Computers, Materials & Continua

Tech Science Press

DOI: 10.32604/cmc.2023.038220
Article

Improving Targeted Multimodal Sentiment Classification with Semantic Description of Images

Jieyu An*, Wan Mohd Nazmee Wan Zainon and Zhang Hao

School of Computer Sciences, Universiti Sains Malaysia, Penang, 11800, Malaysia
*Corresponding Author: Jieyu An. Email: anjieyu@student.usm.my
Received: 02 December 2022; Accepted: 16 March 2023

Abstract: Targeted multimodal sentiment classification (TMSC) aims to identify the sentiment polarity of a target mentioned in a multimodal post. The majority of current studies on this task focus on mapping the image and the text to a high-dimensional space in order to obtain and fuse implicit representations, ignoring the rich semantic information contained in the images and not taking into account the contribution of the visual modality in the multimodal fusion representation, which can potentially influence the results of TMSC tasks. This paper proposes a general model for Improving Targeted Multimodal Sentiment Classification with Semantic Description of Images (ITMSC) as a way to tackle these issues and improve the accuracy of multimodal sentiment analysis. Specifically, the ITMSC model can automatically adjust the contribution of images in the fusion representation through the exploitation of semantic descriptions of images and text similarity relations. Further, we propose a target-based attention module to capture the target-text relevance, an image-based attention module to capture the image-text relevance, and a target-image matching module based on the former two modules to properly align the target with the image so that fine-grained semantic information can be extracted. Our experimental results demonstrate that our model achieves comparable performance with several state-of-the-art approaches on two multimodal sentiment datasets. Our findings indicate that incorporating semantic descriptions of images can enhance our understanding of multimodal content and lead to improved sentiment analysis performance.

Keywords: Targeted sentiment analysis; multimodal sentiment classification; visual sentiment; textual sentiment; social media

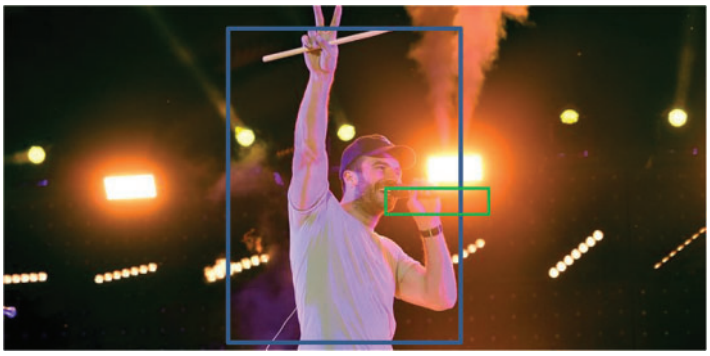
1 Introduction

With the rise in popularity of social media, an increasing number of users use multimodal posts to express their emotions or opinions (i.e., many posts contain both text and related images). Effective sentiment analysis of massive and multimodal social media data can aid in comprehending public sentiment and opinion trends, providing a scientific foundation for government and corporate decision-making [1–3]. In comparison to traditional textual sentiment analysis [4], performing

sentiment analysis utilizing data from different modalities presents a number of opportunities and challenges.

The targeted multimodal sentiment classification (TMSC) is a fine-grained task of natural language processing to extract sentiment polarity (e.g., positive, negative, or neutral) that has become one topic of increasing research interest over the past few years. Automatically identifying the underlying attitude of targeted entities (i.e., aspects) in a sentence and image pair is the goal of targeted multimodal sentiment classification. As shown in Table 1, the targeted entity is expected to be extracted from the multimodal post (i.e., *SamHunt*: *positive*).

Table 1: An example of the TMSC task in the Twitter dataset

Textual modality	# **SamHunt** performs at Stagecoach # MusicFestival 2016!
Visual modality	
Target polarity	SamHunt: *positive*

Many approaches have been proposed in recent years to perform sentiment classification for TMSC and have gained attention [5–8]. Although these studies have shown that combining textual and visual modality information can improve performance on the sentiment classification task, they have the following limitations:

(1) These studies perform sentiment classification by fusing only text and image representation. They do not take into account the visual modality's contribution to the fusion. In other words, there is not always consistency between images and text in terms of sentiment tendencies, which can serve sentiment classification more effectively if they are consistent but can compromise its accuracy if they are inconsistent.

(2) These methods exploit only the representational information of the image, ignoring the supplementary information that comes from the semantic description of the image. By using semantic descriptions of images, we can recognize information such as objects, positions, and actions in images. For instance, based on the image in Table 1, we can generate the following image descriptions: *A man is holding up a microphone to take a picture*. In this description, "*a man*" is aligned with the visual modality, which is indicated by the blue rectangle, and it is also aligned with *SamHunt*, which is indicated by the red underline in the textual modality. According to our hypothesis, the semantic description of images helps us understand what they are about and may tell the model to focus on the parts of the image that match the given target while reducing noise in other parts.

In order to address the above limitations, we propose an Improving Targeted Multimodal Sentiment Classification (ITMSC) model based on the semantic description of images for the TMSC task. The following is a condensed summary of the most important contributions made by this paper:

- To the best of our knowledge, this is the first time that semantic descriptions of images have been used to establish the information interaction with images and text to obtain more semantic information for the TMSC task.
- To adjust the contribution of the visual modality in the fusion representation of different modalities, we propose a method to automatically and dynamically adjust the input of the image based on the similarity between the image description and the text.
- To obtain the finer-grained semantic alignment information between different modalities, we develop three matching modules that effectively reduce redundant data and extract meaningful information.

Experiments conducted on two multimodal sentiment datasets have shown that our ITMSC outperforms superior performance compared to the most advanced models currently available. Furthermore, our model generates insightful and interpretable visualizations that highlight the importance of semantic descriptions of images for the TMSC task.

2 Related Work

2.1 Multimodal Sentiment Analysis

Based on the content generated by users, sentiment analysis attempts to determine the emotional orientation (e.g., negative, neutral, or positive) [9]. It enables machines to comprehend human emotions and react in the proper manner. The need to automate the evaluation of customers' feelings about products or services is on the rise. It has made significant progress in areas such as natural language processing and artificial intelligence.

Sentiment analysis based on textual content alone [10–13] is no longer sufficient in today's social media environment, as users often share and discuss things mostly presented in a multimodal form. As a new part of the field of multimodal machine learning, researchers are paying more attention to multimodal sentiment analysis.

Early multimodal works are predominately handcrafted. Nevertheless, handcrafted features are usually created with limited human knowledge and cannot fully describe the highly abstract nature of emotions, leading to suboptimal results [14]. In the last few years, there have been significant advances in the field of multimodal sentiment analysis due to the emergence of deep learning models [15–18]. Most of these studies also reached similar conclusions, correlating and supplementing the information contained in data from various modalities to achieve a more accurate classification of sentiment than analysis based on a single modality. In general, there is a close association that exists between the text and image in posts from users on social media platforms [19–21]. However, we cannot directly apply these coarse-grained multimodal sentiment classification methods to our targeted sentiment classification tasks. Consequently, fine-grained multimodal sentiment analysis is the primary focus of our work.

2.2 Targeted Multimodal Sentiment Analysis

Targeted multimodal sentiment analysis is a form of fine-grained sentiment analysis task. It is a relatively new field for understanding the sentiment of a particular entity or topic from multiple sources of information, allowing for a more comprehensive view of the sentiment. By leveraging the information from various modalities, targeted multimodal sentiment analysis can be used to gain a better understanding of the sentiment expressed towards a particular entity or topic.

In the literature, this task has attracted considerable interest from researchers. In 2019, Xu et al. [5] proposed a multi-interactive memory network to independently model text and image data and learn the interactive influences of cross-modality data. In 2020, Yu et al. [6] proposed an entity-sensitive attention and fusion network (ESAFN) to study intra-modality and inter-modality interactions in a sentence and image pair for targeted sentiment classification. In addition, with the widespread use of pre-trained models, Yu et al. [7] proposed a target-oriented multimodal bidirectional encoder representation from Transformers (TomBERT) architecture to capture the relationship between target, text, and image for the TMSC task. More recently, to overcome the problem of short texts with little information, Khan et al. [8] introduced a two-stream model in 2021 that first translates images into auxiliary sentences and then fuses the input sentence and auxiliary sentences for the TMSC task. In 2022, Ye et al. [22] introduced a sentiment-aware multimodal pre-training (SMP) framework to address the lack of attention to sentiment signals in most existing multimodal pre-trained models, which mainly focus on general lexical and/or visual information. In addition, Yu et al. [23] proposed a multi-task learning architecture named coarse-to-fine-grained Image-Target Matching network (ITM) in order to capture both coarse-grained and fine-grained image-target matching.

Our research is similar to Khan's approach, as we also employ their image descriptions as auxiliary sentences for the TMSC task. However, we argue that merely combining the input sentence and image description, without incorporating the image feature, may lead to a loss of crucial sentiment-related information. To address this issue, our research focuses on incorporating all three elements—input sentence, image description, and image feature—into the sentiment analysis process. In order to extract in-depth semantic information from the feature data, we further propose an attention-based fusion mechanism, which we experimentally validated. Our results demonstrate a significant improvement over Khan's approach.

3 Methodology

In contrast to existing approaches, we utilize not only text and image information for targeted multimodal sentiment classification but also image description as supplementary information to investigate image and text interactions. In this section, we define the tasks of TMSC and present the overall architecture of the proposed ITMSC model, as shown in Fig. 1. We then elaborate on the details of each module in ITMSC for targeted multimodal sentiment classification. Our research is mostly about social apps where user-generated content is a paragraph of text with an image.

Task Formulation: Given one multimodal sentiment sample M as a pair of sentence and image, it consists of an associated image I, an opinion target with m words $T = (t_1, t_2, \ldots\ t_m)$, and a sentence with n words $S = (s_1, s_2, \ \ldots\ s_n)$. In this paper, we attempt to make a prediction regarding the polarity label y, which can be either *neutral*, *negative*, or *positive*, of each opinion target mentioned in M.

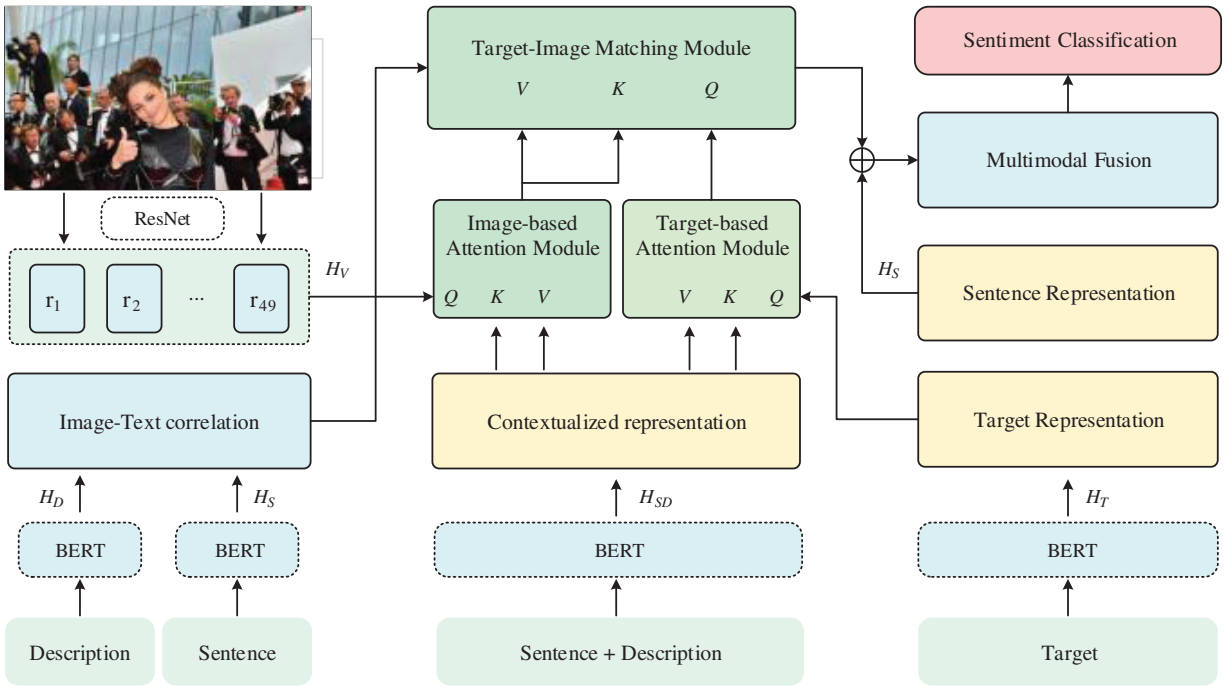


Figure 1: The overall architecture of the proposed ITMSC

3.1 Input Representations Extraction

Textual Representation Extraction. When dealing with textual representation extraction, Bidirectional Encoder Representations from Transformers (BERT) [24] is widely used as a language representation model that can capture the full semantic information of a sentence and also discover association features between words through the context [25]. Following previous research [7], given a sentence S as input, we first extract the input target T from the sentence and replace it with a special token $\$T\$$. To enable BERT to process these text sequences, we add a special classification token [CLS] at the beginning of the sentence and a special segmentation token [SEP] between different text sequences. Then, we utilize the fine-tuned BERT to obtain the hidden representation H_S and H_T, as illustrated in Eqs. (1) and (2). Similarly, using the same fine-tuned BERT we also get the image description presentation H_D of image description D and the concatenated textual presentation H_{SD}, as illustrated in Eqs. (3) and (4).

$$H_S = BERT\left([CLS]\, S\, [SEP]\, T [SEP]\right),\ H_S \in \mathbb{R}^{n\times d} \tag{1}$$

$$H_T = BERT\left([CLS]\, T [SEP]\right),\ H_T \in \mathbb{R}^{t\times d} \tag{2}$$

$$H_D = BERT\left([CLS]\, D [SEP]\right),\ H_D \in \mathbb{R}^{m\times d} \tag{3}$$

$$H_{SD} = BERT\left([CLS]\, S [SEP] T + D [SEP]\right),\ H_{SD} \in \mathbb{R}^{(n+m)\times d} \tag{4}$$

where n, t, and m are the lengths of S, T and D, respectively; d is 768-dimensional hidden state.

Visual Representation Extraction. When it comes to extracting features of the images, we adopt one of the most advanced image recognition models, Residual Network 152 (ResNet) [26], pre-trained on ImageNet [27] classification, to obtain the image representation. Before feeding the image into the

model, we first rescale it to 224 × 224 pixels. The visual representation is then obtained from the last convolutional layer of ResNet:

$$ResNet(I) = \left\{r_j | r_j \in \mathbb{R}^{2048}, j = 1, 2, \ldots, 49\right\}, \tag{5}$$

where 49 is the number of regions with the same size that have been split from the image, and r_j is a 2048-dimensional vector representing each region as depicted in the upper left-hand corner of Fig. 1. Further, we use a linear function to map each region in the image to the same space as the text representation:

$$H_V = W_v ResNet(I), \tag{6}$$

where $W_v \in \mathbb{R}^{d \times 2048}$ is a weight parameter.

3.2 Multimodal Interaction

While image descriptions can provide supplementary semantic information, they are not always advantageous when attempting to classify sentiment. An excess of semantic information can increase background noise and reduce classification accuracy. To address this problem, we propose three main modules: (1) a Target-based Attention Module to capture the target-text relevance; (2) an Image-based Attention Module to capture the image-text relevance; and (3) a Target-Image Matching Module to align the target-text with the image-text to obtain fine-grained semantic information.

Target-based Attention Module. Since the input target is extracted from a sentence, as previously mentioned, we argue that it is unsuitable for direct use in sentiment analysis tasks due to a lack of contextually relevant semantic information. Consequently, we apply the cross-modal Transformer layer [28] to modality interaction between the input target and the concatenated text, where the representations of the input target H_T serve as queries, and the representations of the contextualized text H_{SD} serve as keys and values:

$$H'_T = \text{Cross} - \text{ATT}\left(H_T, H_{SD}, H_{SD}\right), \tag{7}$$

where $H'_T \in \mathbb{R}^{d \times t}$ is the generated target-based attention representation.

Image-based Attention Module. To obtain the semantic information jointly presented by the image and the concatenated text, we use another cross-modal Transformer layer to modality the interaction between the image and the concatenated text, where the representations of the image H_V serve as queries and the representations of the contextualized text H_{SD} also serve as keys and values:

$$H'_V = \text{Cross} - \text{ATT}\left(H_V, H_{SD}, H_{SD}\right), \tag{8}$$

where $H'_V \in \mathbb{R}^{d \times 49}$ is the generated image-based attention representation.

Target-Image Matching Module. Based on the Target-based Attention Module and the Image-based Attention Module, which both work to extract key information and reduce the impact of irrelevant information, the Target-Image Matching Module aims to identify target-based attention representation aligned with image-based attention representation. Specifically, we use target-based attention representation H'_T as queries and image-based attention representation as H'_V keys and values:

$$H_T'' = \text{Cross} - \text{ATT}(H_T', H_V', H_V'), \tag{9}$$

where $H_T'' \in \mathbb{R}^{t \times d}$ is the matched representation.

Although the representation fused in the above way reduces data redundancy, the contribution of image modality to the fusion is not considered. Incorrect correlations between the different modalities could result in the combination of unrelated information, thereby reducing the accuracy of the final classification results. Consequently, our model first measures the relationship between the text and the image by computing the similarity between the text and the image description. Specifically, given that BERT captures a wealth of semantic information [29], we directly create two sentence vectors H_S' and H_D', by averaging all of the word vectors in the final hidden representation layer of H_S and H_D, and then we compute the cosine similarity of the two sentence vectors:

$$similarity\left(H_S', H_D'\right) = \frac{H_S' \cdot H_D'}{\|H_S'\| \left\|H_D'\right\|}. \tag{10}$$

Then, we dynamically regulate the contribution of the image modality to the fusion process according to the degree of similarity. Thus, we construct a visual filter matrix $G \in \mathbb{R}^{d \times 49}$ based on the similarity score in Eq. (10), which indicates the relevant score between the text and the image. Then the filtered image representations can be obtained from the visual filter matrix:

$$H_V' = H_V' \odot G, \tag{11}$$

where $\odot$ represents the element-by-element multiplication. Consequently, H_V' in Eq. (9) requires revision of Eq. (11).

3.3 Multimodal Sentiment Classification

With the representations H_T'' generated from the Target-Image Matching Module, a late feature fusion is performed to concatenate them with sentence representation H_S, and feed them to a self-attention layer for multimodal fusion:

$$H = \text{Self} - \text{ATT}\left(concat\left(H_T'', H_S\right)\right). \tag{12}$$

Finally, the softmax layer receives the representation of the first token $H^{[0]}$ to produce the prediction result y after layer normalization (LN).

$$y = Softmax\left(LN\left(W^T H^{[0]} + b\right)\right). \tag{13}$$

As the classification loss for the purposes of model training, the cross-entropy loss is used by the majority of multimodal sentiment analysis methods:

$$\mathcal{L}^{ITMSC} = -\frac{1}{N}\sum_{i=1}^{N} logP\left(y^i\right), \tag{14}$$

where N is the number of samples for the classification task, y^i is the representation of the probability distribution of the final target classification that was obtained by our model.

4 Experiment

4.1 Datasets

We evaluate our ITMSC model using the Twitter-2015 and Twitter-2017 public multimodal sentiment databases. Both of these databases are freely available to the public. The two Twitter databases that Yu et al. [7] presented include multimodal tweets posted to Twitter. Each multimodal tweet contains a sentence, an accompanying image, the target in the sentence, and the sentiment

polarity of each target. Yu et al. categorized each target as either negative, neutral, or positive. We followed the same partitioning of the dataset as several recent publications [7] and [8] to ensure an equitable evaluation. The characteristics of the datasets are summarized in Table 2.

Table 2: The statistics of the multimodal sentiment databases for Twitter-2015 and Twitter-2017

Label	Twitter-2015			Twitter-2017		
	Train	Validation	Test	Train	Validation	Test
Positive	928	303	317	1508	515	493
Neutral	1883	670	607	1638	517	573
Negative	368	149	113	416	144	168
Total	3179	1122	1037	3562	1176	1234

4.2 Experimental Settings

We obtained the text representation by fine-tuning a BERT-base model, and the image representation was obtained using a pre-trained ResNet152. With the help of Google Colaboratory, we were able to conduct our experiment. We would like to express our gratitude for the opportunity to use this resource, as it made our research and experimentation significantly easier and more efficient. As for the hardware, we used a powerful Tesla Graphics Processing Unit (GPU) with 16 GB of Random Access Memory (RAM). Pytorch was used to realize the framework of the model. Pytorch is an ideal deep learning framework for quickly and accurately building and deploying sophisticated machine learning models. Refer to Table 3 for details on the hyper-parameters used in our model, such as maximum text length, description length, target length, training batch size, and learning rate.

Table 3: Settings of important parameters

Hyperparameters	Twitter-2015	Twitter-2017
Maximum text length	64	64
Maximum description length	64	64
Maximum target length	16	16
Attention head	12	12
Hidden dimension	768	768
Learning rate	2e-5	4e-5
Training batch size	32	16
Training epoch	8	8

4.3 Experimental Results and Analysis

To further validate the efficacy of the proposed ITMSC model for targeted multimodal sentiment classification, we compare our approach to various existing competitive methods. The results of our analysis demonstrate that the proposed model has a strong performance in terms of accuracy and Macro-F1 score. This further validates the efficacy of our proposed model and provides further

evidence that the accuracy of multimodal sentiment classification can be improved using semantic descriptions of images.

We choose three kinds of baselines. The first category is the visual-based ResNet-Aspect model. The second category is some traditional text-based models, including the BERT, the long short-term memory with aspect embedding (AE-LSTM) [30], the deep memory network (MemNet) [31], and the recurrent attention network on memory (RAM) [32]. The third category consists of multimodal models such as the multi-interactive memory network (MIMN) [5], the entity-sensitive attention and fusion network (ESAFN) [6], the target-oriented multimodal bidirectional encoder representation from Transformers (TomBERT) [7], and the exploiting BERT for multimodal target sentiment classification through input space translation (EF-CapTrBERT) [8], the sentiment-aware multimodal pre-training framework (SMP) [22], the coarse-to-fine grained Image-Target Matching network (ITM) [23], and the vision-and-language BERT (ViLBERT) [33].

Table 4 presents a comparison of accuracy and Macro-F1 score for the proposed ITMSC method and other benchmark models on both Twitter-2015 and Twitter-2017 datasets. The baseline models were evaluated using different approaches: TomBERT's performance was predicted using the model generated in [7], while the results of EF-CaTrBERT and ITM were generated by running their provided code. In contrast, the remaining baseline model results were obtained directly from the original papers.

Table 4: Comparison of different methods on two Twitter datasets. Results marked with $*$ represent the average performance over five runs with a seed set ranging from 42 to 46. The marker $\pm$denotes the standard deviation of the results, and the marker † indicates the significant test p-value, which is less than 0.05

Model	Methods	Twitter-2015		Twitter-2017	
		Accuracy	Macro-F1	Accuracy	Macro-F1
Image only	ResNet	59.88	46.48	58.59	53.98
Text only	AE-LSTM	70.30	63.43	61.67	57.97
	MemNet	70.11	61.76	64.18	60.80
	RAM	70.68	63.05	64.42	61.01
	BERT	74.15	68.86	68.15	65.23
Image+Text	MIMN (2019)	71.84	65.69	65.88	62.99
	ViLBERT (2019)	73.76	69.85	67.42	64.87
	TomBERT* (2019)*	76.82 $\pm$ 0.08	71.04 $\pm$ 0.12	70.02 $\pm$ 0.17	67.67 $\pm$ 0.12
	ESAFN (2020)	73.38	67.37	67.83	64.22
	EF-CaTrBERT (2021)*	76.32 $\pm$ 0.85	71.54 $\pm$ 0.73	67.96 $\pm$ 1.16	65.61 $\pm$ 0.94
	SMP (2022)	77.53	72.24	71.15	69.47
	ITM (2022)*	77.38 $\pm$ 0.56	72.43 $\pm$ 1.15	**71.79 $\pm$ 0.32**	**70.38 $\pm$ 0.32**
	ITMSC (Ours)	**78.59 $\pm$ 0.11**†	**74.28 $\pm$ 0.09**†	70.28 $\pm$ 0.06	68.40 $\pm$ 0.07

Table 4 presents a quantitative analysis of our method and its performance compared to other state-of-the-art approaches. Our method demonstrated superior performance on Twitter-2015, outperforming the other methods. Similarly, on Twitter-2017, our method exhibited strong performance, validating the effectiveness of our proposed improvement strategies for the TMSC task. Based on our

careful examination and comparison of the experimental results, we have drawn the following key findings:

Firstly, the method that relies solely on images for sentiment analysis has been shown to perform the worst. This is mostly because images lack the contextual information required for more accurate analysis. Emotions cannot be expressed as explicitly and directly through images as words. As a result, relying solely on the visual modality for sentiment analysis yields poor results. Therefore, it is desirable to utilize more modality information, such as the corresponding textual information, to obtain a more accurate and reliable sentiment analysis.

Secondly, these text-based analysis methods outperform image-based methods. This is due to the fact that emotional cues in textual content are typically more powerful and informative than those in visual data. Additionally, textual data is generally more accessible and easier to process, making it easier to detect and interpret emotional cues. Furthermore, it is evident that BERT consistently outperforms all baselines. We attribute this success to the BERT models' ability to learn from massive datasets, which in turn improves their performance at extracting task-relevant features.

Thirdly, the majority of multimodal approaches outperform their corresponding unimodal baseline approaches by a significant margin. This demonstrates that relying solely on text or images is typically insufficient for sentiment classification. In fact, richer information can be extracted using different modalities, which mutually assist in capturing the implied semantic features through mutual support and fusion between different modalities of data. This can help capture the implied semantic features that are otherwise difficult to access. Ultimately, this can lead to better decision-making and more accurate outcomes.

Fourthly, it's easy to see that MIMN and ESAFN get the worst results when it comes to multimodal methods. This is attributed to the absence of pre-trained models that can extract relevant features from both text and images. Additionally, the pre-trained version of ViLBERT did not perform as well, which may be due to its failure to explicitly model the interaction between text and images at the aspect level. Unlike TomBERT, SMP, and ITM, which rely on extracted image and text representations for sentiment classification, EF-CaTrBERT takes a different approach by combining both text and image descriptions. In contrast, our ITMSC model not only combines text and image but also incorporates image descriptions, resulting in superior performance compared to the several most advanced models currently available. This demonstrates the validity of our hypothesis that image descriptions contain rich semantic information and help comprehend the content of images, thereby enabling the model to align a given target with relevant image regions while reducing noise in irrelevant regions.

4.4 In-depth Analysis

Ablation Study. We performed ablation analysis experiments to assess the individual contributions of the various modules to the performance of our overall ITMSC model. The results of the ablation analysis were used to identify the best combination of modules for the ITMSC model, which helped optimize the performance of the system. Therefore, we removed the Target-based Attention Module, Image-based Attention Module, and Target-Image Matching Module on the basis of the ITMSC model, respectively.

The results presented in Table 5 provide a comprehensive overview of our findings:

(1) The ITMSC model with all modules performs best. The removal of a single module from a system could have a detrimental effect on its accuracy and the F1 score. This is due to the fact

that the model relies on all its components in order to achieve the best possible performance. Without the single module, the model would be missing a critical component, and thus its performance would suffer.

(2) Removing the Target-based Attention Module results to worse performance, demonstrating the importance of extracting target information in the TMSC task. When the Target-based Attention Module is removed, the model loses the ability to extract relevant contextual information from the target. This in turn affects the alignment in the Target-Image Matching Module and leads to a decrease in sentiment accuracy.

(3) The removal of the Image-based Attention Module also leads to a decrease in performance, although the drop in accuracy is relatively small. This further emphasizes that images do not express emotions as directly as the text does. Nevertheless, the Image-based Attention Module is able to capture cross-modal interactions to improve our model's understanding of sentiment across text and visual modalities.

(4) Without the Target-Image Matching Module, performance suffers drastically. By adjusting the image contributions in the fusion representation and performing a fine-grained alignment between the target and the image, it is possible to assist the model in discovering crucial sentiment prediction information.

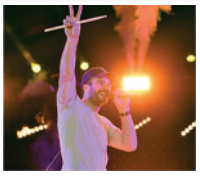
Table 5: Ablation study results on two Twitter datasets

Methods	Twitter-2015		Twitter-2017	
	Accuracy	Macro-F1	Accuracy	Macro-F1
w/o Target-based attention module	74.41	68.57	68.34	65.45
w/o Image-based attention module	77.25	72.96	69.32	66.89
w/o Target-image matching module	75.88	68.69	70.13	67.91
ITMSC	78.59	74.28	70.28	68.40

Case Study. We evaluated the performance of our ITMSC model in targeted multimodal sentiment analysis and compared it with two other models, TomBERT and ITM, which also used images and text for sentiment analysis. Specifically, we conducted experiments on four cases, as presented in Table 6. In case (a), as the image and text are unrelated, there is no apparent alignment between the target *LeBron James* and the image. After analyzing the unrelated image, TomBERT and ITM provided an incorrect prediction. However, our ITMSC model correctly predicted the outcome by obtaining semantic information from the image description. In case (b), the target *Stagecoach* had been given a label for *negative* sentiment. However, TomBERT made an incorrect prediction, probably because it only noticed the facial expression of the person in the image, while our model and ITM model correctly predicted by paying attention to additional image-related information like lights, microphones, and smoke through the image description. In case (c), all three models correctly predicted the target *SteveScalise* with the *negative* sentiment based on the sentiment words and the image representation. Our model was observed to concentrate its attention on the vehicles in the image, which were semantically related to the sentiment expressed in the text. In case (d), where TomBERT made inaccurate predictions because of the inclusion of irrelevant information in the image. Conversely, in the ITM model, the contribution of the visual modality was likely restrained during the feature fusion process due to the discriminative mechanism that suppressed the influence of irrelevant image features. Our proposed model, however, leveraged semantic information provided by image descriptions to

effectively filter out visual noise and concentrate on the relevant features of the hand posture and mouth region, leading to accurate predictions. It is worth noting that despite the incorrect image description labeling the cigarette as a toothbrush, our model still managed to focus on the salient region of the image, demonstrating the efficacy of the Target-Image Matching Module.

Table 6: A case study on some multimodal sentiment samples. The correct and incorrect predictions are denoted by $\surd$ and $\times$, respectively

Visual modality				
Target attention				
Textual modality	**LeBron James** to Produce NBA Documentary.	# SamHunt Performs at **Stagecoach** # MusicFestival 2016.	**SteveScalise** remains in critical condition after shooting at baseball practice.	Petition to have **Jessica Lange** come back for American Horror Story season 6.
Image description	A group of people standing outside of a building.	A man is holding up a microphone to take a picture.	A freeway with a lot of trucks and cars.	A woman is holding a toothbrush in her mouth.
TomBERT	Positive $\times$	Positive $\times$	Negative $\surd$	Positive $\times$
ITM	Positive $\times$	Neutral $\surd$	Negative $\surd$	Positive $\times$
ITMSC	Neutral $\surd$	Neutral $\surd$	Negative $\surd$	Neutral $\surd$
	(a)	(b)	(c)	(d)

5 Conclusion

In this paper, we began by investigating the shortcomings of strategies previously proposed for targeted multimodal sentiment classification. Then we proposed the ITMSC model to improve the targeted multimodal sentiment classification based on the semantic description of the image to address these limitations. The ITMSC model consists of a target-based attention module to capture target-text relevance, an image-based attention module to capture image-text relevance, and a target-image matching module based on the former two modules to properly align the target with the image so that fine-grained semantic information can be extracted. The effectiveness and superiority of our model

are demonstrated by the experimental results and in-depth analysis of two datasets. Our results also demonstrate that the semantic description of the image can provide supplementary information for multimodal sentiment classification, leading to more accurate predictions.

Despite the promising performance, our proposed approach still has several limitations. First, our research has shown that images play an essential role in multimodal sentiment analysis, and the description of the images can provide valuable semantic information to support the analysis of the sentiment expressed. However, some image descriptions do not precisely correspond to the image's content, which can introduce semantic interference and detrimentally influence the accuracy of sentiment analysis. Second, the first limitation affects the results of calculating the semantic similarity between the text and the image description. As a result, the accuracy of sentiment analysis is reduced, thus negatively affecting the overall performance of sentiment analysis.

In future work, we plan to construct a model that employs the advantages of Vision-Language Pre-Trained Models to analyze sentiment accurately and provide more accurate results than existing models. Furthermore, we need to develop an algorithm that can consistently and accurately describe the content of an image. Such an algorithm should consider the visual elements of the image and any associated contextual information to generate a description that accurately reflects the image content. This will ultimately improve the model's performance, resulting in a more reliable and accurate output.

Funding Statement: The authors received no specific funding for this study.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] L. Zhang, S. Wang and B. Liu, "Deep learning for sentiment analysis: A survey," *WIREs Data Mining and Knowledge Discovery*, vol. 8, no. 4, pp. e1253, 2018. https://doi.org/10.1002/widm.1253

[2] M. Imran, F. Ofli, D. Caragea and A. Torralba, "Using AI and social media multimodal content for disaster response and management: Opportunities, challenges, and future directions," *Information Processing & Management*, vol. 57, no. 5, pp. 102261, 2020. https://doi.org/10.1016/j.ipm.2020.102261

[3] A. Gandhi, K. Adhvaryu, S. Poria, E. Cambria and A. Hussain, "Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions," *Information Fusion*, vol. 91, no. 3, pp. 424–444, 2023. https://doi.org/10.1016/j.inffus.2022.09.025

[4] B. Liu, "Sentiment analysis and opinion mining," *Synthesis Lectures on Human Language Technologies*, vol. 5, no. 1, pp. 1–167, 2012.

[5] N. Xu, W. J. Mao and G. D. Chen, "Multi-interactive memory network for aspect based multimodal sentiment analysis," in *Proc. of the AAAI Conf. on Artificial Intelligence*, vol. 33, no. 1, Honolulu, Hawaii, USA, pp. 371–378, 2019.

[6] J. F. Yu, J. Jiang and R. Xia, "Entity-sensitive attention and fusion network for entity-level multimodal sentiment classification," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 429–439, 2020. https://doi.org/10.1109/TASLP.2019.2957872

[7] J. F. Yu and J. Jiang, "Adapting BERT for target-oriented multimodal sentiment classification," in *Electronic Proc. of IJCAI 2019*, Macao, China, pp. 5408–5414, 2019.

[8] Z. Khan and Y. Fu, "Exploiting BERT for multimodal target sentiment classification through input space translation," in *Proc. of the 29th ACM Int. Conf. on Multimedia*, Virtual Event, China, pp. 3034–3042, 2021.

[9] B. Pang and L. Lee, "Opinion mining and sentiment analysis," *Foundations and Trends® in Information Retrieval*, vol. 2, no. 1–2, pp. 1–135, 2008.

[10] A. Agarwal, B. Xie, I. Vovsha, O. Rambow and R. J. Passonneau, "Sentiment analysis of twitter data," in *Proc. of the Workshop on Language in Social Media (LSM 2011)*, Portland, Oregon, pp. 30–38, 2011.
[11] X. Fang and J. Zhan, "Sentiment analysis using product review data," *Journal of Big Data*, vol. 2, no. 1, pp. 1–14, 2015.
[12] G. X. Xu, Y. T. Meng, X. Y. Qiu, Z. H. Yu and X. Wu, "Sentiment analysis of comment texts based on BiLSTM," *IEEE Access*, vol. 7, pp. 51522–51532, 2019.
[13] M. Singh, A. K. Jakhar and S. Pandey, "Sentiment analysis on the impact of coronavirus in social life using the BERT model," *Social Network Analysis and Mining*, vol. 11, no. 1, pp. 1–11, 2021.
[14] T. Zhu, L. D. Li, J. F. Yang, S. C. Zhao, H. T. Liu *et al.,* "Multimodal sentiment analysis with image-text interaction network," *IEEE Transactions on Multimedia*, pp. 1, 2022. https://doi.org/10.1109/TMM.2022.3160060
[15] M. F. Liu, L. M. Zhang, Y. Liu, H. J. Hu and W. Fang, "Recognizing semantic correlation in image-text weibo via feature space mapping," *Computer Vision and Image Understanding*, vol. 163, no. 5, pp. 58–66, 2017. https://doi.org/10.1016/j.cviu.2017.04.012
[16] N. Xu, W. J. Mao and G. D. Chen, "A co-memory network for multimodal sentiment analysis," in *The 41st Int. ACM SIGIR Conf. on Research & Development in Information Retrieval*, New York, NY, United States, pp. 929–932, 2018.
[17] Z. Y. Zhao, H. Y. Zhu, Z. H. Xue, Z. Liu, J. Tian *et al.,* "An image-text consistency driven multimodal sentiment analysis approach for social media," *Information Processing & Management*, vol. 56, no. 6, pp. 102097, 2019.
[18] N. Xu, Z. X. Zeng and W. J. Mao, "Reasoning with multimodal sarcastic tweets via modeling cross-modality contrast and semantic association," in *Proc. of the 58th Annual Meeting of the Association for Computational Linguistics*, Online, pp. 3777–3786, 2020.
[19] T. Chen, D. Y. Lu, M. Y. Kan and P. Cui, "Understanding and classifying image tweets," in *Proc. of the 21st ACM Int. Conf. on Multimedia*, Barcelona, Spain, pp. 781–784, 2013.
[20] T. Chen and H. SalahEldeen, "Velda: Relating an image tweet's text and images," in *Twenty-Ninth AAAI Conf. on Artificial Intelligence*, Austin, USA, vol. 29, pp. 1, 2015.
[21] A. Vempala and D. Preoţiuc-Pietro, "Categorizing and inferring the relationship between the text and image of Twitter posts," in *Proc. of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy, pp. 2830–2840, 2019.
[22] J. J. Ye, J. Zhou, J. F. Tian, R. Wang, J. Y. Zhou *et al.,* "Sentiment-aware multimodal pre-training for multimodal sentiment analysis," *Knowledge-Based Systems*, vol. 258, pp. 110021, 2022. https://doi.org/10.1016/j.knosys.2022.110021
[23] J. F. Yu, J. M. Wang, R. Xia and J. J. Li, "Targeted multimodal sentiment classification based on coarse-to-fine grained image-target matching," in *Proc. of the Thirty-First Int. Joint Conf. on Artificial Intelligence, IJCAI 2022*, Vienna, Austria, pp. 4482–4488, 2022.
[24] J. Devlin, M. W. Chang, K. Lee and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," arXiv preprint arXiv:1810.04805, 2018.
[25] N. Liu and J. H. Zhao, "A BERT-based aspect-level sentiment analysis algorithm for cross-domain text," *Computational Intelligence and Neuroscience*, vol. 2022, pp. 1–11, 2022. https://doi.org/10.1155/2022/8726621
[26] K. M. He, X. Y. Zhang, S. Q. Ren and J. Sun, "Deep residual learning for image recognition," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Las Vegas, USA, pp. 770–778, 2016.
[27] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh *et al.,* "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.
[28] Y. H. H. Tsai, S. J. Bai, P. P. Liang, J. Z. Kolter, L. P. Morency *et al.,* "Multimodal transformer for unaligned multimodal language sequences," in *Proc. of the Conf. Association for Computational Linguistics. Meeting*, Florence, Italy, pp. 6558–6569, 2019.

[29] G. Jawahar, B. Sagot and D. Seddah, "What does BERT learn about the structure of language?," in *ACL 2019-57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy, pp. 3651–3657, 2019.
[30] Y. Q. Wang, M. L. Huang, X. Y. Zhu and L. Zhao, "Attention-based LSTM for aspect-level sentiment classification," in *Proc. of the 2016 Conf. on Empirical Methods in Natural Language Processing*, Austin, USA, pp. 606–615, 2016.
[31] D. Tang, B. Qin and T. Liu, "Aspect level sentiment classification with deep memory network," arXiv preprint arXiv:1605.08900, 2016.
[32] P. Chen, Z. Q. Sun, L. D. Bing and W. Yang, "Recurrent attention network on memory for aspect sentiment analysis," in *Proc. of the 2017 Conf. on Empirical Methods in Natural Language Processing*, Copenhagen, Denmark, pp. 452–461, 2017.
[33] J. S. Lu, D. Batra, D. Parikh and S. Lee, "Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks," *Advances in Neural Information Processing Systems*, vol. 32, pp. 13–23, 2019.