Computer Science and Application 计算机科学与应用, 2025, 15(1), 209-219
Published Online January 2025 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2025.151021

Hans汉斯

基于大模型的引文情感分类问题的研究

孔明辉[1]，赵　兰[2*]

[1]西南交通大学，计算机与人工智能学院，四川 成都
[2]吉利学院，盛宝金融科技学院，四川 成都

收稿日期：2024年12月23日；录用日期：2025年1月21日；发布日期：2025年1月28日

摘　要

针对科学文献影响力排名研究领域中，需要对引文情感极性进行预测的问题，提出了将大语言模型的提示工程(零样本学习以及少样本学习)方法应用在引文情感分类中这一方案，分析当下热门大语言模型如Llama，Gpt-4o-Mini等以及基于Bert的深度学习模型在科学引文情感分类问题上的效果。首先通过基于大语言模型的提示工程方法预测引文情感极性，分析预测效果，再与基于Bert的深度学习模型在这一问题中的表现进行对比分析。实验结果表明，基于Bert的深度学习模型情感分类准确率在90%以上，最高可达94.31%，F1值均在80%以上；基于大语言模型的零样本学习和少样本学习方法分类效果与前者有明显差距，准确率最高可达84.70%，F1值最高仅可达63.65%。和基于Bert的深度学习模型分类效果相比，基于大语言模型的提示工程方法虽然在该任务中准确率受限，但其泛化能力较强，是一种简便且高效的方法，对于任务快速部署和应用非常有用。

关键词

引文情感分类，Llm，深度学习

A Research on Citation Sentiment Classification Based on Large Language Model

Minghui Kong[1], Lan Zhao[2*]

[1]School of Computer and Artificial Intelligence, Southwest Jiaotong University, Chengdu Sichuan
[2]School of SAXO Fintech, Geely University of China, Chengdu Sichuan

Received: Dec. 23rd, 2024; accepted: Jan. 21st, 2025; published: Jan. 28th, 2025

*通讯作者。

文章引用: 孔明辉, 赵兰. 基于大模型的引文情感分类问题的研究[J]. 计算机科学与应用, 2025, 15(1): 209-219.
DOI: 10.12677/csa.2025.151021

Abstract

This paper proposes a method that applies prompt engineering (zero-shot and few-shot learning) from large language models (LLMs) to predict citation sentiment polarity in scientific literature impact ranking research. The study analyzes the performance of popular LLMs, such as Llama and GPT-4o-Mini, and BERT-based deep learning models in the task of scientific citation sentiment classification. The method uses prompt engineering with large language models to predict sentiment polarity. The results are compared with those of BERT-based deep learning models. Experimental results show that BERT-based models achieve sentiment classification accuracy over 90%, with a maximum of 94.31%, and F1 scores above 80%. The zero-shot and few-shot learning methods based on large language models have a significant performance gap. Their maximum accuracy is 84.70%, and the highest F1 score is only 63.65%. Compared to BERT-based models, the prompt engineering method based on large language models has lower accuracy but shows strong generalization ability. It is a simple and efficient method for quick deployment and application.

Keywords

Citation Sentiment Classification, Llm, Deep Learning

1. 引言

研究文献影响力可以帮助研究者了解哪些领域或主题受到关注，哪些文献对该领域的发展具有引领性的作用。这有助于指导研究方向，使研究更加前瞻性和有针对性。影响力高的文献通常会吸引更多的关注，促使学者之间进行更多的学术交流和合作。密切的学术交流有助于形成研究网络，推动学术社区的发展。而文献影响力的研究离不开文献之间的相互引用。

学术界对引文分析的需求和作用正在逐渐增加，然而，传统的引文分析方法主要侧重于定量指标，如引文数量和 H 指数，忽略了引文功能和引文情感的深层信息。Radicchi Filippo [1]和 Baird L M [2]的工作进一步证明了被引用数量的局限性，例如有缺陷或有争议的论文往往会被更高的引用，而被引用数量无法反映这一信息。作为一种著名的排名评估算法，Pagerank [3]-[5]方法已经被广泛有效地用于解决各种排名任务，如网络流量预测和社区发现。而 Jian-Feng Jiang [6]开发了一种链接加权算法，该算法根据不同文章节点之间的实际意义和表示为相应的链接分配权重。

学术文章的作者通过引用概念、方法、结论和实验过程来支持他们的工作，或者通过指出以前工作中的不足来介绍他们自己的工作，从而在不同论文之间建立各种关系。对引文行为的分析和挖掘有助于揭示研究领域内的知识结构、研究热点、研究趋势和学术交流网络。通过对科学引文中作者情感倾向的分析，可以深入了解学术论文的情感色彩，促进学术合作和交流；同时，分析引文有助于改进信息检索和推荐系统，提供更精准的信息推荐。此外，了解科技引文中的情感倾向也对科技发展趋势、科学传播以及决策制定和政策评估具有指导意义。

2009 年以后，对文本情感分类的研究逐渐出现并显著增加。产品评论、社交媒体对话、新闻和博客是最受关注的领域[7]。根据 Yousif 等人[8]的研究，科学引文情感分类最早出现在 2011 年前后。情感词

典、机器学习和深度学习是三种最常见的方法。Small 等人[9]使用一到三个句子作为引文上下文来帮助分析引文情绪，以了解引文的结构和潜在的认知过程。他使用了一个由大量提示词或短语组成的数据集，详细分析了 20 篇论文的功能和情感。Athar [10]使用具有不同引文的 Svm 分类器将引文分为三类：阳性、阴性和中性，并构建了一个包含 8736 个实例的语料库。Poria 等人[11]提出使用 Cnn 从多模态内容中提取特征，并将这些特征提供给多核学习分类器进行情绪检测，这在不同的数据集上也取得了良好的效果。Azhar Ahmed Bilal 等人[12]结合了深度学习和机器学习的方法，提出了一种将深度序列特征与随机森林(Rf)技术相结合的混合二元分类框架，采用深度长短期记忆(Lstm)模型提取悲伤和快乐情绪对应的深度序列特征并利用随机森林(Rf)算法实现了五重交叉验证技术来区分儿童故事悲伤和快乐情绪。Tirthankar Ghosal 等人[13]证明了与被引用文章的附加上下文(标题信息)相关的研究论文的结构信息可以用来有效地分类引文的意图，并提出了一种新的具有三个辅助的深度 Mtl 框架。辅助任务与科学论文的结构性质有关。它们帮助模型将科学文献中可用的结构信息整合到引文意图中。

预训练模型的方法也常用于情感分类问题中，Beltagy 等人[14]使用了一个大型科学语料库，包括 114 万篇生物医学(82%)和计算机科学(12%)的科学论文，而不是一个通用语料库来预训练 Bert。在某种程度上，Scibert 更适合于科学论文的 Nlp 任务，显著提高了科学引文分类的效果。这项研究的重点是将先验知识整合到预先训练的模型中。Dahai Yu 等人[15]提出将情感词典等先验信息与 Bert 模型结合，获得了不错的准确率。周文远等人[16]采用 Scibert 预训练模型得到语料集中句子的语义表示向量，根据文本特点，依次通过 Bigru 神经网络和多尺度卷积神经网络(Multi-Cnn)提取句子中的时序全局特征和局部关键特征，引入注意力机制对提取出的特征重新分配权重，达到突出关键特征的目的，最后通过线性层实现引文情感和引文目的自动分类。Komal Rani Narejo 等人[17]介绍了一种基于 Bert 架构的 Emoji-Enhanced Bert (Eebert)技术。使用嵌入层为表情符号标记化创建的情绪调整因子(Saf)用于分析评论文本内容中的情绪和情感。该模型的准确率达到97.00%，进一步通过5倍交叉验证进行连续测试，平均准确率达到99.21%，支持了 Eebert 模型的可靠性。

Kiana Kheiri 等人[18]将大模型应用到推特评论情感分类问题中，具体方法包括微调大模型，提出了 Prompted-Gpt 以及运用编码模型对语料数据进行分类，取得了良好的效果。该研究在这些策略和单个 Gpt 模型之间产生了详细的比较见解，揭示了它们独特的优势和潜在的局限性。此外，该研究还将这些基于 Gpt 的方法与之前使用的其他基于相同数据集的高性能模型进行了比较。结果表明，Gpt 方法在预测性能方面具有显著的优势，与最先进的方法相比，f1 得分提高了 22%以上。本文还揭示了情感分析任务中的常见挑战，例如理解上下文和检测讽刺。Konstantinos I. Roumeliotis 等人[19]探讨了大型语言模型(Llm)在分析与加密货币相关的新闻文章中的情绪方面的能力，对 Gpt-4、Bert 和 Finbert 等最先进的模型进行微调，以完成这项特定的任务。

这些研究为实现引文情感的自动分类提供了丰富的理论指导，具有重要的参考价值，但对 Llama，Qwen 等当下热门大模型，将其应用在引文情感分类任务中的研究有些匮乏。本文主要的研究工作是探索将上述大语言模型的应用在这一任务中的可能性以用来更快捷更准确的分析引文的情感极性，以便更准确地分析论文的影响力。

2. 研究思路与框架

深度学习和预训练模型的技术已经广泛运用于引文情感分类问题中，但对 Llama，Qwen，Gpt 等当下热门大语言模型在该任务中的应用，并没有充分的研究。本文基于零样本学习以及少样本学习的方法对上述模型进行实验效果分析，同时与基于 Bert 模型预训练的方法进行比对，探讨将上述大语言模型更好的应用在这一任务中的可能性。主要研究框架图如图 1 所示：

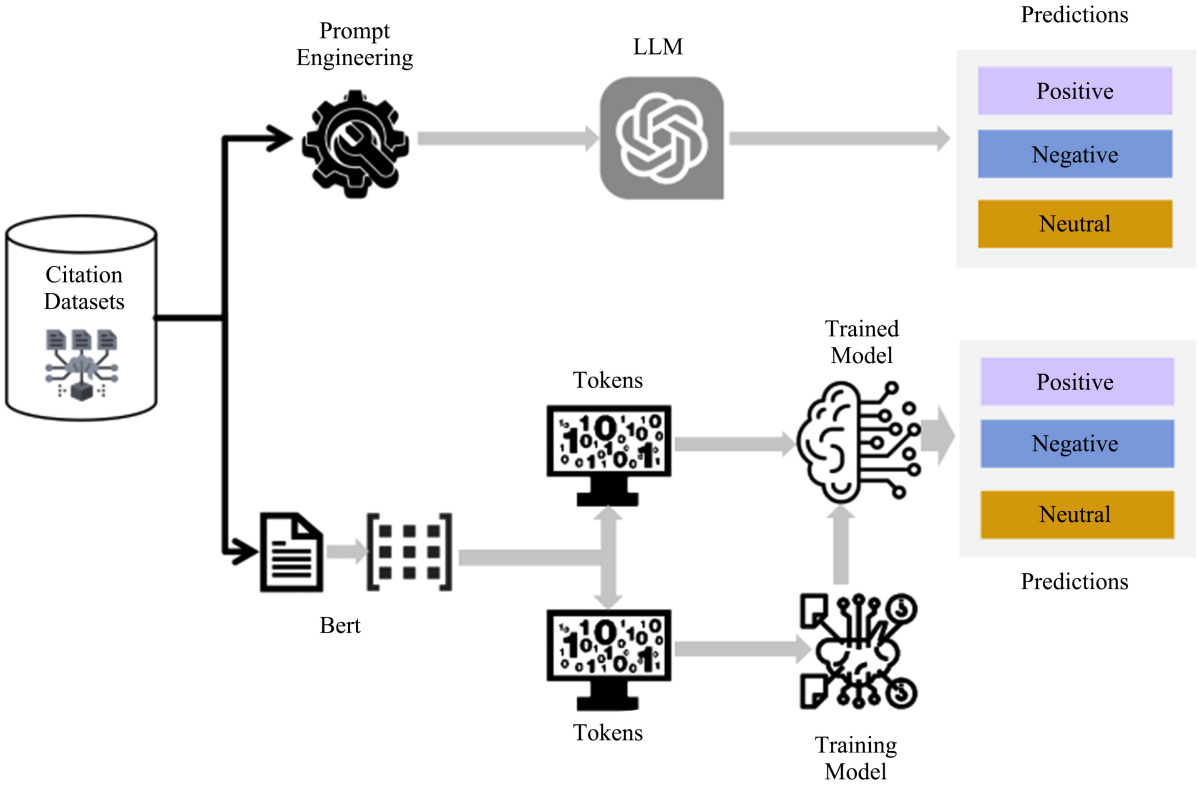

Figure 1. Research structure diagram
图 1. 研究框架图

2.1. 基于 Bert 预训练语言模型的深度学习

运用基于 Bert 的深度学习模型实现引文情感分类，对实验结果进行对比分析。在本任务中，采用了结合 Bert (Bidirectional Encoder Representations From Transformers)模型与其他类型的神经网络(如 Fnn、Lstm、Gru、Rnn)的做法。这些组合利用了 Bert 这一强大的预训练语言模型来提取文本的上下文特征，然后通过不同的神经网络结构对这些特征进行进一步的建模和分类。Bert 网络结构如图 2 所示，E1, E2, ···, En 表示初始字向量，通过基于 Transformer 的双向编码器，得到含有特征的引文文本向量化表示 T1, T2, ···, Tn。

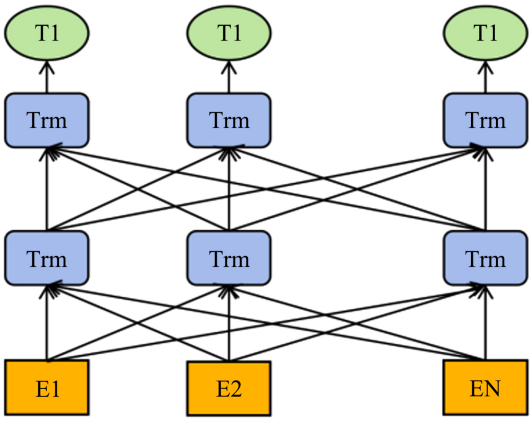

Figure 2. Bert structure diagram
图 2. Bert 网络结构图

Bert 的最终输入表示是将词嵌入(Token Embeddings)，位置嵌入(Position Embeddings)和段嵌入(Segment Embeddings)三者加在一起：词嵌入(Token Embeddings)：每个 Token 对应一个向量表示，通常是通过一个查找表得到的。位置嵌入(Position Embeddings)：每个 Token 根据其在句子中的位置(即 Index)获取对应的向量表示。段嵌入(Segment Embeddings)：区分不同句子的嵌入。对于句子对任务，第一个句子的所有 Token 对应的段嵌入为 0，第二个句子的所有 Token 对应的段嵌入为 1。这三部分的向量相加后，得到每个 Token 的最终输入表示。假设我们有一个 Token 序列$\{t1, t2, \cdots, Tn\}$，其中 n 是序列长度，最终的输入表示是：

$$\text{Input Embedding} = \text{Token Embedding} + \text{Position Embedding} + \text{Segment Embedding} \tag{1}$$

每个 Token 都会被映射为一个 768 维的向量，表示该 Token 在当前上下文中的表示。这个向量包含了该 Token 在整个输入句子中的上下文信息。

1) Bert + Fnn (前馈神经网络)：

Bert 模型通过其 Transformer 结构生成文本的上下文表示后，使用前馈神经网络(Fnn)进行分类。Fnn 通常是一个简单的全连接层或多个全连接层堆叠，具有激活函数(如 Relu)，最后通过 Softmax 进行分类。Fnn 层将 Bert 输出的嵌入传递到前馈神经网络中，进行进一步的非线性变换，然后输出类别的概率分布。Fnn 结构简单，适合在 Bert 的基础上进行分类任务，训练速度相对较快。

2) Bert + Rnn (循环神经网络)：

Rnn 是一种经典的递归神经网络结构，能够处理序列数据，但相比 Lstm 和 Gru，它容易遭遇梯度消失或梯度爆炸的问题。结合 Bert 与 Rnn 时，Rnn 对 Bert 输出的序列进行处理。Rnn 对文本序列建模，捕捉文本的时序特征。最后传到全连接层将 Rnn 的输出用于分类。Rnn 模型结构简单，适合处理不太长的文本。

3) Bert + Lstm (长短期记忆网络)：

Lstm 是 Rnn 的一种改进，能够捕捉长期依赖关系，适合处理序列数据中的上下文关系。其对 Bert 输出的序列进行进一步处理，能够考虑到序列中的长期依赖和上下文信息。最后 Lstm 的输出被传递到全连接层进行最终的分类。Lstm 适合较长文本的情感分析。

4) Bert + Gru (门控循环单元)：

Gru 是 Rnn 的另一种变种，它通过更简洁的门控机制来控制信息流动，通常在较短文本或训练数据较小的场景下比 Lstm 表现得更好。Gru 对 Bert 输出的序列进行建模，处理短期和长期的上下文信息。最后其输出被传递到全连接层进行最终分类。Gru 比 Lstm 更轻量级，能够在较小的训练集上取得较好的性能，且计算速度较快。

2.2. 基于大语言模型的零样本学习及少样本学习

运用基于以下热门大语言模型 Zero-Shot 和 Few-Shot 的方法进行情感分类，对实验结果进行对比分析。

1) Qwen-2:7B 是由 Qwen 团队(清华大学、北京智源研究院等研究机构)开发的一个大规模语言模型。它是 Qwen 系列的一部分，专注于自然语言处理任务，如文本生成、对话、翻译等。Qwen-2:7B 在规模上具有 7 亿个参数，这使它能够在多个语言理解和生成任务中表现出色。

2) Qwen-2.5:7B 是清华大学、北京智源研究院等机构在 Qwen 系列模型基础上推出的改进版大规模语言模型。它是 Qwen-2:7B 的升级版本，继续强化了多任务处理能力，并在多个自然语言处理任务中表现出色。Qwen-2.5:7B 仍然维持在 7 亿参数的规模，但相比前一版本(Qwen-2:7B)，在算法、训练策略和数据处理方面进行了优化，增强了模型的推理能力和多任务适应性。

3) Llama 3.1:8B 是 Meta (前 Facebook)开发的 Llama (Large Language Model Meta Ai)系列中的一个大规模预训练语言模型。Llama 3.1 版本是 Llama 3 系列的一个改进版，拥有 8 亿个参数，相较于更大的模

型(如 Llama 13B 或 Llama 65B)，它的规模较小，计算需求和内存消耗较低，但依然具备出色的语言理解与生成能力，适合中等规模的应用场景。

4) Gpt-4O-Mini 是 Openai 推出的一款较小规模的 Gpt-4 模型变体。这个版本在规模和性能上相较于完整的 Gpt-4 模型进行了优化和精简，目的是在资源受限的环境中提供高效的推理能力，同时保留 Gpt-4 核心模型的强大语言理解和生成能力。Gpt-4O-Mini 是 Gpt-4 系列的一个精简版本，旨在提供高效的自然语言处理能力，并能在资源有限的设备上运行。它继承了 Gpt-4 的核心优势，能够高效处理多种 Nlp 任务，适合用于对性能和计算资源有较高要求的场景，如移动端应用、实时对话系统和低资源环境。Gpt-4O-Mini 的具体参数数量并未公开详细披露，因为它是一个较小的变体。通常情况下，类似的“Mini”版本会在原始模型的基础上减少参数量，以实现更高效的推理和更低的计算资源消耗。

如果参考 Gpt-4 作为对比，Gpt-4 本身有多个版本，参数数量大致在 1700 亿至 10,000 亿(1.7T 至 10T)之间，具体取决于不同的变体。而 Gpt-4O-Mini 应该是一个参数量显著减少的版本，估计它的参数数量可能在几十亿(例如 6B、8B、12B 参数左右)，具体数字需要依赖于官方的详细说明。

零样本学习通过让模型基于自然语言描述和任务定义进行推理，而不是依赖任务特定的训练数据。在大语言模型中，零样本学习的实现通常是通过提示工程(Prompt Engineering)来引导模型对引文进行情感分类，利用模型强大的生成和理解能力，让模型自动完成任务。在引文情感分类任务中，给大语言模型提供以下的信息：

任务描述：You are an assistant for sentiment classification of scientific citations. You will analyze the sentiment of the citation and return a numerical label.

分类标签：

- 0: Negative sentiment, indicating criticism or skepticism about the cited research, pointing out its limitations.

- 1: Neutral sentiment, indicating that the citation is used for background or methodological reference without expressing any strong sentiment.

- 2: Positive sentiment, indicating endorsement or approval of the cited research.

与零样本学习不同的是，少样本学习在给模型的提示里加入了多个输入输出示例，通过示例展示任务的模式和多样性，模型依此进行推理。少样本学习任务泛化能力强，对多样任务模式适用对示例的数量和质量要求较高。通过对实验效果对比分析，最后选择的示例如表 1 所示：

Table 1. Examples of few-shot learning prompt
表 1. 少样本学习任务提示示例

Scientific Citation	Label
The resulting net increase in ATF4 and CHOP is significantly less than that observed with a bona fide ER stress inducer, such as TG.	0
While work on subjectivity analysis in other languages is growing, Chinese data are used in, and German data are used in), much of the work in subjectivity analysis has been applied to English data.	1
For ovarian and other similar cancer cell lines have shown an increase in infectivity through CAR-independent transduction, achieving higher reporter gene expression by several orders of magnitude in the primary tumor cells.	2

然后，基于模型的预训练知识和自然语言理解能力，模型能够自动推断和判断引文的情感。提示工程流程图如图 3 所示：

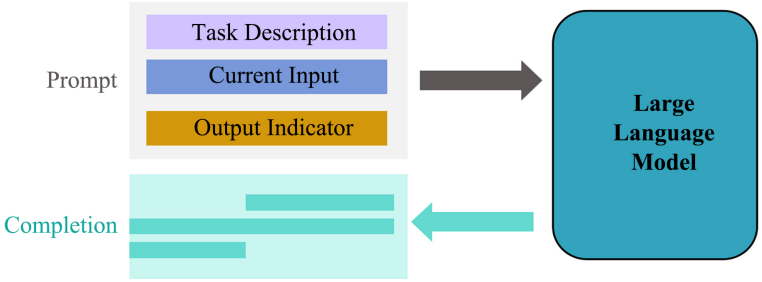

Figure 3. Prompt engineering flowchart
图 3. 提示工程流程图

3. 实验结果及分析

3.1. 数据来源及预处理

科技论文与其他文本如推特评论，影视剧评论的差异是多方面的，包括情感表达的目的、方式、强度、对象和受众。由于缺乏其他更好的解决方案，本研究使用 Dahai Yu [15]提出的数据集处理方案，该数据集来源于 Athar [11]提出的数据集。语料库包含 8736 条数据，每个引用都根据情绪手动标注为积极、消极或中立。这些引文摘自 Acl 文集网络语料库。此外，考虑使用 Arman 等人[20]提出的 Sciite 数据集进行数据补充，并从 Sciite 中提取了大约 1000 个句子来补充 Athar 提出的语料库。对不完整文本，重复标签文本，错误标签文本等数据进行了删除，修复等处理。最后，编译的数据集由 7912 个句子组成，其中 1237 个是肯定的，347 个是否定的，6328 个是中立的。其数据集分布情况如表 2 所示：

Table 2. Dataset distribution
表 2. 数据集分布情况

Sentence label	Number
Positive	1237
Negative	347
Neutral	6328

3.2. 实验参数设置

实验均采用 Anaconda 集成开发环境，编程语言为 Python3.8，深度学习框架为 PyTorch，显卡内存为 16 GB，处理器为 12th Gen Intel(R) Core(TM) i7-12700H 2.30 GHz。

实验 1 (基于 Bert 的深度学习)参数如表 3 所示：

Table 3. Experiment 1 main parameters
表 3. 实验 1 主要参数

实验参数	参数值
Train_Batch_Size	32
Num_Epoch	50
Learning Rate	5e−6
Weight_Decay	0.01
Test_Batch_Size	64

实验 2，3 (基于大语言模型的零样本学习，基于大语言模型的少样本学习)参数如表 4 所示：

Table 4. Experiment 2 main parameters
表 4. 实验 2 主要参数

实验参数	参数值
Max-token	5
Temperature	0

3.3. 评价指标

Accuracy (准确率)：对整个样本空间中的样本分类正确的一个比例。

F1-Score：是统计学中用来衡量二分类模型精确度的一种指标，用于测量不均衡数据的精度。它同时兼顾了分类模型的精确率和召回率。F1-Score 可以看作是模型精确率和召回率的一种加权平均，它的最大值是 1，最小值是 0。在多分类问题中，如果要计算模型的 F1-Score，则有两种计算方式，分别为 Micro-F1 和 Macro-F1。Micro-F1 和 Macro-F1 分数之间的关键区别在于它们在不平衡数据集上的行为。当类不平衡时，Micro F1 分数通常不会返回模型性能的客观衡量标准，而 Macro F1 分数可以。由于本研究所用数据集的不平衡性，所以选择 Accuracy 和 Macro-F1 分数为评价指标。分类结果混淆矩阵如表 5 所示：

Table 5. Confusion matrix
表 5. 混淆矩阵

真实类别	预测类别	
	正例	反例
正例	TP (真正例)	FN (假反例)
反例	FP (假正例)	TN (真反例)

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{2}$$

$$\text{Precison}_i = \frac{\text{TP}_i}{\text{TP}_i + \text{FP}_i} \tag{3}$$

$$\text{Recall}_i = \frac{\text{TP}_i}{\text{TP}_i + \text{FN}_i} \tag{4}$$

$$\text{F1}_i = 2\frac{\text{Precison}_i * \text{Recall}_i}{\text{Precison}_i + \text{Recall}_i} \tag{5}$$

$$\text{F1}_{macro} = \frac{\sum_{i=1}^{n} \text{F1}_i}{n} \tag{6}$$

两项评估指标的具体计算如公式(2)与公式(6)。

3.4. 实验结果

实验 1 (基于 Bert 的深度学习方法)结果如表 6 所示：

Table 6. The results of the performance indicators of each model in experiment 1
表 6. 实验 1 各模型的性能指标结果

Model	Acc	F1
Bert + Fnn	94.31%	84.45%
Bert + Rnn	93.81%	80.26%
Bert + Gru	93.62%	81.76%
Bert + Lstm	93.11%	80.09%

Bert + Fnn 在准确率(94.31%)和 F1 值(84.45%)上表现最好。它的 F1 值明显高于其他模型，意味着它在精确度和召回率的综合表现上最为均衡。Bert + Lstm 在准确率(93.11%)和 F1 值(80.09%)上表现最差，虽然准确率相差不大，但它的 F1 值略低，说明其在任务中对精确度与召回率的平衡有所欠缺可能在该任务上对长期依赖的建模效果不如其他模型。Fnn 在该实验中表现最佳，尤其在 F1 值上的优越性，说明其在综合评价指标上占据优势。Gru 和 Rnn 的表现接近，但 Gru 略有优势。

实验 2 (基于大语言模型的零样本学习)结果如表 7 所示：

Table 7. The results of the performance indicators of each model in experiment 2
表 7. 实验 2 各模型的性能指标结果

Model	Acc	F1
Qwen2:7b	57.52%	50.01%
Qwen2.5:7b	74.35%	57.86%
Llama3.1:8b	75.35%	48.80%
Gpt-4o-Mini	82.57%	63.65%

Gpt-4o-Mini 表现最佳，具有最高的准确率(82.57%)和 F1 值(63.65%)。这表明该模型在任务中表现出色，能够较好地平衡精确度和召回率，适应任务要求。Llama3.1:8b 紧随其后，准确率为 75.35%，但 F1 值相对较低(48.80%)。尽管该模型在准确率上较高，但它的 F1 值较低，可能存在一定的精度或召回率的不平衡，说明它可能在处理某些类别或数据时遇到困难。Qwen2.5:7b 的表现稍逊于 Llama3.1:8b，准确率为 74.35%，F1 值为 57.86%。虽然 F1 值比 Llama3.1:8b 高，但整体表现不如 Gpt-4o-Mini。Qwen2:7b 在准确率(57.52%)和 F1 值(50.01%)上相对较低，说明它在任务中的表现逊色，可能在处理该任务时的泛化能力较弱。

实验 3 (基于大语言模型的少样本学习)结果如表 8 所示：

Table 8. The results of the performance indicators of each model in experiment 3
表 8. 实验 3 各模型的性能指标结果

Model	Acc	F1
Qwen2:7b	66.20%	49.74%
Qwen2.5:7b	79.13%	59.15%
Llama3.1:8b	78.35%	49.32%
Gpt-4o-Mini	84.70%	62.37%

实验三中，Gpt-4o-Mini 仍然表现最佳，具有最高的准确率(84.70%)和 F1 值(62.37%)。Qwen2.5:7b 的表现稍逊于 Gpt-4o-Mini，准确率为 79.13%，F1 值为 59.15%。Llama3.1:8b 准确率为 78.35%，但仍存在 F1 值偏低的问题。Qwen2:7b 在准确率(66.20%)和 F1 值(49.74%)上都相对较低。与实验二相比，基于各个模型的少样本学习方法情感分类准确率都有所提高，但 F1 值只呈现了小范围的上升或下降。

总之，在引文情感分类任务中，大语言模型普遍表现较差，尤其是 Qwen2:7b，而 Qwen2.5:7b 和 Llama3.1:8b 的表现稍有提升，Gpt-4o-Mini 表现最为出色，但其准确率和 F1 分数仍然低于大多数基于 Bert 的深度学习模型。总体来看，基于 Bert 的深度学习模型在该任务上的表现要优于大语言模型。

4. 结语

虽然基于大语言模型的零样本学习方法和少样本学习方法在引文情感分类任务中表现不佳，但与深度学习方法不同的是，该两种方法不需要任务特定的标注数据。大语言模型可以通过自然语言理解直接完成任务。通过修改任务描述和提示，可以将模型应用于不同类型的任务(例如情感分析、文本分类、命名实体识别等)，实现高效的任务扩展，无需为每个任务进行单独的训练。

但使用大语言模型进行零样本学习或是少样本学习都具有一定的局限性。首先其准确度受限：尽管大语言模型有很强的推理能力，但由于缺乏针对特定任务的训练，模型的准确性可能不如专门训练的模型。其次是大语言模型依赖于模型的预训练知识：模型的效果依赖于其在预训练阶段所学习到的知识，可能对一些特定领域或少数样本的处理效果不理想。

总之使用基于大语言模型的零样本学习方法或少样本学习方法进行引文情感分类是一种简便且高效的方法。尽管该方法在某些情境下可能没有传统训练方法准确，但它对于任务快速部署和应用非常有用，特别是在数据稀缺的情况下。未来的研究重点应该是在大语言模型在高效微调的前提下，如何提高模型性能以及其泛化能力，以及更好的结合实际应用问题。

基金项目

吉利学院“一院一品”教学改革项目(2024JG30253)。

参考文献

[1] Radicchi, F. (2012) In Science “There Is No Bad Publicity”: Papers Criticized in Comments Have High Scientific Impact. *Scientific Reports*, **2**, Article No. 815. https://doi.org/10.1038/srep00815

[2] Baird, L.M. and Oppenheim, C. (1994) Do Citations Matter? *Journal of Information Science*, **20**, 2-15. https://doi.org/10.1177/016555159402000102

[3] Chung, F. (2014) A Brief Survey of PageRank Algorithms. *IEEE Transactions on Network Science and Engineering*, **1**, 38-42. https://doi.org/10.1109/tnse.2014.2380315

[4] Page, L. (1998) Te PageRank Citation Ranking: Bringing Order to the Web. Technical Report, Stanford Digital Library Technologies Project.

[5] Bollen, J., Rodriquez, M.A. and Van de Sompel, H. (2006) Journal Status. *Scientometrics*, **69**, 669-687. https://doi.org/10.1007/s11192-006-0176-z

[6] Jiang, J., Xu, S. and You, L. (2023) An Optimization Ranking Approach Based on Weighted Citation Networks and P-Rank Algorithm. *Complexity*, **2023**, Article ID 7988848. https://doi.org/10.1155/2023/7988848

[7] Piryani, R., Madhavi, D. and Singh, V.K. (2017) Analytical Mapping of Opinion Mining and Sentiment Analysis Research during 2000-2015. *Information Processing & Management*, **53**, 122-150. https://doi.org/10.1016/j.ipm.2016.07.001

[8] Yousif, A., Niu, Z., Tarus, J.K. and Ahmad, A. (2017) A Survey on Sentiment Analysis of Scientific Citations. *Artificial Intelligence Review*, **52**, 1805-1838. https://doi.org/10.1007/s10462-017-9597-8

[9] Small, H. (2011) Interpreting Maps of Science Using Citation Context Sentiments: A Preliminary Investigation. *Scientometrics*, **87**, 373-388. https://doi.org/10.1007/s11192-011-0349-2

[10] Athar, A. (2011) Sentiment Analysis of Citations Using Sentence Structure-Based Features. *Proceedings of the ACL* 2011 *Student Session*, Portland, June 2011, 81-87.

[11] Poria, S., Chaturvedi, I., Cambria, E. and Hussain, A. (2016). Convolutional MKL Based Multimodal Emotion Recognition and Sentiment Analysis. 2016 *IEEE* 16*th International Conference on Data Mining* (*ICDM*), Barcelona, 12-15 December 2016, 439-448. https://doi.org/10.1109/icdm.2016.0055

[12] Ahmed Bilal, A., Ayhan Erdem, O. and Toklu, S. (2024) Children's Sentiment Analysis from Texts by Using Weight Updated Tuned with Random Forest Classification. *IEEE Access*, **12**, 70089-70104. https://doi.org/10.1109/access.2024.3400992

[13] Ghosal, T., Varanasi, K.K. and Kordoni, V. (2023) A Deep Multi-Tasking Approach Leveraging on Cited-Citing Paper Relationship for Citation Intent Classification. *Scientometrics*, **129**, 767-783. https://doi.org/10.1007/s11192-023-04811-5

[14] Beltagy, I., Lo, K. and Cohan, A. (2019) SciBERT: A Pretrained Language Model for Scientific Text. *Proceedings of the* 2019 *Conference on Empirical Methods in Natural Language Processing and the* 9*th International Joint Conference on Natural Language Processing* (*EMNLP-IJCNLP*), Hong Kong, November 2019, 3615-3620. https://doi.org/10.18653/v1/d19-1371

[15] Yu, D.H. and Hua, B.L. (2023) Sentiment Classification of Scientific Citation Based on Modified BERT Attention by Sentiment Dictionary. *Joint Workshop of the* 4*th Extraction and Evaluation of Knowledge Entities from Scientific Documents and the* 3*rd AI + Informetrics* (*EEKEAII*2023), Santa Fe, 26 June 2023, 59-64.

[16] 周文远, 王名扬, 井钰. 基于 AttentionSBGMC 模型的引文情感和引文目的自动分类研究[J]. 数据分析与知识发现, 2021, 5(12): 48-59.

[17] Narejo, K.R., Zan, H., Dharmani, K.P., Zhou, L., Alahmadi, T.J., Assam, M., *et al*. (2024) EEBERT: An Emoji-Enhanced BERT Fine-Tuning on Amazon Product Reviews for Text Sentiment Classification. *IEEE Access*, **12**, 131954-131967. https://doi.org/10.1109/access.2024.3456039

[18] Kheiri, K. and Karimi, H. (2023) SentimentGPT: Exploiting GPT for Advanced Sentiment Analysis and Its Departure from Current Machine Learning.

[19] Roumeliotis, K.I., Tselikas, N.D. and Nasiopoulos, D.K. (2024) LLMS and NLP Models in Cryptocurrency Sentiment Analysis: A Comparative Classification Study. *Big Data and Cognitive Computing*, **8**, Article No. 63. https://doi.org/10.3390/bdcc8060063

[20] Cohan, A., Ammar, W., van Zuylen, M. and Cady, F. (2019) Structural Scaffolds for Citation Intent Classification in Scientific Publications. *Proceedings of the* 2019 *Conference of the North American Chapter of the Association for Computational Linguistics*: *Human Language Technologies*, **1**, 3586-3596. https://doi.org/10.18653/v1/n19-1361