首届高校ICT产教融合创新大赛企业命题

命题编号：25

|  |  |
| --- | --- |
| 命题企业 | 中国联通研究院 |
| 命题题目 | 面向社交网络环境的不良言论识别模型 |
| 命题方向 | （请填写命题应用的场景领域）  社交媒体内容监控：在社交媒体平台上，识别并过滤不良言论，防止仇恨言论、虚假信息和网络暴力的传播。  公共服务平台内容过滤：在公共服务平台如在线教育、新闻发布等场景中，不良言论数据识别系统可以确保平台内容的质量，为用户提供安全、健康的信息环境。 |
| （请填写命题涉及的技术方向）  面向AI的训练数据安全治理，具体包括:   * 数据收集与标注 * 预处理与清洗策略 * 不良言论识别模型的开发 |
| 命题内容 | （请详细阐述命题背景、研究内容及输出成果）  命题背景：随着人工智能技术的快速发展，AI模型在各个领域的应用越来越广泛。然而，训练数据的质量直接影响到模型的性能和可靠性。特别是在涉及用户生成内容的场景中，训练数据中可能包含不良言论，如仇恨言论、虚假信息、暴力内容和其他有害言论。这些不良言论不仅会影响AI模型的决策准确性，还可能引发法律和伦理问题，损害用户体验。因此，为了确保AI系统的安全性和合规性，有必要开发一种高效的机制来识别并清理训练数据中的不良言论。  研究内容：基于网络上的开源不良言论训练数据集，完成数据的收集、预处理、清洗，以及模型训练等工作，实现基于AI的不良言论数据识别模型。 |
| 答题所需软硬件资源 | （请写明团队完成命题所需用到的软硬件资源）  输出成果包括但不限于：  1、文档报告（包括但不限于：模型训练过程、相关注释、分析等内容，以及对实战中异常数据的洗清策略的思考）  2、完整的不良言论识别模型源代码、数据集。 |
| 提交材料和评价方法 | （请详细阐述团队所需提供的参赛材料，如对策方案、测试报告、展示实物等；及评价方法和标准，如现场答辩、实物测试验收等）  提交材料：报告（包括但不限于：模型训练过程、相关注释、分析等内容，以及对实战中异常数据的洗清策略的思考）、测试集结果  1、不良言论识别模型文档报告（包括但不限于：模型训练过程、相关注释、分析等）  2、数据处理与清洗策略方法  3、完整的不良言论识别模型源代码、数据集  评价方法：  1、报告的结构内容完整性、语言逻辑的清晰度、方法的创新性和实用性  2、训练模型的复杂度和性能（应尽量保证模型的轻量级）  3、针对测试集的准确率统计 |
| 配套支持 | （企业在参赛团队技术支持、软硬件资源配套、优秀项目成果知识产权转化、优秀学生技术认证、实习和就业等方面能够提供的支持） |
| 其他 | （如有其他意见建议请填写） |