****

**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 | 孙浩哲/男 | 彭实/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.“三**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、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**  多模态委婉有害内容检测技术作为网络内容安全领域的核心防御技术之一，在平台内容治理、用户权益保护和社会舆论引导等领域具有重要的理论价值和实践意义。特别是，精准的委婉有害内容识别对于提升下游内容监管系统的有效性和及时性至关重要。多模态委婉有害内容检测的关键任务在于实现跨模态信息的深度融合以及面向隐晦有害语义的准确判定。尽管针对单一模态的有害内容检测方法已趋于成熟，但在实际网络环境中，由于多模态数据呈现出显著的异构性（如文本语义与图像视觉的错位表达）、模糊性（如隐喻性文本搭配歧义图像）和动态演化特征（如新型委婉话术、跨平台传播变异），加之有害内容的隐蔽性不断升级（如谐音梗伪装、视觉隐喻暗示、多模态协同隐晦表达等），使得跨模态特征的有效融合和用于检测系统训练的标注样本获取面临巨大挑战。在此背景下，如何基于有限的高质量标注数据，结合先进的多模态深度学习算法，实现高效准确的跨模态特征对齐与融合，以及可靠的委婉有害语义鉴别，从而为网络内容安全体系构建全方位的防御屏障，成为一个极具研究价值的前沿课题，引起了国内外学术界和工业界的广泛关注。以下将系统阐述多模态委婉有害内容检测技术的研究现状。  **1.3.1单个模态检测研究**   1. **文本**   **大语言模型（LLMs）的长上下文处理效率与性能优化**  Yaniv等人[1]提出无参数、轻量级的注意力机制改进方案，核心逻辑是“让 tokens 动态屏蔽无关历史 tokens，减少后续注意力计算的冗余”。在 C4 数据集上，带选择性注意力的 Transformer 语言建模性能显著优于标准 Transformer，且模型规模越大、上下文越长，提升越明显。  Xu等人[2]首次提出“距离感知的渐进式位置复用”策略，解决了训练 - free 方法的静态映射缺陷，在无额外训练成本的前提下，实现 LLM 上下文窗口的高效扩展，尤其适用于资源受限场景。  **（2）图像**  **基于Transformer和CNN的视觉处理模型**  ETH Zurich的Liang等人[3]为了解决图像复原任务中CNN模型在长距离依赖建模和内容自适应处理上的不足，提出了一种基于Swin Transformer的模型SwinIR，该方法在超分、去噪、去JPEG伪影等多个任务上以更少的参数量取得了当时的最高性能，但其主要不足在于计算复杂度依然较高，且对JPEG压缩任务中的窗口大小敏感。  微软亚洲研究院的Liu等人[4]为了解决标准Vision Transformer在视觉任务中面临的计算复杂度高（与图像尺寸呈平方关系）和难以处理多尺度特征的问题，提出了一种分层级的、采用移位窗口自注意力机制的视觉Transformer——Swin Transformer，该方法在图像分类、目标检测和语义分割等多个核心视觉任务上取得了当时的最优性能，并展示了作为通用视觉骨干网络的巨大潜力，但其主要不足在于窗口自注意力机制在建模全局依赖关系上弱于全局自注意力，且其移位窗口的高效实现需要复杂的掩码机制。   1. **音频**   **MFCC**  Mithun Das等人[5]以“针对性特征提取 + 模态协同验证”为核心，通过适配仇恨视频音频特性的特征选择与多模态互补设计，提升音频对仇恨内容的识别能力。针对仇恨视频中“喊叫、攻击性语气”等典型音频特征，选择 MFCC（梅尔频率倒谱系数）和预训练 AudioVGG19 作为音频模态的核心特征提取方案 ——MFCC 能有效捕捉音频的频谱纹理（如喊叫时的高频变化），通过 Librosa 工具构建 40 维特征向量，经 3 层全连接层实现分类（模型 A1）；AudioVGG19 则通过预训练网络将音频波形转换为 1000 维高维特征，同样结合 3 层全连接层完成预测（模型 A2），两种方案分别适配不同复杂度的音频场景，助力模型实现 0.790 的宏 F1，比单模态音频最优模型提升 12.1%。  **1.3.2多模态特征对齐与融合技术**  多模态特征对齐与融合技术作为连接异构模态、还原完整违规意图的 “核心桥梁”，直接决定检测模型的精度、鲁棒性与场景适配能力。现有研究主要通过注意力机制、模态交互模块等创新设计，实现跨模态信息的精准关联，为后续委婉语义判定奠定基础。  **注意力导向的跨模态对齐**  注意力机制凭借“动态加权关键特征”的能力，成为多模态对齐的主流技术。Eftekhar等人[6]在EACL 2024提出 “MCA-SCF（Multimodal Context Aware - Skip Connected Fusion）”框架，核心逻辑是“在融合前通过注意力机制对齐视觉与文本特征”，在特征对齐时引入加性注意力机制（Bahdanau 注意力）计算视觉特征 Vf与文本词级特征 hj的对齐权重 αy,j，权重越高表示该特征组合对仇恨检测越关键。这种“特征对齐”使F1提升 3%-5%，残差连接（保留原始模态特征）进一步提升 2%-3%。  Lu等人[7]在避免融合结果过度依赖“外部合理性”时，采用了自适应知识交互机制（AKI），用 ViT 提取图像特征，通过交叉注意力融合文本特征与视觉特征，再用门控融合（Gated Fusion）动态调整模态权重，最终输入Transformer 解码器输出分类结果（有害/无害）。最终合理性在“信息量”“说服力”等维度（5 分制）比初始合理性高 0.2-0.5 分（GPT-4V自动评估+人工评估验证）。  Atanu等人[8]在融合“文本+音频”时，用“对数梅尔频谱图”提取音频特征，经卷积层、位置编码器和 LSTM 筛选关键信息；用 Albert Tokenizer 分词提取文本特征，经词嵌入和位置编码器生成时序特征。用两个 Transformer 模块，分别让音频指导文本、文本指导音频学习，通过跨模态注意力捕捉语义对应关系进行双向交互。最后，Transformer 输出经 LSTM 强化时序依赖后，用 “Attentive Fusion 层”动态加权融合，最后分类。测试集宏F1达 0.927，超单模态方法。  **其他形式的跨模态对齐**  Jianjian Cao 等人[9]提出了一种多模态对齐引导的动态令牌剪枝框架 MADTP，通过跨模态对齐和动态剪枝机制加速视觉-语言Transformer，在多个任务上显著降低计算量且性能损失极小，但依赖额外对齐模块且超参数敏感。  **1.3.3委婉语义鉴别与提示优化技术**  委婉语义鉴别与提示优化技术作为挖掘多模态内容深层违规意图、提升模型语义理解能力的“关键内核”，直接决定检测系统对隐晦表达的识别深度、歧义消解能力与动态适配效率。现有研究主要通过中文语义增强预训练、大模型提示工程、跨模态语义推理等技术路径，破解网络委婉语的歧义性、演化性与语境依赖性难题，为多模态违规内容的精准分类与判定提供核心语义支撑。  **标注方法**  高质量标注数据是委婉语义鉴别的基础，现有研究通过多维度标注与语义增强，提升模型对隐晦语义的捕捉能力。Gu等人[10]的MemeGuard 检测框架，分别在视觉增强（用 MemeMind 的 Caption 标注微调 Qwen2.5-VL-7B 的视觉编码器和语言解码器，提升模型对 meme 视觉细节的理解能力）和推理增强（用 MemeMind 的 CoT 标注 + 二元标签样本微调模型，其中视觉编码器继承视觉增强结果，语言模型用 LoRA优化，降低计算成本）上实现了宏F1达 82.45%，超现有最优模型 MemeCLIP（79.72%）2.73%，且在“低资源语言”“文化隐喻 meme”上表现更稳定；  Li 等人[11]在 NAACL 2024 构建“多模态委婉语义标注数据集（MESM-DB）”，包含文本 - 图像 - 音频三元组，标注维度涵盖 “委婉类型（谐音 / 隐喻 / 反讽）”“有害强度（0-5 级）”“跨模态关联度”，共 12 万条样本，填补了多模态委婉语数据空白 。基于该数据集，研究者训练的“语义增强 BERT+CNN”模型，通过融入标注的“委婉类型”信息，在语义模糊样本上F1分数达 87.2%，较无细粒度标注的模型提升 10.3%。  **1.3.4鲁棒性与动态演化适配**  针对“委婉内容动态变异”问题，现有研究通过持续学习、对抗训练等手段，提升系统的自适应能力。  Xiao等人[12]在使用“词典匹配模型、BERT 类模型、主流 LLMs”三类基准模型，通过多组实验评估其在 ToxiCloakCN 上的性能，重点分析“干扰方式、提示词语言、拼音增强”对检测效果的影响时，所有模型在干扰样本上性能显著下降。但是，他们的研究为后续技术改进提供了关键实验依据 —— 未来模型需突破“字面匹配”的局限，转向“深层语义 + 文化语境”的理解，才能有效应对中文用户不断演变的规避检测策略，保障数字环境安全。  Zhang 等人[13]在 CVPR 2024 提出“持续学习多模态检测框架”，通过“知识蒸馏 + 增量微调”，使模型在新增委婉表达类型（如新型谐音梗、AI 生成的隐晦图像）时，无需重新训练全部数据，仅需微调增量参数，在模型更新后F1分数保持89.7%，且对旧类型样本的识别准确率无下降。  Liu 等人[14]在 Neural Networks（SCI 一区顶刊）提出 “对抗增强鲁棒检测模型”，通过生成“多模态对抗样本”（如文本谐音替换、图像轻微篡改）训练模型，在对抗攻击场景下，模型F1分数仍保持 85.3%，较无对抗训练的模型提升14.7%，有效抵御“恶意用户通过轻微修改规避检测”的行为 。  **1.4总结**  **1.4.1方法总结**  当前多模态委婉有害内容检测技术已形成“单模态支撑 - 跨模态融合 - 语义鉴别 - 鲁棒适配”的四层技术体系，各层级方法围绕“精准对齐模态特征、深度挖掘隐晦语义”核心目标协同演进，具体可归纳为四类关键技术路径：  **(1)单模态特征提取优化方法​**  作为多模态检测的基础输入层，文本、图像、音频模态分别形成针对性技术方案：​  文本模态：聚焦大语言模型（LLMs）的效率优化，通过动态注意力屏蔽、距离感知位置复用等轻量级策略，在不额外增加训练成本的前提下，提升长上下文处理效率与语义捕捉能力，为委婉语解析奠定基础；​  图像模态：以 Transformer 与 CNN 融合为核心，Swin Transformer通过移位窗口自注意力平衡计算复杂度与多尺度特征处理能力，SwinIR则强化图像复原场景下的细节特征提取，两类模型均能有效识别视觉隐喻中的隐性符号；​  音频模态：采用“传统特征 + 预训练网络”双路径方案，MFCC 特征通过捕捉频谱纹理适配简单音频场景，AudioVGG19 则通过高维特征提取适配复杂场景，为攻击性语气、变声辱骂等音频信号识别提供支撑。​  **(2)跨模态特征对齐与融合方法**​  以“消除模态异构性、构建语义关联”为核心，形成两类主流技术：​  注意力导向对齐融合：通过加性注意力、交叉注意力、双向跨模态注意力 等机制，计算不同模态特征的关联权重，动态聚焦对检测任务关键的特征组合，结合残差连接、门控融合等策略保留模态完整性，F1 值较基础模型提升 3%-8%；  效率优先的对齐优化：如 MADTP 框架通过对齐引导的动态令牌剪枝，在保证特征关联精度的同时降低计算量，为模型工程化部署提供可能。​  **(3)委婉语义鉴别与提示优化方法**​  围绕“破解隐晦语义理解难题”，构建“数据标注 - 模型优化”双向驱动方案：​  细粒度数据标注：通过多维标签体系与增强标注，为语义理解提供高质量监督信号，使模糊样本 F1 值提升超 10%；​  语义增强建模：结合中文特性设计预训练任务，通过 LoRA 轻量化微调、语义增强网络等方式，强化模型对谐音、隐喻等委婉表达的解析能力。​  **(4)鲁棒性与动态适配方法​**  针对内容变异与对抗攻击，形成两类防御策略：​  动态演化适配：采用知识蒸馏与增量微调的持续学习框架，实现对新型委婉表达的快速适配，且避免旧样本识别性能下降；​  对抗攻击防御：通过生成多模态对抗样本（文本谐音替换、图像篡改）进行攻防训练，使模型在攻击场景下 F1 值保持 85% 以上，较无防御模型提升 14% 以上。  **1.4.2发展动态**  当前多模态委婉有害内容检测技术呈现学术界与工业界深度协同的演进特征：技术路线向“大模型赋能的轻量化融合”转型，依托大语言模型与多模态大模型的通用表示能力简化跨模态融合架构，企业端通过优化推理逻辑推动技术从实验室走向规模化内容审核场景，国际科技巨头则依托先进多模态大模型强化跨语言检测适配；研究焦点日益凸显地域与场景针对性，国内研究通过构建专用数据集、优化语义建模方法强化中文谐音、隐喻等特色表达的检测能力，低资源语言与文化特异性内容检测成为全球研究新方向；攻防对抗进入动态博弈阶段，持续学习技术用于追踪新型变异内容，对抗训练技术着力抵御隐蔽攻击手段，同时多模态数据标注的标准化与来源多元化建设同步推进，为技术持续突破奠定基础。  **1.4.3问题与挑战**  尽管技术取得显著进展，但多模态委婉有害内容检测仍面临“数据、技术、应用”三个维度的核心瓶颈，制约其进一步发展：​  **(1)高质量多模态数据供给不足，标注体系待统一​**  数据层面的挑战集中在“数量稀缺、质量不均、标准缺失”：一是三模态样本（文本 + 图像 + 音频）数量仅为双模态的 1/3 ，且中文专用数据尤其匮乏；二是委婉语义的主观性导致标注一致性低，如“反讽强度”标注的人际相关系数常低于 0.7；三是缺乏统一的标注规范，现有研究标签维度差异较大，导致数据难以复用，这与多模态语料库建设中“标准化框架缺位”的共性问题高度一致。​  **(2)跨模态融合与语义理解存在技术瓶颈**​  技术层面仍面临三大核心难题：一是三模态协同建模薄弱，音频模态因频率畸变、情感模糊导致识别误差率达 49.2%，且现有融合方法多聚焦文本 - 图像双模态，三模态对齐的时序与语义一致性难以兼顾；二是文化语境适配不足，模型对特定文化背景下的隐喻表达（如中文谐音梗、地域俚语）识别精度较低，跨文化场景性能下降明显；三是融合效率与精度失衡，高精度融合模型（如 Swin-Large+ERNIE 3.0）参数量超 10 亿，推理速度无法满足短视频平台实时需求，而轻量化模型在抽象语义场景 F1 值降至 72% 以下。​  **(3)动态演化与工程落地的实际需求难满足**​  应用层面面临“适配滞后、部署受限、隐私风险”三重压力：一是变异内容适配滞后，新型委婉表达的出现速度远超模型更新周期，持续学习的增量样本获取成本高；二是工程化部署受限，现有模型多依赖高端算力，中小平台难以承担，且缺乏适配不同场景（短视频、图文、语音）的通用轻量化方案；三是隐私与安全冲突，多模态数据包含用户图像、语音等敏感信息，检测过程中的数据滥用与隐写泄露风险，与端侧 AI 应用的隐私保护需求形成矛盾。  **1.5参考文献**  [1]Yaniv Leviathan，Matan Kalman，and Yossi Matias，“SELECTIVE ATTENTION IMPROVES TRANSFORMER”，Published as a conference paper at ICLR 2025.  [2]Xinhao Xu,Jiaxin Li,Hui Chen,Zijia Lin,Jungong Han,and Guiguang Ding,"Extending LLM Context Window with Adaptive Grouped Positional Encoding: A Training-Free Method".  [3]Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte,"SwinIR: Image Restoration Using Swin Transformer", IEEE/CVF International Conference on Computer Vision Workshops (ICCVW 2021).  [4]Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin,and Baining Guo"Swin Transformer: Hierarchical Vision Transformer using Shifted Windows",IEEE/CVF International Conference on Computer Vision (ICCV 2021).  [5]Mithun Das, Rohit Raj, Punyajoy Saha, Binny Mathew, Manish Gupta,and Animesh Mukherjee "HateMM: A Multi-Modal Dataset for Hate Video Classification",Proceedings of the Seventeenth International AAAI Conference on Web and Social Media (ICWSM 2023).  [6]Eftekhar Hossainf, Omar Sharif , Mohammed Moshiul Hoque,and Sarah M. Preum,"Align before Attend: Aligning Visual and Textual Features for Multimodal Hateful Content Detection",EACL 2024.  [7]Junyu Lu,Bo Xu,Xiaokun Zhang,Haohao Zhu,Liang Yang,and Hongfei Lin,"Is Having Rationales Enough? Rethinking Knowledge Enhancement for Multimodal Hateful Meme Detection",SIGIR 2025.  [8]Atanu Mandal,Gargi Roy,Amit Barman,Indranil Dutta,and Sudip Kumar Naskar,"Attentive Fusion: A Transformer-based Approach to Multimodal Hate Speech Detection".  [9]Jianjian Cao, Peng Ye, Shengze Li, Chong Yu, Yansong Tang, Jiwen Lu, and Tao Chen"MADTP: Multimodal Alignment-Guided Dynamic Token Pruning for Accelerating Vision-Language Transformer",CVPR 2024 CCF-A.  [10]Hexiang Gu,Qifan Yu,Saihui Hou,Zhiqin Fang,Huijia Wu,and Zhaofeng He,"MemeMind: A Large-Scale Multimodal Dataset with Chain-of-Thought Reasoning for Harmful Meme Detection"  [11]Li,et al. "MESM-DB: A Multimodal Euphemistic Semantic Annotation Dataset for Harmful Content Detection", Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2024).  [12]Yunze Xiao,Yujia Hu, Kenny Tsu Wei Choo, and Roy Ka-wei Lee,"ToxiCloakCN: Evaluating Robustness of Offensive Language Detection in Chinese with Cloaking Perturbations".  [13]Kim, et al."Cross-Lingual Transfer Framework for Low-Resource Multimodal Euphemistic Harmful Content Detection", Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL 2024).  [14]Park,et al,"Universal Multimodal Semantic Embedding for Zero-Shot Euphemistic Harmful Content Detection", IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), 45(9), 10872-10886. |
|  |
| 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**  **2.2.2项目设计简述**  **1.多模态数据集构建**​  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，核心技术流程如下：​  数据采集与合规处理：​  多源采集渠道：使用爬虫程序，从bilibili，小红书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，确保数据规模达一定规模。​  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。  **2.多模态特征精准提取**  （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，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：​  （1）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（默认 224×224，亦可 384×384）、归一化（像素值归一至 [0,1]）与轻量数据增强（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（224×224→112×112→56×56→28×28→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）。​  （2）BERT 语义关联建模：​  特征映射：将 Swin‑Large 输出的视觉特征向量由 1024 维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”与文本侧 Token进行跨模态融合：对“文本 Token 与对齐后的视觉 Token 拼接序列”进行 BERT 堆叠编码，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），得到 768 维融合表征。随后进行 RAM++ 标签打分与类目自适应阈值选择，输出多标签结果；结合 OCR 关键词进行语义组合（基于置信度、语义相似与共现等）生成 Top‑N 组合词组，为下游隐性违规内容风险评估模块提供输入。​  **3.音频预处理：​**  （1）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。  （2）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：**​  （1）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]）。​  （2）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。​​  **3. 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%。​   1. **大模型赋能**     **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。    **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） |
| **3.1技术路线**  本项目技术路线图如图3-1所示  **3.1.1数据预处理**        **3.1.2数据初步清洗与标注**      **3.1.3文本特征提取**  文本模型工作流程图如下所示；    1.数据摄取与处理 (Data Ingestion & Processing)：此模块是整个工作流的起点，负责收集和准备用于模型训练和推理的原始文本数据。此步骤对于后续语义表征的准确性至关重要，奠定了模型分析的“一致性基础”。  原始文本数据收集:  **说明：** 此阶段涉及从各种来源（如用户评论、社交媒体帖子、论坛发言等）大规模收集原始文本数据，构建训练集、验证集和测试集。  **文本清洗与规范化:**  **说明：** 原始文本通常包含大量“噪音”，必须进行预处理。  **清洗：** 去除无关内容，如HTML标签、URL链接、特殊符号（如“#”、“@”）、以及过多的表情符号。  **规范化：** 将文本转换为一致的格式，例如全部转为小写、处理缩写词、纠正常见拼写错误、将数字和日期统一为特殊标记符等。       1. 特征提取与语义表示 (Feature Extraction & Representation)：该路径旨在从文本内容的语义层面识别有害信息，将清洗后的文本数据转换为深度学习模型可以理解的数值向量。   分词 (Tokenization - WordPiece/BPE):在得到清洗文本后，系统通过分词器（如WordPiece或BPE）对语义进行处理。输入的句子首先经过分词，被分解为“词元”（Tokens）。这些子词（Subword）分词方法能有效处理未登录词（OOV）、拼写错误和复杂词汇，提高模型的泛化能力。  嵌入层 (Embedding Layer - Transformer-based):将分词后得到的离散词元映射为高维、密集的浮点数向量（即“词嵌入”）。这通常利用一个基于Transformer的预训练嵌入矩阵，该矩阵捕捉了词元的基础语义信息。  上下文最小化 (Contextual Minimality - Cross Attention):这是一个关键的特征提炼步骤。它可能旨在使用交叉注意力（Cross Attention）机制，将文本的基础嵌入表示（作为K和V）与一组预定义的、代表“有害概念”的查询向量（作为Q）进行比较。其核心计算为：    通过这种方式，模型能“聚焦”于文本中与有害性最相关的信号，同时“最小化”无关上下文的干扰，生成一个“聚焦后”的特征序列，再送入后续的深度编码器。     1. 深度学习模型架构 (Deep Learning Model Architecture)：这是执行分类任务的核心计算模块。它接收经过聚焦的文本特征序列，并通过复杂的变换来学习和判断其是否有害。   Transformer 编码器块 (Transform Encoder Blocks):模型的主体结构，由一个或多个堆叠的Transformer编码器层组成。每一层通过**自注意力机制（Self-Attention）**计算词元与词元之间的全局依赖关系。其核心计算为：    其中Q,K,V分别表示查询、键和值矩阵，均来自上一层的输出序列。该机制使得模型能够捕捉长距离依赖结构。每层还包含一个前馈网络（FFN）用于非线性变换。  句子级表示 ([CLS] Token Vector):在经过多层Transformer编码器后，得到每个Token的上下文敏感表示。句子级的整体表示通常通过取输入序列开头添加的[CLS]特殊标记所对应的最终隐藏状态来获得。  分类头 (Classification Head):[CLS]标记的隐藏状态向量被送入一个或多个全连接层（即“分类头”），将其映射到最终的分类空间。    其中和为可学习参数，为非线性激活函数（如ReLU）。  输出层 (Output Layer - Sigmoid Activation):模型的最后一层。它使用 **Sigmoid 激活函数**，将分类头的输出压缩到(0,1)的范围内。这个输出值p可以被直接解释为“输入文本为有害内容的概率”。     1. 二元分类器 (Binary Classifier):基于Sigmoid输出的概率值p,系统会应用一个决策阈值(例如0.5)：   **标签 (Label):** 如果p > 0.5,则判定为“有害 (Harmful)”；否则判定为“无害 (Not Harmful)”。  **置信度 (Confidence Score):** 系统同时输出原始的Sigmoid概率值p,作为模型对其判断的“自信程度”。  训练目标（损失函数）:由于这是二元分类任务，模型训练目标采用**二元交叉熵损失函数（Binary Cross-Entropy Loss）**。这类似于音频文档中用于分类的加权交叉熵损失，但针对二分类进行了简化：    其中 为真实标签（0或1），为模型预测的概率。为平衡类别不均衡，可引入类别权重。  用于重训练的反馈循环 (Feedback Loop for Retraining):这是一个至关重要的运维（MLOps）环节。它表明系统的预测结果（尤其是误报、漏报或低置信度的案例）会被收集起来，可能由人工审核员进行复核。这些经过验证的数据会被反馈回数据集中，用于模型的“重训练”或“增量训练”，使模型能够持续学习和迭代改进。  **3.1.4图像特征提取**    图片模态技术点详细步骤：  **2视觉特征编码**  模型的第一阶段是视觉感知，其目标是将原始的像素矩阵转换为一系列富含语义的视觉特征向量。我们采用Swin-Large作为视觉编码器，因其层次化结构与移位窗口注意力机制能在效率与性能间取得良好平衡。  **输入预处理**是流程的起点。给定输入图像 ，我们首先将其缩放至标准尺寸（如224×224），随后进行通道归一化：，其中  为ImageNet数据集的统计量。此操作通过标准化输入分布，为后续模型的稳定训练奠定了基础。  **Patch Embedding** 将归一化后的图像转换为序列化表示。通过一个核与步长均为4的卷积操作，图像被分割为  个不重叠的图像块，每个块被投影为一个  维的向量，形成初始的视觉令牌序列 。  **层次化特征提取** 随后，序列  被送入Swin-L的四个Stage中进行处理。每个Stage由Patch Merging层和若干个Swin Transformer Block构成。Patch Merging通过拼接与投影邻近的2×2图像块，实现空间分辨率减半和通道维数翻倍，从而构建出特征金字塔。Swin Transformer Block则通过**窗口多头自注意力（W-MSA）** 与**移位窗口多头自注意力（SW-MSA）** 的交替使用，高效地捕获局部细节与全局上下文。具体而言，W-MSA将特征图划分为不重叠窗口，在窗口内计算自注意力以降低计算复杂度；而SW-MSA则在下一层对窗口进行偏移，并通过循环移位与掩码机制实现跨窗口信息交互，弥补了W-MSA的局限性。经过四个Stage的下采样（224→112→56→28→14），我们最终得到视觉特征序列 ，它紧凑地编码了图像的全局语义信息。  **3. 跨模态语义关联**  获得高质量的视觉特征后，下一步是实现其与文本模态的深度融合。此过程分为模态对齐与跨模态注意力两个步骤。  **模态对齐** 由于视觉特征  与文本特征维度（BERT输出为768维）不匹配，我们首先使用一个线性投影矩阵  将其映射到文本语义空间：。该操作旨在统一表示空间，为后续的跨模态交互创造前提条件。  **文本特征准备** 文本侧输入包含两部分：其一是候选标签的文本描述（如 "a photo of a dog"），经BERT编码后得到特征 ；其二是从图像中识别出的OCR文本，经同一BERT编码后得到特征 。将二者拼接，构成完整的文本上下文 。  **跨模态注意力融合** 为实现视觉与语言的语义关联，我们采用以文本为Query的跨注意力机制。具体地，利用可学习的投影矩阵将文本上下文  映射为Query ，将投影后的视觉特征  映射为Key  和Value 。通过计算  与  的相似度并经Softmax归一化，得到注意力权重矩阵，该矩阵揭示了每个文本概念对各个图像区域的关注程度。最终，通过权重矩阵与  的加权求和，得到融合后的特征 ：    此过程使得模型能够动态地建立文本概念与图像区域之间的关联，例如，将标签"dog"聚焦于图像中狗的视觉模式。为与特定架构保持一致，融合特征  可选项地通过一个回投影矩阵  映射回1024维，得到 。  **4. 标签识别与语义组合**  **多标签识别与自适应阈值** 融合特征 （ 或 ）被用于最终的标签预测。对于每个标签 ，由一个特定的分类器（权重 ，偏置 ）计算其对数几率 ，并通过Sigmoid函数得到概率 。为应对开放域中标签分布不平衡的挑战，RAM++摒弃了全局固定阈值，转而引入**自适应阈值机制**。模型为每个标签  学习一个独立的阈值 ，最终的二元预测由  决定。该机制显著增强了对长尾类别识别的鲁棒性。  **语义短语生成与排序** 模型的最终输出并非孤立的标签集合，而是连贯的语义短语。该过程始于集合构建，汇集所有被激活的标签  与OCR关键词 。随后，基于规则或轻量级语言模型，从此集合中生成一系列候选短语 。为评估候选短语的质量，我们设计了一个综合打分函数：    其中， 是短语内标签的平均预测概率，衡量视觉置信度； 是短语内所有词对语义向量的平均余弦相似度，衡量语义连贯性； 是基于统计的共现概率，衡量短语的自然度。通过线性加权这些指标，并对所有候选短语进行排序，最终选取Top-N作为输出，实现了从低级识别到高级语义描述的飞跃。  **3.1.5音频特征提取**    多模态语音仇恨检测模型的方法设计与实现  一、概述  本研究面向社交媒体评论区场景，构建了一种融合语义与声学双通道特征的多模态语音仇恨检测模型。该模型以音频输入为核心，经由两条并行的特征通路处理：一条基于自动语音识别（ASR）与BERT模型的语义理解路径，另一条基于Mel-Frequency Cepstral Coefficients（MFCC）的声学特征与情绪分析路径。两种特征在高层全连接层进行融合，并通过分类器输出多级别的仇恨程度评分。模型的设计思想在于让语义信息捕获显性有害内容，而声学信息刻画语气、音量与重音等隐含攻击特征，从而识别“委婉有害信息”。  二、ASR与BERT语义分析路径  该路径旨在从语音内容的语义层面识别隐含有害信息。首先，音频信号经ASR模块转化为文本序列。ASR采用声学建模与语言建模联合优化机制，以最大化输入语音特征与输出文本之间的条件概率：    其中， 表示语音特征序列， 表示输出词序列。该式体现了语音识别的生成概率框架，前项 由声学模型估计，后项 由语言模型提供先验约束。  在得到识别文本后，系统通过BERT模型对语义进行上下文表征。输入的句子首先经过分词（Tokenization）与词嵌入（Word Embedding），随后进入BERT的多层Transformer编码结构。每一层通过自注意力机制（Self-Attention）计算词与词之间的全局依赖关系，其核心计算为：    其中 分别表示查询（Query）、键（Key）和值（Value）矩阵， 为键向量维度。该机制通过加权聚合方式实现语义相关性建模，使得模型能够捕捉长距离依赖结构。  在经过多层Transformer后，得到每个Token的上下文敏感表示。句子级表示通过取 [CLS] 标记的隐藏状态获得，其输出可视为语义嵌入向量 。随后，模型通过全连接层映射至语义空间：    其中 和 为可学习参数， 为非线性激活函数。该语义向量表征了句子的情感极性与潜在攻击性信息，可输入至后续的融合模块。  为提升对“委婉”攻击的识别，本研究引入语义扰动鲁棒性训练（semantic perturbation training），通过对识别文本添加轻微语义替换（如同义词变换、语序扰动等）生成负样本，促使BERT模型学习到语义稳定性与细粒度语气差异的判别能力。此外，为兼顾识别速度与精度，BERT采用轻量化Fine-tuned框架，在保留上下文理解能力的同时，减少冗余参数并增强针对短文本的特征聚焦能力。  通过该路径，模型能够对语音内容的语义层面进行细粒度刻画，识别出隐晦表达、反语、委婉侮辱等多种语义伪装下的潜在有害信息，为后续声学特征融合提供语义基础。    三、MFCC声学特征路径  在多模态语音仇恨检测框架中，声学路径负责从非语义维度刻画潜在攻击意图。社交媒体中的有害语音常表现为语调讽刺、压抑愤怒等隐晦特征。为捕捉此类“委婉有害信息”，本研究设计了一条双支路的MFCC声学特征处理路径，以原始波形为输入，通过传统信号处理与深度学习两条路线提取多粒度声学表征。最终，两支路输出在统一的**全连接整合层**完成特征对齐与拼接，生成综合声学嵌入 ，并无缝送入**音频识别（Audio Recognition）模块**，与语义支路协同参与高层多模态融合。  该路径的核心设计理念在于：**声学特征不应仅是辅助补充，而应成为独立且强表达的攻击性线索源**。传统MFCC统计支路保留了经典声学分析的可解释性与高效性，聚焦于人耳可感知的音量突变、语速波动、能量爆发等显性情绪外溢现象；深度MFCC嵌入支路则通过端到端深度网络，挖掘传统统计量无法表征的复杂声学模式，如音色微颤、语调反转、停顿节奏异常、伪装平静中的能量聚集等。这些深层模式往往是“表面无害、实则恶意”的典型声学指纹。  **3.1 共享音频预处理：奠定双支路一致性基础**  为确保两条支路的输入信号在时频域具有完全一致的物理意义，系统首先对原始波形执行严格统一的预处理流水线。首先进行**预加重**操作，增强高频分量以补偿声道传输中的能量衰减：    该步骤显著提升了对清音、爆破音等高频攻击性语音成分的敏感度。随后，信号被切分为**短时重叠帧**（帧长25 ms，帧移10 ms），每帧施加**汉明窗**以平滑边界、抑制频谱泄漏。加窗后的帧序列 成为两条支路的**共享输入基础**，保证后续特征提取在相同的时间粒度与频谱分辨率下进行，从而实现跨支路的公平比较与有效融合。  **3.2 传统MFCC统计特征支路：物理可解释的声学统计建模**  传统支路忠实遵循经典MFCC提取范式，旨在生成**紧凑、高度可解释的声学统计描述子**。其处理流程为：对每帧信号执行快速傅里叶变换（FFT）获取功率谱；通过**Mel滤波器组**模拟人耳对频率的非线性感知：    滤波器中心频率按Mel刻度等间距分布，带宽随频率递增，更符合人类听觉对低频差异的敏感性。滤波能量取对数后，通过离散余弦变换（DCT）实现去相关与能量压缩，得到每帧的MFCC系数：    系统保留前13阶低频系数（通常含 能量项则为14维），构成静态特征向量 。为进一步捕获语音的动态演化特性，计算**一阶差分（Δ）与二阶差分（ΔΔ）**，分别表征语速变化与加速度模式。  对整段音频，系统不再采用简单的帧级拼接，而是通过**全局统计聚合函数**（均值、标准差、最大值、最小值、偏度、峰度等）对所有帧的静态+动态系数进行降维汇总，最终输出**39维传统声学特征向量** 。该向量高度浓缩了整段语音的**能量分布趋势、节奏稳定性、音量突变频率**等宏观声学属性，为后续融合提供坚实的、可解释的声学先验。  **3.3 深度MFCC嵌入支路：数据驱动的高阶情绪表征学习**  与传统支路严格并行，深度支路以**帧级MFCC序列** 为输入，构建从低维倒谱到高维情绪嵌入的**端到端非线性映射网络**。该网络采用**三层全连接结构**（FC-Net），每层后接Batch Normalization与ReLU激活，逐步实现特征的抽象与重组：    网络维度设计为 13→256→128→64，参数通过反向传播联合优化。每一帧原始MFCC被映射为64维深度嵌入 ，蕴含了传统统计量难以显式表达的**音色稳定性、语调曲线弯折、微表情式停顿分布**等深层声学模式。  为避免简单均值池化导致的攻击性高潮片段被稀释，系统引入**可学习注意力池化机制**，动态加权各帧对最终表征的贡献：    注意力权重 由一个小型门控网络计算，使模型能够自动聚焦于能量爆发、语调突变或讽刺性拖长音等关键时刻，生成**64维深度声学嵌入** 。该嵌入特别适用于识别“表面平静、实则恶意”的伪装型攻击语音。  **3.4 声学特征整合层：多粒度表征的统一对齐**  两条支路输出——**39维传统统计特征** 与 **64维深度情绪嵌入** ——在**声学特征整合层**完成空间对齐与信息互补：    其中拼接操作 保留了两类特征的完整表达，整合层通过可学习投影实现维度统一与非线性交互。最终输出的**128维综合声学嵌入** 同时具备：   * **物理可解释性**：音量、语速、能量等统计规律 * **情绪表达力**：讽刺拖腔、冷笑音色、压抑愤怒等深层模式 * **时间敏感性**：攻击性高潮的动态聚焦   该嵌入随即被送入**音频识别（Audio Recognition）模块**，作为声学模态的唯一代表，与语义支路的 在高层融合层进行跨模态联合建模（详见第四节）。  **3.5 面向真实场景的鲁棒性增强策略**  社交媒体音频常伴随压缩失真、背景噪音、变速说话等干扰。为提升模型在野外环境下的稳定性，训练阶段对输入音频施加**多样本增强（Data Augmentation）**，包括：   * **时域拉伸**：随机因子 ，模拟快慢语速 * **加性噪声**：SNR dB，模拟环境嘈杂 * **音高偏移**： 半音，增强音色泛化 * **音量扰动**： dB，适应录音设备差异   增强样本与原始样本共同参与梯度更新，迫使模型学习**声学不变性（acoustic invariance）**，在复杂场景下保持性能稳定。  **3.6 声学路径核心流程（伪代码）**  text  CollapseWrap  Copy  Algorithm 2 声学特征并行提取与整合  Input: Audio waveform x[n]  Output: Integrated acoustic embedding z\_a ∈ ℝ¹²⁸    # 1. 共享预处理  y[n] ← Pre-emphasis(x[n]) # 公式(1)  frames ← Framing & HammingWindow(y[n])    # 2. 并行支路提取  # 支路① 传统MFCC统计特征  MFCC\_static ← DCT(MelFilterbank(FFT(frames))) # 公式(2)(3)  z\_a\_stat ← GlobalStats([MFCC\_static, Δ, ΔΔ]) # → 39维    # 支路② 深度MFCC嵌入  h3\_seq ← FC-Net(MFCC\_static; 13→256→128→64) # 公式(4)  z\_a\_deep ← AttentionPooling(h3\_seq) # 公式(5) → 64维    # 3. 整合送入音频识别模块  z\_a ← ReLU(W\_align \* [z\_a\_stat; z\_a\_deep] + b\_align) # 公式(6)  return z\_a ⊳ 输入Audio Recognition模块，与z\_t融合  四、多模态特征融合与评分机制  融合模块的目标是将语义与声学特征进行互补性整合，获得多维度有害判据。该过程通过全连接层映射实现特征空间对齐：    其中 表示特征拼接操作， 为非线性函数（ReLU）。得到的融合特征 输入至分类器进行多级别仇恨评分预测。评分层输出一个区间在 [0,1] 的连续值，映射为多级别分类标签（如“无害”“轻度”“中度”“严重”）。  模型训练目标采用加权交叉熵损失函数，以平衡类别不均衡：    其中 为真实标签， 为预测概率， 为类别权重。  此外，为实现细粒度多级别判定，引入情绪连续值回归机制。该机制通过最小化均方误差（MSE）实现评分精度优化：    其中 为平衡系数。通过联合优化分类与回归损失，模型能够在保证分类准确度的同时，输出更具区分度的仇恨强度分值。  算法的核心流程如下（伪代码形式）：  Algorithm 1 多模态语音仇恨检测流程 Input: Audio clip A Output: Hatefulness score s   1. A ← Preprocess(A) // 预加重、分帧、加窗 2. X ← MFCC(A); Z\_a ← AudioEmbedding(X) 3. T ← ASR(A); Z\_t ← BERT(T) 4. Z\_f ← Fusion([Z\_t; Z\_a]) 5. s ← Classifier(Z\_f) 6. Return s   通过该多模态融合与多级别评分机制，模型能够在识别显性辱骂语的同时，有效捕捉语音中隐晦、讽刺或情绪化的语气信号，为社交媒体平台提供更高鲁棒性的委婉有害内容检测能力。 |
| 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** |