MetaOff-Meme: A Metaphor-Enriched Benchmark for Meme Offensiveness Detection

Bo Xu[1], Chenyuan Wang[1], Liang Zhao[1]*, Chuansen Yuan[1], Xinyu Chen[1], Jiuyan Sun[1], Jianshu Cao[1], Xutai Hou[1], Xinchen Xiao[1], Yulun Lin[1], Hongfei Lin[2], Feng Xia[3]*
[1]School of Software, Dalian University of Technology, Dalian, Liaoning, China
[2]School of Computer Science and Technology, Dalian University of Technology, Dalian, Liaoning, China
[3]School of Computing Technologie RMIT University, Melbourne, Victoria, Australia

A Related Work

A detailed comparison between MetaOff-Meme and other datasets is shown in Table 1.

Table 1: Comparison with existing metaphor datasets and offensive meme datasets. Our MetaOff-Meme is enriched with metaphorical content and diverse offensive topics, including group, gender, and political themes.

Datasets	Data Source	Metaphor Annotation(% Metaphor)	Offensive
MultiMET [13]	Social Media, Adv	✔(58%)	✗
MEMECAP [6]	Social Media	✔(89%)	✗
NYK-MS [2]	Cartoon	✔(50%)	✗
V-FLUTE [9]	Cartoon, Social Media	✔(33%)	✗
IRFL [12]	Web	✔(27%)	✗
MEMOTION [10]	Web	✗	✔
MAMI [5]	Social Media	✗	✔
HMC [7]	Social Media	✗	✔
MetaOff-Meme (ours)	Social Media	✔(82%)	✔

B Supplementary Experiments

B.1 Model Information

Since each model has distinct advantages, we select the optimal version of each based on its specific strengths.

- Qwen-VL [1]: Qwen-VL enhances the visual capabilities of VLM through various strategies and improves fine-grained recognition abilities, such as text reading, multilingual text recognition, and object localization, via a three-stage training process. In this study, we utilize the "qwen-vl-chat" version.
- LLaVA-v1.5 [8]: LLaVA-v1.5 systematically explores the construction of VLM, including higher-resolution inputs, compositional capabilities, and mitigation of model hallucination, leading to significant performance improvements. We utilize the "llava-v1.5-7b" version. This version demonstrates outstanding performance across multiple multimodal tasks, particularly excelling in understanding fine-grained image content.
- MiniGPT-v2 [3]: MiniGPT-v2 uses unique identifiers for different tasks during training, enhancing its learning efficiency across tasks. We construct this model using the "llama-2-7b-chat" LLaMA version. This version delivers superior performance across multiple tasks.
- InternVL [4]: InternVL extends its foundational visual model to 6 billion parameters, delivering robust visual capabilities and enabling the completion of multiple general vision-language tasks. We specifically utilize the "InternVL2-8B" version. This version excels in understanding complex scenes.
- MiniCPM-V [11]: MiniCPM-V integrates SOTA architecture, pretraining, and alignment techniques for multimodal large language models (MLLMs), supporting efficient operation on edge devices. We utilize the "MiniCPM-V2.6" version. This version is the most powerful model in the MiniCPM-V series, demonstrating exceptional capabilities in multi-image and video understanding.

C Limitations

When collecting offensive memes, we primarily restricted the scope to certain thematic categories, which may have resulted in relatively limited sample diversity within the dataset. Additionally, during the manual filtering process, background knowledge was required to identify metaphorical information, but some knowledge might have exceeded the cognitive scope of the collection team. As a result, MetaOff-Meme may not fully encompass all types of offensive memes, highlighting its limitations in diversity. In the future, we plan to involve more metaphor experts to further enrich the dataset and enhance the accuracy of the filtering process.

References

[1] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond. *arXiv preprint arXiv:2308.12966* (2023).

[2] Ke Chang, Hao Li, Junzhao Zhang, and Yunfang Wu. 2024. NYK-MS: A Well-annotated Multi-modal Metaphor and Sarcasm Understanding Benchmark on Cartoon-Caption Dataset. *arXiv preprint arXiv:2409.01037* (2024).

[3] Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. 2023. MiniGPT-v2: large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478* (2023).

[4] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. 2024. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 24185–24198.

[5] Elisabetta Fersini, Francesca Gasparini, Giulia Rizzi, Aurora Saibene, Berta Chulvi, Paolo Rosso, Alyssa Lees, and Jeffrey Sorensen. 2022. SemEval-2022 Task 5: Multimedia Automatic Misogyny Identification. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*. 533–549.

[6] EunJeong Hwang and Vered Shwartz. 2023. MemeCap: A Dataset for Captioning and Interpreting Memes. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. 1433–1445.

[7] Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020. The hateful memes challenge: detecting hate speech in multimodal memes. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*. 2611–2624.

[8] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. Improved Baselines with Visual Instruction Tuning. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 26286–26296.

[9] Arkadiy Saakyan, Shreyas Kulkarni, Tuhin Chakrabarty, and Smaranda Muresan. 2024. Understanding Figurative Meaning through Explainable Visual Entailment. *arXiv preprint arXiv:2405.01474* (2024).

[10] 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.

[11] Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, Qianyu Chen, Huarong Zhou, Zhensheng Zou, Haoye Zhang, Shengding Hu, Zhi Zheng, Jie Zhou, Jie Cai, Xu Han, Guoyang Zeng, Dahai Li, Zhiyuan Liu, and Maosong Sun. 2024. MiniCPM-V: A GPT-4V Level MLLM on Your Phone. *arXiv preprint arXiv:2408.01800* (2024).

[12] Ron Yosef, Yonatan Bitton, and Dafna Shahaf. 2023. IRFL: Image Recognition of Figurative Language. In *Findings of the Association for Computational Linguistics: EMNLP 2023*. 1044–1058.

[13] Dongyu Zhang, Minghao Zhang, Heting Zhang, Liang Yang, and Hongfei Lin. 2021. MultiMET: A Multimodal Dataset for Metaphor Understanding. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 3214–3225.