

基于提示的小样本情感分析综述

姜鑫, 马宏伟, 张展峰

(山东建筑大学 计算机科学与技术学院, 山东 济南 250102)

摘要: 随着多媒体平台和大规模语言模型的迅猛发展, 采用基于提示的方法实现小样本情感分析, 对用户需求分析和系统服务改进具有重要意义。基于提示的小样本情感分析研究, 致力于在不同应用场景中合理利用预训练语言模型理解分类任务, 推理情感类别。首先, 阐述了小样本情感分析的问题背景; 其次, 介绍了使用提示—微调、提示调优和上下文学习的方法步骤; 再次, 系统对比了近来基于提示的小样本情感分析的主流技术, 归纳总结了相关语料库、预训练语言模型与提示模板; 最后, 对基于提示的小样本情感分析未来可能的研究方向进行展望, 为句子级文本分类和提示学习相关领域的研究提供参考。

关键词: 小样本情感分析; 预训练语言模型; 提示—微调; 提示调优; 上下文学习

DOI: 10.11907/rjdk.241771

中图分类号: TP391.1

文献标识码: A

文章编号: 1672-7800(2026)001-0213-08

扫描二维码阅读全文:



Review on Prompt-Based Few-Shot Sentiment Analysis

JIANG Xin, MA Hongwei, ZHANG Zhanfeng

(School of Computer Science and Technology, Shandong Jianzhu University, Ji'nan 250102, China)

Abstract: With the rapid development of multimedia platforms and large-scale language models, implementing few-shot sentiment analysis using prompt based methods is of great significance for user demand analysis and system service improvement. The research on prompt based few-shot sentiment analysis aims to make reasonable use of pretrained language models to understand classification tasks and infer sentiment categories in different application scenarios. Firstly, the problem background of few-shot sentiment analysis was presented. Secondly, the steps of using prompt based fine-tuning, prompt tuning, and in-context learning were introduced. Then, the mainstream technologies of prompt based few-shot sentiment analysis in recent years were systematically compared. Additionally, relevant corpora, pretrained language models, and prompt templates were summarized. Finally, the possible future research directions of prompt based few-shot sentiment analysis were discussed, providing a reference for research in the fields of sentence-level text classification and prompt learning.

Key Words: few-shot sentiment analysis; pretrained language model; prompt based fine-tuning; prompt tuning; in-context learning

0 引言

情感分析是在文档、句子和方面级别对给定文本的极性进行分类, 判断样本表达的意见是积极的、消极的还是中性的, 或者区分诸如“愤怒”、“悲伤”和“快乐”等情绪状态^[1-2]。情感分析研究专注于从评论中自动提取用户的情感状态, 应用场景广泛, 包括新闻媒体、在线教育平台、旅游服务系统和电子商务软件等^[3-4]。

在情感分析的极性检测子任务上, IMDB^[5]、SST-2^[6]

等许多基准数据集都已经趋于饱和。然而, 这并不意味着句子级情感分类在所有实际场景中都得到了解决, 因为许多实际应用场景往往缺乏充足的标注数据或有效数据。小样本学习 (Few-Shot Learning, FSL) 指通过少量有标签的监督样本, 或使用这样的样本进行某种方式的训练, 尽可能降低模型结果与真实结果的误差^[7]。

自然语言处理领域正在从“预训练—微调”过渡到“预训练—提示—预测”第四范式阶段^[8]。随着 ChatGPT 引发 AI 浪潮, 全球范围内的科技公司都在积极研究和开发大规模语言模型 (Large Language Model, LLM), 巨量参数的

收稿日期: 2024-12-16

作者简介: 姜鑫 (1999-), 女, 山东建筑大学计算机科学与技术学院硕士研究生, 研究方向为自然语言处理、数据挖掘与机器学习; 马宏伟 (1969-), 男, 博士, 山东建筑大学计算机科学与技术学院教授, 研究方向为自然语言处理、计算机网络; 张展峰 (1997-), 男, 山东建筑大学计算机科学与技术学院硕士研究生, 研究方向为自然语言处理、数据挖掘与机器学习。本文通讯作者: 马宏伟。

模型使任务性能增益的同时也带来很多弊端。其中包括模型的迁移能力下降、训练样本增加和对设备性能要求更高等问题。第四范式通过定义一个新的提示函数,使模型在小样本甚至零样本学习中具有显著优势。基于提示的方法可以通过学习一个语言模型解决以上问题,该语言模型对输入文本 $p(x; \theta)$ 的概率进行建模并预测输出,从而减少或消除对大量标注数据集的需求^[8]。

小样本情感分析是指模型仅利用少量标注数据(如每个标签类别提供 k 个样本的 k -shot 设置),在下游任务中实现情感分类。这一任务与文本分类、自然语言理解任务密切相关,但也具有其独特性。相比之下,小样本情感分析任务的可训练数据更为有限,其标签通常为积极、中立、消极等情感极性或更具体的情绪类型。从提示模板的角度来看,小样本情感分析的任务类型较为单一,可采用格式统一的提示模板;而文本分类和自然语言理解任务则涵盖多种类型,包括新闻主题分类、意图识别、文本匹配、代码纠错等,这些任务的提示模板内容差异较大。在应用场景方面,自然语言理解覆盖的领域最为广泛,其输入甚至能扩展至表格、代码等结构化或非纯文本格式(需结合特定处理技术);文本分类主要聚焦于将文本划分到预设的主题或类别中;而情感分析则用于系统对用户情感类文本的分析,通常被视为一项更基础的任务。尽管三者应用场景不同,但都能借助预训练语言模型(Pretrained Language Model, PLM)进行预测或生成答案。综上,小样本情感分析任务的解决方法可灵活迁移应用于不同领域和任务中。如何充分利用 PLM 的先验知识和语言表征能力,使其更好地适应低资源环境、提高响应速度,并平衡实际生产环境中准确率与效率之间的矛盾,是小样本情感分析面临的核心挑战。

基于提示的小样本情感分析方法通过提示使模型识别情感分析任务类型,并输出对应的情感类别。这类方法可大致分为三类:提示—微调、提示调优(Prompt Tuning, PT)和上下文学习(In-Context Learning, ICL)。该任务常用的预训练语言模型主要分为自编码语言模型(如 BERT^[9] 和 RoBERTa^[10])、序列到序列语言模型(如 T5^[11])与自回归的生成模型(如 GPT-2^[12] 和 OPT^[13])三大类。前者多用于提示—微调 and 提示调优,后两者多用于上下文学习。当模型的下游任务之间类型不一致时(如文本分类、文本蕴含和问题解答等),模型需要自适应地选择不同的模板,即为不同任务分配不同的提示信息。对于不同主题的小样本情感分析任务,提示模板分为突出该领域任务特性的特定模板(如“My review for last nights film: {input} \n The critics agreed that this movie was {good, bad}”)和情感分析任务的通用模板(如“Text: {input} \n It was <mask>”)。不同的提示可能导致最终准确度的巨大差异,当模板固定时,标签词与语义类别匹配得越好,最终的准确率就越高。结合不同特点的预训练语言模型选择合适的训练方法,并

设计合理的自动提示模板优化方法,可以进一步提高任务性能。

本文首先阐述了基于提示的小样本情感分析研究背景,然后围绕常用数据集、预训练语言模型与提示模板,详细分析了提示—微调、提示调优和上下文学习三类小样本情感分析方法,对比其性能表现与优缺点,最后对该研究方向进行总结与展望。

1 基于提示—微调的小样本情感分析

对基于提示学习的小样本情感分类问题进行建模,给定输入空间 X , 输出空间 C , 标签预测词 $V^* \subset \{v_1^*, v_2^*, \dots, v_i^*, \dots, v_n^*\}$, 标签空间 $V \subset \{v_1, v_2, \dots, v_i, \dots, v_n\}$ 是特定于分类任务的,包含所有可能的类别标签。模板(pattern)是指一个特定的文本结构,引导 PLM 执行特定任务,通常包括一个或多个占位符,如[MASK]。Verbalizer是指模型预测[MASK]的内容到最终类别标签之间的映射,即 $v_i^* \rightarrow v_i$ 。举例来说,电影评论“I love this movie!”,此句的输入空间是文本本身,提示模板为“Review: 输入.Sentiment: [MASK]”。输出空间是一个概率分布,表示该评论属于不同情感类别的概率。标签空间包括 positive、neutral 和 negative 这 3 个标签。使用 verbalizer 将具体的标签预测词,如 great、excellent 或 brilliant 映射到最终类别标签 positive。小样本情感分析进行提示—微调的步骤如下:首先,通过不同的提示构建方法得到提示模板;其次,利用小样本情感分析训练集对模型参数进行微调;最后,使用单个或集成的提示模板,对测试样本的输入进行标签词映射,并通过 verbalizer 将[MASK]位置的输出结果映射到具体类别标签上。

为解决小样本情景下,结果过拟合以及灾难性遗忘的问题, Schick 等^[14]利用先验知识人工定义提示模板,将目标分类任务转化为完形填空形式。针对不同的 pattern-verbalizer 对,本研究采用交叉熵损失与语言建模损失的加权和作为总体损失函数的方法,对多个 PLM 的参数进行微调。通过集成这些微调后的模型对未标注数据进行标注,最终将集成模型的预测知识蒸馏至目标输出模型中。然而在实际应用场景中,该方法存在一个明显的局限性:为了达到最优效果,模型不得不依赖广泛测试多个不同的 pattern 和 verbalizer 来构建提示模板,这一过程既耗时间又消耗资源。Gao 等^[15]在 SST-2 等数据集上验证了使用相同模板不同标签词,和相同标签词不同模板对模型效果都有显著影响。利用 <X> 或 <Y> 替换一个或多个连续文本片段, T5 模型会自动生成不同的 pattern,不需要指定离散模板的具体占位标记数目。通过搜索方法获取 verbalizer 并形成提示模板后,经过微调在 CR(Customer Review)小样本数据集上取得了优于传统微调方法的效果^[16]。这一结果表明,优化提示模板以更好地匹配任务语义,对小样本情

感分析性能的提升具有重要作用。虽然离散提示模板能够释放预训练模型知识潜力,不会引入随机初始化参数,且可以避免过拟合,但是结果稳定性差,固定的离散任务模板无法全局优化^[14]。由于少量样本难以保证很好地优化随机初始化的连续提示模板,Gu 等^[17]提出先让连续提示(软提示)在大量无标注预训练语料库进行预训练,然后将其加载至下游任务的 PLM 上进行微调。采用预训练任务的提示来初始化下游任务的提示模板,虽然会导致训练时间有所增加,但在 32-shot 设置下,该方法在 SST-2 和 SST-5^[6]任务的准确率显著超越了传统的全模型微调方法。

综上,提示模板的选择与自动构造都对小样本情感分类任务的性能有很大影响。离散提示模板可读性强,但无法充分利用语言模型,与之相反,连续提示能在一定范围内使模型选出相对最优模板,但不满足人类语言的语法形态,可解释性差^[18]。相比于全模型微调,提示-微调增加了自动构建提示模板的工作,充分利用 PLM 的掩码语言建模能力,有助于提高小样本情感分析任务的准确率。

2 基于提示调优的小样本情感分析

针对不同的情感分析任务,模板定义为 $T=[x][h_0][h_1] \dots [h_m][\text{MASK}]$ 。其中, $[h_m]$ 代表的不是具体的单词,而是一个抽象的 token,本质是向量。小样本情感分析进行提示调优的步骤如下:首先,初始化任务相关的连续提示模板;其次,利用训练样本集对提示参数进行调优;最后,利用得到的提示模板,对测试样本的输入进行标签词映射,将 $[\text{MASK}]$ 位置的输出结果通过 verbalizer 映射到具体类别标签上。具体流程如图 1 所示。

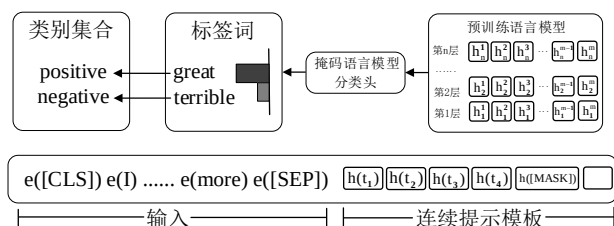


Fig. 1 Prompt-tuning for few-shot sentiment analysis

图 1 使用提示调优进行小样本情感分析

PT 的本质是复用预训练语言模型在预训练阶段所使用的目标和参数,从而避免引入新参数导致过拟合问题。Wang 等^[19]提出从任务转换的角度将每一个实例重新表述为二分类的文本蕴含任务,进而缩小预测空间。尤其是在小样本的多类别情感分析任务 Yelp 中,该方法避免了标签预测概率分散导致的准确度降低^[20]。这表明,设计全新的预训练任务以满足下游情感分析任务的需求是可行方向。

PT 的实现仅需考虑如何设计提示模板,而模型和训练目标都复用了预训练阶段的模型和目标,在训练阶段不添加或添加少量与模板有关的参数。赵文辉等^[21]提出基于提示调优的多领域中文情感分析方法,由连续提示和离散 to-

ken 构成混合提示模板。离散提示用于指示情感文本的所属领域和待选的情感标签,而微调的连续提示用于学习特定领域的文本情感特征。在 32-shot 的中文数据集上,该方法的准确率达到 91.64%,证明了提示调优在小样本学习条件下的优越性,尤其是在需要适应不同领域情感特征时,提示调优能以较少的训练步骤达到与模型全参数微调相当的效果。结果表明,提示模板的可读性不重要,只需要优化模板的 token 组成、插入位置和输出的候选空间,由文本向量空间的点组成提示能更充分地利用原始 PLM,起到更优的小样本学习效果。然而,训练软提示通常需要大量的超参数调整和更长的训练时间才能达到预期的性能^[22]。

由于软提示的收敛时间长,Liu 等^[23]提出使用一个提示生成器为每个输入生成一个独立的提示。引入的提示生成器对隐藏层表征先降维,进行池化和激活,再升维。该提示插入 PLM 的中间层,而非 embedding 层或所有层。在 RoBERTa-large 上,该方法的训练速度可达全模型微调的 2 倍,而且在小样本的 SST-2 和 MR (Movie Review)^[24]数据集上,准确率高于原始的提示调优^[22]。实验结果表明,适当缩短提示的传播距离可以提高提示调优性能,设计提示时需要平衡标签信号到提示向量的传播距离和提示向量对模型输出能力的影响。

为正确预测不同用户的情感标签,Zhong 等^[25]提出了一种针对用户自身数据单独微调的虚拟前缀 token 方法。在前缀调优阶段,用户间共享冻结参数的通用模型^[26]。实验结果表明,只存储和训练微量参数在特定用户的数据上有更好的泛化能力,作为微调 PLM 所有参数的替代方案,在大量新用户加入的情感分析场景中发挥了重要作用。为降低训练和存储成本,同时提高整体性能,Mireshghallah 等^[27]训练了单一的共享个性化模型。通过在每个用户输入文本前添加不可训练的唯一标识字符串,向模型提供与用户相关的提示。同时,以最小化分类的交叉熵损失为目标进行微调,帮助模型学习个人用户行为和偏好,从而实现个性化的响应生成。在 IMDB 和 Yelp 数据集上的实验结果显示,采用随机生成的用户标识符可使模型性能较文献^[25]提升 13 个百分点。该方法特别适用于用户规模有限但需快速实现个性化模型部署的应用场景。这一发现进一步证实,额外添加参数的提示调优在个性化情感分析中具有良好的应用前景。

3 基于上下文学习的小样本情感分析

上下文学习主要面对超大规模的 PLM,使用少量的下游任务数据组成演示(demonstration)。通过加入固定提示模板,将查询的相关演示和测试样本进行拼接后输入 PLM 进行预测^[28]。与模型预训练阶段的监督学习不同,ICL 不涉及模型参数的更新,而是直接利用 PLM 进行预测。这也是 ICL 区别于提示-微调 and 提示调优的特性,在推理时不

需要利用训练样本调整模型参数。给定输入为 x ,样本标签类别集合为 $C = \{c_i, i \in [1, 2, \dots, n]\}$, ICL 预测的是概率分布 $P(c_i|x)$ 。基于上下文学习的小样本情感分析方法,旨在设计演示和评分函数,从而提高模型推理情感类别的准确率。具体流程如图2所示。

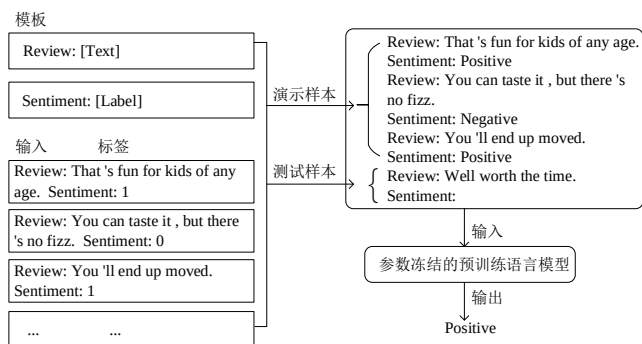


Fig. 2 In-context learning for few-shot sentiment analysis

图2 使用上下文学习进行小样本情感分析

不同的演示样本(包括数量和顺序)、标签词、提示模板和模型规模都会对ICL的性能产生一定影响^[29-31]。其中,提示模板构造方法包括多提示集成和改进提示的答案工程,即设计 verbalizer。多提示集成是指在推理时将多个提示输入的标签预测概率进行加权,利用测试输入选择多种提示模板^[8]。由于人工设计提示模板时标签词覆盖范围有限,Hu等^[32]设计了 verbalizer,通过使用外部知识库为每个标签生成一组标签词。这种外部扩展的标签词集不仅包含互为同义词,还涵盖了不同的粒度和角度。该方法通过将多个词的预测概率映射到某个特定标签,在小样本 Amazon^[33]和 IMDB 数据集上取得了优于全模型微调 and 原始提示调优的平均准确率,验证了有效引入外部知识对小样本情感分析任务的重要性。

为进一步提高超大规模模型的推理能力,Zhou等^[34]设计了情感感知的思维链提示。通过多步提示的中间推理步骤,引导模型从粗粒度(如正面的、不明确和负面的)到细粒度(如感激的、惊讶的和伤心的)进行情感标签预测。考虑到相同类别样本之间的情感相关性,设计相关性加权的软对比损失函数计算正负样本之间的距离,并调整不同情感标签之间的距离。通过优化包括交叉熵损失和软对比损失的总损失函数进行训练,有助于模型更好地区分和理解细粒度情感之间的微妙差别,从而提高情感分析的准确性。在包含28种情感的小样本 GoEmotions^[35]数据集上,该方法的性能优于使用多层感知机的 P-tuning^[36]和 PPT^[17]等提示调优方法。因此,未来可将细粒度情感分类作为研究重点,增强模型对上下文语义关系的理解,进一步提升小样本情感分析的性能。

预训练语言模型对演示样本的选择表现出显著敏感性,其预测结果具有高度可变性^[37]。因此,在ICL中样本选择策略至关重要。根据特定属性,通过使用启发式和无监督的度量,能够有效评估样本增加模型性能的潜力,从

而优化样本选择。当前主流方法主要根据与测试样本的相似性来选择演示样本。Long等^[38]检索与测试样本的余弦相似度最高的跨领域样本子集,引导LLM通过学习目标领域分布和区分性任务信号来适应上下文环境,同时利用增强的跨领域上下文演示样本进行学习。实验结果验证了实现域外自适应在真实的小样本情感分析场景中的必要性。另一种特定属性为样本的信息量或不确定性。Rouzegar等^[39]在IMDB中利用LLM输出的样本置信度分数作为不确定性指标,指导数据进行选择性注释。实验结果表明,结合LLM和人工注释的主动学习策略能够同时降低数据注释成本,并维持或提高模型的准确性,平衡模型的可靠性与效率。此外,部分改进方法通过多属性融合策略进一步提升效果,如Li等^[40]先在过滤阶段通过信息量指标逐步筛选样本,然后在搜索阶段使用多样性函数更新候选样本排列。该方法在SST-2、SST-5、Amazon和MR上的准确率显著超过利用单一属性的样本选择策略。但是,以上方法尚未能够充分挖掘“输入-输出”映射。在小样本情感分析中,上下文学习模型内部工作机制的理论分析仍是未来工作的重要方向。

此外,Zhao等^[41]提出ICL中存在多数标签偏差、近因偏差和常见 token 偏差等问题。为减少上述3种偏差,提出一种基于模型校准的方法,具体可分为两类:一是直接改变模型面向测试集标签概率的计算方法;二是得到预测概率后再调整测试集标签概率分布,间接改变输出结果^[42-43]的计算方法。

在上下文学习中给定输入 x 和输出 y ,类别空间 $C = \{c_i, i \in [1, 2, \dots, n]\}$,直接计算 $p(y|x)$ 是应用最广泛的概率计算方法。在此基础上,Min等^[44]提出了噪声信道模型,以实现 $p(x|y)p(y)$ 最大化。根据贝叶斯规则将 $p(c_i|x)$ 重参数化为 $\frac{p(x|c_i)p(c_i)}{p(x)}$,化简后计算 $p(x|c_i)$ 。在小样本情感分析中,该方法通过解释输入中的每个单词,放大了低数据区中的训练信号。实验结果表明,信道模型对文本分类和小样本设置中的分布转移更具鲁棒性,显著改善了训练数据标签不平衡造成的低性能问题。

当提供与上下文无关的输入“N/A”时,LLM表现出积极预测的趋势,在SST-2任务上可能性超过90%。这类偏差称为模板偏差,即仅使用提示模板促使LLM在没有任何演示样本的情况下生成偏向特定答案的预测^[45]。为减少模板偏差,Zhao等^[41]提出对测试样本的输出概率 p 应用线性变换 $q = \sigma(Wp + b)$,使用提示模板对无内容字符串计算输出概率 p_{cf} ,权重矩阵 W 更新为 $\text{diag}(p_{cf})^{-1}$, b 为零向量。Liu等^[45]结合最大信息增益属性与模型校准,利用GPT-3-175B^[46]在1-shot的SST-2任务上达到96.5%的高准确率。实验结果表明,在提升小样本情感分析的性能方面,结合样本选择策略和模型校准的方法进行上下文学习是一种可行的研究方向,未来可向更长文本的小样本句子级和文

档级情感分析任务拓展。

4 小样本情感分析数据集与方法对比

4.1 常用数据集

表 1 总结了小样本情感分析任务的常用数据集。

Table 1 Common datasets for few-shot sentiment analysis

表 1 小样本情感分析常用数据集

类别	文本	数据集	数据集介绍	下载地址
二元分类	句子级 短文本	SST-2 ^[6]	共 67 349 条训练样本,1 820 条测试样本,872 条验证样本;包括用户对电影的评论	https://nlp.stanford.edu/sentiment
		MR ^[24]	共 10 662 条样本,正负面样本数量相同;包括用户对电影的评论	https://www.cs.cornell.edu/people/pabo/movie-review-data/rt-polaritydata.tar.gz
		CR ^[16]	共 1 775 条训练样本,2K 条测试样本;包括用户对五种电子产品的评论	http://www.cs.uic.edu/~liub/FBS/Customer-ReviewData.zip
多元分类	句子级长文本	IMDB ^[5]	共 50K 条样本,训练集和测试集样本数量相同;包括用户对电影的评论	https://ai.stanford.edu/~amaas/data/sentiment/
	句子级 短文本	SST-5 ^[6]	共 11 855 条样本;包括用户对电影的评论,标签为负面、稍微负面、中立、稍微正面和正面	https://nlp.stanford.edu/sentiment
	句子级 长文本	Amazon ^[33]	共 3 000K 条训练样本和 650K 条测试样本;包括用户对书籍、服装和电子产品等商品的评论,标签为 1 星到 5 星评分	https://snap.stanford.edu/data/web-Amazon-links.html
		Yelp ^[20]	每一类有 650K 条训练样本和 50K 条测试样本;包括用户对餐馆、购物中心和医院等商户的评论,标签为 1 星到 5 星评分	https://www.yelp.com/dataset

Table 2 Comparison of some experiments related to few-shot sentiment analysis

表 2 小样本情感分析相关实验对比

(%)

研究方法	PLM	k-shot	任务性能				优缺点
			(准确率、标准差)				
			SST-2	SST-5	MR	CR	
Gao 等 ^[15]	RoBERTa-Large (355 M) ^[10]	16-shot (每个标签类别包 含 16 个样本)	93.0 (0.6)	49.5 (1.7)	87.7 (1.4)	91.0 (0.9)	优点:在序列到序列模型上创新了自动构建提示模板的生成式方法,首次对小样本情感分析任务先进行提示-微调,再结合上下文学习。缺点:扩展搜索离散模板空间具有限制性
		32-shot (共 32 个训练样 本)	93.5 (0.3)	50.2 (0.7)	—	—	
Mahabadi 等 ^[26]	RoBERTa-Large (355 M) ^[10]	32-shot (16 个训练集样本 和 16 个验证集样 本)	90.7 (1.2)	42.7 (2.9)	86.3 (1.6)	90.0 (1.4)	优点:首次通过插入任务相关的 adapter 层和多 token 的标签 embedding 替换人工设计的提示和 verbalizer,减少训练数据依赖和内存损耗,embedding 与 PLM 词汇表无关,可以从有限的学习中学习更准确的语义表示
Liu 等 ^[23]	RoBERTa-Large (355 M) ^[10]	共 100 个训练样本	92.7 (0.8)	—	88.5 (0.5)	—	优点:通过实证分析表明提示调优性能滞后是因为从标签信号到输入提示的传播路径过长。在模型中间层生成实例依赖的提示,延迟提示调优,其性能与基于 adapter 的调优方法乃至全模型调优方法相当,有效降低训练成本
Aly 等 ^[47]	T0(3 B) ^[48]	32-shot (每个标签类别包 含 32 个样本)	—	48.6 (2.5)	—	92.5 (1.1)	优点:使用指令调优检索合适的任务指令作为提示,同其他无提示方法相比具有有竞争力的性能和效率。缺点:需要一些数据信号来优化答案选择空间,在极低数据量或训练样本高度不平衡的情况下预期效果不佳
Zhang 等 ^[49]	RoBERTa-large (355 M) ^[10]	16-shot (每个标签类别包 含 16 个样本)	91.9 (2.0)	—	88.0 (1.1)	91.1 (1.6)	优点:使用强化学习方法设计了一个灵活的动作空间,允许代理编辑指令、上下文演示和 verbalizer,将连续提示编辑前后的分数差作为奖励,对提示池大小和小样本演示的数量表现出良好的鲁棒性。缺点:需要大量计算资源
Min 等 ^[44]	GPT2-Large (774 M) ^[11]	16-shot (共 16 个训练样 本)	85.0	36.2	80.5	80.8	优点:当训练样本数量较少,训练数据中的标签不平衡,或者需要泛化到未知标签时,信道提示调优结合上下文学习的方法比其他微调方法表现更好。缺点:实验在因果语言模型中效果较好,但是在掩码语言模型中的适用性有待验证
Jiang 等 ^[50]	GPT-2-Large (774 M) ^[11]	8-shot (共 8 个演示样本)	88.17	38.02	82.23	87.96	优点:首次从理论和实证分析上确定 ICL 中的模型偏差是数据分布的标签边缘偏移造成的。缺点:在包含标签类别不平衡的现实场景下并不成立,降低了多元情感分析的准确性

而一些研究为了模拟真实数据分布,放宽了先前工作中的假设,即每个标签有相同数量的训练样本^[15],不采用均匀采样。由表2可知,基于上下文学习的方法在小样本情感分析的二分类任务上取得了较好效果,但对于多分类任务

还有很大的提升空间。

表3针对不同类型的小样本情感分析方法展开了系统性对比分析,从核心思想、应用场景、方法优势以及局限性等多个维度进行了深入剖析。

Table 3 Comparison of some research methods related to few-shot sentiment analysis

表3 小样本情感分析相关研究方法对比

研究方法	核心思想	适用场景	优点	缺点
提示-微调	设计提示模板,训练过程中针对不同的下游任务微调模型参数	适用于特定任务,且适用于具备大量的计算资源和存储能力的环境	①在训练数据充足时,模型预测效果较好;②在较小模型上即表现出良好的域内性能;当使用大模型时,其在域内外的性能均优于ICL方法,且能通过持续训练进一步提升	①训练成本最高;②微调过程中的高性能仍然依赖大量的监督语料;③微调可能会过度地将模型定制为特定任务的格式,潜在影响对其他新任务的适应性,域外泛化性能低
提示调优	设计连续提示模板,有效微调模板中的词向量或模型的部分参数	适用于小模型,且适用于长文本和二分类的小样本情感分析	①即使部署在小模型上也能很好地工作;②模型推理时间更短;③允许无限的训练样本,从样本中获益更多	①模型需要训练(在处理大型模型时,训练代价变高);②训练需要具有一定的专业知识;③模型不具备可重用性
上下文学习	将提示模板、演示样本组成上下文,拼接在要查询的测试样本之前,直接输入到预训练语言模型中进行预测	适用于较少训练数据和大规模预训练语言模型,且适用于短文本和多分类的小样本情感分析	①普通用户能够更便捷地应用这种方法;②模型在不同任务上可重用;③模型不需要训练	①无法直接应用在推理过程复杂和文档级的情感分析任务上;②需要LLM才能工作,低资源语言开发的模型适用性低;③模型推理时间较长;④模型的固定输入长度大小限制了演示样本数量

由表3可知,3种方法的性能很大程度上都依赖于模板中的提示和 verbalizer 的选择,这表明选择合适的训练数据和模板在基于提示的小样本情感分析任务中的重要性。Yang等^[51]在二分类的小样本情感分析任务 Amazon 中,基于相应域内训练集(Amazon 和 Amazon-food)和域外测试集(SST-2 和转化为二分类任务的 Yelp)进行了评估。实验结果表明,采用上下文学习的模型微调方法在这些数据集上的表现均不如原始的微调模型。与生成任务进行对比,分类任务对演示样本更敏感,而且使用上下文学习的模型微调往往比原始微调方法的最终损失更高。这种现象可能源于模型直接从演示样本中复制标签,而非利用相关信息进行推理,进一步说明了模型校准在上下文学习中的关键作用。

5 未来研究方向

小样本情感分析任务是当代新闻、社交及电子商务等多媒体平台中必不可少的组成部分。从微调的角度看,小样本学习具有挑战性;从预训练语言模型的角度看,模型的任务学习能力主要通过模型规模激发,并随着演示样本数量的增加而显著提升。因此,需要探索提示模板构造方法、演示样本选择策略和模型校准方法,以加速连续提示模板收敛,提高模型推理准确率,平衡预测性能和训练成本。当前,多分类的小样本情感分析任务性能仍有较大的提升空间,亟待深入研究。基于提示的小样本情感分析中仍有诸多发展的可能性值得进一步探索,具体包括:

(1)探索构建更优的提示模板,以提升小样本情感分析准确率;研究如何选择最具代表性的演示样本,以增强

预训练语言模型的性能。

(2)针对小样本情感分析中的标签不平衡和多分类挑战,研究有效的模型校准技术,从而提高模型的泛化能力,并优化细粒度情感分类算法。

(3)研究个性化模型和域适应技术,提高算法的计算效率和资源优化;深入分析预训练语言模型的内部机制,增强小样本情感分析的解释性和用户信任。

6 结语

本文系统梳理了基于提示的小样本情感分析的相关研究成果,全面阐述了常用数据集、预训练语言模型和离散提示模板,并对基于提示-微调方法、提示调优方法以及上下文学习方法的性能、优势和局限性进行了深入对比分析。未来,随着预训练语言模型规模不断扩大,基于上下文学习的小样本情感分析将成为情感计算领域的重要研究方向。

参考文献:

- [1] PORIA S, HAZARIKA D, MAJUMDER N, et al. Beneath the tip of the iceberg: current challenges and new directions in sentiment analysis research [J]. IEEE Transactions on Affective Computing, 2023, 14(1): 108-132.
- [2] LIU B. Sentiment analysis and opinion mining [M]. Vermont: Morgan & Claypool, 2012: 6.
- [3] XIAO J M, CHEN S Z, FENG Z Y, et al. An automatic analysis of user reviews method for APP evolution and maintenance[J]. Chinese Journal of Computers, 2020, 43(11): 2184-2202.
肖建茂,陈世展,冯志勇,等.一种基于用户评论自动分析的APP维护和演化方法[J].计算机学报,2020,43(11):2184-2202.

- [4] ZENG H L, LI L, LYU S Y, et al. Risk identification metho for news public opinion driven by prompt learning[J]. Computer Engineering and Applications, 2024, 60(1): 182–188.
曾慧玲, 李琳, 吕思洋, 等. 提示学习驱动的新闻舆情风险识别方法研究[J]. 计算机工程与应用, 2024, 60(1): 182–188.
- [5] MAAS A L, DALY R E, PHAM P T, et al. Learning word vectors for sentiment analysis [C]// Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, 2011: 142–150.
- [6] SOCHER R, PERELYGIN A, WU J, et al. Recursive deep models for semantic compositionality over a sentiment treebank [C]// Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, 2013: 1631–1642.
- [7] WANG Y Q, YAO Q M, KWOK J, et al. Generalizing from a few examples: a survey on few-shot learning [J]. ACM Computing Surveys, 2020, 53(3): 1–34.
- [8] LIU P F, YUAN W Z, FU J L, et al. Pre-train, prompt, and predict: a systematic survey of prompting methods in natural language processing [J]. ACM Computing Surveys, 2023, 55(9): 1–35.
- [9] DEVLIN J, CHANG M W, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding [C]// Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics, 2018: 4171–4186.
- [10] LIU Y H, OTT M, GOYAL N, et al. RoBERTa: a robustly optimized BERT pretraining approach [C]// Proceedings of the 2020 International Conference on Learning Representations, 2020: 1–12.
- [11] RAFFEL C, SHAZEER N, ROBERTS A, et al. Exploring the limits of transfer learning with a unified text-to-text transformer [J]. Journal of Machine Learning Research, 2020, 21(1): 1–67.
- [12] RADFORD A, WU J, CHILD R, et al. Language models are unsupervised multitask learners [EB/OL]. <https://openai.com/index/better-language-models>.
- [13] ZHANG S, ROLLER S, GOYAL N, et al. OPT: open pre-trained transformer language models [DB/OL]. <https://arxiv.org/abs/2205.01068>.
- [14] SCHICK T, HINRICH S. Exploiting cloze-questions for few-shot text classification and natural language inference [C]// Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics, 2021: 255–269.
- [15] GAO T Y, FISCH A, CHEN D Q. Making pre-trained language models better few-shot learners [DB/OL]. <https://arxiv.org/abs/2012.15723>.
- [16] HU M Q, LIU B. Mining and summarizing customer reviews [C]// Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2004: 168–177.
- [17] GU Y X, HAN X, LIU Z Y, et al. PPT: pre-trained prompt tuning for few-shot learning [C]// Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, 2022: 8410–8423.
- [18] YU B H, CAI X Y, WEI J X. Few-shot text classification method based on prompt learning [J]. Journal of Computer Applications, 2023, 43(9): 2735–2740.
于碧辉, 蔡兴业, 魏靖短. 基于提示学习的小样本文本分类方法[J]. 计算机应用, 2023, 43(9): 2735–2740.
- [19] WANG S N, FANG H, KHABSA M, et al. Entailment as few-shot learner[DB/OL]. <https://arxiv.org/abs/2104.14690>.
- [20] TANG D Y, QIN B, LIU T. Document modeling with gated recurrent neural network for sentiment classification [C]// Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, 2015: 1422–1432.
- [21] ZHAO W H, WU X L, LING J, et al. Multi-domain sentiment analysis of Chinese text based on prompt tuning [J]. Computer Engineering and Science, 2024, 46(1): 179–190.
赵文辉, 吴晓鸽, 凌捷, 等. 基于 prompt tuning 的中文文本多领域情感分析研究[J]. 计算机工程与科学, 2024, 46(1): 179–190.
- [22] LESTER B, AL-RFOU R, CONSTANT N. The power of scale for parameter-efficient prompt tuning [C]// Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, 2021: 3045–3059.
- [23] LIU X Y, SUN T X, HUANG X J, et al. Late prompt tuning: a late prompt could be better than many prompts [C]// Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, 2022: 1325–1338.
- [24] PANG B, LEE L. Seeing stars: exploiting class relationships for sentiment categorization with respect to rating scales [C]// Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics, 2005: 115–124.
- [25] ZHONG W J, TANG D Y, WANG J H, et al. UserAdapter: few-shot user learning in sentiment analysis [C]// Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics, 2021: 1484–1488.
- [26] LI X L, LIANG P. Prefix-tuning: optimizing continuous prompts for generation [C]// Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, 2021: 4582–4597.
- [27] MIRESHGHALLAH F, SHRIVASTAVA V, SHOKOUHI M, et al. UserIdentifier: implicit user representations for simple and effective personalized sentiment analysis [C]// Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics, 2022: 3449–3456.
- [28] DONG Q, LI L, DAI D, et al. A survey on in-context learning [C]// Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, 2024: 1107–1128.
- [29] LIU J C, SHEN D H, ZHANG Y Z, et al. What makes good in-context examples for GPT-3? [C]// Proceedings of Deep Learning Inside Out: The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures, 2022: 100–114.
- [30] LU Y, BARTOLO M, MOORE A, et al. Fantastically ordered prompts and where to find them: overcoming few-shot prompt order sensitivity [C]// Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, 2022: 8086–8098.
- [31] MIN S, LYU X, HOLTZMAN A, et al. Rethinking the role of demonstrations: what makes in-context learning work? [C]// Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, 2022: 11048–11064.
- [32] HU S D, DING N, WANG H D, et al. Knowledgeable prompt-tuning: incorporating knowledge into prompt verbalizer for text classification [C]// Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, 2022: 2225–2240.
- [33] MCAULEY J, LESKOVEC J. Hidden factors and hidden topics: understanding rating dimensions with review text [C]// Proceedings of the 7th ACM Conference on Recommender Systems, 2013: 165–172.
- [34] ZHOU J Y, ZHOU J, ZHAO J B, et al. A soft contrastive learning-based prompt model for few-shot sentiment analysis [C]// Proceedings of the 2024 IEEE International Conference on Acoustics, Speech and Signal Processing, 2024: 10016–10020.
- [35] DEMSZKY D, MOVSHOVITZ-ATTIAS D, KO J, et al. GoEmotions: a

- dataset of fine-grained emotions [C]// Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020: 4040–4054.
- [36] LIU X, ZHENG Y, DU Z, et al. GPT understands, too[J]. AI Open, 2024, 5: 208–215.
- [37] PECHER B, SRBA I, BIELIKOVA M. On the effects of randomness on stability of learning with limited labelled data: a systematic literature review [DB/OL]. <https://arxiv.org/abs/2312.01082>.
- [38] LONG Q Y, WANG W Y, PAN S. Adapt in contexts: retrieval-augmented domain adaptation via in-context learning [C]// Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, 2023: 6525–6542.
- [39] ROUZEGAR H, MAKREHCHI M. Enhancing text classification through LLM-driven active learning and human annotation [C]// Proceedings of the 18th Linguistic Annotation Workshop, 2024: 98–111.
- [40] LI X N, QIU X P. Finding support examples for in-context learning [C]// Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, 2023: 6219–6235.
- [41] ZHAO T, WALLACE E, FENG S, et al. Calibrate before use: improving few-shot performance of language models [C]// Proceedings of the 38th International Conference on Machine Learning, 2021: 12697–12706.
- [42] ZHAO Y, NASUKAWA T, MURAOKA M, et al. A simple yet strong domain-agnostic de-bias method for zero-shot sentiment classification [C]// Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, 2023: 3923–3931.
- [43] ZHANG H L, ZHANG Y F, YU Y D, et al. A study on the calibration of in-context learning [C]// Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics, 2024: 6118–6136.
- [44] MIN S, LEWIS M, HAJISHIRZI H, et al. Noisy channel language model prompting for few-shot text classification [C]// Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, 2022: 5316–5330.
- [45] LIU H F, WANG Y. Towards informative few-shot prompt with maximum information gain for in-context learning [C]// Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, 2023: 15825–15838.
- [46] BROWN T, BENJAMIN M, RYDER N, et al. Language models are few-shot learners [C]// Proceedings of the 34th International Conference on Neural Information Processing Systems, 2020: 1877–1901.
- [47] ALY R, SHI X J, LIN K X, et al. Automated few-shot classification with instruction-finetuned language models [C]// Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, 2023: 2414–2432.
- [48] SANH V, WEBSON A, RAFFEL C, et al. Multitask prompted training enables zero-shot task generalization [C]// The 10th International Conference on Learning Representations, 2022: 1–216.
- [49] ZHANG T J, WANG X Z, ZHOU D, et al. TEMPERA: test-time prompt editing via reinforcement learning [C]// Proceedings of the 2023 International Conference on Learning Representations, 2023: 1–10.
- [50] JIANG Z T, ZHANG Y Z, LIU C, et al. Generative calibration for In-context Learning [C]// Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, 2023: 2312–2333.
- [51] YANG H R, ZHANG Y M, XU J Q, et al. Unveiling the generalization power of fine-tuned large language models [C]// Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics, 2024: 884–899.

(责任编辑:陈维捷)