计算机研究与发展
Journal of Computer Research and Development
DOI: 10.7544/issn1000-1239.202330960
61(5): 1206−1217, 2024

基于多模态大语言模型的攻击性模因解释生成方法

林　萌[1,2]　戴程威[1,2]　郭　涛[1]
[1]（中国科学院信息工程研究所　北京　100019）
[2]（中国科学院大学网络空间安全学院　北京　100049）
（linmeng@iie.ac.cn）

A Method for Generating Explanations of Offensive Memes Based on Multimodal Large Language Models

Lin Meng[1,2], Dai Chengwei[1,2], and Guo Tao[1]
[1]（*Institute of Information Engineering*, *Chinese Academy of Sciences*, *Beijing* 100019）
[2]（*School of Cyber Security*, *University of Chinese Academy of Sciences*, *Beijing* 100049）

Abstract　With the advancement of 5G technology, offensive speech has increasingly proliferated across social networks in the form of multimodal memes. Consequently, the detection and interpretive generation of offensive memes play a crucial role in enhancing content moderation effectiveness and maintaining a harmonious and healthy public discourse environment. Existing studies on the interpretive generation of offensive memes focus solely on the targets and content of offense, neglecting the societal background knowledge and metaphorical expressions embedded in memes. This oversight limits the ability to comprehensively and accurately interpret the meaning of offensive memes, thus constraining the applicability of interpretations. To address this challenge, we propose a method based on multimodal large language model for generating interpretations of offensive memes. By augmenting elements such as offense targets, the content of the offense, and metaphor recognition into the instruction tuning data, we can effectively improve the multimodal large model's proficiency in interpretively generating offensive memes through instruction tuning. The experimental outcomes validate three key strengths of our method: first, it achieves a notable 19% enhancement in the BERTScore evaluation metric over baseline models; second, it incorporates comprehensive background knowledge pertinent to offensive metaphorical expressions within its interpretations; third, it exhibits strong generalization capabilities when handling previously unseen meme data.

Key words　offensive meme; explanation generation; multi-modal LLM; data augment; instruction fine-tuning

摘　要　随着5G的发展，攻击性言论逐渐以多模态的方式在社交网络上广泛传播.因此，攻击性模因的检测与解释生成对于提高内容审核效果、维护和谐健康的舆论场环境有着重要的作用.现有的攻击性模因解释生成研究只关注于攻击对象和攻击内容，忽略了模因包含的社会背景知识和隐喻表达手法，无法全面、准确地解释攻击性模因的含义，大大限制了解释的应用范围.为了应对这一挑战，提出一种基于多模态大模型的攻击性模因解释生成方法，通过增强攻击目标、攻击内容和隐喻识别等多种指令数据，利用其微调多模态大模型，以提升大模型对攻击性模因的解释生成能力.实验结果证实，该方法生成的解释具有3点优势：一是相比基线模型在BERTScore评估指标上提高了19%；二是解释中包含了攻击性隐喻表达的相关背景知识；三是在处理未见的模因数据时也表现出良好的泛化性能.

关键词　攻击性模因；解释生成；多模态大语言模型；数据增强；指令微调

收稿日期：2023-11-30；**修回日期**：2024-03-12

通信作者：郭涛（guotao@iie.ac.cn）

中图法分类号 TP183

多模态数据由于其丰富的表现形式，已经成为社交媒体上主要的信息传播载体. 其中由图像与文本结合而成的模因数据，逐渐成为社交媒体上散布攻击性言论的主要手段之一. 攻击性模因往往包含歧视和暴力倾向等负面内容，可能会引发虚假信息的传播[1]，损害社会公众的认知和判断能力，助长网络欺凌的滋生，进一步加剧社会矛盾. 面对海量的模因数据，人工审核方式给社交平台带来了沉重的经济负担，亟需通过多模态内容理解技术，实现对攻击性模因的自动识别. 2020 年，Facebook 举办的多模态仇恨模因检测[2]的比赛，引发了对攻击性模因检测的研究高潮. 研究者从多个角度提出了不同的识别方法，并取得了良好的识别效果[3-4].

虽然仇恨言论识别的研究取得了较大的进展，但关于攻击性模因解释生成的研究目前仍处于起步阶段. 生成攻击性模因的解释有助于内容审核员更深入地理解模因的仇恨内涵，从而制定有效的审核策略. 同时详尽的解释也有助于提高分类模型对于潜在语义的理解，提升识别效果. 生成的解释还有助于减少由于文化偏差引发的用户无意传播，引导用户更加谨慎地参与社交媒体，进一步减少恶意内容的传播.

近年来，研究人员逐步认识到攻击性模因解释生成任务的重要性. Hee 等人[5]在 2020 年仇恨模因分类数据集上，采用“贬低性动词+攻击目标+攻击内容”的模板方式，人工标注生成了简明扼要的仇恨模因解释，公布了目前第一个也是唯一一个开源的攻击性模因解释生成数据集 HatReD，为攻击性模因解释生成研究提供了评测基准. 之后，Sharma 等人[6]提出为攻击性模因中的语义角色生成解释的方法，该方法首先识别模因中的语义角色，再对其中的每一个语义角色生成模板化的简明解释.

已有研究生成的解释主要聚焦于攻击目标和攻击内容，而忽略了对攻击性模因隐喻表达方式和社会背景知识的解释，没有相关背景知识的人员理解其含义依然有一定门槛. Scott[7]的研究证明隐喻是模因尤其是攻击性模因的主要表达方式之一. 因此，识别攻击性模因中的隐喻表达手法和相关联的社会背景知识，对于生成完整准确、易于理解的解释至关重要. 如图 1 所示，只有结合德国纳粹在二战期间屠杀犹太人等背景知识，识别出“烹饪”隐喻了“屠杀”行为，才能正确理解这个模因所表达的深刻含义以及展现出的极端攻击性.

为了生成全面翔实的攻击性模因解释，本文构建了一个面向模因解释生成的指令微调数据集，并提出了一个基于多模态大语言模型的攻击性模因解释生成方法. 首先，基于已有的 6 个公开数据集（3 个隐喻模因数据集和 3 个攻击性模因数据集），利用 GPT-3.5 丰富的常识知识与强大的语义理解能力，为模因数据生成相应的解释. 其次，根据增强后的数据，构建一个多模态指令微调数据集. 该指令既包括了攻击性模因的攻击性属性，比如分类标签、攻击对象、攻击内容等，也包含了隐喻相关属性，比如隐喻分类、隐喻标签和相关的社会背景知识等. 最后，利用构建的指令微调数据集对多模态大语言模型进行微调. 通过这 3 步，从而使本文方法能够深入学习模因中的隐喻宣传技巧，同时精确匹配视觉和文本中的仇恨特征，进而有效地生成攻击性模因全面且准确的解释.

综上，本文的贡献主要在于 3 个方面：

1）基于 GPT-3.5 构建了一组模因解释生成指令数据集，该数据集包含 3 个攻击性模因数据集和 3 个隐喻模因数据集，为攻击性模因解释生成研究提供了评测数据集和评测基准.

2）提出了一种针对攻击性模因解释生成任务的多模态大语言模型训练方法，该方法不仅学习到模因固有的隐喻表达方式，也对齐跨模态的视觉和文



HatReD人工标注答案：
Disrespects the jews by mocking the killing of jews in gas chambers during the holocaust.
本方法生成的解释：
【隐喻解释】The metaphor used in this meme is that of cooking as a metaphor for the Holocaust.
【社会背景知识】Joking about the smell of cooking as a metaphor for the Holocaust is a form of minimization and trivialization of the atrocities committed by the Nazis during the Holocaust.
The use of this metaphor is offensive and insensitive, and it is important to be mindful of the impact of such jokes on survivors and their families.

Fig. 1 Example of interpretive generation of offensive memes

图 1 攻击性模因解释生成示例

本攻击性特征，从而生成一个全面准确的攻击性模因解释.

3)在 HatReD 数据集上进行实验，实验结果表明本文方法在 BERT-Score-F 评测指标上较已有方法有 19% 的提升，并具备较强的隐喻表达识别能力与背景知识关联能力. 另外，本文模型在处理新的模因数据时同样表现出良好的泛化能力.

1 相关工作

1.1 攻击性模因的识别问题

随着移动互联网的快速发展，攻击言论呈现出多样化趋势，特别是多模态形式的攻击言论日渐增多，使得其检测工作成为研究领域的新热点. Kiela 等人[2] 发布了一个精心设计的多模态仇恨模因数据集，该数据集结合了图像和文本 2 种模态，并特意加入了一些难以辨识的样本(单一模态下看似正常，但跨模态结合后则成为仇恨模因)，这大大增加了检测的难度. 基于 Kiela 的工作，部分研究者[8-9] 通过微调多模态预训练模型来增强特征表示能力，以提升有害模因检测的效果. 还有研究通过数据增强[10]、解耦表征[11] 和模型融合[12] 等方法来增强多模态模型在仇恨言论检测方面的能力

近年来，预训练大语言模型领域取得了重大进展，同时视觉基础模型在感知能力上也有显著的进步. 基于对这些模型互补优势的认识，应用预训练大语言模型解决多模态任务成为一种新的研究方向[13]. Gupta 等人[14] 利用 GPT-3 进行视觉推理任务，通过大语言模型的上下文学习能力生成类似 Python 的模块化程序，并执行这些程序以获得结果. Shao 等人[15] 探索了如何引导 GPT-3 进行多模态视觉问答的问题，他们首先在 VQA 数据集上训练 VQA 模型，生成 2 种类型的回答启发式信息，再将这些信息与问题和图像描述整合成格式化的提示，指导 GPT-3 预测答案. 在攻击性模因识别领域，Cao 等人[16] 构建了简单的提示并结合上下文示例，利用预训练 RoBERTa[17] 语言模型的隐式知识进行仇恨模因分类. Ji 等人[18] 指出，在传统的预训练模型中，同时使用文本和图像作为输入可能会在特征空间造成显著的语义差异，这样的差异可能对模型的性能造成不利影响. 鉴于此，他们提出首先生成视觉数据的描述性文本和关键属性信息，以此将多模态数据转化为纯文本模式. 随后，这些文本数据被融合进基于提示的预训练语言模型中，从而有效实现对多模态有害模因的检测. 这些研究主要聚焦于提升攻击性模因检测性能，并未涉足攻击性模因的解释生成研究.

1.2 多模态模因的可解释性研究

由于模因通常包含大量隐喻的修辞手法，当前的图像描述生成方法主要致力于准确描述图像场景，而无法解释其蕴含的深层含义. 随着对多模态模因研究的深入，研究者们进一步探索模因的生成背景和传播环境，提出了模因的解释生成研究领域.

Hee 等人[5] 构建了一个名为 HatReD 的多模态仇恨模因解释生成数据集. 该数据集经过人工精心标注，旨在深化对仇恨模因背后含义的理解. 此外，他们利用先进的预训练语言模型对这一任务进行了基准性能评估，发现即便是当前领先的多模态预训练模型，在生成仇恨模因的解释方面，与人类专家相比也存在显著差距. 特别是在面对新领域的模因时，生成的解释也常常缺乏准确性和相关性. 这项研究为未来攻击性模因解释生成的领域研究提供了一个重要的基准和研究起点. Hwang 等人[19] 构建了一个名为 MemeCap 的数据集. 该数据集的每个模因都标注了模因的文字信息、图像描述、隐喻表达以及人工生成的模因解释. Hwang 等人使用了最新的多模态模型对模因解释生成任务进行了广泛的实验. 实验结果表明，现有的模型在理解视觉隐喻方面仍然存在困难，并且性能远远低于人类水平. 模因的解释生成，尤其是攻击性模因的解释生成目前仍处于研究起步阶段，随着大模型的发展，该领域也逐渐受到更多关注.

1.3 多模态大语言模型

多模态大语言模型是指利用大语言模型强大的推理能力，通过调整模型结构使其接收多模态信息并对其进行推理. 由于多模态大语言模型更符合人类感知世界的方法，由 GPT-4 展示的惊人效果，引发了多模态大语言模型的研究热潮.

Zhu 等人[20] 提出了 MiniGPT-4 模型，该模型由一个冻结的 ViT 和 Q-Former 的视觉编码器、一个线性投影层和冻结的 Vicuna 组成. MiniGPT-4 只需要训练线性层来对齐视觉特征和 Vicuna. Zhang 等人[21] 提出 LLaMA-Adapter 的多模态模型，该模型基于门控的零初始化注意力机制，自适应地将图像特征注入 LLaMA，同时有效地保留其预训练的知识，该模型在下游任务中表现优越. 在 LLaMA-Adapter 基础上，Gao 等人[22] 提出了 LLaMA-Adapter V2 的模型，该模型是一种高效调参的视觉指令跟随多模态大模型，在微调参数更少的情况下取得了比 LLaMA-Adapter 更好的下游

任务识别效果. Liu 等人[23]提出 LLaVA 模型，通过请求 GPT-4 构造指令微调训练数据，然后在训练中加入视觉令牌进行指令微调，利用 GPT-4 的回复实现与用户意图的对齐. Horawalavithana 等人[24]提出了 LLaMA-SciTune模型. 构建了科学任务领域的专用多模态大语言模型，在科学问题回答任务上超越了人类的表现.

1.4 数据增强

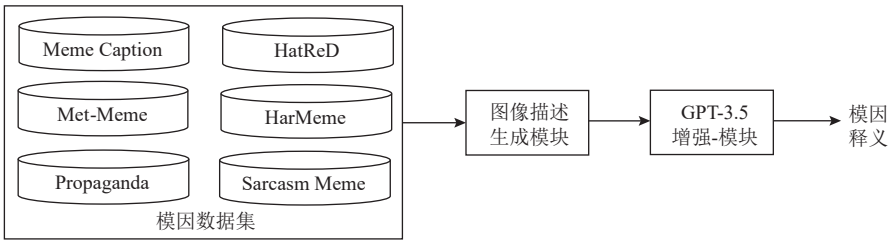
数据增强方法通常用于防止过度拟合并提高深度神经网络模型的泛化能力，是解决样本量有限任务的有效策略. 早期，Wei 等人[25]提出用于提高文本分类任务性能的简单数据增强（easy data augmentation，EDA）技术. EDA 由 4 种简单但功能强大的操作组成：同义词替换、随机插入、随机交换和随机删除. EDA 对于较小的数据集表现出特别强大的结果. 后来，预训练语言模型 PLMs 受到广泛关注后，Wu 等人[26]提出 CBERT 方法，通过掩码填词任务构建与原始输入文本具有高语义相关性的增强样本. Kumar 等人[27]进一步探索了条件数据增强的自回归模型 GPT2[28]、自动编码器模型 BERT[29]和 seq2seq 模型 BART[30]. 最近大语言模型展现出了强大的数据增强性能，Dai 等人[31]提出 ChatAug 方法，利用 ChatGPT 将训练样本中的每个句子重新表述为多个概念相似但语义不同的样本，来增强下游模型训练.

2 本文方法

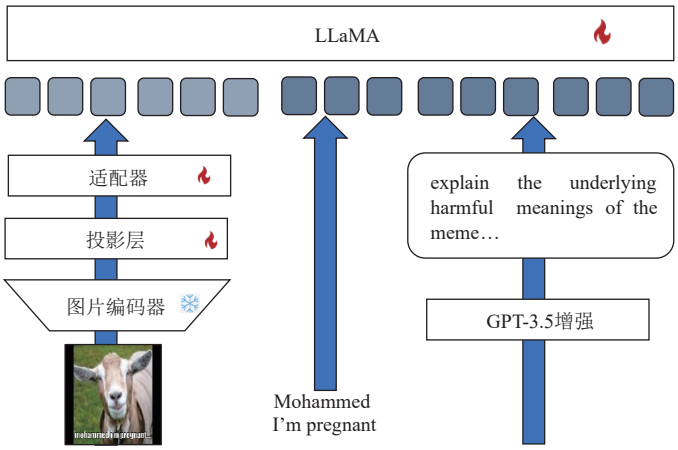
本文方法的模型结构图如图 2 所示，主要包含 3 个模块，分别是多模态大模型基座、基于 GPT-3.5 的攻击性模因解释增强和多数据集指令微调. 多模态大模型基座提供了基础的多模态信息处理和解释生成能力. 基于 GPT-3.5 的模因解释增强模块为多个模因数据集提供了模因解释，以解决模因解释数据集的缺失的问题. 多数据集指令微调训练则进一步增强了通用多模态大模型基座对模因隐喻表达的理解，同时提高了跨模态仇恨识别能力.

2.1 多模态大模型基座

LLaMA-Adapter V2[22]是一种基于适配器（adapter）的向预训练大语言模型注入图像知识的方法，在多个下游任务中取得了良好的效果. 由于该模型是一种轻量的高效参数微调方法，本文方法选择其作为训练的大模型基座. 图片编码器使用训练好的 CLIP[32]提取多尺度的全局特征，然后将这些多尺度特征拼接起来，经过一个投影层得到全局的信息表征. 在整个训练过程中冻结图片编码器的参数，通过调整投影层和适配器的参数，使图片特征注入到



(a) 阶段1: 指令集构建



(b) 阶段2: 指令微调

Fig. 2 Illustration of the proposed model

图 2 所提模型的示意图

LLaMA[33]模型中. 由于攻击性模因解释生成任务较为复杂，为了达到更好的生成效果，除了调整投影层和适配器的参数，本文方法也调整了 LLaMA 模型最后 8 层 Transformer 的参数.

2.2 基于 GPT-3.5 的攻击性模因解释增强

模因解释数据集由于构建难度大和构建成本高，目前可以获得的开源数据集只有 2 个，分别是 HatReD[5]仇恨模因解释生成数据集(3 228 个样本)和 MemeCap[19]模因隐喻解释生成数据集(6 384 个样本). 有限的训练数据可能不足以让大模型学到模因的隐喻表达手法和多模态互补的攻击性特征. 而其他的开源模因数据集主要针对分类任务构建，缺少相应的模因解释，比如对隐喻方式进行分类的 MET-Meme[34]数据集和 Propaganda[35]数据集，以及识别有害模因的 HarMeme[36]数据集和讽刺模因的 Sarcasm Meme[37]数据集. 这些数据集有人工标注的不同属性特征，比如对应分类任务的类别、情感和攻击对象等特征. 本文结合人工标注的属性特征和多模态模型生成的模因图片描述信息，利用 GPT-3.5 强大的文本推理分析能力，自动生成开源模因数据集的文本解释，使其与模因的隐喻识别和攻击性概念对齐. 具体的数据集列表如表 1 所示.

由于 GPT-3.5 不接受多模态的输入，因此本文将模因的图片特征转换为文本形式的图像描述作为 GPT-3.5 的输入. LLaMA Adapter V2[22]和 Blip-2[38]是目前效果较优的图像描述生成模型，我们在实验中发现 LLaMA Adapter V2 对图像细节的描述，比如表情、动作等方面刻画得比较丰富，但是在图像重要实体识别上有很多幻觉. Blip-2 对图像重要实体的识别比较精准，但是基本上没有图像细节的描述. 在模因解释生成的过程中，实体特征和图像细节都很重要. 因此本文生成的图像描述拼接了 Blip-2 和 LLaMA Adapter V2 的生成结果. 图 3 给出了利用 GPT-3.5 生成模因解释的示例. 在 Prompt 构建上，本文方法加入了图像描述、文字内容以及数据集原有的人工标注的各类属性标签信息，来提示 GPT-3.5 去分析输入模因所要表达的深层含义. 其中，加入各类属性标签是为了更好地利用人工标注的信息，使得生成的内容与模因包含的隐喻和攻击性概念对齐. 从 GPT-3.5 生成结果来看，在图像描述完整如实地反映图像特征的情况下，结合模因原有的文本信息，基本可以推理出模因的内在含义，为后续指令微调提供扩充的训练数据.

2.3 面向攻击内容和隐喻表达的指令微调训练

多模态大语言模型在预训练阶段所使用的训练数据中，攻击性模因的数据量有限. 因此通用多模态大语言模型在识别隐喻以及生成攻击性模因解释的能力相对较弱. 为了增强这一能力，本文利用 GPT-3.5 增强后的 6 个模因解释数据集对 LLaMA-Adapter V2 多模型大模型基座进行指令微调，以提高通用多模态大语言模型生成攻击性模因解释的质量. 指令微调的训练目标包括让多模态大语言模型充分理解模因的隐喻表达手法，并生成相关的攻击性解释.

2.3.1 指令模板

6 类主流的模因数据集如表 1 所示，本文分别从任务相关的攻击性标签、任务相关的隐喻标签和生成的解释 3 个方面，将其转换为指令跟随数据集. 构建的指令 S 包括指令模板 S_{I}、模因图像 x_{V} 和模因文本 x_{T}：

$$S=\{S_{\mathrm{I}},x_{\mathrm{V}},x_{\mathrm{T}}\}. \tag{1}$$

具体的指令模板如图 4 所示. 指令模板共包含 3 类指令：1)模因解释生成指令，此类指令与具体数据集无关；2)攻击性识别指令，此类指令与 3 个攻击性模因识别任务相关，对应 HatRed、HarMeme 和 Sarcasm Meme 数据集；3)隐喻识别指令，此类指令与 3 个模因隐喻识别任务相关，对应 MemeCap、MET-Meme 和 Propaganda 数据集. 通过融合这 3 类指令，包括判断、生成和选择等多种问题类别，基本覆盖了攻击性模因解释生成的主要要素，使得模型生成的模因解

Table 1 Statistical Information of the Datasets

表 1 数据集统计信息

数据集	模因数量	数据集说明	隐喻标签	模因解释	是否攻击性模因
HatReD[5]	3 228	仇恨模因解释生成	无	有	是
MemeCap[19]	6 384	模因解释生成	有	有	否
MET-Meme[34]	10 045	模因隐喻分类	有	有（增强后）	否
Propaganda[35]	950	模因隐喻分类	有	有（增强后）	否
HarMeme[36]	3 544	有害模因分类	无	有（增强后）	是
Sarcasm Meme[37]	24 635	讽刺模因分类	无	有（增强后）	是



Sentiment: Sorrow
Sentiment degree: Very
Intention: Expressive
Offensive: Non
Metaphor: Complimentary
Target domain: Skull
Source domain: Flower

Prompt:

I have a textual description of a meme, and I'd like your help in explaining its underlying meaning or message. Please analyze the meme based on the provided description:

任务描述

The description of the picture generated by BLIP2 is: Thepurpleflower. The description of the picture generated by LLaMA is: The painting depicts a scene of a pink rose with a skull in the middle of it. The rose is in full bloom, and the skull is placed in the center of the flower, creating a unique and intriguing composition. The skull is positioned in such a way that it appears to be looking at the viewer, adding an eerie and mysterious atmosphere to the painting.

图像描述

The text is: death in bloom. The sentiment category of the meme is sorrow, and the sentiment degree is very. The intention of the meme is expressive, the offensive degree is non-offensive. The metaphor category is complementary. The target domain is skull, the source domain is flower.

Could you break down the metaphor used in the meme and provide an explanation of what it conveys?

数据属性

GPT-3.5 output:
the meme uses the juxtaposition of a beautiful, blooming rose and a skull to symbolize the coexistence of life and death. It conveys a message of the impermanence of life, the inevitability of death, and the sorrow associated with this reality. The contrast between the two elements prompts viewers to reflect on the fleeting nature of beauty and existence.

Fig. 3 Augmenting explanatory examples with GPT-3.5

图 3 GPT-3.5 解释增强示例

攻击性识别指令集：
“Is this a hateful meme?”
“Is this a sarcasmmeme?”
“Is this a harmful meme?”
“Which one is the offensive meme?”
“What is the offensive meme targets at?”
隐喻识别指令集：
“What is the metaphor used in the meme?”
“Does the meme use metaphor?”
“What is the sentiment category of the meme?”
解释生成指令集：
“Explain the underlying meanings of the meme.”
“Uncover the latent meanings embedded within the meme.”
“Provide an analysis of the meme's hidden messages.”

Fig. 4 Instruction set

图 4 指令集

释具备隐喻分析和攻击性分析的能力.

2.3.2 损失函数

训练的损失函数如式(2)所示：

$$L_{\mathrm{CE}} = -\sum_{i=1}^{N}\sum_{j=1}^{\ell}\log\rho_\theta(r_j^i|x_{\mathrm{T}}^i, x_{\mathrm{V}}^i, r_i^i, \cdots, r_{j-1}^i), \tag{2}$$

其中x_{T}^i表示第i个模因的文本信息，x_{V}^i表示第i个模因的图像信息，r^i表示第i个模因的仇恨解释. 由于r^i由一段文字序列组成，可以将其表示为$r^i = r_1^i, \cdots, r_\ell^i$，其中$\ell$表示序列的长度，$\theta$表示多模态大模型的可调参数，$N$表示训练集的模因个数.

3 实验结果与分析

本节先对实验设置进行介绍，包括数据集、基线方法、实验参数以及实验结果评估方法；之后对实验结果进行对比分析.

3.1 数据集

本文一共使用了 6 个数据集，分别是与攻击性相关的数据集 HatReD[5], HarMeme[36] 和 Sarcasm Meme[37]. 以及与隐喻相关的数据集 MEMECAP[19], MET-Meme[34] 和 Propaganda[35]. 使用 HatReD 数据集的训练集部分和其他 5 个数据集来训练模型，保留 HatReD 的测试集来测试模型效果. 具体的划分情况如表 2 所示.

HatReD 是人工标注的多模态仇恨模因解释生成数据集，它提供了基于固定模板创建的解释. 该数据集包括 3 228 个已标注的仇恨模因，其中训练集有 2 982 个，测试集则有 246 个. 在本研究中，我们将 HatReD 的训练集并入我们的训练过程，并保留其测试集用于验证模型效果.

MemeCap 是一个模因解释生成任务数据集. 该任务要求模型根据给定的模因和模因标题，生成一个简洁的描述来解释模因的含义，这个任务涉及理解模因中的视觉隐喻，并结合跨模态信息进行解释. 该数据集有 6 384 个模因样本.

MET-Meme 包含了 10 045 个模因，并人工标注了隐喻类别、情感倾向、模因意图和冒犯程度等属性.

Table 2 Dataset Division Results

表 2 数据集划分结果

数据集	训练集	测试集
HatReD	2 982	246
MemeCap	6 384	0
MET-Meme	10 045	0
Propaganda	950	0
HarMeme	3 544	0
Sarcasm Meme	24 635	0
总计	48 540	246

该数据集考虑了模因研究中的隐喻特征并实验证明了隐喻特征的重要性.

Propaganda 数据集主要关注模因的隐喻手法识别，包括贬低性称呼、夸大和反讽等 22 种攻击隐喻宣传手段. 该数据集标注了 950 条数据.

HarMeme 是一个多模态的有害模因识别数据集，包含 3 544 个与 COVID-19 相关的有害模因. HarMeme 数据集经过了两阶段注释过程：在第 1 阶段，将模因标记为非常有害、部分有害或无害；在第 2 阶段，进一步注释了每个有害模因所指向的目标类型.

Sarcasm Meme 是一个多模态的讽刺模因分类数据集，经过人工标注以确定每个模因是否具有讽刺特征. 该数据集包含 24 635 个模因，主要从 Twitter 社交媒体平台中收集.

3.2 评价指标

本文采用了自动评估和 GPT-4 专家评估相结合的方法. 自动评估指标通过可计算的量化方式评估生成文本和参考文本之间的相似度. 专家评估指标从生成解释的正确性、流畅性、背景知识分析能力、隐喻识别能力和客观性 5 个方面评估生成质量.

3.2.1 自动评估指标

在自动评估中，采用了自然语言生成任务中常见的度量标准，其中包括基于词语相似性的 N-gram 匹配以及基于嵌入的语义相似性度量. 对于 N-gram 匹配度量标准，本文采用了平均 BLEU[39] 和 ROUGE-L[40] 的评估指标. BLEU 的得分范围在 0~1，越接近 1 表示生成的结果与目标文本越相似. ROUGE-L 是基于生成文本与参考文本之间的最长公共子序列的重叠度量，旨在衡量内容的保留程度. 基于嵌入的语义相似性度量上，本文采用 BERTScore[41] 评价方法. BERTScore 通过计算 2 个句子在 BERT 模型中的编码向量之间的余弦相似度，来评估生成文本和参考文本的相似性. 特别是对于语义相似但表达不同的句子，BERTScore 更能捕捉到它们之间的相似性. 自动评估使用的参考文本是 HatReD 人工标注答案.

3.2.2 专家评估指标

由于攻击性模因需要丰富的背景知识帮助理解，且包含大量隐喻相关的表达方式，因此在评估生成解释的时候，除了要关注正确性和流畅性指标 2 个通用指标外，还需要关注社会背景和隐喻识别的解释生成情况，因此本文列出 5 类指标进行评估：

1)正确性. 生成的解释需准确无误地反映模因所蕴含的攻击对象与攻击内容，从而避免产生误解.

2)流畅性. 生成的解释应当明晰、逻辑严密且易于理解.

3)背景知识分析能力. 生成的解释应全面考量模因在其创建和传播过程中所处的社会政治、历史及文化环境，并考虑它们在理解模因深层意义时所起的关键作用.

4)隐喻识别能力. 生成的解释应阐明隐喻元素如何与模因的主题相互作用，并分析它们对模因整体信息或意图的影响.

5)客观性. 生成的解释应保持中立和尊重，避免宣传或支持任何形式的仇恨内容，确保公平对待涉及的所有个体或群体.

本文采用 GPT-4 专家评估的方法，采用 5 分量表法进行打分，分越高表示生成的解释越满足该项指标. 1 分表示生成的解释与该指标不符，5 分表示该生成的解释充分满足该项指标.

3.3 基线方法

本文对比了 2 类方法，分别是基于预训练语言模型方法和多模态预训练模型的解释生成方法. 对于预训练语言模型的解释生成方法，本文对比了 Hee 等人[5] 提出的基于 T5-large[42] 的方法、基于 RoBERTa[17] 的方法和基于 RoBERTa-base-GPT2[28] 的方法. 由于预训练语言模型仅接受文本输入，图像特征需转为由 Blip-2 生成的文字图像描述和原始文本信息. 对于多模态预训练模型的解释生成方法，本文对比了 Hee 等人[5] 提出的VL-T5-VL-T5, VisualBERT-GPT2 和VisualBERT-RoBERTa 方法.

3.4 参数设置

实验参数设置为：批次大小(batch size)为 16，共进行 20 个训练周期(epoch). 每个批次的梯度更新次数为 1，基础学习率为 5E−4. 在前 10 个训练周期内，采用学习率的 warm up 策略，之后的周期中采用半周期余弦(half-cycle cosine)学习率调度策略. 本文使用了 AdamW 优化策略，其中衰减率 β_1 设定为 0.9，β_2 设定为 0.95. 权重衰减(weight decay)设置为 0.02，用于对模型参数进行正则化控制. 模型接受输入的最大长度序列长度为 128. GPT-3.5 使用的版本是 GPT-3.5-turbo-0613.

3.5 实验结果

3.5.1 自动评估指标实验结果

在 HatReD 测试集上的实验结果如表 3 所示. 在所有 6 个基线模型中，采用 GPT2-base 解码器的 2 个基线模型(包括文本模型和多模态模型)效果相对较差，其余 4 个模型(包括文本模型和多模态模型)效果差别不大. 综上所述，文本模型和多模态模型在自

动评估指标上差别不大.

本文模型的效果在基于 BERTScore 的评估标准上大大优于基线模型. BERT-P 达到了 0.689, 比基线模型提升了 0.181; BERT-R 达到了 0.668, 比基线模型提升了 0.195; BERT-F 达到了 0.677, 比基线模型提升了 0.197. 基于 BERTScore 的结果表明, 本文方法生成的模因解释在语义上更接近人工标注的解释, 证明了本文方法的有效性. 在基于 N-gram 的评价标准上, 本文方法没有取得最优的性能. 原因在于本文方法生成的解释长度较长, 平均长度为 62.17 个词; 生成的解释不仅包括模因的攻击对象和攻击点, 也深入分析了模因中多模态的隐喻手法和背景知识, 可以更准确更全面地理解模因的含义. 而人工标注的解释依据模板生成, 十分简短, 平均长度为 13.62 个词.

3.5.2 GPT-4 专家评价实验结果

本文采用 GPT-4 专家评分方法来对 HatReD 测试集的人工标注答案、本文方法以及 T5-large 模型生成的内容进行评分, 实验结果如表 4 所示. 在正确性和客观性这 2 个指标上, 人工标注答案的表现最佳. 正确性指标主要用于评估生成解释的攻击目标和攻击内容是否正确, 是理解攻击性模因的关键指标. GPT-4 在评分中对人工标注答案给出了最高分(4.66/5), 显著优于其他方法. 同时, 由于人工标注在处理负面内容时更加谨慎, 因此在客观性指标上也获得了最高分(4.9/5). 这 2 个评价指标表明 GPT-4 和人工标注在理解攻击目标和内容方面基本一致. 然而, 由于 HatReD 数据集人工标注答案没有考虑到隐喻识别和背景知识, 因此这 2 项指标的得分明显低于本文方法. 本文方法在文本流畅度、背景分析能力和隐喻识别能力方面均表现出色, 尽管在正确性指标上低于人工标注答案, 但明显优 T5-large 模型, 验证了本文方法的有效性.

3.5.3 消融实验

消融分析实验旨在进一步探究各个模块对攻击性模因解释生成性能的影响. 具体设置为:

1)w/o 数据增强. 使用原本有解释的 2 个数据集 HatReD 和 MemeCap 做解释生成任务, 而其他 4 个数据集只做分类任务. 2)w/o 隐喻数据集. 在训练集中去掉 MemeCap, Propaganda 和 MET-Meme 3 个关注隐喻分析的数据集. 3)w/o 有害数据集. 在训练集中去掉 Sarcasm Meme 和 HarMeme 数据集. 由于 HatReD 数据集人工标注答案的局限性, 特别是其模板化的解释生成方式, 没有包含对模因隐喻的解释以及相关背景知识. 因此, 为了更加客观地评估所生成的解释, 本文主要采用了基于 GPT-4 的专家评估方法进行消融实验. 此外, 将 HatReD 数据集的人工标注答案作为自动评估的参考文本, 提供补充评估参考.

消融实验的 GPT-4 专家评分结果如表 5 所示. 当

Table 3 Experimental Results from Automated Evaluations

表 3 自动评估实验结果

模型			N-gram 匹配		基于 BERTScore 的评估		
类别	编码器	解码器	BLEU	ROUGE-L	BERT-P	BERT-R	BERT-F
文本模型	RoBERTa-base	GPT2-base	0.068	0.222	0.112	0.327	0.218
	RoBERTa-base	RoBERTa-base	0.177	0.389	<u>0.508</u>	0.453	<u>0.480</u>
	T5-large	T5-large	**0.190**	**0.392**	0.485	<u>0.473</u>	0.479
图片-文本模型	VisualBERT	GPT2-base	0.065	0.219	0.100	0.342	0.219
	VisualBERT	RoBERTa-base	0.179	<u>0.391</u>	0.499	0.449	0.474
	VL-T5	VL-T5	<u>0.180</u>	0.378	0.472	0.409	0.446
本文方法			0.098	0.250	**0.689**	**0.668**	**0.677**
本文方法与其他方法的差值			−0.092	−0.142	0.181	0.195	0.197

注: 加粗表示最优结果, 下划线表示次优结果.

Table 4 Experimental Results from GPT-4 Evaluations

表 4 GPT-4 评估实验结果

方法	流畅度	正确性	背景分析能力	隐喻识别能力	客观性
本文方法	**4.90**	<u>3.74</u>	**4.42**	**3.86**	<u>4.74</u>
T5-large	4.70	2.08	2.34	2.48	4.32
HatReD 人工标注答案	<u>4.80</u>	**4.66**	<u>3.22</u>	<u>3.00</u>	**4.90**

注: 加粗表示最优结果, 下划线表示次优结果.

移除由 GPT-3.5 生成的增强数据后, 在正确性、背景分析能力以及隐喻识别能力方面均出现了明显的下降, 尤其是背景分析能力的下降最为显著, 证明了 GPT-3.5 增强数据在生成模因解释过程中扮演了关键要素. 相比之下, 移除有害数据集比移除隐喻数据集对正确性、背景分析能力和隐喻识别能力的影响更为显著, 这表明在攻击性模因解释生成任务中, 使用丰富的同类数据进行训练能够带来更加明显的效果提升. 表 6 展示了消融实验的自动评估结果, 其与 GPT-4 专家评分的结果大体一致. 然而, 由于自动评估方法采用的是 HatReD 数据集人工标注答案作为评估标准, 并未覆盖模因的隐喻解释和背景知识. 因此, 当移除 GPT-3.5 增强数据时, 生成的解释更加简短且形式上更接近 HatReD 的标准答案, 从而导致 BLEU 和 ROUGE-L 得分较高. 但由于缺乏其他数据的增强, 基于语义的 BERTScore 分数出现了较大的下降, 其 BERT-F 值比本文方法降低了 0.094.

3.5.4 对攻击性模因分类任务的影响

为了验证生成解释的有效性, 我们将生成的解释作为提示加入到原生的 LLaMA-Adapter V2, 验证生成的解释是否对于 zero-shot 的分类有增益. 实验结果如表 7 所示. 可以看出增加了生成的解释后, 模型 zero-shot 的分类效果有明显的提升, HateMeme 数据集的准确度(accuracy, ACC)提升了 14.6 个百分点, 受试者曲线面积(area under the receiver operating characteristic curve, AUROC)提升了 2.1 个百分点. 实验结果验证了本文方法生成解释对攻击性模因的分类有较大的增益效果.

Table 5 Ablation Experimental Results from GPT-4 Evaluations

表 5 消融实验 GPT-4 评估结果

方法	流畅度	正确性	背景分析能力	隐喻识别能力	客观性
本文方法	**4.90**	**3.74**	**4.42**	**3.86**	4.74
w/o 数据增强	4.82	2.46	2.52	2.94	4.48
w/o 隐喻数据集	4.86	3.12	3.74	3.24	4.36
w/o 有害数据集	4.82	2.94	3.36	3.18	4.78

注: 加粗表示最优结果, 下划线表示次优结果.

Table 6 Ablation Experimental Results from Automated Evaluations

表 6 消融实验自动评估结果

方法	N-gram 匹配		基于 BERTScore 的评价		
	BLEU	ROUGE-L	BERT-P	BERT-R	BERT-F
本文方法	0.098	0.250	**0.689**	**0.668**	**0.677**
w/o 数据增强	**0.108**	**0.335**	0.604	0.569	0.583
w/o 隐喻数据集	0.090	0.318	0.661	0.649	0.653
w/o 有害数据集	0.074	0.279	0.581	0.635	0.604

注: 加粗表示最优结果, 下划线表示次优结果.

3.5.5 跨领域数据实验结果

为了评估本文方法的泛化性能, 将本文方法在新数据集 MAMI[43] 上进行攻击性解释生成实验. MAMI 数据集主要收集针对女性的仇恨模因. 由于 MAMI 是面向分类任务的数据集, 没有人工编写的解释, 因此本文采用 GPT-4 评估生成效果方法. 表 8 展示了在 MAMI 数据集上随机抽样 50 个模因生成解释的专家评估结果. 实验结果说明本文方法具有良好的泛化能力, 面向新的数据也能做出全面和准确的攻击性解释生成.

3.5.6 案例分析

除了定量分析的方法, 本文通过实例说明本文方法生成解释的优势. 如图 5 所示, 本文方法生成的解释既包括攻击的目标和攻击内容, 也包括背景知识和隐喻解释, 如图 5 的下划线所示内容, 内容全面易懂. 相比之下, 人工标注的解释比较简洁, 采用“负面谓词+攻击目标+攻击手段”的模板化格式生成方式, 并未探讨模因的意图、象征意义或背后的社会文化背景, 理解依然有一定的难度. 而 T5-large 方法在理解攻击目标和攻击内容上存在一些错误. 比如图 5 的第 1 个例子中, 人工标注的结果中, 关于“coon”的说明仅有“derogatory term(诬称词)”, 如果读者没有相关的背景知识依然无法准确理解其含义. 而 T5-large 生成的解释错误地将对“blacks(黑人)”的攻击解释为针对“Chinese(中国人)”的攻击. 本文方法生成的解释, 在隐喻释义生成部分详细说明了“coon”的含义, 解释了模因如何通过隐喻的手段来强化对黑人的负面刻板印象, 展示了本文方法对模因解释的全面性和正确性.

Table 7 Experimental Results of Offensive Memes Classification Task

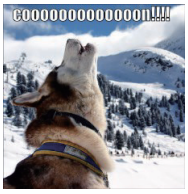
表 7 攻击性模因分类任务结果 %

数据集	无生成原因		有生成原因	
	ACC	AUROC	ACC	AUROC
HateMeme	33.4	48.5	48.0	50.6

Table 8 Experimental Results on Unseen Data

表 8 在未见数据集上的实验结果

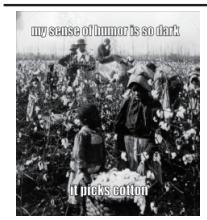
方法	流畅度	正确性	背景分析能力	隐喻识别能力	客观性
本文方法	4.90	3.54	4.26	4.12	4.85
T5-large	4.32	2.48	2.73	2.78	4.60



人工标注结果：disrespects the blacks by the use of the deragatory term coon.
T5-large 方法生成的解释：mocks the Chinese for eating dogs using derogatory term.
本文方法生成的解释：The metaphor used in the meme is comparing the howling of the dog to the sound of a crying black person. The crying black person is often used as a metaphor for someone who is perceived as weak, vulnerable, and in need of help. By howling, the dog is implying that it is in a state of distress or discomfort, much like a crying black person would be. The use of the snowy environment and mountain range could also symbolize the idea of isolation and vulnerability, as the howling dog is surrounded by a vast and potentially harsh landscape. Overall, the meme conveys a message of white supremacy and reinforces harmful stereotypes about black people.



人工标注结果：disrespects the jews by downplaying the gassing of jews during the holocaust.
T5-large 方法生成的解释: disrespects the jews by making fun of the gassing of the jews in world war 2.
本文方法生成的解释：The metaphor used in this meme is "going full gas," which is a reference to the Nazi regime's use of gas chambers to kill Jews during the Holocaust. The jews depicted in the painting are being herded into the chambers, which have been described as "killing factories." The use of this metaphor conveys a message of support for the Nazi regime and its ideology, which is a hateful and dangerous message. The jews in the painting are shown as happy and content, which is a false representation of the jews who were actually being killed in these chambers. The use of this metaphor is a clear example of how hate memes can spread harmful messages and promote hate speech.



人工标注结果：dehumanizes the blacks by suggesting that they should be ensalved.
T5-large 方法生成的解释：disrespects the blacks by suggesting that they pick cottons.
本文方法生成的解释：The metaphor used in the meme is \"my sense of humor is so dark it picks cotton,\" which is a play on words based on the history of black cotton-picking slaves. The slaves were often dark-skinced man and woman who were forced to pick cotton in the cotton fields. The phrase suggests that the person's sense of humor is so dark and sarcastic that it is similar to the pain and suffering experienced by the black cotton-picking slaves. The underlying message of the meme is highly offensive as it diminishes the historical and ongoing pain caused by slavery and systemic racism.

Fig. 5 Comparison of three types of explanations with examples

图 5 举例说明 3 种解释的比较

4 结 论

本文提出了一种基于多模态大语言模型的攻击性模因解释生成方法，生成的解释不仅包括攻击性模因的攻击目标和攻击内容，而且涵盖了其隐喻表达方式和社会背景知识，从而形成了一个更加全面完整的解释. 实验结果表明，本文方法生成的解释在 BERTScore 评估标准下相较于基线模型提高了 19%. 基于 GPT-4 的专家评估方法表明本文方法生成的解释在处理背景知识分析和隐喻识别方面超越了人工标注结果. 此外，实验结果也表明本文方法对新数据也展现出了很好的泛化性.

作者贡献声明：林萌提出了论文的核心思想，设计算法进行实验并撰写论文；戴程威实现了论文对比方法；郭涛提出指导意见并修改论文.

参 考 文 献

[1] Hu Songlin, Zhao Jun, Tang Jie, et al. Preface to the special issue on fake information detection[J]. Journal of Computer Research and Development, 2021, 58(7): 1351−1352 (in Chinese)
(虎嵩林, 赵军, 唐杰, 等. 虚假信息检测专题前言[J]. 计算机研究与发展, 2021, 58(7): 1351−1352)

[2] Kiela D, Firooz H, Mohan A, et al. The hateful memes challenge: Detecting hate speech in multimodal memes[J]. Advances in Neural Information Processing Systems, 2020, 33: 2611−2624

[3] Zhang Linhao, Jin Li, Sun Xian, et al. TOT: Topology-aware optimal transport for multimodal hate detection[C]//Proc of the AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2023, 37(4): 4884−4892

[4] Cao Rui, Hee MS, Kuek A, et al. Pro-Cap: Leveraging a frozen vision-language model for hateful meme detection[C]//Proc of the 31st ACM Int Conf on Multimedia. New York: ACM, 2023: 5244−5252

[5] Hee M S, Chong W H, Lee R K W. Decoding the underlying meaning of multimodal hateful memes[C]//Proc of the 32nd Int Joint Conf on Artificial Intelligence. Freiburg: IJCAI, 2023: 5995−6003

[6] Sharma S, Agarwal S, Suresh T, et al. What do you MEME? Generating explanations for visual semantic role labelling in memes[C]//Proc of the AAAI Conf on Artificial Intelligence. Washington, DC: AAAI, 2023, 37(8): 9763−9771

[7] Scott K. Memes as multimodal metaphors: A relevance theory analysis[J]. Pragmatics & Cognition, 2021, 28(2): 277−298

[8] Pramanick S, Sharma S, Dimitrov D, et al. MOMENTA: A multimodal framework for detecting harmful memes and their targets[C]//Findings of the Association for Computational Linguistics: EMNLP 2021. Stroudsburg, PA: ACL, 2021: 4439−4455

[9] Zhu Ron. Enhance multimodal Transformer with external label and in-domain pretrain: Hateful meme challenge winning solution[J]. arXiv preprint, arXiv: 2012.08290, 2020

[10] Yang Chuanpeng, Zhu Fuqing, Liu Guihua, et al. Multimodal hate speech detection via cross-domain knowledge transfer[C]//Proc of the 30th ACM Int Conf on Multimedia. New York: ACM, 2022:

4505−4514

[11] Lee R K W, Cao Rui, Fan Ziqing, et al. Disentangling hate in online memes[C]//Proc of the 29th ACM Int Conf on Multimedia. New York: ACM, 2021: 5138−5147

[12] Velioglu R, Rose J. Detecting hate speech in memes using multimodal deep learning approaches: Prize-winning solution to hateful memes challenge[J]. arXiv preprint, arXiv: 2012.12975, 2020

[13] Yin Shukang, Fu Chaoyou, Zhao Sirui, et al. A survey on multimodal large language models[J]. arXiv preprint, arXiv: 2306.13549, 2023

[14] Gupta T, Kembhavi A. Visual programming: Compositional visual reasoning without training[C]//Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2023: 14953−14962

[15] Shao Zhenwei, Yu Zhou, Wang Meng, et al. Prompting large language models with answer heuristics for knowledge-based visual question answering[C]//Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2023: 14974−14983

[16] Cao Rui, Lee R K W, Chong W H, et al. Prompting for multimodal hateful meme classification[C]//Proc of the 2022 Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2022: 321−332

[17] Liu Yinhan, Ott M, Goyal N, et al. RoBERTa: A robustly optimized BERT pretraining approach[J]. arXiv preprint, arXiv: 1907.11692, 2019

[18] Ji Junhui, Ren Wei, Naseem U. Identifying creative harmful memes via prompt based approach[C]//Proc of the ACM Web Conf 2023. New York: ACM, 2023: 3868−3872

[19] Hwang E J, Shwartz V. MemeCap: A dataset for captioning and interpreting memes[J]. arXiv preprint, arXiv: 2305.13703, 2023

[20] Zhu Deyao, Chen Jun, Shen Xiaoqian, et al. MiniGPT-4: Enhancing vision-language understanding with advanced large language models[J]. arXiv preprint, arXiv: 2304.10592, 2023

[21] Zhang Renhui, Han Jiaming, Liu Chris, et al. LLaMA-Adapter: Efficient fine-tuning of language models with zero-init attention[J]. arXiv preprint, arXiv: 2303.16199, 2023

[22] Gao Peng, Han Jiaming, Zhang Renrui, et al. LLaMA-Adapter V2: Parameter-efficient visual instruction model[J]. arXiv preprint, arXiv: 2304.15010, 2023

[23] Liu Haotian, Li Chunyuan, Wu Qingyang, et al. Visual instruction tuning[J]. Advances in Neural Information Processing Systems, 2024. DOI: 10.48550/arXiv.2304.08485

[24] Horawalavithana S, Munikoti S, Stewart I, et al. SciTune: Aligning large language models with scientific multimodal instructions[J]. arXiv preprint, arXiv: 2307.01139, 2023

[25] Wei J, Zou Kai. EDA: Easy data augmentation techniques for boosting performance on text classification tasks[C]//Proc of the 2019 Conf on Empirical Methods in Natural Language Processing and the 9th Int Joint Conf on Natural Language Processing (EMNLP-IJCNLP). Stroudsburg, PA: ACL, 2019: 6382−6388

[26] Wu Xing, Lv Shangwen, Zang Liangjun, et al. Conditional BERT contextual augmentation[C]// Proc of 19th Int Conf on Computational Science(CCS 2019). Berlin: Springer, 2019: 84−95

[27] Kumar V, Choudhary A, Cho E. Data augmentation using pre-trained Transformer models[C]//Proc of the 2nd Workshop on Life-long Learning for Spoken Language Systems. Stroudsburg, PA: ACL, 2020: 18−26

[28] Radford A, Jeffrey W, Child R, et al. Language models are unsupervised multitask learners[J]. OpenAI Blog, 2019, 1(8): 9−9

[29] Kenton J D M W C, Toutanova L K. BERT: Pre-training of deep bidirectional Transformers for language understanding[C]//Proc of Annual Conf of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg, PA: ACL, 2019: 4171−4186

[30] Lewis M, Liu Yinhan, Goyal N, et al. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension[C]//Proc of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2020: 7871−7880

[31] Dai Haixing, Liu Zhengliang, Liao Wenxiong, et al. ChatAug: Leveraging ChatGPT for text data augmentation[J]. arXiv preprint, arXiv: 2302.13007, 2023

[32] Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision[C]//Proc of Int Conf on Machine Learning. New York: PMLR, 2021: 8748−8763

[33] Touvron H, Lavril T, Izacard G, et al. Llama: Open and efficient foundation language models[J]. arXiv preprint, arXiv: 2302.13971, 2023

[34] Xu Bo, Li Tingting, Zheng Junzhe, et al. MET-Meme: A multimodal meme dataset rich in metaphors[C]//Proc of the 45th Int ACM SIGIR Conf on Research and Development in Information Retrieval. New York: ACM, 2022: 2887−2899

[35] Dimitrov D, Ali B B, Shaar S, et al. Detecting propaganda techniques in memes[C]//Proc of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th Int Joint Conf on Natural Language Processing (ACL-IJCNLP 2021). Stroudsburg, PA: 2021: 6603−6617

[36] Pramanick S, Sharma S, Dimitrov D, et al. MOMENTA: A multimodal framework for detecting harmful memes and their targets[C]//Findings of the Association for Computational Linguistics: EMNLP 2021. Stroudsburg, PA: ACL, 2021: 4439−4455

[37] Cai Yitao, Cai Huiyu, Wan Xiaojun. Multi-modal sarcasm detection in twitter with hierarchical fusion model[C]//Proc of the 57th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2019: 2506−2515

[38] Li Junnan, Li Dongxu, Savarese S, et al. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models[J]. arXiv preprint, arXiv: 2301.12597, 2023

[39] Papineni K, Roukos S, Ward T, et al. BLEU: A method for automatic evaluation of machine translation[C]//Proc of the 40th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2002: 311−318

[40] Lin Chin-Yew. Rouge: A package for automatic evaluation of summaries[C]//Text Summarization Branches Out. Stroudsburg, PA: ACL, 2004: 74−81

[41] Zhang Tianyi, Kishore V, Wu Felix, et al. BERTScore: Evaluating text generation with BERT[J]. arXiv preprint, arXiv: 1904.09675, 2019

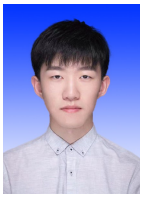
[42] Raffel C, Shazeer N, Roberts A, et al. Exploring the limits of transfer learning with a unified text-to-text transformer[J]. The Journal of Machine Learning Research, 2020, 21(1): 5485−5551

[43] Fersini E, Gasparini F, Rizzi G, et al. SemEval-2022 Task 5: Multimedia automatic misogyny identification[C]//Proc of the 16th Int Workshop on Semantic Evaluation (SemEval-2022). Stroudsburg, PA: ACL, 2022: 533−549



Lin Meng, born in 1991. PhD candidate. Her main research interests include multi-modal hate speech detection and multimodal semantic understanding.

林　萌, 1991 年生. 博士研究生. 主要研究方向为多模态仇恨言论检测、多模态语义理解.



Dai Chengwei, born in 2000. Master candidate. His main research interests include model extraction and large language model distillation.

戴程威, 2000 年生. 硕士研究生. 主要研究方向为模型提取、大模型蒸馏.



Guo Tao, born in 1974. PhD, Professor, PhD supervisor. His main research interests include cybersecurity, vulnerability analysis and risk assessment.

郭　涛, 1974 年生. 博士, 教授, 博士生导师. 主要研究方向为网络空间安全、漏洞分析与风险评估.