****

**技术报告**

**基于MindSpore和mPLUG-Owl的多模态情感计算与分析在网络暴力检测中的应用**

参赛队名： 笨萝卜炖牛肉

队 长： 张嘉慧

联系方式： 18927008880

组员名单： 王诗雁

指导老师： 刘寿强

参赛单位：华南师范大学阿伯丁数据科学与人工智能学院

作品类别**：** 项目贡献赛道

作品链接：https://gitlink.org.cn/weixian/MindSpore.git

**摘要**：基于MindSpore框架和先进的mPLUG-Owl多模态大规模预训练模型[1]，本项目旨在开发一种高效且全面的网络暴力检测系统。与仅能处理文本信息的传统方法相比，本系统能够理解和分析图像和文本的复杂交互，大大提高了网络暴力检测的准确性和全面性。本项目具有重要的社会和公益价值，符合国家和党的网络治理政策和方向，为网络暴力信息治理提供了参考。特别地，本项目进一步丰富了在多模态网络暴力检测方面相对有限的国内外研究，拓宽了大规模预训练模型在社会应用场景中的应用范围，同时也提高了公众对精神暴力和心理问题的认识和重视。

**关键词：** Mindspore, mPlug-Owl, 多模态 ，网络暴力

**目录**

[1 设计方案 1](#_Toc147240003)

[1.1 需求分析 1](#_Toc147240004)

[1.2 国内外研究现状 2](#_Toc147240005)

[1.3 设计目标 3](#_Toc147240006)

[2 实现方案 4](#_Toc147240007)

[2.1 项目设计流程图 4](#_Toc147240008)

[2.2.2 网络脚本分析 5](#_Toc147240009)

[2.2.3 网络脚本开发 6](#_Toc147240010)

[2.2.4 网络执行调试 7](#_Toc147240011)

[2.2.5 网络精度性能调优 8](#_Toc147240012)

[3 实验数据与结果 8](#_Toc147240013)

[3.1 项目代码链接 8](#_Toc147240014)

[3.2 实验展示 9](#_Toc147240015)

[3.3 预测作品展示 9](#_Toc147240016)

[4特色创新 10](#_Toc147240017)

[5 总结与展望 10](#_Toc147240018)

[5.1 总结 10](#_Toc147240019)

[5.2 展望 11](#_Toc147240020)

[5.2.1短期展望 11](#_Toc147240021)

[5.2.2中期展望 11](#_Toc147240022)

[5.2.3长期展望 11](#_Toc147240023)

[6 参考文献 12](#_Toc147240024)

# 1 设计方案

## 1.1 需求分析

1. 功能需求
2. **多模态检测能力**: 系统需要能够同时处理文本和图像信息，从而更准确地识别网络暴力。
3. **高准确率和全面性**: 系统应具备较高的检测准确率，并能全面地分析和识别各种形式的网络暴力，包括但不限于言辞攻击、恶意图片和欺凌行为。
4. **数据安全与合规性**: 在数据收集和处理过程中，需要遵守相关的数据保护和隐私法律。
5. **自动识别与报警**: 随着技术的进步，数字/社交媒体的使用日益增加。 二十一世纪的人们是在一个拥有社交媒体的互联网世界中长大的。只需单击一下按钮即可进行通信。 尽管数字媒体有很多机会，但人们往往会滥用它。 人们在社交网络中传播对某人的仇恨。 网络欺凌从不同方面影响人们。 它不仅仅影响健康，还有更多不同的方面会对生命造成威胁。 网络欺凌是一种世界性的现代现象，人类无法百分百避免，但可以预防[2]。需要设计自动网络欺凌检测方法，以避免网络骚扰造成的不良后果[3]。系统的发展方向应能自动识别网络暴力行为并实时发出警报，以便采取及时的干预措施。
6. 性能需求
7. **实时处理**: 系统需要具备实时处理能力，以满足网络环境中信息传播速度快的特点。
8. 可扩展性: 考虑到网络环境的复杂性和多样性，系统应具备良好的可扩展性。
9. 高可用性: 系统应能在高负载下稳定运行，具备故障自恢复能力。
10. 社会与政策需求
11. 符合政策法规: 项目需要符合《网络暴力信息治理规定（征求意见稿）》[4]和其他相关政策。
12. **教育与宣传**: 网络骚扰因其强烈的负面社会影响而备受关注[5]。除了技术应用外，项目还应当有助于提高公众对网络暴力，特别是精神和心理暴力的认识和重视。

## 1.2 国内外研究现状

1. 国内研究现状
2. 研究焦点：国内关于网络暴力检测的研究主要集中在**自然语言处理（NLP）的传统领域**。这些研究大多采用机器学习或深度学习算法，如支持向量机（SVM）和卷积神经网络（CNN），用于文本的语义或情感分析。
3. 技术局限：国内研究普遍**没有涉及大型自然语言模型（LLM）或预训练模型的应用**。这意味着目前的方法通常仅基于手工特征工程和规则匹配，缺乏模型的泛化能力和多模态处理能力。
4. 研究空白：随着社交网络成为人们生活的重要组成部分，网络欺凌成为青少年在社交网络平台上最常见的风险，并引起社会的严重关注。 在过去的几十年里，大多数有关网络欺凌的现有工作都集中在文本分析上。 但网络欺凌已呈现多目标、多渠道、多形式的发展趋势。 传统的文本分析方法无法满足社交网络中欺凌数据的多样性[6]。当前，在国内几乎没有公开的研究成果明确地涉及到使用LLM或预训练模型在网络暴力检测方面的应用。因此，有**大量的研究空白**等待填补，特别是在多模态网络暴力检测和自动识别方面。
5. 国外研究现状
6. 研究趋势：国外有若干研究尝试使用LLM或预训练模型，如BERT，来进行网络暴力检测。最近，基于深度神经网络的模型在检测网络欺凌方面比传统模型有了显着的改进。 此外，新的、更复杂的深度学习架构正在开发中，事实证明这些架构在各种 NLP 任务中非常有用。 谷歌研究人员最近开发了一种名为 BERT 的语言学习模型，它能够生成上下文嵌入，还能够生成用于分类的任务特定嵌入。 提出了一种新的社交媒体平台网络欺凌检测方法，该方法使用新颖的预训练 BERT 模型，顶部有单个线性神经网络层作为分类器，该方法改进了现有结果[7]。这些研究相对于国内研究在模型复杂性和技术前瞻性方面有所先进。
7. 技术瓶颈：然而，这些研究大多仅支持**单模态（**主要是文本）的输入，并且缺乏自动识别和实时干预的功能。此外，很多研究仍然使用**相对陈旧的预训练模型**，如BERT，没有充分挖掘LLM的潜能。
8. 研究局限：国外的研究在技术应用方面尚未成熟，多数还停留在理论探究或初步的应用阶段。尤其是在多模态数据处理和实时网络暴力识别方面，仍有大量的研究空间。

总体而言，无论是国内还是国外，当前的研究都存在**明显的局限性**和不足，尤其是在**多模态网络暴力检测**和自动识别方面。这为本项目提供了宝贵的研究和应用机会，具有**重要的社会价值和技术前景**。

## 1.3 设计目标

1. 保护受害者
2. 实时识别与响应: 利用高度准确的多模态暴力检测技术，对网络暴力行为进行实时识别，并即刻触发警报系统。
3. 个性化支持与援助: 为受害者提供定制化的心理支持和法律援助资源，降低网络暴力对其造成的伤害。
4. 遏制恶意信息传播
5. 信息过滤与屏蔽: 通过先进的算法，在信息发布前或发布后进行审核，有效阻止含有谣言、辱骂、恶意攻击等不良内容的信息传播。
6. 跟踪与追责: 为法律部门提供准确的数据支持，帮助追踪并处罚网络暴力行为的始作俑者。
7. 维护良好网络环境
8. 用户教育与宣传: 利用多媒体教育材料和在线培训，提升公众对网络文明和网络安全的认知。
9. 社区激励机制: 通过积分、徽章等激励机制，鼓励用户举报网络暴力和参与维护网络环境。
10. **对接国家政策与社会需求**
11. 法规依据与合规性: 确保项目的所有操作和流程都符合《网络暴力信息治理规定（征求意见稿）》及其他相关法规。
12. 贯彻党和国家指导方针: 通过本项目的实施，积极响应**党的二十大**和2023年“清朗”系列专项行动的精神，为形成良好网络生态作出贡献。
13. **理论与实践推进**
14. 推动理论创新: 攻击性语言检测对于维护文明的社交媒体平台和部署预先训练的语言模型越来越重要。 然而，这项中文任务仍在探索中[8]。**填补多模态网络暴力检测在国内外研究中的空白**，为网络暴力研究提供新的理论基础。
15. 实践意义与应用拓展: 通过本项目，不仅在社会实践中找到有效解决方案，还为大规模预训练模型在其他**社会应用场景中的推广**提供了借鉴。

通过达成以上设计目标，本项目旨在全方位、**多维度**地解决网络暴力问题，具有高度的实用性和前瞻性。

# 2 实现方案

## 2.1 项目设计流程图

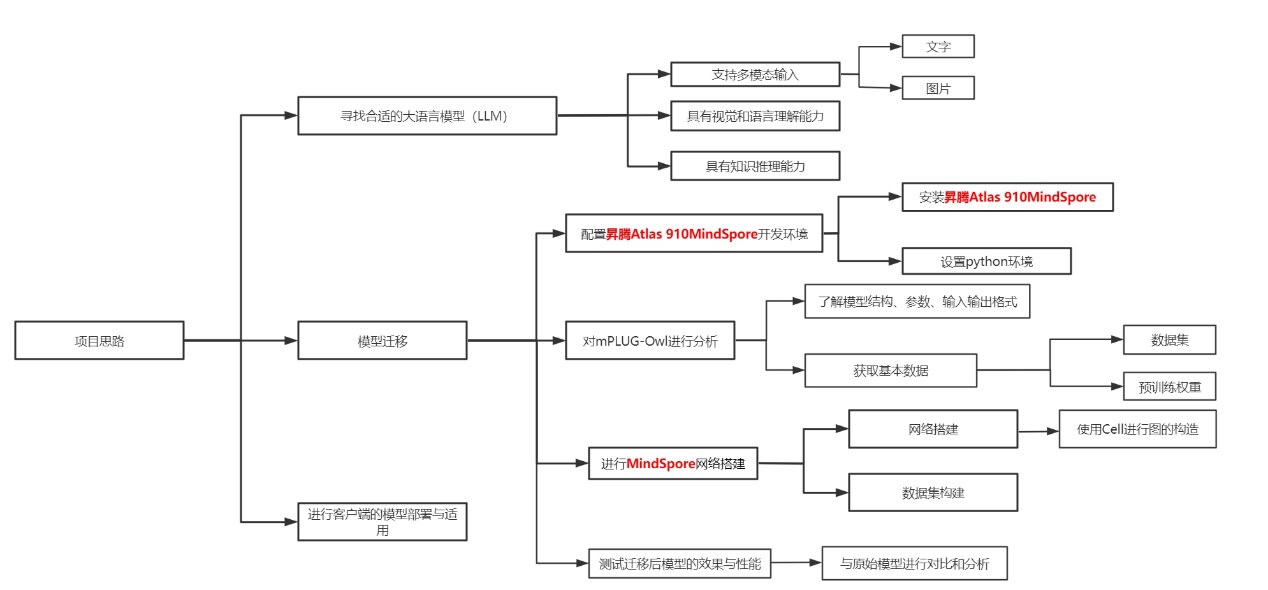


图1：项目设计流程图

本项目设计分为三个步骤，首先我们经过多次单论对话，多轮对话等语言质量比对，选取了mPLUG-Owl预训练大语言模型[1]，该模型支持多模态输入，具有视觉和语言的理解能力，并具有知识推理能力，在网络暴力图片与文字检测上表现良好。接着，我们对基于pytorch的模型迁移[9]至mindspore神经网络框架。我们需要配置晟腾Atlas910Mindspore开发环境[10]，然后对mPLUG-Owl进行分析[11]，包括模型结构，参数以及API覆盖度等，对模型进行基于Mindspore的网络搭建。最后测试迁移模型的效果并部署在客户端。

### 2.2.2 网络脚本分析

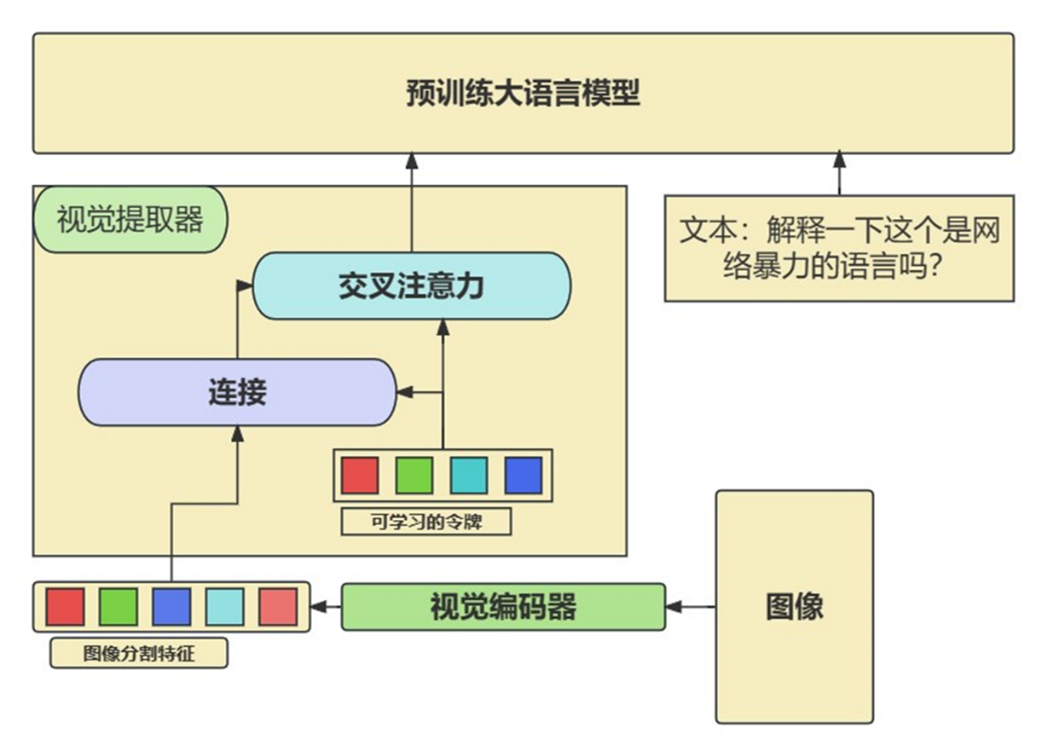


图2：项目脚本分析

它由视觉基础模块[1]（开源的ViT-L）、视觉抽象模块以及大型语言模型（LLaMA-7B）组成。

视觉抽象模块将较长的、细粒度的图像特征概括为少量可学习的 Token，从而实现对视觉信息的高效建模。

生成的视觉 Token 与文本查询一起输入到语言模型中，以生成相应的回复。

### 2.2.3 网络脚本开发

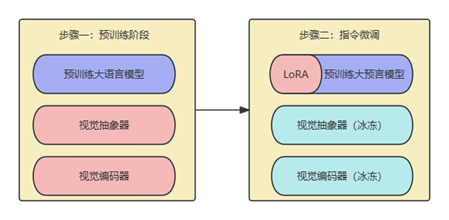


图3：项目脚本开发过程

项目脚本开发过程分为以下几步：

首先，其使用多模态数据训练视觉模块，并冻结语言模块。这样可以让视觉特征贴合语言特征。

然后，其使用多模态和单模态数据联合调整语言模块的LoRA参数，同时冻结视觉模块。

这样，模型可以学习多样化的单模态和多模态指令，同时具备单模态和多模态多轮对话能力。

### 2.2.4 网络执行调试

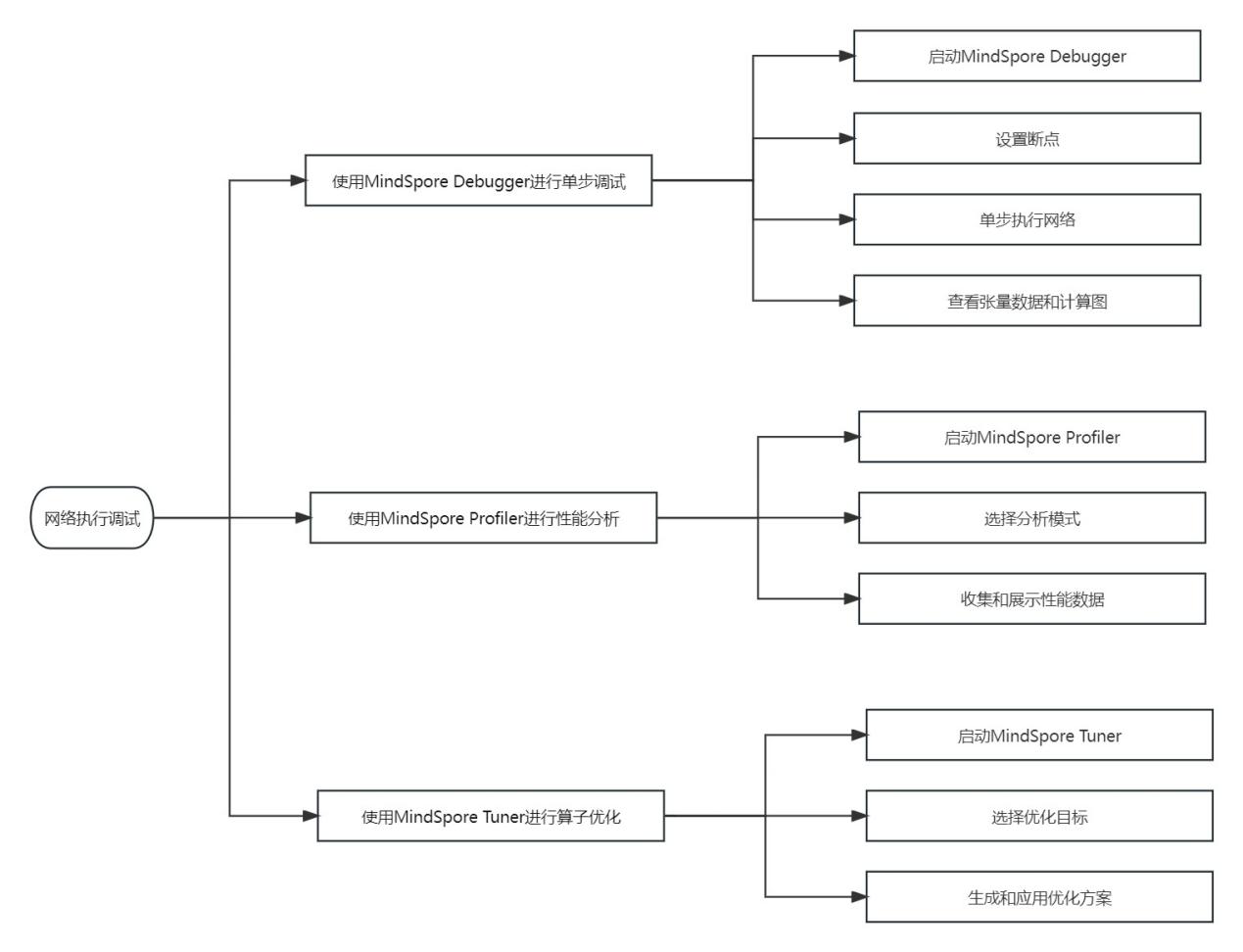


图4：网络执行调试过程

网络执行调试分为以下几步[9]：

1. 单步调试：使用MindSpore Debugger工具进行单步调试，可以查看网络中每个节点或算子的输入输出张量数据和计算图结构，以及设置断点、条件断点、观察点等功能，方便用户定位和解决网络中的错误或异常。
2. 性能分析：使用MindSpore Profiler工具进行性能分析，可以收集并展示网络运行时的各种性能数据，如运行时间、内存占用、算力利用率、数据传输等，以及提供多种分析模式，如时间线分析、算子分析、数据并行分析等，帮助用户评估和提升网络的性能和效率。
3. 算子优化：使用MindSpore Tuner工具进行算子优化，可以根据用户指定的优化目标，如运行时间、内存占用、算力利用率等，自动搜索并生成最优的算子实现方案，并应用到网络中，实现算子级别的优化。

### 2.2.5 网络精度性能调优

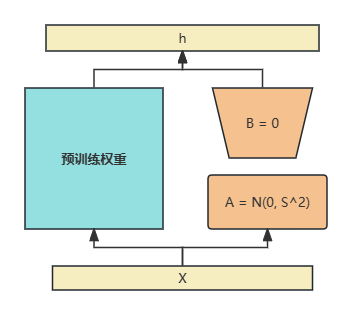


图5：Lora方式微调示意图

Lora[12]是一种低秩自适应（Low-Rank Adaptation）的方法，它可以用少量的参数来微调大型的预训练语言模型，从而适应不同的下游任务，如文本分类、生成、推理等。

Lora的优点是：

1. 它可以显著减少微调所需的参数量和显存占用，例如，对于一个175亿参数的GPT-3模型，Lora只需要更新约200万个参数。
2. 它不会增加推理延迟，因为在推理过程中，只需要计算 W=W0​+ΔW 即可。
3. 它可以保持预训练模型的泛化能力和知识，因为它不会改变预训练模型的参数。

Lora的效果是：

1. 在多个自然语言理解和生成任务上，Lora可以达到与全参数微调（Fine-Tuning）相当甚至更高的性能，而且使用更少的数据和时间。

# 3 实验数据与结果

## 3.1 项目代码链接

https://gitlink.org.cn/weixian/MindSpore.git

## 3.2 实验展示

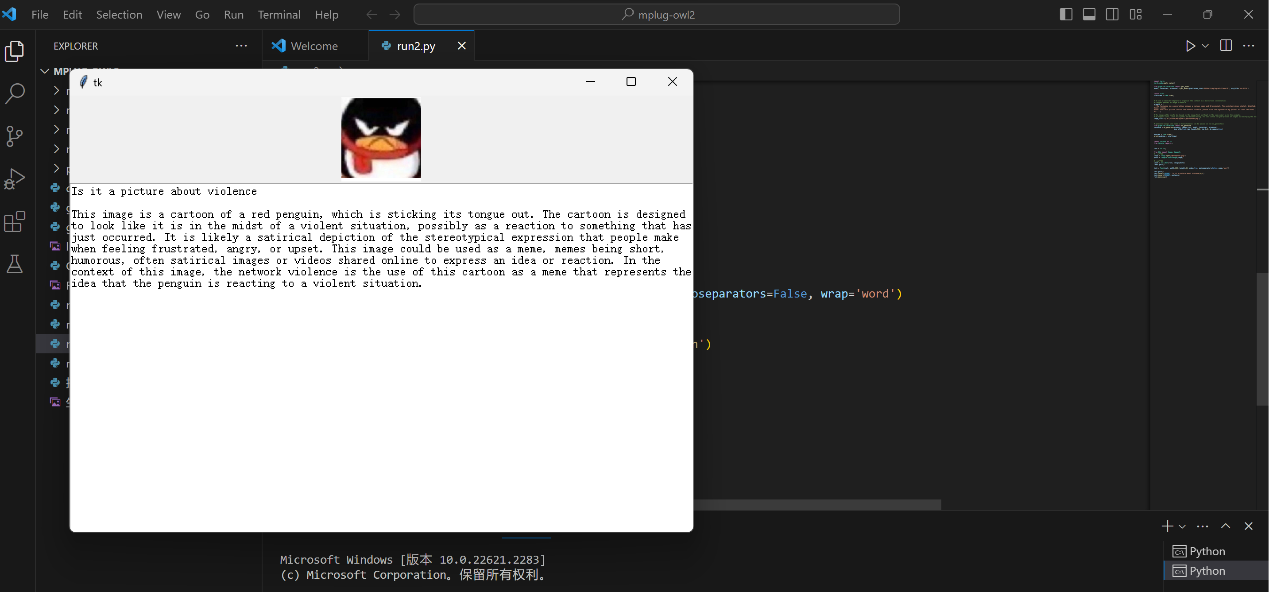


图6：实验演示示意图

## 3.3 预测作品展示



图7：预测作品展示示意图

# 4特色创新

1. 创新性应用**多模态大规模预训练模型** (LLM)

高级模型选用: 通过采用mPLUG-Owl模型，项目实现了先进的视觉和语言信息处理，以及多轮对话和知识推理能力，使得网络暴力检测更为精准和全面。

1. 高效稳定的系统架构

框架迁移: 利用**MindSpore框架**的高性能和易用性，大大提高了网络暴力检测系统的运行效率和稳定性。

1. **研究空白的填补和扩展**

多模态网络暴力检测: 项目填补了目前在国内外网络暴力检测研究中多模态信息处理的空白，利用图像等多种非文字信息辅助网络暴力判断，从而拓宽了网络暴力检测的方法论。

1. 应用场景和社会影响的扩展
2. LLM技术在**社会应用的推广**: 本项目不仅提供了使用大规模预训练模型进行网络暴力检测的实证案例，也为LLM技术在其他社会应用场景提供了有力的借鉴。
3. 重视非物理性暴力: **不同于主流研究**多聚焦于**物理暴力**，本项目通过构建网络暴力检测系统，显著提升了公众对心理和**精神暴力**问题的认识和重视。

总体来说，本项目在技术应用、研究领域和社会影响三方面都表现出高度的创新性和前瞻性，为解决网络暴力这一日益严重的社会问题提供了有效的解决路径。

# 5 总结与展望

## 5.1 总结

通过应用先进的多模态LLM技术和MindSpore框架，本项目不仅在网络暴力检测方面实现了技术创新，也成功拓宽了大规模预训练模型在国内应用的范畴。它突破了传统单模态检测的局限，实现了更全面和精准的暴力行为识别。此外，该项目强调了心理暴力与身体暴力同等重要，提升了社会对网络暴力多维性的认识和警觉性。

该项目不仅符合最新的《网络暴力信息治理规定（征求意见稿）》和党的二十大报告中提出的网络治理方针，还具有强烈的社会责任感和实践价值，为推动我国网络环境向好提供了有力的技术支持。

## 5.2 展望

### 5.2.1短期展望

1. 技术层面
2. **LoRA方式指令微调**：计划使用LoRA（Low-Rank Approximation）技术对预训练模型进行指令微调。通过确定范围的网络暴力数据，涵盖种族、性别和地区等不同主题[5]，进行模型调优，以提高网络暴力检测的准确性和准确度。
3. **自动检测与实时监控**：将完成的模型部署到客户端或云端，实现实时的网络暴力信息监测和过滤。
4. **个性化设置**：根据用户需求和偏好，实现自动检测功能的个性化设置，如敏感度调节、反馈方式选择、处理方式等。
5. 社会层面
6. 推广与合作：计划与各大社交媒体平台、学校、企业和政府机构合作，将项目的成果推广到更广泛的应用场景。
7. 公众教育与宣传：通过线上线下的讲座、培训和社会活动，提升公众对网络暴力问题，尤其是心理暴力问题的认识。

### 5.2.2中期展望

1. 模型优化与扩展：计划进一步研究和开发，例如引入**时间序列分析**以**跟踪和预测网络暴力趋势**，或者引入**地理位置数据**以分析和防止**特定区域的网络暴力行为**。
2. 执法与政策合作：积极参与和推动与网络暴力相关的执法工作，以科技手段辅助和加强法律制度的落实。具体内容包括：
3. 与有关部门沟通协调，建立**网络暴力案件的信息共享**、快速反馈、联合处置机制，提高网络暴力案件的受理、审理、执行效率。
4. 与**网络服务提供者**合作，推动其履行信息网络安全管理义务，加强对网络暴力信息的**监测、过滤、删除**等工作，规范网络平台的申诉渠道和惩处机制。

### 5.2.3长期展望

1. 国际化：将研究成果和技术推广到全球范围，与国际组织合作，共同对抗网络暴力。
2. 可持续性和伦理：在保证技术不断进步的同时，**注重其社会、伦理和环境影响**，以实现项目的**可持续发展**。
3. 综合治理：计划将网络暴力检测技术**与其他社会治理手段相结合**，例如心理咨询、教育培训等，以形成一个更加全面和高效的网络暴力治理体系。具体来说，为了实现这一目标，我们将：
4. 与政府部门、司法机构、教育机构、心理机构等建立合作关系，共同打造一个**涵盖预防、监测、干预、惩戒、恢复等环节的网络暴力治理平台**。
5. 我们将利用我们的技术**为网络暴力受害者提供及时有效的帮助**，如删除或屏蔽有害信息、提供法律援助、开展心理疏导等。
6. 我们还将利用我们的技术**为网络暴力施加者提供必要的教育和引导**，如提醒或警告其不良行为、推送正向信息或课程、转介专业机构进行干预等。
7. 此外，我们还将利用我们的技术**为广大网民提供网络安全教育和素养培训**，如普及网络法律知识、提高网络道德意识、增强网络自我保护能力等。

通过这些短期、中期和长期的展望，我们希望本项目不仅能有效地解决当前网络暴力的问题，而且能在未来为维护网络安全和社会和谐做出更多贡献。

# 6 参考文献

1. Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, Chenliang Li, Yuanhong Xu, Hehong Chen, Junfeng Tian, Qian Qi, Ji Zhang, Fei Huang. mPLUG-Owl: Modularization Empowers Large Language Models with Multimodality, 2023.
2. Mohamed Mostafa Soliman; Mohamed Hussein Kamal; Mina Abd El-Massih Nashed. Violence Recognition from Videos using Deep Learning Techniques, 2019.
3. Perera, Andrea; Fernando, Pumudu. Accurate Cyberbullying Detection and Prevention on Social Media [J]. Procedia Computer Science, 2021, 181:605-611.
4. 《最高人民法院、最高人民检察院、公安部关于依法惩治网络暴力违法犯罪的指导意见（征求意见稿）》
5. Md Manowarul Islam, Md Ashraf Uddin, Selina Sharmin, Arnisha Akter, Linta Islam, Uzzal Kumar Acharjee. Cyberbullying Detection on Social Networks Using Machine Learning Approaches[C]. In: 2020 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE). IEEE, 2020. DOI: 10.1109/CSDE50874.2020.9411601.
6. Wang, K., Xiong, Q., Wu, C., Gao, M., & Yu, Y. 2020. Multi-modal cyberbullying detection on social networks. State Key Laboratory of Power Transmission Equipment and System Security and New Technology, School of Big Data and Software Engineering, Chongqing University. Email: kgwang@cqu.edu.cn, [cquxqy@163.com](mailto:cquxqy@163.com).
7. Yadav, J., Kumar, D., & Chauhan, D. Cyberbullying Detection using Pre-Trained BERT Model [C]. In Proceedings of the International Conference on Electronics and Sustainable Communication Systems (ICESC 2020), IEEE Xplore Part Number: CFP20V66-ART; ISBN: 978-1-7281-4108-4. Delhi Technological University, 2020. Email: jdp02041997@gmail.com, deveshkumarkblock@gmail.com, chauhan.dheerajll@gmail.com.
8. Jiawen Deng, Jingyan Zhou, Hao Sun, Chujie Zheng, Fei Mi, Helen Meng, Minlie Huang. COLD: A Benchmark for Chinese Offensive Language Detection. The CoAI group, DCST, Institute for Artificial Intelligence, State Key Lab of Intelligent Technology and Systems, Beijing National Research Center for Information Science and Technology, Tsinghua University, Beijing, China; Dept. of Systems Engineering & Engineering Management, The Chinese University of Hong Kong; Huawei Noah’s Ark Lab, 2022.
9. Mindspore网络迁移: <https://www.mindspore.cn/docs>
10. 海思晟腾：<https://www.hisilicon.com/cn/products/Ascend/Ascend-910>
11. Mindsporw模型库<https://www.mindspore.cn/docs/zh-CN/r2.1/note/official_models.html>
12. Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen. LoRA: Low-Rank Adaptation of Large Language Models, 2021.
13. LoRA: <https://zhuanlan.zhihu.com/p/646791309>, <https://zhuanlan.zhihu.com/p/611557340>