

Supplementary Materials: Sentiment-oriented Sarcasm Integration for Video Sentiment Analysis Enhancement with Sarcasm Assistance

Junlin Fang

Southwest Jiaotong University
School of Computing and Artificial Intelligence
Chengdu, China
jlfang@my.swjtu.edu.cn

Guosheng Lin

Nanyang Technological University
School of Computer Science and Engineering
Singapore, Singapore
gslin@ntu.edu.sg

Wenya Wang

Nanyang Technological University
School of Computer Science and Engineering
Singapore, Singapore
wangwy@ntu.edu.sg

Fengmao Lv*

Southwest Jiaotong University
School of Computing and Artificial Intelligence
Chengdu, China
fengmaolv@swjtu.edu.cn

A Experiments on the text-only setting

In order to validate the scalability of the proposed PS2RI framework, we also conduct experiments on the text-only setting by considering only the textual modality of the MUSTARD dataset. Specifically, we use the pre-trained ALBERT encoder composed of 12 layers as the base model \mathcal{M} . The \mathcal{M} -SEN model applies a linear layer on top of the base model \mathcal{M} and is only trained under the supervision of sentiment labels. Similar to Section 5.3, we implement \mathcal{M} -MTL by using two classification heads on top of the base model \mathcal{M} to respectively generate sarcasm and sentiment predictions. \mathcal{M} -MTL is trained under the supervision of both sarcasm and sentiment labels. As the input only involves the text-modality information, we implement PS2RI by removing the multimodal interaction or fusion operations (i.e., removing MTF in the sarcasm feature encoder and the gate unit in SOSR blocks, as well as replacing MTF layers in the SASL module with conventional transformer layers). The results are reported in Table A1. We can see that \mathcal{M} -MTL is still significantly inferior to the \mathcal{M} -SEN base model on Subset 2 consisting of non-sarcastic samples. The observation is consistent with our motivation (i.e., the sentiment analysis task suffers the negative interference of the sarcasm detection task within the MTL framework). From the last row, we can see that our proposed PS2RI approach achieves performance improvements on both Subset 1 and Subset 2. The above results demonstrate that our proposed PS2RI mechanism can be also effective when applied into the text-only setting.

B Scalability with other sentiment-related tasks

In order to explore the effect of other sentiment-related sub-tasks, we utilize the UR-FUNNY benchmark [2] focusing on human humor and the Memotion benchmark [4] focusing on offense intentions to improve the sentiment recognition task. Specifically, we replace the Sarcasm Feature Encoder with a Sentiment-related Feature Encoder which is trained under the supervision of humor and offense labels from the UR-FUNNY and Memotion datasets, respectively. Then, we train the PS2RI framework on the MUSTARD benchmark with the Sentiment-related Feature Encoder trained with different

Table A1: Results on the text-only setting of MUSTARD. The results are reported in terms of the weighted-F1.

Benchmark	Method	Entire testing set	Subset 1	Subset 2
MUSTARD [1]	\mathcal{M} -SEN	39.81	43.16	37.93
	\mathcal{M} -MTL	40.11	46.79	34.17
	PS2RI (ours)	42.07	47.16	38.44

Algorithm 1 The forward propagation procedure of Progressive Sentiment-oriented Sarcasm Refinement and Integration.

Input: the unimodal features from encoders: $H_{l,v,a}^{sen} \in \mathbb{R}^{(T_m^u+T_m^c) \times d^H}$; the sarcasm features from Sarcasm Feature Extraction module: $Z_{sar} \in \mathbb{R}^{(T_m^u+T_m^c) \times d^H}$; the number of layers: J .

Output: the multimodal sentiment features Z_{sen} .

- 1: Initialize the sarcasm features: $Z_{sar}^0 = Z_{sar}$;
- 2: Initialize the sentiment features: $H_m^0 = H_m^{sen}$, where $m \in \{l, v, a\}$;
- 3: $j = 1$;
- 4: **while** $j \leq J$ **do**
- 5: Update the sarcasm features:
 $Z_{sar}^j = \text{SOSR}^j(Z_{sar}^{j-1}, H_l^{j-1}, H_v^{j-1}, H_a^{j-1})$;
- 6: Produce the sarcasm-aware sentiment features:
 $\hat{H}_m^j = \text{SI}_m^j(Z_{sar}^j, H_m^{j-1})$;
- 7: Integrate information from multiple modalities:
 $H_m^j = \text{MTF}_m^j(\hat{H}_l^j, \hat{H}_v^j, \hat{H}_a^j)$;
- 8: $j = j + 1$;
- 9: **end while**
- 10: Produce the final sentiment features Z_{sen} by Eq. (??) with the input of H_l^J, H_v^J and H_a^J ;
- 11: **return** Z_{sen} .

sentiment-related information. To simplify the notation, we refer to the Sentiment-related Integration block as the SI block, which performs similar operations to the original Sarcasm Integration block. Furthermore, the Sentiment-Oriented Sentiment-related Refinement block is denoted as the SOSR block and has the similar

*Corresponding author.

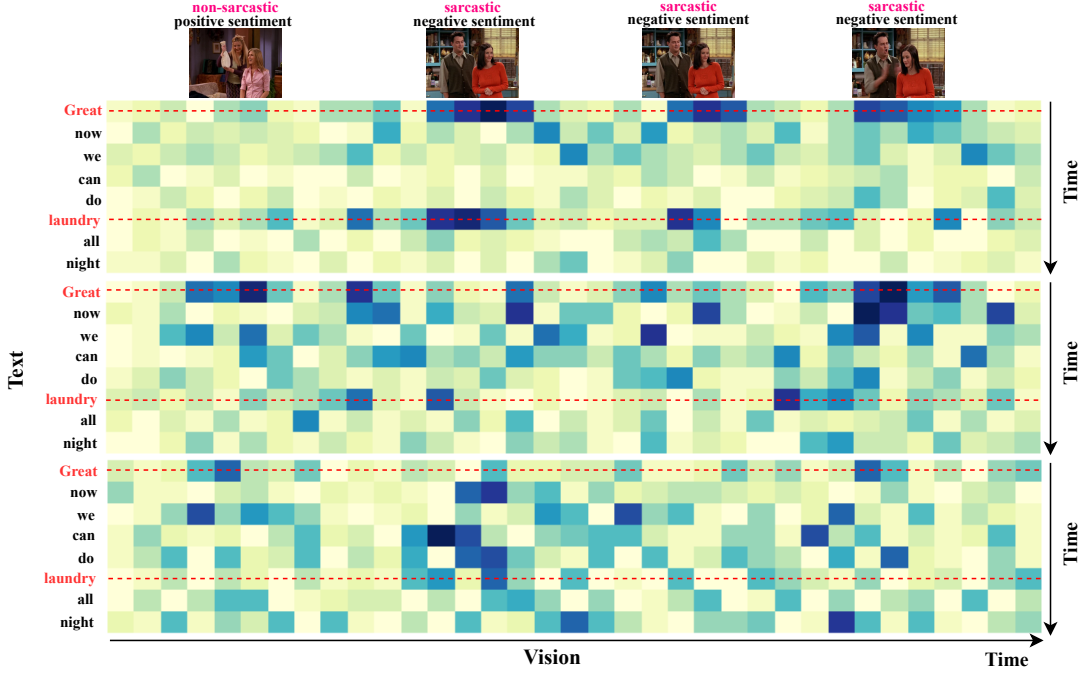


Figure A1: Visualization for the crossmodal correlations on the MUSTARD benchmark. The visualization samples of our approach, DMD, and \mathcal{F} -MTL are displayed in the upper part, the middle part, and the bottom part, respectively. We conduct the visualization by observing the crossmodal attention weights of the MTF unit in the fourth SASL layer. The textual words which are closely related to sarcastic information are displayed in red. The textual words above the video clips are the corresponding spoken words.

operations with the original Sentiment-Oriented Sarcasm Refinement block. The results are reported in Table A2. The first row of each source benchmark shows the performance of the full model with sentiment-related information enhancement. In the second row of each source dataset, we remove the SI block, leading to the base \mathcal{F} -SEN model which only involves the supervision of sentiment labels. Compared to the first row, the consistent performance drops demonstrate that PS2RI can boost the sentiment recognition task with other sentiment-related information. In the third row of each source dataset, we remove the SOSR blocks from the full model by directly integrating the original sentiment-related feature generated from the Sentiment-related Feature Encoder into sentiment features. The performance degradation compared to that of the first row indicates that the original sentiment-related feature can bring negative interference for sentiment recognition. This observation further validates the effectiveness of modeling sentiment-oriented features via the SOSR blocks, which is consistent with our motivation.

C Visualization

Figure A1 displays the visualization for the crossmodal interaction between elements in our study. Specifically, the upper, middle and lower parts show visualization samples for our approach, DMD [3], and \mathcal{F} -MTL, respectively. In the example, the textual words “great” and “laundry” form the sarcastic sentiment. We can see that our approach can correlate the sarcasm-related textual

Table A2: Experimental results of utilizing the sentiment-related information to enhance PS2RI on the MUSTARD benchmark. The sentiment-related information respectively includes the humor information from the UR-FUNNY benchmark and the offence information from the Memotion benchmark. The results are reported in terms of the weighted-F1.

Source Benchmark	Method	Entire testing set	Subset 1	Subset 2
UR-FUNNY [2]	PS2RI (with humor)	56.03	63.11	52.49
	w/o SI	54.12	62.89	47.58
	w/o SOSR	53.76	59.14	50.37
Memotion [4]	PS2RI (with offense)	55.97	63.01	51.25
	w/o SI	54.12	62.89	47.58
	w/o SOSR	53.17	58.87	49.61

words with sarcasm-related video clips well, although they are not well aligned in time. In contrast, the visualization samples of both DMD and \mathcal{F} -MTL fail to model sarcasm-related crossmodal interactions. This observation qualitatively demonstrates that our approach can encourage the model to attend to more sarcastic signals across modalities.

References

- [1] Santiago Castro, Devamanyu Hazarika, Verónica Pérez-Rosas, Roger Zimmermann, Rada Mihalcea, and Soujanya Poria. 2019. Towards Multimodal Sarcasm Detection (An *Obviously* Perfect Paper). In *ACL*, Anna Korhonen, David R. Traum, and Lluís Màrquez (Eds.), 4619–4629.

- [2] Md. Kamrul Hasan, Wasifur Rahman, AmirAli Bagher Zadeh, Jianyuan Zhong, Md. Iftekhar Tanveer, Louis-Philippe Morency, and Mohammed (Ehsan) Hoque. 2019. UR-FUNNY: A Multimodal Language Dataset for Understanding Humor. In *EMNLP*. 2046–2056.
- [3] Yong Li, Yuanzhi Wang, and Zhen Cui. 2023. Decoupled Multimodal Distilling for Emotion Recognition. In *CVPR*. 6631–6640.
- [4] Chhavi Sharma, Deepesh Bhageria, William Scott, Srinivas PYKL, Amitava Das, Tanmoy Chakraborty, Viswanath Pulabaigari, and Björn Gambäck. 2020. SemEval-2020 Task 8: Memotion Analysis- the Visuo-Lingual Metaphor!. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*. 759–773.