

MODALITY-DEPENDENT SENTIMENTS EXPLORING FOR MULTI-MODAL SENTIMENT CLASSIFICATION

Jingzhe Li¹²³, Chengji Wang^{123*}, Zhiming Luo⁴, Yuxian Wu¹²³, Xingpeng Jiang^{123*}

¹Hubei Provincial Key Laboratory of Artificial Intelligence and Smart Learning

²National Language Resources Monitoring and Research Center for Network Media

³School of Computer Science, Central China Normal University, Wuhan, China

⁴Department of Artificial Intelligence, Xiamen University, Xiamen, China

ABSTRACT

Recognizing human feelings from image and text is a core challenge of multi-modal data analysis, often applied in personalized advertising. Previous works aim at exploring the shared features, which are the matched contents between images and texts. However, the modality-dependent sentiment information (private features) in each modality is usually ignored by cross-modal interactions, the real sentiment is often reflected in one modality. In this paper, we propose a Modality-Dependent Sentiment Exploring framework (MDSE). First, to exploit the private features, we compare shared features with original image or text features, identifying previously overlooked unimodal features. Fusing the private and shared features can make the model more robust. Second, in order to obtain unified sentiment representations, we treat unimodal features and multi-modal fused features equally. We introduce a Modality-Agnostic Contrastive Loss (MACL) that performs contrastive learning between unimodal features and multi-modal fused features. The MACL can fully exploit sentiment information from multi-modal data and reduce the modality gap. Experiments on four public datasets demonstrate the effectiveness of our MDSE compared with existing methods. The full codes are available at <https://github.com/royal-dargon/MDSE>.

Index Terms— Multi-modal sentiment analysis, Private feature learning, Contrastive learning

1. INTRODUCTION

Multi-modal sentiment analysis is commonly employed to discern human emotions from diverse data sources. With the development of social networking platforms, e.g., Twitter and Facebook, people like to share their feelings on social networks. How to mine sentiments in social media data efficiently and correctly has broad applications, e.g. personalized advertising, community opinion mining. It has become a popular research topic and attracted attention from both the industrial and academic communities [1, 2]. In this paper, we focus on text-image sentiment analysis in social media data.

Compared to the unimodal sentiment analysis, the complementary between text and image can help the model analyze the real sentiment of the multi-modal data. On the one hand, images and texts can express the same sentiment with different data forms; this is shown as the matched contents between images and texts. We call them shared features. On the other hand, images and texts have their own private features that cannot be found in another modality. The private features encode modality-dependent sentiments, which can help detect the real sentiment more accurately. Distinguishing between shared and private features among modalities can improve the accuracy of multi-modal sentiment analysis. Works [3, 4] find that both audio, image, and text contain private information; this kind of information is crucial for multi-modal sentiment analysis.

Previous text-image sentiment analysis works [5–9] focused on exploring the shared features. They used cross-modal interaction to explore the matched sentiment information from images and texts. There are several cross-modal interaction methods: word-image interaction [10] exploits the correlation between image and words; word-object interaction [8] captures word-object correspondence to investigate the relationship between affective image regions and words for multi-modal sentiment analysis. Works [6, 7] utilize Transformer to model the implicit relationships between image patches and words. The cross-modal interaction is based on the attention mechanism; it uses cross-modal attention to perform multi-modal feature interaction and fusion. These methods [6–8, 10] treat each modal semantic as a whole; they do not distinguish shared features from private features. In other words, existing text-image sentiment analysis works ignore the important role of private features.

In this paper, we propose a Modality-Dependent Sentiments Exploring framework for multi-modal sentiment classification, which addresses the above issue in two ways. Firstly, cross-modal interaction usually employs a cross-attention mechanism, which tends to explore corresponding features from two modalities. We calculate a similarity map between the features before and after the interaction, with

*Corresponding Author: {wcj,xpjjiang}@ccnu.edu.cn

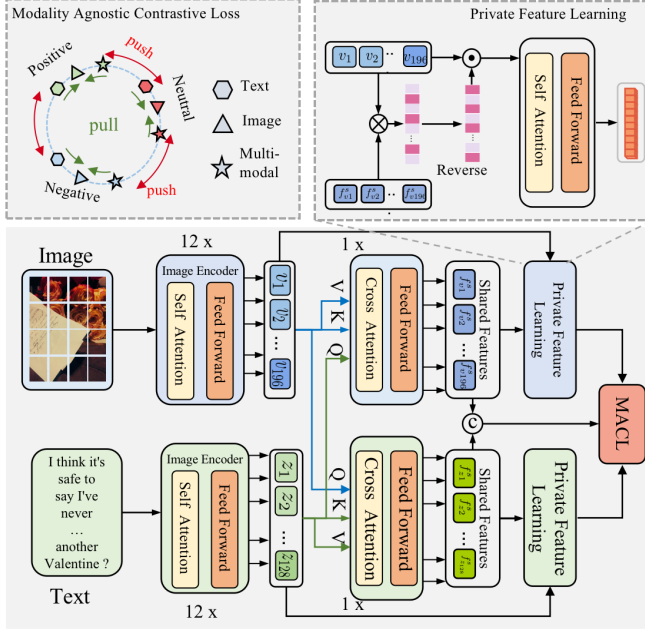


Fig. 1. Overview of the proposed MDSE framework

private features exhibiting a lower response. We propose a private feature learning module that reverses the similarity map to pinpoint which features are ignored. We then multiply the reversed similarity map with the original unimodal features to extract private features from each modality. Secondly, in order to exploit the unimodal sentiment information, we propose a Modality-Agnostic Contrastive Loss to align unimodal and multi-modal features. MACL concurrently handles unimodal and multi-modal sentiment classification. This approach treats unimodal and multi-modal features as different samples, enabling the mining of positive and negative samples across three feature types.

2. METHODOLOGY

The framework of our proposed method is shown in Fig. 1. A pre-trained ViT model [11] is adopted to obtain image patch embeddings $\{v_i \in R^d\}_{i=1}^M$, the image is divided into M patches. For the text branch, we take a pre-trained RoBERTa model [12] to extract word embeddings $\{z_i \in R^d\}_{i=1}^N$, the maximum sentence length is N . The dual cross-modal interaction module utilizes the cross-attention block to extract multi-modal shared features. In private feature learning, overlooked unimodal features are extracted by reversing the similar map. The modality-agnostic contrastive loss is designed to enhance the learning of more robust, modality-invariant sentiment embeddings.

2.1. Dual Cross-Modal Interaction

To comprehensively explore multi-modal shared features, it is essential to model the fine-grained correspondence between images and text. As is shown in Fig. 1, the dual cross-modal interaction module has one co-attentional transformer layer [13]. A co-attentional transformer layer contains a cross-attention block (CA) and a feed-forward layer (FF).

Given unimodal features v and z , we adopt a co-attentional transformer layer to obtain intermediate image-text features:

$$\begin{aligned} f_v^s &= FF(CA(z, v, v)) \\ f_z^s &= FF(CA(v, z, z)) \end{aligned} \quad (1)$$

where $f_v^s = \{f_{v_i}^s\}_{i=1}^M$ are the image intermediate features, $f_{v_i}^s \in R^d$, $f_z^s = \{f_{z_i}^s\}_{i=1}^N$ are the text intermediate features, $f_{z_i}^s \in R^d$. We implement the CA block by multi-head attention [14]. The three inputs in CA are the query, key, and value of multi-head attention. Following the cross-attention block, we can explore the relationships between image patches and words, which helps our MDSE effectively explore the fine-grained correspondence.

We concatenate the tokens from two modalities and perform an average pooling to obtain the final shared feature $f^s \in R^d$:

$$f^s = avg([f_v^s, f_z^s]) \quad (2)$$

where $[\cdot, \cdot]$ is concatenation operation and $avg()$ stands for average pooling.

2.2. Private Feature Learning

As is shown in Fig. 1, we propose a private feature learning module. Given unimodal image embeddings v and the intermediate features f_v^s , we obtain the similarity matrix S by computing the similarities between every image patch embedding v_i and intermediate image patch feature $f_{v_j}^s$:

$$S_{ij} = v_i^T f_{v_j}^s \quad (3)$$

where $S \in R^{M \times M}$. We calculate the sum of every column of S and the normalized vector $\bar{S} \in R^{1 \times M}$ reflects the importance of different patches:

$$\bar{S} = softmax(\sum_j (S_{ij})) \quad (4)$$

We reverse the importance vector \bar{S} and multiply it with the original image embeddings to obtain the ignored features:

$$v^p = (1 - \bar{S}) \cdot v \quad (5)$$

The ignored features also contain noise information. We use a self-attention layer and a feed-forward layer to filter the useless features:

$$f_v^p = avg(FF(SA(v^p, v^p, v^p))) \quad (6)$$

where avg is global average pooling. $f_v^p \in R^d$ is the private image feature. Private feature and shared feature are complementary; we fuse private feature and shared feature to generate the final image embedding:

$$f_v = FF([f_v^s, f_v^p]) \quad (7)$$

where $f_v \in R^d$.

In the same way, given unimodal word embeddings z and the intermediate features f_z^s , we extract the private text feature $f_z^p \in R^d$, and the final text embedding is represented by $f_z \in R^d$.

We combine the image and text embeddings to obtain the final multi-modal sentiment representation:

$$f = FF([f_v, f_z]) \quad (8)$$

Table 1. Experimental results of the model on the MVSA-S and MVSA-M datasets (%).

Models	MVSA-S		MVSA-M	
	ACC	W-F1	ACC	W-F1
MultiSentiNet [15]	69.84	69.63	68.86	68.11
Co-MN-Hop [16]	70.51	70.01	68.92	68.83
MFF [17]	71.44	71.06	69.62	69.35
MGNNs [18]	73.77	72.70	72.49	69.34
TBNMD [19]	75.22	73.46	70.72	67.94
CLMLF [7]	75.33	73.46	72.00	69.83
MDSE(Base-VGG-19)	74.33	74.38	70.17	68.74
MDSE(VGG-19)	74.22	73.20	71.23	68.10
MDSE(Base-ResNet)	73.33	72.74	70.29	67.65
MDSE(ResNet)	75.93	74.83	71.97	69.92
MDSE(Base-ViT)	73.77	72.52	71.00	67.83
MDSE(ViT)	76.22	75.71	72.31	70.12

where $f \in R^d$.

We feed the final multi-modal feature into a multi-layer perceptron for classification, and a cross-entropy loss is used for sentiment classification:

$$L_{SC} = CrossEntropy(MLP(f)) \quad (9)$$

2.3. Modality-Agnostic Contrastive Loss

The modal gap is a major challenge for multi-modal sentiment analysis; the representation of individual modalities and multi-modal representations should have the same ability to represent sentiments. So, the unimodal and multi-modal fused features should be treated equally. As is shown in Fig. 1, we propose a modality-agnostic contrastive loss to learn modality-invariant sentiment representations. MACL considers the image, text, and image-text pair as three samples; it performs contrastive learning among them.

Given a mini-batch of image-text pairs with B samples, there are three sentiment labels (image sentiment label, text sentiment label, and multi-modal sentiment label). We obtain sentiment representations $e \in R^{3B \times d}$ (consisting of B image embeddings, B text embeddings, and B multi-modal embeddings), and their respective sentiment labels are $l \in R^{3B}$. We then calculate the cosine distance between them:

$$C_{ij} = \cos(e_i, e_j) \quad (10)$$

We hope samples with the same label to be as similar as possible and samples with different labels as far away as possible:

$$L_{MACL} = -\log \frac{\sum_{l_i=l_j} C_{ij}}{\sum_{i,j} C_{ij}} \quad (11)$$

3. EXPERIMENTS

3.1. Experiment Settings

Dataset. We evaluate our method on MVSA-S [20], MVSA-M [20], TWITTER-15 [21] and TWITTER-17 [21]. All

Table 2. Experimental results of the model on the Twitter datasets (%).

Models	TWITTER-15		TWITTER-17	
	ACC	M-F1	ACC	M-F1
TomBERT [21]	77.15	71.75	70.50	68.04
CapTrBERT [22]	78.01	73.25	72.30	70.20
TBNMD [19]	76.73	71.19	71.52	70.18
CLMLF [7]	78.11	74.60	70.98	69.13
MDSE(Base-VGG-19)	73.44	73.14	70.58	70.52
MDSE(VGG-19)	75.77	73.22	70.98	71.02
MDSE(Base-ResNet)	75.04	73.75	70.91	70.92
MDSE(ResNet)	78.05	74.92	71.93	71.88
MDSE(Base-ViT)	75.60	74.28	70.98	70.91
MDSE(ViT)	78.49	75.29	72.77	72.63

datasets consist of paired text and image data collected from the Twitter platform, including three different types of sentiments, positive, neutral and negative. All the categories for both datasets are splitted as CLMLF [7].

Implementation Details. We use PyTorch to implement our model. The self attention and cross attention has 8 heads. We apply image augmentation techniques, including horizontal flipping, random cropping with padding, and random erasing. We use synonym replacement for text augmentation. Images are resized to 224×224. The maximum length of text sequence $N = 128$. Batch size is 64. The feature dimension is 768. We train the model using the AdamW optimizer for 50 epochs with a learning rate of 2e-5. Experiments are conducted on one NVIDIA 4090 GPU.

3.2. Overall Result

Table 1 and Table 2 illustrate the performance comparison of our MDSE model with state-of-the-art methods. We adopt the same evaluation metrics as CLMLF [7]. In MVSA-S and MVSA-M, we use ACC and Weighted-F1 (W-F1). In TWITTER-15 and TWITTER-17, we use ACC and Macro-F1 (M-F1). We have the following observations: Firstly, our model is competitive with other strong models on all four datasets. This demonstrates that the proposed modality-dependent sentiment exploring framework is effective for multi-modal sentiment analysis. To the best of our knowledge, our work is the first text-image sentiment analysis approach to address the private feature extraction issue. Comparing with CLMLF [7] multi-modal fusion module has 88.4M parameters, our proposed MDSE introduces only 17.7M. The MDSE model demonstrates improvement in the metrics, highlighting the significance of private features. Secondly, we also evaluate our MDSE with ResNet [23] and VGG-19 [24]. There is a performance degradation by replacing ViT with ResNet, but MDSE(ResNet) has still achieved competitive results. Using a strong backbone (ViT vs. ResNet) only results in a 0.5% improvement. This result shows that our method is effective. Furthermore, we com-

Table 3. Ablation analysis of MDSE (%).

Components			MVSA-S		TWITTER-15	
iPFL	tPFL	MACL	ACC	W-F1	ACC	M-F1
			73.77	75.52	75.60	78.28
✓			74.07	72.83	76.08	74.52
	✓		74.66	73.50	75.60	75.02
✓	✓		75.44	74.91	77.85	75.12
✓		✓	75.55	74.52	76.69	75.14
	✓	✓	75.33	74.31	77.07	75.15
✓	✓	✓	76.22	75.71	78.49	75.29

pare our MDSE(VIT) with baselines (MDSE(Base-VIT)). Compared to the baseline, the MDSE has shown significant performance improvements.

3.3. Ablation Study

We conduct the ablation experiments to distinguish the contribution of each component. As shown in Table 3, we evaluate both private feature learning (PFL) and modality-agnostic contrastive loss (MACL) on MVSA-S and TWITTER-15. iPFL stands for image private feature learning, and tPFL is text private feature learning. We can observe that either private feature learning or modality agonist contrastive loss hurts the model’s performance, which indicates both PFL and MACL are useful for sentiment prediction.

Table 4. Ablation results of feature adaption in PFL (%).

Method	MVSA-S		TWITTER-15	
	ACC	W-F1	ACC	M-F1
PFL w/o ada	75.33	74.36	77.25	73.95
PFL	76.22	75.71	78.49	75.29

The effectiveness of private feature learning. We evaluate the contributions of iPFL and tPFL, respectively. From Table 3, we can observe that both iPFL and tPFL can significantly improve model performance. When removing either iPFL or tPFL, we observe a significant performance degradation. The results in Table 3 signify the importance of private feature learning. In fact, both images and texts contain rich private sentiment information. Cross-modal interaction focuses more on the common features between the two modalities; and does not efficiently capture private features. In private feature learning, we use a feature adaptation block (ada, Eq. 6) to filter out useless features. In Table 4, we show its contributions. The feature adaption block plays an essential role in private feature learning. It works in two ways: 1) filtering out useless features; and 2) making private and shared features have similar distributions. In Fig. 2, we compare the sentiment classification scores. In the first image-text data pair, the labels of image and text are neutral and positive, respectively, and the final label is positive. From the result, if the text private feature learning(tPFL) module is ignored, the final result will be wrong, indicating that the part marked by the green box in the text data is important to the result. In the second data pair, the image part is marked with a green box as the private feature of the image, and it can be found that if ignored, the result will become neutral.

The effectiveness of modality-agnostic contrastive loss. MACL performs contrastive learning among image embed-



Image	Text	Label	MDSE	-tPFL	-iPFL
 (Neutral)	Minister of Transportation Steven DelDuca says "these are the investments that will help all of us" of LPC plan! (Positive)	Positive	0.8077	0.2629	0.6903
		Neutral	0.0999	0.2842	0.1030
		Negative	0.0924	0.4529	0.2067
 (Negative)	Big-spending election campaign set to ramp up airwaves war. (Neutral)	Positive	0.0250	0.2973	0.0286
		Neutral	0.3073	0.2105	0.5083
		Negative	0.6677	0.4922	0.4676

Fig. 2. Visualizations of sentiment classification scores.

ding f_v , text embedding f_z and multi-modal representation f . Table 3 shows the effectiveness of our MACL. Whether used in combination with iPFL or tPFL, MACL delivers steady performance gains. The results in Table 3 demonstrate that MACL can help models learn modality-invariant sentiment representations. In Table 5, we show the results of MACL after removing multi-modal features. We can observe that simply reducing modal differences is not enough. By performing contrastive learning between f_v and f_z , the model cannot obtain effective multi-modal sentiment representations. It is more effective to perform contrastive learning between multi-modal fused features and unimodal features.

Table 5. Ablation results of MACL (%).

Method	MVSA-S		TWITTER-15	
	ACC	W-F1	ACC	M-F1
f_v, f_z	75.83	74.03	77.18	74.73
f, f_v, f_z	76.22	75.71	78.49	75.29

4. CONCLUSIONS

In this paper, we introduce modality-dependent sentiments exploring framework for multi-modal sentiment classification(MDSE). MDSE aims to address the issue of overlooking modality-dependent private features in previous methods, resulting in learning of unrobust multi-modal sentiment representations. To address the issue, we propose a private feature learning module. It extracts modality-dependent private features by mining unfocused unimodal features. We further propose a modality-agnostic contrastive loss to perform contrastive learning among unimodal features and multi-modal features. These modules collaborate to learn robust multi-modal sentiment representations. Significant performance gains on four popular benchmark datasets prove the superiority and effectiveness of our proposed MDSE framework. We believe that private feature learning will play an important role in multi-modal analysis.

5. ACKNOWLEDEGMENT

This work was partially supported by the National Language Commission Key Research Project (ZDI145-56), the National High-end Foreign Expert Cooperation Project (G2022158003L), the China Postdoctoral Science Foundation (2023M741305), the Fundamental Research Funds for the Central Universities (CCNU23XJ001) and Postdoctoral innovation research post in Hubei Province.

6. REFERENCES

- [1] Ringki Das, Singh, and Thoudam Doren, "Multimodal sentiment analysis: A survey of methods, trends, and challenges," *ACM Computing Surveys*, vol. 55, no. 270, 2023.
- [2] Ankita Gandhi, Kinjal Adhvaryu, Soujanya Poria, Erik Cambria, and Amir Hussain, "Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions," *Information Fusion*, vol. 91, pp. 424–444, 2023.
- [3] Yang Wu, Zijie Lin, Yanyan Zhao, Bing Qin, and Li-Nan Zhu, "A text-centered shared-private framework via cross-modal prediction for multimodal sentiment analysis," in *Findings of ACL-IJCNLP*, 2021, pp. 4730–4738.
- [4] Songning Lai, Xifeng Hu, Yulong Li, Zhaoxia Ren, Zhi Liu, and Danmin Miao, "Shared and private information learning in multimodal sentiment analysis with deep modal alignment and self-supervised multi-task learning," *arXiv preprint arXiv:2305.08473*, 2023.
- [5] Tian Li, Xiang Chen, Shanghang Zhang, Zhen Dong, and Kurt Keutzer, "Cross-domain sentiment classification with contrastive learning and mutual information maximization," in *ICASSP*, 2021, pp. 8203–8207.
- [6] Yan Ling, Jianfei Yu, and Rui Xia, "Vision-language pre-training for multimodal aspect-based sentiment analysis," in *ACL*, 2022, pp. 2149–2159.
- [7] Zhen Li, Bing Xu, Conghui Zhu, and Tiejun Zhao, "CLMLF: a contrastive learning and multi-layer fusion method for multimodal sentiment detection," in *Findings of NAACL*, 2022, pp. 2282–2294.
- [8] Tong Zhu, Leida Li, Jufeng Yang, Sicheng Zhao, Hantao Liu, and Jiansheng Qian, "Multimodal sentiment analysis with image-text interaction network," *IEEE TMM*, 2022.
- [9] Haidong Zhu, Zhaoheng Zheng, Mohammad Soleymani, and Ram Nevatia, "Self-supervised learning for sentiment analysis via image-text matching," in *ICASSP*, 2022, pp. 1710–1714.
- [10] Ashima Yadav and Dinesh Kumar Vishwakarma, "A deep multi-level attentive network for multimodal sentiment analysis," *ACM TOMM*, vol. 19, no. 1, pp. 1–19, 2023.
- [11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al., "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.
- [12] 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," *arXiv preprint arXiv:1907.11692*, 2019.
- [13] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee, "Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks," *NeurIPS*, vol. 32, 2019.
- [14] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," *NeurIPS*, vol. 30, 2017.
- [15] Nan Xu and Wenji Mao, "Multisentinet: A deep semantic network for multimodal sentiment analysis," in *CIKM*, 2017, pp. 2399–2402.
- [16] Nan Xu, Wenji Mao, and Guandan Chen, "A co-memory network for multimodal sentiment analysis," in *SIGIR*, 2018, pp. 929–932.
- [17] Kang Zhang, Yushui Geng, Jing Zhao, Jianxin Liu, and Wenxiao Li, "Sentiment analysis of social media via multimodal feature fusion," *Symmetry*, vol. 12, no. 12, pp. 2010, 2020.
- [18] Xiaocui Yang, Shi Feng, Yifei Zhang, and Daling Wang, "Multimodal sentiment detection based on multi-channel graph neural networks," in *ACL*, 2021, pp. 328–339.
- [19] Siqi Li, Weihong Deng, and Jiani Hu, "Momentum distillation improves multimodal sentiment analysis," in *PRCV*. Springer, 2022, pp. 423–435.
- [20] Teng Niu, Shiai Zhu, Lei Pang, and Abdulmotaleb El Saddik, "Sentiment analysis on multi-view social data," in *MMM*, 2016, pp. 15–27.
- [21] Jianfei Yu and Jing Jiang, "Adapting bert for target-oriented multimodal sentiment classification," in *IJCAI*, 2019, pp. 5408–5414.
- [22] Zaid Khan and Yun Fu, "Exploiting bert for multimodal target sentiment classification through input space translation," in *ACM MM*, 2021, pp. 3034–3042.
- [23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.
- [24] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.