HMUM：面向仇恨模因检测的多阶段多模态理解模型

裴淑娟　左家莉*　何乐　万剑怡　王明文
江西师范大学　数字产业学院　江西　上饶
Email: {psj, zjl, hele, mwwang}@jxnu.edu.cn,wanjianyi@aliyun.com

摘要

随着社交媒体的广泛普及，模因(meme)已成为信息传播与舆论引导的重要载体，其中蕴含的仇恨内容对网络生态与公共安全构成威胁,尤其是通过图像暗示、文化隐喻或社会符号等方式表达的隐性仇恨模因，具有更强的隐蔽性与误导性，给仇恨模因检测任务带来显著挑战。针对上述问题，本文提出了一种仇恨模因理解模型(Hateful Meme Understanding Model，HMUM)，在Qwen2.5-VL-72B-Instruct模型基础上引入LoRA微调，并设计了一种多模态多阶段的提示学习框架。该框架通过阶段性引导模型依次完成文本识别、情绪建模与仇恨性推理，逐步增强其对模因语义与情感的理解能力，从而有效提升模型在中文语境下检测语义隐晦、情绪复杂仇恨模因的准确性。在公开数据集ToxiCN MM上的实验结果表明，HMUM(Qwen)在整体任务中取得了显著性能提升，在隐性仇恨模因子集检测方面，相较于基线模型表现出更强的优势。为评估其在更广泛隐性场景中的检测能力，本文构建了以隐性仇恨模因为主的数据集ITTD-220，实验结果显示，HMUM(Qwen)在该数据集上的检测性能同样优于现有模型，验证了其出色的泛化能力。

关键词： 多模态大语言模型；仇恨模因检测

HMUM: A Multi-Stage Multimodal Understanding Model for Hateful Meme Detection

Shujuan Pei　Jiali Zuo*　Le He　Jianyi Wan　Mingwen Wang
School of Digital Industry　Jiangxi Normal University　shangrao　Jiangxi
Email: {psj, zjl, hele, mwwang}@jxnu.edu.cn,wanjianyi@aliyun.com

Abstract

With the widespread adoption of social media, memes have emerged as a prominent medium for information dissemination and public opinion shaping. However, hateful content embedded in memes poses significant risks to online ecosystems and public safety. In particular, implicit hateful memes—conveying harmful messages through visual insinuation, cultural metaphors, or symbolic cues—exhibit high levels of semantic concealment and misdirection, making them especially challenging to detect. To address this issue, we propose the Hateful Meme Understanding Model (HMUM), a multi-stage, multimodal prompt-based framework built upon the Qwen2.5-VL-72B-Instruct model and fine-tuned using LoRA. HMUM guides the model through three semantic reasoning stages: text recognition, affective context modeling, and hatefulness inference. This progressive prompting strategy incrementally enhances the model's

第二十四届中国计算语言学大会论文集，第783页-第791页，济南，中国，2025年8月11日至14日。

(c) 2025 中国中文信息学会计算语言学专业委员会

capability to comprehend meme content and evaluate underlying emotional and ideological bias—especially in Chinese memes with implicit hateful expressions. Experiments conducted on the public ToxiCN MM dataset demonstrate that HMUM(Qwen) achieves notable performance improvements over strong baseline models, particularly on the subset of implicit hateful memes. To further assess generalization, we introduce a newly constructed dataset, ITTD-220, focused on implicit hate scenarios. Results show that HMUM(Qwen) maintains superior performance on this dataset as well, confirming its robustness and practical effectiveness in real-world multimodal hate detection tasks.

Keywords: Multimodal large language model , hateful meme detection

1 引言

模因作为一种典型的多模态传播载体，近年来在社交媒体平台中广泛应用，已成为信息传播和情感表达的重要形式 (Eftekhar, 2024)。模因通常由图像与简洁文本构成，其视觉素材多源自于名人肖像、影视画面、漫画或流行语录，通过图文协同机制传达特定情感或观点 (Ma, 2023)，其中积极的模因能有效促进文化传播，而仇恨模因则可能激化社会矛盾、煽动群体对立，甚至引发系统性风险。因此，仇恨模因的检测已成为学科研究的重要课题。

仇恨模因通过针对特定社会群体、个人、组织或社区传递贬低、歧视或敌意信息，从而对社会整体的和谐与团结构成威胁 (Lu et al., 2024)，模因检测任务的核心在于判定图文组合是否包含仇恨内容 (Wang, 2024)。值得关注的是，近年来隐性仇恨模因(Implicitly Hateful Meme)的检测任务逐渐得到重视，这类模因可能缺乏明显的仇恨文本内容或表面上不具备直接攻击性，但通过特定的文化背景、语境暗示或象征手法，依然可能传递具有伤害性的隐含意义 (Adak et al., 2025)。传统模因检测模型主要依赖显性文本特征，难以有效通过识别讽刺手法、象征符号或文化隐喻等形式掩饰的隐性仇恨内容。因此，针对隐性仇恨模因的检测方法需要深入研究和完善 (Adak et al., 2025)。

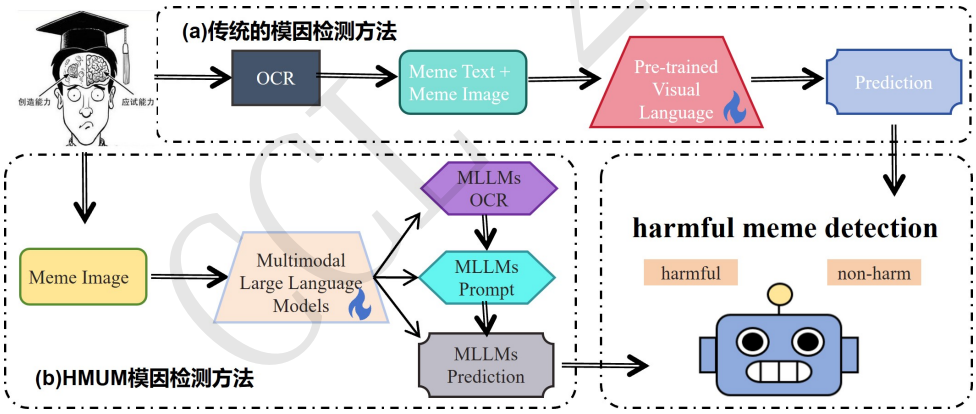

Figure 1: (a)传统模因检测方法与(b)HMUM方法之间的比较示意图

现有研究主要聚焦于英文仇恨模因检测，中文仇恨模因检测研究相对较少。Lu (2024)发布首个中文仇恨模因数据集ToxiCN MM，涵盖丧文化、一般冒犯、性暗示等多种仇恨模因类型，共12,000条模因数据。尽管该数据集填补了中文模因领域的空白，但其中隐性模因样本数据较少，难以评估模型在复杂语境下的检测能力。为弥补隐性仇恨模因数据量的不足，本文构建并标注了一个共计220条样本、以隐性仇恨模因为主的数据集，用于评估模型对中文隐性仇恨模因的检测能力。如图1(a)所示，传统仇恨模因检测方法通常使用双流架构与特征融合策略进行分类检测任务 (Huang, 2025)，以ViLBERT (Lu, 2019)和LXMERT (Tan, 2019)为代表的

©2025 中国计算语言学大会
根据《Creative Commons Attribution 4.0 International License》许可出版
基金：国家自然科学基金(62466027,62266023)
通讯作者：左家莉

典型双流架构模型，首先通过独立编码器处理各模态信息，再经交叉Transformer结构 (Ashish Vaswani et al., 2017)进行特征融合，这种架构的模型难以有效捕捉隐性仇恨内容。图1(b)展示的HMUM框架模型以Qwen2.5-VL-72B-Instruct为基础，通过LoRA微调技术与AdamW优化器 (Loshchilov, 2019)实现高效的参数更新。在此基础上，HMUM模型设计了“文本识别—情绪建模—仇恨性推理”的三阶段提示学习框架：(i)使用模型OCR提取图像中的嵌入文字(ii)进行情绪语义建模，补充模因的情感语境；(iii) 综合多模态语义与情绪信息完成仇恨性推理。

本研究的主要贡献如下：

1. 提出了结合LoRA与多阶段Prompt框架的模型HMUM，在ToxiCN MM数据集上取得了当前最优的检测结果，实现了中文仇恨模因的高效检测；

2. 为弥补ToxiCN MM数据集中隐性仇恨模因样本不足，本文构建并标注了ITTD-220数据集，为中文隐性仇恨模因检测研究提供了新的数据支持。

2 相关工作

2.1 仇恨模因检测

近年来，多模态仇恨模因检测研究取得了显著进展。Facebook AI发起的Hateful Memes Challenge (Kiela et al., 2020) 极大地推动了该领域的发展，催生了多种创新方法。早期研究主要使用特征融合策略，例如Lippe(2020)提出的的双流架构和Cao (2020)设计的外部知识增强方法。随着研究的深入，Pramanick (2021)构建的MOMENTA框架通过全局-局部双视角分析显著提升了检测性能，而Sharma (2022)设计的的DISARM模型创新性地整合了语义理解技术，在定向攻击检测方面表现突出。当前，基于预训练模型的方法展现出显著优势，Cao (2022)推出的PromptHate通过创新的提示学习策略取得突破，后续的工作进一步拓展了跨模态理解能力。随着大规模多模态预训练模型的发展，使用多模态大语言模型(MLLMS)进行仇恨模因检测逐渐成为研究热点 (Li et al., 2023; Yang et al., 2023)。

2.2 多模态大语言模型与高效微调

随着多模态大语言模型（Multimodal Large Language Models, MLLMs）的快速发展，主流模型如GPT-4 (Lyu et al., 2023)、LLaVA (Liu et al., 2023)和LLaMA (**?**)在跨模态理解任务中展现出卓越性能。面向中文场景的开源模型也取得显著进展，包括GLM (Zeng et al., 2024)、DeepSeek (DeepSeek, 2024)和Qwen (Bai et al., 2025)等，这些模型均具备图文建模和复杂语义解析能力。在此技术背景下，研究者开始探索MLLMs在仇恨模因检测中的应用。最新工作表明，通过参数高效微调方法可有效提升模型对文化特定语义的理解。例如，Cao (2024)提出的适配框架创新性地使用低秩自适应技术，通过构建任务导向的适配模块，使MLLMs在小样本场景中保持优异的领域泛化能力。该工作为平衡模型通用表征学习与领域特定需求提供了有效解决方案。

3 方法

3.1 问题定义

本研究聚焦于中文多模态模因的仇恨性检测任务。由图像I和与之关联的文本T共同构成模因样本可记作$M = (I, T)$。检测任务旨在判断该模因是否包含仇恨内容，我们将该问题形式化为一个二分类任务。

$$\hat{y}_j = f(\{x_{j^i}^I, x_{j^t}^T\}), \quad \hat{y}_j \in \{0, 1\} \tag{1}$$

其中，X_j^I和X_j^T分别表示第j个模因的图像模态和文本模态输入，$\hat{y}_j$为模型预测结果，若$\hat{y}_j = 1$，则表示该模因为仇恨模因；若$\hat{y}_j = 0$，则表示其为无害模因。

3.2 基于提示的三阶段模因理解框架

为提升多模态大语言模型在中文仇恨模因检测任务中的表现，本文设计了一种基于提示学习的三阶段模因理解框架。该框架整体流程如图 2所示，围绕图像与文本之间的语义交互关系，分阶段引导模型依次完成文本识别、情绪建模与仇恨性推理，从而逐步构建对模因语义的层次化理解，并输出最终的分类判断。

(c) 2025 中国中文信息学会计算语言学专业委员会

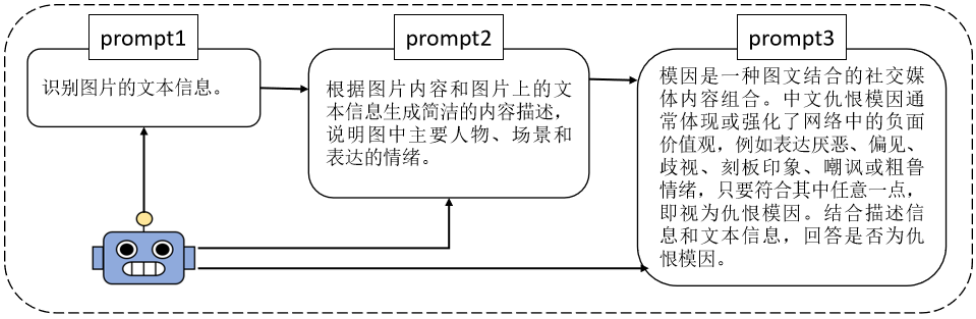

Figure 2: 三阶段模因理解流程示意图

阶段一：嵌入文本识别。本阶段利用多模态大语言模型对模因图像中的文本信息进行提取，得到文本表示：

$$T_{ocr} = \text{MLLM}_{ocr}(\text{meme}_j, \text{OCR}_{prompt}) \tag{2}$$

阶段二：情绪与语义建模。本阶段旨在获取模因所蕴含的情绪特征，以增强模型对潜在语义倾向的理解能力。通过设计的提示指令，引导多模态大语言模型生成简洁的描述，该描述涵盖图像中的核心人物、场景要素以及其所表达的情绪倾向，作为后续推理的情感语境基础。该阶段的输出记为：

$$D_{img} = \text{MLLM}_{dis}(\text{meme}_j, T_{ocr}, \text{DIS}_{prompt}) \tag{3}$$

阶段三：仇恨性推理与分类判断。该阶段综合模因图像I、嵌入文本T_{ocr} 与情绪语义描述D_{img}，通过多模态语义联合推理，判断模因是否包含仇恨内容，完成最终的分类预测：

$$\hat{y}_i = \text{MLLM}_{detect}(\text{meme}_j, T_{ocr}, D_{img}, X_{prompt}) \tag{4}$$

4 实验

4.1 数据集分析

实验使用ToxiCN MM模因数据集，该数据集中每条样本均被标注了是否为仇恨模因。在完成对数据集整体的仇恨模因检测任务之后，进一步关注具有隐含仇恨倾向的模因类型，从ToxiCN MM测试集中筛选出图像与文本在单独分析时均不构成仇恨内容，但在图文结合后被标注为仇恨的模因数据，这类数据往往以图像暗示、双关语、文化隐喻、社会符号等方式表达仇恨。如表1所示，ToxiCN MM子集中模因的仇恨内容通常通过图像与文本之间复杂的语义映射与隐含关联进行表达，具有较强的语义隐蔽性，给检测任务带来了更大的挑战。

本文从百度贴吧、新浪微博等社交媒体平台采集并构建了以隐性仇恨模因为主的ITTD-220数据集。在标注过程中，由两名具备相关背景的研究生对每条模因数据分别进行标注，标注类别包括“隐性有害”、“显性有害”与“无害”，仅保留两位标注者意见一致的样本，最终构建了由220条模因样本组成的数据集。该数据集主要涵盖中文语境中的教育歧视、性别歧视和地域歧视等类型的仇恨模因，而ToxiCN MM数据集中此类模因的占比较低。ITTD-220数据集旨在评估所提出模型HMUM在处理多样化及隐性仇恨表达场景中的泛化能力。

数据集	split	无害	显性仇恨	隐性仇恨	总计
ToxiCN MM	train	6538	1984	1078	9600
	test	1635	595	270	2400
ITTD-220	test	14	16	190	220

Table 1: 数据集分布

4.2 实验设置

通过系统性对比实验，评估所提方法HMUM在中文仇恨模因检测任务中的效果。对比模型涵盖当前主流的单模态与多模态模型，包括文本编码模型RoBERTa (Cui et al.,

(c) 2025 中国中文信息学会计算语言学专业委员会

2019)，图像视觉模型ViT (Dosovitskiy et al., 2021)，以及图文联合预训练模型CLIP (Radford et al., 2021)。此外，本文选取了近期表现突出的多模态大语言模型Qwen (Bai et al., 2025)[1]、DeepSeek (DeepSeek, 2024)[2]和GLM (Zeng et al., 2024)[3]，并在与HMUM框架一致的LoRA参数配置下对上述模型进行微调。在本实验中，设定LoRA的秩$r = 8$，学习率为$1e{-}4$，优化器使用AdamW (Loshchilov, 2019)。表2给出了所有对比模型的具体版本信息。在评估指标方面，本文使用Precision(P)、Recall(R)与F1值作为主要性能评价标准，从多角度衡量各模型在仇恨模因检测任务中的检测能力与鲁棒性。

模型	版本
RoBERTa	chinese-roberta-wwm-ext-base
ViT	vit-base-patch16-224
CLIP	chinese-clip-vit-base-patch16
Qwen	Qwen2.5-VL-72B-Instruct
DeepSeek	deepseek-vl-7b-chat
GLM	glm-4v-9b-chat

Table 2: 模型的具体版本

4.3 ToxiCN MM总体结果

表3展示了各类模型在ToxiCN MM测试集上的仇恨模因检测性能，其中纯文本模型(如RoBERTa)与图像模型(如ViT)在该任务中的表现相对较低，F1值分别为76.7%和67.6%。这一结果表明，文本在模因中往往承担更多与仇恨性判断相关的语义表达。在多模态模因检测任务中，CLIP、MKE以及ViT-RoBERTa等典型的图文融合模型整体表现稳定，F1值主要分布在78-80%区间，体现出较强的跨模态协同建模能力。

为评估多模态大语言模型在仇恨模因检测任务中的理解能力，选取GLM、DeepSeek和Qwen三个主流模型进行对比实验。结果显示，三者在Precision、Recall 及F1 等指标上整体优于其他对比模型。值得注意的是，GLM在性能上略优于Qwen，这一差异可能与二者模型规模的差异有关。较大规模的模型通常需要更多的训练轮数（epoch）以实现充分收敛，而在相同训练设置下，Qwen 的性能可能因此未能完全释放。

模态	模型	模型大小	精确率	召回率	F1 值	F1 值(har)
Text	RoBERTa	<1B	78.2	75.9	76.7	69.2
Image	VIT	<1B	67.1	68.3	67.6	54.1
Multimodal	CLIP	<1B	77.4	79.8	78.4	69.3
	MKE	<1B	80.8	80.2	80.5	73.3
	VIT-Roberta	<1B	78.2	75.9	76.7	69.2
	GLM	9B	82.3	81.2	81.7	74.3
	Deepseek	7B	75.7	74.9	75.3	65.5
	Qwen	72B	82.1	80.6	81.2	73.6
	HMUM(GLM)	9B	84.2	82.0	80.1	74.5
	HMUM(Qwen)	72B	**88.4**	**86.0**	**87.1**	**81.7**

Table 3: 各模型在仇恨模因检测任务中的性能表现，加粗表示最优值

[1]https://huggingface.co/Qwen/Qwen2.5-VL-72B-Instruct
[2]https://www.modelscope.cn/models/deepseek-ai/deepseek-vl-7b-chat
[3]https://modelscope.cn/models/ZhipuAI/glm-4v-9b

(c) 2025 中国中文信息学会计算语言学专业委员会

为验证多阶段提示框架的有效性，本文在多个主流多模态大语言模型中引入该框架。在实验中，我们选取微调效果最优的GLM作为基础模型，在引入多阶段提示框架后，HMUM(GLM)在P、R、F1和F1(har)指标上分别达到84.2%、82.0%、80.1%和74.5%。尽管F1值略有下降，其余指标均有一定提升，但整体提升幅度有限。相比之下，基于Qwen构建的HMUM(Qwen)在所有评估指标上均取得最优结果，F1(har)提升至81.7%。这一提升得益于三阶段提示框架对Qwen模型先验知识的有效调动，增强了其对仇恨模因语义与情感的理解能力。相较于对比模型MKE(F1(har) = 73.30%)，HMUM(Qwen)在仇恨模因检测任务中的性能提升明显，F1(har) 提高了8.4%。因此，本文最终选择Qwen2.5-VL-72B-Instruct作为HMUM的基础模型。

模型变体	精确率	召回率	F1 值	F1 值(har)
HMUM(Qwen)	**88.4**	**86.0**	**87.1**	**81.7**
w/o dis	82.5	84.1	83.2	77.4
w/o ocr	84.1	81.3	82.5	75.1
w/o def	82.8	84.5	83.6	77.8
w/o dis+def	83.9	79.6	81.69	73.8
w/o ocr+def	82.1	80.5	81.3	74.0
w/o dis+ocr	83.7	81.3	82.3	75.0
w/o dis+ocr+def	82.1	80.6	81.2	73.6

Table 4: 消融实验结果，加粗表示最优值

在ToxiCN MM测试集上开展的消融实验系统评估了HMUM(Qwen)各功能模块对整体性能的贡献，实验结果如表4所示。移除情绪建模模块(w/o dis)后，模型的F1(har)值下降至77.4%；去除仇恨性推理模块(w/o def)同样导致明显性能下降，F1(har)值降低至77.8%；图文识别模块的取消(w/o ocr)也带来明显性能退化，F1(har)降至75.1%。在多模块联合去除的情形下，性能退化更为显著。在三模块全部去除的情形下(w/o dis+ocr+def)，模型性能降至最低，F1值与F1(har) 值分别降为81.2%与73.6%，与完整模型相比，分别下降5.9%和8.1%。总体来看，任一模块的去除均对性能造成不利影响，特别是多模块协同缺失的情况下，检测效果显著下降，说明各模块在仇恨模因检测中协同发挥重要作用，共同支撑模型对多模态语义的准确判断。

4.4 隐性模因检测能力评估

在ToxiCN MM数据集的隐性仇恨子集上，HMUM(Qwen)模型展现出较强的检测能力。实验结果如表5所示，HMUM(Qwen)模型在该子集上取得了$R = 70.6\%$和$F_1 = 82.8\%$的检测结果，整体性能优于其他对比模型。相较于多模态大语言模型(如GLM、deepseek)，HMUM(Qwen)凭借其三阶段提示策略，在建模语义模糊与表达隐晦的模因内容方面表现出更强的理解能力。

模型	R (har)	F1 (har)
CLIP	53.0	69.2
MKE	61.5	76.1
vit-roberta	37.8	54.8
GLM	59.6	74.7
DeepSeek	50.4	67.0
Qwen	58.5	73.8
HMUM(Qwen)	**70.6**	**82.8**

Table 5: ToxiCN MM隐性数据子集评估结果

HMUM(Qwen)在自构建的ITTD-220数据集上进行了补充评估，结果如表6所示，模型在该

(c) 2025 中国中文信息学会计算语言学专业委员会

数据集上同样超过了其余基线模型。HMUM(Qwen)在两个来源不同的模因测试集中均保持性能领先，说明模型在复杂语境下处理仇恨模因的稳健性与跨域泛化能力。

模型	召回率(har)	F1 (har)
MKE	64.8	77.7
GLM	63.7	76.5
Qwen	70.6	81.6
HMUM(Qwen)	**73.5**	**83.6**

Table 6: ITTD-220评估结果

4.5 模型推理能力与误判分析

表7展示了MKE与HMUM(Qwen)在多个典型隐性仇恨模因样本上的分类结果对比，用以分析两种模型的性能差异。MKE在处理带有刻板印象、语义反讽或图文错配的模因时，易受到文本表层“非攻击性”语义的干扰，导致其未能识别潜藏的歧视意图。尤其是在隐性模因中，攻击性往往以图像暗示、双关语、文化隐喻、社会符号等表达方式呈现，MKE缺乏语境整合与多层次语义建模能力，易产生误判。相比之下，HMUM(Qwen)能够更有效地捕捉图像与文本间的深层语义联系。

模因组件	示例1	示例2	示例3
图像			
文本	18 岁的我，51 岁的Jennifer Aniston	错误的搭讪方式、害。正确的搭讪方式、哇	表面，内心
真实标签	`hate`	`hate`	`hate`
MKE	`non-hate`	`non-hate`	`non-hate`
HMUM(Qwen)	`hate`	`hate`	`hate`

Table 7: MKE与HMUM(Qwen)在隐性仇恨模因样本上的分类结果

图3展示了HMUM(Qwen)在处理仇恨模因任务中的三阶段提示推理过程，并以一则涉及招聘性别歧视的模因示例具体说明模型检测流程。第一阶段中，模型依据提示提取图像中嵌入文本（如“男”“女”“公开招聘”），实现OCR文本识别，获取图文组合中的关键信息。第二阶段通过生成图像主旨与情绪场景描述，明确指出图像反映了性别歧视语境及其所蕴含的社会批判立场，为后续判断提供情感语境支持。最终阶段，模型结合“模因定义”提示，依据前两阶段获取的图文语义与情绪信息完成语境推理，识别出该模因通过图文协同构成对特定群体的负面刻板印象表达，从而判断是否为仇恨模因。整个流程通过多阶段提示机制，在逐层抽象与引导中实现多模态语义的深度整合，有效增强了模型在检测中文模因中隐性仇恨内容时的理解能力与判别鲁棒性。

(c) 2025 中国中文信息学会计算语言学专业委员会

Figure 3: HMUM的三阶段prompt流程

5 总结

本文针对中文仇恨模因检测这一任务，提出了一种基于多模态多阶段提示学习的大语言模型方法HMUM。实验结果表明，所提出模型在公开数据集与自建数据集上均取得了优异的检测结果，尤其在隐性模因检测方面相较于现有方法展现出更高的准确性。进一步的消融实验与案例分析验证了多阶段提示在提升模型语义推理能力方面的显著成效。综上所述，本文方法在中文模因检测场景中具有良好的应用潜力，并为多模态语义理解任务提供了可行的技术路径。未来工作将致力于扩展高质量隐性仇恨模因数据的规模，并深入探索Qwen模型的参数微调策略，以优化HMUM模型的性能效果，提升其泛化能力与实际应用价值。

参考文献

Adak S., Banerjee S., Mandal R., et al. 2025. *MemeSense: An Adaptive In-Context Framework for Social Commonsense Driven Meme Moderation*. arXiv:2502.11246.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. *Attention is All You Need*. In Advances in Neural Information Processing Systems (NeurIPS 2017), pages 5998–6008.

Bai S., Chen K., Liu X., et al. 2025. *Qwen2.5-VL Technical Report*. arXiv:2502.13923.

Cao J., Gan Z., Cheng Y., et al. 2020. *Behind the scenes: Revealing the secrets of pre-trained vision-and-language models*. In ECCV 2020. Springer: 565–580.

Cao R., Lee R. K. W., Chong W. H., et al. 2022. *Prompting for Multimodal Hateful Meme Classification*. In EMNLP 2022: 321–332.

Cao R., Hee M. S., Kuek A., et al. 2023. *Pro-Cap: Leveraging a frozen vision-language model for hateful meme detection*. In Proceedings of the 31st ACM International Conference on Multimedia. ACM: 5244–5252.

Cao R., Lee R. K. W., Jiang J. 2024. *Modularized Networks for Few-shot Hateful Meme Detection*. arXiv:2402.11845.

Cui Y., Che W., Liu T., et al. 2019. *Pre-Training with Whole Word Masking for Chinese BERT*. arXiv:1906.08101.

DeepSeek-VL Team. 2024. *DeepSeek-VL: Towards Real-World Vision-Language Understanding*. arXiv:2405.04147.

Dosovitskiy A., Beyer L., Kolesnikov A., et al. 2021. *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*. In ICLR 2021.

(c) 2025 中国中文信息学会计算语言学专业委员会

Eftekhar Hossain, Omar Sharif, Mohammed Moshiul Hoque, and Sarah Masud Preum. 2024. *Deciphering Hate: Identifying Hateful Memes and Their Targets*. In Proceedings of ACL 2024, pages 8347–8359.

Hu E., Shen Y., Wallis P., et al. 2021. *LoRA: Low-rank adaptation of large language models*. arXiv:2106.09685.

Huang J., Pan J., Wan Z., Lyu H., Luo J. 2025. *Evolver: Chain-of-Evolution Prompting to Boost Large Multimodal Models for Hateful Meme Detection*. In COLING 2025: 7321–7330.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, et al. 2023. *LLaMA: Open and Efficient Foundation Language Models*. arXiv preprint arXiv:2302.13971.

Haotian Liu, Chunyuan Zhang, Yinan Xu, et al. 2023. *LLaVA: Visual Instruction Tuning*. arXiv preprint arXiv:2304.08485.

Kiela D., Firooz H., Mohan A., et al. 2020. *The hateful memes challenge: Detecting hate speech in multimodal memes*. arXiv:2005.04790.

Li C., Gan Z., Yang Z., Yang J., Li L., Wang L., Gao J. 2023. *Multimodal foundation models: From specialists to general-purpose assistants*. arXiv:2309.10020.

Lippe P., Holla N., Chandra S., et al. 2020. *A multimodal framework for the detection of hateful memes*. arXiv:2012.12871.

Loshchilov I., Hutter F. 2019. *Decoupled Weight Decay Regularization*. In ICLR 2019.

Lu J., Batra D., Parikh D., Lee S. 2019. *ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks*. In NeurIPS 2019.

Lu J., Xu B., Zhang X., et al. 2024. *Towards comprehensive detection of Chinese hateful memes*. arXiv:2410.02378.

Lyu H., Huang J., Zhang D., Yu Y., Mou X., Pan J., Yang Z., Wei Z., Luo J. 2023. *GPT-4v(ision) as a social media analysis engine*. arXiv:2311.07547.

马志羽. 2023. 基于多任务学习的多模态仇恨模因检测. 云南大学.

Pramanick S., Sharma S., Dimitrov D., et al. 2021. *MOMENTA: A multimodal framework for detecting hateful memes and their targets*. In Findings of EMNLP 2021. ACL: 4439–4455.

Radford A., Kim J. W., Hallacy C., et al. 2021. *Learning Transferable Visual Models From Natural Language Supervision*. arXiv:2103.00020.

Sharma S., Akhtar M. S., Nakov P., Chakraborty T. 2022. *DISARM: Detecting the victims targeted by hateful memes*. In Findings of ACL: NAACL 2022. ACL: 1572–1588.

Tan H., Bansal M. 2019. *LxmerT: Learning cross-modality encoder representations from transformers. Chinese Chemical Letters*, 30(11): 2255–2258.

王文智. 2024. 多模态中文仇恨言论检测研究. 广西大学. DOI:10.27034/d.cnki.ggxiu.2024.001381.

Yang Z., Li L., Lin K., Wang J., Lin C.-C., Liu Z., Wang L. 2023. *The dawn of LMMs: Preliminary explorations with GPT-4v(ision)*. arXiv:2309.17421.

Zeng A., Yang X., Liu Z., et al. 2024. *ChatGLM: A Family of Large Language Models from GLM-130B to GLM-4 All Tools*. arXiv:2403.07830.

CCL 2025

(c) 2025 中国中文信息学会计算语言学专业委员会