


1. 实验概览

- 模型架构: BERT-base-uncased (本地部署)
- 数据集: IMDB影评数据集 (单句分类任务)
- 训练参数: 随机种子42, 未使用GPU加速
- 执行攻击类型:
  - 安全攻击: 对抗样本攻击 (AdvAttack, 已执行)
  - 隐私攻击: 模型窃取攻击 (ModelStealingAttack, 已执行)
  - 未执行攻击: 后门/投毒/模型反演/梯度反演 (executed=false)

2. 基准性能分析

- 正常训练准确率: 90.000%, F1分数: 0.859
- 基准对比: BERT-base在IMDB典型准确率约92%-94% , 当前结果低于文献基准2-4个百分点, 可能因训练超参差异或本地数据集版本导致

3. 安全攻击分析

- \*\*对抗样本攻击(AdvAttack)\*\*:
  - 平均攻击成功率: \*\*65.0%\*\* (显著高于40%行业风险阈值)
  - 准确率变化: 攻击前100.0% → 攻击后35.0% (**暴跌65个百分点**)
  - 攻击分布: 成功13次/失败7次/跳过0次 (失败率35%表明防御机制部分生效)
  - 异常点: 攻击前准确率100%异常偏高, 可能因测试集过小或数据泄露

4. 隐私攻击分析

- \*\*模型窃取攻击(ModelStealingAttack)\*\*:
  - 模型性能对比:
    - 受害者模型准确率: 88.62%
    - 窃取模型准确率: 50.20% (**性能差距38.42个百分点**)
  - 模型相似度: 52.84% (表明窃取模型仅能部分复现受害者行为)
  - 训练动态:

迭代 0: 损失0.4808 | 训练acc 86.25% | 验证acc 90.00%

迭代 9: 损失0.1456 | 训练acc 92.71% | 验证acc 90.00%

- 关键发现: 训练acc上升26.8%但验证acc始终不变, 表明**严重过拟合**

5. 横向对比分析

安全攻击特性对比

评估维度	对抗样本攻击	后门攻击	数据投毒攻击
性能影响	高 (65%↓)	未测试	未测试
检测难度	中 (35%失败)	未测试	未测试
时效特性	即时	未测试	未测试
缓解成本	高	未测试	未测试

隐私攻击特性对比

评估维度	模型反演攻击	梯度反演攻击	模型窃取攻击
信息质量	未测试	未测试	低 (50.2%acc)
泄露范围	未测试	未测试	模型级
实施复杂度	未测试	未测试	中 (10次迭代)
防御可行性	未测试	未测试	高

关键风险评估：

- 最大业务威胁：** 对抗样本攻击使模型准确率骤降至35%，导致分类系统失效
- 最高合规风险：** 模型窃取中52.84%的相似度可能引发知识产权争议
- 最紧急漏洞：** 对抗样本防御接口（TextFooler攻击下35%防御失败率）

6. 防御建议

- 对抗样本防御：**
  - 增强文本清洗：部署同义词替换检测模块
  - 实时监控：建立准确率波动阈值告警（>30%下降即触发）
- 模型窃取防御：**
  - 查询限制：实施API调用频率限制（当前实验500次查询过高）
  - 输出扰动：添加置信度掩码，返回Top-3标签替代原始分数
- 架构改进：**
  - 集成对抗训练（Adversarial Training）
  - 部署模型水印技术

7. 可视化需求

- 图1：对抗样本攻击影响矩阵（攻击成功率vs准确率下降）
- 图2：模型窃取训练曲线（损失/训练acc/验证acc三轴图）
- 图3：防御效能对比柱状图（启用防御vs未启用防御）

指标解释

- 攻击成功率：** 成功误导模型的攻击样本占比
- \*\*模型相似度(agreement)\*\*：** 受害者模型与窃取模型输出一致性
- 过拟合：** 训练准确率持续提升但验证准确率停滞
- \*\*词错误率(WER)\*\*：** 语音/文本反演中错误词占比（未触发）
- ROUGE-L：** 最长公共子序列相似度（未触发）