2024 IEEE International Conference on Multimedia and Expo (ICME) | 979-8-3503-9015-5/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICME57554.2024.10688334

Enhancing Multimodal Sentiment Analysis via Learning from Large Language Model

Ning Pang[†], Wansen Wu[†], Yue Hu, Kai Xu, Quanjun Yin, Long Qin[*]

College of Systems Engineering
National University of Defense Technology
Changsha, China
pangning14@nudt.edu.cn; wuwansen14@nudt.edu.cn; huyue11@nudt.edu.cn;
xukai09@nudt.edu.cn; yin_quanjun@163.com; qldbx2007@sina.com

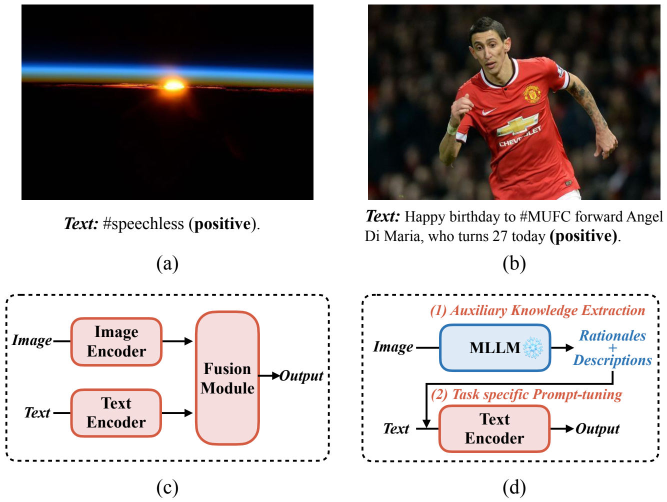
***Abstract*—Multimodal sentiment analysis (MSA) detects human sentiments by understanding data from multiple modalities, such as text and images. Existing research primarily strives for an effective multimodal fusion framework to derive informative representations. However, these methods neglect the necessity of exploiting external knowledge to aid in analyzing sentiments. As a result, the lack of external commonsense embarrasses these models when the opinion cues come in an implicit and obscure manner. To address the limitation, in this paper, we propose an Auxiliary Rationale Knowledge enhanced framework, namely ARK, which improves MSA models via learning from a multimodal large language model (MLLM). Specifically, based on text-image pairs, we employ Chain-of-Thought prompting to generate image descriptions and rationales from the MLLM as auxiliary knowledge, thus enriching the original samples with commonsense knowledge encoded within the MLLM. By combining the source text with image descriptions, we are able to effectively handle MSA through a Text+Text paradigm. In this paradigm, smaller pre-trained language models (LMs) can be tasked for sentiment classification via prompt-tuning. Besides, rationales are leveraged as additional supervision to facilitate the learning of reasoning abilities by LMs. Experimental results demonstrate that our proposed method outperforms current state-of-the-art approaches across four datasets. Our data and code are available at https://github.com/ningpang/ArkMSA.**

***Index Terms*—Multimodal sentiment analysis, large language model, prompt-tuning, rationale knowledge**

I. Introduction

Multimodal sentiment analysis (MSA) is a burgeoning field that combines multiple modalities, such as text, images, and audio, to understand and interpret human sentiments. MSA has garnered considerable interest due to its diverse range of applications, including the comprehension of user sentiment in product reviews and the examination of emotion-rich multimedia content within social networks [1].

In this paper, following previous works [3], [4], we mainly focus on the task of MSA where the data comprises two modalities, *i.e.,* text and images. To take advantage of features from different modalities, current state-of-the-art (SOTA) MSA methods [3], [4] model the interactions and fusion between text and images, the workflow of which is depicted in Figure 1(c). Despite the dominant performance, it is worth noting that previous methods still suffer from two inherent problems: 1) **Lack of external commonsense knowledge.** In several cases where the expression of sentiment is very implicit, a certain level of common knowledge is necessary for inference. Take Figure 1(a) as an example, both the image "sunrise with a black background" and the text "speechless" convey an ambiguous sentiment polarity [1]. A human can easily determine the positive states based on their commonsense knowledge since the focus of the image is "sunrise" rather than "black background" which leaves humans "speechless" due to the breathtaking beauty of nature. Nevertheless, traditional MSA methods are insufficient to detect such implicit expressions. 2) **Uneven distribution of sentiment information among modalities.** Usually, images contain less sentimental


Text: #speechless (**positive**).

(a)

Text: Happy birthday to #MUFC forward Angel Di Maria, who turns 27 today **(positive)**.

(b)

Image
Image Encoder
Text
Text Encoder
Fusion Module
Output
(c)
(1) Auxiliary Knowledge Extraction
Image
MLLM
Rationales + Descriptions
(2) Task specific Prompt-tuning
Text
Text Encoder
Output
(d)

Fig. 1. (a-b) Two examples from `TWITTER-15` [2] dataset. (c) Conventional MSA pipelines. (d) Our proposed method.

[†] Equal contribution, [*]Corresponding author.

This work is supported by National Natural Science Foundation of China under Grants 62103420, 62103425, 62103428, 62306329, Hunan Natural Science Foundation under Grants 2023JJ40676, 2021JJ40697, 2021JJ40702, Postgraduate Scientific Research Innovation Project of Hunan Province (CX20220011).

[1] For the image, the sunrise represents a beautiful natural scene (positive) while the black background may evoke a sense of oppression (negative). The text "speechless" actually means being unable to speak possibly due to surprise (positive) or anger (negative).

979-8-3503-9015-5/24/$31.00 ©2024 IEEE

information than text. As shown in another example in Figure 1(b), the image “a running footballer” has no sentiment polarity, while the text ”happy birthday” conveys a distinct positive sentiment. Besides, this observation is empirically supported by the comparison between text-only and image-only methods in Table II. However, existing multimodal fusion networks treat them equally, thereby encountering difficulties in accurately discerning their unequal contributions.

Recently, multimodal large language models (MLLMs) have demonstrated a strong understanding of textual and visual inputs, as well as possessing certain levels of common sense and reasoning abilities [5]. Inspired by the success of MLLMs, we propose a novel *Auxiliary Rationale Knowledge* enhanced framework (dubbed ARK) to address the above two problems, which improves the MSA models by learning auxiliary knowledge generated from an MLLM. The architecture of ARK is illustrated in Figure 1(d). Specifically, an MLLM (*i.e.*, the open-sourced MiniGPT4 [6]) is utilized to analyze original text-image pairs and generate auxiliary knowledge, including image descriptions and *rationales* [2], via Chain-of-Thought (CoT) prompting [7]. By converting raw images into descriptions in text format, data of image modality is enriched with commonsense knowledge in MLLM, and thus the uneven distribution of sentiment information is alleviated. By leveraging original text and image descriptions, we can handle MSA in a Text+Text paradigm, where smaller pretrained language models (LMs) are prompt-tuned for sentiment analysis. Besides, a masked language modeling task is incorporated into our MSA models to recover rationales, thus facilitating an understanding of the reasoning process. This simple yet effective framework eliminates the need for multimodal fusion networks. Additionally, the MLLM serves as a powerful external knowledge base to provide the model with common knowledge for enhanced reasoning capabilities.

Our contributions can be summarized as follows:

- We are among the first to introduce the MLLM as an external knowledge base to provide auxiliary information for solving the MSA tasks.
- We design a CoT strategy for auxiliary knowledge extraction from the MLLM and a prompt-tuning approach to enhance text-only MSA models with rationales.
- Our experiments on four MSA datasets demonstrate that our framework outperforms state-of-the-art methods and performs stably in few-shot scenarios.

II. Related Work

In this section, we review the related work from two aspects: multimodal sentiment analysis and large language models.

Multimodal sentiment analysis (MSA) can be divided into two types of research, target-oriented MSA [2], [4] and target-free MSA [8], where target-oriented MSA works necessitate a greater emphasis on capturing fine-grained information. Additionally, recent research has shifted focus towards target-oriented multimodal sentiment analysis [2], which necessitates a greater emphasis on capturing fine-grained information. Almost all prevalent approaches involve designing multimodal interaction and fusion strategies for combining information from different modalities, *e.g.,* VAuLT [3], EF-CaTrBERT [4]. Besides, recent work M^{2}CL [8] can learn enhanced global representations for different modalities with two-stream multimodal fusion components via contrastive learning.

Large language models (LLMs) [9], [10] exhibit exceptional abilities in natural language processing research [11], [12]. The introduction of GPT4 [5], an MLLM, has brought about a paradigm shift by enabling the processing of multimodal inputs. Nevertheless, this function is presently inaccessible to the public. To unravel the underlying mechanisms, MiniGPT4 [6] was open-sourced, which is built upon Vicuna [13] and BLIP-2 [14]. Owing to the high cost of LLMs, some endeavors utilize CoT prompting [7] to extract rationales from LLMs and leverage this rational knowledge to train smaller task-specific models [15].

Our work is fundamentally different from previous MSA works that design complex multimodal fusion networks. Instead, we leverage an MLLM to convert multimodal signals into the textual format to handle. Surprisingly, even the vanilla language models such as BERT [16] or RoBERTa [17] can achieve state-of-the-art performance on MSA tasks.

III. Methodology

We propose a novel *Auxiliary Rationale Knowledge* enhanced framework, namely ARK, to handle MSA under a Text+Text paradigm. Our proposed ARK leverages the reasoning ability of MLLM to derive *rationale* knowledge which is further employed to train smaller but more effective MSA models. As depicted in Figure 2, our paradigm consists of two crucial steps: (1) Auxiliary knowledge extraction from MLLM (cf. Section III-A) and (2) Task-specific LMs prompt-tuning with rationales (cf. Section III-B).

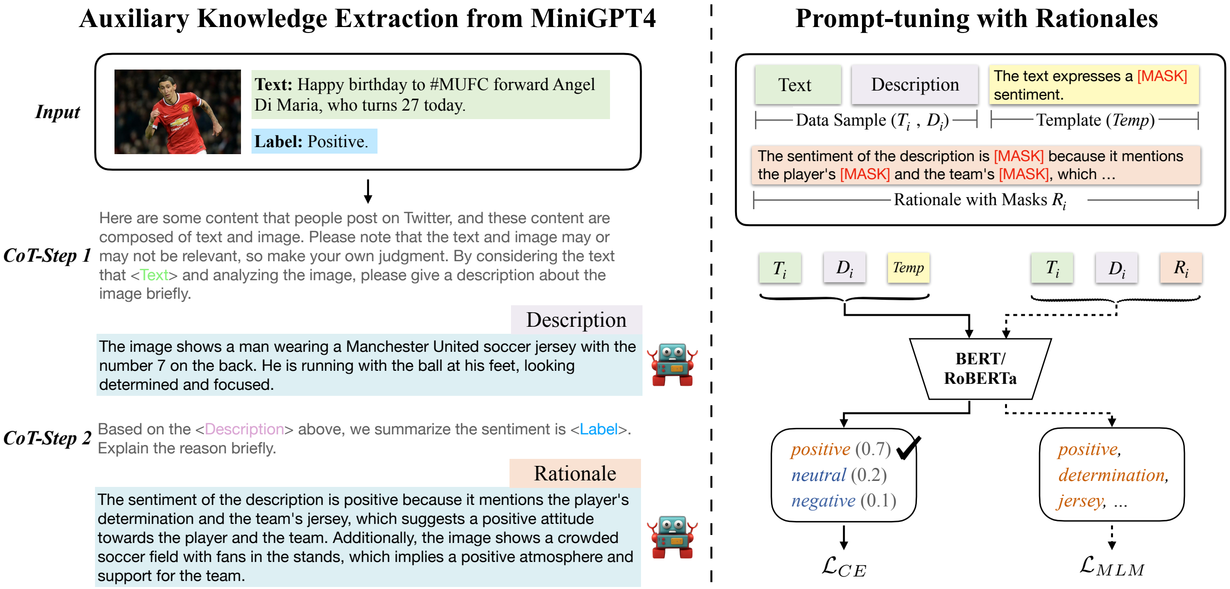
A. Auxiliary knowledge extraction from MLLM

Recent research endeavors have predominantly concentrated on eliciting the inherent cognition in LLMs to proficiently tackle a multitude of tasks requiring logical deduction [7], [18]. Inspired by the success of this paradigm, we exploit auxiliary knowledge extracted from an MLLM to enhance MSA.

Formally, we denote the original training dataset for MSA as $\mathcal{D} = \{(T_i, I_i, y_i)\}_{i=1}^{N}$, where each instance is composed of a text-image pair (T_i, I_i) and an assigned sentiment label y_i. Previous MSA approaches [4], [8] resort to learning a mapping function $F : (T_i, I_i) \subset y_i$ with the input consisting of a text-image pair. In contrast, this work prompts an MLLM to derive image descriptions and rationales as auxiliary knowledge to train smaller MSA models.

Specifically, for each sample $(T_i, I_i, y_i) \in \mathcal{D}$, we obtain image description D_i and rationale knowledge R_i from MiniGPT4 via exploiting *Chain-of-Thought* (CoT) prompting. As illustrated in the left part of Figure 2, we construct the two-hop prompts as follows.

[2] Rationales are textual explanations that offer support for the assigned label in training data (see Figure 2).



Text: Happy birthday to #MUFC forward Angel Di Maria, who turns 27 today.

Label: Positive.

Here are some content that people post on Twitter, and these content are composed of text and image. Please note that the text and image may or may not be relevant, so make your own judgment. By considering the text that <Text> and analyzing the image, please give a description about the image briefly.

The image shows a man wearing a Manchester United soccer jersey with the number 7 on the back. He is running with the ball at his feet, looking determined and focused.

Based on the <Description> above, we summarize the sentiment is <Label>. Explain the reason briefly.

The sentiment of the description is positive because it mentions the player's determination and the team's jersey, which suggests a positive attitude towards the player and the team. Additionally, the image shows a crowded soccer field with fans in the stands, which implies a positive atmosphere and support for the team.

The text expresses a [MASK] sentiment.

The sentiment of the description is [MASK] because it mentions the player's [MASK] and the team's [MASK], which ...

Fig. 2. An illustration of our ARK framework for auxiliary knowledge extraction from MiniGPT4 (left) and prompt-tuning with rationales for task-specific language models (right).

Step 1. We first instruct MiniGPT4 to generate the image description D_i by analyzing counterpart tweet text T_i and image I_i on its own judgment.

Step 2. Based on image I_i, text T_i and image description D_i, we ask MiniGPT4 to answer the underlying rationale R_i why the data sample is categorized as the assigned sentiment y_i.

After extracting auxiliary knowledge from MLLM, we augment the original training dataset with image descriptions and rationales into $\mathcal{D} = \{T_i, D_i, R_i, y_i\}_{i=1}^{N}$. Based on $\mathcal{D}$, we train a smaller but more effective task-specific LM for MSA via prompt-tuning with rationales.

B. Task-specific LM prompt-tuning with rationales

Prompt-tuning originates from the *masked language modeling* (MLM) task of masked LMs, such as BERT [16] and RoBERTa [17]. This pre-training task randomly replaces some percentage of the input tokens with a special `[MASK]` token, and then predicts those masked tokens. Next, we first present our prompt-tuning framework and subsequently expand upon it to integrate rationales during training.

Standard prompt-tuning for MSA. In prompt-tuning for MSA, we augment each data sample with a prompt template that holds a `[MASK]` position, and an LM is required to predict the `[MASK]` token. For instance, given a binary sentiment classification task, the LM decides whether it is more appropriate to fill in "*positive*" or "*negative*". In this paper, we manually define the prompt template as $\texttt{P}(T_i, D_i) =$ `[CLS]` T_i `[SEP]` D_i `[SEP]` The texts express a `[MASK]` sentiment for the target `[Target]`. [3].

To detect target sentiments, we introduce a verbalizer that establishes a map between sentiment labels and specific tokens in the vocabulary of an LM. The mapping function is denoted as $\mathcal{M}(\mathcal{Y}) \subset \mathcal{V}$ where $\mathcal{V}$ is the vocabulary of the LM. Instead of adhering to a mapping between one sentient label and one existing token in the vocabulary, we expand the LM with a set of learnable sentiment tokens that encapsulate the implicit semantics of sentiment labels [4]. In this way, we can treat our task as an MLM task, and model the probability of predicting $y \in \mathcal{Y}$ as:

$$\begin{aligned} p(y_i \mid (T_i, D_i)) &= p(\texttt{[MASK]} = \mathcal{M}(y_i) \mid \texttt{P}(T_i, D_i)) \\ &= \frac{\exp(\mathbf{w}_{\mathcal{M}(y_i)} \cdot \mathbf{h}_{\texttt{[MASK]}})}{\sum_{y_j \mathcal{Y}} \exp(\mathbf{w}_{\mathcal{M}(y_i)} \cdot \mathbf{h}_{\texttt{[MASK]}})}, \end{aligned} \tag{1}$$

where $\mathbf{w}_{\mathcal{M}(y_i)}$ denotes the embedding of golden sentiment token, and $\mathbf{h}_{\texttt{[MASK]}}$ is the hidden vector of `[MASK]` position encoded by an LM. Given the training set $\mathcal{D}$, the cross-entropy loss is leveraged to minimize the prediction loss between the predicted and target tokens,

$$\mathcal{L}_{CE} = -\frac{1}{N} \sum_{i=1}^{N} \log p(y_i \mid (T_i, D_i)). \tag{2}$$

Multi-task learning with rationales. Apart from standard prompt-tuning, we utilize rationales as additional supervision to allow the LM for MSA to understand the reason why a data sample belongs to a specific sentiment. By enhancing the comprehension capability of LMs, the performance of sentiment identification can be improved. We adopt the MLM task to assist the task-specific LM in building the connection between data samples and rationales. Specifically, we

[3] In target-oriented MSA, we replace the `[Target]` token with the actual objective mentioned in the counterpart text. However, the `[Target]` token is directly eliminated in target-free scenarios.

[4] Take a binary sentiment classification task for example, we introduce two sentiment tokens to represent sentiment labels "*positive*" and "*negative*". The embedding of each sentiment token is initialized with the average embedding of synonymous words for the corresponding sentiment label. For instance, the sentiment token for the label "*positive*" is initialized with the average embedding of words {optimistic, encouraging, promising, good}.

augment the prompt-tuning input with additional rationales, *i.e.*, $\bar{\mathrm{P}}_i = \mathrm{P}(T_i, D_i)\texttt{[SEP]}R_i\texttt{[SEP]}$, and mask a ratio of tokens in the sequence at random. The objective is to predict the masked words to recover the original input. Suppose M masked tokens in the input sample, we compute the cross-entropy loss to recover the original sample as:

$$\mathcal{L}_{MLM} = -\frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M}\log p(v \mid \texttt{[MASK]}_j \in \bar{\mathrm{P}}_i), \qquad (3)$$

where the probability of j-th $\texttt{[MASK]}$ in $\bar{\mathrm{P}}$ filled with $v \in \mathcal{V}$ is derived similar to Equation (1).

Finally, we harness the joint loss to optimize our task-specific LM model (*e.g.,* BERT [16] or RoBERTa [17]),

$$\mathcal{L} = \mathcal{L}_{CE} + \lambda\mathcal{L}_{MLM}, \qquad (4)$$

where λ is a hyper-parameter to weight the importance of $\mathcal{L}_{MLM}$ and set empirically in the implementation.

Sentiment inference. During inference, for a test sample $(T\ , I\ )$, we merely need to use an image caption generation model to obtain an image description $D\ $ without the necessity to generate the corresponding rationale. The sentiment prediction can be determined by Equation (1).

IV. Experiments

A. Experimental settings

Datasets. To comprehensively evaluate our model, we conduct experiments in both target-oriented and target-free scenarios. For target-oriented MSA, we conduct experiments on `TWITTER-15` and `TWITTER-17` [2]. Besides, `MVSA-Single` and `MVSA-Multiple` are employed for evaluation in target-free MSA scenarios and we use the same method as [8] to preprocess data. The statistics information of the four datasets is shown in Table I.

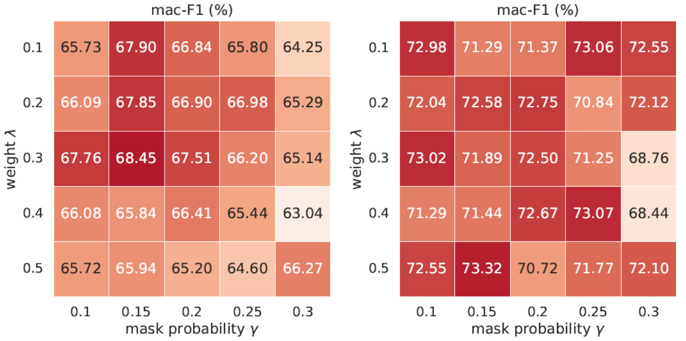
Dataset	Train	Dev	Test	Total
`TWITTER-15`	3,176	1,122	1,037	5,335
`TWITTER-17`	3,559	1,176	1,234	5,969
`MVSA-Single`	3,608	451	451	4,510
`MVSA-Multiple`	13,620	1,702	1,702	17,024

TABLE I
Statistics on the four benchmark datasets.

Implementation details. In implementation, we employ `bert-base-uncased` [16] and `roberta-large` [17] pre-trained models as our backbone task-specific LMs to handle MSA. We adopt AdamW [23] as the optimizer with a 1e-5 learning rate. Three typical criteria, *i.e.*, accuracy (Acc), macro F1 (mac-F1), and weighted F1 (w-F1) are utilized for evaluation. To facilitate the reproduction of our experimental results, the source code of our proposed framework and detailed hyper-parameters are provided on GitHub.

Baselines. We compare our method with baselines in three categories: 1) For **text-only** methods, BERT and RoBERTa serve as strong baselines to handle sentiment classification tasks. 2) For **image-only** methods, we use the VGG16 [19], ResNet34 [20] and ViT-B/16 [21] as strong baselines. 3) For **multimodal** methods, ViLT [3], VAuLT [3], EF-CaTrBERT [4], ITIN [24] and M^{2}CL [8] design different multimodal fusion strategies to learn global representations from two modalities for sentiment detection.

B. Hyper-parameter study



(a) BERT (b) RoBERTa

Fig. 3. Hyper-parameter analysis on `TWITTER-17`.

In the implementation, there are two hyper-parameters, i.e., the weight λ for balancing two losses and the mask probability γ for input tokens, that have a significant impact on the model performance. Therefore, in this set of experiments, we analyze the sensitivity of our model to these two hyper-parameters. We selected $\lambda \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$ and $\gamma \in \{0.1, 0.15, 0.2, 0.25\}$. We conducted our analysis on the development set of `TWITTER-17` using both the BERT and RoBERTa as backbone LMs. The results are shown in Figure 3, where deeper color indicates better performance. It can be observed from the results that our framework is robust across different settings. We selected $(\lambda = 0.3, \gamma = 0.15)$ (resp. $(\lambda = 0.5, \gamma = 0.15)$) for the following evaluation when employing BERT (resp. RoBERTa) as the backbone LM for MSA due to the optimal performance.

C. Comparison with the state-of-the-art

The performance comparison of our method with baselines and ablation experiments are presented in Table II. It reads from the results that: 1) Our proposed ARK, employing the BERT LM, almost outperforms all baseline methods across four different datasets. When Roberta LM is harnessed, the performance can be further improved. 2) The removal of rationales gives rise to performance degradation for both ARK $_{\text{BERT}}$ and ARK $_{\text{RoBERTa}}$, indicting that auxiliary rationales knowledge can enhance the performance of MSA. 3) In the comparison, text-only baselines evidently outperform image-only methods. This finding supports our argument that image data contains less sentiment information than text data.

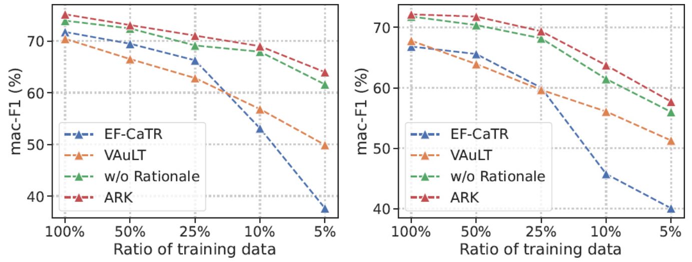
D. Analysis of few-shot scenario

In this set of experiments, we compare ARK with ARK w/o rationale, EF-CaTR, and VAuLT on `TWITTER-15` and `TWITTER-17` to observe the performance variations across different sizes of training data. Specifically, we randomly sample 50%, 25%, 10%, and 5% of training data to train the

Model	TWITTER-15			TWITTER-17			MVSA-Single			MVSA-Multiple		
	Acc	mac-F1	w-F1	Acc	mac-F1	w-F1	Acc	mac-F1	w-F1	Acc	mac-F1	w-F1
Text-only												
BERT‡ [16]	71.55	68.85	71.67	68.15	66.22	68.10	74.28	62.93	73.51	68.74	53.28	66.08
RoBERTa‡ [17]	75.02	71.53	74.93	70.66	70.26	70.66	75.17	62.79	73.74	66.51	55.22	66.46
Image-only												
VGG16‡ [19]	58.53	24.61	43.22	46.43	21.14	29.45	59.65	47.89	58.44	65.75	32.61	57.03
ResNet34‡ [20]	52.84	38.81	50.70	46.60	38.90	45.18	64.97	52.15	63.71	66.33	37.95	59.31
ViT-B/16‡ [21]	63.31	41.64	51.28	50.15	41.92	49.47	69.62	52.19	66.50	66.63	39.04	60.05
Text + Image												
ViLT‡ [3]	69.53	61.07	68.96	64.34	59.09	63.81	74.50	66.24	74.25	67.33	50.72	65.28
VAuLT‡ [3]	75.80	70.43	75.56	69.45	67.80	69.41	76.50	64.19	75.43	69.80	57.49	68.75
EF-CaTrBERT‡ [4]	76.62	71.75	76.20	68.78	66.84	67.86	75.39	66.21	75.43	65.22	52.10	64.04
ITIN [22]	-	-	-	-	-	-	75.19	-	74.97	73.52	-	73.49
M^{2}CL [8]	-	-	-	-	-	-	75.5	-	74.2	73.2	-	70.5
ARK $_{\text{BERT}}$	78.11	73.82	78.23	70.02	68.90	70.39	77.61	68.89	77.32	73.97	60.13	71.65
w/o Rationale	77.14	72.38	77.03	69.21	67.53	69.02	76.05	67.33	75.82	72.76	60.05	70.84
ARK $_{\text{RoBERTa}}^{\triangle}$	79.85	75.16	78.95	73.74	72.10	73.54	80.71	70.75	79.95	75.02	62.91	73.93
w/o Rationale	78.78	73.94	78.42	72.93	70.52	72.68	78.49	69.69	77.89	74.32	62.85	73.64

TABLE II

The experiment results (%) on four multimodal MSA datasets in terms of accuracy, macro F1 and weighted F1 metrics. ‡ denotes the re-implementation results by ourselves. $^\triangle$ indicates a statistically significant improvement over baselines for $p < 0.05$. We highlight the best and second best performance in the red and blue colors.



(a) TWITTER-15 (b) TWITTER-17

Fig. 4. The performance changes of MSA methods across different training data sizes.

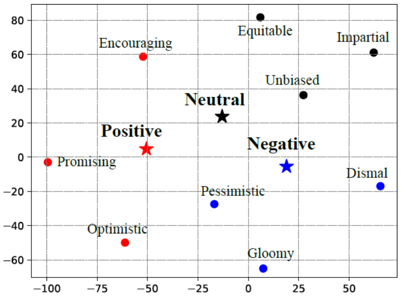


Fig. 5. Visualization of sentiment tokens and their synonymous words in a 2-D embedding space.

MSA models and keep the test set unchanged. Pre-trained RoBERTa LM is employed as our backbone MSA model for evaluation. The macro F1 values of MSA methods are illustrated in Figure 4. From the results, we can observe that: 1) Under different settings, our ARK demonstrates significant superiority over the two comparison methods, and the inclusion of rationales consistently enhances the performance of MSA. 2) Owning to the prompt-tuning strategy in our model, ARK exhibits better learning capabilities in low-data scenarios. Consequently, with the gradual decrease in training data sizes, the advantages of our model become more pronounced.

E. Visualization of verbalizer

In this set of experiments, we conducted analysis to investigate whether these expanded sentiment tokens in Section III-B can reflect the corresponding labels. To be specific, we derive the embeddings of sentiment tokens from the embedding layer of a task-specific RoBERTa LM for MSA. We then search the top-3 synonyms in the vocabulary that are closest to each sentiment token using the L2 distance between their embeddings. As shown in Figure 5, by visualizing these embeddings in a 2-D space, we observe that each expanded sentiment token comprehensively expresses the sentiment polarity of its corresponding label.

V. Conclusion

In this paper, we propose a novel method dubbed ARK to handle the task of MSA in the Text+Text paradigm. Specifically, we first extracted image descriptions and rationales from the multimodal large language model MiniGPT4. Afterward, we harnessed the auxiliary knowledge to train smaller pre-trained language models for MSA. The results of extensive experiments demonstrated the effectiveness of our framework which enhances MSA via learning from a large language model. In future work, it is of interest to explore the integration of visual embeddings directly encoded from images into our method.

References

[1] Ramandeep Kaur and Sandeep Kautish, “Multimodal sentiment analysis: A survey and comparison,” *Int. J. Serv. Sci. Manag. Eng. Technol.*, vol. 10, no. 2, pp. 38–58, 2019.

[2] Jianfei Yu and Jing Jiang, “Adapting BERT for target-oriented multimodal sentiment classification,” in *IJCAI*. 2019, pp. 5408–5414, ijcai.org.

[3] Georgios Chochlakis, Tejas Srinivasan, Jesse Thomason, and Shrikanth Narayanan, "Vault: Augmenting the vision-and-language transformer with the propagation of deep language representations," *arXiv preprint arXiv:2208.09021*, 2022.

[4] Zaid Khan and Yun Fu, "Exploiting bert for multimodal target sentiment classification through input space translation," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 3034–3042.

[5] OpenAI, "GPT-4 technical report," *CoRR*, vol. abs/2303.08774, 2023.

[6] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny, "Minigpt-4: Enhancing vision-language understanding with advanced large language models," *CoRR*, vol. abs/2304.10592, 2023.

[7] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou, "Chain-of-thought prompting elicits reasoning in large language models," in *NeurIPS*, 2022.

[8] Yang Xiaocui, Feng Shi, Wang Daling, Hong Pengfei, and Poria Soujanya, "Multiple contrastive learning for multimodal sentiment analysis," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2023, pp. 1–5.

[9] OpenAI, "Introducing chatgpt," *https://openai.com/blo g/chatgpt*, 2022.

[10] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample, "Llama: Open and efficient foundation language models," *CoRR*, vol. abs/2302.13971, 2023.

[11] Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo, "Are emergent abilities of large language models a mirage?," *CoRR*, vol. abs/2304.15004, 2023.

[12] Chunkit Chan, Jiayang Cheng, Weiqi Wang, Yuxin Jiang, Tianqing Fang, Xin Liu, and Yangqiu Song, "Chatgpt evaluation on sentence level relations: A focus on temporal, causal, and discourse relations," *CoRR*, vol. abs/2304.14827, 2023.

[13] UC Berkeley, Stanford Cmu, and UC San, "Vicuna: An open-source chatbot impressing gpt-4 with 90.

[14] Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi, "BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models," in *ICML*. 2023, vol. 202, pp. 19730–19742, PMLR.

[15] Jinyuan Li, Han Li, Zhuo Pan, and Gang Pan, "Prompt chatgpt in MNER: improved multimodal named entity recognition method based on auxiliary refining knowledge from chatgpt," *CoRR*, vol. abs/2305.12212, 2023.

[16] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," in *NAACL-HLT (1)*. 2019, pp. 4171–4186, Association for Computational Linguistics.

[17] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov, "Roberta: A robustly optimized BERT pretraining approach," *CoRR*, vol. abs/1907.11692, 2019.

[18] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa, "Large language models are zero-shot reasoners," in *NeurIPS*, 2022.

[19] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," in *ICLR*, 2015.

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[21] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *ICLR*. 2021, OpenReview.net.

[22] Tong Zhu, Leida Li, Jufeng Yang, Sicheng Zhao, Hantao Liu, and Jiansheng Qian, "Multimodal sentiment analysis with image-text interaction network," *IEEE Transactions on Multimedia*, 2022.

[23] Diederik P. Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," in *ICLR (Poster)*, 2015.

[24] Tong Zhu, Leida Li, Jufeng Yang, Sicheng Zhao, Hantao Liu, and Jiansheng Qian, "Multimodal sentiment analysis with image-text interaction network," *IEEE Trans. Multim.*, vol. 25, pp. 3375–3385, 2023.

Appendix

The section provides more details about the training procedure and hyper-parameter settings. We utilized Pytorch to implement our framework and conducted the experiments with 1 Nvidia 3090 GPU. All optimizations are performed with an AdamW optimizer.

Hyper-parameter settings.

The settings of hyper-parameters are shown as follows:

- number of epochs: 10
- batch size: 8 (for BERT as 16)
- learning rate for overall parameters: 1e-5
- max sequence length: 400
- gradient accumulation steps: 1
- weight λ: 0.5 (for BERT as 0.3)
- mask probability: 0.15

Prompt template design.

For `MVSA-Single` and `MVSA-Multiple` datasets, we use the following two-step prompts:

> *CoT-Step1: Here are some content that people post on Twitter, and these content are composed of text and image. Please note that the text and image may or may not be relevant, so make your own judgment. By considering the text that <text> and analyzing the image, please give a description about the image briefly.*
>
> *CoT-Step2: Based on the <description> above, we summarize the sentiment is <label>. Explain the reason briefly.*

For `TWITTER-15` and `TWITTER-17` datasets, we use the following two-step prompts:

> *CoT-Step1: Here are some content that people post on Twitter, and these content are composed of text and image. Please note that the text and image may or may not be relevant, so make your own judgment. By considering the text that <text> and analyzing the image, please give a description about the image within <20> words.*
>
> *CoT-Step2: Based on such description, what is the sentiment polarity towards <target>? Summarize the sentiment polarity, and return only one of these words: [<Negative>, <Neutral>, <Positive>]. Make the answer format like: [Sentiment: Reason]*

<text> represents the original text in the four MSA datasets. *<description>* is filled with the image description generated after the first CoT step. *<target>* is the target mention in a piece of text in `TWITTER-15` and `TWITTER-17` datasets.