

Supplementary Materials: Aspects are Anchors: Towards Multimodal Aspect-based Sentiment Analysis via Aspect-driven Alignment and Refinement

Anonymous Authors

1 GENERALIZATION CAPABILITY ON MULTIMODAL SENTIMENT ANALYSIS

To verify the generalization capability of our ADAR on other tasks, we conduct experiments on two datasets, MOSI [10] and MOSEI [11], for multimodal sentiment analysis. MOSI is a collection of YouTube monologues consisting of 2,199 movie samples, which are separated into 1,284 training samples, 229 validation samples, and 686 testing samples. MOSEI is an improvement over MOSI with a total of 23,453 video clips, spanning 1,000 distance speakers. To comprehensively validate the performance of our ADAR, we make a comparison with several advanced methods, including LMF [4], MFM [6], MuT [5], MISA [3], Self-MM [9], MMIM [2], FDMER [8], and DBF [7]. In order to fit our method for the task, we employ the pre-trained WavLM [1] to extract the audio features and conduct optimal transport similar to the text and visual features.

As shown in Table 1, we find that our model still maintains a competitive performance over all comparative baselines. It is worth noting that our model obtains the second best results on the metrics of *Acc-2* and *F1* on both datasets, demonstrating its excellent generalization ability. This performance gain is largely due to our alignment and refinement method, which aligns disparate modalities using aspect-driven optimal transport and augments their synergy through a dual-layer refinement process. Moreover, ADAR also shows an effective fine-grained analysis ability for its superior performance on *Acc-7*. In summary, the study presents compelling evidence of ADAR’s consistent ability to enhance performance in Multimodal Sentiment Analysis.

REFERENCES

- [1] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al. 2022. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing* 16, 6 (2022), 1505–1518.
- [2] Wei Han, Hui Chen, and Soujanya Poria. 2021. Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis. *arXiv preprint arXiv:2109.00412* (2021).
- [3] Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. 2020. Misa: Modality-invariant and-specific representations for multimodal sentiment analysis. In *Proceedings of the 28th ACM international conference on multimedia*. 1122–1131.
- [4] Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. 2018. Efficient low-rank multimodal fusion with modality-specific factors. *arXiv preprint arXiv:1806.00064* (2018).
- [5] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the conference. Association for computational linguistics. Meeting*, Vol. 2019. NIH Public Access, 6558.
- [6] Yao-Hung Hubert Tsai, Paul Pu Liang, Amir Zadeh, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2018. Learning factorized multimodal representations. *arXiv preprint arXiv:1806.06176* (2018).
- [7] Shaoxing Wu, Damai Dai, Ziwei Qin, Tianyu Liu, Binghui Lin, Yunbo Cao, and Zhifang Sui. 2023. Denoising Bottleneck with Mutual Information Maximization for Video Multimodal Fusion. *arXiv preprint arXiv:2305.14652* (2023).
- [8] Dingkan Yang, Shuai Huang, Haopeng Kuang, Yangtao Du, and Lihua Zhang. 2022. Disentangled representation learning for multimodal emotion recognition.

In *Proceedings of the 30th ACM International Conference on Multimedia*. 1642–1651.

- [9] Wenmeng Yu, Hua Xu, Ziqi Yuan, and Jiele Wu. 2021. Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 35. 10790–10797.
- [10] Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. 2016. Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. *IEEE Intelligent Systems* 31, 6 (2016), 82–88.
- [11] AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2236–2246.

Methods	MOSI					MOSEI				
	MAE	Corr	Acc-7	Acc-2	F1	MAE	Corr	Acc-7	Acc-2	F1
LMF	0.917	0.695	33.20	-/82.50	-/82.40	0.623	0.677	48.00	-/82.00	-/82.10
MFMM	0.877	0.706	35.40	-/81.70	-/81.60	0.568	0.717	51.30	-/84.40	-/84.30
MuT	0.861	0.711	40.00	81.50/84.10	80.60/83.90	0.580	0.703	51.80	-/82.50	-/82.30
MISA	0.804	0.764	42.30	80.79/82.10	80.77/82.03	0.568	0.724	52.20	82.59/84.23	82.67/83.97
Self-MM	0.712	0.795	45.79	82.54/84.77	82.68/84.91	0.529	0.767	53.46	82.68/84.96	82.95/84.93
MMIM	<u>0.700</u>	<u>0.800</u>	46.65	84.14/ <u>86.06</u>	84.00/ <u>85.98</u>	<u>0.526</u>	<u>0.772</u>	<u>54.24</u>	82.24/85.97	82.66/ <u>85.94</u>
FDMER	0.724	0.788	44.10	-/84.60	-/84.70	0.536	0.773	54.10	-/ <u>86.19</u>	-/85.80
DBF	0.693	0.801	44.80	85.10/86.90	85.10/86.90	0.523	<u>0.772</u>	54.20	84.30/86.40	84.80/86.20
ADAR (ours)	0.703	0.790	<u>46.04</u>	<u>84.22</u> /85.93	<u>84.32</u> /85.44	0.533	0.771	54.11	<u>83.33</u> /85.92	<u>84.10</u> /85.56

Table 1: Performance on MOSI and MOSEI for multimodal sentiment analysis. The best results are in bold and the second best ones are underlined. Acc-2 denotes the accuracy over negative/non-negative, and F1 corresponds to negative/positive.