|  |  |
| --- | --- |
|  | 学校代码： 10246 |
|  | 学 号： 21262010070 |
|  |  |



|  |
| --- |
| 硕 士 学 位 论 文 |

**（专业学位）**

|  |
| --- |
| **基于大语言模型生成检测电商恶意评论的研究** |

|  |
| --- |
| **Research on Generating and Detecting Malicious E-commerce Reviews Based on LLMs** |

院 系： 软件学院

专业学位类别（领域）： 软件工程

姓 名： 张士鹏

指 导 教 师： 曾剑平

# 目 录

[摘要 4](#_Toc21812)

[Abstract 5](#_Toc1653)

[第一章 绪论 7](#_Toc16000)

[1.1 研究背景 7](#_Toc18666)

[1.2 本文主要工作 8](#_Toc4159)

[1.3 本文章节结构 9](#_Toc19448)

[1.4 本章小结 9](#_Toc21331)

[第二章 相关理论与技术基础 11](#_Toc32413)

[2.1 使用GAN生成文本数据的研究现状与问题 11](#_Toc9698)

[2.2 大语言模型与相关微调技术 11](#_Toc11543)

[2.3 本章小结 11](#_Toc30488)

[第三章 基于Qwen2大语言模型生成对抗网络 12](#_Toc13867)

[3.1 引言 12](#_Toc11746)

[3.2 方法 12](#_Toc17243)

[3.2.1 底座模型选择 12](#_Toc31424)

[3.2.2 qwenGan模型结构 12](#_Toc25615)

[3.2.3 参数高效微调 12](#_Toc21724)

[3.3 训练算法 12](#_Toc10939)

[3.4 本章小结 12](#_Toc25852)

[第四章 实验 13](#_Toc9354)

[4.1 实验设置 13](#_Toc29981)

[4.1.1 实验环境与模型参数 13](#_Toc1819)

[4.1.2 数据集与baseline 13](#_Toc29101)

[4.2 验证方法 13](#_Toc3398)

[4.3 实验结果 13](#_Toc12730)

[4.4 本章小结 13](#_Toc786)

[第五章 总结与展望 14](#_Toc9628)

[5.1 本文总结 14](#_Toc16954)

[5.2 不足与展望 14](#_Toc18398)

[参考文献 15](#_Toc11248)

[学位论文使用授权声明 17](#_Toc19052)

# 摘要

近年来，由于互联网尤其是移动互联网的高速发展，线上购物逐渐成为消费者选购商品的主要平台。随着消费逐步升级，消费者对产品的品质要求越来越高；同时电商需要了解用户喜好和观点，来调整采购、营销的策略。用户评论是获取反馈的最重要的渠道之一。消费者往往会根据商品的用户评论来判断商品的品质，决定是否购买；电商平台也会根据用户评论来针对性地采购商品、改进商品、提升质量。随之而来的层出不穷的恶意评论（包括但不限于刷差评、刷好评、刷误导性的评价等），给消费者和电商都带来困扰。主流的检测电商恶意评论的解决方案主要是情感分析（sentiment analysis）和关键词提取 (Key phrase Extraction，KPE) ，需要依赖大量样本数据集来达到良好的检测效果，但用户评价符合长尾效应，大部分评价内容内容简短、信息量低，有价值、信息量高的文本很少。

为了解决样本过少的问题，主流的解决方案是采用生成对抗网络（GAN）通过对抗训练来生成文本数据。生成对抗网络（GAN）最初是为生成图像设计的，但其原理也可以扩展到文本生成。由于文本数据的离散性，直接将 GAN 用于文本生成面临各种挑战。

近年来，大语言模型迅猛发展，在nlp尤其是文本生成领域表现突出。本文研究了基于大语言模型构建生成对抗任务来生成电商恶意评论数据样本。本文设计了一个生成对抗网络模型，采用 Qwen-2 作为基座模型充当生成器（Generator），利用 Qwen-2 强大的文本生成能力，生成多样化的恶意评价；采用bert作为判别器（Discriminator）来判断文本是真实还是生成的。训练目标是：通过微调 Qwen-2 并结合对抗训练，可以高效实现类似 GAN 的生成对抗任务。最终系统可以生成多样化的恶意评论样本，同时提升判别器对恶意评论的检测能力。

**关键字：**电商评论；恶意评论检测；生成对抗网络；大语言模型；

# Abstract

In recent years, the rapid development of the internet, particularly mobile internet, has transformed online shopping into a primary platform for consumers to purchase goods. As consumer demand evolves, there is an increasing emphasis on product quality. Simultaneously, e-commerce platforms must understand user preferences and opinions to adjust their procurement and marketing strategies effectively. Among various feedback channels, user reviews serve as one of the most crucial sources of information. Consumers often rely on product reviews to assess quality and make purchase decisions, while e-commerce platforms use these reviews to guide product procurement, improve offerings, and enhance quality.

However, the proliferation of malicious reviews—such as fake negative reviews, fake positive reviews, and misleading evaluations—has created challenges for both consumers and e-commerce platforms. Mainstream solutions for detecting malicious reviews primarily rely on sentiment analysis and keyphrase extraction (KPE), which depend heavily on large-scale datasets to achieve effective results. Yet, user reviews exhibit a long-tail distribution: most reviews are short and low in information density, while highly valuable and information-rich reviews are scarce.

To address the issue of limited data samples, a common approach is to utilize Generative Adversarial Networks (GANs) to generate synthetic text data through adversarial training. Originally designed for image generation, GANs have been adapted for text generation, despite challenges arising from the discrete nature of text data.

In recent years, the rapid advancement of large language models (LLMs) has demonstrated exceptional performance in natural language processing (NLP), particularly in text generation. This study explores leveraging LLMs to construct adversarial tasks for generating malicious e-commerce review data samples. Specifically, we designed a GAN-like framework, where Qwen-2 serves as the generator, utilizing its robust text generation capabilities to produce diverse malicious reviews, and BERT acts as the discriminator to distinguish between authentic and generated texts. The training objective involves fine-tuning Qwen-2 and integrating adversarial training to efficiently achieve GAN-like adversarial tasks. The final system is capable of generating diverse malicious review samples while enhancing the discriminator's ability to detect such reviews.

**Keywords**: E-commerce Reviews; Malicious Review Detection; Generative Adversarial Networks; Large Language Models

# 第一章 绪论

## 1.1 研究背景

由于互联网尤其是移动互联网的高速发展，线上购物逐渐成为我们消费者选购商品的主要平台。随着消费逐步升级，消费者对产品的品质要求越来越高；同时电商需要了解用户喜好和观点，来调整采购、营销的策略。用户评论是获取反馈的最重要的渠道之一。消费者往往会根据商品的用户评论来判断商品的品质，决定是否购买；电商平台也会根据用户评论来针对性地采购商品、改进商品、提升质量。随之而来的层出不穷的恶意评论（来自买家或者竞争对手，包括但不限于刷差评、刷好评、刷误导性的评价等），给消费者和电商都带来困扰。主流的检测电商恶意评论的解决方案主要是情感分析（sentiment analysis）和关键词提取 (Key phrase Extraction，KPE) 。

情感分析（sentiment analysis）是一种常见的自然语言处理任务，研究人们在文本中 （如产品评论、博客评论和论坛讨论等）“隐藏”的情绪。主流的方式是将其转换为二分类或者多分类问题，利用正则匹配、机器学习或者深度学习来解决。近些年随着self attention架构的流行，学术界开始采用transformer乃至bert等模型进行情感分析，或者传统的深度学习网络架构基于预训练word2vec或者大规模语料库上预训练模型（如GloVe）进行嵌入层(embedding)表示来完成任务。

关键词提取 (Key phrase Extraction，KPE) 任务用于提取文档中能够概括核心内容的短语。由于对文档进行标注需要耗费大量资源且缺乏大规模的关键词提取数据集，无监督关键词抽取方法在实际应用中更为广泛。常见的算法有基于词袋模型（Bag-of-Words）加权的TF-IDF算法、考虑词关联网络的TextRank算法以及结合语义编码的KeyBert算法等。近年来，命名实体识别(Named Entity Recognition,NER)被用来作为关键词提取的重要方案，用来从文本中识别并标注出具有特定意义的实体。

这些恶意评价的检测任务依赖高质量的数据集。而用户评论较为特殊，相关的数据集通常表现出明显的长尾效应：文本丰富、逻辑清晰的高价值评论较少，更多的是很随意、很简短的评论。生成对抗网络（GAN）可以完成生成样本数据集的任务，一个GAN包含两个主要组件：生成器（Generator）：负责生成“伪造”的文本内容；判别器（Discriminator）：评估生成的文本是否与真实文本相似，并给生成器反馈。生成器的目标是生成看起来尽可能真实的文本，以骗过判别器；而判别器的目标是正确地区分生成的文本和真实文本。生成对抗网络（GAN）最初是为生成图像设计的，在文本生成方面有诸多挑战。主要难点在于：1.离散性问题：文本是离散的符号序列，生成器的输出很难通过连续空间进行梯度回传；2.长序列生成：文本通常是长序列，生成过程中需要考虑上下文一致性；3.评估复杂性：判别器需要判断整个句子的语义和流畅性，而非局部特征。

为了解决上述问题，研究者对 GAN 进行了多种优化和扩展：包括SeqGAN(使用强化学习策略解决文本生成中的离散性问题)、TextGAN（引入了目标分布的最大似然估计（MLE）以增强生成文本的语言质量）、RelGAN（判别器不直接判断生成文本的真假，而是比较生成文本与真实文本的相对距离）等。

然而，目前已有的这些样本生成方法在小样本生成拟合真实数据方面依然有较多缺陷。近两年来,大语言模型发展迅速。大语言模型通过预训练掌握了丰富的语言

知识和领域信息,能够深入理解不同文本的复杂语境和语义表达。与传统

模型相比,大语言模型在深度和广度方面都表现更大出色。因此，基于大语言模型生成检测电商恶意评论的研究具备重要的实用价值。

## 1.2 本文主要工作

为了提升生成恶意评价的数据质量、拟合真实的用户数据，本文基于生成对抗网络的理论基础，提出了一个基于Qwen2大模型的生成对抗网络框架（QwenGan）。具体工作如下：

* 采用qwen2大模型作为生成器，通过微调，让 Qwen-2 生成符合特定需求的对抗性文本。本文研究对比指令微调（Instruction Tuning）和LoRA微调等方式的生成效果，使用SwanLab来监控整个训练过程，并评估最终的模型效果。
* 采用bert作为判别器，区分真实文本和生成文本，通过通过交叉熵损失优化判别器，不断优化模型。相对于qwen2，BERT 是轻量化的、任务专用的预训练模型，非常适合作为判别器使用。BERT是为理解任务设计的，擅长句子分类和语义判断，电商评论等通常是短文本，BERT的输入长度和架构足够应对这些任务。
* 循环训练：

使用 Qwen-2 生成对抗文本；

判别器对生成文本和真实文本进行分类；

更新判别器参数；

利用判别器的反馈优化 Qwen-2 的生成能力。

* 本研究在国内主要的电商平台的用户评价数据集上做了可行性验证。通过在包括JD.com E-Commerce Data和中文淘宝评论等多个真实数据集上进行实验。结果表明，与传统的文本类生成对抗网络相比，qwen2大语言模型在拟合真实用户评论方面表现更优秀。

本文的主要创新点在于：利用 Qwen-2 的强大生成能力，生成多样化、高质量的对抗样本。同时针对Qwen-2训练成本高、推理速度慢的缺点，提出采用bert作为判别器，有助于降低整体计算成本和在实际项目种快速迭代和部署。

## 1.3 本文章节结构

本文一共分为六个章节,具体安排如下:

第一章:本文的研究背景及意义。通过对比现有的相关工作总结出现存的不足与挑战，提出基于Qwen2大模型的生成对抗网络框架（QwenGan），阐述基于大语言模型生成电商恶意评论的研究的研究意义和现实价值。然后介绍了本文的创新点,最后概括了本文的整体结构。

第二章:相关工作。主要介绍生成对抗模型在文本生成领域的相关工作,一方面介绍了现有的一些生成模型,以及它们生成文本的主要原理及流程,另一方面总结了基于不同生成模型的一些现有的文本数据生成方法,并对一些比较典型的方法进行了简要介绍。

第三章:QwenGan模型的整体架构。从现有模型存在的问题及挑战出发,引出了本文的研究动机,并分析了现有的生成模型框架,总结出了本文提出的模型QwenGan的框架基础。然后对本文方法框架中的细节展开进行介绍,分别为整体框架流程、生成器架构、判别器架构、损失函数和训练算法。

第四章:本章是论文的实验部分,详细介绍了本文的软硬件环境及选用的数据集和baseline。然后展开介绍了本文选用的各种评价指标,最后根据实验结果分析验证了本文提出的模型QwenGan的先进性。

第六章:该章节主要对全文内容进行了总结,最后展望在QwenGan以及现有其他生成对抗模型的基础上用户评论数据生成的研究方向。

## 1.4 本章小结

本章为绪论部分。第一节首先介绍了本文的研究背景及意义,强调了研究电商平台用户恶意评价数据生成的重要性。然后简单阐述了目前文本数据生成算法和研究现状,总结出了现有算法存在的缺陷与不足,提出采用大模型来优化生成用户评价的任务。第二节介绍了本文提出基于Qwen2大模型的生成对抗网络框架（QwenGan）网络结构以及创新点 ,最后第三节概括了本文的整体结构。

# 第二章 相关理论与技术基础

## 2.1 使用GAN生成文本数据的研究现状与问题

## 2.2 大语言模型与相关微调技术

## 2.3 本章小结

# 第三章 基于Qwen2大语言模型的生成对抗网络

## 3.1 引言

## 3.2 方法

### 3.2.1 底座模型选择

（介绍选取qwen2-instruct的原因）

### 3.2.2 qwenGan模型结构

（介绍生成器和判别器）

### 3.2.3 参数高效微调

（介绍多种微调方法，LoRA、Instruction Tuning等）

参数高效微调 即PEFT(Parameter-Efficient Fine-Tuning)，它仅更新模型中的小部分参数，保持大部分预训练权重不变，在保持模型性能的同时减少所需的计算资源和存储空间，可在有效避免过拟合问题的同时，还有助于保留模型在广泛任务上的通用知识（即：泛化能力）。

## 3.3 训练算法

## 3.4 本章小结

# 第四章 实验

## 4.1 实验设置

### 4.1.1 实验环境与模型参数

### 4.1.2 数据集与baseline

（介绍JD.com E-Commerce Data和中文淘宝评论等多个真实数据集）

## 4.2 验证方法

## 4.3 实验结果

## 4.4 本章小结

# 第五章 总结与展望

## 5.1 本文总结

## 5.2 不足与展望

# 参考文献