刘结, 陈梅, 刘江越. 基于多语言模型词汇增强的低资源情感分析[J]. 智能计算机与应用,2024,14(12):82-89. DOI:10.20169/j.issn.2095-2163.24080202

基于多语言模型词汇增强的低资源情感分析

刘 结, 陈 梅, 刘江越
(乌鲁木齐职业大学, 乌鲁木齐 830002)

摘 要: 针对多语言情感分析任务中低资源语言模型词汇量稀少的问题,本文提出一种多语言模型词汇增强的框架。该框架基于齐普夫定律对低频但信息量丰富的词汇进行选择,以扩充低资源语言中的少见词。并结合加权熵对模型词汇表进行优化,以扩充与特定情感任务相关的特异词。然后利用多语言模型预训练和微调进行情感分类。实验结果表明,在印地语和印地语-英语混合语言任务上,提出的框架显著提升了低资源情感分析的性能。本方法不仅改善了低资源语言情感分析的性能,还提高了多语言情感分析的整体适应性。

关键词: 低资源语言; 情感分析; 词汇增强; 多语言模型
中图分类号: TP391.3 **文献标志码**: A **文章编号**: 2095-2163(2024)12-0082-08

Low-resource sentiment analysis based on multilingual model with vocabulary augmentation

LIU Jie, CHEN Mei, LIU Jiangyue
(Urumqi Vocational University, Urumqi 830002, China)

Abstract: Aiming at the limited vocabulary of low resource language model in multilingual sentiment analysis, a framework is proposed for vocabulary augmentation of multilingual model. This framework gains low-frequency but informative words based on Zipf's law to expand rare words in low-resource languages. Moreover, it optimizes the vocabulary with the weighted entropy to expand the specific words related to sentiment analysis. Then, the multilingual model is pretrained and finetuned for sentiment analysis. Experimental results show that this framework significantly improves the performance of low-resource sentiment analysis on both Hindi and Hindi-English mixed language. This framework not only improves the performance of low-resource language sentiment analysis, but also enhances the overall adaptability of multilingual sentiment analysis.

Key words: low resource; sentiment analysis; vocabulary augmentation; multilingual model

0 引 言

mBERT 或 XLM-R 等多语言模型(Multilingual Language Model, MLLM)[1],通过对大型多语言语料库的学习已具备丰富的语言知识。这些模型虽支持众多低资源语言(Low-Resource Language, LRL)情感分析任务[2],但低资源语言与英语等高资源语言(High-Resource Language, HRL)相比,在词汇数量级上仍存在着很大差异。表 1 展示了 mBERT 词汇字典中各类印度语言与英语、中文的词汇数量差异。这种差异可能导致低资源语言任务[3-4]遇到一系列挑战。

表 1 mBERT 词汇表中各种语言的统计

Table 1 Statistics of various languages in the vocabulary of mBERT

语言	词汇量	百分比/%
孟加拉语	946	0.79
印地语	1 852	1.55
坎那达语	653	0.55
泰米尔语	832	0.70
特拉古语	887	0.74
中文	13 542	11.32
英语	47 464	39.70

基金项目: 新疆维吾尔自治区社会科学基金项目(2023BGL076); 广东省高等教育学会研究课题(22GYB065); 中国科学院"西部之光"人才培养计划项目(2021-XBQNXZ-032)。

作者简介: 刘 结(1973—),女,副教授,主要研究方向:信息技术,职业技术教育。Email:liujie13319801502@163.com; 陈 梅(1979—),女,教授,主要研究方向:信息技术,物联网应用技术; 刘江越(1981—),男,副教授,主要研究方向:人工智能,职业技术教育。

收稿日期: 2024-08-02

首先，当 LRL 的单词不能被 MLLM 词表有效地分解为词片时，就可能会与未知标记混淆，影响模型的理解和生成能力。其次，尽管 MLLM 的细碎化词片可以组合成任何 LRL 词，从而避免了未知标记的直接出现，但这些词片可能会与 HRL 中的语义无关词汇产生混淆，使其在语境中的整合难以实现准确的 LRL 词嵌入。

虽然可以通过大量人力和计算资源投入来建立针对情感分析任务的大规模 LRL 语料库以扩充 MLLM 词汇，但这种做法往往代价昂贵。因此，在预训练阶段就提升对 LRL 的支持能力，并在微调时减少对大量特定语料的依赖，成为了优化多语言模型性能的关键探索方向。

本研究针对低资源情感分析任务，提出了基于齐普夫定律与加权熵扩充（Zipf's Law and Weighted Entropy-enhanced Expansion，ZLWEE）的多语言模型词汇增强方法。该方法结合了齐普夫定律进行低频词语的选择，并采用加权熵增强单词及其组成碎片的任务特异性，然后利用多语言模型进行训练和微调，实现低资源语言的情感分析。通过利用 ZLWEE 对低资源语言词汇的增强，加强了预训练模型对 LRL 的适应能力，显著提升了该模型在处理 LRL 情感分析任务时的效果。

1　相关工作

1.1　低资源语言情感分析

由于大语言模型对大规模训练数据的依赖，相关研究开始关注低资源环境下的情感分析[5-7]。低资源环境通常指的是训练数据不足的有监督情感分析任务[8]。相关研究往往需要设定一个阈值来判断当前目标任务是否属于低资源环境。根据任务类型的不同，低资源环境的阈值也有所不同。例如，Ke 等学者[9]认为产品评论领域情感分析任务的低资源环境阈值为 200 条标记的训练数据。当前低资源情感分析通常采用数据增强、迁移学习和微调等方法。其中，数据增强通过对训练数据进行各种转换，在保留标签的同时修改数据特性[10]。例如，Xie 等学者[11]通过回译进行数据增强，提升了模型在 IMDB 数据集上的情感分类性能。迁移学习通过传输学习到的表示减少了对标记数据的需求。对预训练语言模型进行微调[12]，能够充分利用预训练语言模型的知识，对于低资源任务非常有效。然而，微调整个预训练模型代价较高，导致高昂的计算时间和内存成本。

1.2　多语言模型词汇增强

多语言模型词汇增强包括非词汇扩充和词汇扩充方法。其中，非词汇扩充的方法致力于解决稀有或词汇表外（OOV）词汇所引发的问题[13]。例如，Purkayastha 等学者运用 UROMAN 工具实现 UTF-8 到拉丁文的转写，增强了多语言预训练模型（mPLMs）对多种低资源语言的兼容性[14]。Liu 等学者[15]在预训练及微调过程中加入了嵌入生成模块，旨在缩小词汇间的差异。这些方法在一定程度上缓解了问题，但未能根本解决词汇表内令牌的特定领域和语言的挑战。在词汇扩充的方法中，Poerner 等学者[16]通过整合领域特定词汇到预训练模型中，使模型更加紧密地贴合特定领域的需求。Chung 等学者[17]探讨了基于语言群体创建多语言词汇表的方法，为理解语言的多样性提供了见解。Nag 等学者[18]开发的基于熵的语言模型对词汇进行了扩充。这些方法表明词汇扩充能够显著提升模型性能，但是单词选择依赖于词频，未能充分考虑模型内令牌可能出现的表征偏差。

2　方法

2.1　方法概述

本研究对低资源语言进行词汇选择和扩充，以增强多语言模型的词汇表，并利用多语言模型进行情感分析，算法框架如图 1 所示。图 1 中，在左上角的虚线框中，基于齐普夫定律扩充的词汇、如词汇 8 和词汇 9 代表少见词，词汇 1 和词汇 2 表示其他常见低资源语言词汇。若少见词的平均位置超过一定的阈值，则将该词加入词汇表。左下角的虚线框中展示了基于加权熵的词汇扩充，综合了词汇在不同情感类别中的分布和情感类别的分布，熵值低于阈值的词汇被扩充入词汇表，表明该词汇与该特定情感分类任务高度相关。右侧展示了扩充词汇表后的分词情况。例如，句子 राम का अच्छापन उसे सभी का प्रिय बनाता है। 表达了正面情感，如果错误地将 अच्छापन 分词为 अच्छा 和 पन，可能会误导情感类别判定。然后利用多语言模型 mBERT 的预训练和微调进行情感分类。设类别的集合为 C，其中 $c \in C$ 表示其中的任一类别。

2.2　多语言模型词汇增强

词汇增强包括 2 个主要模块：基于齐普夫定律的少见词增强（Zipf Selection，ZS），和基于加权熵的任务特异性词汇增强（Weighted Entropy，WE）。通过分析词汇与少见词之间的关系，使用 ZS 选择词汇

并将其加入词汇表中,以对词汇表进行少见词的扩充。并使用 WE 评估词汇在情感分析任务中的加权熵,来选择词汇并将其加入词汇表中,以对词汇表进行任务特异性词汇的扩 0 充。

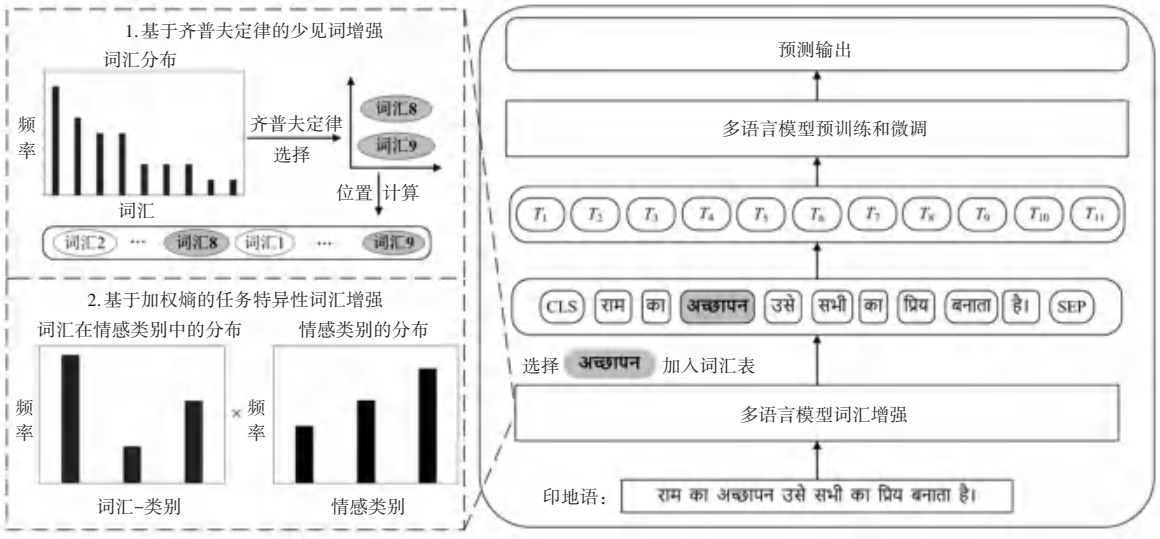


图 1 算法的整体框架

Fig. 1 Framework of the proposed algorithm

2.2.1 基于齐普夫定律的少见词增强

齐普夫定律(Zipf's Law)指出,在特定语料库中,词汇的出现频率与其排名成反比关系:排名靠前的词汇出现频率较高,而排名靠后的词汇出现频率较低。齐普夫定律的数学表达式如下:

$$f(w)=\frac{k}{rank(w)} \tag{1}$$

其中, $rank(w)$ 表示单词 w 在所有词汇中按出现频率降序排列的位置, k 是常数。通过识别频率低于某一预设阈值 ϕ 的词汇,可以揭示可能被忽视的低频词汇。

通常,带有情感极性的少见词倾向于出现在句子的起始或结束部分,由于这些词汇的低频特性,则可能未被包含在多语言模型的词汇表中。为了发现和情感极性有关的候选少见词,进一步计算每个词在句子中的相对位置百分比:

$$l(w)=\left(\frac{pos(w)}{len(s)}\right)\times 100 \tag{2}$$

其中, $pos(w)$ 表示单词 w 在句子中的位置, $len(s)$ 表示句子的总长度。每个单词的平均位置百分比的计算公式见如下:

$$\bar{l}(w)=\frac{\sum_{i=1}^{N} l_i(w)}{N} \tag{3}$$

其中, N 表示单词 w 出现的总次数。

基于上述分析,选择平均位置百分比在特定阈值范围内的词汇集合,这些阈值通过参数 α 和 β 定义,分别代表最高和最低的平均位置百分比。具体而言,选择平均位置百分比高于 α 和低于 β 的词汇作为候选少见词。这些词汇在句子结构和情感表达中通常发挥重要作用,尽管这些词汇却可能因出现频率较低而未被纳入现有的多语言模型词汇表中。

2.2.2 基于加权熵的任务特异性词汇增强

针对特定的情感分析任务,基于加权熵的词汇增强方法,评估低资源语言词汇对多语言词汇表扩充的有效性。为了避免词汇在分词时被错误拆分,直接将完整的低资源语言词汇加入词汇表。如果某个词汇 w 被分词器 T 拆分为词片序列 $s_1,\cdots,s_T$, 则需要考虑其在情感分析任务的不同类别中的出现频率。

给定情感分析任务的类别集合 C, 可以计算单词 w 和词片 s_t 在每个类别 $c\in C$ 中的归一化频率:

$$p(c\mid\cdot)=\frac{n(\cdot,c)}{\sum_{c'\in C} n(\cdot,c')} \tag{4}$$

其中, "·"可以是单词 w 或词片 s_t, $n(\cdot,c)$ 表示在类别 c 中单词"·"的出现次数。

然后,基于类别频率归一化计算每个类别的权重:

$$\alpha(c)=\frac{1/f(c)}{\sum_{c'} 1/f(c')} \tag{5}$$

其中, $\alpha(c)$ 表示每个类别的权重, $f(c)$ 表示每个类别中样本的数量。然后使用每个类别的权重定义加权熵 $H(\cdot)$:

$$H(\cdot) = -\sum_{c \in C} p(c \mid \cdot) \times \log(p(c \mid \cdot)) \times \alpha(c) \tag{6}$$

这个熵值不仅反映了词汇在不同类别中的分布情况，还考虑了不同类别的重要性。如果单词 w 的熵值很低，意味着该词与特定类别高度相关。相反，如果词片 s_t 的熵值很高，说明该词片在多个任务中被广泛使用，可能没有任务特异性。基于该加权熵可以更好地评估低资源语言词汇对多语言词汇表的影响，以及其在特定任务中的有效性。

2.3　基于多语言模型的情感分类

基于增强后的多语言词汇表，利用多语言模型 mBERT 的预训练和微调进行情感分类。mBERT 是基于 Transformer 的预训练语言模型，能够处理多语言文本。mBERT 的预训练主要包括以下几个部分：

（1）嵌入层：对于输入序列 X，其词嵌入表示为 E，位置嵌入表示为 P，则输入的嵌入表示为：

$$H_0 = E + P \tag{7}$$

（2）多层 Transformer 编码器：mBERT 模型包含多个堆叠的 Transformer 编码器层，每个编码器层由多头自注意力机制（multi-head self-attention mechanism）和前馈神经网络（feed-forward neural network）组成。这些编码器层能够捕捉输入序列中的长距离依赖关系，并生成上下文敏感的表示。对于每个编码器层 l，其输出表示为 H_l，其中每个编码器层的操作可以表示为：

$$H_l' = LayerNorm(H_{l-1} + MultiHeadAttention(H_{l-1})) \tag{8}$$

$$H_l = LayerNorm(H_l' + FeedForward(H_l')) \tag{9}$$

（3）池化层：在编码器层的顶部，mBERT 通常会应用池化操作来聚合序列中的信息，用[CLS]标记的表示作为整个序列的表示。

（4）输出层：对于情感分类任务，mBERT 的输出层通常是一个全连接层加上一个 Softmax 层，将[CLS]标记的表示映射为情感类别的概率分布。数学公式为：

$$y = \text{Softmax}(W_c h_{[\text{CLS}]} + b_c) \tag{10}$$

其中，$h_{[\text{CLS}]}$ 为[CLS]标记的表示；W_c 和 b_c 分别表示全连接层的权重和偏置；y 表示情感分类的预测结果。

在微调阶段，利用标注了情感类别的训练数据，预训练的 mBERT 模型被进一步训练以适应特定的情感分类任务。损失函数公式具体如下：

$$L_{CE} = -\sum_{i=1}^{N} \hat{y}_i \log y_i \tag{11}$$

其中，y_i 表示样本 i 的预测结果，$\hat{y}_i$ 表示实际类别。

3　实验与分析

3.1　数据集

在印地语数据集（IITP 产品评论分析）和印地语-英语混合数据集（GLUECos 情感分析）上进行了低资源情感分析实验，验证了 ZLWEE 扩充词汇表后的影响。低资源多语言数据集的详细信息见表 2。

表 2　低资源语言情感分析数据集的信息统计

Table 2　Statistics of low-resource sentiment analysis datasets

任务名称	LRL(s)	训练集	测试集
IITP 产品评论分析	Hindi	4 182	523
GLUECos 情感分析	Hindi-English 混合	10 079	1 260

3.2　实验设置

使用标准差为 0.02 的截断正态分布对多语言 BERT 基础模型的权重进行初始化，偏差设置为 0。在 NVIDIA A100 40 GB GPU 上开展实验，详细参数设置见表 3。为了加速模型在训练期间的收敛，使用 MLLM 词汇表中现有的低资源语言词片及其对应的英语翻译令牌来初始化新低资源语言令牌的嵌入。采用宏 $F1$ 和准确率评估低资源语言情感分析的性能。

表 3　mBERT 模型的超参数设置

Table 3　Parameter settings of mBERT

超参数	值
mBERTversion	bert-base-multilingual-cased
批次大小	16
Epoch	15
学习率	2×10^{-5}
最大序列长度	128
θ	1
γ	25
ϕ, α, β	20%

3.3　基准方法

（1）微调（Lees 等学者[19]）：不通过词汇表扩充来增强模型对语言的理解，而是依托已有词汇基础，对小型低资源语言任务进行参数微调，以便更好地适应情感分析。

（2）FLOTA（Hofmann 等学者[20]）：在分词过程中优先挑选可能的最长令牌，有效保持了词汇原有的形态结构，显著降低了因过度分词造成的信息流

失。此外,通过偏好长令牌,FLOTA 显著提升了分词过程对空白符噪声的对抗能力,有效减少了错误分词的发生。

(3)EVALM(Nag 等学者[18]):通过计算单词的熵值来评价词汇的信息含量和预测难度,以 LRL 词汇到词片的熵值增加作为判定词汇分解风险的依据,从而准确找出 LRL 中最易受到破坏的单词。词片的较高熵值表示在各个类别中的分布更加均衡,这可能会导致词汇的过度破碎化。

3.4 数据集碎片化情况分析

数据集碎片化是指词汇表不匹配导致单词被细分为更小的词片单位。该现象在表 4 中的词/词片比率中得到量化,将特定单词数量除以总词片数量得出该比率。较低的比率指示了较高的碎片化程度,意味着默认词汇表未能充分覆盖文本中的单词。表 4 中,IITP 产品评论数据集的 Hindi 语言内容表现出适中的碎片化水平,词/词片比率略低于 0.5。而 Hindi-English 混合的 GLUECos 情感分析数据集显示出稍高的碎片化水平,比率约为 0.6。这些比率反映了各数据集对原生词汇表的不同依赖程度。具体而言,GLUECos 情感分析数据集的碎片化程度较低,表明其对原生词汇表的依赖程度较高,即词汇表能够更好地覆盖该数据集的词汇。相对较低的比率则表明数据集的碎片化程度较高,如 IITP 产品评论数据集,暗示其对原生词汇表的依赖程度较低,需要更多的词汇表优化措施以提高处理效果。

表 4 数据集的碎片信息统计

Table 4 Statistics of word pieces in the datasets

数据集	词/词片比率			
	语言	训练集	验证集	测试集
IITP 产品评论分析	Hindi	0.50	0.49	0.50
GLUECos 情感分析	Hindi-English 混合	0.59	0.60	0.60

3.5 实验对比分析

将 ZLWEE 与 3 种基准方法在 2 个低资源情感分析任务上进行了仿真实验对比分析。对比实验结果见表 5。表 5 中,ZLWEE 在 2 个任务中的性能和平均性能都超越了所有基线,实现了最优异的性能。同时,表 5 中的实验结果揭示了各种方法在多语言任务上的性能差异,其中 ZLWEE 表现突出。ZLWEE 结合了 ZS 和基于频率加权熵的方法 WE,这两者共同优化了模型的词汇表。ZS 通过选择低频但信息量丰富的词汇,确保了模型对稀有且有意义的词汇有较好的覆盖。WE 则通过计算词汇的加权熵,进一步筛选和优化词汇表,使得模型在处理多样化语言环境时能够更有效地捕捉语义信息。这种方法的优越性在于其综合考虑了词汇的频率和信息量,从而在低资源语言环境中提供了更好的词汇覆盖和语义理解能力。实验对比分析可知,通过整合词汇表优化和加权熵分析可知,ZLWEE 不仅加强了模型在低资源语言上的性能,还提高了在多语言任务中的整体准确性和适应性。

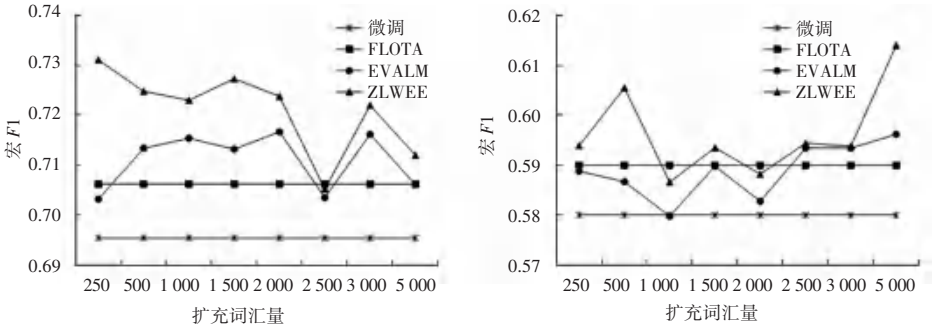
表 5 对比实验结果

Table 5 Performance comparison

	微调		FLOTA		EVALM		ZLWEE	
	准确率/%	宏 $F1$/%	准确率/%	宏 $F1$/%	准确率/%	宏 $F1$/%	准确率/%	宏 $F1$/%
IITP 产品评论分析(印地语)	72.21 (±0.39)	69.54 (±0.24)	74.19 (±0.90)	70.62 (±0.85)	75.21 (±0.46)	71.67 (±0.96)	**76.06 (±0.11)**	**73.11 (±0.87)**
GLUECoS 情感分析(印地语-英语混合)	59.74 (±0.60)	58.14 (±0.42)	60.87 (±0.66)	59.41 (±0.67)	61.30 (±1.09)	59.62 (±0.72)	**61.92 (±0.27)**	**61.40 (±0.68)**
平均	65.98 (±0.51)	63.84 (±0.34)	67.53 (±0.79)	65.02 (±0.77)	68.26 (±0.84)	65.65 (±0.85)	**68.99 (±0.21)**	**67.26 (±0.78)**

3.6 词汇表规模的影响

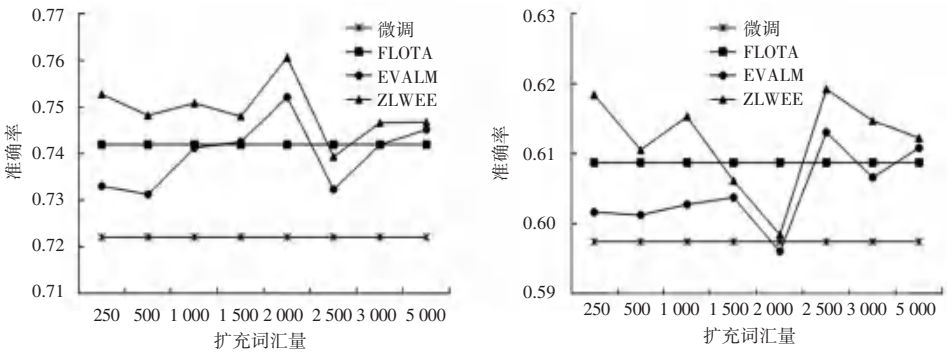
图 2 和图 3 展示了 2 个任务的宏 $F1$ 和准确率与词汇扩充量之间的相关性,确保词汇增加在规模上是可比的。微调和 FLOTA 的词汇表规模保持不变,因此微调和 FLOTA 不受规模增大的影响。EVALM 和 ZLWEE 随着词汇表扩充量的增加表现出相似的趋势,并且 ZLWEE 总是比 EVALM 表现优越。具体地,EVALM 和 ZLWEE 的性能与词汇表规模的扩展并未显示出明显的正相关性,这表明加入词汇表的扩充单词可能与情感分析任务的直接关联度不足,导致其性能波动。词汇量的增加引入一些不那么相关的词汇,暂时降低了模型对情感分析任务的适应性。随后,当词汇量继续增加,更多相关的词汇被加入,模型性能得以恢复并提升。这些发现突出了通过词汇增强策略显著提升模型性能的潜力,同时也揭示了词汇选择策略持续优化的必要性。



(a) IITP 产品评论分析　　(b) GLUECos 情感分析

图 2　宏 *F*1 与多语言模型词汇表扩充量的关系

Fig. 2　Relationship between macro-*F*1 and augmented vocabulary for MLLM



(a) IITP 产品评论分析　　(b) GLUECos 情感分析

图 3　准确率与多语言模型词汇表扩充量的关系

Fig. 3　Relationship between precision and augmented vocabulary for MLLM

3.7　案例研究

通过实际案例分析，展示了在不同语言数据集中，ZLWEE 方法对复杂语境情感识别的有效性。表 6 详细列出了从 2 个数据集中选取的几条评论及其情感倾向，所有条目均被 ZLWEE 正确识别。例如“आईस पड़ ने से आचा तो मै nfs most wanted खेल लेटा .. 3rd blacklist को हरना जो है _ ड”译文为“相比下雪，我更喜欢玩《NFS Most Wanted》.. 第三名黑名单要击败”。这句话的情感极性更倾向于中性，ZLWEE 把“खेल लेटा”这样的动词短语识别为一个整体，译为“玩”，不带有明显的情感极性。而 EVALM 把“खेल लेटा”分词为“खेल”和“लेटा”译为了“游戏”和“躺下”，从而翻译成“我更喜欢游戏躺下《NFS Most Wanted》”，从而可能导致情感极性被错误评估为了“积极”。从 GLUECoS 的印地语-英语混合到 IITP 产品评论中的纯印地语的文本，ZLWEE 均能准确识别不同情感倾向的评论。这种能力显示了其对复杂语境和多语言环境的高度适应性，特别是具有挑战性的混合语言和复杂脚本的语言。

表 6　不同方法在不同数据集上的情感识别案例

Table 6　Case study with different methods in different datasets

评论内容	真实标签	ZLWEE	EVALM	数据集来源
आईस पड़ ने से आचा तो मै nfs most wanted खेल लेटा 3rd blacklist को हरना जो है _ ड	中性的	中性的	积极的	GLUECoS 情感分析(印地语-英语混合)
फिल्म का स् टोरी mast ही और आप भी मस्त हो	积极的	积极的	中性的	GLUECoS 情感分析(印地语-英语混合)
[illegible]	积极的	积极的	积极的	IITP 产品评论分析(印地语)
ये दोनों टैबलेट एंड्रॉइड के 4.2.2 वर्जन ऑपरेटिंग सिस्टम पर काम करेंगे।	中性的	中性的	积极的	IITP 产品评论分析(印地语)
पॉपुला सिंगर अंकित तिवारी भी निराश करते है।	消极的	消极的	消极的	IITP 产品评论分析(印地语)

3.8 消融实验

通过消融实验探讨了 2 种关键技术 WE 和 ZS，见表 7 和表 8。在 IITP 产品评论(印地语)任务中，根据表 7 的数据，对比各个词汇表规模下应用 WE(ZLWEE)和未应用 WE(w/o WE)的性能差异，突显了 WE 在调整和优化词汇表以提升对特定任务敏感性方面的重要性。同时，对比各个词汇表规模下应用 ZS(ZLWEE)和未应用 ZS(w/o ZS)的性能差异，表明 ZS 对模型性能的提升具有积极作用。在 GLUECoS(印地语-英语混合)任务的消融实验中也观察到了类似的趋势。不仅展示了 WE 和 ZS 各自的重要性，还强调了两者在提升整体模型性能中的互补作用。ZLWEE 方法通过结合这 2 种技术，能够在不同的词汇表规模下优化性能，并根据具体任务动态调整词汇表，适应各种复杂的语言环境。

表 7 不同词汇表扩充规模下 IITP 产品评论分析(印地语)的消融实验

Table 7 Ablation study with different augmented vocabularies for IITP product review analysis (Hindi)

方法	指标	扩充规模							
		250	500	1 000	1 500	2 000	2 500	3 000	5 000
w/o ZS	准确率/%	73.30	73.12	74.12	74.25	75.21	73.23	74.19	74.51
	宏 *F*1/%	70.32	71.34	71.54	71.32	71.67	70.35	71.62	70.61
w/o WE	准确率/%	72.21	72.21	72.21	72.21	72.21	72.21	72.21	72.21
	宏 *F*1/%	69.54	69.54	69.54	69.54	69.54	69.54	69.54	69.54
ZLWEE	准确率/%	75.27	74.82	75.08	74.80	76.06	73.93	74.66	74.68
	宏 *F*1/%	73.11	72.48	72.30	72.73	72.38	70.52	72.20	71.20

表 8 不同词汇表扩充规模下 GLUECoS 情感分析(印地语-英语混合)的消融实验

Table 8 Ablation study with different augmented vocabularies for GLUECoS sentiment analysis (Hindi-English code-mix)

方法	指标	扩充规模							
		250	500	1 000	1 500	2 000	2 500	3 000	5 000
w/o ZS	准确率/%	60.16	60.12	60.27	60.37	59.60	61.30	60.66	61.08
	宏 *F*1/%	58.88	58.67	57.98	58.98	58.28	59.34	59.34	59.62
w/o WE	准确率/%	59.74	59.74	59.74	59.74	59.74	59.74	59.74	59.74
	宏 *F*1/%	58.14	58.14	58.14	58.14	58.14	58.14	58.14	58.14
ZLWEE	准确率/%	61.84	61.05	61.53	60.61	59.84	61.92	61.47	61.22
	宏 *F*1/%	59.39	60.55	58.66	59.35	58.82	59.44	59.38	61.40

4 结束语

针对多语言情感分析任务中低资源语言模型词汇量稀少的问题，提出了基于 ZS 和 WE 的 ZLWEE 方法，成功将低资源语言的相关词汇纳入到多语言模型词汇表中。ZS 通过选择低频但信息量丰富的词汇，扩充了低资源语言中的少见词，确保了模型对稀有且有意义词汇的较好覆盖。WE 通过计算词汇的加权熵，增强了词汇和特定情感任务的相关性，使得模型在处理多样化语言任务时能够更有效地捕捉语义信息。这种方法的优越性在于其综合考虑了词汇的频率和信息量，从而在低资源语言环境中提供了更好的词汇覆盖和语义理解能力。不仅加强了模型在低资源语言上的性能，还提高了在多语言任务中的整体准确性和适应性。但是存在性能波动，并且未考虑大模型的应用。因此在未来工作中，将考虑借助大模型强大的外部知识库和推理能力，提升低资源词汇和情感分析任务的语义相关性。

参考文献

[1] ALEXIS C, KARTIKAY K, NAMAN G, et al. Unsupervised cross-lingual representation learning at scale[C]// Proceedings of Annual Meeting of the Association for Computational Linguistics (ACL). ACL,2020: 8440-8451.

[2] 赵天锐. 基于深度学习的韩国语文本情感分类[J]. 智能计算机与应用, 2021, 11(5): 82-87.

[3] HEDDERICH M A, LANGE L, ADEL H, et al. A survey on recent approaches for natural language processing in low-resource scenarios[C]// Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT). ACL,2021: 2545-2568.

[4] 苏杭, 胡亚豪, 潘志松. 利用提示调优融合多种信息的低资源

事件抽取方法[J]. 计算机应用研究，2024，41(2)：381-400.
[5] CHEN Zhuang，QIAN Tieyun. Description and demonstration guided data augmentation for sequence tagging[J]. World Wide Web，2022，25(1)：175-194.
[6] DING Bosheng，LIU Linlin，BING Lidong，et al. DAGA：Data augmentation with a generation approach for low-resource tagging tasks[C]// Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing(EMNLP). ACL，2020：6045-6057.
[7] HSU T W，CHEN C C，HUANG H H，et al. Semantics-preserved data augmentation for aspect-based sentiment analysis[C]// Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing(EMNLP). ACL，2021：4417-4422.
[8] 张涵，王晶晶，罗佳敏，等. 针对低资源场景下连续情感分析任务的持续注意力建模[J/OL]. 软件学报.[2024-01-04]. https://link.cnki.net/urlid/11.2560.TP.20240103.1327.002.
[9] KE Zixuan，LIU Bing，MA Nianzu，et al. Achieving forgetting prevention and knowledge transfer in continual learning[C]// Proceedings of the Conference on Neural Information Processing Systems (NeurIPS). Montreal，Canada：dblp，2021：22443-22456.
[10] JASON W，ZOU Kai. EDA：Easy data augmentation techniques for boosting performance on text classification tasks[C]// Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing(EMNLP-IJCNLP). ACL，2019：6381-6387.
[11] XIE Qizhe，DAI Zihang，EDUARD H，et al. Unsupervised data augmentation for consistency training[J]. arXiv preprint arXiv，1904.12848，2019.
[12] 陈壮，钱铁云，李万理，等. 低资源方面级情感分析研究综述[J]. 计算机学报，2023，46(7)：1445-1472.
[13] ELENA S，LEONARDO R，MICHELE M，et al. Lacking the embedding of a word? Look it up into a traditional dictionary[C]// Findings of the Association for Computational Linguistics (ACL). ACL，2022：2651-2662.
[14] VALENTIN H，JANET B，HINRICH S. Superbizarre is not superb：Derivational morphology improves BERT's interpretation of complex words[J]. arXiv preprint arXiv，2101.00403，2021.
[15] LIU Xin，YANG Bangsong，LIU Dayiheng，et al. Bridging subword gaps in pretrain-finetune paradigm for natural language generation[J]. arXiv preprint arXiv，2106.06125，2021.
[16] POERNER N，ULLI W，HINRICH S. Inexpensive domain adaptation ofpretrained language models：Case studies on biomedical NER and Covid-19 QA[J]. arXiv preprint arXiv，2004.03354，2020.
[17] CHUNG H W，DAN G，KIAT C，et al. Improving multilingual models with language-clustered vocabularies[C]// Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing(EMNLP). ACL，2020：4536-4546.
[18] NAG A，BIDISHA S，ANIMESH M，et al. Entropy-guided vocabulary augmentation of multilingual language models for low-resource tasks[C]// Findings of the Association for Computational Linguistics(ACL). ACL，2023：8619-8629.
[19] LEES A，JEFFREY S，IAN K. Jigsaw@ AMI and HaSpeeDe2：Fine-tuning a pre-trained comment-domain BERT model[C]// Proceedings of the 7th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian(EVALITA). Torino：Accademia University Press，2020：40-47.
[20] HOFMANN V，HINRICH S，JANET B. An embarrassingly simple method to mitigate undesirable properties ofpretrained language model tokenizers[C]// Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics(ACL). ACL，2022：385-393.