基于标签感知注意力的短文本分类方法

李大帅,叶成荫

(辽宁石油化工大学 信息与控制工程学院,辽宁 抚顺 113001)

摘 要: 针对目前短文本分类只是将分类标签作为分类结果判断依据,而忽略了分类标签文本中所蕴含的语义信息这一问题,提出以大规模预训练语言模型为基础的基于标签感知注意力的短文本分类方法。该方法通过大规模预训练语言模型将文本数据表征为分布式向量形式以获得更丰富的语义信息;同时将分类标签信息融入到文本数据训练过程中,通过注意力机制使文本数据感知与分类最相关的信息;使用CNN网络和最大池化层提取局部词级向量特征,以更好地解决英文文本中的双重否定、比较级否定等语义问题;使用残差连接将句级向量与词级向量融合,以有效缓解文本信息衰减问题。在R8、R52和MR 3个公共英文数据集上进行测试,实验结果表明,所提方法在R8和R52数据集上的精度分别为98.51%和97.10%,优于DeBERTa和BertGCN。

关键词: 短文本分类;CNN;标签感知;注意力;预训练

DOI: 10. 11907/rjdk. 231951 **开放科学(资源服务)标识码(OSID):**

中图分类号: TP181 **文献标识码:** A **文章编号:** 1672-7800(2024)009-0110-06

Short Text Classification Method Based on Label Awareness Attention

LI Dashuai, YE Chengyin

(*School of Information and Control Engineering, Liaoning Petrochemical University, Fushun* 113001, *China*)

Abstract: Aiming at the problem that current short text classification only uses classification labels as the basis for judging classification results, and ignores the semantic information contained in the classified label text, a short text classification method based on label aware attention is proposed, which is based on a large-scale pre trained language model. This method represents text data in distributed vector form through large-scale pre trained language models to obtain richer semantic information; At the same time, incorporating classification label information into the text data training process, using attention mechanisms to make the text data perceive and classify the most relevant information; Using CNN networks and max pooling layers to extract local word level vector features, in order to better address semantic issues such as double negation and comparative negation in English texts; Using residual connections to fuse sentence level vectors with word level vectors effectively alleviates the problem of text information decay. Tests were conducted on three common English datasets, R8, R52, and MR, and the experimental results showed that the proposed method achieved accuracies of 98.51% and 97.10% on the R8 and R52 datasets, respectively, which are better than DeBERTa and BertGCN.

Key Words: short text classification; CNN; label awareness; attention; pre-trained

0 引言

在互联网飞速发展的时代,各种应用程序和网站中充斥着大量短文本数据,使人们应接不暇,对其进行分类有助于快速找到所需信息[1]。因此,如何有效对短文本进行分类,并高效地从中提取关键信息成为自然语言处理领域的重要任务。目前,短文本分类多用于搜索引擎[2]、情感分析[3]、问题解答[4]和新闻检索[5]等场景。近年来,注意力机制在短文本分类任务中被广泛应用并取得显著效果[6]。基于注意力机制的大规模预训练语言模型能有效解决静态词向量无法处理的短文本数据噪声和一词多义问题,使短文本分类效果得到一定提高[7]。如何在大规模预训练语言模型上进一步提升短文本分类效果是目前的研究热点。

收稿日期: 2023-09-05

扫描二维码阅读全文:

作者简介: 李大帅(1990-),男,辽宁石油化工大学信息与控制工程学院硕士研究生,研究方向为自然语言处理和深度学习;叶成荫(1977-),男,博士,辽宁石油化工大学信息与控制工程学院副教授、硕士生导师,研究方向为网络控制和决策、5G网络优化和机器学习。本文通讯作者:李大帅。

1　相关研究

1.1　文本向量表示

短文本数据具有稀疏性、高特征性、高噪声性和歧义性等特点。要应对以上问题，必须要解决短文本数据的表征问题[8]。早期多采用词袋模型（Bag of Word，BoW）对短文本数据进行表征，然而该模型将每个单词都看成独立的个体，忽略了文本之间的顺序和词语之间的共现关系[9]。Word2Vec 能有效表征文本的上下文语义关系，然而其是静态词向量，每个单词只有唯一的一个词向量与其对应，无法处理一词多义问题，且无法处理短文本中包含拼写错误的噪声数据[10]。BERT 等预训练语言模型可动态利用上下文语义产生词向量，有效解决了一词多义和噪声问题[11]。本文便是基于 BERT 和 RoBERTa 两种预训练模型进行短文本数据的向量表征[12]。

1.2　基于神经网络的文本分类

循环神经网络（Recurrent Neural Network，RNN）由于考虑到输入数据的前后时序关系而被广泛应用于文本分类任务中。长短期记忆网络（Long Short-term Memory，LSTM）通过遗忘门、更新门和输出门缓解 RNN 的长距离依赖问题。LSTM 可以很好地利用时序信息对文本进行分类，但捕捉局部信息的能力不足[13]。卷积神经网络（Convolutional Neural Network，CNN）多用于图像处理任务，但由于其易于提取局部特征，在文本分类也有广泛应用。例如，TextCNN 是 CNN 在文本分类领域中的具体应用，其使用多个过滤层提取特征，结构简单，但固定了过滤层感受野的大小，对于较长序列无法建模，并且调参困难[14]；TextRCNN 将 CNN 与 RNN 相结合，既利用 CNN 提取特征的能力，又发挥 RNN 上下文依赖的特性，但对于过长的文本仍存在梯度爆炸或消失问题[15]。注意力机制（Attention）对特征进行注意力加权求和，缓解了梯度消失问题，增强了模型的可解释性。Zhou 等[16]将注意力机制与 LSTM 相结合，既能获取文本时序信息，又能缓解长距离依赖问题。

1.3　基于标签的文本分类

在文本分类任务中，通常直接将训练数据输入模型中进行文本训练，然后经过全连接层和 Softmax 层进行分类，这往往忽略了分类标签数据携带的重要语义信息[17]。为了更好地利用分类标签信息，一些基于分类标签的文本分类方法被提出[18-19]。例如，LEAM（Label-Embedding Attentive Model）方法是一种在文本分类任务中表现出色的深度学习模型，其使用 GloVe 静态词向量作为文本表示特征[20]，定义一个标签嵌入层表示标签分类，但在处理噪声数据和歧义数据方面会受到很大限制。在多标签多分类领域，LSAN（Label Specific Attention Network）模型使用双向长短时记忆网络（Bidirectional Long Short-Term Memory，Bi-LSTM）以更好地获得文本的上下文依赖信息，使用两个注意力机制进行特征提取并动态进行加权，但是需要两个注意力机制都有好的表现才能有显著效果[21]；LguidedLearn 模型使用预训练模型获取文本词向量，利用 Bi-LSTM 和相似度进行分类，更关注时序信息，但局部信息捕获不足；SLDC（Simulated Label Distribution Method Based on Concepts）是一种基于图神经网络的标签概念分类方法，利用概念图将扩充的标签概念生成图神经网络，然而概念图的质量对模型性能影响较大[22]。在文本分类应用研究方面，Chen 等[23]在电商产品分类场景中应用标签指导方法，使用高概率空间嵌入层次结构的产品标签，取得了良好的分类效果。

为了充分利用分类标签文本语义信息，本文提出基于标签感知注意力的短文本分类方法（Label Awareness Attention Method，LAAM）。该方法首先通过大规模预训练语言模型将文本数据表征为分布式向量形式；其次将分类标签信息融入到文本数据训练过程中，通过注意力机制使文本数据感知与分类最相关的信息；再次使用 CNN 网络提取词级向量特征，充分融合单词局部上下文信息以更好地解决英文双重否定和比较级否定等问题；最后通过残差连接结合句级向量和词级向量克服信息衰减问题。在 3 个公共英文数据集上进行实验，实验结果证实了本文方法的有效性，尤其在 R8 和 R52 数据集上的精度分别为 98.51% 和 97.10%，比目前已知最好结果，即 DeBERTa[24] 在 R8 的 98.45% 和 BertGCN[25] 在 R52 的 96.6% 分别高出 0.06% 和 0.5%。

2　基于注意力的标签感知方法

2.1　基于预训练的文本特征表示层

本文方法基于 BERT 等大规模预训练语言模型进行设计，充分利用动态词向量根据上下文动态生成文本表示的特性。BERT 模型使用 Transformer 的 Encoder 部分进行堆叠构建，使用多头自注意力机制，采用掩码语言模型（Masked Language Models，MLM）进行深度双向表示[26]。RoBERTa 是 BERT 模型的变体，相较 BERT 模型训练时间更长、批次更大、训练数据更多。RoBERTa 没有使用 BERT 的静态掩码技术，而是采用动态掩码技术，并且去掉了 BERT 中的 NSP（Next Sentence Prediction）任务，性能更佳。

本文采用 BEET 和 RoBERTa 两种预训练语言模型对短文本数据和标签文本数据进行特征表示。如图 1 所示，将短文本数据和标签文本数据分别输入到 BERT 模型中获取向量表示。二者在预处理过程中进行不同操作，其中短文本数据中保留 BERT 模型的[CLS]标记，因为[CLS]标记可以代表句子级别向量，是对输入文本的一个整体表示；而标签文本数据输入到 BERT 模型中会去掉[CLS]标记信息，因为标签之间相互独立并不构成句子关系，无需使用[CLS]标记进行整句语义的提取。本文直接使用[SEP]标

识对标签进行分割，以使BERT模型更好地学习标签的边界信息。使用共享权重对短文本数据和标签文本数据进行训练。

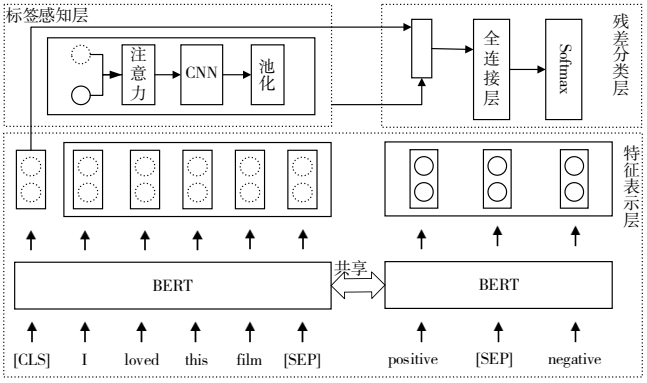

Fig. 1 Overall architecture of the model

图 1 模型整体架构

BERT模型会给输入的文本数据添加位置编码信息，主要用于解决使用多头自注意力机制导致的文本前后位置顺序信息缺失问题。为了避免模型记住标签文本数据的位置信息而不是根据内容进行正确分类，本文对每个批次输入的标签数据进行随机拼接，以减少固定编码位置对模型训练的影响。例如，对于二分类标签文本positivie和negative，每个批次的输入标签文本有[positive，[SEP]，negative]或[negative，[SEP]，positive]两种随机选择。

2.2 基于注意力的标签感知层

为了充分利用分类标签语义信息，本文对分类标签信息直接进行编码向量表示，然后与训练文本向量进行注意力分数计算，进而对训练文本向量进行权重加权，以更有针对性地提取关键信息。将加权后的文本向量经过CNN网络进行局部信息融合，以避免过度关注。标签感知层结构如图2所示。首先从BERT模型输出中获取训练文本词向量和分类标签词向量；然后对训练文本向量与标签文本向量进行点积注意力计算。本文采用的注意力计算公式为：

$$G = \frac{X^T V}{\sqrt{d}} \tag{1}$$

$$S = \max(G) \tag{2}$$

$$\beta = \frac{\exp(S)}{\sum \exp(S)} \tag{3}$$

$$Q = \beta X \tag{4}$$

式中：X为文本特征向量，V为标签特征向量，d为调整数值参数，G为注意力向量，S为注意力计算后的最关注向量，β为注意力分数，Q为经过注意力打分后的文本特征向量。在计算训练文本向量与标签文本向量注意力分数的过程中，采用最大值函数进行向量提取，进而获取到与文本特征最相关的标签向量；然后使用Softmax函数进行权重分数计算获取到权重值，再根据权重值对原训练文本向量进行加权处理，以提升重要文本的权重比例。

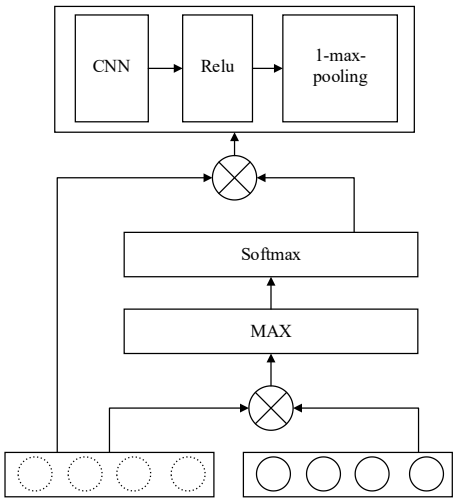

Fig. 2 label awareness layer

图 2 标签感知层

CNN在自然语言处理中的应用类似于N-gram语法，能够充分融合当前单词与前后单词的语义信息。在计算注意力分数时，本文选取结果最大的向量作为目标向量，但是对于双重否定、比较级否定等语句容易造成错误判定。例如，在英语中，“The movie is not bad”这种双重否定表示肯定的句子很常见，模型通过注意力机制可能只关注到了not或bad单词，然而只关注一个单词不能对整个句子意思作出正确判断，需要两个单词同时关注才能正确判断是消极情绪还是积极情绪。因此，需要通过CNN网络对整个句子信息的特征进行局部融合提取，以便更好地对句子进行分类。本文CNN采用单层卷积网络，卷积核大小设置为5，通过Relu激活函数和最大池化层完成局部特征提取。

2.3 基于残差连接的文本分类层

在进行短文本分类前首先进行词级向量与句级向量的融合，通过残差连接进行特征处理；然后经过全连接层进行特征降维，采用Softmax函数进行分类。本文通过注意力机制获取到的为词级向量，BERT模型第一个[CLS]标识可以表示句级向量。在标签感知层处理之后，将词级向量与句级向量进行残差连接，既可以整合两种向量包含的有效信息，又可以防止由于信息衰减而无法获得预期效果。

在获取到特征融合的向量数据后，首先要经过一个全连接层，将特征维度降维到与分类大小维度相同；然后经过Softmax函数进行分类，并使用交叉熵损失函数计算函数损失。采用均值方式计算每个批次的平均损失。公式为：

$$x = w^T v_i + b_i \tag{5}$$

$$l_n = -\sum_{c=1}^{c} w_c \log\left(\frac{\exp(x_{n,c})}{\sum_{i=1}^{c}\exp(x_{i,c})}\right) y_{n,c} \tag{6}$$

$$l(x,y) = \frac{\sum_{n=1}^{N} l_n}{N} \tag{7}$$

式中：i表示词向量分量；v_i表示经过特征融合后的向量；w表示权重参数；b_i表示偏置参数；c表示分类数量；N表示批次数量，$n \in N$；y表示真实标签分类信息。

3　实验方法与结果分析

3.1　实验环境

硬件环境：GPU为Tesla P40，显存为24 GB，内存为60 GB，CPU为12核Intel（R）Xeon（R）CPU E5-2650 v4 @ 2.20GHz。软件环境：开发语言为Python（3.8版本），深度学习框架为Pytorch（1.10版本），开发工具使用PyCharm社区版进行代码编写。模型超参数设置如表1所示。

Table 1　Model super parameter setting

表1　模型超参数设置

Parameter	Value
学习率	2e-5
批次大小	16
Dropout	0.1
优化器	AdamW
权重衰减	1e-2
梯度修剪	10

3.2　数据集

选择R8、R52、MR 3个英文公共数据集进行实验。其中，R8是一个包含8种分类的主题数据集，R52是一个包含52种主题的分类数据集，二者均为Reuters数据集的子集，平均长度相近，训练集和测试集比例也大致相同，但R52的分类更加精细，分类难度也更大。MR是一个二分类的电影评论情感分析数据集，每条评论只有一个句子，分类结果为积极或消极。各数据集基本信息如表2所示。可以看出，MR数据集中平均每条句子的长度相较R8和R52更短，包含的有效分类信息更少。

Table 2　Basic information of data set

表2　数据集基本信息

数据集	训练集长度	测试集长度	类别	平均长度
R8	5 485	2 189	8	102.37
R52	6 532	2 568	52	109.56
MR	7 108	3 554	2	20.39

3.3　数据处理

本文方法基于BERT系列模型进行文本表征，因此首先将数据处理成该类模型可以接受的输入格式。在文本数据开头添加[CLS]标签以表示整个句子的分类信息，在句子结尾添加[SEP]标签代表结束信息。每个批次的文本数据会根据批量大小不同程度地填充[PAD]标志，目的是统一数据长度。将所有类别数据进行拼接，拼接标签使用[SEP]标志。分类数据主要用于后期计算注意力分数，因此没有必要在开头和结尾添加[CLS]和[SEP]。BERT和RoBERTa使用的分割符不同，RoBERTa使用</s>作为分割符。不论是BERT还是RoBERTa，都会对数据顺序进行学习表征。为减少顺序标注对结果的影响，每个批量数据都会对分类数据进行随机拼接，以克服模型记住顺序而造成的预测偏差。

3.4　比较算法

采用以下模型与本文模型进行比较：①Bi-LSTM。一个双向的LSTM，学习前向和后向两个方向的文本语言特征；②BERT[11]。基于Transformer中Encoder的大规模预训练语言模型；③RoBERTa[12]。BERT模型的变体，采用更大规模数据进行训练，去掉BERT模型的句子相似度匹配；④LSTM[13]。一个RNN的变体，用于处理时间序列，缓解RNN遗忘问题；⑤CNN[14]。使用多个卷积核的神经网络进行文本分类，类似于N-Gram语法，抽取局部上下文信息；⑥LEAM[18]。已知第一个基于标签的文本分类算法，使用自定义变量进行标签表示，通过注意力机制进行文本分类；⑦LguidedLearn[19]。基于预训练模型进行训练，使用Bi-LSTM进行特征提取，然后计算标签相似度进行文本分类；⑧SLDC[22]。通过图神经网络将概念库引入模型训练过程中，在预训练模型下使用Bi-LSTM网络训练文本，并计算与标签的相似度进行文本分类；⑨Bert-GCN[25]。结合了大规模预训练模型与图卷积网络的文本分类模型，RoBERTaGCN是在RoBERTa模型下进行训练，其他与BertGCN相同；⑩BertGAT[25]。结合图注意力机制和大规模预训练模型，实现对不同邻居权值的自动，RoBERTaGAT是在RoBERTa下进行训练，其他与BertGAT相同；⑪Fasttext[27]。一个快速文本分类方法，不仅可以进行文本分类，还可以训练词向量，网络模型简单，由Facebook提出；⑫DeBERTa是由微软提出的自然语言处理模型，使用两种新技术改进了BERT和RoBERTa模型，同时还引入虚拟对抗训练方法以提高模型的泛化能力。

3.5　评价指标

使用准确率作为评价指标，公式如下：

$$accuracy = \frac{K}{N} \tag{8}$$

式中：K为预测正确的样本个数，N为预测样本总数。

3.6　实验结果

各模型在R8、R52、MR 3个英文公共数据集上的准确率结果如表3所示，其中带有*的表示本文实验得出的数据，其余数据均来自于文献；加粗数据表示最好结果。可以看出，本文模型在两个数据集都能达到最好结果，在MR数据集上接近最好结果。与Fasttext、CNN和LSTM等深度学习模型相比，本文模型只使用BERT或RoBERTa进行微调便能获得更好的分类性能；BertGCN和BertGAT等基于图神经网络的分类模型虽然能使用全局上下文信息，但对于局部重要信息的捕捉能力不足，而本文模型使用文本标签集中捕获关键信息，分类效果更显著；LEAM、SLDC和Lguided-BERT等基于标签的方法采用自定义向量代表标签分类情况，没有真正将标签引入到训练过程中，而本文模型将标签文本与数据一起训练，有效提升了短文本分类效果。

Table 3 Comparison of accuracy of various models on three datasets

表 3 各模型在 3 个数据集上的准确率比较 %

模型	R8	R52	MR
Fasttext	96.13	92.81	75.14
CNN	94.02	85.37	74.98
LSTM	93.68	85.54	75.06
Bi-LSTM	96.31	90.54	77.68
BERT	97.80	96.40	85.70
RoBERTa	97.80	96.20	89.40
BertGCN	98.10	96.60	86.00
RoBERTaGCN	98.20	96.10	89.70
BertGAT	97.80	96.50	86.50
RoBERTaGAT	98.00	96.10	89.20
SLDC	97.21	94.72	82.16
LEAM	93.31	91.84	76.95
Lguided-BERT-1	97.49	94.26	81.03
Lguided-BERT-3	98.28	94.32	81.06
DeBERTa	98.45	–	**90.21**
BERT-LAAM	98.25 *	96.92 *	87.03 *
RoBERTa-LAAM	**98.51** *	**97.10** *	89.31 *

本文模型对 MR 数据集分类精度的提升效果优于其他两个数据集，说明其不仅具有良好的复杂标签分类能力，而且在简单标签分类方面也更加优秀。此外，本文模型在 RoBERTa 预训练模型下表现更好，主要是由于 RoBERTa 训练数据量更大，时间更长，对下游数据的增强更显著。

3.7 消融实验

为验证本文模型各模块的有效性，在 R8 数据集上进行消融实验，预训练模型选择 BERT。结果如表 4 所示，表中 RC 代表残差连接。可以看出，在只使用 BERT 和标签注意力的情况下，模型准确率提升了 0.37%；加入 CNN 网络后准确率由 98.17% 上升至 98.22%，再加入残差连接后准确率进一步提升 0.03%。实验结果表明，各模块对准确率均有提升效果，其中标签注意力的提升效果更为显著。

Table 4 Ablation experiment result

表 4 消融实验结果 %

模型	准确率
BERT	97.80
LAAM w/o (CNN + RC)	98.17
LAAM w/o RC	98.22
LAAM	98.25

3.8 结果可视化

在 R8 数据集上对分类结果进行可视化，以更加直观的方式展示本文模型分类效果与其他模型的不同。采用 T 分布随机近邻嵌入(T-Distribution Stochastic Neighbour Embedding，T-SNE)方法对高维隐藏层进行降维处理，以便使用二维图像进行展示。图 3 为使用 CNN 训练后的可视化效果(彩图扫 OSID 码可见，下同)；图 4 为采用 RoBERTa 模型直接进行降维的效果；图 5 为本文模型在 RoBERTa 下的训练效果。可以看出，CNN 和 RoBERTa 模型降维后的数据分类比较分散，而本文模型经过降维后的同类别数据集中在一起。

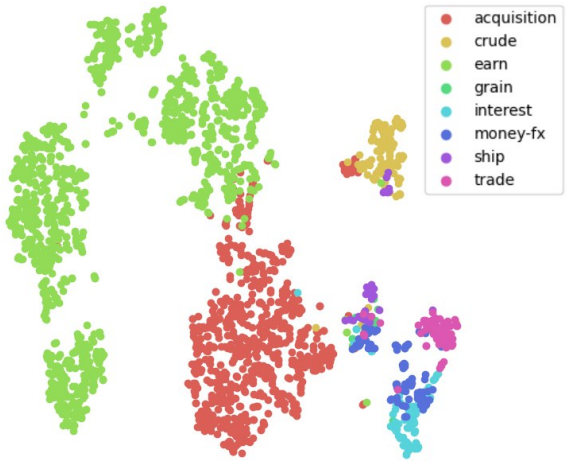

Fig. 3 CNN adopting T-SNE to reduce dimensions

图 3 CNN 采用 T-SNE 降维

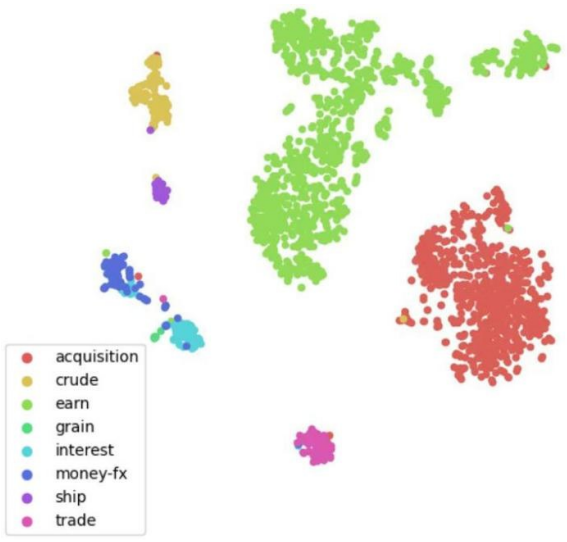

Fig. 4 RoBERTa adopting T-SNE to reduce dimensions

图 4 RoBERTa 采用 T-SNE 降维

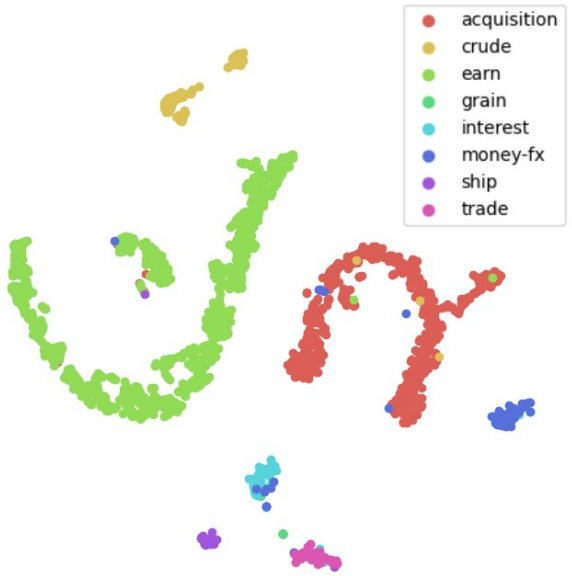

Fig. 5 RoBERTa-LAAM adopting T-SNE to reduce dimensions

图 5 RoBERTa-LAAM 采用 T-SNE 降维

4　结语

本文针对短文本分类问题中稀疏性与歧义性问题,提出基于大规模预训练模型的注意力的标签感知方法。该方法利用BERT和RoBERTa的动态性和上下文相关性解决歧义性问题,利用注意力的标签感知方法在数据有限的情况下快速提取重要信息,再经过CNN融合词的上下文信息解决双重否定等问题,最后将词级向量与句级向量相结合防止信息衰减。与同样基于标签的方法和一些基线方法相比,本文方法的分类效果更佳。然而目前引入的标签数据有限,未来可尝试通过同义词、概念库等方式扩充分类标签数量进而提升短文本分类效果。

参考文献:

[1] GAN Y T, AN J Y, XU X. Survey of short text classification methods based on deep learning[J]. Computer Engineering and Applications, 2023, 59(4):1-13.
淦亚婷, 安建业, 徐雪. 基于深度学习的短文本分类方法研究综述[J]. 计算机工程与应用, 2023, 59(4): 1-13.

[2] LI X W. Chinese language and literature online resource classification algorithm based on improved SVM[J]. Scientific Programming, 2022, 2022: 1-7.

[3] BRAUWERS G, FRASINCAR F. A survey on aspect-based sentiment classification[J]. ACM Computing Surveys, 2021,55(4):1-37.

[4] ZHANG S, ZHANG X. Does QA-based intermediate training help fine-tuning language models for text classification[C]//Proceedings of Annual Workshop of the Australasian Language Technology Association, 2021: 158-162.

[5] ZHANG H F, ZENG C, PAN L, et al. News topic text classification method based on BERT and feature projection network[J]. Journal of Computer Applications, 2022, 42(4): 1116-1124.
张海丰, 曾诚, 潘列, 等. 结合BERT和特征投影网络的新闻主题文本分类方法[J]. 计算机应用, 2022, 42(4): 1116-1124.

[6] CHEN L C, QIN J, LU W D, et al. Short text classification method based on self-attention mechanism [J]. Computer Engineering and Design, 2022, 43(3):728-734.
陈立潮, 秦杰, 陆望东,等. 自注意力机制的短文本分类方法[J]. 计算机工程与设计, 2022,43(3):728-734.

[7] BI W, GAO J, LIU X J, et al. Fine-grained sentence functions for short-text conversation [C]//Proceedings of Annual Meeting of the Association for Computational Linguistics,2019: 3984-3993.

[8] TAN Y Y, WANG J L, ZHANG C B. Review of text classification methods based on graph convolutional network[J]. Computer Science, 2022, 49(8): 205-216.
檀莹莹, 王俊丽, 张超波. 基于图卷积神经网络的文本分类方法研究综述[J]. 计算机科学, 2022, 49(8): 205-216.

[9] ZHANG Y, JIN R, ZHOU Z H. Understanding bag-of-words model: a statistical framework[J]. International Journal of Machine Learning and Cybernetics, 2010, 1(1-4): 43-52.

[10] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient estimation of word representations in vector space [DB/OL]. https://arxiv. org/abs/1301. 3781.

[11] DEVLIN J, CHANG M W, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding[C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics, 2019: 4171-4186.

[12] LIU Y, OTT M, GOYAL N, et al. RoBERTa: a robustly optimized BERT pretraining approach [DB/OL]. https://arxiv. org/pdf/1907. 11692. pdf.

[13] LIU P F, QIU X P, HUANG X J. Recurrent neural network for text classification with multi-task learning [C]//Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, 2016: 2873-2879.

[14] KIM Y. Convolutional neural networks for sentence classification[C]// Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, 2016: 2873-2879.

[15] LAI S W, XU L H, LIU K, et al. Recurrent convolutional neural networks for text classification[C]// Proceedings of the 29th AAAI Conference on Artificial Intelligence, 2015: 2267-2273.

[16] ZHOU P, SHI W, TIAN J, et al. Attention-based bidirectional long short-term memory networks for relation classification[C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, 2016: 207-212.

[17] ZHONG G F, PANG X W, SUI D. Text classification method based on Word2Vec and AlexNet-2 with improved attention mechanism[J]. Computer Science, 2022, 49(4): 288-293.
钟桂凤, 庞雄文, 隋栋. 基于Word2Vec和改进注意力机制AlexNet-2的文本分类方法[J]. 计算机科学, 2022, 49(4): 288-293.

[18] WANG G, LI C, WANG W, et al. Joint embedding of words and labels for text classification[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, 2018: 2321-2331.

[19] LIU X, WANG S, ZHANG X, et al. Label-guided learning for text classification[DB/OL]. https://arxiv. org/pdf/2002. 10772. pdf.

[20] PENNINGTON J, SOCHER R, MANNING C. Glove: global vectors for word representation[C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, 2014: 1532-1543.

[21] XIAO L, HUANG X, CHEN B, et al. Label-specific document representation for multi-label text classification[C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing , 2019: 466-475.

[22] LI H, HUANG G, LI Y, et al. Concept-based label distribution learning for text classification[J]. International Journal of Computational Intelligence Systems, 2022, 15(1): 85.

[23] CHEN L, MIYAKE H. Label-guided learning for item categorization in e-commerce[C]//Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics, 2021: 296-303.

[24] KARL F, SCHERP A. Transformers are short text classifiers: a study of inductive short text classifiers on benchmarks and real-world datasets [DB/OL]. https://arxiv. org/pdf/2211. 16878. pdf.

[25] LIN Y X, MENG Y X, SUN X F, et al. BertGCN: transductive text classification by combining GNN and BERT [C]//Findings of the Association for Computational Linguistics, 2021: 1456-1462.

[26] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need [C]// Proceedings of International Conference on Neural Information Processing Systems, 2017: 6000-6010.

[27] BOJANOWSKI P, GRAVE E, JOULIN A, et al. Enriching word vectors with subword Information[J]. Transactions of the Association for Computational Linguistics, 2017, 5: 135-146.

(责任编辑:尹晨茹)