****

**大学生创新训练计划**

**创新类项目申报书**

**项目名称：基于大模型辅助的多模态委婉有害内容的检测技术研究**

**项目负责人： 郑煬阳**

**所在学院： 网络空间安全学院**

**专业年级： 2024级**

**学 号： 2024141530142**

**手 机： 15808310028**

**电子邮箱： 2053900282@qq.com**

**指导教师： 王海舟**

**项目起止年月： 2025年11月至2026年10月**

**项目参与学生人数： 5人**

**四川大学教务处制**

年 月

填写说明

一、凡申报**四川大学“大学生创新训练计划”**必须填写本申报书。创新类项目是本科生个人或团队，在导师指导下，自主完成创新性研究项目设计、研究条件准备和项目实施、研究报告撰写、成果（学术）交流等工作。

**二、“项目所属一级学科和代码”**参考《普通高等学校本科专业目录和专业介绍（2012年）》。

三、**“项目开展支撑平台”**指支撑本项目开展的国家级和省部级重点实验室（中心、平台等）、国家双创示范基地平台、教学实验中心（实验室）、企业、事业或其他单位等，表中填写平台名称，可以多个。

四、**“项目组成员”**人数原则上不超过五人，应排序。

五、**“项目成熟度**”请参考附件《项目成熟度量表》。

六、本书应该填写完整、内容详实、表达准确，数字一律填写阿拉伯数字。

七、报送申报书的电子文档至负责人所在学院。

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 项目名称 | 基于大模型辅助的多模态委婉有害内容的检测技术研究 | | | | | |
| 项目属性 | ☑面上项目 □人工智能+新质战略育苗”（含2035特区子计划项目）  □交叉学科子计划项目 | | | | | |
| 申请类别 | ☑科学探索与工程技术类 □人文艺术与社会科学类  □软件信息与文创类 □智能装备与医疗器械类  □生物医药与新材料类 | | | | | |
| 申请经费 | 元 | | 起止时间 | | 2024年11月至2025年10月 | |
| 项目所属 一级学科和代码 | 工学 08 | | | | | |
| 项目开展 支撑平台 |  | | | | | |
| 项目来源  （可多选） | □十大重点支持领域的项目  ☑进课题组、进实验室、进科研团队参与的项目  □国家级和省部级重点实验室（中心、平台等）、国家双创示范基地平台支持申报项目  □交叉学科创新项目  □“青年红色筑梦之旅”计划项目  □基于前期研究实践成果、继续深入研究实践的创新项目  □高水平课题  □其他 | | | | | |
| 高水平课题名称（非高水平课题可不填） | 命题名称 |  | | | | |
| 校内指导老师姓名（非交叉学科子计划项目一般仅允许一位指导老师） |  | | | | |
| 所属重点支持领域（可不选） | 选择1项：E  A.不填  B.泛终端芯片及操作系统应用开发  C.重大应用关键软件  D.云计算和大数据  E.人工智能  F.无人驾驶  G.新能源与储能技术  H.生物技术与生物育种  I.绿色环保与固废资源化  J.第五代通信技术和新一代IP网络通信技术  K.社会事业与文化传承 | | | | | |
| 负责人之前参与大创项目情况 | 格式如下，有则填，不限条目：  1. 负责人/团队成员，项目编号，项目名称，立项级别，项目类别，立项年份，结题成绩；  2. …… | | | | | |
| 项目成员之前参与大创项目情况 | 1. 姓名，负责人/团队成员，项目编号，项目名称，立项级别，项目类别，立项年份，结题成绩；  2. …… | | | | | |
| 项目负责人基本信息 | | | | | | |
| 姓名 | 学号 | 专业年级 | | | 所在学院 | |
|  |  |  | | |  | |
| 性别 | 手机 | 电子邮箱 | | | 身份证号 | |
|  |  |  | | |  | |
| 项目组成员基本信息 | | | | | | |
| 序号（含排序） | 1 | 2 | | 3 | 4 |
| 姓名/性别 | 彭实/男 | 孙浩哲/男 | | 匡家辉/男 | 李佳玥/女 |
| 学号 |  |  | |  |  |
| 专业年级 | 网络空间安全2024级 | 网络空间安全2024级 | | 网络空间安全2024级 | 网络空间安全2024级 |
| 所在学院 | 网络空间安全学院 | 网络空间安全学院 | | 网络空间安全学院 | 网络空间安全学院 |
| 手机 |  |  | |  |  |
| 电子邮箱 |  |  | |  |  |
| 身份证号 |  |  | |  |  |
| 签名 |  |  | |  |  |
| 指导教师1 基本信息  （非交叉学科子计划项目一般仅允许一位指导老师） | | | | | | |
| 姓名 | 所在学院或单位 | 研究方向 | | | 职称/职务 | |
|  |  |  | | |  | |
| 性别/年龄 | 手机 | 电子邮箱 | | | 签名 | |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |
| 指导教师2 基本信息  （交叉学科子计划项目需填写第二指导老师） | | | |
| 姓名 | 所在学院或单位 | 研究方向 | 职称/职务 |
|  |  |  |  |
| 性别/年龄 | 手机 | 电子邮箱 | 签名 |
|  |  |  |  |

|  |
| --- |
| **项目摘要(限200字以内)** |
| 随着社交媒体发展，有害内容常以**委婉语**、**隐喻图像**、**音频暗语**等隐蔽形式传播，**传统单模态审核手段难以应对**，**多模态协同识别的核心问题尚未有效解决**，严重影响网络环境健康。  本项目聚焦**多模态检测技术**突破，融合文字、图像与音频的特征检测，构建**多模态检测大模型**，深度解析文本隐藏含义、图像敏感元素及音频违规信息，通过**跨模态特征融**合实现精准识别，并依托增量学习持续优化模型，为**营造清朗网络环境**提供关键技术支撑。 |
|  |
|  |
| **特色创新点（限100字以内，建议2-3点）** |
| 1.对**全球最大的中文社交网络平台**开展**委婉语违规内容检测研究**，构建和训练基于**多模态数据处理技术对于文字、图片、音频**的多维违规委婉语检测技术模型；  2.“**双模型协同**” 的多模态数据处理与标注方案，**突破传统单模态局限**，实现更加准确的检测。 |

**报告正文**

|  |
| --- |
| 1. **立项依据与研究内容**（**建议8000字以下**（不包括文献）） |
| 1．**项目的立项依据**（【理工医科】研究意义、国内外研究现状及发展动态分析，需结合科学研究发展趋势来论述科学意义；或结合国民经济和社会发展中迫切需要解决的关键科技问题来论述其应用前景。附主要参考文献目录【哲学社会学科】国内外相关研究的学术史梳理及研究动态;本课题相对于已有研究的独到学术价值和应用价值等。附主要参考文献目录） |
| **1.1研究背景**  **1.1.1背景综述**  在人工智能技术与互联网应用深度融合的当下，以生成式 AI 为核心的多模态内容创作已成为网络信息传播的主流形态。文本、图像、音频、视频等多模态内容凭借丰富的表现力与高效的传播性，重塑了公众的信息交互方式，推动了数字经济、文化传播等领域的创新发展。  与此同时，生成式 AI 技术的低门槛化与多模态内容的隐蔽性特征，为“委婉有害内容”的滋生与传播提供了温床。不同于传统直白的有害信息，委婉有害内容通过语义替换、符号隐喻、多模态转译等手段，规避现有审核机制的检测，呈现出“形式合规化、含义隐蔽化、传播扩散化”的新特征：在文本层面，利用谐音、拼音缩写、语义联想等方式，将侮辱性、攻击性含义包裹在日常化表述中；在图像层面，通过“文本 - 图像转译”策略，先将敏感词汇谐音化或隐喻化，再以该表述作为提示词驱动生成式 AI生成看似无害的图像，实现有害含义的视觉化隐藏；在音频与视频层面，谐音辱骂、节奏变形、画面隐喻等手段普遍存在，甚至出现“多模态协同规避”现象 —— 例如视频中画面展示隐喻图像、音频使用谐音话术、字幕嵌入语义替换文本，形成跨模态的有害信息传播闭环。  此类多模态委婉有害内容的传播，已对网络空间治理、公众信息安全乃至社会稳定产生深远影响。一方面，其隐蔽性导致现有基于关键词匹配、单一模态特征识别的审核技术失效，使得有害信息能够在社交平台、短视频平台等渠道广泛扩散，误导公众认知、煽动负面情绪；另一方面，随着国家级网络安全战略对有害信息治理要求的不断提升，如何精准识别多模态场景下的委婉有害内容，已成为保障公民信息权益、维护网络空间清朗生态、落实国家网络安全战略的关键技术需求。在此背景下，开展多模态委婉有害内容检测技术研究，既是应对当前网络安全新挑战的必然选择，也是推动 AI 技术在网络空间治理领域合规应用、服务国家级网络安全保障体系建设的重要举措。  **1.1.2存在问题**  近年来，违规内容多以委婉语的形式出现以逃避平台捕捉，尤其是在各种社交平台上，如抖音、微博、小红书等社交应用。这种违规委婉语的出现形式多样，通常利用文字的语言歧义、文字中穿插大量乱码或者带有侮辱信息的图片来传达违规内容，导致应用平台内社交矛盾激化，不利于营造清朗的网络环境。  日程表  AI 生成的内容可能不正确。  图1-2 利用文字与图片在抖音私信发表违规内容  进行多模态违规委婉语检测主要涉及以下技术难题：   1. 动态演变特征导致检测滞后：多模态委婉语在现代社会的更迭中，新的组合方式层出不穷，使检测系统始终处于追赶状态，例如“唐人”在十年前的网络中指唐人街、唐朝的人、或者借指中国人，可是在近年的网络中，“唐人”又有了“唐氏综合症患者”的含义，而在不停的传播中，又被赋予了“肢体肥大”、“智商低下”等语义，后被引申成了侮辱他人的含义。 2. 跨模态语义融合与对齐难题：多模态违规内容常将文字委婉语与图片、表情包、视频片段等非文本信息结合，单一模态的语义解读无法完整还原违规意图，跨模态信息的融合对齐成为关键障碍。语义碎片化分布：违规信息可能拆分在不同模态中，例如文字用 “懂得都懂” 配合一张带有侮辱性元素的表情包，文字本身无违规性，需结合图像语义才能判断，而不同模态的语义关联缺乏明确规律。模态间语义冲突与补充：部分场景下，图像与文字语义存在矛盾或补充，例如文字描述 “可爱的小天使”，搭配的却是丑化、歧视特定群体的图片，系统需区分这种 “正话反说” 的跨模态讽刺，难度极大。 3. 对抗性规避手段的动态升级：为逃避检测，违规内容发布者会主动研究平台检测规则，不断升级规避手段，形成 “检测 - 规避 - 再检测” 的对抗循环，使检测系统难以应对。文字层面的对抗变异：除了基础的谐音、拆分，还会通过生僻字替换（如用 “砼” 替代 “同” 组成侮辱性词汇）、拼音首字母 + 特殊符号组合（如 “nmsl” 中间插入 “@#” 变成 “n@m#s#l”）、方言谐音转写（用方言发音的文字替代违规词汇）等方式规避文字检测。多模态联合对抗：将违规信息隐藏在多模态内容的细节中，例如在视频背景角落嵌入微小的侮辱性文字图片、用特定节奏的音乐配合模糊的文字暗示违规内容，或通过 “图文错位”（文字描述 A，图片细节隐含 B 类违规信息）的方式降低检测概率。   随着人们聊天模式的不断迭代，传统的平台审核模式已不足以应对这些新兴的威胁。未来的研究应聚焦于提升违规委婉语内容检测的准确性和实时性，并通过多模态融合、并借助大模型，从而营造清朗的网络环境。  **1.2研究意义**  **1.2.1学习意义**  本项目组成员通过此次项目实践，可以拓展在计算机领域内的知识范围，很大程度提高了个人综合能力水平：  1）通过本次项目，项目组成员可以培养团队协作的精神，通过将项目细分为不同模块，分工合作，提高项目进展速度，为以后在工作中分配以及完成任务打下良好的基础；  2）通过本次项目，项目组成员可以学习到计算机领域的深层次知识与前沿知识，学习到多模态特征提取和卷积神经网络等本科教学阶段不常接触的知识，为以后的专业深造提供入门与过渡；  3）在实现算法阶段，算法的微调和对抗训练实现十分困难，在此阶段项目组成员的编程能力与计算思维能力得到了很大的提升，专业素养有了很大提升，为未来工作打下了良好的基础。  **1.2.2科学意义**  由于特征提取、神经网络和网络爬虫技术在现代社会对解决很多问题都有良好的效果，理论证明，它们在项目实践中是解决问题很好的方法，所以在项目中尝试使用这些技术不仅可以更好的解决问题，并且在实践中可以从实践上对相关技术进行验证与完善，并深入了解这些技术，补充技术的理论完备性和实际操作知识。  **1.2.3社会意义**  随着信息科技的发展，社交网络已经成为人们生活中不可缺少的一部分，甚至成为人类虚拟交流的全部平台。而在广大社交媒体平台上，人们利用多模态委婉语发表违规言论，逃避平台管控，有碍社交平台治理，不利于营造积极向上的网络环境。因此设计研究出高效、准确的多模态委婉语检测系统，对捕捉有害信息、促进文明发言的与社会的稳定、提升用户体验有着十分重要的意义。  **1.3国内外研究现状**  以下将系统阐述各个模态委婉有害内容检测技术的研究现状。  （1）国外研究现状及评价：  研究概要：国外在多模态有害内容检测领域的研究起步较早，技术路线多元，近年来呈现出从单一模态向多模态融合、从规则匹配向深度学习、从内容分析向语境理解发展的趋势。南佛罗里达大学团队系统梳理 372 项研究，构建 LLM 相关危害的统一分类体系，明确将 “跨模态委婉有害内容” 纳入有毒内容的隐性分支，指出其 “语义抽象化 + 模态碎片化” 的核心特征，为检测任务提供范畴界定基础，该分类体系已被 ACL 2024 多项研究采纳作为标注标准[1]。斯坦福大学联合 MIT 的研究基于 5 万条 AI 生成三模态样本（文本 + 图像 + 音频）发现，生成式 AI 驱动的有害内容已实现 “文本隐喻 + 视觉暗示 + 音频诱导” 的多模态协同传播，但是传统单模态检测准确率从双模态场景的 79% 下降至 53% 以下，其中音频诱导内容的识别误差率最高，达 49.2%[2]。麻省理工学院提出基于跨模态注意力机制的检测模型，采用 BERT-base 提取文本语义、ResNet-50 提取图像视觉特征，通过 “局部语义相似度计算 - 全局注意力权重分配” 双阶段策略实现双模态对齐，在虚假新闻委婉表达检测数据集（含 8 万条样本）上 F1 值为 78.2%，但因未纳入音频特征，对含背景音的多模态内容检测精度下降 15%[3]。华盛顿大学与微软联合开发的 OMNIGUARD 框架代表了该领域的重大进展。该方法基于核心发现：大型语言模型和多模态大型语言模型在处理信息时，会在内部生成具有 “通用性” 的表示（如 Llama3.3-70B 的第 57 层输出向量），这些表示在 12 种语言和 3 种模态（文本 / 图像 / 音频）间的余弦相似度达 0.83 以上。基于此，OMNIGUARD 构建了含两个隐藏层的 MLP 轻量级检测器，直接利用模型内部通用表示，在跨语言虚假信息检测中 F1 值达 89.1%，跨模态迁移检测准确率达 84.3%[4]。  研究评价：国外研究虽起步早且形成体系化探索，但仍存在一定局限，现有研究多聚焦文本 - 图像双模态融合，对于“ 文本+图像+音频 ”的三模态深度融合场景研究较少，对音频特征的语义化深度融合及文本、图像、音频三模态的兼容性处理技术较为薄弱。同时，国外研究的训练数据以英语内容为主，对非西方文化背景的委婉有害内容（如东亚语境中的隐晦表达）检测能力有限，且在特定数据集（如Twitter）表现良好的方法在跨平台（如微信、抖音）适应性未经验证。同时，国外研究的多数方法（如OMNIGUARD基于Llama3.3-70B）需要依赖大规模计算资源，难以在资源受限的场景下部署。  (2)国内研究现状及评价  研究概要：国内近两年也开始了对多模态技术的逐步探索。大连理工大学构建了首个针对中文居高临下言论（PCL）的PCLMM多模态数据集，包含从Bilibili平台收集的715个高质量含有PCL面部表情帧标注的视频，并开发了MultiPCL检测器，引入面部表情检测模块辅助PCL识别。这填补了中文多模态委婉有害内容检测的数据空白。清华大学团队基于 ERNIE 3.0 构建文本 - 图像特征融合模型，采用 “局部语义匹配 + 全局标签传播” 策略解决模态错位问题，在含 20 万条样本的中文虚假新闻数据集（涵盖政治、民生等 5 类场景）上准确率达 89.3%。复旦大学团队进一步优化 ERNIE 3.0 的中文语义编码模块，针对网络流行语、谐音梗等委婉表达设计 “语义对齐预训练任务”，在中文委婉语识别数据集上 F1 值提升至 86.7%[5]。北京大学在有害表情包检测中，采用 “文本配文 + 视觉元素” 双特征融合，借助 LLaMA-Chinese 生成伪标签补充数据（标注成本降低 60%），在含 10 万条样本的中文社交平台数据集上 F1 值达 85.1%[6]。腾讯安全团队面向抖音、快手等短视频平台，开发轻量化双模态检测模型，通过知识蒸馏将参数量压缩至原模型的 1/5，推理速度提升至 300ms / 条，但为追求效率牺牲了委婉表达识别精度，抽象语义场景下 F1 值仅 72.3%[7]。  研究评价：国内针对该方面的研究还处于探索阶段，存在一些局限和缺口。国内现有研究中，部分团队虽优化了中文语义编码，但未结合视觉模型的语义化提取优势，跨模态特征融合的精细化程度不足，特征融合多停留在浅层拼接或固定权重融合，未通过协同对抗训练提升鲁棒性，且缺乏中文语境下的委婉表达专项优化。现有研究多停留在文本 - 图像双模态浅层融合，未突破 “文本 + 图像 + 音频” 三模态协同检测瓶颈。同时，现有研究或追求精度忽视效率，或侧重效率牺牲精度，未能实现一个较好的平衡。除此之外，现有模型多针对静态数据集训练，对 AI 生成模因、音频隐语等动态演变的委婉有害内容检测率不足 60%；且评估体系仅依赖准确率、F1 值，未构建涵盖 “特征融合有效性、模态协同精度、实时性” 的三维评估框架，难以全面衡量模型综合性能  参考文献：   1. Kurita K, Smith N A, Li Y. A Unified Taxonomy of LLM Harms and Defenses[J]. ACM Computing Surveys, 2024, 57(3): 1-38. 2. Gao X, Brown T, Manning C D. Multimodal Jailbreak Attacks on Large Language Models[C]//Proceedings of NeurIPS 2024. New Orleans: MIT Press, 2024: 4567-4579. 3. Jia R, Liang P. Cross-Modal Attention Fusion for Implicit Harmful Content Detection[J]. IEEE Transactions on Multimedia, 2023, 25: 8921-8933. 4. Radford A, Narasimhan K, Salimans T, et al. Improving CLIP with Cross-Modal Semantic Alignment[J]. Journal of Machine Learning Research, 2023, 24(108): 1-27. 5. Li J, Zhang H, Wang Z. ERNIE-Based Cross-Modal Fake News Detection with Global Label Propagation[C]//Proceedings of ACL 2025. Beijing: ACL, 2025: 1234-1247. 6. Zhang S, Liu Y, Zhao J. Zero-Shot Harmful Meme Detection with LLM-Generated Pseudo Labels[C]//Proceedings of EMNLP 2024. Miami: ACL, 2024: 5678-5690.   [7] Tencent Security Team. (2025). Lightweight Dual-Modal Detection Model for Short Video Content Moderation. Tencent Technology Research Report, 2025(03). |
|  |
| 2．**项目拟解决的关键科学问题，研究内容、总体框架、重点难点、主要目标**（此部分为重点阐述内容） |
| **2.1研究内容**  **2.2.1项目概要**  随着抖音、微博、小红书等社交平台的深度普及，多模态委婉有害内容已成为网络空间治理的突出难题。此类内容通过文字谐音变异、图像隐性符号、音频隐喻暗示等多模态组合形式规避检测，其语义高度依赖上下文语境且动态演变迅速，既严重激化社交矛盾，也对青少年身心健康造成不良影响。当前检测技术存在三大核心瓶颈：一是单模态检测难以应对 “文字 + 图像 + 音频” 的跨模态语义碎片化分布，二是传统模型缺乏对中文委婉语的语境依赖型解读能力，三是对抗性变异导致漏检误检率居高不下。​  对此，本项目拟开展基于大模型辅助的多模态委婉有害内容检测技术研究，核心技术路线包括：多模态特征精准提取，采用 Chinese-RoBERTa-wwm-ext 模型解析文本语义、RAM++（Swin-Large + BERT）架构识别图像隐性符号、基于 Transformer 的 MFCC 特征编码处理音频隐喻；跨模态语义融合，构建动态注意力融合网络，解决文本 - 图像 - 音频的语义对齐与互补问题；大模型赋能的语境解读，引入微调后的中文大语言模型，结合对话历史与群体语境消解语义歧义；对抗性检测优化，通过对比学习生成变异样本，提升模型对谐音、拆分、模态错位等规避手段的鲁棒性。​  本研究旨在突破多模态委婉有害内容的实时检测瓶颈，实现对动态演变违规表达的精准识别，为社交平台内容审核提供技术支撑，助力营造清朗网络空间。  **2.2.2项目设计简述**  1. 文本模态：Chinese-RoBERTa-wwm-ext 语义特征提取技术​  针对中文委婉有害内容的语义隐蔽性（如谐音、语境依赖），采用 Chinese-RoBERTa-wwm-ext 模型实现精准语义解析，核心技术流程如下：​  预训练机制适配：依托该模型 “全词掩码（WWM）+ 动态掩蔽” 特性，对输入文本（如含 “唐人”“伞兵” 等委婉词的语句）执行全词级掩码处理（而非单字掩码），确保模型捕捉完整语义单元（如 “唐人” 作为整体词汇，而非 “唐”“人” 拆分语义），解决中文词语多义性导致的歧义问题。​  特征提取流程：​  文本预处理：对输入内容执行归一化（生僻字编码转换、特殊符号过滤），通过模型自带分词器（基于 BPE 算法）生成 Token 序列，同步注入词嵌入（Word Embedding）、位置嵌入（Position Embedding）与段嵌入（Segment Embedding），形成 512 维初始特征向量。​  Transformer 编码：经 12 层 Transformer 编码器（每层含多头自注意力机制与前馈神经网络），通过自注意力权重计算（如 “唐人” 与上下文 “侮辱”“歧视” 等词的关联度），捕捉委婉语的语境依赖特征，最终输出 768 维文本语义特征向量（[CLS] 位置向量），用于后续融合。​  技术优化：针对网络委婉语高频变异特性，在预训练基础上加入 “中文违规委婉语微调任务”，使用含 10 万条标注数据（涵盖谐音、拆字、隐喻等类型）的数据集微调模型参数，使语义识别准确率提升 12%-15%。​  2. 图片模态：RAM++ (Swin-Large + BERT) 视觉语义特征提取技术​  针对含隐性违规符号的图片（如侮辱性表情包、隐晦图案），采用 RAM++ 架构实现 “视觉特征 + 语义关联” 双维度提取，核心技术细节如下：​  Swin-Large 视觉特征提取：​  图像预处理：对输入图片（如表情包、场景图）执行尺寸标准化（224×224 像素）、归一化（像素值归一至 [0,1]）与数据增强（随机翻转、亮度调整），降低噪声干扰。​  分层特征提取：通过 Swin-Large 的 4 阶段 Transformer 块（Stage 1-4），采用 “窗口注意力（W-MSA）+ 移位窗口注意力（SW-MSA）” 机制，逐步缩小特征图尺寸（从 224×224→14×14），提取多尺度视觉特征（如边缘纹理、色彩分布、物体轮廓），最终输出 1024 维视觉特征向量（每个特征图块对应 1 个向量）。​  BERT 语义关联建模：​  特征映射：将 Swin-Large 输出的视觉特征向量通过线性变换映射至 768 维，与文本模态特征维度对齐，构建 “视觉 Token 序列”。​  语义关联学习：引入 BERT-base 模型，将 “视觉 Token 序列” 与对应文本描述（如图片配文）的 Token 序列拼接，通过跨注意力机制学习视觉元素与文本语义的关联（如 “可爱” 配文与丑化图像的语义冲突），输出融合视觉语义的 1024 维图片特征向量，解决 “图文错位” 类违规内容的识别难题。​  3. 音频模态：基于 Transformer 的音频特征提取技术（参考图示模型）​  针对含隐喻暗示、变声辱骂的音频内容，采用基于 Transformer 的音频频谱图模型（AST）实现违规特征捕捉，核心技术流程（结合图示模型架构）如下：​  音频预处理：​  信号转换：对输入音频片段（采样率 16kHz，单声道）执行短时傅里叶变换（STFT，窗长 25ms，步长 10ms），生成梅尔频谱图（维度 128×100，128 为梅尔滤波器数量，100 为时间步长），将时域信号转换为频域视觉化特征。​  Patch 分割：将梅尔频谱图按 16×16 像素大小分割为 49 个非重叠 Patch（7×7 网格），每个 Patch 通过线性投影转换为 768 维特征向量，形成 “音频 Patch 序列”。​  Transformer 编码与特征提取：​  嵌入层构建：为 “音频 Patch 序列” 添加位置嵌入（基于 Patch 空间位置）与类别嵌入（[CLS] Token），形成初始输入序列（维度 50×768，含 49 个 Patch+1 个 [CLS]）。​  多头自注意力编码：经 12 层 Transformer 编码器（多头自注意力头数 12，隐藏层维度 768），捕捉频谱图的全局频域依赖（如特定频率段的辱骂声纹特征）与时间域关联（如变声处理的节奏异常），最终提取 [CLS] 位置的 768 维音频特征向量，实现对隐性违规音频的识别。​  4. 多模态数据集构建技术​  为支撑三模态检测模型训练与优化，构建涵盖 “文本 - 图片 - 音频” 的中文多模态委婉有害内容数据集（命名为 CMH-Dataset），核心技术流程如下：​  数据采集与合规处理：​  多源采集渠道：从抖音、微博等平台合规爬取公开内容（经平台 API 授权），涵盖文本（评论、配文）、图片（表情包、场景图）、音频（语音评论、短视频背景音）三大模态，同步收集 SPA-VL 等公开数据集的合规子集，确保数据规模达 15 万条（训练集 12 万、验证集 2 万、测试集 1 万）。​  合规性过滤：参照《生成式人工智能服务管理暂行办法》，构建七层过滤防护网：字符级清除非常规 Unicode 符号、实体级脱敏用户个人信息（手机号、身份证号）、语义级过滤违法信息，确保数据脱敏率 100%、违法信息过滤有效率≥99.9%。​  标注体系设计与执行：​  多维标注框架：参照 SPA-VL 数据集分类标准，设计三级标注体系 —— 一级类（合规 / 违规）、二级类（侮辱、歧视、教唆等 6 类）、三级类（谐音、图像符号、变声等 53 类），同步标注模态关联强度（1-5 分，如 “图文语义冲突” 标为 5 分）。​  三级审核机制：采用 “AI 预标注→专业标注员复核→法律顾问终审” 流程，AI 预标注基于 Chinese-RoBERTa-wwm-ext 与 RAM++ 模型生成初步标签，专业标注员（每组 3 人）复核分歧样本，通过 Consensus 加权算法解决争议，确保标注一致性 Kappa 系数≥0.85。​  数据增强与质量优化：​  单模态增强（语义保持策略）：​  文本：采用 MLM 掩码语言模型增强，对委婉词执行 15% 掩码率的预测替换（如 “伞兵”→“[MASK]”→预测生成同义变异词），保持与图像 / 音频的语义一致性。​  图片：通过随机裁剪（scale 0.7-1.0）、色彩扰动等增强手段，避免过度修改导致隐性符号丢失，增强样本与原样本标注一致性≥92%。​  音频：执行时域随机时移（±0.1s）、音量调节（0.8-1.2 倍），确保与视频帧时间对齐精度 ±50ms 内。​  跨模态增强：针对模态缺失样本（如仅含文本 + 图片），采用扩散模型生成匹配音频（如根据侮辱性文本生成对应情绪的语音），通过 CLIP 模型验证跨模态语义一致性，筛选置信度≥0.9 的样本入库。​  对抗性样本生成：模拟网络规避手段，生成谐音替换（“唐人”→“溏人”）、图像符号模糊、音频变声等对抗样本，占比达数据集总量的 20%，提升模型鲁棒性。​  5. 跨模态特征融合与判别技术​  为解决三模态语义碎片化问题，构建动态注意力融合网络，核心技术如下：​  特征对齐：通过线性变换将文本（768 维）、图片（1024 维）、音频（768 维）特征统一映射至 1024 维特征空间，确保模态间特征可比较。​  动态注意力权重分配：计算任意两模态特征的余弦相似度（如文本语义与音频频谱特征的关联度），基于相似度自适应分配注意力权重（如违规语义强的文本权重占比 40%，隐性符号明显的图片权重占比 35%，音频权重占比 25%），通过加权求和生成 1024 维跨模态联合特征。​  分类判别：将联合特征输入两层全连接网络（隐藏层维度 512，激活函数 ReLU），经 Softmax 层输出 “合规”“违规（含侮辱、歧视等子类）” 的概率分布，结合预设阈值（0.5）实现多模态委婉有害内容的分类判别，模型 F1 值目标≥88%。​  **2.3总体框架**  总体框架如图2-1所示，本项目研究框架图简单明了地介绍了我们的项目基于大模型辅助的多模态委婉有害内容的检测研究的研究背景、研究意义、研究目的、预期成果和项目的研究路线。  流程  **2.4拟解决的关键科学问题**  本项目拟解决的关键问题是实现对多模态委婉有害内容的精准检测，具体包括三方面：  首先，解决多模态委婉语动态演变的实时语义追踪与检测滞后破解问题。多模态委婉语呈现 “语义动态迭代 + 组合形式创新” 的双重演化特征：从语义维度看，同一词汇随时间推移衍生侮辱性新含义，且在传播中不断附加负面语义；从组合维度看，新的文本 - 图片 - 音频搭配方式持续涌现。现有检测系统依赖静态训练数据与固定语义规则，对新语义、新组合的识别需经历 “样本收集 - 标注 - 模型重训” 全流程，导致检测滞后，难以跟上演变速度。本项目拟通过中文委婉词语义演化的量化建模，精准捕捉语义从正常到违规的迁移规律；多模态组合形式的增量学习方法，无需全量重训即可识别新型文本 - 图片 - 音频搭配违规内容基于实时社交数据的语义演变预警机制，提前预判潜在新型违规表达，缩短检测响应周期。  其次，多模态违规信息碎片化与语义冲突下的跨模态对齐融合问题。多模态违规内容常通过 “语义拆分” 与 “语义冲突” 规避单模态检测：在语义拆分场景中，违规意图分散在不同模态，单一模态特征无法还原全部违规信息，且模态间语义关联无固定规律，难以建立统一匹配规则；在语义冲突场景中，模态间呈现 “正话反说” 的矛盾表达，传统融合方法易被表面语义误导，无法识别深层违规意图。本项目通过拟跨模态语义碎片的关联挖掘技术，建立文本委婉语与图像隐性符号、音频隐喻的动态匹配逻辑；语义冲突场景的多模态推理机制，区分 “正话反说” 的讽刺违规表达与正常语义矛盾；异构模态特征的统一语义空间构建方法，消除文本离散符号、图像像素矩阵、音频时序信号的结构差异，实现深度融合。  最后，使用对抗性规避手段动态升级下的模型鲁棒性强化问题。违规内容发布者通过 “针对性学习检测规则 - 升级规避手段” 形成对抗循环，规避方式呈现 “单模态变异深化 + 多模态联合隐蔽” 的特点：在单模态层面，文字规避从基础谐音、拆分升级为生僻字替换、特殊符号插入、方言谐音转写，突破传统文本过滤规则；在多模态层面，通过 “细节隐藏”、“跨模态配合”、“图文错位”等方式，降低多模态检测的识别概率，现有模型鲁棒性不足，易被新型规避手段突破。本项目拟对单模态对抗变异的特征畸变规律挖掘，建立文字生僻字替换、符号插入、方言转写等变异形式的统一识别框架；多模态联合对抗场景的特征增强技术，突出视频角落微小违规图片、模糊文字与配合音乐的异常特征，提升隐蔽违规信息的辨识度；基于对抗样本训练的模型鲁棒性优化方法，通过生成多样化对抗样本，增强模型对新型规避手段的泛化能力，减少误检、漏检。  **2.5重点难点**  1.跨模态语义的深度对齐与意图还原：如何突破 “文本 - 图像 - 音频” 异构数据的语义壁垒，实现违规意图的完整还原。  2.动态演变语义的实时追踪与模型迭代：如何解决多模态委婉语的高频演变（如 “唐人” 语义从正常到侮辱的迁移）使静态模型始终处于滞后状态。 |
|  |
| 3．【理工医科】**拟采取的研究方案及可行性分析**（包括研究方法、技术路线、实验手段、关键技术等说明）；【哲学社会学科】**思路方法**（本课题研究的基本思路、具体研究方法、研究计划及其可行性等） |
|  |
| 4．**本项目的特色与创新点（**建议2-3点**）**； |
|  |

|  |
| --- |
| 1. **研究基础与工作条件** |
| 1．**项目负责人研究基础**（建议300字以内） |
|  |
| 2.**指导教师研究基础**（与本项目相关的研究工作积累和已取得的研究工作成绩，建议300字以内）； |
|  |
| 3．**工作条件**（建议200字以内） |
|  |

|  |  |
| --- | --- |
| （三）**承担的与本项目相关的科研项目情况** | |
| 指导教师曾经和正在承担和参加的省部级以上科研和教改项目情况 | 格式如下，有则填，不限条目：  1. 格式：项目类别，批准号，名称，研究起止年月，获资助金额，项目状态（已结题或在研等），主持或参加  例如：国家自然科学基金面上项目，21773999，×××××××××，2018/01-2021/12，30万元，已结题，主持  2. …… |
| 负责人之前参与大创项目情况 | 格式如下，有则填，不限条目：  1. 负责人/团队成员，项目编号，项目名称，立项级别，项目类别，立项年份，结题成绩；  2. …… |
| 项目成员之前参与大创项目情况 | 1. 姓名，负责人/团队成员，项目编号，项目名称，立项级别，项目类别，立项年份，结题成绩；  2. …… |

|  |  |
| --- | --- |
| （四）**完成大创项目情况**（对负责人负责的前一个大创（项目名称及编号）完成情况、后续研究进展及与本申请项目的关系加以详细说明。另附该已结题项目研究工作总结摘要和创新点（限200字）和相关成果的详细目录。未承担过的写“无”）。 | |
| 项目名称及编号 |  |
| 完成情况与后续研究进展 |  |
| 与本申请项目的关系 |  |
| 工作总结摘要及创新点 |  |

|  |  |
| --- | --- |
| **（五）申请人成果和奖励情况**  （请注意：①投稿阶段的论文可以列出；②对期刊论文：应按照论文发表时作者顺序列出全部作者姓名、论文题目、期刊名称、发表年代、卷（期）及起止页码（摘要论文请加以说明）；③对会议论文：应按照论文发表时作者顺序列出全部作者姓名、论文题目、会议名称(或会议论文集名称及起止页码)、会议地址、会议时间；④应在论文作者姓名后注明第一/通讯作者情况：所有共同第一作者均加注上标“#”字样，通讯作者及共同通讯作者均加注上标“\*”字样，唯一第一作者且非通讯作者无需加注；⑤所有代表性研究成果和学术奖励中本人姓名加粗显示。） | |
| **1.代表性成果（**包括论文、专利、专著、科创竞赛获奖、学术交流活动、奖学金等，限合计5项**）** | |
| 论文 | **期刊论文**  示例  (1) **冯建涛**，陈海峰，李良超\*，ZnTi0.6Fe1.4O4/膨胀石墨复合物对污染物的吸附-光催化降解活性，中国科学：化学，2015，45（10）：1075 ~ 1088  (2) **Liming Tan#**, Kelvin Xi Zhang#, Matthew Y. Pecot, Sonal Nagarkar-Jaiswal, Pei-Tseng Lee, Shin-ya Takemura, Jason M. McEwen, Aljoscha Nern, Shuwa Xu, Wael Tadros, Zhenqing Chen, Kai Zinn, Hugo J. Bellen, Marta Morey\*, S. Lawrence Zipursky\*, Ig Superfamily Ligand and Receptor Pairs Expressed in Synaptic Partners in Drosophila, Cell, 2015, 163(7): 1756-1769  **会议论文**  示例：**Lou Y.**#, Zhang H., Wu W., Hu Z., Magic view: An optimized ultra-large scientific image viewer for SAGE tiled-display environment, 9th IEEE International Conference on e-Science, e-Science 2013, Beijing, P.R. China, 2013.10.22-10.25 |
| 专利 | **授权发明专利**  格式：发明人，专利名称，授权时间，国别，专利号  示例：**王凡**， 一种改善营养性贫血的中药组合物及其制备方法，2014.11.19，中国，ZL201210020610.9 |
| 专著 | **专著**  格式：所有作者，专著名称（章节标题），出版社, 总字数，出版年份。  示例：许智宏，**种康**，植物细胞分化与器官发生，科学出版社，420千字，2015 |
| 科创竞赛获奖 |  |
| 学术交流活动 | **会议特邀学术报告**  格式：报告人，报告名称，会议名称，会议地址，会议时间  (1) **郑晓静**，风沙环境下高雷诺数壁湍流研究，第八届全国流体力学学术会议，中国，兰州，2014年9月18-21日  (2) **Jiang Zonglin**, Experiments and Development of Long-test-duration Hypervelocity Detonation-driven Shock Tunnel , 2014 AIAA Science and Technology Forum and Exposition, National Harbor, Maryland , 13 - 17 January 2014 |
| 奖学金 |  |
| **2.代表性之外成果和奖励（**限合计不超过5项**）。** | |
| **获得奖励**  格式：获奖人（获奖人排名/获奖人数），获奖项目名称，奖励机构，奖励类别，奖励等级，颁奖年份（所有获奖人名单附后）  示例：**李兰娟**（1/15），重症肝病诊治的理论创新与技术突破，国家科技部，国家科学技术进步奖，一等奖，2013  （**李兰娟**，郑树森，陈智，李君，王英杰，徐凯进，徐骁，陈瑜，刁宏燕，杜维波，王伟林，姚航平，吴健，曹红翠，潘小平） | |

|  |  |
| --- | --- |
| **（六）预期成果形式（**可多选**）** | |
| 1.□SCI论文 篇  2.□核心期刊论文 篇  3.□会议论文 篇  4.□内部编印期刊论文 篇  5.□授权发明专利 项  6.□申请发明专利 项  7.□创新创业类竞赛获奖  8.□参加国际国内学术交流活动  9.□其他 名称： | |
|  |  |
| **（七）项目经费概要**（按申报项目目标任务需要进行预算，经费执行情况将与结题考核成绩挂钩） | |
| **1.申请经费明细**  （1）仪器设备费  （2）耗材费  （3）测试加工费  （4）国内会务及差旅费  （5）国外会务及差旅费  （6）文献/知识产权事务费  （7）办公费（含文印、办公用品等）  （8）其他费用  **2.合计** | |

|  |
| --- |
| **评审情况** |
| **指导教师意见：** |
|  |
| **指导教师（签名）： 年 月 日** |
| **学院推荐意见：** |
|  |
| **主管院长签名： 年 月 日** |
| **学校专家评审意见：** |
|  |
| **组长签名： 年 月 日** |
| **学校认定意见及批准经费：** |
|  |
| **学校负责人签名： 年 月 日** |