Modern Linguistics 现代语言学, 2025, 13(8), 1127-1138
Published Online August 2025 in Hans. https://www.hanspub.org/journal/ml
https://doi.org/10.12677/ml.2025.138940

Hans汉斯

通过大型语言模型增强训练平衡小数据集中的稀有语言特征

李扬天[1]，杨泽娟[2]

[1]南加州大学维特比工程学院，美国 洛杉矶
[2]江西行政学院党史党建教研部，江西 南昌

收稿日期：2025年7月4日；录用日期：2025年8月18日；发布日期：2025年8月29日

摘　要

稀有语言特征(如隐性线索、结构性否定或低频修饰语)由于在训练语料中的频率极低，长期以来是自然语言处理中的一大挑战。为此，本研究提出一种基于大语言模型(LLMs)的特征导向数据增强方法，旨在有效捕捉并强化训练集中代表性不足的语言现象。以隐性否定为具体案例，我们设计了一个双阶段的数据增强流程：(1) 围绕稀有否定线索生成结构多样化的训练样本；(2) 构造反事实句对以抑制背景偏差，从而凸显模型对隐性否定等关键语言特征的敏感性。在CONDAQA数据集上的实验结果表明，通过LLMs增强训练集，显著提高了RoBERTa模型对隐性及结构复杂否定表达的识别性能。本研究进一步证实，LLMs可作为一种可控且高效的数据增强工具，在低资源情境下有效地再平衡稀有语言现象的训练数据。

关键词

隐性否定，数据增强，大型语言模型，语言泛化，低资源语言现象

Balancing Rare Linguistic Features in Small Datasets through LLM-Augmented Training

Yangtian Li[1], Zejuan Yang[2]

[1]Viterbi School of Engineering, University of Southern California, Los Angeles, USA
[2]Department of Party History and Party Building, Jiangxi Administrative College, Nanchang Jiangxi

Received: Jul. 4th, 2025; accepted: Aug. 18th, 2025; published: Aug. 29th, 2025

Abstract

Rare linguistic features—such as implicit cues, structural negation, or low-frequency modifiers—present a persistent challenge in NLP due to their sparsity in training corpora. This study proposes a feature-targeted data augmentation framework leveraging large language models (LLMs) to surface and

文章引用: 李扬天, 杨泽娟. 通过大型语言模型增强训练平衡小数据集中的稀有语言特征[J]. 现代语言学, 2025, 13(8): 1127-1138. DOI: 10.12677/ml.2025.138940

amplify such underrepresented language phenomena. Taking implicit negation as a case study, a two-stage augmentation pipeline is introduced: (1) generating structurally diverse training samples centered on rare negation cues, and (2) constructing counterfactual sentence pairs to mitigate spurious background biases, thereby enhancing model sensitivity to critical linguistic features. Experiments with RoBERTa on the CONDAQA dataset demonstrate that our LLM-augmented training significantly improves the model's ability to recognize implicit and structurally complex negation. These findings confirm that LLMs can serve as controllable and efficient tools for rebalancing the data distribution of rare linguistic phenomena, particularly in low-resource settings.

Keywords

Implicit Negation, Data Augmentation, Large Language Models, Linguistic Generalization, Low-Resource Language Phenomena

1. 引言

稀有语言特征(如隐性线索、结构性否定)因其在训练语料中的低频性与分布不均，长期以来是自然语言处理(NLP)模型面临的重要挑战。此类特征出现频率远低于主流语言模式，导致模型易过拟合浅层语言信号，难以泛化到语义细腻且结构复杂的表达中。

否定现象尤其是隐性否定，因其缺乏明确的词汇标记、依赖抽象语义和多样句法结构，是研究语言泛化难题的理想诊断任务。因此，否定检测既是研究语言泛化的典型任务，也为探索如何在语料层面对稀有语言特征进行系统性平衡提供了理想案例。本文将否定检测作为代理任务，以深入探讨在低资源条件下平衡稀有语言特征的通用策略。

传统语料如 SFU Review 多以显性否定(如 not、never)为主，结构单一；而 CONDAQA 则包括 hardly、unaddressed 等更丰富的隐性否定形式，但数量有限，难以有效训练模型。尽管已有研究尝试通过简单的极性翻转或词汇替换缓解数据不足问题，但这些方法常引入语义噪声，无法捕捉深层结构特征。

为此，本研究提出一种基于大型语言模型(LLMs)的结构化数据增强框架，包括线索扩展与反事实构造两个阶段，以增强训练数据中隐性否定的频率和结构多样性。基于 RoBERTa 模型的实验结果显示，LLM 驱动的增强方法显著提高了模型对隐性否定表达的泛化性能，证明 LLM 在小规模、不平衡语料环境下，能够有效平衡和放大稀有语言特征。

2. 文献综述

稀有语言特征，如形态结构和隐性语义线索，在自然语言理解任务中提出了独特挑战。这些特征在标准数据集中普遍代表性不足，模型训练时往往忽视其重要性。Henning 等[1]对深度学习驱动的自然语言处理系统中的类别不平衡问题进行了系统综述，指出其对模型泛化能力和公平性均有不利影响。Hofmann 等[2]进一步证明，标准预训练语言模型在处理形态复杂或派生词形式时表现不佳，凸显出模型训练中结构性稀疏的问题。Gururangan 等[3]揭示了自然语言推理(NLI)数据中的注释伪影，这些伪影会扭曲语言特征的分布，使模型更依赖表层相关性而非深层语言线索。上述研究共同强调了一个核心问题：文本数据中具有语言丰富性的信号在频率和分布上均呈现出严重不均，亟需更有针对性的策略加以解决。

为缓解数据不平衡并增强模型对稀有语言现象的鲁棒性，研究者提出了多种数据增强方法。传统策略如 EDA [4]与 HotFlip [5]主要引入词汇层面的扰动，但往往缺乏句法与语义的忠实性。Kaushik 等[6]提出了反事实增强策略，通过在保持语法与语义连贯的前提下更换样本标签，显著提升了模型的鲁棒性。

近年来，大型语言模型(LLMs)逐渐被用于生成高质量、多样化的训练样本[7] [8]。这类方法超越了表层扰动，能够生成具有语义控制性的增强样本，包括对比性或结构导向的变体。Gururangan 等[9]进一步表明，若在特定任务数据分布上进行持续预训练，可提升模型对稀有或特定领域模式的感知能力。在这一趋势基础上，Dai 等[10]提出了 AugGPT，利用 ChatGPT 生成语义一致、结构多样的释义样本。类似地，Qu 等[11]提出了 CoDA，一种对比正则化增强框架，通过受控转换(如回译)促进特征感知的多样性。这些方法共同强调，在泛化稀有语言特征时，结构、对比与线索级控制的重要性不容忽视。

否定，尤其是隐性否定，是一种语言表达丰富却频率极低的典型挑战。在 NLP 中，否定作为稀有语言结构的代表，非常适合用于检验各类增强策略对低代表性线索的建模能力。Hossain 等[12]系统分析了多个主流语料中的否定现象，指出隐性否定线索常被遗漏标注，且建模方式不一致。Poliak 等[13]与 Ravichander 等[14]则指出，模型可能通过利用数据集伪影在否定相关任务中取得意外的高分，而并未真正掌握其背后的语言结构。Shaitarova 与 Rinaldi [15]研究了跨语言、零样本条件下的否定范围解析，发现通用语言表示往往难以识别否定范围边界。Fancellu [16]对多语言否定范围识别进行了深入探讨，进一步佐证了否定在不同语言和语料中均具有结构上的难以建模性。与此同时，Truong 等[17]提出一种结合掩码与增强的否定感知预训练策略，以提升模型对隐性否定的敏感性。

尽管已有研究在理解和解决数据不平衡问题方面取得了进展，但如何系统性地利用大型语言模型提高模型对结构隐性语言线索的泛化能力尚未得到充分探索。本研究旨在填补这一研究空白，提出一种以否定为研究案例、基于 LLM 的结构化数据增强方法，期望有效提升模型对稀有语言现象的识别与泛化能力，并探索语言模型如何超越表面词汇分布，捕捉深层次的语言结构信号。

3. 研究方法与隐性否定建模策略

3.1. 不平衡语料中隐性否定的建模挑战

自然语言中的否定表达呈现出丰富多样的语言线索。其中，显性否定(如 not、never、no)具有明确的句法标记，在训练数据中频繁出现；而隐性否定在词汇表达上更多样化，语义更为细腻，但在语料中的分布严重不足。这种不平衡导致模型训练时易过拟合于显性表达，难以有效泛化至结构与表达方式不同但语义相同的否定形式。此前研究已指出类似的不平衡问题在自然语言推理任务中存在[2] [3]。

为探讨此类结构偏斜的影响，本文选取两个在否定分布上存在显著差异的语料库进行分析：SFU Review 语料库[18]与 CONDAQA 数据集[14]。

SFU 语料包含超过 16,000 条带注释的影评句子，其中约 18%含有否定表达，主要为显性否定(如 not、never、no)。图 1 清晰展示了线索分布的高度单一，反映其结构多样性极为有限。

相比之下，CONDAQA 数据集在结构与否定形式上更加多元化，包含了 1289 对对比句，涉及多种否定策略，包括隐性形式如 hardly、unaddressed、absence of 等。图 2 显示了其在否定线索上的相对平衡。

为开展数据增强实验，本文从 CONDAQA 中筛选出 100 个 SFU 中未出现的隐性否定线索，并进行人工校验。这些线索涵盖多个语义领域，作为种子词被用于借助大型语言模型生成约 5000 条合成句子(具体过程参见下文方法介绍)。

以下例句展示了两个语料在否定表达方面的对比：

SFU 例句(多为显性否定)：

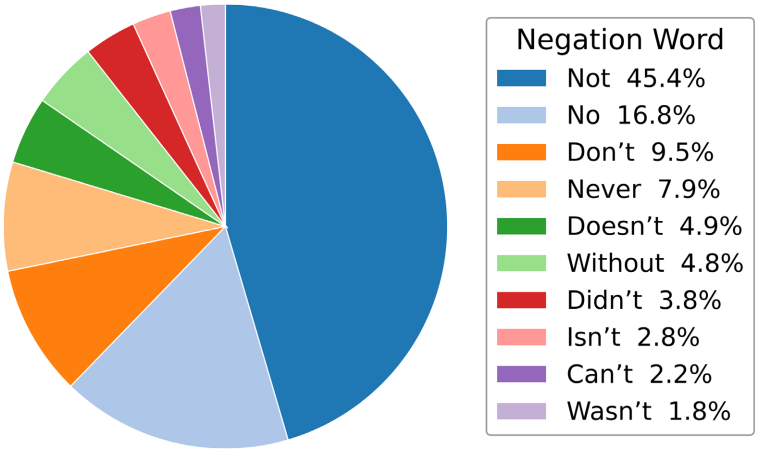

Figure 1. Distribution of negation cues in SFU corpus
图 1. SFU 语料否定线索分布情况

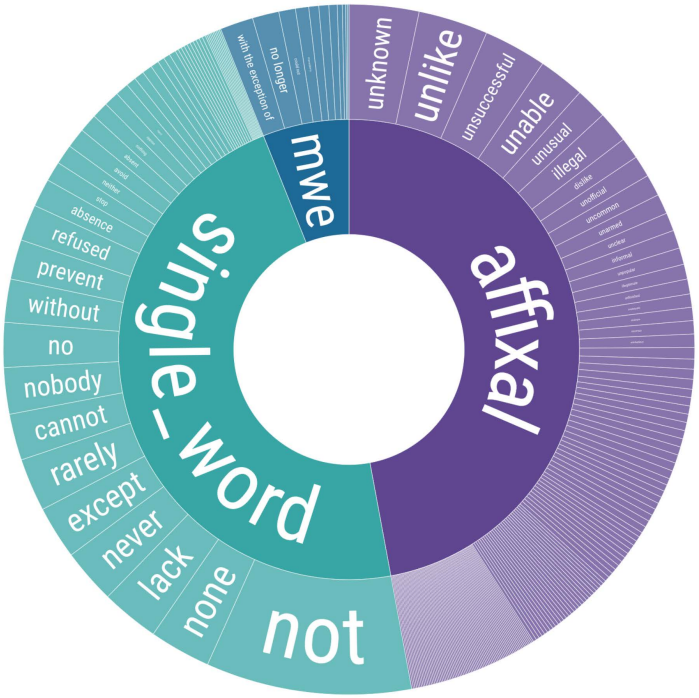

Figure 2. Distribution of negation cues in CONDAQA corpus
图 2. CONDAQA 语料否定线索分布情况

- There is no mystery to the story.
- Damon and Jannie have never been described as anything but angelic.

CONDAQA 例句(显性 + 隐性否定)：

- However, Machias Seal Island... was unaddressed by the commission.
- ...absence from work due to sickness and elevated risk of future disability pension.

综上所述，本节突出了 SFU 与 CONDAQA 两个语料库在否定表达上的分布性与结构性不均衡问题，这一差异可通过表 1 更直观地呈现。此类不平衡性进一步突显了构建针对性增强策略的必要性。区别于表面改写类增强方法，本文提出的增强框架旨在通过语义控制与多样化的否定线索设计，解决训练数据

中结构性稀缺的问题。下一节将具体阐述该增强框架。

Table 1. Corpus statistics and negation coverage
表 1. 语料库统计及否定分布

Dataset	Size	Negation Type
SFU	16,944	18% negation (mostly explicit)
CONDAQA	1289	100% negation (balanced)

3.2. 基于大型语言模型的特征导向数据增强框架

为解决训练语料中隐性否定表达不足且结构单一的问题，本研究提出一种基于大型语言模型(LLMs)的结构化数据增强框架。该框架通过一个两阶段流程，系统性地提升语料中隐性、低频否定表达的频率和多样性，以强化模型对深层语言特征的泛化能力。

阶段一：线索驱动的句子生成：

此阶段旨在扩充包含稀有否定线索的训练样本，在两个维度上重新平衡了代表性不足的语言特征：(1) 通过增加低频线索的出现频率，实现频率上的补偿；(2) 通过多样的句法配置与领域特定表达，实现结构上的补偿。

1) 种子线索提取：首先，从 CONDAQA 数据集中筛选出 100 个在 SFU Review 语料中未出现的隐性否定线索(如 unaddressed、hardly、absence)，以确保引入新的语言结构并避免数据污染。

2) 线索语义扩展：随后，利用 GPT-4 对每个种子线索进行语义扩展，生成 5 个语义相近的替代词(例如，将 unlike 扩展为 untenable、unfathomable 等)，最终获得 500 个扩展线索，极大地丰富了否定表达的词汇库。

3) 多样化句子生成：以 CONDAQA 语料中的句子作为结构模板，利用这 500 个扩展线索，引导 GPT-4 生成约 5000 条新的训练句子。生成过程覆盖了教育、科技、文化等多个语义领域，确保了样本的领域多样性。

4) 质量控制：所有生成的句子都经过严格的质量筛选，包括使用自动化工具(如 Grammarly API)进行语法检测和人工审核，以确保最终纳入训练集的样本在句法和语义上均是高质量的。

阶段二：结构化的反事实生成：

此阶段旨在通过生成极性相反的对比句对，增强模型对否定线索功能的敏感性，抑制其对数据集中表面统计偏差的依赖。我们为整个训练集(包括原始 SFU 样本和 LLM 生成的样本)都生成了反事实版本。为确保生成质量，我们采用了一种受控的重写策略，该策略借鉴了 Plyler 等人的理据引导思想，并施加了以下四个核心约束：

- 约束 1：理据识别：从原句中识别出承载核心极性信息的最小片段(约占 15%~20%，可不连续)，作为“理据”。若句中存在否定线索，则必须包含在理据中。
- 约束 2：极性反转重写：对识别出的理据进行重写，使其极性反转(如否定转为肯定)。重写过程必须移除所有原始否定标记，且不得引入新的否定词。
- 约束 3：受控重嵌入：将重写后的理据重新嵌入原句的句法结构中，并最大限度地减少对句子其余部分的改动，以保持语义上下文的连贯性。
- 约束 4：结构保真度：生成的反事实句在字符数上必须控制在原句长度的±10%范围内，并确保语言流畅自然。

以下示例展示了在该约束机制下完成的一次受控重写：

- 原语料：*This was later adopted in Ancient Greece as the “gamos” and “engeysis” rituals, although unlike in Judaism the contract made in front of witness was only verbal.*
- 大语言模型生成：*This was later adopted in Ancient Greece as the “gamos” and “engeysis” rituals, similar to Judaism, the contract made in front of witnesses was simply oral.*

通过这一系列明确的约束，LLM 能够生成高质量、语义连贯且结构清晰的反事实句对。这种方法为模型提供了丰富且平衡的对比监督信号，有效促进了其对否定语言特征的深度泛化能力。

3.3. 基线增强方法

为验证我们提出的结构化增强框架的有效性，我们设置了两种基线增强方法作为对比：一种是基于自监督的“理据重写”框架，另一种则是传统的简单加入否定词。

3.3.1. 自监督反事实生成

自监督方法如图 3 所示，该方法源于 Plyler 等[19]的研究成果。该研究提出了一种自动化的、基于理据(rationale-based)的反事实生成框架。具体而言，给定一个带有情感极性标注的句子，模型首先识别出与原始标签有因果关联的最小词元片段(理据)。随后，利用掩码语言模型(Masked Language Model, MLM)对这些片段进行修改，以生成具有对比性的版本，从而实现句子极性的翻转，同时保持语言流畅性和句法结构的完整。

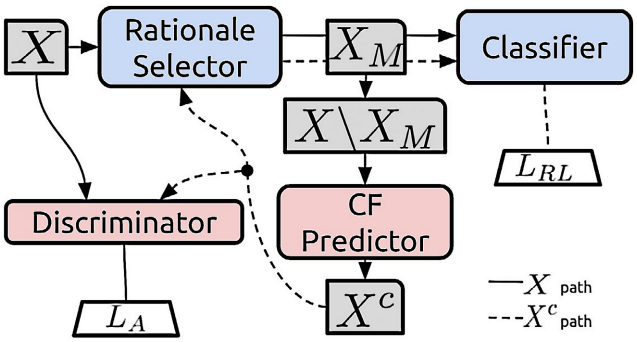

Figure 3. Counterfactual data generation workflow
图 3. 反事实数据生成工作流

尽管这一策略高效且无需额外标注，但可能存在句法多样性不足的问题，且在极性反转中过分依赖如 *not* 或 *never* 等显性否定标记，从而限制其对语义细腻现象的建模能力。

3.3.2. 简单否定词插入

此方法是一种更朴素的基线，通过在原始肯定句中随机插入显性否定词(如 not)来生成否定样本。该基线旨在验证仅仅增加否定信号的数量，而缺乏结构和语义多样性，是否足以提升模型的泛化性能。

4. 实验设置

为系统评估所提出的数据增强框架的有效性，我们基于 roberta-base 模型设计了六种实验配置。所有模型均使用 Hugging Face Transformers 库进行微调，并通过操控训练数据的构成来评估不同增强策略对模型性能的影响。

1) SFU：仅使用原始 SFU 语料库的 60% (3117 条否定句和 7049 条肯定句)进行微调。

2) SFU + LLM：在 SFU 数据基础上，加入 5000 条由本文提出的阶段一方法生成的隐性否定样本。

Table 2. Hyperparameter settings
表 2. 超参数设置

参数类别	超参数	值
模型与架构	基础模型	roberta-base
	最大序列长度	256 tokens
训练方案	训练轮数(Epochs)	20
	每设备批量大小(Batch Size)	16
	梯度累积步数	1
	早停策略(Early Stopping)	监控指标：eval_f1，容忍轮数：2
优化器与调度器	优化器	AdamW
	峰值学习率	2×10^{-5}
	AdamW Betas ($\beta 1, \beta 2$)	(0.9, 0.999)
	AdamW Epsilon (ϵ)	1×10^{-8}
	权重衰减(Weight Decay)	0.01
	学习率调度策略	带预热的线性衰减
	预热步数(Warmup Steps)	500
硬件与精度	GPU	1× RTX4090 (24GB)
	混合精度训练	FP16

3) SFU + LLM + CF (SS)：在配置 2 的基础上，额外加入通过自监督反事实生成方法(见 3.3.1 节)产生的反事实样本。

4) SFU + LLM + CF (GPT-4)：在配置 2 的基础上，额外加入通过本文提出的阶段二方法(见 3.2 节)生成的结构化反事实样本。

5) SFU + Not-inserted：在 SFU 语料中，加入 5000 条通过简单否定词插入方法(见 3.3.2 节)生成的合成样本。

6) NegBERT：直接在 CONDAQA 测试集上评估公开发布的预训练 NegBERT 模型，不进行任何额外微调，作为领域内先进方法的参考。

训练数据由 SFU 语料与不同策略生成的增强样本构成。在引入反事实样本的配置中，我们确保正负类别完全平衡。测试集统一使用 CONDAQA 数据集，该数据集包含 1289 对语义相反的肯定与否定句。所有报告的评估结果均为三次独立实验的平均值，评估指标包括准确率(Accuracy)、精确率(Precision)、召回率(Recall)与 F1 分数。表 2 详细列出了本次微调实验所使用的全部超参数。

5. 实验结果与分析

本研究以多种数据增强配置对 RoBERTa 模型进行了微调，并在 CONDAQA 衍生的测试集上评估模型表现。该测试集包含 1289 对肯定与否定句，评估指标包括准确率(Accuracy)、F1 分数与 AUC 值，具体结果如表 3 所示。

未经微调的 RoBERTa 基线模型性能接近随机水平，说明其难以有效捕捉否定相关的语言特征模式。以 SFU 语料微调后，模型表现有明显提升(准确率为 60.76%，F1 分数为 0.5792)，但由于过度依赖显性否定线索，限制了其对更复杂隐性否定形式的泛化能力。当加入 5000 条由 LLM 生成的隐性否定样本(SFU

+ LLM 配置)时，模型的表现显著提高(准确率达 67.31%，F1 分数为 0.6712)，验证了结构多样、特征导向的数据增强对模型性能的积极影响。

Table 3. Model performance comparison on CONDAQA dataset
表 3. 模型在 CONDAQA 测试集上的性能表现

	Accuracy	F1	AUC
RoBERTa (w/o FT)	50.06%	0.3340	0.3484
SFU	60.76%	0.5792	0.6379
SFU + LLM	67.31%	0.6712	0.7467
SFU + LLM + CF (SS)	68.15%	0.6805	0.7734
SFU + LLM + CF (GPT4)	75.20%	0.7513	0.7855
NegBERT	60.17%	0.5932	0.6745
SFU + Not	60.56%	0.5774	0.6483

进一步引入反事实增强策略(Counterfactual Augmentation)后，模型性能进一步提升。在自监督变体(SFU+ LLM + CF (SS))中，F1 分数小幅但稳定上升至 0.6805，说明即使是自动生成的对比样本，也有助于引导模型关注与极性相关的语言片段。但此方法主要依赖浅层变换和表面语言线索，因此效果仍存在一定局限。

表现最优的是使用 GPT-4 生成反事实样本的配置(SFU + LLM + CF (GPT4))，其准确率为 75.20%，F1 分数为 0.7513，AUC 达 0.7855。这一结果凸显了将特征导向的数据增强与语义连贯的极性转换相结合，在有效建模低频语言特征方面的优势。与自监督方法相比，GPT-4 生成的反事实句子在结构多样性和语义流畅性上更具优势，显著增强了模型的泛化能力。

值得一提的是，更简单的数据增强策略(如随机插入“not”或直接使用未微调的 NegBERT 模型)未能表现出显著优势，这些方法缺乏必要的语义控制与深层次的结构变化。这进一步强调了否定识别任务的特征导向性质，模型表现强烈依赖于训练过程中所提供的否定线索的质量与多样性。

另一个有趣的现象是，仅使用 LLM 生成的反事实样本在 CONDAQA 测试集上取得了近 90%的高准确率。然而，在 SFU 数据集上的初步测试(未列入表中)则显示性能大幅下降，与未经微调的 RoBERTa 基线模型表现相当。这一现象表明模型可能过拟合于合成数据中浅层否定线索，而未真正学习到否定的泛化特性。这也再次凸显了构建跨领域、线索多样的训练数据的重要性。

综上，实验结果验证了本研究的核心假设：通过 LLM 引导、结构多样的数据增强，有助于显著提升模型对稀有与隐性否定模式的敏感性。同时，不同反事实策略之间的对比表明，提升模型泛化能力不仅依赖于对比监督的存在，更取决于其语义与句法层面的质量控制。

上述发现契合本文的研究目标：通过结构化、由 LLM 驱动的数据增强机制，增强模型对代表性不足、隐性表达语言特征的建模能力。

6. 结论与未来工作

本文提出了一种基于大型语言模型(LLMs)的特征导向数据增强框架，以解决小规模、不平衡语料中稀有语言特征表示不足的问题。以否定检测为诊断性案例，本研究证明 LLMs 可通过隐性线索扩展与结构化的反事实重写生成高质量的训练样本，显著提升模型对隐性否定等上下文敏感语言特征的识别与泛化能力。

实验结果显示，通过 LLM 驱动的增强方法，RoBERTa 模型在隐性否定识别任务上的性能明显提升，说明结构丰富且语义细腻的训练数据能有效克服传统增强方法的不足。此外，本研究还指出，现有数据集中显性否定表达数量远超过隐性否定表达，盲目合并数据可能导致隐性线索的进一步稀释，未来应有意识地避免此类数据建设问题。

未来研究可进一步拓展本文提出的方法，应用于其他同样具有稀疏特征问题的语言现象，如情态表达、指代消解或修辞语言等。这些特征往往无法仅通过表层特征建模，需要结合结构和语义导向的增强策略。另一方面，将增强策略拓展至跨领域、多文体的语料中，以提升模型在实际应用场景中的泛化能力，也将是未来重要的研究方向。此外，深入探索经特征平衡数据训练的模型在情感分析、问答系统与事实验证等更广泛的下游任务中的表现与泛化特性，具有显著的研究与实践价值。

参考文献

[1] Henning, S., Beluch, W., Fraser, A., *et al.* (2023) A Survey of Methods for Addressing Class Imbalance in Deep-Learning Based Natural Language Processing. arXiv: 2210.04675. http://arxiv.org/abs/2210.04675

[2] Hofmann, V., Pierrehumbert, J.B. and Schütze, H. (2021) Superbizarre Is Not Superb: Derivational Morphology Improves BERT's Interpretation of Complex Words. arXiv: 2101.00403. http://arxiv.org/abs/2101.00403

[3] Gururangan, S., Swayamdipta, S., Levy, O., *et al.* (2018) Annotation Artifacts in Natural Language Inference Data. arXiv: 1803.02324. http://arxiv.org/abs/1803.02324

[4] Wei, J. and Zou, K. (2019) EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks. arXiv: 1901.11196. http://arxiv.org/abs/1901.11196

[5] Ebrahimi, J., Rao, A., Lowd, D., *et al.* (2018) HotFlip: White-Box Adversarial Examples for Text Classification. arXiv: 1712.06751. http://arxiv.org/abs/1712.06751

[6] Kaushik, D., Hovy, E. and Lipton, Z.C. (2020) Learning the Difference that Makes a Difference with Counterfactually-Augmented Data. arXiv: 1909.12434. http://arxiv.org/abs/1909.12434

[7] Min, S., Lyu, X., Holtzman, A., *et al.* (2022) Rethinking the Role of Demonstrations: What Makes In-Context Learning Work? arXiv: 2202.12837. http://arxiv.org/abs/2202.12837

[8] Zhou, C., Liu, P., Xu, P., *et al.* (2023) LIMA: Less Is More for Alignment. *Proceedings of the* 37*th International Conference on Neural Information Processing Systems*, New Orleans, 10-16 December 2023, 55006-55021.

[9] Gururangan, S., Marasović, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D., *et al.* (2020) Don't Stop Pretraining: Adapt Language Models to Domains and Tasks. *Proceedings of the* 58*th Annual Meeting of the Association for Computational Linguistics*, 5-10 July 2020, 8342-8360. https://doi.org/10.18653/v1/2020.acl-main.740

[10] Dai, H., Liu, Z., Liao, W., Huang, X., Cao, Y., Wu, Z., *et al.* (2025) AugGPT: Leveraging ChatGPT for Text Data Augmentation. *IEEE Transactions on Big Data*, **11**, 907-918. https://doi.org/10.1109/tbdata.2025.3536934

[11] Qu, Y., Shen, D., Shen, Y., *et al.* (2020) CoDA: Contrast-Enhanced and Diversity-Promoting Data Augmentation for Natural Language Understanding. arXiv: 2010.08670. http://arxiv.org/abs/2010.08670

[12] Hossain, M.M., Chinnappa, D. and Blanco, E. (2022) An Analysis of Negation in Natural Language Understanding Corpora. arXiv: 2203.08929. http://arxiv.org/abs/2203.08929

[13] Poliak, A., Naradowsky, J., Haldar, A., Rudinger, R. and Van Durme, B. (2018) Hypothesis Only Baselines in Natural Language Inference. *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, New Orleans, 5-6 June 2018, 180-191. https://doi.org/10.18653/v1/s18-2023

[14] Ravichander, A., Gardner, M. and Marasović, A. (2022) CONDAQA: A Contrastive Reading Comprehension Dataset for Reasoning about Negation. arXiv: 2211.00295. http://arxiv.org/abs/2211.00295

[15] Shaitarova, A. and Rinaldi, F. (2021) Negation Typology and General Representation Models for Cross-Lingual Zero-Shot Negation Scope Resolution in Russian, French, and Spanish. *Proceedings of the* 2021 *Conference of the North American Chapter of the Association for Computational Linguistics*: *Student Research Workshop*, 7 June 2021, 15-23. https://doi.org/10.18653/v1/2021.naacl-srw.3

[16] Fancellu, F. (2018) Computational Models for Multilingual Negation Scope Detection. Master's Thesis, University of Edinburgh. http://hdl.handle.net/1842/33038

[17] Truong, T.H., Baldwin, T., Cohn, T., *et al.* (2022) Improving Negation Detection with Negation-Focused Pretraining. arXiv: 2205.04012. http://arxiv.org/abs/2205.04012

[18] Cruz, N.P., Taboada, M. and Mitkov, R. (2015) A Machine-learning Approach to Negation and Speculation Detection for Sentiment Analysis. *Journal of the Association for Information Science and Technology*, **67**, 2118-2136. https://doi.org/10.1002/asi.23533

[19] Plyler, M., Green, M. and Chi, M. (2021) Making a (Counterfactual) Difference One Rationale at a Time. *Advances in Neural Information Processing Systems* 34 (*NeurIPS* 2021), Online, 6-14 December 2021, 28701-28713. https://proceedings.neurips.cc/paper/2021/hash/f0f800c92d191d736c4411f3b3f8ef4a-Abstract.html

附　录：提示词设计

提示词 1：生成含隐性否定的句子

以下提示词用于引导模型从 CONDAQA 语料中生成包含隐性否定线索的陈述句：

You are now playing the role of a professional writing assistant. Your task is to help construct well-written declarative sentences that showcase linguistic features related to negation.

- Below is a source sentence that uses the implicit negation cue “unlike”:

This was later adopted in Ancient Greece as the “gamos” and “engeysis” rituals, although unlike in Judaism the contract made in front of witnesses was only verbal.

- Please generate 10 new declarative sentences that retain a similar syntactic and rhetorical structure but shift to different topics or domains (e.g., law, science, education, or culture).
- In each sentence, replace “unlike” with a semantically similar implicit negation cue (e.g., unsubstantiated, untenable, unfathomable, unprecedented).
- Use formal language, and ensure each sentence is coherent and contextually meaningful.

提示词 2：反事实重写(否定→肯定)

以下提示词用于引导大型语言模型对否定句进行反事实重写，将其转化为肯定句：

Hi, ChatGPT. Please help me to transform a sentence containing the negation cue “unlike” into an fully affirmative version under the following constraints:

- Character count of the result must be within ±10% of the original.
- 15%~20% of the original words must be changed (via replace, delete, or insert).
- Remove all negation markers. Do not introduce new negation.
- Output must be fluent and natural.

For any input, reply with:

1. Original metrics (character and word count)
2. Changed tokens (words replaced/added/removed)
3. Final rewritten sentence

Example input:

This was later adopted in Ancient Greece as the “gamos” and “engeysis” rituals, unlike in Judaism the contract made in front of witness was only verbal.

Example output:

This was later adopted in Ancient Greece as the “gamos” and “engeysis” rituals, similar to Judaism, the contract made in front of witnesses was simply oral.

Your help are deeply appreciated.

提示词 3：反事实重写(肯定→否定)

以下提示词用于引导大型语言模型对肯定句进行反事实重写，将其转化为否定句：

Hi ChatGPT, please create a counterfactual version of the following affirmative sentence under these rules:

- Character count of the result must be within ±10% of the original.
- 15%~20% of the original words must be changed (via replace, delete, or insert).

- The result should remain negated, but do not introduce any explicit “no”, “not” or similar explicit structural negation tokens.
- Output must be fluent and natural.

For any input, reply with:

1. Original metrics (character and word count)
2. Changed tokens
3. Final sentence

Example input:

She always arrives early to every meeting.

Example output:

She hardly comes early to every meeting.