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电商恶意评论的研究

Research on Generating 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

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，大语言模型不需要调整即可直接完成很多 NLP 任务。

不过，大模型生成文本时，往往偏向于生成概率较高的样本（如语言模型预测的高概率词序列），这可能导致生成结果较为单一或模式化。大模型生成的评论：可能倾向于使用固定模式，如“产品很好，值得推荐”；而通过生成对抗训练的模型可能生成多样化表达，如“整体质量满意，但包装有待改进”。因此，基于大语言模型构建生成对抗网络进行文本生成任务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 Qwen-2 大模型的生成对抗网络框架（QwenGAN）。具体工作如下：

- 采用 Qwen-2 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。本文研究对比各种微调方式（如 LoRA 等）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 Qwen-2，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、单纯的大语言模型相比，本文提出的 QwenGAN 在拟合真实用户评论方面表现更优秀。
- 设计、开发一个基于 QwenGAN 的用户评价数据生成 web 系统，通过可视化的操作界面，支持上传样本、运行模型生成、预览、下载生成的数据。

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。同时提出采用 BERT 作为判别器，进行对抗训练，提升生成文本的拟真程度。

1.3 本文章节结构

本文一共分为五个章节,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 Qwen-2 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。

第五章:本章是论文的系统设计部分,详细介绍了基于 QwenGAN 设计、开发的用户评价数据生成 web 系统，通过可视化的操作界面，展示 QwenGAN 的实用性。

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 Qwen-2 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

本章主要介绍使用GAN生成文本数据在行业内的发展与主要工作，然后介绍大语言模型在生成文本方面的优势以及相关的微调技术。首先介绍一系列GAN生成文本数据的模型，如SeqGAN、TextGAN、RelGAN，并通过对比总结出各个模型的特点。然后介绍几种主流的大语言模型以及微调技术。

2.1 使用GAN生成文本数据的研究现状与问题

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。

2.1.1 SeqGAN

SeqGAN (Sequence Generative Adversarial Network[1])是由Lantao Yu等人提出的首个将生成对抗网络（GAN）应用于序列生成任务的模型，提出了如何解决 GAN 在生成离散序列（如文本）时的梯度问题。传统 GAN 适合生成连续数据，而文本是离散的，这导致无法直接进行梯度传播。SeqGAN 通过强化学习（Reinforcement Learning, RL）解决了这一难题。SeqGAN 的核心方法是，生成器基于 RNN（通常是 LSTM 或 GRU），逐步生成序列中的每个 token。生成过程被视为一个马尔科夫决策过程（MDP），生成器的目标是最大化整个序列的奖励值。判别器通过 CNN 或其他分类网络区分生成序列和真实序列。它为生成器

提供训练信号，即序列是否接近真实分布。强化学习框架：判别器输出的评分被用作生成器的奖励信号。SeqGAN 使用策略梯度（Policy Gradient）方法优化生成器，使其生成的序列能够骗过判别器。SeqGAN 的特点是通过将序列生成问题转化为强化学习任务，避免了因文本离散性导致的梯度消失问题。缺点是生成序列时对采样的依赖较强，训练效率较低；随序列长度增加，生成器容易陷入模式崩溃（mode collapse）。

2.1.2 TextGAN

TextGAN[2] 针对文本生成任务提出了一种基于 GAN 的端到端解决方案。它的目标是通过最小化生成样本分布与真实数据分布之间的差异，实现高质量的文本生成。TextGAN 的核心方法是生成器基于 LSTM 结构，生成完整的文本序列，直接建模文本的联合分布。判别器通过一个特征提取器（如卷积层或 RNN）捕捉文本的语义特征，并利用最大均值差异（Maximum Mean Discrepancy, MMD）衡量生成样本与真实样本的分布差异。TextGAN 优化了特征匹配目标的方法，不仅仅依赖判别器的二分类结果，还利用 MMD 对生成样本的全局分布进行优化，从而提升文本生成的多样性和质量。TextGAN 的特点是全局分布对齐，使用 MMD 衡量生成样本和真实样本在高维特征空间的分布差异；生成样本质量高，相较于 SeqGAN，TextGAN 的生成器能更好地学习到真实数据的复杂分布。缺点是 MMD 的计算复杂度较高，训练时间较长；对短文本生成效果较好，但在长文本生成任务中表现有限。

2.1.3 RelGAN

RelGAN[3] 是一种针对文本生成任务优化的 GAN 模型，旨在解决 SeqGAN 和 TextGAN 存在的训练效率低下及生成样本多样性不足的问题。它通过引入基于关系的生成器和优化策略，显著提升生成质量。RelGAN 使用一个基于关系建模的生成器（Relation-aware Generator），通过注意力机制建模序列中 token 之间的关系，从而生成更符合上下文逻辑的序列。判别器使用多分类结构，能够输出多维得分，用于衡量生成样本的质量。RelGAN 提出一个优化策略，即结合强化学习和对抗训练，将生成任务分解为多个子任务（如局部生成和全局生成），提高训练效率和稳定性。RelGAN 的特点是基于关系建模，注意力机制显著提升了生成文本的上下文一致性。与 SeqGAN 相比，RelGAN 通过强化学习优化长序

列生成，避免了模式崩溃问题。RelGAN 的多分类判别器提升了生成样本的多样性。RelGAN 的缺点是需要更多的超参数调节，模型复杂度较高。

2.2 预训练语言模型

Vaswani 等人在 2017 年的论文《Attention Is All You Need[4]》中提出 Transformer 架构，它克服了 RNN 和 CNN 在处理长序列时的局限性，彻底改变了自然语言处理（NLP）和其他序列任务的研究方向。Transformer 的核心思想是通过自注意力机制（Self-Attention）建模序列中任意位置的依赖关系，并完全摒弃了循环(RNN)或卷积(CNN)结构。Transformer 主要分为两个部分：encoder 编码器和 decoder 解码器，两个部分各自包含 6 个 block，每个 block 包含多头自注意层。Encoder 负责将输入序列编码成上下文感知的表示，decoder 则使用这个表示来生成目标序列，自注意力机制则在这两个部分种起到了帮助模型捕捉长距离依赖关系的作用。

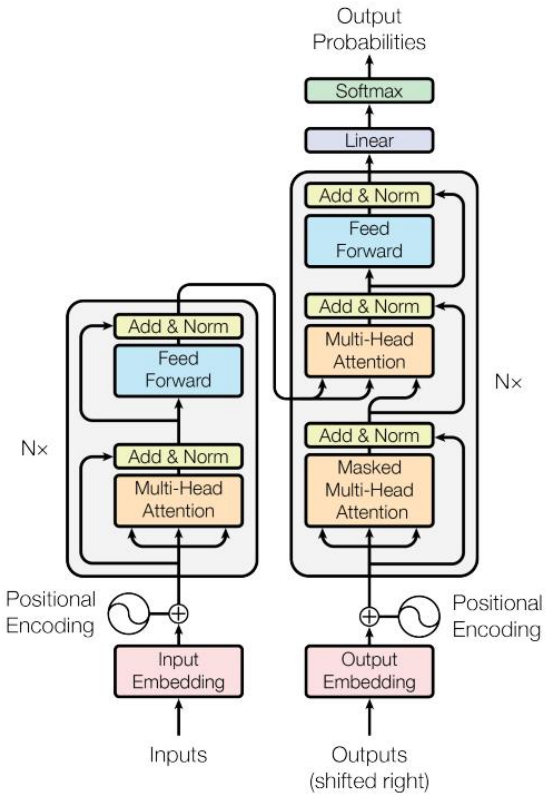

图 2-1 Transformer 架构图

Transformer 问世在 NLP 任务上取得了巨大的成功，其效果远远超越过去的 CNN 和 RNN。此后最热门的语言模型 BERT、GPT 系列、RoBERTa、BART 以及风靡一时的大语言模型如 chatgpt、chatGLM、Qwen-2 等都采用了 Transformer 架构。

2.2.1 BERT

BERT[5] (Bidirectional Encoder Representations from Transformers)是一种专注于语言理解的双向预训练模型，其核心基于 Transformer 的编码器架构，能够同时建模句子中前后文的语义关系。BERT 的预训练通过两个任务完成：1.掩码语言模型 (Masked Language Model, MLM)：随机掩盖输入中的部分 token，模型需要基于上下文预测被掩盖的 token，从而学习双向语义；2.下一句预测 (Next Sentence Prediction, NSP)：通过预测两个句子是否连续，建模句间关系，适用于文本对任务。BERT 的设计使其在多种下游任务中表现优异，如文本分类、命名实体识别和问答系统等。然而，由于其核心在于理解语言，而非生成语言，BERT 通常不直接用于生成对抗任务。BERT 因其易用性和优良的表现，使得其在很多中小型项目中广泛应用，也因此出现了多种变体，如 RoBERTa(优化了 BERT 的训练策略，性能更高)、DistilBERT（轻量化版本，适合资源受限环境）。本文采用 BERT 作为判别器网络模型。

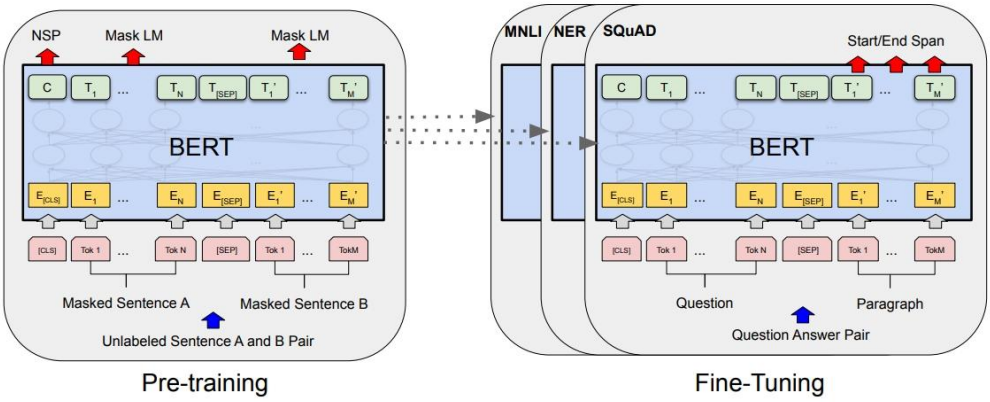

图 2-2 BERT 预训练和微调过程

2.2.2 GPT 系列

GPT[6] (Generative Pre-trained Transformer) 系列模型由 OpenAI 开发，是以生成任务为核心的预训练语言模型系列。这些模型的发布标志着自然语言处理（NLP）领域从特定任务定制模型向通用预训练模型的转变。GPT-1 基于 Transformer 的解码器结构，由 12 层 Transformer 堆叠而成，采用自回归语言建模任务，目标是预测当前 token 的下一个 token。GPT-1 首次在生成任务中引入大规模预训练 + 小样本微调的范式。GPT-2 提升了模型规模与生成质量。GPT-3 是通用语言模型的飞跃，规模大幅增长，拥有 1750 亿参数，成为当时最大的语言模型之一。支持 Few-shot、Zero-shot 学习，可以通过提示（Prompt）直接完成任务，无需微调，大幅降低定制化成本。GPT 系列的技术核心：1. 基于自回归生成方式，从左到右逐词生成文本，确保语言生成的自然流畅性；2. 预训练 + Prompt 学习，GPT 模型在大规模无监督语料上预训练，再通过 Prompt 指令或少量示例实现特定任务的泛化；3. Transformer 解码器架构，多头注意力机制捕获序列中的长距离依赖关系，位置编码建模词序列的顺序；4. Few-shot 学习与大规模模型，模型规模与任务泛化能力呈正相关，通过扩大参数量，显著提升任务表现。

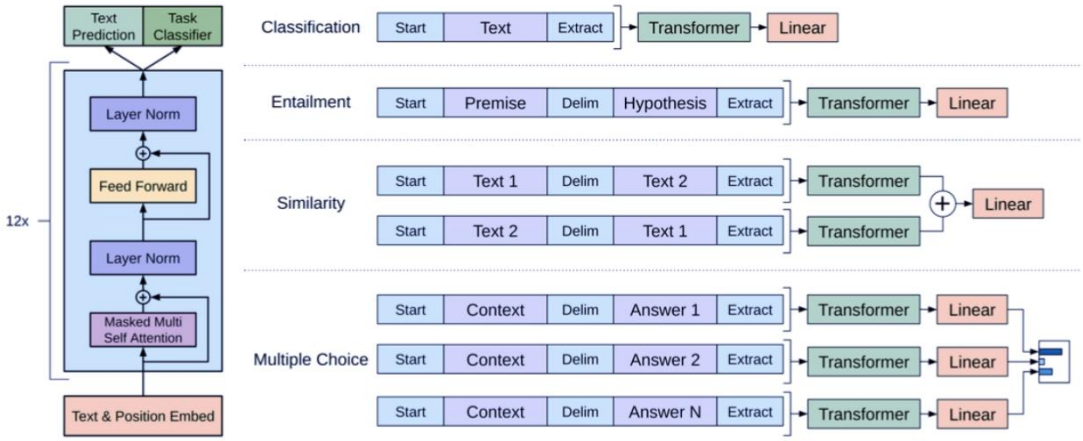

图 2-3 GPT-1 模型架构图

GPT 系列从 GPT-1 到 GPT-4 实现了从基础语言生成到通用任务执行的跨越，其生成能力和泛化性已广泛应用于内容生成、智能对话、代码生成等领域。GPT 系列的成功为语言模型的研发提供了重要方向，也为后续模型（如 ChatGLM、Qwen-2）的发展奠定了基础。

2.2.3 ChatGLM 与 Qwen-2

自 chatgpt 火爆全球后，各种大语言模型（Large Language Models，LLMs）如雨后春笋般涌现。Google 的 T5、meta 的 LLaMA 是其中的佼佼者，他们在 NLP 任务如机器翻译、摘要生成、智能客服中表现非常亮眼。但这些模型并未将中文语料加入预训练，因此在中文任务上的表现远逊于英文。由于本文主要着眼于国内电商平台，因此优先选择国内的开源 LLMs，其中清华大学和智谱联合开发的 ChatGLM、阿里巴巴集团 Qwen 团队研发的 Qwen（通义千问）系列最为流行。

ChatGLM-6B 使用了和 ChatGPT 相似的技术，针对中文问答和对话进行了优化。经过约 1T 标识符的中英双语训练，辅以监督微调、反馈自助、人类反馈强化学习等技术的加持，62 亿参数的 ChatGLM-6B 已经能生成相当符合人类偏好的回答。Qwen-2 大语言模型最新版本已升级至 Qwen-2.5 版本，无论是语言模型还是多模态模型，均在大规模多语言和多模态数据上进行预训练，并通过高质量

数据进行后期微调以贴近人类偏好。Qwen 系列一个明显的优势是与阿里云深度集成，可以便捷地整合到应用服务落地使用，是国内互联网电商的首选。

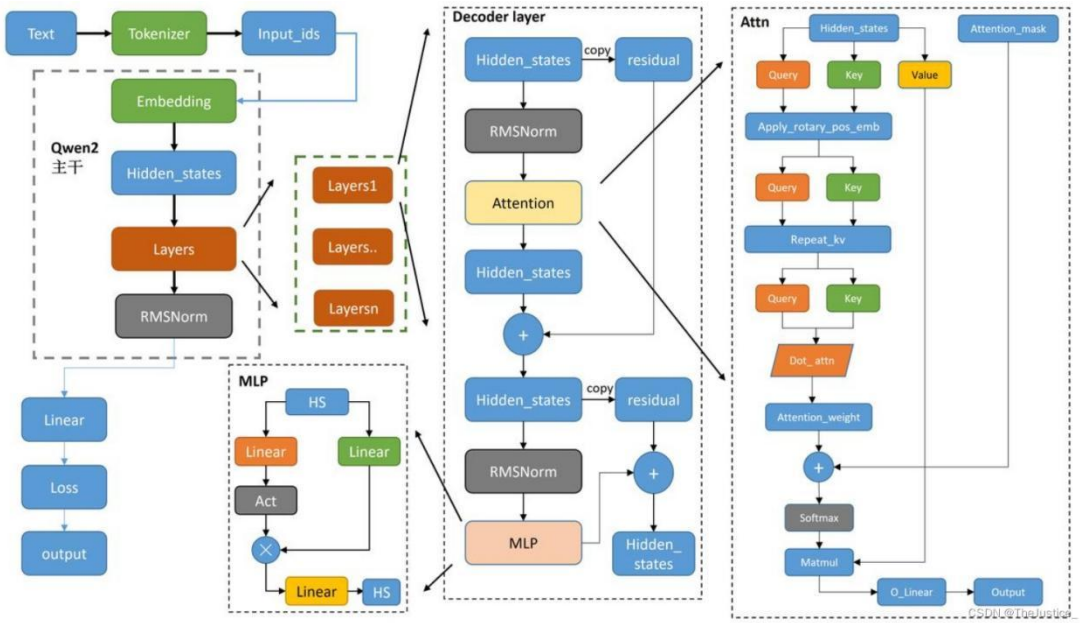

图 2-4　Qwen-2 模型架构图

2.3 大语言模型微调技术

大模型微调（Fine-Tuning）是指在预训练的大语言模型（LLMs）的基础上，通过使用特定任务的数据进行进一步训练，以优化模型的表现，使其适应某一特定领域或任务。其核心思想是，预训练大模型已经掌握了通用语言知识，但可能缺乏领域知识或对某些任务的精确理解，微调引入领域数据可弥补这些不足。从资源利用角度来说，微调比从头训练一个模型更高效，显著节约时间和计算资源。大模型微调方式有全参数微调 (Full Fine-Tuning，FFT)和参数高效微调 (Parameter-Efficient Fine-Tuning，PEFT)两种。由于 LLMs 模型规模庞大，在民用显卡上进行 FFT 越来越变得不切合实际，并且对一项下游任务做 FFT，无论是部署还是存储，成本都非常高昂。因此，目前主流的微调方式和研究方向都在 PEFT 方向，并且最先进的 PEFT 取得的性能表现，与 FFT 相差无几。以下介绍几种主流的 PEFT 方式。

2.3.1 LoRA

LoRA (Low-Rank Adaptation)基本原理是利用权重矩阵的低秩分解假设，通过在 Transformer 的权重矩阵中插入一个低秩的可训练模块，并冻结原始权重，显著减少需要调整的参数量。具体来说，将权重矩阵 WWW 分解为两个低秩矩阵 A 和 B：W′ =W+ΔW=W+A· B 。其中 A 和 B 是可训练的低秩矩阵。训练中，只需要微调新增的低秩参数 A 和 B，减少内存和存储需求。图 2-5 展示了 Lora 微调的重新参数化过程。在这个过程中,我们只训练低秩矩阵 A 和 B,而保持预训练权重矩阵 W 不变。微调后参数模块可独立保存，适用于多任务切换。LoRA 适用于资源受限的任务。

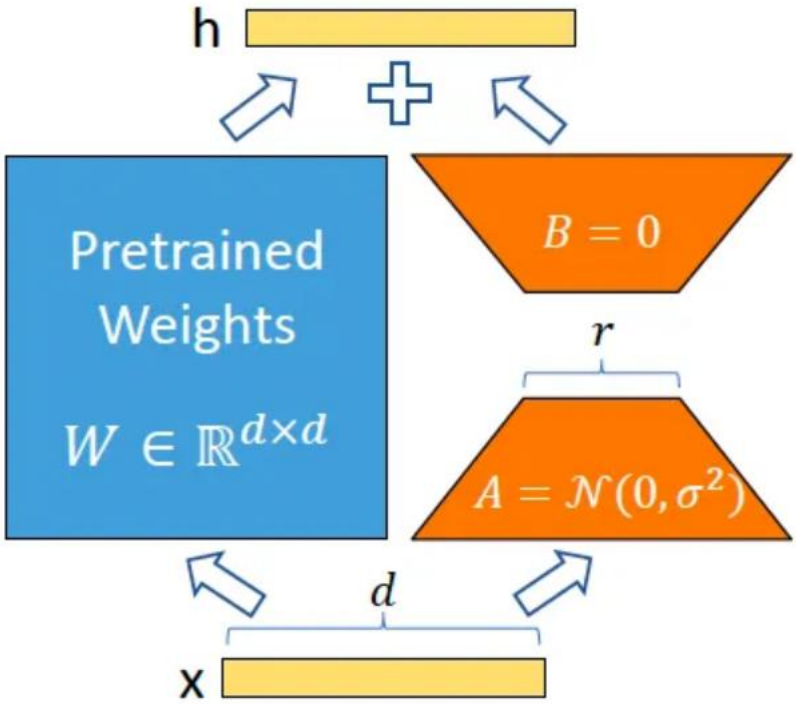

图 2-5 LoRA 微调简介图

以 Qwen-2-7B-Instruct 为例，模型参数量 70 亿，LoRA 在指定 lora_rank=8，target_modules=[query_key_value]的时候，仅需要训练参数量为 200 万，所以只需要相对较少的硬件资源即可完成任务。

2.3.2 Adapter Tuning

Adapter Tuning（论文：Parameter-Efficient Transfer Learning for NLP[7]）的作者希望使用预训练模型的时候不需要重新训练整个模型，因此提出了一种 transfer learning 的方法。常规的 NLP 中做 transfer learning 的技术主要是 feature-based transfer 和 fine-tuning，作者提出了另一种方式 adapter module。该方法设计了 Adapter 结构，并将其嵌入 Transformer 的结构里面，针对每一个 Transformer 层，增加了两个 Adapter 结构(分别是多头注意力的投影之后和第二个 feed-forward 层之后)，在训练时，固定住原来预训练模型的参数不变，只对新增的 Adapter 结构和 Layer Norm 层进行微调，从而保证了训练的高效性。每当出现新的下游任务，通过添加 Adapter 模块来产生一个易于扩展的下游模型，从而避免全量微调与灾难性遗忘的问题。

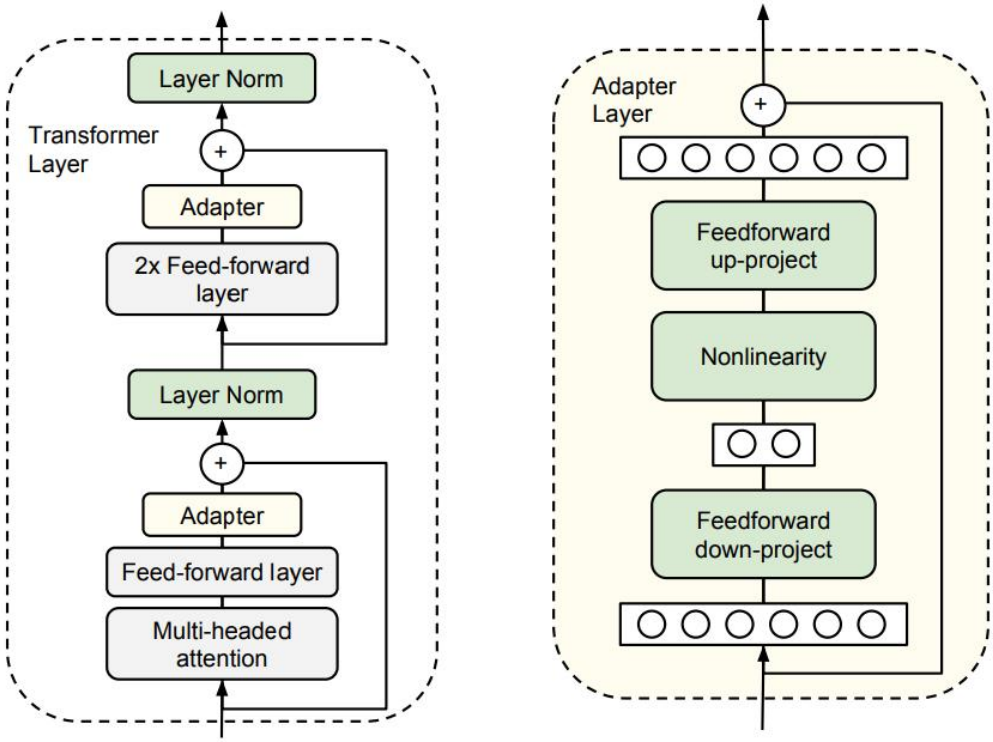

图 2-6 Adapter Tuning 网络结构

2.3.3 P-Tuning

大模型的 Prompt 构造方式严重影响下游任务的效果。比如：GPT-3 采用人工构造的模版来做上下文学习（in context learning），但人工设计的模版的变化特别敏感，加一个词或者少一个词，或者变动位置都会造成比较大的变化。同时，近来的自动化搜索模版工作成本也比较高，以前这种离散化的 token 的搜索出来的结果可能并不是最优的，导致性能不稳定。基于此 Xiao Liu 等人提出 P-Tuning（论文：GPT Understands, Too[8]），该方法设计了一种连续可微的 virtual token（同 Prefix-Tuning 类似，不过 P-Tuning 加入的可微的 virtual token，但仅限于输入层，没有在每一层都加），将 Prompt 转换为可以学习的 Embedding 层，并用 MLP+LSTM 的方式来对 Prompt Embedding 进行一层处理。从对比实验证实看出，P-Tuning 获得了与全参数一致的效果。甚至在某些任务上优于全参数微调。P-Tuning v2[9]在 P-Tuning 基础上进一步做了优化，该方法在每一层都加入了 Prompts tokens 作为输入，更多可学习的参数（从 P-tuning 和 Prompt Tuning 的 0.01%增加到 0.1%-3%），同时，也足够参数高效。

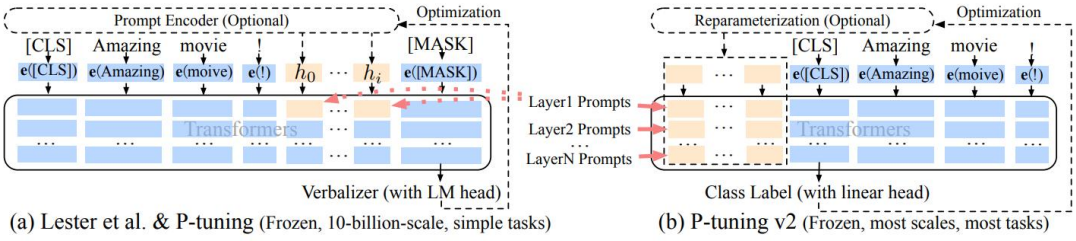

图 2-7 P-Tuning 与 P-Tuning v2 任务流程图

2.4 本章小结

本章主要介绍了使用 GAN 生成文本数据的研究现状以及生成式网络的发展。重点介绍了 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）等网络。然后介绍了大语言模型相对于传统模型在的优势，列举了几种预训练模型（尤其是国内的中文大语言模型）以及它们在生成式任务中的优秀表现。最后

介绍了预训练大语言模型的几种微调方式，为后续的本文提出的模型以及实验提供了理论基础。

第三章 基于预训练大语言模型的生成对抗网络

本章提出用 Qwen-2 做生成器和用 BERT 做判别器构建一个 QwenGAN 的生成对抗网络架构。Qwen-2 与 BERT 的组合充分发挥了生成器与判别器各自的优势。在生成器端，Qwen-2 强大的生成能力与多样化表达保障了高质量的恶意评论样本生成；在判别器端，BERT 的深度语言理解能力和鲁棒性为生成对抗网络提供了可靠的判断基础。这种组合不仅能够显著提升生成对抗任务的性能，还能够通过多样化样本生成与精确判别，全面优化电商恶意评论检测的效果。

3.1 方法

3.1.1 底座模型选择

在生成对抗网络的框架中，生成器的主要任务是生成多样化且逼真的文本数据。本文选择 Qwen-2 作为生成器模型的底座，其具体原因如下：

1. 强大的文本生成能力。Qwen-2 是一个专为中文及多语言场景优化的大语言模型，具备强大的语言理解和生成能力。得益于其大规模的预训练数据集和先进的 Transformer 架构，Qwen-2 在文本生成的流畅性、语义一致性以及上下文理解方面表现卓越。对于恶意评论生成任务，Qwen-2 能够生成多样化、高信息量的恶意评论，覆盖不同场景和风格。

2. 高效微调支持。Qwen-2 提供了一系列参数高效微调技术，如 LoRA、Prefix Tuning 等。通过这些技术，可以在保持基础模型强大能力的同时，以较低的资源成本微调生成器，使其更贴合电商恶意评论的特定任务需求。此外，这些微调技术还可以显著提升模型在小样本场景中的性能。

3. 适配性强与资源友好。Qwen-2 的架构经过优化，能够高效适配生成对抗任务所需的大量训练迭代，同时对算力需求相对较低。最新的 Qwen-2.5-7B-Instruct 拥有 70 亿参数，采用 LoRA 仅需要 16GB～24GB 显存，采用 P-Tuning 和 Adapter Tuning 可以将 GPU 显存要求降低到 16GB。Qwen 由阿里巴巴团队研发，与阿里云深度集成，可以非常方便地满足互联网企业的项目落地运行。

在生成对抗网络中，判别器的核心作用是区分文本是真实的还是生成的。无论是 Qwen 系列还是其他生成式预训练大语言模型，他们的参数量过大，可能会记住生成器的模式，而不是学会泛化；在生成对抗网络训练过程中，生成器和判别器需要频繁交替训练，资源消耗过大；相比轻量模型，Qwen-2 等生成式预训练大语言模型在推理时的速度较慢，尤其在实时或大规模应用场景下，不利于后续的部署和落地应用。本文选择 BERT 作为判别器的底座模型，主要基于以下考虑：

1. 强大的文本表示能力。BERT 是一种双向 Transformer 结构，通过 Masked Language Model（MLM）和 Next Sentence Prediction（NSP）的预训练目标，能够捕捉文本的全局上下文信息。这种特性使得 BERT 在捕捉复杂语义和深层次语言特征方面表现突出，非常适合作为恶意评论检测任务中的判别器。

2. 鲁棒性与通用性。BERT 在多种文本分类任务中表现出卓越的鲁棒性和通用性，其对噪声文本的识别能力较强。在恶意评论检测中，BERT 能够高效捕捉生成文本的语言模式，从而为生成对抗网络提供准确的判别信号。

3. 丰富的社区支持与优化技术。BERT 已成为自然语言处理领域的基础模型之一，其优化和部署技术非常成熟。借助丰富的开源资源和技术文档，可以大幅降低模型开发和调试的难度。

3.1.2 生成器的设计与策略

Qwen-2.5-7B-Instruct 在生成器中承担文本生成任务，其核心策略包括：

1.多样性文本生成。Qwen-2.5-7B-Instruct 通过微调适配于恶意评价场景，在生成过程中引入随机性（如采样策略、Top-k 或 Top-p 采样）以生成多样化的文本。模型以电商评论模板和负面语义词汇为输入上下文，通过扩展性强的语言生成能力生成恶意评论。例如：

> 输入提示：“生成一个关于物流延迟的负面评价”
> 输出评论：“物流真的太慢了，完全超出预期，客服态度也很敷衍！”

2. 风格化生成。利用电商场景中的领域语料对生成器进行微调，使其生成的文本更加符合实际用户的语言风格。例如：

某些恶意评论可能模仿真实用户的语气，如：“从来没有见过这么差的质量，退货都不给！”

3. 条件生成控制。在生成过程中引入条件控制变量（Condition Variables），如商品类别、评论情感强度等，确保生成文本的多样性和相关性。例如：

输入条件：商品类别 =“电子产品”；负面程度 =“高”。
生成评论：“这个耳机质量太差了，没用两天就坏了，真的不值这个价格！”

3.1.3 判别器的设计与策略

BERT 在判别器中承担检测生成文本真实性的任务，核心策略包括：

1. 文本真实性判别。判别器通过区分生成文本和真实文本来反馈生成器，优化生成器的生成质量。训练过程中，BERT 被微调以捕捉生成文本中可能存在的语义不一致性、逻辑漏洞或语言模式异常。

2. 特征提取与分类。BERT 通过其双向 Transformer 架构，捕捉评论的语义特征和上下文关系，从而精确判断评论是否由生成器生成。输出层采用二分类策略，给出真实文本或生成文本的概率。

3. 对抗训练增强。在训练过程中，判别器与生成器不断对抗。判别器提升自身对生成文本的识别能力，生成器则通过改进生成策略，生成更加逼真的文本。

3.1.4 生成对抗网络的整体策略

生成对抗网络结合生成器和判别器，通过交替训练实现优化：

1. 初始训练阶段。为生成器提供一组带标注的电商评论数据（包括真实和恶意评论），利用 Qwen-2.5-7B-Instruct 生成初始恶意评论样本。判别器 BERT 在这批数据上进行初始训练，学会区分真实与生成的文本。

2. 对抗训练阶段。生成器优化：生成器生成恶意评论，并通过判别器反馈调整生成策略，使生成文本更加逼真且多样化。判别器优化：判别器在生成器生成的样本和真实数据上继续训练，提高其对复杂生成文本的识别能力。

3. 损失函数设计。生成器损失函数：用判别器的反馈概率值作为生成器的优化目标，最大化判别器判断为“真实”的概率。判别器损失函数：使用交叉熵损失（Cross-Entropy Loss）区分真实文本和生成文本。图 3-1 展示损失函数的公式：

$$L_D = -\mathbb{E}_{x_r \sim p_{\text{real}}}[\log D(x_r)] - \mathbb{E}_{x_g \sim p_{\text{gen}}}[\log(1 - D(x_g))]$$

$$L_G = -\mathbb{E}_{x_g \sim p_{\text{gen}}}[\log D(x_g)]$$

图 3-1 生成对抗网络损失函数

模式崩溃是 GAN 的常见问题，表现为生成器仅能生成有限的模式，无法覆盖真实分布的多样性。为了解决这一问题，在判别器中引入梯度惩罚项（由 WGAN-GP 提出[10]），确保生成对抗网络满足 Lipschitz 连续性，有助于生成器更全面地学习真实数据分布。

4. 收敛目标。当生成器生成的恶意评论难以被判别器区分时，模型达到收敛状态，表明生成的文本具有高质量和多样性。

3.2 训练算法

QwenGAN 的算法训练如图 3-2 所示，算法的输入是一份真实数据 Dreal ，算法的输出是一个训练好的生成器 G θ 。训练的流程为首先初始化生成器和判别器的参数，然后固定生成器去训练判别器，每训练一次生成器，判别器就需要训练 k 次。在判别器训练过程中需要先处理采样随机噪声 B 和真实数据，然后根据随机噪声用生成器得到生成样本，之后更新判别器参数。训练完判别器之后再训练生成器，固定判别器的参数，利用随机采样得到噪声向量 z，然后利用生成器得到生成数据，使用判别器来判别这些生成数据的结果，然后更新生成器的参数。一个完整的 Epoch 训练完，返回生成器的参数。

算法具体流程如下：

1. 初始化：

- 生成器 G θ 基于 Qwen-2.5-7B-Instruct，通过预训练和微调为生成评论提供强大的自然语言生成能力。

- 判别器 Dϕ 基于 BERT，通过分类任务的预训练初始化，用于判断评论的真实性。

2. 对抗训练：
   - **判别器更新**：判别器通过优化交叉熵损失，学习如何区分真实评论和生成评论。
   - **生成器更新**：生成器通过对抗性学习，提高生成评论的多样性和真实性，使其难以被判别器区分。
3. 正则化约束：
   - 在判别器中引入梯度惩罚项，确保生成对抗网络满足 Lipschitz 连续性，防止模式崩溃。
4. 循环训练
   - k 步判别器更新后执行一次生成器更新，确保判别器有足够的能力对生成样本进行判断。
5. 输出：最终优化后的生成器 G_θ 可用于生成多样化的电商恶意评论样本。

Algorithm 1: QwenGan Training for E-Commerce Reviews

Input: Dataset $D_{\text{real}} = \{x_i\}_{i=1}^{N}$ containing real e-commerce reviews
Output: Trained generator G_θ

T: Number of training iterations
k: Number of discriminator updates per generator update
η_D: Discriminator learning rate
η_G: Generator learning rate
B: Batch size
P_z: Noise distribution
λ: Regularization coefficient for gradient penalty

Initialize generator G_θ and discriminator D_ϕ with pre-trained weights;
for $t = 1$ *to* T **do**
  for k *steps* **do**
    Sample B noise vectors $\{z_1, z_2, \ldots, z_B\} \sim P_z$;
    Sample B real reviews $\{x^1_{\text{real}}, x^2_{\text{real}}, \ldots, x^B_{\text{real}}\}$ from D_{real};
    Generate fake reviews $x_{\text{fake}} = \{G_\theta(z_1), G_\theta(z_2), \ldots, G_\theta(z_B)\}$;
    Update discriminator parameters ϕ:
    $\phi \leftarrow \phi - \eta_D \cdot \nabla_\phi \left[\frac{1}{B} \sum_{i=1}^{B} \left(\log D_\phi(x^i_{\text{real}}) + \log(1 - D_\phi(x^i_{\text{fake}})) \right) \right]$;
    Enforce gradient penalty for Lipschitz constraint:
    $\phi \leftarrow \phi - \eta_D \cdot \lambda \cdot E_x \left[\left(\|\nabla_x D_\phi(x)\|_2 - 1 \right)^2 \right]$;
  Sample B noise vectors $\{z_1, z_2, \ldots, z_B\} \sim P_z$;
  Generate fake reviews $x_{\text{fake}} = \{G_\theta(z_1), G_\theta(z_2), \ldots, G_\theta(z_B)\}$;
  Update generator parameters θ:
  $\theta \leftarrow \theta - \eta_G \cdot \nabla_\theta \left[\frac{1}{B} \sum_{i=1}^{B} \log D_\phi(G_\theta(z_i)) \right]$;
return G_θ;

图 3-2 QwenGAN 训练流程图

3.3 本章小结

本章介绍 Qwen-2 做生成器和用 BERT 做判别器构建一个 QwenGAN 的生成对抗网络架构。首先介绍了用 Qwen-2 做生成器模型和用 BERT 做判别器模型的各自优势。然后介绍了生成器、判别器以及整个生成对抗网络的策略和设计，之后介绍整个网络的损失函数。最后详细介绍了网络的算法训练流程。

第四章 实验设计

本章介绍实验的软硬件环境、数据集以及实验需要比较的基准模型等实验设置。随后详细说明本文研究所采用的验证方法，以评估 QwenGAN 模型的方法可行性和有效性。

4.1 实验设置

4.1.1 实验环境

实验的硬件环境目前为：Gen Intel(R) Core(TM) i5-12600KF CPU 3.70 GHz，64GB 运行 RAM，GPU NVIDIA GeForce RTX 4070（TODO 后续会更换到更高性能的 gpu 环境）；软件环境为：操作系统 Windows 11 专业版，NVIDIA CUDA 12.6.20，编程语言 Python 3.10.14，深度学习框架 Pytorch 2.4.0。

4.1.2 数据集与 baseline

本文会使用多个数据集进行实验，目前包括 JD.com E-Commerce Data[11]和中文淘宝评论两个真实数据集（TODO 由于电商用户评价属于公司的重要信息资产，故很难找到高质量的数据集来源，后续会在找更多的数据集进行实验）。JD.com E-Commerce Data 如图 4-1 所示，包括 52 万件商品，1100 多个类目，142 万用户，720 万条评论/评分数据。中文淘宝评论数据集[12]包含食品、鞋子、儿童服装、女性服装、珠宝首饰、男性服装、户外用品、建筑材料、办公用品、行李箱等多个分类的评价数据。这些数据集都包含了热门的电商商品品类的用户真实评论。如前文所述，其中高质量的数据并不多，因此对 QwenGAN 的模型效果表现要求更高。

用户ID	商品ID	评论时间戳	评论标题	评论内容
1013654	PRODUCT_176056	1382025600	东西不错	大三元之一 东西看上去不错,包装也都很好,关键是价格比京东便宜很多。 还没试过,回去试一下。 不足是不能开增票。
99935	PRODUCT_130680	1296144000	这么丰富的经	这么丰富的经历没写出来,对于我们以后上哪玩挺有帮助,作为游记一般吧。
307768	PRODUCT_323370	1303142400	很喜欢 支持原	很喜欢 支持离歌 支持饶雪漫~~
152011	PRODUCT_383545	1313510400	内容空洞,不值	内容很空洞,有炫富意味,其它的倒还真没看出什么所以然来。很后悔买了这本书。完全想废纸一样。
1070630	PRODUCT_346185	1272556800	爱自己多一点	这个书的内容总的来说不错的,书名有点夸张,但看了内容后,发现真的很实实在在的,一点也不夸大。本人特别喜欢后面部
1133263	PRODUCT_247806	1336060800	易懂,好用	程博士写的书易懂好用!
42055	PRODUCT_82381	1324742400	火机油	收到时外包装没问题,但奇怪的是里面瓶身上角有些挤变形了,还好没破,没有泄漏。除去包装外,满意。
1433	PRODUCT_457115	1338134400	不错的书	不错的书,价格合适,质量还行
650346	PRODUCT_348453	1337097600	翻译它最大	很喜欢里面的翻译讲解,用四步定位来解决每一个翻译题,屡试屡爽!
1033284	PRODUCT_184929	1358611200	吊丝女自强指	是不是看这种书的女人都是每人爱的剩女?恐龙?凭一本书真的就能改变命运吗?
155035	PRODUCT_512582	1364486400	还行	感觉像是一本过去的教科书,书的封面质量很一般。因为是公司需要买,所以无所谓
648574	PRODUCT_20267	1368547200	最满意的一次	包装很好,日期也很新,到2016年。价格也实惠,特价是买的,很开心,不过还没用,不知道效果咋样,涂开来是没有味道的我很
2117	PRODUCT_263736	1320249600	快递很给力,书	印刷排版很精美,内容比较言简意赅,是本不错的小励志,值得读一读
572968	PRODUCT_462423	1282665600	非常好	非常满意,看到这样的好书
1226714	PRODUCT_311882	1300636800	还好	还好,就是没想象的那么多,解析比较详细
75472	PRODUCT_81456	1382544000	不错的书,经典	书的内容和纸张质量都没得说
89091	PRODUCT_237923	1323705600	鼠标一般般	鼠标刚到了两天,一般般。觉得确实宽了些有些不适应,中间的滚轮容易被按下去,没有图片看上去那样有质感,其他的暂时
194708	PRODUCT_495102	1365609600	考过了,哈哈	六级好好准备就会过的。
38757	PRODUCT_445933	1324569600	通俗易懂	书中讲到了很多概念性的东西,但是老雕并没有做过深的讲解。
1249974	PRODUCT_328938	1244908800	品质上乘,颜色	收音机很快送来了。 在订单中注明了希望是黑色的,本以为如果不是黑色,至少也应该是灰色的吧。打开包装时就想着千万
892	PRODUCT_207816	1304092800	每一次相遇都	写得很美很真实,很羡慕作者能游历各地,观赏美景,寻找自己心中的宁静。 强烈推荐这本书,在现代的繁忙中找到旅游的情
5442	PRODUCT_88730	1279555200	经典	现代人写的前言特搞笑。毛泽东,马克思,受不了。
535170	PRODUCT_511127	1318262400	喜欢	从这本书可以了解到另一层面的社会热点,真的就好像是一问一世界,每一个问题都会把你带入一个不一样的视角去了解世
1311203	PRODUCT_421436	1331481600	不太好	书没打折,不值!配送慢!送不到就别显示送达时间!书封面有折痕,不太满意!
355382	PRODUCT_279945	1374508800	好的。	印刷质量好,满意!!
11532	PRODUCT_50605	1331222400	不错的选择	书很不错 发货特别快 强力推荐卓越亚马逊
726231	PRODUCT_341915	1348761600	满意	因为我之前一直用手提电脑,而这个键盘打字的手感就跟手提差不多,所以我不会觉得有什么不适应的,所以我建议你如果你
443899	PRODUCT_151508	1383926400	纸张好	内容好,包装好,送货快
716359	PRODUCT_44501	1365350400	有道理	她的文字会让你自然地联系到自己在生活中所遇到过的情况,可能会恍然大悟,当时自己这么做或者那么做原来是因为什么
10167	PRODUCT_390407	1368892800	真心贵,画得也	非常非常贵,而且看起来就像是别人看过的旧书,而且画得也不好,跟我在网上看的不一样,而且第一页中三只小熊是共同坐
4597	PRODUCT_124001	1322668800	很不错	很不错很不错很不错很不错很不错很不错很不错很不错很不错很不错很不错很不错
78218	PRODUCT_325599	1351008000	还可以	煮过一顿饭,觉得挺方便的,口感也不错。之前没用过电压力锅,它工作时有不大不小的哨子似的声音,会从限压阀冒些水泡

图 4-1 JD.com E-Commerce Data

本文验证模型的 Baseline 分两个方法：

1. 选取对文本生成对抗表现比较好的几种算法模型与 QwenGAN 进行比较，包括 SeqGAN、TextGAN、RelGAN。这三个模型在文本生成领域表现出色，但也存在一些问题没用解决。与这三个模型对比，评估多样性、真实度和任务适应性，可以证明 QwenGAN 的算法可行性。

2. 与大语言模型直接生成结果对比。即不使用对抗训练，而直接通过 Qwen-2.5 或类似模型微调生成恶意评价。验证对抗训练是否提升了生成数据的质量或增强了模型对恶意评论的检测能力。

4.1.3 SwanLab 可视化训练

SwanLab 对训练细节的监控十分强大，提供可视化界面，直观呈现训练过程。SwanLab 本身不能直接支持 Qwen-2.5-7B-Instruct 的训练过程，但通过模型转换或与其他工具联合使用，可以间接实现训练的监控与优化。在生成对抗网络的训练过程中，SwanLab 在以下方面特别有用：

1. 模式崩溃检测：生成器输出单一模式时，通过判别器损失曲线分析问题。

2. 梯度惩罚验证：确保模型满足 Lipschitz 连续性，尤其在使用梯度惩罚时，检测是否产生了过拟合或梯度消失问题。

3. 样本分布对比：通过对生成样本的特征分布分析，评估生成器输出的质量和多样性。

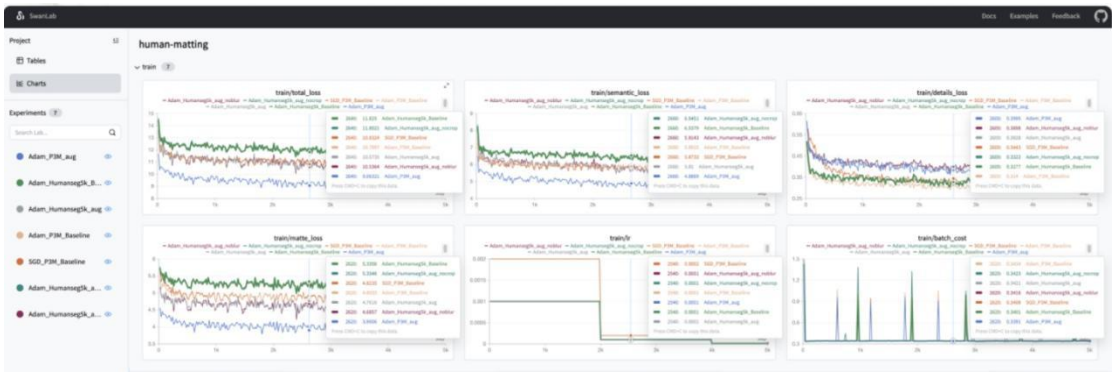

图 4-2 SwanLab 可视化训练界面

4.2 验证方法

4.2.1 机器学习性能验证方法

机器学习任务一般分为分类任务与回归任务两种，针对这两种任务的评价方法基本一致，区别在于不同的任务细节指标不尽相同。在评估文本生成在机器学习任务的性能表现时，首先将真实数据分割为训练（train_data_set）数据和测试数据（test_data_set）；然后在训练数据上训练生成模型并输出与训练数据等量的生成数据，然后在生成数据上执行分类任务；最后在测试数据上计算模型准确性，通过在生成数据上训练的算法模型在测试数据上表现的性能来评估数据的质量。对于回归任务会用到的指标主主要有 Fl-score、R2、MSE 等,对于分类

任务的指标主要有准确率、召回率、ROC 曲线、AUC 等。这两种在真实数据和虚假数据上的评价指标结构流程如图 4-3 所示。

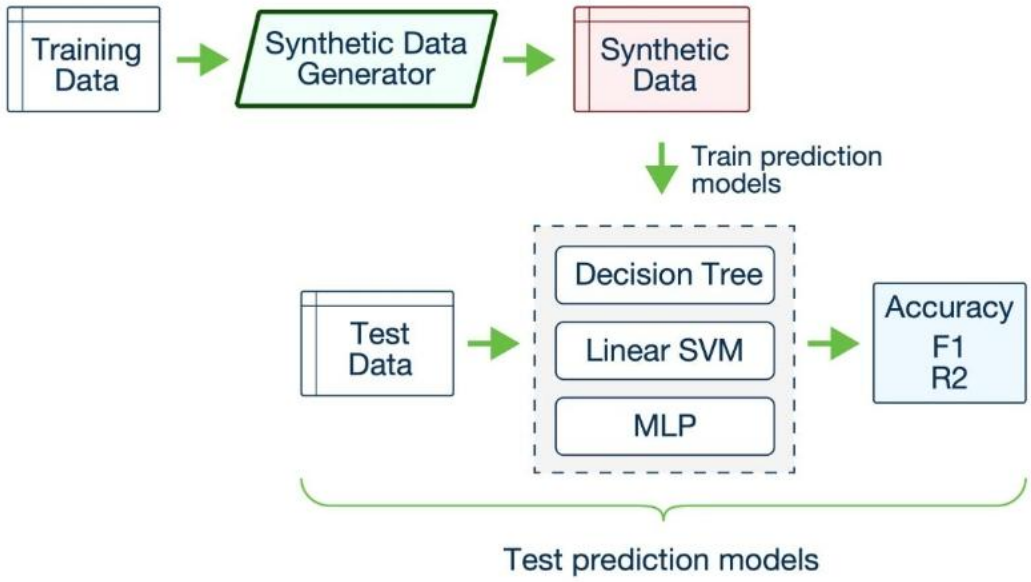

图 4-3 机器学习性能评估结构

4.2.2 生成数据质量验证

1. 真实性和可读性。主要靠人类评价（Human Evaluation），通过对生成文本的流畅性(Fluency)、逻辑性(Coherence)、语义合理性(Semantic Relevance)多个维度进行评估。

2. 多样性。可以通过去重率（Distinct-n，公式参考图 4-4）：计算生成文本中不同 n-gram（例如 bi-gram 和 tri-gram）占比。更高的去重率表明文本更具多样性。

$$\text{Distinct-n} = \frac{\text{Unique } n\text{-grams in generated samples}}{\text{Total } n\text{-grams in generated samples}}$$

图 4-4 Distinct-n 公式

3. 与真实数据的相似性。指标包含 BLEU、ROUGE、BERTScore 等。其中 BLEU 和 ROUGE 计算生成文本与真实数据在词汇和短语上的相似性;BERTScore 利用预训练的 BERT 模型，基于语义嵌入计算生成文本与真实文本的相似度。

4.3 实验结果

4.3.1 机器学习性能实验结果

TODO 目前实验证明本文提出的方法可行，不过效果不达预期，还在优化。实验过程中因为显存等问题出现一些错误。后续会更换更好的硬件进行实验，完成算法模型训练和测试，包含生成器和判别器的微调等等，以及几种算法的对比。

4.3.2 消融实验

每次移除或修改一个组件，保持其他部分不变，并记录对生成数据质量的影响。本文设计的消融实验包括以下几组：

1. 移除梯度惩罚项。实验设置：在判别器的损失函数中移除梯度惩罚项。假设：没有梯度惩罚项时，模型可能失去 Lipschitz 连续性，导致生成器崩溃（如模式崩溃或生成样本单一）。评估指标：①FID/BERTScore：衡量生成数据与真实数据的相似性。②模式崩溃率（Mode Collapse Rate）：生成样本的多样性降低率。

2. 替换生成器。将生成器替换为未经过微调的 Qwen-2.5-7B-Instruct。假设：未微调的生成器或较小的模型可能生成的文本质量较差，表现为语义不连贯或生成结果不多样化。评估指标：人类评价和 Distinct-n。

3. 替换判别器。将判别器替换为未进行微调的 BERT。假设：判别器的能力下降会导致生成器无法学习到有效的生成策略。评估指标：判别器的分类准确率和生成文本的真实度。

TODO 实验完成过程中会进行消融实验验证。

4.4 本章小结

本章为实验部分，首先介绍了实验的硬件和软件环境。之后介绍数据集和 SwanLab 可视化训练工具。接着列举了验证模型性能表现的一些指标，包括常规的机器学习下游任务的性能评估（如准确率、召回率、ROC 曲线、AUC）和生成数据质量验证评估（如真实性和可读性、多样性、与真实数据的相似性）。实验结果表明 QwenGAN 生成的用户评论数据相比其他算法模型生成的数据实验效果更好。最后通过消融实验，验证模型架构中各模块的必要性以及量化不同组件对最终生成质量和任务性能的贡献。

第五章 系统设计

本文提出的 QwenGAN 算法模型旨在生成拟真的恶意用户评价数据，扩充真实数据集，为下游任务提供数据样本支持。因此会设计一个系统，用户可以上传一份数据，然后系统会生成一份生成的样本数据，并支持下载到本地。

5.1 系统概述

基于 QwenGAN 开发一个包含前后端的 web 端电商恶意用户评价数据生成系统。该系统集成了 QwenGAN 的模型训练、推理生成样本数据等功能。系统的总体流程是：系统集成了 TabTransGAN 的模型训练、数据合成等功能。系统的总体流程为:(1)用户上传需要生成的评论数据,支持上传 csv 文件格式，系统会展示数据的预览(2)用户选择可以选择系统内的数据集或者上传的数据集进行样本生成(3)基于已经选择的数据集点击生成操作,服务器自动生成数据并提供前端页面列表预览展示,用户也可以点击下载按钮获取合成数据集。

系统设计到的技术主要有:编程语言 Python 3.10.14,深度学习框架 Pytorch 2.4.0、Flask2.3.2、、Node 10.16.0、Vue 4.5.14、axios 1.5.0 等。

5.2 系统页面展示

TODO 页面会包含文件上传、列表选择、通过按钮生成、任务进度展示、生成数据列表预览、文件下载等功能。系统设计还在开发完善中。

5.3 本章小结

本章介绍了基于QwenGAN的用户评价数据生成web系统。该系统为通用的web前后端技术开发。通过页面交互而不是枯燥的命令行程序，提升了用户使用体验。首先介绍了系统设计的概况以及主要模块，随后对系统的整体流程进行分别详细介绍，最后展示系统的实际运行效果，包含文件预览、通过按钮生成数据、数据集预览等功能，用户也可以直接在页面下载生成的数据集。

第六章 总结与展望

6.1 本文总结

NLP 任务中的情感分析和关键词提取可以解决常见的电商平台恶意用户评价检测问题，但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，所以生成、扩充高质量样本是非常重要的任务。GAN 网络在 CV 领域生成、合成图片表现非常优异，因此很多研究者选择利用 GAN 网络的思路来生成文本，诸如 SeqGAN、TextGAN、RelGAN 等网络都做了这方面的尝试，但都有各自的不足。

最近几年，大语言模型的兴起将 NLP 的文本生成任务带到了新的高度。大语言模型可以执行通用任务，对自然语言的理解远超过以前的模型。不过，大模型生成文本时，结果较为单一或模式化。本文研究了如何利用预训练大语言模型构建一个生成对抗网络，旨在以较少的数据集，生成拟真的电商恶意用户评价数据，提升平台的恶意评价的检测能力。本文首先介绍了几种文本生成对抗网络的方法以及特点和不足，然后介绍几种预训练大语言模型以及微调方法。然后提出一种以 Qwen-2 为生成器、BERT 为判别器的 QwenGAN 文本生成网络，介绍了模型底座的选择、网络结构和整体策略、数据集和算法性能评价体系，并通过实验证明该模型算法的可性能与有效性，通过消融实验证明微调生成器模型和判别器模型的意义。

最后，基于上述研究，本文设计并实现了一个包含前后端的 web 端电商恶意用户评价数据生成系统，方便用户上传、下载、预览、管理样本文件。

6.2 不足与展望

本文提出的通过预训练大语言模型构建生成对抗网络生成电商恶意评论，依然有很大的改进空间，包括训练的硬件性能要求导致训练的成本较高，在新的数据集上如果表现不佳，则需要继续训练。此外，本文选择 Qwen-2 作为生成器的基座模型是因为它背靠阿里云，实用价值更高，国内大预言模型发展日新月异，今后会有更好更合适的模型基座替代。

参考文献

[1] SeqGAN: Sequence Generative Adversarial Nets with Policy Gradient Lantao Yu, Weinan Zhang, Jun Wang, Yong Yu https://arxiv.org/abs/1609.05473

[2] Evaluating Text GANs as Language Models Guy Tevet, Gavriel Habib, Vered Shwartz, Jonathan Berant https://arxiv.org/abs/1810.12686

[3] ReLGAN: Generalization of Consistency for GAN with Disjoint Constraints and Relative Learning of Generative Processes for Multiple Transformation Learning Chiranjib Sur https://arxiv.org/abs/2006.07809

[4] Attention Is All You Need Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin https://arxiv.org/abs/1706.03762

[5] BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova https://arxiv.org/abs/1810.04805

[6] Generative Pre-trained Transformer: A Comprehensive Review on Enabling Technologies, Potential Applications, Emerging Challenges, and Future Directions Gokul Yenduri, Ramalingam M, Chemmalar Selvi G, Supriya Y, Gautam Srivastava, Praveen Kumar Reddy Maddikunta, Deepti Raj G, Rutvij H Jhaveri, Prabadevi B, Weizheng Wang, Athanasios V. Vasilakos, Thippa Reddy Gadekallu https://arxiv.org/abs/2305.10435

[7] Parameter-Efficient Transfer Learning for NLP Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, Sylvain Gelly https://arxiv.org/abs/1902.00751

[8] GPT Understands, Too Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, Jie Tang https://arxiv.org/abs/2103.10385

[9] P-Tuning v2: Prompt Tuning Can Be Comparable to Fine-tuning Universally Across Scales and Tasks Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng

Lam Tam, Zhengxiao Du, Zhilin Yang, Jie Tang
https://arxiv.org/abs/2110.07602
[10] Wasserstein GAN with Gradient Penalty Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, Aaron Courville
https://arxiv.org/abs/1704.00028v3
[11] WWW.JD.com E-Commerce Data
https://aistudio.baidu.com/datasetdetail/96333
[12] 中文淘宝评论数据 https://aistudio.baidu.com/datasetdetail/94812