任务书-神经网络模型投毒训练

背景

随着人工智能技术的广泛应用，其在提升业务智能化水平的同时，也带来了不容忽视的安全隐患。在水果电商领域，人工智能模型正被深度应用于核心业务场景，如通过情感分析技术处理消费者对水果商品的海量评价，支撑智能客服响应、个性化推荐、市场情绪监测以及供应链动态调整等关键决策。然而，这些模型面临诸如对抗样本攻击与模型投毒攻击等数据安全挑战。一旦模型被恶意干扰或训练数据遭篡改，可能导致推荐失准、舆情误判、用户满意度下降，甚至引发运营风险与品牌信任危机。因此，系统性地评估并增强AI模型在复杂环境下的数据安全性，已成为保障水果电商平台稳定运行和用户信赖的重要基础。

场景描述

请选手以某领先水果电商平台的安全技术专家身份，参与平台用户评论情感分析系统的安全评估工作。该平台依赖深度神经网络模型对消费者评论进行情感极性判断（正面/负面），其输出结果直接驱动商品排序、促销策略制定与客户服务响应机制。在本次挑战中，您需模拟真实网络环境下的安全威胁，聚焦模型投毒训练攻击两类典型风险，深入探究模型在恶意输入和污染数据下的脆弱性表现，全面检验当前情感分析系统的防御能力，并提出切实可行的技术改进方案，以提升模型在实际业务场景中的鲁棒性与可靠性。

题干

神经网络模型投毒训练

【题干描述】

通过构造并注入精心设计的投毒样本（不限制投毒样本数量），重新训练神经网络情感识别模型，使得模型在关键验证集上的情感预测错误率、标签翻转率最大化。

你的任务是根据提供的在线神经网络数据投毒模型训练平台，构造并上传投毒数据样本，注入至原始训练数据中并调优训练神经网络超参数，使最终训练的模型在1000条验证数据上达到最大情感标签翻转效果，导出并提交训练好的模型output.zip文件至竞赛平台验证，根据标签翻转率进行评分。

【平台提供】

附件：

任务书.pdf

web环境：

访问地址：http://答题平台提供的“环境url”地址

一个神经网络模型投毒训练平台，提供上传投毒样本、数据预处理、超参数设置、投毒效果验证、模型文件导出等功能。

注意： web训练平台，多个浏览器或者同时训练一个地址，会出现冲突，请仅用单个浏览器训练，训练过程中会出现CPU资源大量占用的情况。

【答案提交】

提交从模型投毒训练沙盒平台导出的output.zip文件（文件名必须为output.zip）

【评分标准】

平台使用选手提交的zip文件，提取模型文件进行与神经网络模型投毒训练沙盒平台一致的1000条验证样本推理，计算标签翻转率（即预测标签与真实标签不一致的样本比例）。

最终得分 = 标签翻转率 × 当前题目总分。

例如：

标签翻转率：0.7321，当前题目总分：500分。

最终得分 = 0.7321 × 500 = 366.05

【注意事项】

- 每位选手提交平台的次数为10次，提交格式失败会返回错误原因，不得分，扣除次数。
- 提交前请勿篡改神经网络模型投毒训练沙盒平台所导出zip压缩包，仅修改成规定的文件名即可。
- 若因随意篡改zip内容，导致竞赛平台解析失败，损失提交次数，后果选手自负。
- 在提交答案期间，平台将在最多10次成功提交中选取最高分作为最终成绩。
- 选手提交write up时，请选手在文档最后附上投毒样本csv的完整内容。