

# Face-Sensitive Image-to-Emotional-Text Cross-modal Translation for Multimodal Aspect-based Sentiment Analysis

Hao Yang, Yanyan Zhao\* and Bing Qin

Harbin Institute of Technology

hyang@ir.hit.edu.cn, yyzhao@ir.hit.edu.cn, qinb@ir.hit.edu.cn

## Abstract

Aspect-level multimodal sentiment analysis, which aims to identify the sentiment of the target aspect from multimodal data, recently has attracted extensive attention in the community of multimedia and natural language processing. Despite the recent success in textual aspect-based sentiment analysis, existing models mainly focused on utilizing the object-level semantic information in the image but ignore explicitly using the visual emotional cues, especially the facial emotions. How to distill visual emotional cues and align them with the textual content remains a key challenge to solve the problem. In this work, we introduce a face-sensitive image-to-emotional-text translation (FITE) method, which focuses on capturing visual sentiment cues through facial expressions and selectively matching and fusing with the target aspect in textual modality. To the best of our knowledge, we are the first that explicitly utilize the emotional information from images in the multimodal aspect-based sentiment analysis task. Experiment results show that our method achieves state-of-the-art results on the Twitter-2015 and Twitter-2017 datasets. The improvement demonstrates the superiority of our model in capturing aspect-level sentiment in multimodal data with facial expressions<sup>1</sup>.

## 1 Introduction

As an important task of multimodal sentiment analysis, multimodal aspect-based sentiment analysis(MABSA) aims to classify the sentiment polarity expressed in the sentence-image pair towards a specific aspect. In contrast to textual aspect-level sentiment analysis, mining and utilizing emotional clues for the aspect in visual content is the core problem.

Early works (Xu et al., 2019; Yu and Jiang, 2019; Wang et al., 2021) on the MABSA task treat the

\* Corresponding author

<sup>1</sup>Our code is publicly available at: <https://github.com/yhit98/FITE>.



Figure 1: Examples of multimodal aspect-based sentiment analysis with image caption.

visual features extracted by unimodal pre-trained models as equal to the textual features. These works focus on utilizing the attention mechanism to implicitly align and fuse the semantic information and the emotional information in the two modalities. More recently, Yu et al. (2019) takes the object-level visual semantic information into account in the feature-based multimodal fusion stage. As shown in Figure 1(a) and (b), we observe that the facial expression in the image is important to identify the sentiment of the target aspect “antonellaRoccuzzo” and “Kate Middleton”. But due to the limitations of the dataset and weak supervision of cross-modal alignment in the MABSA task, the methods that implicitly capture emotional information in images are more likely to learn the non-emotional cues bias.

Khan and Fu (2021) introduce an object-aware transformer to translate images to captions for cross-modal fusion. The cross-modal translation effectively solves the problem of multimodal feature fusion from different feature spaces. As shown in Figure 1, the introduced image caption represents semantic information of visual content. But we can observe that the sentiment polarities of almost image captions are neutral, which indicates the method ignores almost all emotional cues from visual content. Existing research (Fan et al., 2018b)

has shown that the face region in the images is a strong emotional indicator. The work found that (1) when human observers watch visual content, the emotional object attracts human attention more than the neutral object and (2) this emotion prioritization effect is stronger for human-related objects than objects unrelated to humans. According to statistics, there are more than 50% of the images in Twitter data contain facial expressions. Therefore, we believe that the human facial expression cannot be ignored in the MABSA task, and propose explicitly using face information as visual emotional cues to transform visual emotional signals into text for cross-modal fusion.

Moreover, compared with the image with a single face that the emotion reflected in the image is highly consistent with the facial expression. For the image with multiple faces, especially those faces with different expressions, it's important to match the target aspect to the relevant face expression. For example, in Figure 1(b), the expression of the face-3 is different from the face-1 and face-2.

Inspired by the success of facial emotion recognition research (Dalal and Triggs, 2005; Li et al., 2017) in the computer vision field, we propose a simple but effective face-sensitive image-to-emotional-text cross-modal translation method, which textualizes the sentiment contained in the image by generated facial expression description. Our approach does not require additional training in emotional image caption generation models. Furthermore, in the aspect-sensitive alignment and modification stage, while considering the relationship between scene information and facial expression, we adopt one pre-trained Vision-Language model CLIP (Radford et al., 2021) to selectively retrieve facial sentiment cues in images that are most relevant to the target aspect. It effectively solves the problem of matching target aspect and image expression descriptions in multi-face scenes. In the final stage, we apply the gate mechanism to fuse and denoise the multimodal features.

In our extensive set of experiments, we show (a) that our method achieves state-of-the-art performance on Twitter-2015 and Twitter-2017 datasets. And the performance on the Twitter-face dataset demonstrates the model's ability to classify the sentiment polarity of fine-grained aspects in multimodal data. The Twitter-face dataset consists of data that contains facial expressions in the Twitter-2015 and Twitter-2017 datasets. And (b) through a

series of ablation experiments, we proved that our FITE model can effectively capture the emotional cues in the image, and align the visual emotional cues with the textual target aspect. Meanwhile, (c) case study proves that our method also has a significant effect on the aspect that is not directly human-related.

## 2 Related Work

Early works on aspect-based sentiment analysis only focused on the text (Wang et al., 2016; Xue and Li, 2018; Hu et al., 2019; Zhu et al., 2019; Li et al., 2020). While for multimodal data, the goal becomes to identify the aspect in multimodal text-image pairs. In 2019, Xu et al. (2019) proposed the aspect-based multimodal sentiment analysis task and proposed a novel Multi-Interactive Memory Network (MIMN) model based on BiLSTM. Yu and Jiang (2019) proposed a BERT-based multimodal architecture TomBERT for target-oriented multimodal sentiment classification task (TMSC). Yu et al. (2019) proposed a entity-sensitive attention and fusion network for multimodal target-based sentiment classification. Ju et al. (2021) proposed a multimodal joint learning approach with auxiliary cross-modal relation detection for multimodal aspect-level sentiment analysis. However, compared with other multimodal tasks such as image and text retrieval, the sentiment annotation used in the MABSA task lack strong supervision signals for cross-modal alignment. This issue makes it difficult for most existing MABSA models to learn cross-modal interactions and causes models to learn the bias brought by the image.

In an effort to align the multimodal features from different semantic spaces and learn cross-modal representations with visual fine-grained object information, Khan and Fu (2021) propose a new method to utilize visual modalities, the image caption generation module in their model undertakes the task of cross-modal alignment. They convert images into text descriptions based on the idea of cross-modal translation. The success of their work on the MABSA task benefits from the image caption generation model and the powerful context modeling ability of the pre-training language model for textual content. However, according to statistics, nearly 98% image captions used in their work are with neutral sentiment polarities, which indicates that their cross-modal translation module ignores almost all sentiment cues from visual con-

Template	Example	Facial Attribute & Confidence				Face Description
		Age	Race	Gender	Emotion	
A [Age]-year-old [Race] [Gender]with a [Emotion] expression.	(a) face-2	38 1.0000	white 0.4425	man 1.0000	neutral 0.9904	A 38-year-old white man with a neutral expression.

Table 1: The example of generate face description for the face (2) in Figure 1(a) by using template.

tent. Most recently, Ling et al. (2022) propose a task-specific Vision-Language Pre-training framework for MABSA (VLP-MABSA), which used three types of task-specific pre-training tasks.

### 3 Method

#### 3.1 Task Formulation

The task of MABSA can be formulated as follows: given a set of multimodal samples  $S = \{X_1, X_2, \dots, X_{|S|}\}$ , where  $|S|$  is the number of samples. And for each sample, we are given an image  $V \in \mathbb{R}^{3 \times H \times W}$  where 3,  $H$  and  $W$  represent the number of channels, height and width of the image, and an  $N$ -word textual content  $T = (w_1, w_2, \dots, w_N)$  which contains an  $M$ -word sub-sequence as target aspect  $A = (w_1, w_2, \dots, w_M)$ . Our goal is to learn a sentiment classifier to predict a sentiment label  $y \in \{\text{Positive}, \text{Negative}, \text{Neutral}\}$  for each sample  $X = (V, T, A)$ .

#### 3.2 Overview

As shown in Figure 2, the proposed model consists of three modules: face-sensitive image-to-emotional-text translation module, aspect-sensitive alignment and modification module, and gate-based multimodal fusion module. For the given multimodal tweet  $X = (V, T, A)$ , we take the visual input  $V$  into the face-sensitive image-to-emotional-text translation module to generate face descriptions  $D = \{D_1, D_2, \dots, D_I\}$  where  $I$  is the number of faces that contained in the visual input  $V$  and  $D_i = (w_1, w_2, \dots, w_K)$  represent a  $K$ -word sentence. This module focuses on extracting and textualizing facial expressions in the image which contain abundant emotional cues. Subsequently, since the visual input  $V$  may contain multiple facial expressions, it is necessary to match the target aspect  $A$  with the relevant facial description  $D_A$ . In the aspect-sensitive alignment and modification module, we calculate the cosine similarity between each face description spliced with aspect  $A$  and the image input  $V$ . And then we select and rewrite the face description  $D_A$  according to these similarity scores. In addition, considering that image scenes

can supplement additional semantic information, we adopt the caption transformer used in CapBERT (Khan and Fu, 2021) to generate image caption for scene  $C = (w_1, w_2, \dots, w_J)$ , where  $J$  denotes the length of the image caption. Finally, in the gate-based multimodal fusion module, we utilize two pre-trained language models to model the face description and the image caption of the scene, and then we adopt the gating mechanism for feature fusion and denoising. The output of the gated unit passes through a linear layer for aspect sentiment prediction. In the following subsections, we will introduce each module in detail.

#### 3.3 Face-Sensitive Image-to-Emotional-Text Translation

This module is proposed to address two inherent challenges in MABSA. One challenge is that due to the image in the multimodal tweet without any category restrictions, it's difficult to distill the object-level emotional cues in complex images. The other one is that if emotional cues are extracted from images, how to translate the affective image content into textual modality in a low-resource setting.

To approach the first challenge, as mentioned above, exploiting the rich facial expressions in images is a direct and effective way to extract visual emotional cues. We firstly apply the face recognize tool (Serengil and Ozpinar, 2020) to recognize all the faces  $F = \{F_1, F_2, \dots, F_I\}$ , where  $I$  is the number of faces and  $F_i \in \mathbb{R}^{3 \times H_F \times W_F}$  denotes a face region with 3 channels,  $H_F$  height and  $W_F$  width. And then we take these faces as input of four pre-trained face-based classifiers (Serengil and Ozpinar, 2021) which can provide facial attribute analysis, including age, gender, facial expression (including angry, fear, neutral, sad, disgust, happy, and surprise) and race (including Asian, Black, White, Middle Eastern, Indian, and Latino) predictions.

For the second challenge, we want to translate facial expressions in images to textual content without additionally training a new emotional image caption model. Therefore, after the facial attribute analysis, we filter the obtained facial attributes according to the prediction confidence. We filter out the face attributes with confidence below the thresh-

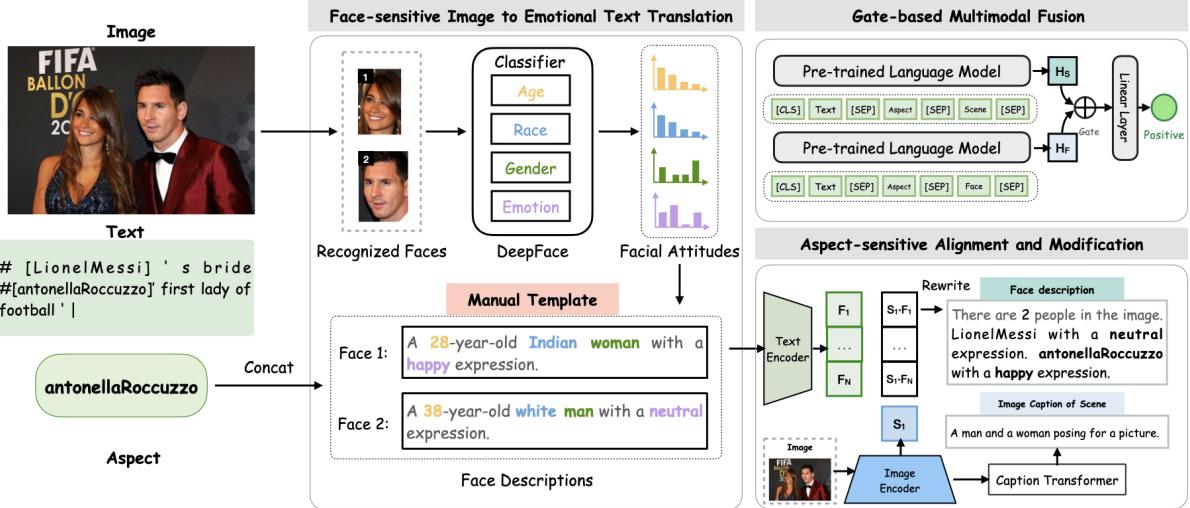


Figure 2: The overview of face-sensitive image-to-emotional-text cross-modal translation model architecture.

old  $\theta = 0.4$ . To generate fluent nature language emotional face descriptions, we manually design a face description generated pattern which is consist of the facial attributes. The example of generating a face description is shown in Table 1.

### 3.4 Aspect-Sensitive Alignment and Modification

Consider the multi-face example in Figure 1(b), the facial expressions in the image are different, and the children’s angry facial expression is not helpful for predicting the sentiment polarity of the aspect “*Kate Middleton*”. However, the irrelevant facial expressions introduce noise and downgrade the performance. Hence, it is necessary to accurately align the facial expressions in the image with the target aspect. This module mainly focuses on the fine-grained alignment of facial expressions and the target aspect and rewrites the corresponding facial descriptions.

Given that the MABSA task does not contain direct image-text alignment supervision, and the size of the datasets for the MABSA task constrains the model to learn fine-grained alignments through contrastive learning, it is necessary to introduce external image-text alignment knowledge. To this end, we apply the CLIP model to perform such fine-grained alignment. we use the text encoder and image encoder of the CLIP model which is pre-trained on the large-scale image-text pair datasets to encode the face descriptions  $D$  connected with aspect  $A$  and the image  $V$  respectively. The feature

embeddings of face descriptions and images:

$$H_{D\&A} = \text{Text\_Encoder}(\text{concat}(D, A)) \quad (1)$$

$$H_V = \text{Image\_Encoder}(V) \quad (2)$$

After that, we project the output feature embeddings into the same feature space. And with L2-normalization, we calculate the cosine similarity  $L$  of these feature embeddings. After that, we select and rewrite the face description with the highest similarity to the current image as the textualized visual emotional cues for the current aspect. The rewirited face description only retains the target aspect and the expression from predicted facial attributes.

$$H'_{D\&A} = \text{L2\_Normalize}(H_{D\&A} \cdot W_{D\&A}) \quad (3)$$

$$H'_V = \text{L2\_Normalize}(H_V \cdot W_V) \quad (4)$$

$$L = (H'_V \cdot (H'_{D\&A})^T) * e^t \quad (5)$$

where  $W_{D\&A}$  and  $W_V$  are learnable weight, and  $t$  is the temperature scaling in CLIP model. More details are presented in the appendix.

In addition, considering the impact of visual scene information on multimodal semantics, we follow the transformer-based image caption model from CapBERT to generate a neutral overall description of the image.

$$C = \text{Caption\_Transformer}(V) \quad (6)$$

Finally, we obtain the aligned face descriptions and image captions of the scene and feed them into the next module as input.

Split	Twitter-2015						Twitter-2017					
	#POS	#Neutral	#NEG	Total	#Aspects	#Len	#POS	#Neutral	#NEG	Total	#Aspects	#Len
Train	928	1883	368	3179	1.34	16.72	1508	416	1638	3562	1.41	16.21
Valid.	303	679	149	1122	1.33	16.74	515	144	517	1176	1.43	16.37
Test	317	607	113	1037	1.35	17.05	493	168	573	1234	1.45	16.38

Table 2: Statistics of two benchmark datasets for multimodal aspect-based sentiment analysis task.

Split	Twitter-face					
	#POS	#Neutral	#NEG	Total	#Aspect	#Len
Train	1285	1531	408	3204	1.37	16.52
Valid.	449	514	137	1100	1.37	16.54
Test	442	494	156	1092	1.39	16.53

Table 3: Statistics of Twitter-face dataset.

### 3.5 Gate-Based Multimodal Fusion Module

In this module, we aim to fuse the text input and target aspect with the generated face description  $D_{all}$  and the image caption  $C$  of the scene in the text-modal feature space. To take advantage of the powerful textual context modeling capabilities of the pre-trained language model, we concatenate the face description and the image caption of sense with text  $T$  and target aspect  $A$  to form two new sentences respectively:

$$[CLS]w_1^T, \dots, w_N^T [SEP] w_1^A, \dots, w_M^A [SEP] w_1^{D_{all}}, \dots, w_K^{D_{all}} [SEP] \quad (7)$$

$$[CLS]w_1^T, \dots, w_N^T [SEP] w_1^A, \dots, w_M^A [SEP] w_1^C, \dots, w_J^C [SEP] \quad (8)$$

Then we feed the new sentences into two pre-trained language models and fine-tune the language models to obtain the pooler outputs of the  $[CLS]$  token  $H_D^{[CLS]} \in \mathbb{R}^{768}$  and  $H_C^{[CLS]} \in \mathbb{R}^{768}$ . Given the performance of the generation methods, the generated face descriptions and image captions of the scene contain non-negligible noise. To alleviate the noise, we utilize the gate mechanism to denoise the feature representations  $H_D^{[CLS]}$  and  $H_C^{[CLS]}$ . The fused feature representations are passed through a linear classification layer to obtain sentiment prediction results.

$$g_t = \sigma(W_D H_D^{[CLS]} + W_C H_C^{[CLS]} + b_g) \quad (9)$$

$$H = g_t H_D^{[CLS]} + (1 - g_t) H_C^{[CLS]} \quad (10)$$

$$p(y|H) = softmax(WH + b) \quad (11)$$

where  $W_D \in \mathbb{R}^{768 \times 768}$ ,  $W_C \in \mathbb{R}^{768 \times 768}$ ,  $W \in \mathbb{R}^{768 \times 3}$ ,  $b_g \in \mathbb{R}^{768}$  and  $b \in \mathbb{R}^3$  are learnable parameters, and  $\sigma$  is the non-linear transformation function  $tanh$ .

We use the standard cross-entropy loss to optimize all the parameters in this module.

$$\mathcal{L} = -\frac{1}{|D|} \sum_{l=0}^{|D|} \log p(y^{(l)} | H^{(l)}) \quad (12)$$

## 4 Experiment

### 4.1 Experimental Setup

We trained our model and measured its performance on the Twitter-2015 and Twitter-2017 datasets. These two datasets consist of multimodal tweets that are annotated the mentioned aspect in text content and the sentiment polarity of each aspect. Each multimodal tweet is composed of an image and a text that contains the target aspect. Since our method focuses more on examples containing faces, we extract the examples containing faces in the above two datasets to form the Twitter-face dataset and verify the superiority of our model on this dataset. The detailed statistics of the three datasets are shown in Table 2 and Table 3. In addition, we set the model learning rate as 5e-5, the pre-trained model attention head as 12, the dropout rate as 0.1, the batch size as 16 and the fine-tuning epochs as 8, and the maximum text length is 256. We report the average results of 5 independent training runs for all our models. And all the models are implemented based on PyTorch with two NVIDIA TeslaV100 GPUs.

### 4.2 Compared Baselines

In this section, we compared with the following models and reported the accuracy and Macro-F1 score in Table 4.

We compare the method in the image-only setting: the Res-Target model which directly uses the visual feature of the input image from ResNet (He et al., 2016). As well as the text-only models: (1) LSTM. (2) MGAM, a multi-grained attention network (Fan et al., 2018a) which fuses the target and text in multi-level. (3) BERT, the representative pre-trained language model (Devlin et al., 2019), which has strong text representation ability and can learn

Method	Twitter-2015		Twitter-2017	
	Acc	Macro-F1	Acc	Macro-F1
Image Only				
Res-Target	59.88	46.48	58.59	53.98
Text Only				
LSTM	70.30	63.43	61.67	57.97
MGAN	71.17	64.21	64.75	61.46
Bert	74.25	70.04	68.88	66.12
Text and Image				
MIMN	71.84	65.69	65.88	62.99
ESAFN	73.38	67.37	67.83	64.22
VilBERT	73.69	69.53	67.86	64.93
TomBERT	77.15	71.15	70.34	68.03
CapBERT	78.01	73.25	69.77	68.42
CapBERT-DE	77.92	73.90	72.30	70.20
VLP-MABSA	78.60	73.80	73.80	71.80
FITE	78.49	73.90	70.90	68.70
FITE-DE	78.64	74.30	72.98	71.97
FITE-DE-Large	<b>78.76</b>	<b>74.79</b>	<b>73.87</b>	<b>73.03</b>

Table 4: Experiment results for multimodal aspect-based sentiment analysis.

alignment between two arbitrary inputs. Moreover, the multimodal compared baselines include: (1) MIMN, The Multi-Interactive Memory Network ([Xu et al., 2019](#)) learn the interactive influences in cross-modality and self-modality. (2) ESAFN, a entity-sensitive attention and fusion network ([Yu et al., 2019](#)). (3) VilBERT, a pre-trained Vision-Language model ([Lu et al., 2019](#)), the target aspect is concatenated to the input text. (4) TomBERT, the TomBERT ([Yu and Jiang, 2019](#)) models the inter-modal interactions between visual and textual representations and adopts a Target-Image (TI) matching layer to obtain a target-sensitive visual. (5) CapBERT, a BERT-based method ([Khan and Fu, 2021](#)) which translates the image to caption and fuses the caption with input text-aspect pair through the auxiliary sentence. (6) CapBERT-DE, which replaces BERT to BERTweet ([Nguyen et al., 2020](#)) in CapBERT. (7) VLP-MABSA ([Ling et al., 2022](#)), which is a task-specific pre-training vision-language model for MABSA.

### 4.3 Experimental Results and Analysis

We compare our methods with the above baselines on Twitter-2015 and Twitter-2017 datasets, where FITE-DE is the model that replaces BERT with BERTweet-base, and FITE-DE-Large is with BERTweet-Large. The experimental results are shown in Table 4. The best scores on each metric are marked in bold.

Our methods perform better in these two datasets compared with all the image-only, text-only and

Modalities	Method	Twitter-face	
		Acc	Macro-F1
Text	BERT	67.02	63.32
	CapBERT	67.52	64.33
Text+Image	FITE	69.50 (+1.98%)	66.89 (+2.56%)
	FITE-DE	74.20(+6.68%)	72.12(+7.79%)

Table 5: Experiment results on Twitter-face dataset.

Method	Twitter-2015		Twitter-2017	
	Acc	Macro-F1	Acc	Macro-F1
FITE	<b>78.49</b>	<b>73.90</b>	<b>70.90</b>	<b>68.70</b>
w/o Gate Mechanism	77.47	72.70	68.93	67.12
w/o Fine-grained Alignment	77.89	72.90	70.03	68.32
w/o Image Caption of Scene	76.56	72.28	70.01	67.95

Table 6: Ablation study of our FITE model.

multimodal baselines. This demonstrates the effectiveness of the proposed face-sensitive image-to-emotional-text translation method. FITE-DE and FITE-DE-Large method show improvements over typical approaches and outperform the SOTA method CapBERT-DE on the macro-f1 score by about 0.4% and 0.9% on the Twitter-2015 dataset, 1.8% and 2.8% on Twitter-2017 dataset, respectively. And our methods also show competitive performances compared with the VLP-MABSA model. The performance of the FITE-DE-Large model is better than that of the FITE-DE model and the BERT-based FITE model, showing that our model performs better with a stronger language model, which illustrates that the context modeling ability of the language model has a great influence in the fusion stage. And compared with the baseline model, the improvement of the FITE-DE model on the Twitter-2017 dataset is more significant than the improvement on the Twitter-2015 dataset. We conjecture this is because the Twitter-2017 dataset contains more data with facial emotions, the proportion of images that contains face is 15% higher than that of the Twitter-2015 dataset. While this phenomenon is not obvious in the model comparison of the base version, we speculate that it is due to the weak text context modeling ability of the base version pre-trained language model.

As shown in Table 5, on the Twitter-face dataset, the text-only BERT model performs poorly. Compared with the baseline model CapBERT, our models achieve a significant improvement. This proves that our model has the ability to model the facial expressions in the image and also proves the importance of explicitly modeling emotional cues in visual content.

Label	Positive	Positive	Negative	Neutral
Image				
Text	(a) Some of that Dodger baseball * @ [alyssajacinto] Positive.	(b) # Lionel Messi 's bride # [antonellaRoccuzzo] Positive first lady of football	(c) Bill Clinton Fired [FBI] Negative Director One Day B4 Vince Fosters Death # PJNET	(d) Nancy Ajram during the [Beirut Cultural Festival] Neutral; beautiful as always .
Image Caption	(a) Two women in a field with a dog.	(b) A man and a woman pose for a picture.	(c) An old man in a suit and tie is giving a speech.	(d) A woman is talking on her cell phone.
Face Description	(a) There are 2 people in the image. A 33-year-old Woman with a happy expression. alyssajacinto with a happy expression.	(b) There are 2 people in the image. Lionel Messi with a neutral expression. antonellaRoccuzzo with a happy expression.	(c) There are 1 people in the image. Bill Clinton with a angry expression.	(d) There are 1 people in the image. Nancy Ajram with a happy expression.
BERT	Neutral (✗)	Neutral (✗)	Negative (✗)	Positive (✗)
CapBERT	Neutral (✗)	Neutral (✗)	Neutral (✗)	Neutral (✗)
Ours	Positive (✓)	Positive (✓)	Negative (✗)	Neutral (✓)

Figure 3: Case analysis on BERT, CapBERT and our FITE model.

#### 4.4 Ablation Study

To further study the influence of the individual components of our method, we perform comprehensive ablation analysis using the BERT-base version FITE on Twitter-2015 and Twitter-2017 datasets. The results are shown in Table 6. Firstly, without the gate mechanism, we concatenate the pooler outputs of language models as input of a linear classification layer to predict the sentiment label of the target aspect. The performance drops a lot due to the noise in the image-to-emotional-text generation stage. The accuracy and macro-f1 score drop by about 1% on the Twitter-2015 dataset, and the accuracy drops by 1.97%, the macro-f1 score drops by 1.58% on the Twitter-2017 dataset. This verifies that the gate mechanism helps reduce noise and extract better features. Secondly, we can find that removing the fine-grained alignment module leads to a decline of about 1%. This observation indicates the alignment of the aspect with visual emotional clues is essential. Thirdly, we also explore the influence of removing the image caption of the scene, and the model significantly performs worse. This validates that image-to-text translation helps to promote image-text fusion.

**Face Description.** We study the impacts of the different face descriptions generated patterns and the contribution of the different facial attributes on our FITE model. Table 7 depicts the results of different face descriptions. We test 2 settings for

Method	Twitter-2015		Twitter-2017	
	Acc	Macro-F1	Acc	Macro-F1
single face	76.60	71.19	68.53	66.92
No Pattern	77.33	72.43	69.36	67.54
FITE	<b>78.49</b>	<b>73.90</b>	<b>70.90</b>	<b>68.70</b>
w/o Age	78.20	73.28	70.62	68.44
w/o Race	77.82	73.44	70.25	67.97
w/o Gender	77.82	73.07	69.93	67.57
w/o Emotion	76.22	71.44	68.80	66.35

Table 7: Results of different face descriptions.

the face description generated pattern. First, we only use the face with the highest emotional prediction confidence in multi-face cases to generate face descriptions. In the next, we remove the manual pattern and directly use all the predicted facial attributes as face descriptions. As we can see, the single-face setting drops the performance by about 2.5%, and the next setting drops by nearly 1.5%. As for the facial attributes, we remove one of the four attributes and generate face descriptions to test the influence. We find that without the emotion of face leads to the most significant decline of more than 2.1%, and removing any attribute leads to a decline in performance. As such, the design of the face description generated pattern must be treated with care.

**Visual Feature.** To verify the strength of our face-sensitive image-to-emotional-text cross-modal translation method compared with the

Method	Twitter-2015		Twitter-2017	
	Acc	Macro-F1	Acc	Macro-F1
Bert	74.25	70.04	68.88	66.12
Res-BERT+BL	75.02	69.21	69.20	66.48
Res-BERT_face+BL	76.37	71.34	69.25	67.42
FITE	<b>78.49</b>	<b>73.90</b>	<b>70.90</b>	<b>68.70</b>
FITE+face feature	77.37	73.01	69.35	67.32
FITE+image feature	76.54	71.83	69.25	67.24

Table 8: Ablation study of visual feature.

feature-level fusion strategies, we also study the performance of the visual feature of the image and face on our models. The Res-BERT+BL model and the FITE+image feature model fuse the image feature from ResNet with the text feature from language models. And the Res-BERT\_face+BL model and FITE+face feature only use the face regions in the image. We can infer from Table 8 that adding the visual feature brings improvement compared to the text-only BERT model, demonstrating that adding visual input can bring additional effective features. Specifically, our method outperforms the models with the visual feature. We argue that it is because we have explicitly used the visual emotional cues by image-to-emotional-text translation. This results in the noise impact outweighing the information gain after adding visual features that lack cross-modal alignment.

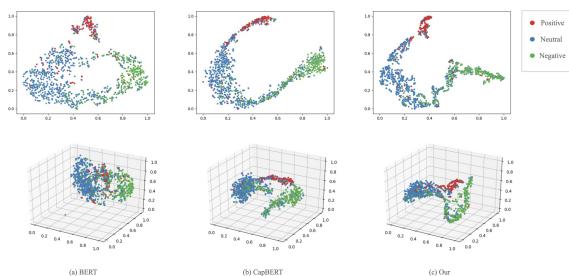


Figure 4: Visualization for distributions of different sentiments in learnt embedding space. The red, blue and green dots represent positive, neutral and negative sentiment respectively.

## 4.5 Case Study

Figure 3 shows the comparison between the predictions of the BERT, CapBERT, and our model on four samples. Since both our model and CapBERT are well interpretable, we also show the generated textual image captions and face descriptions in the four samples. First, in sample (a), we can see with the help of the face description that includes positive emotional word *happy*, our method

can correctly predict the positive sentiment polarity while the other model with neutral text and image caption makes a wrong prediction. Likewise, in sample (b), there are multiple aspects and multiple faces with inconsistent facial expressions. For such examples, the fine-grained alignment of an aspect and its matched emotional face description is particularly important. Through the fine-grained alignment and modification module in our method, the generated expression description can directly align the correct aspect “antonellaRoccuzzo” to the happy facial emotion and is helpful for the aspect sentiment classification. Moreover, in sample (c), the aspect to be judged “FBI” is a non-human object (belonging to the name of the institution), but our method can still identify helpful emotional cues from the image input with human faces. Similarly, in sample(d), benefitting from the help of the gate mechanism, our model is also able to filter out the emotional impact of non-current matching aspects. These four samples further confirm our motivations is generally useful in one-face and multiple faces cases. And our proposed method can capture emotional cues in the image and cross-modal align the emotional cues with target aspects.

## 4.6 Visualization of Embedding Space

We provide a visualization for distributions of sentiments in the embedding space where the (a) sub-figure on Figure 4 illustrates the embedding space learned by BERT while the (b) sub-figure learned by CapBERT and the (c) sub-figure learned by our method. We selected the hidden layer vector that removed the final classification layer and used the T-SNE algorithm to transform the 768-dimensional vector into two-dimensional and three-dimensional feature points. From the comparison chart we can find that compared to the Bert and CapBERT model, our method can make the clusters of different categories more distinct and make the point’s in the same sentiment cluster closer. In contrast, it is clear that the embedding space learned by our method can distinguish positive, neutral, and negative sentiments effectively.

## 5 Conclusion

In this paper, we propose the face-sensitive image-to-emotional-text translation method for multi-modal aspect-based sentiment analysis by firstly introducing the facial emotions in images as the visual emotional cues. We identify the challenge

that the multimodal aspect-based sentiment analysis with weak supervision for fine-grained text-image alignment, and propose a direct and effective method to align the text-modal target aspect and the facial emotions in visual content. Our method achieve state-of-the-art performance on Twitter-2015 and Twitter-2017 datasets. And we also build a new Twitter-face aspect-level sentiment dataset to evaluate our model. The results show that our method outperforms a series of benchmark models and demonstrate the superiority of our method in capturing the visual emotional cues and cross-modal alignment on multimodal sentiment data.

## Acknowledgments

We thank the anonymous reviewers for their insightful comments and suggestions. This work was supported by the National Key RD Program of China via grant 2020AAA0106501 and the National Natural Science Foundation of China (NSFC) via grant 62176078.

## Limitations

The major limitation is that our method is not suitable for the sample without facial emotion in the visual modality. One of the main reasons is that it is difficult and lacks interpretability to identify visual emotional cues for the images without faces. Since Twitter images in the open domain are all-encompassing, the emotional impact of the visual objects appearing in the images is affected by many factors. Even for humans, the emotional perception of the same visual object will have a large deviation due to their cognitive levels. This leads to the lack of a clear definition of image sentiment in the field of sentiment analysis, and there are no other datasets related to Twitter image sentiment analysis. Our method can utilize facial expressions, which is a relatively obvious and strong visual emotional signals with a fine emotional consistency. In the future, we will conduct further research from the perspectives of aesthetics, common sense knowledge, etc. Another limitation is that the generated face description is based on artificial templates. We have also tried to use the generative model to fuse the textual context and expression information, but it is limited by the performance of the generative model and the lack of public emotional image-to-text datasets.

## Ethics Statement

Our work complies with Twitter’s data policy, and all the codes and datasets used in our work comply with the ethics policy. There is no difference in our code for facial expressions of different ages, genders, and races.

## References

- Navneet Dalal and Bill Triggs. 2005. Histograms of oriented gradients for human detection. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR’05)*, volume 1, pages 886–893. Ieee.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. *BERT: Pre-training of deep bidirectional transformers for language understanding*. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Feifan Fan, Yansong Feng, and Dongyan Zhao. 2018a. 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.
- Shaojing Fan, Zhiqi Shen, Ming Jiang, Bryan L Koenig, Juan Xu, Mohan S Kankanhalli, and Qi Zhao. 2018b. Emotional attention: A study of image sentiment and visual attention. In *Proceedings of the IEEE Conference on computer vision and pattern recognition*, pages 7521–7531.
- 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.
- 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.
- 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.

- 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.
- Shan Li, Weihong Deng, and JunPing Du. 2017. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2852–2861.
- Yuncong Li, Cunxiang Yin, and Sheng-hua Zhong. 2020. Sentence constituent-aware aspect-category sentiment analysis with graph attention networks. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 815–827. Springer.
- 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), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 2149–2159. Association for Computational Linguistics.
- Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32.
- 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*, pages 9–14.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Satsky, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR.
- Sefik Ilkin Serengil and Alper Ozpinar. 2020. Lightface: A hybrid deep face recognition framework. In *2020 Innovations in Intelligent Systems and Applications Conference (ASYU)*, pages 23–27. IEEE.
- Sefik Ilkin Serengil and Alper Ozpinar. 2021. Hyper-extended lightface: A facial attribute analysis framework. In *2021 International Conference on Engineering and Emerging Technologies (ICEET)*, pages 1–4. IEEE.
- Jiawei Wang, Zhe Liu, Victor Sheng, Yuqing Song, and Chenjian Qiu. 2021. Saliencybert: Recurrent attention network for target-oriented multimodal sentiment classification. In *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*, pages 3–15. Springer.
- Yequan Wang, Minlie Huang, Xiaoyan Zhu, and Li Zhao. 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.
- Nan Xu, Wenji Mao, and Guandan Chen. 2019. 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 *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2514–2523, Melbourne, Australia. 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-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.
- Peisong Zhu, Zhuang Chen, Haojie Zheng, and Tieyun Qian. 2019. Aspect aware learning for aspect category sentiment analysis. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 13(6):1–21.

## A Appendix

### A.1 Face Description Rewriting

The aspect-sensitive alignment and modification module is only used for cases with multiple facial expressions. For the cases with only one face and one aspect, we directly use the aspect to modify the face description generated in the face-sensitive image-to-emotional-text translation module and add a sentence “There are 1 people in the image.”, and take the new face description as input to gate-based multimodal fusion module. For the cases with multiple faces, for example, the case in Figure 2, for the aspect “antonellaRoccuzzo”, we connected the aspect with face descriptions: “antonellaRoccuzzo, A 28-year-old Indian woman with a happy expression” and “antonellaRoccuzzo, A 38-year-old white man with a neutral expression”. We take the connected new sentences as the textual input of CLIP’s text encoder and the image as the visual input of CLIP’s image encoder, and calculated the cosine similarities of visual features with

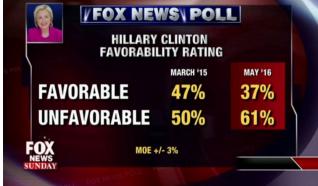
Label	Negative	Positive	Negative
Image			
Text	(a) # MTVStars Lady Gaga # MTVStars Lady Gaga Fox News Poll : [HillaryClinton] Negative favorability rating .	(b) # TheComeback reunion panel at @ ATXFestival with @ LisaKudrow , Michael Patrick King @danbucatinsky [Laura Silverman] Positive	(c) Donald Trump ' s victory proves [Republican] Negative voters want resentful nationalism , not principled conservatism ...
Image Caption	(a) A sign with a map and a a picture of a clock.	(b) A group of people sitting at a table with a sign.	(c) A man in a suit and tie holding a cell phone.
Face Description	(a) There are 0 people in the picture.	(b) There are 3 people in the picture. Michael Patrick King with an angry expression. Laura Silverman with a fear expression. danbucatinsky with a angry expression.	(c) There are 1 people in the picture. Donald Trump with a happy expression.
Prediction	Positive	Negative	Neutral

Figure 5: Error cases for FITE model.

the two textual features. After that, we can select the first face description with a higher score and modify the face description. Through our experiments and case analysis, we found that CLIP has the ability to distinguish aspect and face description according to the image, and in this case, we can get the score of the first face description is 0.615 and the second is 0.385 by CLIP-ViT-B/16 model. Thus, we can obtain the fine-grained alignment of aspect and faces(face description) in cases with multiple faces.

Since we generate a fixed face description for each image, there are different aspects corresponding to the same image-text pair in the datasets. For example, the case in Figure 2 and the fourth example in Figure 3 contain multiple aspects in the textual input, each aspect can be matched with its corresponding face description. If multiple aspects point to the same face, the aspect will be connected with the face description respectively, and as inputs to the CLIP model, the aspect with the highest score will be selected for rewriting.

The aspect-sensitive alignment and modification module is only used for cases with multiple facial expressions. For the cases with only one face and one aspect, we directly use the aspect to modify the face description generated in the face-sensitive image-to-emotional-text translation module and add a sentence “There are 1 people in the image.”, and take the new face description as input to gate-based multimodal fusion module. And we gener-

ated the sentence “There are 0 people in the image.” as the face description for the cases without a face. As for those images with only scenes, logos, or buildings, we tried to annotate them manually and we found that it’s hard to identify their sentiment polarities even for humans. Therefore, we only choose to supplement the semantic information that no one appears in the image and avoid introducing wrong emotional noises.

## A.2 Analysis of Cases with Multiple Faces

According to the statistics, there are 23.2% images in the Twitter-2015 dataset with one face and 14.7% images with multiple faces, and there are 28.9% images in the Twitter-2017 dataset with one face and 22.4% images with multiple faces. Our FITE-base model achieves accuracy: 66.97% and F1-score:63.88% on the examples in the Twitter-face dataset with one face and accuracy: 73.21% and F1-score: 71.86% with multiple faces. And the CapBERT model achieves accuracy: 65.13% and F1-score: 62.82% on the examples in the Twitter-face dataset with one face and accuracy: 67.50% and F1-score: 65.72% with multiple faces. The experiment results can show the effectiveness of the aspect-sensitive alignment and modification module in cases with multiple faces.

## A.3 Error Analysis

We conducted an error analysis for our main model FITE. Figure 5 shows some failed examples that

are categorized into three types: (1) No useful face in the image. (2) Wrong predictions of facial attributes. (3) In specific situations, facial expressions cannot reflect the complete visual sentiment. The example (a) in Figure 5 shows a failed example that resulted from no useful face in the image. Due to no facial information being detected, emotional clues (“Clinton’s support declines”) in the image are ignored. And the example (c) in Figure 5 mistakenly identifies Trump’s expression as “happy”, which misleads the prediction of sentiment. The above two types of errors are limited by the image quality and facial expression recognition tools. Especially for blur images, it is difficult to extract correct facial expressions from blur images. The example (b) in Figure 5 shows a special scene. The text describes the reunion of old friends, but everyone in the image has a solemn expression. The sentiment polarities in the image and the text are conflicted, as well as some cases of sarcasm. In such cases, visual sentiment is affected by text, facial expressions cannot reflect the complete visual emotional cues. The loss of information caused by transferring images to text makes cross-modal fusion incompletely.