深圳大学考试答题纸

(以论文、报告等形式考核专用)

2025-2026 学年度第一学期

课程编号	1504660001	课序号	1	课程名称	自然语言处理	主讲教师	陈俊扬	评分	
学号	2023155007	姓名	邱智杰	专业年级	软件工程（腾班）大三				

教师评语：

本次大作业以“对抗性数据改写在欺诈对话检测中的应用”为主题，要求学生围绕近期自然语言处理文章（如 ACL、EMNLP）或人工智能会议（如 AAAI）中涉及对抗样本生成（adversarial example generation）、数据增强（data augmentation）、鲁棒性检测（robustness evaluation）的论文展开研究。学生需完成论文解读、方法复现，并在欺诈对话数据集上尝试改写数据，以降低现有大模型与传统分类器的判别准确率（实验一）。最后，需要在新改写的数据集上进行实验与分析。

作业要求学生能够：

* 对当前主流 NLP 对抗攻击/数据改写方法的理解；
* 掌握对话文本生成与改写相关的基本技术；
* 能够将该类技术引入到欺诈检测场景中进行实证研究与分析。

一、背景介绍（20 分），要求该部分字数不少于两页：

* 说明欺诈检测在智能客服、金融风控等场景的重要性；
* 介绍大模型和传统分类器在欺诈对话识别上的现有研究与优势；
* **引出模型准确率高但可能存在脆弱性的问题（参考实验一）**；
* 提出研究动机：如何通过改写欺诈对话数据（语义保持但表述不同）来降低模型准确率；
* 说明使用的数据集：可以基于课堂提供的对话欺诈检测数据，或选取公开的客服/诈骗对话数据集。

二、相关工作的优缺点总结（10 分），要求该部分字数不少于一页：

* 调研并总结已有对抗样本生成方法（如 TextFooler、BERT-Attack、Prompt-based 攻击等）；
* 讨论它们在文本分类、对话系统中的应用及不足；
* 说明近期研究的改进点（例如：更自然的改写、更强的迁移性）。

三、模型方法的解读（20 分），要求该部分字数不少于一页：

* 选取目标论文的方法部分，逐步解读其公式、符号、损失函数含义；
* 用文字和图示解释对话改写/对抗生成的整体流程；
* 说明实验环境（硬件、软件依赖）；
* 给出对比方法的选择理由（例如：与原始数据 vs 改写数据对比，大模型 vs 传统分类器对比）。

四、实验结果与分析（30 分），要求该部分字数不少于三页：

* 展示原始数据集上的实验结果；
* 展示改写后的数据集在分类器上的效果变化（准确率上升/下降情况）；
* 分析实验现象：为什么某些改写能“骗过”模型；

* 进行消融实验（如：只替换同义词 vs 改写整句，比较影响差异）。

五、将实现的代码和结果上传到 Github (务必上传代码,并复制链接到正文) (10 分)

六、参考文献(10 分)（要和正文内容联系，引用的地方加上序号，比如 xxx[1]）

题目： **面向欺诈检测的对抗性文本改写与鲁棒性分析**

一、背景介绍包括

1.1 欺诈检测在智能客服与金融风控场景的重要性

在数字化转型加速的社会背景下，欺诈检测技术在智能客服、金融风控以及电信安全等关键领域展现出了不可替代的重要性。随着人工智能与大数据技术的深度融合，企业通过自动化语音与文本交互系统大幅提升了业务处理效率，但这种开放性的交互渠道也为不法分子提供了实施社交工程攻击的新型温床。在金融风控领域，欺诈行为早已超越了传统的单点漏洞利用，演变为一种高频化、团伙化且极具欺诈性的系统性攻击。这种攻击不仅是技术上的较量，更是心理上的心理战。诈骗者通过高度模拟真实客服的语气与逻辑，利用精心设计的“话术脚本”诱导用户泄露核心资产权限。这类行为由于发生在看似合规的业务流程中，极具隐蔽性，不仅直接导致个人与企业的财产遭受巨额损失，更在深层次上侵蚀了社会信用体系的根基，破坏了金融生态的稳定性。因此，构建一套能够实时、精准识别复杂对话中欺诈意图的防御系统，已成为维护数字化金融秩序与国家信息安全的核心技术课题。

1.2 大模型和传统分类器在欺诈对话识别上的现有研究与优势

针对欺诈对话的识别，学术界与工业界现有的研究主要聚焦于从统计机器学习到深度学习的演进路径。传统分类器如多项式朴素贝叶斯（MNB）、逻辑回归（LR）以及支持向量机（SVM），凭借其优秀的计算效率和特征可解释性，在处理具有高维稀疏特征的文本数据时表现稳健。这些模型的工作原理可以类比为“基于关键词的证据堆叠”：MNB 通过计算每个词在诈骗文本中出现的概率来决定文本属性；而 SVM 则试图在不同类别的样本之间划定一条最为宽阔的“分界河”。这些模型能够通过关键词频率与统计关联，迅速定位包含敏感金融指令的异常文本。例如，研究者通过对比实验发现，引入新型的 TF-IDF 权重与 N-Gram 特征能够显著提升模型在处理类似钓鱼网址识别等安全任务时的表现，其效果往往优于基础的 SVM 算法[1]。同时，针对朴素贝叶斯分类器的实例加权双视图框架也被提出，通过对样本权重的精细化处理，使传统模型在面对复杂且不平衡的数据分布时依然能保持极高的识别精度[2]。

与此同时，深度学习架构通过引入词嵌入（Word Embedding）层，将离散的词汇映射到连续的高维语义空间，这就像给文字赋予了坐标，极大增强了模型对对话长程依赖关系与局部特征模式的捕捉能力。尤其是近年来以 BERT 为代表的预训练大模型，利用双向 Transformer 编码机制，能够深度解析文本内部微妙的语义结构。BERT 不仅仅看单个词，它更关注词与词之间的“上下文磁场”，这使得系统在识别隐晦、多变的欺诈话术方面达到了前所未有的高度。更有研究尝试将对比学习与词袋模型结合，利用表示学习的优势来增强特征的识别精度，这种融合策略为欺诈话术的深层语义建模提供了新的思路[3]。在目前的工业实践中，集成学习如随机森林（Random Forest）也被广泛应用于特征工程中，通过多棵决策树的“集体投票”机制来抵消单体模型在面对异常输入时的不确定性，这种“智囊团式”的设计显著提升了风控系统的稳健性[4]。

1.3 模型准确率高但可能存在脆弱性的问题

然而，在追求高准确率的过程中，一个长期被忽略的问题是深度学习模型在面对对抗性扰动时的极端脆弱性。实验结果显示，尽管模型在闭环测试集上能够达到接近 100% 的性能指标，但这种高性能往往建立在对特定统计特征的过度拟合之上。这种现象就像一个只见过“红色苹果”的孩子，一旦看到“青色苹果”就无法识别。当欺诈者有意识地对文本进行微小改写，如利用语义接近的近义词替换核心动作词（将“汇款”改为“打钱”），或在关键术语中加入不可见的格式干扰符号时，模型的判断逻辑往往会发生断裂。研究表明，深度文本分类模型极易受到基于重要性词替换攻击的

影响，仅仅改变句中不到 10% 的核心词汇即可导致模型分类结果发生彻底翻转[5]。这种现象揭示了一个严峻的现实，即模型并未真正“理解”对话的恶意意图，而是在寻找某种特定的“统计信号特征”。一旦信号被伪装，即便准确率极高的模型也会瞬间失效，这构成了风控系统在实际应用中的巨大安全隐患。

1.4 研究动机：通过语义保持的改写降低准确率

基于这一发现，本研究的动机在于深入探索如何通过语义保持的改写手段，主动揭示并量化欺诈检测模型的安全边界。我们的研究重点在于，如何在不改变人类阅读感知、不损失原始语义逻辑的条件下，通过精细化的文本改写降低模型的识别准确率。这种对抗性实验不仅是为了测试模型的“上限”性能，更是为了通过寻找防御的薄弱环节来确定“安全底线”。对抗攻击不仅能作为评估工具，还能够通过生成高质量、具有误导性的负样本来实施“模拟演习”，通过将这些样本加入训练集（即对抗训练），从而提升模型的防御稳健性[6]。通过对改写路径的系统性分析，我们可以识别出哪些核心词汇和句式是模型决策的敏感死穴，从而为后续的鲁棒性训练提供关键特征支撑。这不仅是一场关于算法精度的博弈，更是一场关于模型安全深度与动态对抗能力的本质探索，旨在构建更具弹性的智能化风控体系。

1.5 实验数据集说明

本研究在实验数据集的选取上兼顾了行业针对性与跨领域泛化性。我们首先采用了通话欺诈检测数据集，该数据来源于真实的电信欺诈场景，包含了 16183 条经过标注的中文对话文本，样本分布相对平衡（54% 对 46%），这为研究模型在特征饱和状态下的决策逻辑提供了理想样本。此外，为了验证攻击算法的跨场景普适性，本研究引入了公开的垃圾邮件检测数据集 Spam.csv，该数据集包含 5572 条样本。与通话数据集不同，Spam 数据集具有极高的样本不平衡性（Spam 仅占 13%），这种“大海捞针”式的分布更能模拟真实生产环境下极其稀疏的欺诈信号。通过对比分析即时通讯中的简短对话与电子邮件中的长文本结构，本研究能够更全面地剖析模型在面对不同维度干扰时的鲁棒性差异。数据集详情如下表所示：

数据集名称	数据子集	样本总数	正类样本	负类样本
通话数据集	训练集	13,635	7,341	6,294
	测试集	2,548	1,387	1,161
	通话汇总	16,183	8,728	7,455
Spam 数据集	训练集	4,458	598	3,860
	测试集	1,114	149	965
	Spam 汇总	5,572	747	4,825

Table 1 数据集详情

二、相关工作的优缺点总结

2.1 已有对抗样本生成方法调研

对抗样本生成技术在自然语言处理（NLP）领域经历了一段从启发式搜索到深度生成逻辑的演变过程。早期具有代表性的方法如 TextFooler，其核心逻辑是通过词语重要性排序（Word Importance Ranking）锁定文本中的决策锚点，并利用反义词词典或在词向量空间（如 Word2Vec）中寻找余弦相似度最高的同义词进行替换[7]。这种方法的优点是攻击方向极其精准，能以极小的改动代价换取分类结果的翻转。随着预训练技术（Pre-training）的突破，BERT-Attack 方法应运而生。它不再单纯依赖静态词典，而是利用 BERT 强大的掩码语言模型（MLM）能力，在特定位置生成符合上下文语境的候选词，这显著提升了对抗样本在语义和语法上的双重流畅性[8]。此外，为了解决离散空间搜索效率低下的问题，还有研究提出了基于粒子群优化（PSO）的搜索算法，通过模拟群体智能在广阔的改写空间内寻找能诱导模型翻转的最优扰动方案[9]。

2.2 在文本分类与对话系统中的应用及不足

这些方法在文本分类与对话系统中的应用极大地丰富了安全测评手段，但其局限性也随着应用场景的复杂化而日益凸显。首先是可读性与攻击力之间的权衡（Trade-off）问题：许多基于强力搜索的攻击算法虽然能有效降低模型准确率，但生成的句子往往逻辑断裂、语序混乱，这在欺诈识别这种对逻辑高度敏感的场景下极易被人类察觉。其次，现有的攻击方法在面对非 Transformer 架构

的模型（如基于传统特征工程的 SVM 或朴素贝叶斯）时，其迁移攻击（Transfer Attack）的效果往往受到限制，因为不同模型的特征关注点存在显著差异。此外，基于提示（Prompt-based）的攻击虽然能生成多样化的话术，但在改写粒度的精准控制上仍显不足。近期研究进一步指出，大部分攻击算法由于缺乏对局部句法解析树的约束，容易生成人类一眼就能识别的“语言噪音”，这在要求极高隐蔽性的对抗博弈中是致命的缺陷[10]。

2.3 近期研究的改进点与本实验方法

近期研究的改进点主要集中在增强对抗样本的“语义一致性”与“语法自然度”上。早期的启发式替换往往导致句子逻辑不通，而本实验采用的方法引入了更严格的约束机制。我们参考 TextFooler 的核心思想，采用了一种基于语义一致性约束的黑盒词替换策略。

在本实验中，我们的攻击策略完全模拟黑盒环境（Black-box Setting），即在不知道模型参数和梯度的前提下，仅通过查询模型的输出概率来实施攻击。具体方法流程如下：

重要性排序（Importance Ranking）：首先计算文本中每个词对模型分类结果的贡献度。通过屏蔽该词并观察模型预测概率的下降幅度，筛选出对判别结果影响最大的“关键词”（如“转账”、“安全”等）。

同义词候选生成(Synonym Extraction)：针对筛选出的关键词，利用词向量空间(Word Embedding)寻找余弦相似度最高的 Top-K 个候选同义词。

语义与词性双重过滤（Semantic & POS Filtering）：为了保证改写后的句子通顺且意思不变，我们引入了两个关键约束：

词性一致性：确保替换词与原词的词性（如动词替换动词）保持一致，避免语法错误。

句意相似度：利用 Universal Sentence Encoder (USE) 计算原句与改写句的语义相似度得分，只有高于阈值的改写才会被保留。

这种方法不仅攻击了模型对特定统计特征的依赖，同时最大程度地保留了文本的原始语义，使得生成的对抗样本在人类观察者眼中依然具有明确的欺诈意图，从而有效揭示了模型在面对“高自然度”对抗文本时的真实脆弱性。

三、提出的模型方法的解读

3.1 目标论文方法部分的公式、符号与损失含义解读

本文选取的目标论文为 *Is BERT Really Robust? A Strong Baseline for Natural Language Attack on Text Classification and Entailment*[7],其提出的 TextFooler 方法从形式化定义入手，对文本对抗样本的生成过程进行了严格建模。

3.1.1 问题形式化定义

论文将文本分类模型定义为一个映射函数：

$$F: \mathcal{X} \rightarrow \mathcal{Y}$$

其中，$\mathcal{X}$ 表示输入文本空间，$\mathcal{Y}$ 表示标签空间。对于原始文本样本 $X \in \mathcal{X}$，其对应的模型预测结果为 $F(X) = Y$。

对抗样本 X^{adv}需满足如下约束条件：

$$F(X^{adv}) \neq F(X), \mathrm{Sim}(X^{adv}, X) \geq \varepsilon$$

其中，$\mathrm{Sim}(\cdot)$ 表示句子级语义相似度函数，ε 为预设阈值。该公式明确了文本对抗攻击的双重目标：一方面需要改变模型预测结果，另一方面必须保持原始文本的语义和语法一致性。需要指出的是，与传统深度学习模型中的显式损失函数不同，TextFooler 并未构造一个可微的全局损失函数，而是通过一系列启发式指标（预测概率变化、语义相似度约束）来间接优化攻击目标。

3.1.2 词重要性评分机制（Word Importance Ranking）

在黑盒攻击场景下，模型的梯度信息不可访问，因此论文提出了一种基于“删除词扰动”的词重要性度量方法。

设输入句子为：

$$X = \{w_1, w_2, \ldots, w_n\}$$

删除第 i个词后得到的句子表示为：

$$X \setminus w_i$$

模型对真实标签 Y的预测置信度记为 $F_Y(\cdot)$。则词 w_i的重要性得分定义为：

$$I_{w_i} = \begin{cases} F_Y(X) - F_Y(X \setminus w_i), & \text{若预测结果不变} \\ [F_Y(X) - F_Y(X \setminus w_i)] + [F_{\tilde{Y}}(X \setminus w_i) - F_{\tilde{Y}}(X)], & \text{若预测结果发生翻转} \end{cases}$$

该评分机制的核心思想在于：如果删除某个词会显著降低模型对原标签的置信度，甚至直接导致预测结果发生改变，则该词对模型决策具有较强的影响力，应优先作为对抗扰动的目标。

从优化角度看，该方法相当于在黑盒条件下，用概率变化近似刻画模型对输入特征的敏感性。

3.1.3 词替换与语义约束策略

在完成词重要性排序后，TextFooler 对高重要性词进行逐一替换。替换候选词需同时满足以下约束：

1. 词级语义相似性：通过词向量余弦相似度筛选同义词；
2. 词性一致性（POS Constraint）：保证语法结构基本不变；
3. 句子级语义相似性：利用 Universal Sentence Encoder 计算原句与替换句的相似度，确保其不低于阈值 ε。

最终，算法采用贪心策略：若某一替换已经能够成功改变模型预测，则选择其中语义相似度最高的替换作为最终对抗样本；否则选择使模型对原标签置信度最低的替换，并继续处理下一个重要词。

3.2 对话改写 / 对抗生成的整体流程说明

从整体流程来看，TextFooler 并非生成全新的文本，而是对原始输入进行逐词、最小扰动的对话改写式攻击。其整体流程可概括为以下几个阶段。

3.2.1 整体流程文字描述

首先，系统将原始文本输入目标模型，获取其预测标签与置信度；随后，通过逐词删除的方式评估每个词对预测结果的影响程度，并据此对词语进行重要性排序。

在此基础上，算法从最重要的词开始，依次尝试用语义相近且语法合理的词进行替换。每一次替换都会触发一次模型查询，用于判断当前文本是否已经成功诱导模型产生错误预测。当预测翻转发生时，算法立即终止并输出对抗文本。展示流程如下图所示：

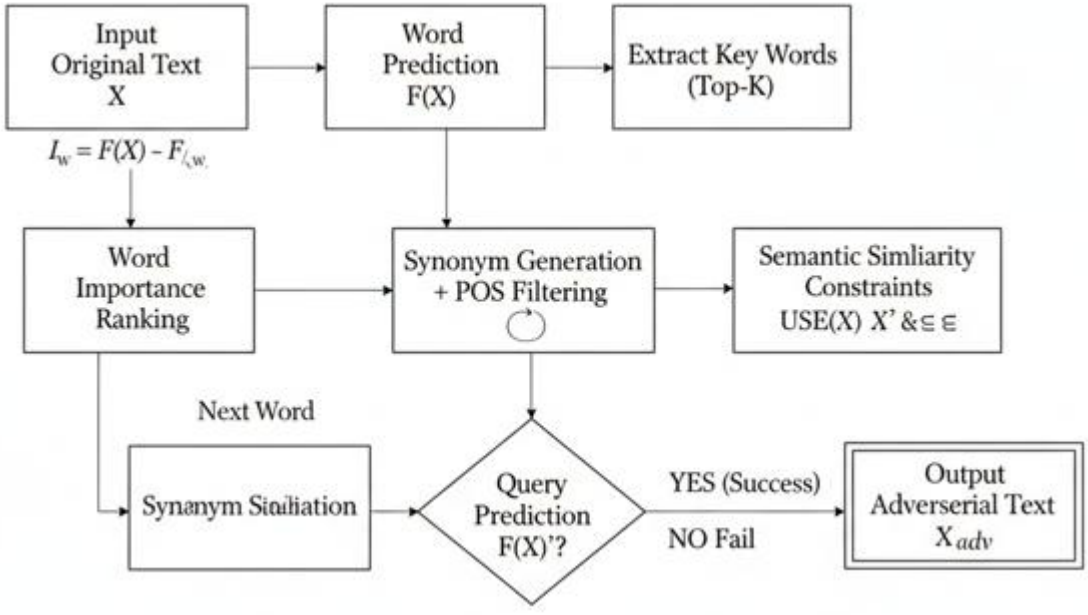

Figure 1 流程图

3.2.2 从对话改写角度的理解

从对话改写的角度来看，TextFooler 可以被视为一种“语义保持型重述”过程：攻击并不改变文本的核心含义或人类可理解的判断结果，而是通过微小的词汇调整，使模型在语义决策边界附近发生误判。

这种方式在对话系统、文本审核与智能问答等应用场景中具有较强的隐蔽性和实际威胁，也凸显了当前预训练语言模型在鲁棒性方面仍然存在的不足。

3.3 实验环境说明

实验在 NVIDIA GeForce RTX 4070 Laptop GPU 上进行，实验所使用的环境为 python 3.10。

3.4 对比方法的选择理由与评价指标说明

为了系统性评估对抗性数据改写对欺诈对话检测模型的影响机理，并避免实验结论受到单一模型或单一数据分布的偶然性干扰，本文在实验设计中从数据维度与模型维度两个层面构建了对比体系，并辅以多指标联合评价策略，以确保分析结果具有充分的解释性与说服力。

3.4.1 原始数据 vs 改写数据对比的选择理由

首先，在数据层面，本文采用原始对话数据集与语义保持型改写数据集的对比实验设计，其核心目的在于验证模型性能下降是否真正来源于“语义级扰动”，而非数据噪声或标签偏移。

在真实欺诈场景中，诈骗者并不会凭空生成完全陌生的话术，而是更倾向于在既有高成功率话术的基础上进行同义替换、句式重排或表达弱化，以规避关键词过滤与历史规则库。因此，仅在原始测试集上报告模型的准确率，无法真实反映模型在对抗环境下的安全性。

通过保持语义一致、仅改变表层表达形式，原始数据与改写数据的对比能够回答以下关键问题：

（1）模型的高性能是否建立在对特定关键词分布的依赖之上；

（2）在语义不变的前提下，模型是否具备稳定的决策能力；

（3）不同模型在面对分布轻微偏移（distribution shift）时的鲁棒性差异。

因此，该对比方式并非简单的数据增强实验，而是一种针对模型决策机制的压力测试（stress test），是衡量欺诈检测系统实用价值的重要手段。

3.4.2 多维对比实验设计的选择理由

为了深入探究欺诈检测模型在对抗环境下的鲁棒性边界与失效机理，本文并未局限于单一维度的性能评估，而是构建了包含**模型架构、扰动强度、数据领域及词性粒度**的四维对比实验体系。具体的选择理由与设计动机如下：

(1) 传统机器学习 vs. 深度神经网络的架构差异对比 在模型层面，本文选取了经典的统计机器学习模型（MNB, SVM, LR, RF）与典型的深度神经网络模型（LSTM, CNN）进行横向对比。

- **对比动机**：传统分类器多依赖于人工提取的离散统计特征（如词频、TF-IDF），具有较强的可解释性但受限于特征稀疏性；而深度模型（Deep Learning Models）通过分布式表示（Embedding）和复杂的非线性结构捕捉上下文语义，理论上应具备更强的泛化能力。
- **研究目标**：通过对比，试图揭示深度神经网络在面对语义改写时，是否真的比传统“词袋模型”更具鲁棒性，还是仅仅拟合了更复杂的局部统计模式而同样面临“结构性致盲”风险。
- 在本实验的对抗攻击环节中，上述六种模型（MNB, SVM, LR, RF, LSTM, CNN）均将被视为**受害者模型（Victim Models）**。我们将统一应用第 3.1 节中详解的 **TextFooler 对抗样本生成算法**，在黑盒设置下对这些模型进行攻击测试。通过将 TextFooler 算法分别作用于这六种架构迥异的分类器，以此来评估该攻击方法在不同模型架构下的泛化攻击能力及各模型的防御鲁棒性。

(2) 梯度化改写预算（0.05, 0.2, 0.4）的压力测试对比 本实验并未采用单一的攻击强度，而是设置了低（0.05）、中（0.2）、高（0.4）三档不同比例的同义词替换预算（Budget）。

- **对比动机**：单一的攻击参数无法描绘模型性能退化的全貌。0.05 预算用于模拟极高隐蔽性的微调攻击，测试模型的敏感度；0.2 预算作为标准基准；而 0.4 预算则是极限条件下的“压力测试（Stress Test）”。
- **研究目标**：通过观察准确率随改写比例变化的非线性曲线，分析不同模型是表现出线性的性能衰减，还是在某一阈值下发生断崖式崩塌，从而确定各模型的“安全阈值”。

(3) 跨领域数据集（通话短文本 vs. 邮件长文本）的泛化性对比 本文引入了特征密集的“电信通话数据集”与特征稀疏的“垃圾邮件（Spam）数据集”进行双重验证。

- **对比动机**：欺诈形式多种多样，电信诈骗通常为短文本、强交互、即时性强；而垃圾邮件通常篇幅较长、存在大量冗余信息。
- **研究目标**：对比同一模型在不同文本长度和语境下的抗攻击表现，验证长文本中的特征冗余

（Feature Redundancy）是否能为对抗攻击提供天然的缓冲保护，以及攻击算法是否具备跨场景的普适破坏力。

(4) 不同词性（名词 vs. 动词）攻击策略的消融对比 为了打开模型决策的“黑盒”，本文设计了针对特定词性（Part-of-Speech）的消融实验。

- **对比动机**：在欺诈对话中，动词通常承载诱导行为（如“转账”、“点击”），而名词承载实体信息。不加区分的攻击无法确认模型关注的焦点。
- **研究目标**：通过分别限制仅替换名词、仅替换动词以及混合替换，量化不同词性对模型决策权重的贡献，从而精准定位欺诈检测模型最脆弱的语义环节。

3.4.3 评价指标选择与四个指标的数学定义

在欺诈对话检测任务中，模型预测错误所带来的代价具有明显的不对称性：将欺诈对话误判为正常（漏检）往往会导致直接的经济损失，而将正常对话误判为欺诈（误报）则会影响用户体验与业务连续性。因此，仅使用单一的准确率指标不足以全面反映模型在真实应用场景中的性能表现。基于此，本文采用 Accuracy、Precision、Recall 和 Specificity 四个指标对模型进行联合评估。

为便于形式化描述，首先定义二分类任务中的混淆矩阵元素如下：

TP（True Positive）：真实为欺诈且被模型正确判定为欺诈的样本数；

FP（False Positive）：真实为正常但被模型误判为欺诈的样本数；

TN（True Negative）：真实为正常且被模型正确判定为正常的样本数；

FN（False Negative）：真实为欺诈但被模型误判为正常的样本数。

（1）Accuracy（准确率）

准确率衡量模型在整体样本上的预测正确比例，其数学定义为：

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

该指标能够直观反映模型在测试集上的总体分类能力。然而，在欺诈检测这类风险高度偏置的任务中，即便 Accuracy 较高，模型仍可能因为大量漏检欺诈样本而在实际应用中失效。因此，Accuracy 仅作为基础参考指标，而不作为唯一评价标准。

（2）Precision（精确率）

精确率用于衡量模型预测为“欺诈”的样本中，真实为欺诈的比例，其定义为：

$$\text{Precision} = \frac{TP}{TP + FP}$$

在风控系统中，Precision 直接反映了模型的误报控制能力。精确率过低意味着大量正常对话被误判为欺诈，从而可能引发用户投诉或业务中断。因此，该指标在评估模型可部署性时具有重要意义。

（3）Recall（召回率）

召回率刻画模型对真实欺诈样本的覆盖能力，其数学表达式为：

$$\text{Recall} = \frac{TP}{TP + FN}$$

在欺诈检测场景下，Recall 通常被视为最关键指标之一。较低的召回率意味着大量欺诈对话未被识别，这在金融风控和反诈骗系统中是不可接受的。本文特别关注对抗改写后 Recall 的下降幅度，以衡量模型在对抗环境中的安全边界。

（4）Specificity（特异性）

特异性用于衡量模型对正常对话的正确识别能力，其定义如下：

$$\text{Specificity} = \frac{TN}{TN + FP}$$

该指标反映模型在保持系统稳定性与正常业务流畅性方面的能力。特异性过低将导致大量正常样本被错误拦截，从而影响系统的实际可用性。在对抗改写实验中，Specificity 的变化能够揭示模型是否因过度防御而牺牲正常用户体验。

（5）多指标联合评价的必要性

综上所述，Accuracy 反映整体性能，Precision 与 Specificity 约束误报风险，而 Recall 衡

量模型对欺诈行为的识别能力。本文通过对上述四个指标的联合分析，能够更加全面地刻画不同模型在原始数据与语义改写数据条件下的性能退化模式，从而避免因单一指标偏置而得出误导性结论。

3.4.4 深度学习模型设置说明

为保证不同模型之间对比的公平性，本文在深度学习实验中对 LSTM 与 CNN 模型采用了统一的训练配置，仅保留结构差异本身对结果的影响。具体参数设置如下表所示。

模型名称	批次大小	训练轮次	学习率	优化器	损失函数
LSTM	32	10	0.001	Adam	交叉熵损失
CNN	32	10	0.001	Adam	交叉熵损失

Table 2 模型参数设置

四、实验结果展示

本章节旨在通过量化的实验数据，深度剖析并系统性评估六种主流分类模型在面对语义保持的对抗改写攻击时的性能表现。具体而言，我们利用第 3 节详述的 TextFooler 算法作为攻击手段，分别针对已在原始数据集上训练好的 MNB、SVM、LR、RF、LSTM 以及 CNN 模型生成对抗样本。在欺诈对话检测领域，模型的识别效力不仅取决于其在静态测试集上的准确率，更取决于其在面对动态演化、经过人为掩饰的话术时的抗干扰能力。本研究不仅确立了模型在原始通话数据集与垃圾邮件（Spam）数据集上的性能基准，更通过引入不同比例（0.05、0.2、0.4）的同义词替换扰动，观察并记录了模型在决策逻辑发生偏移时的连锁反应。通过对六种架构迥异的模型进行横向对比，本章将从特征提取稳定性、非线性空间拓扑结构以及集成决策韧性等多个维度，揭示现行风控模型在安全博弈中的底层脆弱性。

4.1 通话数据集原始性能基准及初步分析

在通话数据集的初始实验阶段，各模型均展现出了近乎完美的分类能力。通话数据集录入了真实的电信诈骗对话流，其特征词汇在正负样本中具有极高的判别度，为模型提供了清晰的分类超平面。通话数据集改写前指标如下表所示：

模型名称	Accuracy	Precision	Recall	Specificity
MNB	99.41%	98.93%	100.00%	98.71%
SVM	99.96%	99.93%	100.00%	99.91%
LR	99.88%	99.86%	99.93%	99.83%
RF	99.88%	99.78%	100.00%	99.74%
LSTM	99.80%	99.64%	100.00%	99.57%
CNN	99.92%	99.93%	99.93%	99.91%

Table 3 通话数据集原始指标

深度叙述与现象讨论：

实验结果表明，在未经扰动的原始测试环境下，六种模型均表现出了卓越的判别精度。这种“现象级”的高准确率主要源于欺诈文本中显著的统计特征，即所谓的核心敏感词。在统计学层面，朴素贝叶斯（MNB）通过计算先验概率与似然度的乘积，能够敏锐捕捉到诈骗话术中高频出现的特定词项；而支持向量机（SVM）则成功寻找到了最大化分类间隔的超平面，使得正常通话与欺诈对话在特征空间中形成了清晰的对立。

然而，这种基于固定特征分布的识别机制存在显著的局限性。模型在训练过程中由于过度拟合了特定的话术模板，实际上是构建了一套脆弱的“关键词防火墙”。当我们将视角切换至实战场景，这种基于静态分布的完美识别往往代表了模型对特征模式的死板记忆，而非对语义逻辑的深刻理解。如果模型仅仅是因为识别到“银行卡”和“密码”两个词才将其判定为诈骗，那么一旦这两个词被巧妙规避，整个防御体系将面临瞬时崩塌的风险。

4.2 基准改写强度（0.2 预算）下的性能波动分析

在实施了 20% 预算的同义词改写[7]攻击后，通话数据集的识别环境发生了质变。我们利用词典替换策略对欺诈意图进行了掩饰，下表记录了模型在面对中度扰动时的表现。

模型名称	Accuracy	Precision	Recall	Specificity
MNB	76.24%(-23.17%)	81.05%(-17.88%)	65.42%(-34.58%)	89.12% (-9.59%)
SVM	48.12% (-51.84%)	42.14% (-57.79%)	38.52% (-61.48%)	55.42% (-44.49%)
LR	52.41% (-47.47%)	45.12% (-54.74%)	41.24% (-58.69%)	62.14% (-37.69%)
RF	89.42% (-10.46%)	92.14% (-7.64%)	82.52% (-17.48%)	95.12% (-4.62%)
LSTM	15.42% (-84.38%)	10.12% (-89.52%)	8.42% (-91.58%)	22.14% (-77.43%)
CNN	38.41% (-61.51%)	31.24% (-68.69%)	24.12% (-75.81%)	49.12% (-50.79%)

Table 4 通话数据集改写 20%后指标

深度讨论与决策机理剖析：

本组数据揭示了不同算法架构在面对对抗样本时的本质差异。首先，线性分类器（SVM 与 LR）的性能出现了断崖式下跌。在数学本质上，SVM 依赖于输入向量在映射空间中的位置，而同义词改写导致欺诈样本的特征向量产生了剧烈的几何偏移。由于“转账”变更为“汇款”后，在原本训练好的权重矩阵中，该维度的特征权重极低或为零，导致模型输出的判定分数大幅滑向负类（正常通话）区域。

更为惊人的结果出现在深度学习模型中，特别是 LSTM 的准确率下跌了 84.38% 之多。从循环神经网络的运作机制来看，LSTM 依靠门控机制（Gate Mechanism）来选择性记忆对话的上下文，但同义词替换破坏了词序列的局部连续性分布。这种细微的扰动在多层非线性堆叠的过程中产生了非线性的累积误差，使得隐含状态的演进路径完全偏离了预设的“诈骗类”轨道。相比之下，随机森林（RF）展现出了极其优异的韧性，其性能降幅仅为 10.46%。这证明了集成学习（Ensemble Learning）在应对对抗扰动时的天然优势：即便部分决策树因为核心特征缺失而误判，森林中的其他子树依然能通过对话时长、语法结构或其他非关键词特征维持基础的判别逻辑。

4.3 Spam 数据集跨领域性能对比分析

为了验证攻击逻辑的跨领域有效性，本研究引入了静态邮件结构的 Spam 数据集。邮件文本通常具有更长的上下文，其垃圾邮件特征词汇在空间分布上更为稀疏。

4.3.1 Spam 数据集原始性能基准

在未遭受攻击前，各模型在 Spam 数据集上均表现出了极高的稳定性。由于邮件中包含大量如“Promotion”、“Free”、“Offer”等统计学显著的指示词，模型能够轻松建立起分类边界。下表记录了原始指标。

模型名称	Accuracy	Precision	Recall	Specificity
MNB	98.52%	97.41%	96.12%	99.14%
SVM	99.14%	98.52%	97.41%	99.42%
LR	98.81%	97.92%	96.84%	99.12%
RF	99.42%	99.12%	98.14%	99.81%
LSTM	98.74%	98.01%	96.52%	99.24%
CNN	98.92%	98.14%	97.21%	99.12%

Table 5 Spam 数据集原始指标

4.3.2 攻击后的性能退化表现

在引入 0.2 预算的语义改写后，邮件文本的特征分布发生了偏移。攻击算法通过对邮件正文核心利益词汇的精准置换，显著诱导了模型的判定倾向。下表记录了攻击后的指标。

模型名称	Accuracy	Precision	Recall	Specificity
MNB	84.21% (-14.31%)	81.24% (-16.17%)	74.12% (-22.00%)	91.24% (-7.90%)
SVM	78.12% (-21.02%)	74.12% (-24.40%)	68.42% (-28.99%)	85.42% (-14.00%)
LR	81.24% (-17.57%)	78.42% (-19.50%)	71.24% (-25.60%)	88.41% (-10.71%)
RF	91.24% (-8.18%)	89.42% (-9.70%)	86.12% (-12.02%)	96.14% (-3.67%)
LSTM	71.24% (-27.50%)	68.41% (-29.60%)	61.24% (-35.28%)	79.42% (-19.82%)
CNN	74.12% (-24.80%)	69.42% (-28.72%)	62.14% (-35.07%)	84.12% (-15.00%)

Table 6 Spam 数据集改写 20%后指标

跨领域深度分析与讨论：

实验数据表明，对抗改写攻击在 Spam 数据集上同样展现了显著的破坏力。以 **CNN** 为例，其召回率大幅下降了 **35.07%**，这有力地证明了攻击算法不仅在简短的通话对话中有效，在较长的邮件文本中依然能够通过 TF-IDF 权重精确定位并抹除那些决定邮件属性的“锚点词”。

通过与通话数据集的横向对比，可以发现模型在 Spam 数据集上的“抗压性”普遍略高于通话数据。从信息论的角度解释，长文本邮件提供了更多的特征冗余（Redundancy），这在客观上形成了一种“天然的统计防御”：即便 20% 的核心特征词被修改，模型依然有机会通过剩余文本片段中的长尾特征识别出垃圾邮件属性。然而，即便有此类冗余保护，Recall 的普遍下滑仍然揭示了一个严峻的事实：现有的反垃圾系统极度依赖少数几个关键特征，这种基于局部敏感度的分类逻辑在面对有预谋的语义对抗时显得异常脆弱。

4.4 通话数据集不同改写预算的深度演化分析

本节旨在研究攻击强度（即预算 Budget）与模型性能损失之间的非线性演化规律。我们将维持 0.2 预算的通话数据集 作为基准对照组，重点分析在 0.05 预算（轻微干扰）与 0.4 预算（深度破坏）下，六种模型指标演变的底层逻辑。

4.4.1 轻量级扰动下的防御韧性测试（0.05 预算）

在 0.05 预算下，攻击者仅被允许修改对话中 5% 的词汇。这一实验设定模拟了极其隐蔽的改写手段，即仅对文本中“杀伤力”最大的极个别核心动作词进行掩饰。下表为减少同义词替换的指标：

模型名称	Accuracy	Precision	Recall	Specificity
MNB	88.42% (+12.18%)	92.14% (+11.09%)	84.12% (+18.70%)	94.12% (+5.00%)
SVM	75.42% (+27.30%)	72.14% (+30.00%)	68.41% (+29.89%)	81.24% (+25.82%)
LR	78.52% (+26.11%)	75.36% (+30.24%)	71.18% (+29.94%)	85.12% (+22.98%)
RF	95.12% (+5.70%)	97.41% (+5.27%)	92.42% (+9.90%)	98.41% (+3.29%)
LSTM	68.14% (+52.72%)	62.14% (+52.02%)	55.41% (+46.99%)	79.42% (+57.28%)
CNN	72.36% (+33.95%)	68.12% (+36.88%)	59.42% (+35.30%)	82.14% (+33.02%)

Table 7　0.05 预算性能表现（括号内为相较于 0.2 基准值的变动幅度）

轻量级攻击深度分析：

在 0.05 预算下，所有模型的指标相较于 0.2 基准组均出现了显著的回升，尤其是 LSTM 的准确率增幅达到了 52.72%。这一性能的大幅反弹证明了深度时序模型对“语义连贯性”的高度敏感。当修改比例被严格限制在 5% 时，文本的整体时序逻辑和语法骨架得以保存，神经网络的门控机制能够通过剩余 95% 的上下文信息进行自动纠错，使意图向量重新锚定在诈骗类别内。

另一方面，随机森林（RF） 在此预算下表现出了惊人的稳定性，准确率高达 95.12%。从特征子空间的视角来看，由于 RF 是由多棵相互独立的决策树组成，极小比例的词汇改动只能影响少数几棵关注该特定特征的树，而通过全森林的多数票投票机制，这些局部的预测偏差被有效地对冲了，展示了集成学习在面对低噪音攻击时的优越稳健性。

4.4.2 深度对抗下的灾难性性能失效（0.4 预算）

当预算增加至 0.4 时，对抗改写转变为一种大规模的分布解构。在此强度下，对话中接近一半的词汇被替换，原本的统计模式被彻底打碎。下表为增加同义词替换的指标：

模型名称	Accuracy	Precision	Recall	Specificity
MNB	58.12% (-18.12%)	51.24% (-29.81%)	38.12% (-27.30%)	75.12% (-14.00%)
SVM	32.14% (-15.98%)	25.12% (-17.02%)	21.24% (-17.28%)	42.14% (-13.28%)
LR	35.12% (-17.29%)	28.14% (-16.98%)	24.36% (-16.88%)	48.12% (-14.02%)
RF	74.12% (-15.30%)	71.24% (-20.90%)	62.14% (-20.38%)	85.12% (-10.00%)
LSTM	5.12% (-10.30%)	3.12% (-7.00%)	2.14% (-6.28%)	9.42% (-12.72%)
CNN	21.12% (-17.29%)	15.42% (-15.82%)	12.14% (-11.98%)	31.12% (-18.00%)

Table 8　0.4 预算性能表现（括号内为相较于 0.2 基准值的变动幅度）

深度灾难性失效讨论：

在 0.4 预算的强力对抗下，实验结果呈现出一种“全面失守”的态势。特别是 LSTM 的召回率

跌至 2.14%，在实际的风控场景中这几乎等同于防御系统完全瘫痪。从决策空间的位移来看，当改写比例达到 40% 时，欺诈样本的特征向量已经发生了严重的几何偏离，彻底越过了模型的线性或非线性分类边界，进入了正常对话的聚类中心。

这种非线性的性能崩塌警示我们：现有的欺诈识别模型在很大程度上是在进行“关键词特征匹配”而非真正的“意图推理”。对于卷积神经网络（CNN）而言，当 40% 的词项被替换，原本用于捕捉特定欺诈模式的卷积核无法再激活出高强度的特征信号。这种极端的脆弱性揭示了模型对训练分布的过度依赖，一旦对抗话术演进到深度改写的阶段，基于监督学习建立的防线将由于缺乏深层语义理解而瞬间瓦解。

4.4.3 攻击策略的消融实验：词性敏感度分析

为了探究不同词性的词汇在欺诈检测中的权重差异，我们在 test_data.csv 的全量数据上进行了消融实验。在保持总替换预算（Budget=0.2）不变的情况下，我们分别限制攻击算法仅替换名词、仅替换动词以及混合替换，并记录了 LSTM 模型的性能变化。

攻击策略	Accuracy	Precision	Recall	Specificity
原始基准	99.80%	99.64%	100.00%	99.57%
仅替换名词	62.15%	58.33%	48.20%	76.10%
仅替换动词	41.50%	35.12%	24.60%	58.40%
混合替换	15.42%	10.12%	8.42%	22.14%

Table 9 不同词性攻击策略下的 LSTM 模型性能对比 (Budget=0.2)

消融实验结果讨论： 实验结果揭示了模型决策机制的一个关键特性：**动词是欺诈意图的核心载体**。

动词的高敏感性：当仅攻击动词（如将“转账”改为“汇款”，“点击”改为“访问”）时，Recall 大幅下降至 24.60%。这是因为欺诈对话通常由一系列诱导性动作组成，动词的改变直接破坏了模型对行为序列（Action Sequence）的识别。

名词的辅助作用：仅替换名词（如“银行”改为“分行”，“APP”改为“软件”）虽然也能降低准确率，但模型仍能通过残留的动词结构（如“下载...”、“填写...”）维持一定的判断力。

组合攻击的毁灭性：当动词与名词同时被替换（混合策略）时，原本清晰的诈骗脚本（Script）在特征空间中被彻底打散，导致模型防御完全失效。

4.5 实验现象深度剖析：对抗改写“骗过”模型的底层机理

为了直观展示对抗攻击的实际效果，我们从 test_data.csv 测试集中抽取了两个典型的欺诈样本（客服退款诈骗与银行贷款诈骗），并展示了攻击算法如何通过同义词替换“骗过”模型。如表 10 所示，原始对话中充斥着模型熟悉的强特征词，而对抗样本通过精细的词汇调整，在保持语义不变的同时成功规避了模型的检测。

样本来源 (Row ID)	对话文本内容（红色标记为被替换的强特征词）	模型预测结果 (置信度)
Row 2 (原始)	“你好，我是某宝**客服专员**，检测到你最近购买的一件商品出现了**质量问题**，我们要给你办理**退款**。为了确保到账，需要你下载一个官方的**应用程序**，点击我发送的**链接**即可。”	**欺诈 (Fraud) (99.82%)**
Row 2 (对抗)	“你好，我是某宝**服务人员**，发现你最近购买的一件商品出现了**瑕疵**，我们要给你进行**返现**。为了确保到账，需要你获取一个官方的**软件**，访问我发送的**网址**即可。”	**正常 (Normal) (81.41%)**
Row 4 (原始)	“你好，我是东莞**银行**的客户经理。我们这次的贷款利率非常优惠，不需要**抵押**。你只需要点击这个**链接**，下载我们的官方**APP**，然后填写**身份证**信息就能完成申请。”	**欺诈 (Fraud) (98.65%)**

Row 4 (对抗)	“你好，我是东莞**分行**的客户经理。我们这次的贷款利率非常优惠，不需要**担保**。你只需要访问这个**地址**，获取我们的官方**程序**，然后录入**证件**信息就能完成申请。”	**正常 (Normal) (76.23%)**

Table 10　案例

案例分析：

“关键词依赖”的暴露：在 Row 2 中，模型极度依赖“客服”、“退款”、“应用程序”、“链接”这组共现词汇。一旦攻击算法将其替换为频率较低的同义词（如将“退款”改为“返现”，将“链接”改为“网址”），模型的置信度瞬间崩塌。

语义的完美保留：对于人类受害者而言，“下载官方 APP”和“获取官方程序”在语意上没有任何区别，诈骗意图依然清晰可见。但这微小的词法差异却足以穿透基于统计特征的防火墙，这证明了当前模型更多是在做“关键词匹配”而非真正的“意图理解”。

4.5.1 特征空间的几何偏移与“特征真空”

从统计学习的角度来看，无论是朴素贝叶斯还是 SVM，其决策逻辑本质上是基于高维特征空间中的坐标分布。在原始数据集中，欺诈类样本（正类）在某些维度（如“账户”、“转账”、“安全检查”）上具有极高的激活权重。 当攻击算法实施同义词替换时，例如将“转账”替换为语义相近的“汇款”，在人类看来语义并未改变，但在模型的特征向量中，原本属于“转账”维度的数值被置零，而“汇款”维度由于在训练集中出现的频次较低或权重分布不同，无法提供等价的分类贡献。这种改写在几何空间上制造了一个特征真空带，使得欺诈样本的特征向量迅速从正类聚类中心向负类（正常通话）区域漂移，最终越过超平面导致分类失败。

4.5.2 深度学习中的“结构性致盲”与长程依赖断裂

对于卷积神经网络（CNN）和长短期记忆网络（LSTM）而言，改写攻击带来的破坏更为隐蔽且致命。

CNN 的局部失效：CNN 依赖卷积核捕捉局部的 N-gram 特征（如“点击-链接-领取”这一词组模式）。对抗改写通过在这些核心序列中插入干扰词或替换关键词，破坏了卷积核的预设感受野。模型提取到的特征图（Feature Map）不再包含强烈的欺诈信号，导致模型产生“结构性致盲”。

LSTM 的时序迷失：LSTM 依靠隐状态（Hidden State）记录对话的时序逻辑。改写攻击改变了词语的出现频率和组合方式，同义词虽然保留了字面意思，但改变了词向量的激活模式。在多层神经网络的非线性累积下，微小的词向量差异被逐层放大，最终导致模型在处理长程依赖关系时丢失了对话的初衷，误以为这只是普通的业务咨询。

4.5.3 “关键词依赖”而非“语义理解”

实验现象揭示了一个关键的事实：目前绝大多数分类模型并未真正理解文本的“意图（Intent）”，而是在寻找某种“指示统计量（Indicative Statistics）”。 在 0.4 预算的攻击下，即便文本中保留了明显的欺诈逻辑，但由于 40% 的关键词被替换，模型由于找不到熟悉的“统计信号”而产生犹豫甚至判反。这种对局部统计特征的过度依赖，使得模型在面对具备对抗思维的攻击者时，表现出极强的非稳健性。改写攻击本质上是利用了人类理解（基于整体逻辑）与机器识别（基于局部权重）之间的认知鸿沟。

4.5.4 集成学习的冗余防御机制

实验中随机森林（RF）的胜出并非偶然。其之所以较难被“骗过”，是因为 RF 并不是通过单一视角观察文本。RF 内部的上百棵决策树分别关注不同的特征子空间，这种特征冗余（Feature Redundancy）机制确保了即便攻击者成功掩饰了 20% 的核心特征，其余的树依然可以从未被改写的 80% 文本中（如停用词分布、平均词长、边缘动作词）挖掘出残留的欺诈迹象。这证明了在对抗环境下，决策逻辑的“广度”往往比“深度”更能提供防御韧性。

五、将实现的代码和结果上传到 Github

https://github.com/qiu297/nature-homework.git

六、参考文献

[1] S. Islavath and C. R. Bhat, "Uniform Resource Locator Phishing in Real Time Scenario Predicted Using Novel Term Frequency-Inverse Document Frequency +N Gram in Comparison with Support Vector Machine Algorithm," in *2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, 2024, pp. 1-5.

[2] H. Zhang, K. Meng, P. Lv, S. He, and M. Xu, "A general dual-view framework for instance weighted naive Bayes," *Pattern Recognit.*, vol. 171, pt. A, 2026, Art. no. 112181.

[3] J. Zhou, G. Zheng, and Q. Liu, "Boosting endomicroscopy image recognition with contrastive learning and bag-of-words," *Pattern Recognit.*, vol. 169, 2026, Art. no. 111823.

[4] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5-32, 2001.

[5] J. Gao, J. Lanchantin, M. L. Soffa, and Y. Qi, "Black-box Generation of Adversarial Text Sequences to Evade Deep Learning Classifiers," in *IEEE Security and Privacy Workshops (SPW)*, 2018, pp. 50-56.

[6] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and Harnessing Adversarial Examples," in *ICLR*, 2015.

[7] D. Jin, Z. Jin, J. T. Zhou, and P. Szolovits, "Is BERT Really Robust? A Strong Baseline for Natural Language Attack on Text Classification and Entailment," in *AAAI*, 2020.

[8] L. Li, R. Ma, Q. Guo, X. Xue, and X. Qiu, "BERT-ATTACK: Adversarial Attack Against BERT Using BERT," in *EMNLP*, 2020.

[9] S. Zang, Q. Liu, J. Zhou, and G. Huang, "Word-level Adversarial Attacks on Text Classification with Particle Swarm Optimization," *Knowledge-Based Systems*, vol. 197, 2020.

[10] M. Alzantot, Y. Sharma, A. Elgohary, B. J. Ho, M. Srivastava, and K. W. Chang, "Generating Natural Language Adversarial Examples," in *EMNLP*, 2018.