F-046

Multi-modal aspect based sentiment analysis: Generating Explanation from Large Language Model.

Jun Cao Jiyi Li Ziwei Yang
University of Yamanashi, Kofu, Yamanashi, Japan
{G22tka12, jyli, G22tka18}@yamanashi.ac.jp

Abstract

Recent approaches to multimodal aspect-based sentiment analysis (MABSA) are based on pre-training small models to capture information from text-image pairs and fine-tuning them. However, small language models (SLMs) have limited capabilities and knowledge and often fail to accurately identify meaning, aspects, and sentiment in text and images. Large language models (LLMs) can mine fine-grained information in multimodal data. Based on these findings, we leverage LLMs to generate interpretations of text and images to enhance the ability of SLMs to identify aspects and sentiment. Extensive experiments demonstrate the superiority of our approach over existing methods.

1. Introduction

Multimodal Aspect-Based Sentiment Analysis (MABSA) has garnered increasing attention as a challenging fine-grained task in the field of sentiment analysis[1,2]. Multimodal Aspect-Based Sentiment Analysis (MABSA), which entails jointly extracting all aspect terms and predicting their sentiment polarity from the image-text pair[2].

Despite significant advancements, aspect extraction and sentiment polarity prediction in this fine-grained sentiment analysis scenario with multimodal information remain challenging for current models. Firstly, the semantic complexity of sentences poses difficulties in identifying aspects and comprehending their corresponding sentiment. Secondly, images often contain abundant detailed information, and models frequently struggle to focus on all key information in the image or mistakenly pay attention to irrelevant areas, introducing noise. Lastly, it is challenging to capture the overall relationship between images and text, particularly when judging the connection between more specific image regions and textual vocabulary.

Recent approaches often employ pretrained small language models (SLMs) to understand image-text pairs and provide representations[3,4]. These approaches also incorporate various modules to assist the small models in comprehending information from image-text pairs. Other enhancement methods include contrastive learning and image-text matching to align semantics and mitigate modality gaps[5,6]. While SLMs have demonstrated improvements, their knowledge and limitations also impact further enhancement of the models. It is necessary to incorporate external knowledge to strengthen the SLMs' understanding of image-text pairs. Some methods utilize image captions generated by CLIPCAP as image prompts to assist image comprehension[7]. However, this external information is often simplistic and crude, as image captions usually lack a relationship with sentiments and provide limited assistance in fine-grained aspect extraction and sentiment judgment.

Aspects	Kris	Soho	U-Know	Bangkok
Sentiment	Positive	Positive	Positive	Positive

Figure 1: An example of the LLM processing image-text pairs to generate the explanation of images and text.

Large Language Models (LLMs) possess powerful abilities to understand, track, and generate complex language[8,9,10]. They have made significant impressions in various tasks[11] and are considered promising solutions for general-purpose tasks[12]. LLMs are typically trained on large-scale text corpora, granting them richer knowledge and the ability to mine fine-grained information, subtle patterns, and relationships from data. So by employing SLMs' flexible task-specific learning and utilizing LLMs' generated information as informative rationales, we can combine the strengths of small and large models. LLM leverages its robust knowledge repository, superior information retrieval, as well as reasoning capabilities to generate insightful rationales. These rationales serve as external sources that the small model can refer to for making informed judgments. With these considerations, we propose a novel framework LRSA that combines the decision-making capabilities of SLMs with the additional information provided by LLMs for MABSA. Specifically, as shown in figure 1, we use LLMs to generate rationales for understanding image-text pairs and their connection as the injected information in SLMs as the connection of the SLMs and LLMs. The images and texts features refer to the rationale features in the fusion process to enhance them, making the fused feature representation more comprehensive and accurate, thus improving the understanding ability and predictive performance of the SLM. In the result generation stage, we connect the fused feature of the image-text pairs with the rationale feature to provide more information for the SLMs to make judgments.

Our contributions can be summarized as follows:

1) We propose a novel framework LRSA that combines the decision-making capabilities of SLMs with the additional

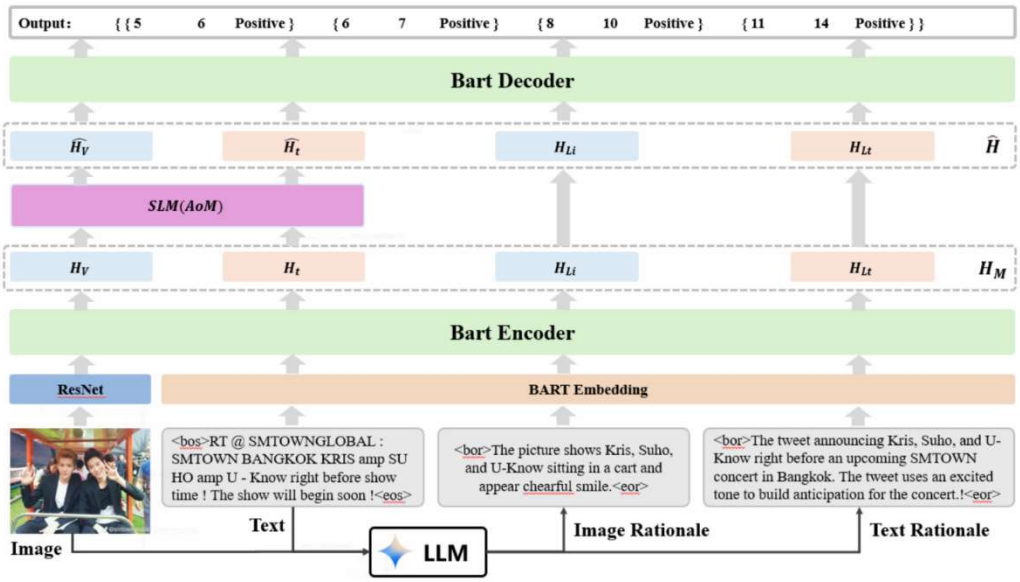

Figure 2: The overview of our proposed aspect-oriented model LRSA.

information provided by LLMs for MABSA, by using rationales generated by LLMs for SLMs to reference. We deeply integrate LLMs in the field of MABSA.

2) Experimental results on two benchmark datasets demonstrate that our method outperforms previous approaches.

2. Related Work

2.1 Aspect-Based Sentiment Analysis

Aspect-Based Sentiment Analysis (ABSA) focuses on predicting the sentiment polarity of aspect terms within sentences as a fine-grained task. Thanks to the excellent performance of language models such as BERT[13] and RoBERTa [14] in NLP, most recent works use pre-trained language models to model the semantic relationship between a given aspect and its context[15,16,17].

In addition, constructing aspect-oriented dependency trees can help establish the connection between aspect words and opinion words, and learn the syntactic feature representation of aspects to further improve the performance of language models[18,19,20].

More recently, recent studies have also adopted end-to-end models to extract all aspects and sentiment elements in triplets[21,22,23,24].

2.2 Multimodal Aspect-Based Sentiment Analysis

With the rapid proliferation of social media, Multimodal Aspect-Based Sentiment Analysis has garnered increasing research attention in recent years. Ju et al.[2] first implemented MABSA in a unified framework, processing two subtasks in an end-to-end manner to jointly extract aspect terms and their corresponding emotions. They also designed an auxiliary cross-modal relation detection to control the rational use of visual information. To capture cross-modal alignment, Ling et al.[3] built a generative multimodal architecture based on BART for vision-language pre-training and downstream MABSA tasks. Yang et al.[6] introduced a cross-modal multi-task transformer (CMMT) for MABSA, which dynamically controls the contribution of visual information to different aspects by taking into account the contribution of images when the confidence of the pure text prediction result is lower. Zhou et al.[4] proposed an aspect-oriented method to detect aspect-related semantic and sentiment information to mitigate visual and textual noise in complex image-text interactions. Yang et al.[7] explored the MABSA task in the case of few samples and proposed a novel generative multimodal cue model for MABSA, performing three MABSA-related tasks with a relatively small number of labeled multimodal samples. Yang et al.[5] introduced a face-sensitive image to sentiment text translation method that focuses on capturing visual sentiment cues through facial expressions and selectively matching and fusing them with target aspects in the textual modality. The above studies mainly focus on how to effectively obtain text and image representations and capture the interactive fusion between text and image representations. However, pre-training models alone are not enough. The existing traditional pre-training models have limited capabilities and knowledge and often cannot accurately identify the meaning, aspects and sentiment in text and images. The understanding of the connection between text and image is also lacking. It needs to be strengthened through auxiliary modules. To tackle the aforementioned issues, we propose a novel framework LRSA that combines the decision-making capabilities of SLMs with the additional information provided by LLMs for MABSA, use the image-text rationale generated by LLM to improve the understanding and prediction performance of SLM.

3. Proposed Approach

As shown in Figure 2, our RASA framework endeavors to incorporate the rationales produced by the LLMs, encompassing both text explanations and image understandings, into the SLMs which builds on an encoder-decoder architecture based on BART[25]. To accomplish this objective, we concatenate the image-text pairs and the image rationale and text rationales, which are then fed into the encoder. Ultimately, the fused image-text pairs after being processed by the SLMs are once again concatenated with the rationales of the image and text and fed into

	Methods	Twitter2015			Twitter2017		
		P	R	F1	P	R	F1
Text-based	SPAN (Hu et al. 2019)	53.7	53.9	53.8	59.6	61.7	60.6
	D-GCN (Chen, Tian, and Song 2020)	58.3	58.8	59.4	64.2	64.1	64.1
	BART (Yan et al. 2021)	62.9	65.0	63.9	65.2	65.6	65.4
Mutilmodal	RpBERT-collapse (Sun et al. 2021)	49.3	46.9	48.0	57.0	55.4	56.2
	UMT-collapse (Yu et al. 2020)	61.0	60.4	61.6	60.8	60.0	61.7
	JML (Ju et al. 2021)	65.0	63.2	64.1	66.5	65.5	66.0
	VLP-MABSA (Ling, Yu, and Xia 2022)	65.1	68.3	66.6	66.9	69.2	68.0
	AoM (Zhou et al. 2023)	65.3	67.3	66.3	66.5	67.3	66.9
	Ours (LRSA)	**66.9**	**68.3**	**67.6**	**67.2**	**69.3**	**68.2**

Table 1: Results of different methods for MABSA on the two Twitter datasets. The best results are bold-typed .

the decoder for result prediction. we will illustrate the details of our proposed model.

3.1 Overview

Task Definition. Formally, given a tweet that contains an image V and a sentence with n words $S = (w1, w2, \cdots, wn)$, our goal is to acquire the sequence Y representing all aspects and their associated sentiment polarities. We formulate the output of MABSA as $Y = [a_1^s, a_1^e, s_1, \cdots, a_i^s, a_i^e, s_i, \cdots, a_k^s, a_k^e, s_k]$, where a_i^s, a_i^s and s_i depict the start index, end index of the i -th aspect and its sentiment polarity in the tweet, and k is the number of aspects.

Data Processing. We input the given images and text into LLMs to obtain the description for the images $L_i$and the explanation for the text L_t . This is then combined with the image rationale and text rationale and inputted into the model as image-text pairs. This will be input into the SLMs as rationale of images and rationale of texts, with the image-text pairs.

3.2 Multimodal Encoder

In this section, we design the multimodal encoder to capture multimodal representations. We first conduct feature extraction on image-text pairs and image-text rationales. the embeddings of image E_V are obtained by preprocessing via ResNet. The embeddings of text E_T and the embeddings of rationale for images E_{Li} and texts E_{Lt} generated by the LLMs are obtained using the pre-trained BART. We add <img> and </img> before and after the image features, <bos> and <eos> for the textual features, <bol> and <eol> for the rationale features. Then, we concatenate the multimodal features as $X = [< img >, E_V, </img >, < bos >, E_T, < eos >, < bol >, E_{Li}, < eol >, < bol >, E_{Lt}, < eol >]$, which is the input of BART encoder.

In the end, we feed X into the BART encoder to obtain the multimodal representation. We believe that integrating the image-text pairs and image-text rationales into the same encoder can share parameters and leading to a more efficient and cohesive feature fusion.

$$H_M = MBART_E(X), H_M \in R^{l_m \times d}$$

where $l_m = l_i + l_t + l_{Li} + l_{Lt}$, l_i is the number of image slots that reserve initial image representation, l_t is the length of text , l_{Li} is the length of image rationale , l_{Lt} is the length of text rationale , and d is the hidden dimension.

3.3 Decoder And Prediction

After obtaining the hidden state H_M, we extract the image and text features from H_M and update them inside the SLMs to obtain the updated image features $\widehat{H_V}$ and text features $\widehat{H_t}$. In order to improve the quality of the decoder generation results and increase the accuracy of the prediction, we concatenate the final image features $\widehat{H_V}$ and text features $\widehat{H_t}$ with the image rationale features H_{Li} and text rationale features H_{Lt}, and the obtained results will be used together with the previous decoder output $Y_{<t}$ as the input of the BART decoder .

$$\widehat{H} = [\widehat{H_V}, \widehat{H_t}, H_{Li}, H_{Lt}], \widehat{H} \in R^{l_m \times d}$$
$$H^D = Decoder(\widehat{H}; Y_{<t})$$
$$\widetilde{H} = (W + \widehat{H_t})/2$$
$$P(y_t) = softmax([\widetilde{H}; S^d]H^D)$$

W denotes the embeddings of input tokens. S^d means the embeddings of the [positive, neutral, negative, <eos>]. The loss function is as follows:

$$\mathcal{L} = -E_{X \sim D} \sum_{t=1}^{o} logP(y_t | Y_{<t}, X).$$

4. Experiments

4.1 Experimental settings

Datasets. Our two benchmark datasets are Twitter2015 and Twitter2017[26]. Based on this foundation, we added the description of images as the rationale of image, and the understanding of texts as the rationales of text generated by LLMs to our dataset.

Implementation Details. Our LLM is Google's latest multimodal large language model "Gemini-1.5 pro"[27]. It can read images and text at the same time and give corresponding answers based on prompts. We input image-text pairs and use specific prompts to generate interpretations of images and understanding of text. We will detail the choice of our prompts in the case study.

Our SLM is based on BART[25], MABSA task is trained for 35 epochs with batch size 16, The learning rates are both 7e-5 and hidden sizes are 768.

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

4.2 Main Results

Results of MABSA Task: The results obtained from the MABSA are presented in Table 1. Firstly, our LRSA model significantly surpasses all text-based models, indicating the substantial contribution of incorporating image modality information and conducting text analysis of image modality in our model. LRSA consistently outperforms all other multimodal methods across all evaluation metrics. we conducted experiments with the second-best model AoM, using the parameters provided by the authors. The comparison revealed that LRSA exhibited both 1.9\% improvement in F1 metric compared to AoM, highlighting the effectiveness of augmenting the small model with additional information about image-text pairs generated by the large model in boosting the overall model performance.

5. Conclusion

In this Paper, we propose a novel framework LRSA that combines the decision-making capabilities of SLMs with the additional information provided by LLMs for MABSA, by using rationales generated by LLMs for SLMs to reference. Experimental results on two benchmark datasets demonstrate that our method outperforms previous approaches.

Acknowledgement

This work was partially supported by JSPS KAKENHI Grant Number JP23K28092.

Reference

[1] Yanxia Lv, Fangna Wei, Lihong Cao, Sancheng Peng, 611 Jianwei Niu, Shui Yu, and Cuirong Wang. Aspect-level sentiment analysis using context and aspect memory network. Neurocomputing, 428:195–205（2021）.

[2] Xincheng Ju, Dong Zhang, Rong Xiao, Junhui Li, Shoushan Li, Min Zhang, and Guodong Zhou. 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. (2021).

[3] Yan Ling, Jianfei Yu, and Rui Xia. Vision language pre-training for multimodal aspect based sentiment analysis. arXiv preprint arXiv:2204.07955 . (2022)

[4] Ru Zhou, Wenya Guo, Xumeng Liu, Shenglong Yu, Ying Zhang, and Xiaojie Yuan. Aom: Detecting aspect-oriented information for multimodal aspect-based sentiment analysis. arXiv preprint arXiv:2306.01004.(2023)

[5] Hao Yang, Yanyan Zhao, and Bing Qin. Face sensitive image-to-emotional-text cross-modal translation for multimodal aspect-based sentiment analysis. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 3324–3335 (2022)

[6] Li Yang, Jin-Cheon Na, and Jianfei Yu. Cross modal multitask transformer for end-to-end multimodal aspect-based sentiment analysis. Information Processing & Management, 59(5):103038. (2022)

[7] Xiaocui Yang, Shi Feng, Daling Wang, Sun Qi, Wenfang Wu, Yifei Zhang, Pengfei Hong, and Soujanya Poria. Few-shot joint multimodal aspect-sentiment analysis based on generative multimodal prompt. arXiv preprint arXiv:2305.10169. (2023)

[8] OpenAI. Chatgpt: Optimizing language models for dialogue. Accessed: 2024-06-13.(2022)

[9] Anthropic. Model card and evaluations for claude.(2023)

[10] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971.(2023)

[11] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. arXiv preprint arXiv:2206.07682.(2022)

[12] Yubo Ma, Yixin Cao, YongChing Hong, and Aixin Sun. Large language model is not a good few-shot information extractor, but a good reranker for hard samples! arXiv preprint arXiv:2303.08559.(2023)

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova.Bert: Pre-training of deep bidirectional transformers for language understand- ing. arXiv preprint arXiv:1810.04805.(2018)

[14] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Man- dar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining ap- proach. arXiv preprint arXiv:1907.11692.(2019)

[15] Alexander Rietzler, Sebastian Stabinger, Paul Opitz, and Stefan Engl. Adapt or get left behind: Do- main adaptation through bert language model finetun- ing for aspect-target sentiment classification. arXiv preprint arXiv:1908.11860.(2019)

[16] Fang Ma, Chen Zhang, Bo Zhang, and Dawei Song. Aspect-specific context modeling for aspect- based sentiment analysis. In CCF International Con- ference on Natural Language Processing and Chi- nese Computing, pages 513–526. Springer.(2022)

[17] Bin Liang, Hang Su, Lin Gui, Erik Cambria, and Ruifeng Xu. Aspect-based sentiment anal- ysis via affective knowledge enhanced graph con- volutional networks. Knowledge-Based Systems, 235:107643.(2022)

[18] Kai Wang, Weizhou Shen, Yunyi Yang, Xiaojun Quan, and Rui Wang. Relational graph attention network for aspect-based sentiment analysis. arXiv preprint arXiv:2004.12362. (2020)

[19] Yuanhe Tian, Guimin Chen, and Yan Song. Aspect-based sentiment analysis with type-aware graph convolutional networks and layer ensemble. In Proceedings of the 2021 conference of the North American chapter of the association for computa- tional linguistics: human language technologies, pages 2910–2922.(2021)

[20] Shuo Liang, Wei Wei, Xian-Ling Mao, Fei Wang, and Zhiyong He. Bisyn-gat+: Bi-syntax aware graph attention network for aspect-based sentiment analysis. arXiv preprint arXiv:2204.03117. (2022)

[21] Haiyun Peng, Lu Xu, Lidong Bing, Fei Huang, Wei Lu, and Luo Si. Knowing what, how and why: A near complete solution for aspect-based sentiment analysis. In Proceedings of the AAAI conference on artificial intelligence, volume 34, pages 8600–8607.(2020)

[22] Hongjie Cai, Rui Xia, and Jianfei Yu. Aspect- category-opinion-sentiment quadruple extraction with implicit aspects and opinions. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 340–350. (2021)

[23] Chen Zhang, Lei Ren, Fang Ma, Jingang Wang, Wei Wu, and Dawei Song. Structural bias for aspect sentiment triplet extraction. arXiv preprint arXiv:2209.00820.(2022)

[24] Hao Chen, Zepeng Zhai, Fangxiang Feng, Ruifan Li, and Xiaojie Wang. Enhanced multi-channel graph convolutional network for aspect sentiment triplet extraction. In Proceedings of the 60th Annual Meeting of the Association for Computational Lin- guistics (Volume 1: Long Papers), pages 2974–2985. (2022)

[25] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. Bart: De- noising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. arXiv preprint arXiv:1910.13461. (2019)

[26] Jianfei Yu and Jing Jiang. Adapting bert for target-oriented multimodal sentiment classification. IJCAI. (2019)

[27] Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Fi- rat, Julian Schrittwieser, et al. Gemini 1.5: Un-locking multimodal understanding across millions of tokens of context. arXiv preprint arXiv:2403.05530. (2024)