modal dialogue generation.

Acknowledgements

We would like to thank the anonymous reviewers for their constructive comments. This work was supported by the National Natural Science Foundation of China (No. 61936012 and 61976114).

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Damian Borth, Rongrong Ji, Tao Chen, Thomas Breuel, and Shih-Fu Chang. 2013. Large-scale visual sentiment ontology and detectors using adjective noun pairs. In *Proceedings of the 21st ACM international conference on Multimedia*, pages 223–232.

Chenhua Chen, Zhiyang Teng, and Yue Zhang. 2020. Inducing target-specific latent structures for aspect sentiment classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5596–5607, Online. Association for Computational Linguistics.

Peng Chen, Zhongqian Sun, Lidong Bing, and Wei Yang. 2017. Recurrent attention network on memory for aspect sentiment analysis. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pages 452–461.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL 2019*, pages 4171–4186. Association for Computational Linguistics.

Feifan Fan, Yansong Feng, and Dongyan Zhao. 2018. Multi-grained attention network for aspect-level sentiment classification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3433–3442.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Ruidan He, Wee Sun Lee, Hwee Tou Ng, and Daniel Dahlmeier. 2018. Exploiting document knowledge for aspect-level sentiment classification. In *ACL 2018*, pages 579–585. Association for Computational Linguistics.

Geoffrey E. Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2012. Improving neural networks by preventing co-adaptation of feature detectors. *CoRR*, abs/1207.0580.

Mengting Hu, Shiwan Zhao, Li Zhang, Keke Cai, Zhong Su, Renhong Cheng, and Xiaowei Shen. 2019. CAN: Constrained attention networks for multi-aspect sentiment analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4601–4610, Hong Kong, China. Association for Computational Linguistics.

Binxuan Huang and Kathleen Carley. 2019. Syntax-aware aspect level sentiment classification with graph attention networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5469–5477, Hong Kong, China. Association for Computational Linguistics.

Zaid Khan and Yun Fu. 2021. Exploiting BERT for multimodal target sentiment classification through input space translation. In *MM '21: ACM Multimedia Conference, Virtual Event, China, October 20 - 24, 2021*, pages 3034–3042. ACM.

Xin Li, Lidong Bing, Wai Lam, and Bei Shi. 2018. Transformation networks for target-oriented sentiment classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 946–956.

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.

Bing Liu. 2012. *Sentiment Analysis and Opinion Mining*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

Dehong Ma, Sujian Li, Xiaodong Zhang, and Houfeng Wang. 2017. Interactive attention networks for aspect-level sentiment classification. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, pages 4068–4074. ijcai.org.

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. Bertweet: A pre-trained language model for english tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, EMNLP 2020 - Demos, Online, November 16-20, 2020*, pages 9–14. Association for Computational Linguistics.

Bo Pang, Lillian Lee, et al. 2008. Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval*, 2(1–2):1–135.

Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 2014. Semeval-2014 task 4: Aspect based sentiment analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation*, pages 27–35. The Association for Computer Linguistics.

Kai Sun, Richong Zhang, Samuel Mensah, Yongyi Mao, and Xudong Liu. 2019. Aspect-level sentiment analysis via convolution over dependency tree. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5679–5688, Hong Kong, China. Association for Computational Linguistics.

Duyu Tang, Bing Qin, Xiaocheng Feng, and Ting Liu. 2016a. Effective lstms for target-dependent sentiment classification. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3298–3307.

Duyu Tang, Bing Qin, and Ting Liu. 2016b. Aspect level sentiment classification with deep memory network. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. The Association for Computational Linguistics.

Hao Tang, Donghong Ji, Chenliang Li, and Qiji Zhou. 2020. Dependency graph enhanced dual-transformer structure for aspect-based sentiment classification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6578–6588, Online. Association for Computational Linguistics.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 6558–6569. Association for Computational Linguistics.

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*.

Jiawei Wang, Zhe Liu, Victor Sheng, Yuqing Song, and Chenjian Qiu. 2021. Saliencybert: Recurrent attention network for target-oriented multimodal sentiment classification. In *Pattern Recognition and Computer Vision - 4th Chinese Conference, PRCV 2021, Beijing, China, October 29 - November 1, 2021, Proceedings, Part III*, volume 13021 of *Lecture Notes in Computer Science*, pages 3–15. Springer.

Kai Wang, Weizhou Shen, Yunyi Yang, Xiaojun Quan, and Rui Wang. 2020. Relational graph attention network for aspect-based sentiment analysis. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3229–3238, Online. Association for Computational Linguistics.

Yequan Wang, Minlie Huang, Li Zhao, et al. 2016. Attention-based lstm for aspect-level sentiment classification. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pages 606–615.

Hu Xu, Bing Liu, Lei Shu, and Philip Yu. 2019a. BERT post-training for review reading comprehension and aspect-based sentiment analysis. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2324–2335, Minneapolis, Minnesota. Association for Computational Linguistics.

Lu Xu, Lidong Bing, Wei Lu, and Fei Huang. 2020. Aspect sentiment classification with aspect-specific opinion spans. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3561–3567, Online. Association for Computational Linguistics.

Nan Xu, Wenji Mao, and Guandan Chen. 2019b. Multi-interactive memory network for aspect based multimodal sentiment analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 371–378.

Wei Xue and Tao Li. 2018. Aspect based sentiment analysis with gated convolutional networks. In *ACL 2018*, pages 2514–2523. Association for Computational Linguistics.

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 2019, Macao, China, August 10-16, 2019*, pages 5408–5414. ijcai.org.

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.

Chen Zhang, Qiuchi Li, and Dawei Song. 2019. Aspect-based sentiment classification with aspect-specific graph convolutional networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4568–4578, Hong Kong, China. Association for Computational Linguistics.

Mi Zhang and Tieyun Qian. 2020. Convolution over hierarchical syntactic and lexical graphs for aspect level sentiment analysis. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3540–3549, Online. Association for Computational Linguistics.

Zhe Zhang, Zhu Wang, Xiaona Li, Nannan Liu, Bin Guo, and Zhiwen Yu. 2021. Modalnet: an aspect-level sentiment classification model by exploring multimodal data with fusion discriminant attentional network. *World Wide Web*, pages 1–18.

Fei Zhao, Zhen Wu, and Xinyu Dai. 2020. Attention transfer network for aspect-level sentiment classification. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 811–821, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Jie Zhou, Jiabao Zhao, Jimmy Xiangji Huang, Qinmin Vivian Hu, and Liang He. 2021. Masad: A large-scale dataset for multimodal aspect-based sentiment analysis. *Neurocomputing*, 455:47–58.