The Thirty-Ninth AAAI Conference on Artificial Intelligence (AAAI-25)

Hateful Meme Detection through Context-Sensitive Prompting and Fine-Grained Labeling (Student Abstract)

Rongxin Ouyang,[1] Kokil Jaidka,[1,2] Subhayan Mukerjee,[1,2] Guangyu Cui[2]

[1] Department of Communications and New Media, National University of Singapore
[2] Centre for Trusted Internet & Community, National University of Singapore
rongxin@u.nus.edu, jaidka@nus.edu.sg, mukerjee@nus.edu.sg, cui_gy@nus.edu.sg

Abstract

The prevalence of multi-modal content on social media complicates automated moderation strategies. This calls for an enhancement in multi-modal classification and a deeper understanding of understated meanings in images and memes. Although previous efforts have aimed at improving model performance through fine-tuning, few have explored an end-to-end optimization pipeline that accounts for modalities, prompting, labeling, and fine-tuning. In this study, we propose an end-to-end conceptual framework for model optimization in complex tasks. Experiments support the efficacy of this traditional yet novel framework, achieving the highest accuracy and AUROC. Ablation experiments demonstrate that isolated optimizations are not ineffective on their own.

Code — https://github.com/reycn/multi-modal-scale

Datasets — https://ai.meta.com/blog/hateful-memes-challenge-and-data-set/

Supplementary Information — https://osf.io/ujnrd/

Introduction

Recent years have seen a significant increase in visual content on social media (Peng, Lu, and Shen 2023; Heley, Gaysynsky, and King 2022), particularly visual misinformation (Yang, Davis, and Hindman 2023). Up to 30% of the content on platforms like X includes images or videos (Pfeffer et al. 2023), highlighting the need for a multi-modal research on social media.

However, while increasingly more studies have recognized the visual moderation challenge (González-Aguilar, Segado-Boj, and Makhortykh 2023; Solea and Sugiura 2023), most prior work has either unimodal (Muddiman, McGregor, and Stroud 2019), or focused on fine-tuning only (Lippe et al. 2020; Hermida and Santos 2023). Prior work on prompt engineering (Furniturewala et al. 2024) indicates the relative advantage of multi-stage prompts that act to pre-empt biases in Large Language Model (LLM) outputs. Yet, these studies focus on unimodal content, and it is unclear whether using multi-stage prompts suffices to improve the classification accuracy in Vision Language Model (VLM) outputs for multimodal input. In the case of Large-

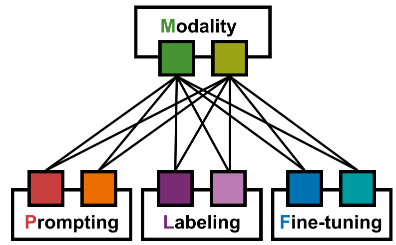


Figure 1: A Conceptual Framework

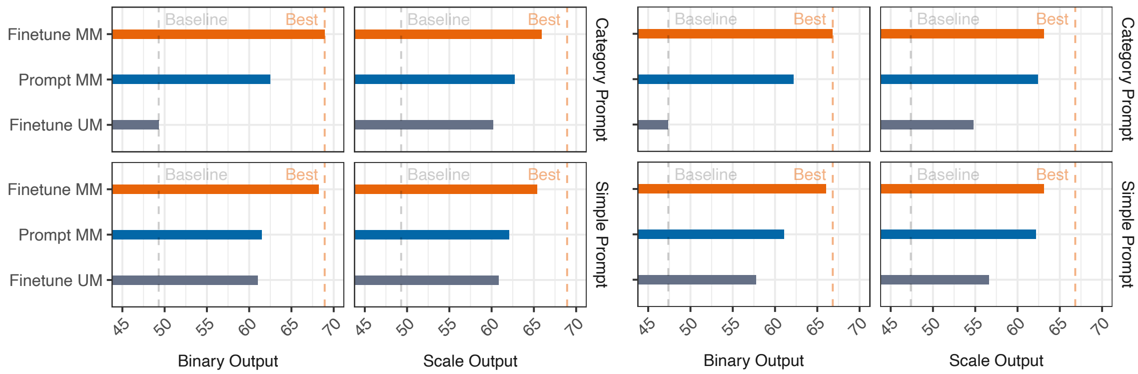
and Multimodal Language Models, it is unclear whether fine-grained categories for prompting and labeling outweigh the effect of fine-tuning. To address these research gaps, our experiment design systematically evaluates the contributions of each factor to determine whether combining them enhances performance in multimodal hate speech detection, with applications for content moderation.

Methodology

The principle underlying the proposed framework is captured by Equation 1: we consider performance (δ) as a multivariate optimization problem dependent on modalities (M), prompting (P), labeling (L), and fine-tuning (F).

$$\delta = f(M, P, L, F) \tag{1}$$

More specifically, we first compare modalities, between a visual model, InternVL 2 (Chen et al. 2023) (8B), with another text-based model, DistilBERT (Sanh et al. 2020) (66M); and expect a better performance of the multi-modal approach for more information recognized. Second, we use both prompting and fine-tuning on the same model, InternVL 8B, with identical prompting and labeling strategies. Last we compiled a 2×2 matrix of construct by both prompting (simple question or categories defined in details) and labeling (binary output or outputs in an interval scale). A simple prompt asks a plain question while categories provide detailed definitions of sub-categories of hateful content (see SI). To get labels in scales for fine-tuning, we used GPT4-o-mini to generate answers in scales and excluded incorrectly annotated cases in training according to the ground truth, ensuring the quality of the extended annotations. More details



(a) Accuracy (b) AUROC

Figure 2: Model Performance of Combinations (%, MM=multi-modal, UM=unimodal)

Category	Model	Size	Prompt	Label	Accuracy	Precision	Recall	F1-score	AUROC
Prompting	InternVL	8B	Simple$^{(a)}$	Binary$^{(c)}$	61.500	68.041*	40.408	50.704	61.086
(multi-modal)				Scale$^{(d)}$	62.100	60.491	65.306	62.807	62.163
			Category$^{(b)}$	Binary	62.500	66.571	47.143	55.197	62.199
				Scale	62.700	66.387	48.367	55.962	62.419
Fine-tuning	InternVL	8B	Simple	Binary	68.233	63.811	53.468	58.183	66.052
(multi-modal)				Scale	65.367	59.673	50.000	54.410	63.097
			Category	Binary (**M**)	68.933	64.695	54.677	59.266	66.827
				Scale	65.933	61.498	47.016	53.291	63.139
Fine-tuning	Distil-Bert	66M	Simple	Binary	61.033	53.958	39.032	45.297	57.783
(unimodal)				Scale	60.833	58.668	56.612	55.657	56.612
			Category#	Binary	49.300	38.103	36.290	37.174	47.378
				Scale	60.167	57.818	54.793	36.975	54.793

Note. * Best results are underlined. # The loss was decreasing slowly, but we maintained the same parameters for comparison.

Table 1: Experimental Results (%)

about prompts hyper-parameters are documented in Supplementary Information.

Based on combinations of those strategies, we conducted 12 experiments with a $3 \times 2 \times 2$ ablation design. These combinations of settings (see Table 1) ablate modalities (unimodal or multi-modal), prompting strategies (simple or category), labeling strategies (binary or scaled outputs) and fine-tuning process. To supplement the ground-truth of scales for fine-tuning, we used GPT-4o-mini and selected the correct ones. More details are provided in SI.

We used the Facebook Hateful Memes dataset (Kiela et al. 2020) for experiments. It includes more than 10k images, human captions and binary labels of hatefulness for training; as well as 3k entries for evaluation. Performance were evaluated by ACCU and AUROC.

Results and Future Work

As shown in Table 1, the best model is not the one with highest complexity, highlighting the necessity of our framework. Among all the models, the model **M** achieves the highest accuracy (68.933%, +19.611 p.p.) and AUROC (66.827%, +19.449 p.p.). Comparisons of ablation show that this improvement results from fine-tuning, categorical prompting, and binary labels. However, components beneficial to model **M** do not universally enhance performance (e.g., scaled outputs generally improve performance but not always), underscoring the need for the end-to-end framework we proposed.

In summary, our study introduces an end-to-end optimization pipeline for complex, multi-modal tasks like hateful meme detection. Our experiments demonstrate the effectiveness of a global optimization strategy within this framework. Moreover, our ablation studies indicate that isolated optimizations are not better by themselves (e.g., scales improve performance in most settings but not on the best model). We therefore argue that this traditional wisdom is both beneficial and necessary for such complicated, modern tasks.

Acknowledgments

This work was supported by the Singapore Ministry of Education AcRF TIER 3 Grant (MOE-MOET32022-0001). We gratefully acknowledge invaluable comments and discussions with Shaz Furniturewala and Jingwei Gao.

References

Chen, Z.; Wu, J.; Wang, W.; Su, W.; Chen, G.; Xing, S.; Zhong, M.; Zhang, Q.; Zhu, X.; Lu, L.; Li, B.; Luo, P.; Lu, T.; Qiao, Y.; and Dai, J. 2023. InternVL: Scaling up Vision Foundation Models and Aligning for Generic Visual-Linguistic Tasks. *arXiv preprint arXiv:2312.14238*.

Furniturewala, S.; Jandial, S.; Java, A.; Banerjee, P.; Shahid, S.; Bhatia, S.; and Jaidka, K. 2024. Thinking Fair and Slow: On the Efficacy of Structured Prompts for Debiasing Language Models. *arXiv preprint arXiv:2405.10431*.

González-Aguilar, J. M.; Segado-Boj, F.; and Makhortykh, M. 2023. Populist Right Parties on TikTok: Spectacularization, Personalization, and Hate Speech. *Media and communication*, 11(2): 232–240.

Heley, K.; Gaysynsky, A.; and King, A. J. 2022. Missing the bigger picture: The need for more research on visual health misinformation. *Science communication*, 44(4): 514–527.

Hermida, P. C. D. Q.; and Santos, E. M. D. 2023. Detecting Hate Speech in Memes: A Review. *Artificial Intelligence Review*, 56(11): 12833–12851.

Kiela, D.; Firooz, H.; Mohan, A.; Goswami, V.; Singh, A.; Ringshia, P.; and Testuggine, D. 2020. The hateful memes challenge: Detecting hate speech in multimodal memes. *Advances in neural information processing systems*, 33: 2611–2624.

Lippe, P.; Holla, N.; Chandra, S.; Rajamanickam, S.; Antoniou, G.; Shutova, E.; and Yannakoudakis, H. 2020. A Multimodal Framework for the Detection of Hateful Memes. arXiv:2012.12871.

Muddiman, A.; McGregor, S. C.; and Stroud, N. J. 2019. (Re) claiming our expertise: Parsing large text corpora with manually validated and organic dictionaries. *Political Communication*, 36(2): 214–226.

Peng, Y.; Lu, Y.; and Shen, C. 2023. An Agenda for Studying Credibility Perceptions of Visual Misinformation. *Political Communication*, 40(2): 225–237.

Pfeffer, J.; Matter, D.; Jaidka, K.; Varol, O.; Mashhadi, A.; Lasser, J.; Assenmacher, D.; Wu, S.; Yang, D.; Brantner, C.; et al. 2023. Just another day on Twitter: a complete 24 hours of Twitter data. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 17, 1073–1081.

Sanh, V.; Debut, L.; Chaumond, J.; and Wolf, T. 2020. DistilBERT, a Distilled Version of BERT: Smaller, Faster, Cheaper and Lighter. arXiv:1910.01108.

Solea, A. I.; and Sugiura, L. 2023. Mainstreaming the Blackpill: Understanding the Incel Community on TikTok. *European Journal on Criminal Policy and Research*, 1–26.

Yang, Y.; Davis, T.; and Hindman, M. 2023. Visual misinformation on Facebook. *Journal of Communication*, jqac051.