大语言模型生成文本的有害性检测

王鑫[1)]

[1)](西安交通大学电子与信息学部 西安 710000)

摘 要 本研究致力于解决大语言模型生成文本的有害性检测难题，为此精心构建了一个既包含人类生成的有害文本，又包含大语言模型生成的有害文本的综合数据集——HateAI-100。本研究选用DistilBERT作为预训练模型，并通过IMDB数据集对其进行微调，旨在提升其检测文本有害性的能力。实验结果显示，微调后的DistilBERT模型在人类生成文本的有害性检测任务中表现卓越，准确率高达88%。然而，在处理大语言模型生成的有害文本时，其性能却出现显著下滑，准确率仅为72%。这一发现不仅揭示了大语言模型在生成复杂且隐蔽有害内容方面的强大能力，同时也凸显了当前针对大语言模型生成文本的有害性检测技术的短缺。展望未来，本研究建议深化对大语言模型生成的有害文本的理解，持续扩充包含大语言模型生成的有害文本的数据集，并积极探索更多先进的深度学习模型与算法用于大语言模型生成文本的有害性检测。

关键词 大语言模型；文本生成；有害性检测；预训练模型；微调

Harmfulness Detection of Text Generated by Large Language Models

WANG Xin[1)]

[1)](Department of Electronics and Information, Xi'an Jiaotong University, Xi'an 710000)

Abstract This research is dedicated to solving the problem of harmfulness detection of text generated by large language models. To this end, a comprehensive data set——HateAI-100, is carefully constructed that contains both harmful text generated by humans and harmful text generated by large language models. This study selected DistilBERT as a pre-training model and fine-tuned it through the IMDB data set, aiming to improve its ability to detect harmful text. Experimental results show that the fine-tuned DistilBERT model performs excellently in the harmfulness detection task of human-generated text, with an accuracy rate as high as 88%. However, when dealing with harmful text generated by large language models, its performance dropped significantly, with an accuracy of only 72%. This finding not only reveals the powerful ability of large language models to generate complex and covertly harmful content, but also highlights the current shortage of harmfulness detection technology for text generated by large language models. Looking to the future, this study recommends deepening the understanding of harmful text generated by large language models, continuing to expand the data set containing harmful text generated by large language models, and actively exploring more advanced deep learning models and algorithms for harmfulness detection of text generated by large language models.

Keywords large language model; text generation; harmfulness detection; pre-trained model; fine-tuning

王鑫，女，2004年生，本科，E-mail: 2983099412@qq.com.

1 引言

随着人工智能技术的日新月异，大语言模型在文本生成、翻译、问答以及代码生成等多个领域展现出了卓越的能力和广泛的应用前景。这些模型不仅能够以惊人的速度生成大量文本，而且在语义理解、上下文连贯性等方面也取得了显著的进步。然而，正如任何技术都有其潜在的风险和挑战一样，大语言模型在文本生成的过程中，同样存在生成有害或不合规内容的可能性。

文本的有害性，指的是文本内容中包含的、可能对个人、群体或社会造成负面影响或伤害的特性。这些有害内容可能包括情感伤害、误导信息、仇恨言论、色情内容、暴力倾向等，它们不仅可能损害个人的名誉和利益，还可能破坏社会的和谐与稳定。因此，针对大语言模型生成文本的有害性检测研究，不仅是保障人工智能技术应用安全性的重要环节，也是维护社会道德和法律秩序的必要手段。

尽管目前已有众多研究聚焦于文本的有害性检测，但大多数研究主要集中在人类生成文本的有害性检测任务上。这些研究通常基于人类编写的文本，通过机器学习或深度学习算法来识别有害内容。然而，专门针对大语言模型生成文本的有害性检测的研究尚显匮乏。由于大语言模型生成的文本与人类编写的文本在风格、语言特征等方面存在差异，因此现有的检测文本有害性的算法可能无法有效地应用于大语言模型生成的文本。这一现状使得大语言模型在文本生成时可能带来的风险与挑战未能得到充分应对，从而增加了其在实际应用中的潜在风险。

因此，本研究旨在对大语言模型生成的文本进行有害性检测，以填补这一研究领域的空白。本研究期望能够为人工智能技术的安全应用提供有力保障，推动大语言模型在文本生成领域健康发展。

2 背景

2.1 文本的有害性

文本的有害性是指文本中潜藏的那些可能对个人、集体或整个社会施加负面效应或造成伤害的特质。这类有害元素可能涵盖情感层面的创伤、误导性的资讯、煽动仇恨的言辞、色情描绘、以及暴力倾向等。它们不仅可能侵害个人的声誉与权益，还可能危及社会的和谐与安宁。

2.2 文本的有害性检测

文本的有害性检测即对文本的有害性进行检测，是指利用自然语言处理技术和机器学习算法，对文本内容进行分析和评估，以识别并过滤掉具有潜在有害性的文本的过程。这一过程通常包括数据预处理、特征提取、模型训练和评估等步骤。

2.3 大语言模型

大语言模型（Large Language Models, LLMs）是指具有大量参数和复杂结构的神经网络模型，它们能够生成连贯、自然的文本。然而，这些模型也可能生成包含有害内容的文本，尤其是在未经充分训练或监管的情况下。

2.4 BERT

BERT（Bidirectional Encoder Representations from Transformers）是一种预训练语言模型（Pre-trained Language Model，PLM）。它由Google于2018年推出，在自然语言处理领域取得了重大突破。BERT是一种基于Transformer架构的模型，通过在大规模文本数据上进行预训练，能够捕捉到语言的深层双向表征。这种双向性使得BERT能够更好地理解文本的上下文和语义关系，从而在多种自然语言处理（Natural Language Processing，NLP）任务中表现出色。

2.5 微调技术

微调是机器学习中的一种技术，特别是在深度学习领域，它涉及在一个已经预训练好的大型神经网络模型的基础上，对其特定层或部分参数进行微小的调整，以适应新的任务或数据集。微调的主要目的是利用预训练模型在大规模数据集上学习的通用特征，减少对新任务的训练时间和数据需求。

2.6 相关工作

尽管目前已有众多研究聚焦于人类生成文本的有害性检测，但针对大语言模型生成文本的有害性检测却相对较少。现有研究更多地是利用大语言模型的文本生成能力在人类生成的有害文本检测任务中起到辅助作用，而专门针对大语言模型生成文本的有害性检测研究尚显匮乏。这一领域的研究亟待加强，以更好地应对大语言模型在文本生成过程中可能带来的风险与挑战。

人类生成文本的有害性检测是一个备受广泛关注和深入研究的领域。具体来说，仇恨言论检测是其中一个主要的研究方向，代表性工作包括Waseem等人[1]采用的基于n-gram和性别特征的有监督机器学习方法检测仇恨言论文本；Kwok等人[2]采用的基于unigram和词袋模型的有监督机器

学习方法检测反黑人言论文本；Sarwar等人[3]首次将无监督域适应方法应用于仇恨言论文本检测，并结合了数据增强和半监督学习方法；Basile等人[4]介绍了SemEval-2019 Task 5任务的目标（移民和女性仇恨言论检测）、数据集、评估方法以及参与团队的结果，旨在推动多语言仇恨言论文本检测；Kiela等人[5]提出了一个包含了“良性干扰因素”表情包挑战集，展示了多种包括视觉和文本信息的单模态和先进的多模态方法在挑战集上的表现。

攻击言论检测是另一个主要的研究方向。代表性工作包括Davidson等人[6]指出区分仇恨言论与攻击性语言是自动检测的主要挑战，研究采用众包方式和Hatebase词库收集并分类数据，训练了一个多类文本分类器以准确区分仇恨言论、攻击性语言和无关内容；Mandl等人[7]介绍了FIRE 2019 HASOC赛道任务，包括印地语、德语和英语仇恨言论和冒犯性文本识别；Kumar等人[8]介绍了LREC 2020 TRAC-2的攻击和性别攻击文本识别任务结果，涵盖孟、印、英三种语言；Fortuna等人[9]提出了合并不同数据集可以提高攻击性文本分类模型的性能；Mittos等人[10]采用了关键词搜索、自然语言处理和视觉工具等方法从Reddit和4chan社交平台收集关于遗传测试的文本，发现其与种族主义、仇恨言论等社会问题相关联；

此外，还存在其它多种类别的人类生成文本的有害性检测研究。代表性工作包括Dadvar等人[11]提出了结合用户上下文可以显著提升网络欺凌检测效果；Chen等人[12]提出了使用词汇句法特征模型检测攻击性语言和用户潜在攻击性；Ortega等人[13]提出了通过传播正面和负面的信任信息来计算用户的信任度排名，以检测恶意用户。

近两年来，人类生成文本的有害性检测这个研究领域在顶级会议所发表的论文[14-27]数量相对有限。多数研究聚焦于低资源语言环境下人类生成文本的有害性检测任务，以及如何利用大语言模型在自然语言处理方面的卓越理解能力，通过引入提示词的策略，将这些模型转化为高效的人类生成文本的有害性检测工具。

3 实验

3.1 数据集

3.1.1 MultiJail

本研究采用了一个由人类生成的有害文本数据集MultiJail[28]，该数据集基于一项针对大型语言模型越狱风险的深入分析研究构建而成。数据集的构建过程严格遵循系统化和标准化的流程，主要包括以下关键步骤：首先，通过一套严格制定的筛选标准，从大量英文查询中筛选出一系列具有潜在危害性的样本；随后，这些样本由具备相关语言专业背景的母语专家进行人工翻译，涵盖九种非英语语言。这九种语言的选择旨在实现从高资源语言到低资源语言的全面覆盖与均衡分布，从而构建了首个多语言越狱数据集——MultiJail。该数据集具有较高的灵活性与广泛的适用性，既可直接用于模拟无意触发越狱的场景，也可通过与英文恶意指令的巧妙结合，模拟有意触发越狱的复杂情境。

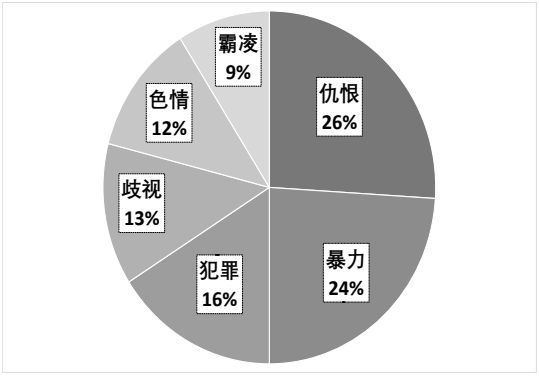

图1 不同有害类型在HateAI-100数据集中的分布

3.1.2 HateAI-100

本研究从MultiJail中精选了50个英文有害查询样例，并设计了一套提示词策略，以引导大型语言模型“文心一言”生成相应的同义表述。通过这一过程，最终构建了由人类与语言模型分别生成的合计100个有害样例，从而形成了一个全新的有害文本数据集——HateAI-100。如图1所示，HateAI-100包含多种常见的有害文本类型。该数据集的构建为检测大型语言模型生成文本的有害性提供了重要的数据支持和实验基础。

3.2 模型

3.2.1 预训练模型

本研究在模型训练过程中采用了DistilBERT作为预训练模型，这一选择是基于其多方面的优势。DistilBERT作为BERT[29]的一个轻量级版本，通过知识蒸馏这一先进的技术手段，从BERT这一强大的模型中提炼出了关键信息。这种提炼过程不仅保留了BERT模型的核心性能，如准确性、泛化能力等，还显著降低了对计算资源的需求。

具体来说，DistilBERT通过减少模型参数的数量和简化模型结构，实现了计算效率的大幅提升。

这意味着在相同的硬件条件下，DistilBERT能够更快地处理数据，完成训练任务。同时，由于计算资源的减少，DistilBERT也能够在资源受限的环境下，如移动设备或嵌入式系统中，实现高效的自然语言处理任务。

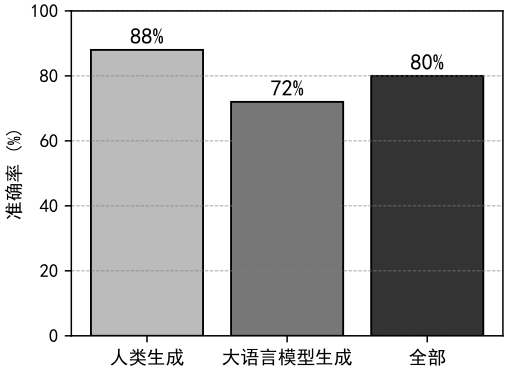

图2 微调后的DistilBERT在HateAI-100数据集上的检测结果

这一特性使得DistilBERT在许多自然语言处理任务中成为了优选。例如，在文本分类、命名实体识别、情感分析等任务中，DistilBERT都能够以较低的计算成本取得与BERT相当甚至更好的性能。此外，DistilBERT的轻量级特性还使得它更易于部署和维护，降低了模型在实际应用中的门槛。

3.2.2 微调

IMDB数据集是一个包含大量电影评论的数据集，每条评论都标记为正面或负面情感。为了检测文本是否有害，本研究在DistilBERT的基础上添加了一个用于二分类任务的分类头：使得模型对于输入的文本分析之后将输出一个“positive”或“negative”的标签。这里，本研究将IMDB数据集中的正面情感理解为“无害”，负面情感理解为“有害”。通过对DistilBERT在IMDB数据集上进行微调，我们能够使其适应文本的有害性检测任务。在微调之前，IMDB数据集经过了预处理，包括文本清洗、分词等步骤，以确保模型能够准确理解输入文本并分为训练集、验证集和测试集。这一步骤不仅为后续的生成文本的有害性检测任务提供了可靠的数据支持，还通过情感标签的转换，为检测任务的实现奠定了基础。

3.2.3 参数

训练过程调用了Hugging Face的Trainer API，这是一个功能强大的工具，能够简化模型训练、评估和部署的流程。在训练过程中，本研究设置了学习率为2e-5、每设备批量大小为16、训练轮数为2等超参数。此外，为了优化数据批处理的效率，还采用了动态填充技术（Dynamic Padding），该技术能够适配不同长度的输入序列，从而提高了训练效率。同时，为了防止模型过拟合，本研究还引入了权重衰减（Weight Decay）策略。

3.2.4 评价标准

在训练过程中，本研究使用准确率作为性能评估指标。通过验证集的准确率对模型进行评估，可以及时了解模型的性能表现，并在必要时调整训练策略。为了进一步提高推理性能，本研究还在每个训练周期结束时保存模型检查点，并在训练结束后加载最佳模型进行最终的评估。最终，微调后的模型在IMDB数据集上（通过情感标签转换为有害性标签）展现了良好的文本的有害性检测能力，为后续生成文本的有害性检测任务提供了高效、准确的深度学习模型。

3.3 实验结果与分析

3.3.1 实验结果

将微调后的DistilBERT模型作为检测文本有害性的检测模型应用于数据集HateAI-100。它由100个有害样例组成：50个由人类生成的样例和50个由大型语言模型生成的同义样例。为了评估模型的性能，本研究采用了准确率作为核心评价指标。准确率的具体定义如下：若检测模型针对某个样例输出了“negative”标签（即判定为有害），则视为检测正确；反之，若模型未能正确识别样例的有害性即输出“positive”标签，则视为检测失败。通过计算所有样例中检测正确的比例，最终得到了模型的准确率。检测结果如图2所示，从图中可以清晰地观察到以下结论：

1) 对于由人类生成的有害样例，微调后的DistilBERT模型展现出了出色的检测能力，准确率高达88%。这一结果表明，模型在识别和理解人类语言中的有害性方面具有较高的可靠性。

2) 对于由大型语言模型生成的有害样例，模型的检测准确率略低，为72%。这可能是由于大语言模型生成的文本在语法和结构上与人类生成的文本相似，但可能包含更隐蔽或复杂的有害内容，从而增加了检测的难度。

总体而言，微调后的DistilBERT模型在HateAI-100数据集上的检测准确率为80%。这一结果不仅验证了检测模型在文本的有害性检测任务中的有效性，也为我们后续的研究提供了重要的参考和依据。

3.3.2 结果分析

检测模型在检测大语言模型生成的有害文本时，相较于人类生成的有害文本，展现出了一定程度的准确率下降。产生这个现象可能有以下几个原

因：

1) 语言生成特性的差异：大语言模型与人类作者在文本生成方面展现出了显著的差异性。尽管这些模型能够生成在语法和结构上与人类文本高度相似的文本，但它们在词汇选择、句子结构以及写作风格上的细微差别，可能对模型的检测能力构成了挑战。特别是，大语言模型可能生成出看似无害但实则含有潜在有害性的内容，这些内容往往以更为微妙或复杂的方式表达，增加了模型识别的难度。

2) 有害内容定义的模糊：有害内容的定义本身存在一定的主观性和模糊性。人类标注者能够基于自身经验，较为准确地识别出人类作者所表达的有害意图。然而，对于大语言模型生成的文本，由于这些文本可能采用与人类作者不同的表达方式，因此人类标注者的理解可能变得更加困难。此外，大语言模型可能生成出具有潜在有害性的内容，这些内容可能涉及到复杂的语境、双关语或隐晦的讽刺等，这些都可能超出模型的检测能力范围。

3) 训练数据的局限性：模型训练数据的多样性对于其泛化能力至关重要。本研究推测训练数据中缺乏足够多的大语言模型生成的文本，可能影响了检测模型在这些文本上的检测性能。如果检测模型在训练过程中未能充分接触到这些文本，那么它可能无法有效地学习到这些文本的特性，从而导致了检测性能的下降。

4) 微调策略与方法：微调策略和方法的选择也可能对检测模型的检测性能产生显著影响。本研究考虑了微调过程中可能采用的策略和方法，如超参数设置、学习率调整等，这些都可能对检测模型在检测任务上的性能产生影响。如果微调策略过于偏向人类生成的文本而忽略了大语言模型生成的文本，那么检测模型可能在这些文本上的表现不佳。

为了进一步提高检测模型在大语言模型生成文本的有害性检测任务上的检测性能，本研究提出了以下改进措施：首先，增加训练数据中大语言模型生成文本的比例，以提高模型的泛化能力；其次，尝试采用更复杂的特征提取和表示方法，以更好地捕捉大模型生成文本中的潜在有害性；最后，引入外部知识库或利用多模态信息，可能有助于提升检测模型的检测性能。

4 结论和未来工作

当前研究揭示了大语言模型生成文本的有害性检测方面所面临的挑战。具体而言，尽管检测器在人类生成的有害文本上表现出色，但在处理大语言模型生成的类似文本时，其性能显著下滑。这一发现不仅强调了大语言模型在生成复杂、隐蔽有害内容方面的能力，也凸显了当前检测技术的局限性。针对这一现状，未来的研究应着重于以下几个方面：

1) 深化对大语言模型生成的有害样本的理解：为了更有效地检测大语言模型生成的有害文本，我们需要更深入地理解这些文本的特性和生成机制。这包括分析它们的语法结构、语义内容以及潜在的有害性模式。通过深化理解，我们可以为检测器的设计提供更准确的指导。

2) 扩充大语言模型生成的有害样本数据集：鉴于大语言模型生成文本的多样性，我们需要不断扩充有害样本库，以涵盖更多类型的有害内容。这将有助于检测器学习到更广泛的有害特征，从而提高其泛化能力。

3) 探索更多深度学习模型与算法：除了现有的深度学习模型外，我们还应探索其他可能更适合处理大语言模型生成文本的有害性检测任务的模型与算法。这包括但不限于基于注意力机制的模型、图神经网络等。通过对比不同模型与算法的性能，我们可以找到更优的解决方案。

4) 加强跨领域合作与知识共享：鉴于生成文本的有害性检测问题的复杂性，我们需要加强跨领域合作，如自然语言处理、计算机科学、心理学等。通过共享知识、资源和经验，我们可以共同推动这一领域的发展。

5) 构建可解释与可信赖的检测器：在追求检测器性能的同时，我们还应关注其可解释性和可信赖性。通过设计透明的检测流程和提供可验证的结果，我们可以增强用户对检测器的信任度，并为其在实际应用中的推广提供有力支持。

综上所述，未来的研究应致力于构建更加高效、准确且可信赖的文本的有害性检测器，以应对由大语言模型生成的有害文本所带来的挑战。这不仅需要我们在技术层面进行不断创新和优化，还需要我们加强跨领域合作与知识共享，共同推动这一领域的发展。

参 考 文 献

[1] Waseem Z, Hovy D. Hateful symbols or hateful people? predictive features for hate speech detection on twitter//Proceedings of the NAACL student research workshop. San Diego, USA, 2016: 88-93.

[2] Kwok I, Wang Y. Locate the hate: detecting tweets against blacks//Proceedings of the AAAI Conference on Artificial Intelligence. Bellevue, USA, 2013, 27(1): 1621-1622.

[3] Sarwar S M, Murdock V. Unsupervised domain adaptation for hate speech detection using a data augmentation approach//Proceedings of the International AAAI Conference on Web and Social Media. Atlanta, USA, 2022, 16: 852-862.

[4] Basile V, Bosco C, Fersini E, et al. Semeval-2019 task 5: multilingual detection of hate speech against immigrants and women in twitter//Proceedings of the 13th international workshop on semantic evaluation. Minneapolis, USA, 2019: 54-63.

[5] Kiela D, Firooz H, Mohan A, et al. The hateful memes challenge: detecting hate speech in multimodal memes. Advances in neural information processing systems, 2020, 33: 2611-2624.

[6] Davidson T, Warmsley D, Macy M, et al. Automated hate speech detection and the problem of offensive language//Proceedings of the international AAAI conference on web and social media. Montréal, Canada, 2017, 11(1): 512-515.

[7] Mandl T, Modha S, Majumder P, et al. Overview of the hasoc track at fire 2019: hate speech and offensive content identification in indo-european languages//Proceedings of the 11th annual meeting of the Forum for Information Retrieval Evaluation. Kolkata, India, 2019: 14-17.

[8] Kumar R, Ojha A K, Malmasi S, et al. Evaluating aggression identification in social media//Proceedings of the second workshop on trolling, aggression and cyberbullying. Marseille, French, 2020: 1-5.

[9] Fortuna P, Ferreira J, Pires L, et al. Merging datasets for aggressive text identification//Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying. Santa Fe, USA, 2018: 128-139.

[10] Mittos A, Zannettou S, Blackburn J, et al. "And we will fight for our race!" a measurement study of genetic testing conversations on Reddit and 4chan//Proceedings of the International AAAI Conference on Web and Social Media. Atlanta, USA, 2020, 14: 452-463.

[11] Dadvar M, Trieschnigg D, Ordelman R, et al. Improving cyberbullying detection with user context//Proceedings of the 35th European Conference on IR Research, Moscow, Russia, 2013.

[12] Chen Y, Zhou Y, Zhu S, et al. Detecting offensive language in social media to protect adolescent online safety//Proceedings of 2012 international conference on privacy, security, risk and trust and 2012 international confernece on social computing. Amsterdam, Netherlands, 2012: 71-80.

[13] Ortega F J, Troyano J A, Cruz F L, et al. Propagation of trust and distrust for the detection of trolls in a social network. Computer Networks, 2012, 56(12): 2884-2895.

[14] Jia M, Xie C, Jing L. Debiasing multimodal sarcasm detection with contrastive Learning//Proceedings of the AAAI Conference on Artificial Intelligence. Vancouver, Canada, 2024, 38(16): 18354-18362.

[15] Delbari Z, Moosavi N S, Pilehvar M T. Spanning the spectrum of hatred detection: a Persian multi-label hate speech dataset with annotator rationales//Proceedings of the AAAI Conference on Artificial Intelligence. Vancouver, Canada, 2024, 38(16): 17889-17897.

[16] Zhang J, Wu Q, Xu Y, et al. Efficient toxic content detection by bootstrapping and distilling large language models//Proceedings of the AAAI Conference on Artificial Intelligence. Vancouver, Canada, 2024, 38(19): 21779-21787.

[17] Balestriero R, Cosentino R, Shekkizhar S. Characterizing large language model geometry helps solve toxicity detection and generation//Proceedings of the forty-first International Conference on Machine Learning. Vienna, Austria, 2024.

[18] Lu J, Xu B, Zhang X, et al. Towards comprehensive detection of Chinese harmful memes//Proceedings of the Thirty-Eighth Annual Conference on Neural Information Processing Systems. Vancouver, Canada, 2024.

[19] Zhu Z, Zhuang X, Zhang Y, et al. Tfcd: Towards multi-modal sarcasm detection via training-free counterfactual debiasing//Proceedings of the International Joint Conference on Artificial Intelligence. Jeju, South Korea, 2024.

[20] Liu J, Feng Y, Chen J, et al. Prompt-enhanced network for hateful meme classification//Proceedings of the International Joint Conference on Artificial Intelligence. Jeju, South Korea, 2024.

[21] Lin H, Luo Z, Gao W, et al. Towards explainable harmful meme detection through multimodal debate between large language models//Proceedings of the ACM on Web Conference 2024. Singapore, Singapore, 2024: 2359-2370.

[22] Waldis A, Birrer J, Lauscher A, et al. The Lou dataset–exploring the impact of Gender-Fair language in German text Classification//Proceedings of the Conference on Empirical Methods in Natural Language Processing. Miami, USA, 2024.

[23] Masud S, Singh S, Hangya V, et al. Hate personified: investigating the role of LLMs in content moderation//Proceedings of the Conference on Empirical Methods in Natural Language Processing. Miami, USA, 2024.

[24] Kumar S, Mondal I, Akhtar M S, et al. Explaining (sarcastic) utterances to enhance affect understanding in multimodal dialogues//Proceedings of the AAAI Conference on Artificial Intelligence. 2023, Washington DC, USA, 37(11): 12986-12994.

[25] De la Peña Sarracén G, Rosso P, Litschko R, et al. Vicinal risk minimization for few-shot cross-lingual transfer in abusive language detection//Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. Singapore, 2023: 4069-4085.

[26] Maity K, Jain R, Jha P, et al. GenEx: A commonsense-aware unified nenerative framework for explainable cyberbullying detection//Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. Singapore, 2023: 16632-16645.

[27] Albanyan A, Blanco E. Pinpointing fine-grained relationships between hateful tweets and replies//Proceedings of the AAAI Conference on Artificial Intelligence. Philadelphia, USA, 2022, 36(10): 10418-10426.

[28] Deng Y, Zhang W, Pan S J, et al. Multilingual jailbreak challenges in large language models//Proceedings of the International Conference on Learning Representations. Vienna, Austria, 2024

[29] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: pre-training of deep bidirectional Transformers for language understanding //Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis, USA, 2019: 4171–4186.