A Knowledge-enhanced Framework for Target-Oriented Multimodal Sentiment Classification

Fei Zhao

Natural Language Processing Laboratory
Nanjing University

September 17, 2022

Target-oriented Multimodal Sentiment Classification

- Target-oriented multimodal sentiment classification(TMSC): determine the sentiment polarity of the opinion target mentioned in a (sentence, image) pair.

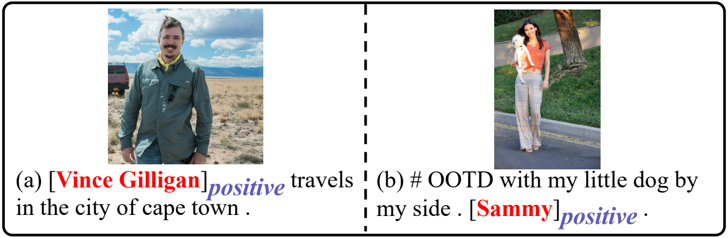


Figure 1: Two examples of TMSC task. Opinion targets and their corresponding sentiment polarities are highlighted in the sentence.

Motivation

- These methods easily fail to align two modalities because of the granularity gap of opinion target across text and image.

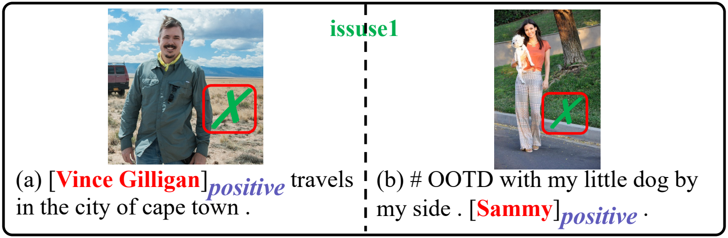


Figure 2: The first issue of TMSC task. The red bounding box denotes the visual clues that the opinion target focuses on.

Motivation

- Even though it is captured, diversified visual representations expressing the same mood also bring challenges for sentiment prediction.

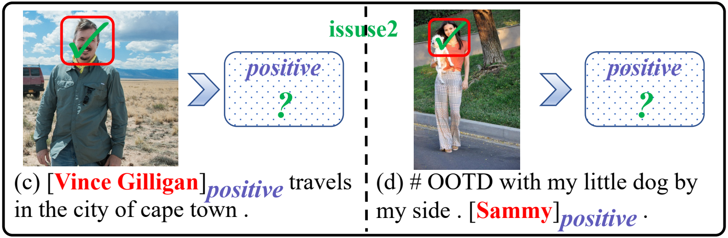


Figure 3: The second issue of TMSC task. The red bounding box denotes the visual clues that the opinion target focuses on.

Intuition

- For the first issue, we observed that the nouns of ANPs are also coarse-grained concepts, so an intuitive idea is to map a fine-grained opinion target (e.g. "*Vince Gilligan*") to a coarse-grained noun (e.g. "*man*") in ANPs.

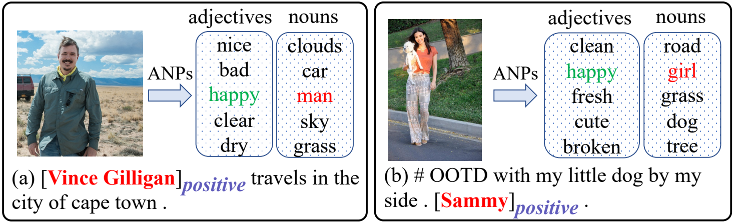
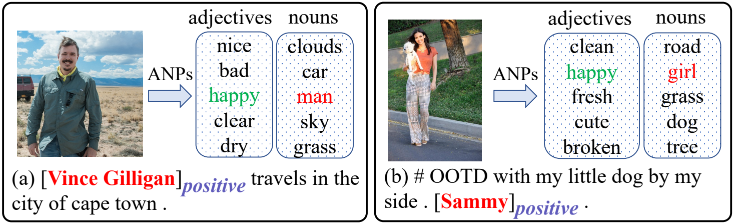


Figure 4: Extract Top-5 adjective-noun pairs (ANPs) from each image in our Twitter datasets.

Intuition

- For the second issue, we observed that ANPs can usually extract the same adjectives from different visual content expressing the same mood, so an intuitive idea is to map diversified visual representations (e.g., smiling faces) to the same adjective (e.g., "*happy*").



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

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

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

1. Introduction
2. Motivation & Intuition
3. The proposed model
4. Experiment
5. Conclusions

Knowledge-enhanced Framework

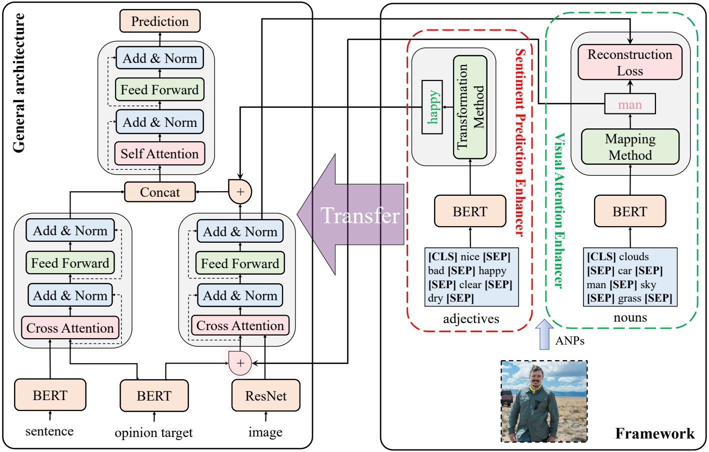


Figure 6: The overview of our KEF framework.

General Attention Architecture

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

$$H_{T \to V} = \text{Cross-ATT}(H_T, H_V), \quad (1)$$
$$H_{T \to C} = \text{Cross-ATT}(H_T, H_C), \quad (2)$$

- We feed the first token H^0 of the multimodal representation to a softmax layer for the sentiment classification:

$$p(y|H^0) = \mathrm{softmax}(W_M^\top H^0), \quad (3)$$

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

$$\mathcal{L}_t = -\frac{1}{|\mathcal{D}|} \sum_{i=1}^{|\mathcal{D}|} \log p(y^i | H^0). \quad (4)$$

Visual Attention Enhancer

Challenge: most of the nouns extracted from the image are target-independent, so we cannot use them directly.

Mapping Method

- We first measure the strength of target-noun relevance by calculating the semantic similarity between noun representation and target representations:

$$\alpha^i = \cos(H_T, H_N^i), \quad (5)$$

where $\cos(\cdot)$ is a cosine function and α^i means the similarity score.

Visual Attention Enhancer

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

$$\alpha^m = \max_{i=1}^{l}(\alpha^i), \quad (6)$$

where H_N^m denotes the noun representation corresponding to the highest similarity score α^m.

- Next, we aggregate them together as complementary information for the opinion target to capture the corresponding visual representations $H_{T \to V}$. Formally, we update H_T in Eq. 1 by:

$$\widetilde{H}_N = \alpha^m H_N^m, \quad (7)$$

$$H_T = H_T + \lambda_N \widetilde{H}_N, \quad (8)$$

Visual Attention Enhancer

Reconstruction Loss

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

$$\mathcal{L}_a = -\frac{1}{|\mathcal{D}|}\sum_{i=1}^{|\mathcal{D}|}(\widetilde{H}_N - H_{T\rightarrow V})^2, \quad (9)$$

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

Sentiment Prediction Enhancer

Challenge: the adjective most relevant to visual representations is unknown, we need to find it explicitly.

Transformation Method

- Since an adjective is a modifier of a noun, the adjective corresponding to this noun is also most relevant to target-aware visual representations.
- We use it as the complementary information of visual representations to reduce the difficulty of sentiment prediction:

$$H_{T\to V} = H_{T\to V} + \lambda_A H_A^m. \tag{10}$$

where H_A^m denotes the adjective representation corresponding to the noun representation H_N^m.

Datasets

- We carry out experiments on two public multimodal datasets TWITTER-15 and TWITTER-17. General information for both datasets is presented in Table 1.

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

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

Compared Methods

We choose three kinds of baselines.

- **The first** is a frequently-used visual-based model *ResNet-Target*.
- **The second** is some classical text-based models, including *AE-LSTM* [WHZ+16], *MemNet* [TQL16], *RAM* [CSBY17], *MGAN* [FFZ18], *BERT* [DCLT19].
- **The third** is the recent multi-modal models, including *Res-MGAN*, *MIMN* [XMC19], *ESAFN* [YJX19], *MMAP* [ZZH+21], *mPBERT* [YJ19], *ModalNet-BERT* [ZWL+21], EF-CapTrBERT [KF21], *TomBERT* [YJ19] and *Saliencybert* [WLS+21].

Main Results

Model	TWITTER-15 Acc	TWITTER-15 Macro-F1	TWITTER-17 Acc	TWITTER-17 Macro-F1
Visual				
Res-Target	59.88	46.48	58.59	53.98
Text				
AE-LSTM	70.30	63.43	61.67	57.97
MemNet	70.11	61.76	64.18	60.90
RAM	70.68	63.05	64.42	61.01
MGAN	71.17	64.21	64.75	61.46
BERT	74.15	68.86	68.15	65.23
Text + Visual				
Res-MGAN	71.65	63.88	66.37	63.04
MIMN	71.84	65.69	65.88	62.99
ESAFN	73.38	67.37	67.83	64.22
MMAP♣	73.50	66.53	67.31	64.34
mPBERT	75.79	71.07	69.61	67.12
ModalNet-Bert♣	76.71	70.93	69.55	67.28
EF-CapTrBERT★	77.01	71.79	69.00	66.71
Our Framework				
SaliencyBERT	77.03	72.36	69.69	67.19
KEF-SaliencyBERT	**78.15**$^{\dagger}\pm0.33$	**73.54**$^{\dagger}\pm0.55$	**71.88**$^{\dagger}\pm0.21$	**68.96**$^{\dagger}\pm0.14$
Δ	+1.12	+1.18	+2.19	+1.77
TomBERT	77.15	71.75	70.50	68.04
KEF-TomBERT	**78.68**$^{\dagger}\pm0.30$	**73.75**$^{\dagger}\pm0.27$	**72.12**$^{\dagger}\pm0.15$	**69.96**$^{\dagger}\pm0.25$
Δ	+1.53	+2.00	+1.62	+1.92

Table 2: Test accuracy on the TWITTER-15 and TWITTER-17 datasets

Ablation Study

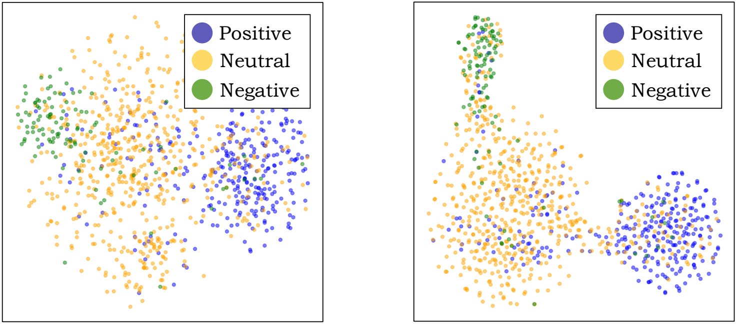
Effects of Knowledge-enhanced Framework

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

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

Ablation Study

Analysis over components of Visual Attention Enhancer



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

Figure 7: Visualization of multimodal output representations for *TomBERT+VAE* and *KEF-TomBERT*.

Parameter Analysis

Effect of the number of ANPs

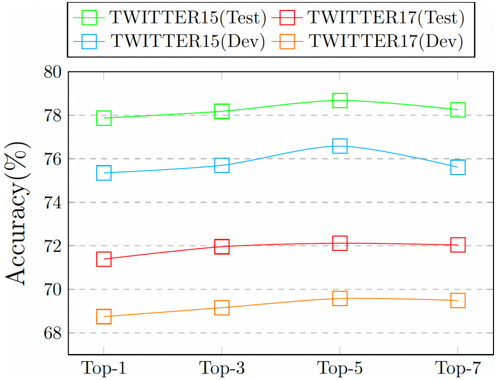


Figure 8: The results of *KEF-TomBERT* under different numbers of ANPs. Dev is short for development set.

Case Study

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

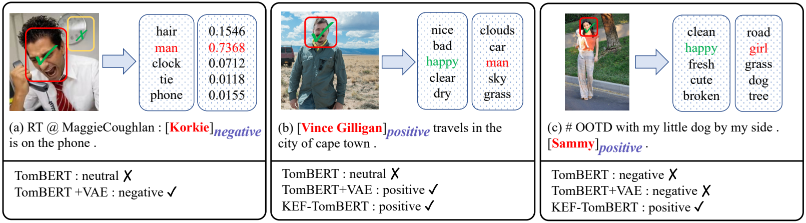


Figure 9: Predictions of *TomBERT*, *TomBERT+VAE* and *KEF-TomBERT* on three samples. The yellow/red bounding box are the visual clues that the opinion target focuses on under different methods.

- In this paper, we propose a novel knowledge-enhanced Framework (*KEF*) for the TMSC task.
- We design two novel knowledge enhancers, Visual Attention Enhancer and Sentiment Prediction Enhancer, to improve the visual attention capability and sentiment prediction capability of the TMSC task.
- Results from numerous experiments indicate that our model achieves better performance than other state-of-the-art methods. Our code and datasets are available at `https://github.com/1429904852/KEF`.

Thanks!

[CSBY17] Peng Chen, Zhongqian Sun, Lidong Bing, and Wei Yang.
Recurrent attention network on memory for aspect sentiment analysis.
In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pages 452–461, 2017.

[DCLT19] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova.
BERT: pre-training of deep bidirectional transformers for language understanding.
In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *NAACL 2019*, pages 4171–4186. Association for Computational Linguistics, 2019.

[FFZ18] Feifan Fan, Yansong Feng, and Dongyan Zhao.
Multi-grained attention network for aspect-level sentiment classification.
In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3433–3442, 2018.

[KF21] Zaid Khan and Yun Fu.
Exploiting BERT for multimodal target sentiment classification through input space translation.
In Heng Tao Shen, Yueting Zhuang, John R. Smith, Yang Yang, Pablo Cesar, Florian Metze, and Balakrishnan Prabhakaran, editors, *MM '21: ACM Multimedia Conference, Virtual Event, China, October 20 - 24, 2021*, pages 3034–3042. ACM, 2021.

[TQL16] Duyu Tang, Bing Qin, and Ting Liu.
Aspect level sentiment classification with deep memory network.
In Jian Su, Xavier Carreras, and Kevin Duh, editors, *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. The Association for Computational Linguistics, 2016.

[WHZ+16] Yequan Wang, Minlie Huang, Li Zhao, et al.
Attention-based lstm for aspect-level sentiment classification.
In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pages 606–615, 2016.

[WLS+21] Jiawei Wang, Zhe Liu, Victor Sheng, Yuqing Song, and Chenjian Qiu.
Saliencybert: Recurrent attention network for target-oriented multimodal sentiment classification.
In Huimin Ma, Liang Wang, Changshui Zhang, Fei Wu, Tieniu Tan, Yaonan Wang, Jianhuang Lai, and Yao Zhao, editors, *Pattern Recognition and Computer Vision - 4th Chinese Conference, PRCV 2021, Beijing, China, October 29 - November 1, 2021, Proceedings, Part III*, volume 13021 of *Lecture Notes in Computer Science*, pages 3–15. Springer, 2021.

[XMC19] Nan Xu, Wenji Mao, and Guandan Chen.
Multi-interactive memory network for aspect based multimodal sentiment analysis.
In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 371–378, 2019.

[YJ19] Jianfei Yu and Jing Jiang.
Adapting BERT for target-oriented multimodal sentiment classification.
In Sarit Kraus, editor, *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, pages 5408–5414. ijcai.org, 2019.

[YJX19] Jianfei Yu, Jing Jiang, and Rui Xia.
Entity-sensitive attention and fusion network for entity-level multimodal sentiment classification.
IEEE/ACM Transactions on Audio, Speech, and Language Processing, 28:429–439, 2019.

[ZWL+21] Zhe Zhang, Zhu Wang, Xiaona Li, Nannan Liu, Bin Guo, and Zhiwen Yu.
Modalnet: an aspect-level sentiment classification model by exploring multimodal data with fusion discriminant attentional network.
World Wide Web, pages 1–18, 2021.

[ZZH+21] Jie Zhou, Jiabao Zhao, Jimmy Xiangji Huang, Qinmin Vivian Hu, and Liang He.
Masad: A large-scale dataset for multimodal aspect-based sentiment analysis.
Neurocomputing, 455:47–58, 2021.