doi:10.11835/j.issn.1000-582X.2024.12.010

一种基于半监督的句子情感分类模型

苏　静,Murtadha Ahmed
(西北工业大学 计算机学院,西安 710072)

摘要:句子情感分类致力于挖掘文本中的情感语义,以基于BERT(bidirectional encoder representations from transformers)的深度网络模型表现最佳。这类模型的性能极度依赖大量高质量标注数据,而现实中标注样本往往比较稀缺,导致深度神经网络(deep neural network,DNN)容易在小规模样本集上过拟合,难以准确捕捉句子的隐含情感特征。尽管现有的半监督模型有效利用了未标注样本特征,但对引入未标注样本可能导致错误逐渐累积问题没有有效处理。半监督模型在对测试数据集进行预测后不会重新评估和修正上次的标注结果,无法充分挖掘测试数据的特征信息。研究提出一种新型的半监督句子情感分类模型。该模型首先提出基于K-近邻算法的权重机制,为置信度高的样本分配较高权重,尽可能减少错误信息在模型训练中的传播。接着,采用两阶段训练策略,使模型能对测试数据中预测错误的样本进行及时修正,通过多个数据集的测试,证明本模型在小规模样本集上也能获得良好性能。

关键词:句子情感分类;半监督学习;K-近邻;transformer
中图分类号:TP311　　**文献标志码**:A　　**文章编号**:1000-582X(2024)12-100-14

A semi-supervised model for sentence-level sentiment classification

SU Jing, MURTADHA Ahmed
(School of Computer Science, Northwestern Polytechnical University, Xi'an 710072, P. R. China)

Abstract: Sentence sentiment classification is an important task for extracting emotional semantics from text. Currently, the best tools for sentence sentiment classification leverage deep neural networks, particularly BERT-based models. However, these models require large, high-quality labeled datasets to perform effectively. In practice, labeled data is usually limited, leading to overfitting on small datasets and difficulties in capturing subtle sentiment features. Although existing semi-supervised models utilize features from large unlabeled datasets, they still face challenges from errors introduced by pseudo-labeled samples. Additionally, once test data is labeled, these models often do not adapt by incorporating feature information from test data. To address these issues, this paper proposes a semi-supervised sentence sentiment classification model. First, a K-nearest neighbors-based weighting mechanism is designed, assigning higher weights to high confidence samples to minimize error propagation during parameter learning. Second, a two-stage training mechanism is implemented, enabling the model to correct misclassified samples in the test data. Extensive experiments on multiple datasets show that this method achieves strong performance on small datasets.

Keywords: sentence-level sentiment classification; semi-supervised learning; K-nearest neighbors; transformer

收稿日期:2023-12-11　　**网络出版日期**:2024-04-24
基金项目:国家自然科学基金资助项目(62172335)。
Supported by the National Natural Science Foundation of China (62172335).
作者简介:苏静(1987—),女,博士,主要从事自然语言处理和人工智能方向研究,(E-mail)sujing@mail.nwpu.edu.cn。

句子级情感分类任务主要对整个句子的情感趋向进行分析,常见的如电商网站上对商品的评价、投资平台上金融机构对股市风险评论、社交媒体中对热点事件和政策的评价等。对这些评价内容挖掘情感信息蕴含着巨大的商业价值,如企业可以利用这些信息来研发新产品或优化服务;金融机构可以据此进行投资预警;政府可以根据这类信息来制定或调整政策。然而,大规模高质量带标签的句子在实际应用场景中很难获取,因为往往需要耗费巨大的人力和时间成本对其进行标注。因此,这篇文章主要针对在只有少量可用带标签样本的场景下进行情感分类学习。众所周知,在句子情感分类任务中,目前最好的模型是基于transformer架构的深度神经网络模型(如BERT[1],RoBERTa[2],XLNet[3]等),通过在大规模wikipedia文档数据集上进行训练捕获自然语言中的语义知识,生成1个预训练模型,很好地应用于特定下游任务。但这些深度网络模型依赖训练样本集的大小和标注质量,当训练集较小时,容易出现过拟合现象,难以捕捉文本中的隐含情感特征,无法学习到多样化的情感特征。虽然标注数据难以获得,但未标注数据非常丰富且易于获取,不需要支付大量的人力和物力成本。本文旨在充分利用大量未标注数据,提高小样本的学习性能。虽然传统的半监督学习模型,如基于教师—学生模型的自训练和结合已标注与未标注数据的协同训练,尝试利用未标注数据解决小样本问题,这些方法都致力于挖掘未标注数据的特征,同时减少未标注数据训练引入的噪声。尽管后续研究提出了一些优化策略,如只选择一部分高置信度的伪标签样本参与训练,但这些方法通常基于深度神经网络(DNN)预测概率的信息熵来计算置信度,仍然存在累积错误的风险,因为DNN对预测错误的样本也可能给出高概率预测,不能准确反映真实的预测置信度。因此,笔者提出了一种基于K-近邻的损失加权机制。该机制在模型训练过程中,对每个样本实例,找到其在同批次中距离最近的K个样本。通过比较这些近邻的预测标签与当前样本的预测标签,计算相同标签的数量比例,作为该样本预测正确的权重。这个权重随后用于加权散度损失,参与训练和模型参数的优化。通过给予高置信度样本较高权重,低置信度样本较低权重,有效降低噪声的影响,通过在损失函数中设置权重,直接影响模型的学习过程。此外,现有的半监督学习模型主要关注如何充分学习和利用未标注数据的特征,但在处理测试数据集时,一旦为测试数据分配了标签,就不再考虑对这些预测标签进行修改。这些模型通常未能充分利用测试数据中的特征信息,仅将测试数据作为评估模型准确率的工具。本文提出一种新方法,旨在通过学习测试数据集上的特征信息来修正测试数据上已有的标注标签。不仅关注如何利用未标注数据集的特征,还探索如何有效使用测试数据集中的特征。本文的方法允许模型在获取测试数据上的预测标签后,继续从测试数据中选择一部分预测准确度较高的数据,将这些数据加入训练集共同参与训练。这有助于修正测试数据集中的错误标签,挑战深度学习模型传统上依赖的独立同分布(i.i.d)假设。现实情况中,训练数据集和测试数据集的特征分布存在差异,特征分布不完全对齐[4]。如果仅使用训练集训练的模型参数直接预测测试数据集的标签,会导致预测偏差。为减少这种误差,必须尽可能学习测试数据集上的特征,缓解训练数据集和测试数据集之间的数据不对齐问题。通过优化模型参数并修正先前的预测结果,可减少分布偏差导致的错误标注。因此,本文的方法不仅利用了未标注数据的特征,还进一步利用测试数据集的特征,提高模型的泛化能力和准确率。

为进一步阐释本研究所提方法与现有相关方法之间的差异,提供以下说明:

1)当前的半监督学习方法主要通过筛选出噪声较少的部分未标注数据参与训练过程,筛选基于深度神经网络(DNN)对样本的预测概率的准确性。然而,这些方法往往没有充分考虑DNN对样本预测的误差。本文提出的方法通过将目标样本与其近邻样本的信息结合起来,计算目标样本的置信度,全面考虑DNN对样本预测的准确性。基于K-近邻加权的损失机制从新的角度选择高置信度样本参与训练,展示了该方法的创新性。

2)现有的半监督情感分类研究未能进一步探索和利用测试数据集的特征,仅限于挖掘未标注数据的特征。本研究提出的2阶段优化模式,通过在模型训练的第1阶段采用K-近邻加权方式,对可能预测错误的样本赋予较小权重,对可能预测正确的样本赋予较大权重,最大程度减少错误累积。随后,在第2阶段的自训练过程中,通过Teacher模型和Student模型交替标注测试集数据,利用已标注测试数据集的特征作为训练集特征,参与下1轮的特征学习。

笔者提出2阶段优化模式如图1所示,采用K-近邻加权的方式在模型第1阶段训练过程中给予最可能预测错误的样本较小权重,给予最可能预测正确的样本较大权重,尽可能缓解错误累积。接着通过第2阶段的

self-training，Teacher模型和Student模型交替标注测试集数据，已标注的测试数据集作为训练集用于下1轮特征学习过程。

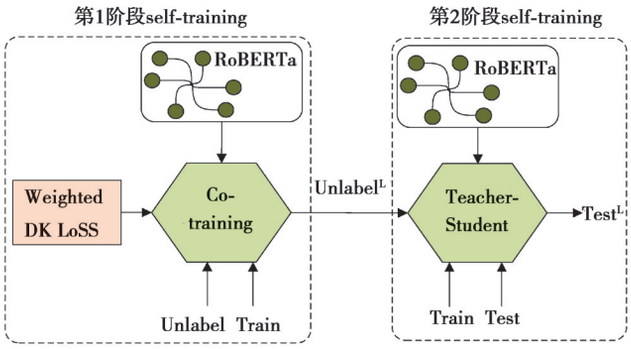

图1 本文所提出的2阶段self-training框架图

Fig. 1 The diagram of the two-stage self-training framework proposed in this paper.

综上所述，本文的主要贡献包括3个方面：

1)提出一种基于K-近邻正则化样本权重机制的损失机制，在模型训练过程中有效降低预测错误样本产生的误差累积；

2) 提出一种小样本上基于2阶段的半监督情感分类模型，学习测试数据集上的特征信息且对测试数据集已经标注的样本标签进行再修改；

3) 进行各种对比实验验证，且验证了该方案的有效性。

1 相关工作

文本情感分类依据文本粒度大小分为文档级情感分类、句子级情感分类和方面级情感分类。本文聚焦于句子级情感分类任务，该任务在情感分析领域占有重要地位。最有效的方法依赖于各种深度网络模型，例如，S2SAN[5]，3W-CNN[6]，SR-LSTM[7]，CHL-PRAE[8]。近期随着BERT系列模型的出现，自然语言处理领域出现了巨大发展。由于BERT通过预训练和微调（finetuning）的方法，不仅使自然语言理解任务变得更加规范，而且通过预训练过程得到的嵌入向量自然地融合了丰富的语义知识，使模型能灵活适应各种上下文环境。只需对特定问题领域的数据进行微调，就能在目标领域实现最佳性能，简化了模型的使用和适应过程。随后，众多研究致力于对BERT体系结构进行优化，提升模型性能和效率。例如，ALBERT[9]模型通过减少参数量来加快训练速度并提高模型效果；DeBERTa[10]改进了注意力机制增强掩码解码器；RoBERTa[2]优化了预训练语言模型(pre-trained language models，PLM)，将静态掩码改为动态掩码，移除了下1句预测任务；XLNet引入了双流自注意力机制。此外，还有研究通过将BERT模型与外部知识融合进一步提升模型性能。例如文献[11]提出如何将词性信息融入DNN模型中，这要求在融合外部知识后重新进行预训练，再进行微调。SKEP[12]模型则是将情感词融入预训练过程中。文献[13−14]提出为了将语言知识集成到预训练模型中，设计了新的预训练任务，在给定句子级情感标签的情况下，预测单词、词性标签和掩码位置的情感倾向。

上述研究主要集中于如何充分利用带标签的训练数据。在现实应用场景中，获取大量带标签数据往往是困难的。特别是在小样本的情况下，这些深度学习模型的表现通常不佳，容易发生过拟合，对超参数（如迭代次数、批大小和学习率等）的选择极为敏感。

为了处理目前DNN在有限标签下的预测性能，根据带标签数据的分布特征与未带标签数据的分布特征是不同的假设[15]，半监督文本分类尝试利用未标注数据来蒸馏多样化知识[16]。目前半监督情感分类模型主要采用以下2种策略：

1)教师—学生(Teacher-Student)结构的交替训练模式。Teacher-Student结构的交替训练模式是构建2个独立的模型，（Teacher和Student）来捕获未标注数据的特征，逐步选取置信度高的伪标注数据加入训练集。例如，CEST[17]利用提升的相似度图在self-training过程中更有效地利用数据。SRIFT[18]将Teacher-Student作

为 Stackelberg 游戏,应用经济学中的 Stackelberg 策略优化整个过程。文献[19]通过使用 2 个 Teachers 分别在 labeled 数据和有抖动的 labeled 数据上提取特征。文献[16]提出优化选择带伪标签数据的过程。文献[15]探索在半监督关系抽取中不同模型的不一致性。文献[20]提出 ASTRA,使用弱规则聚合 Student 的伪标签。文献[21]提出一种使用多个分类器参与分类预测,设置不同子分类器的情感贡献权重得到分类的情感置信度,选出置信度高的样本扩大训练集。文献[22]提出 TS-Aug,能结合数据增强到交替训练过程中。

2)协同训练模式(Co-training)。协同训练模式主要通过把全部未标注数据带入训练过程中,根据已标注数据和未标注数据分别设计不同的损失函数用于营造一种区分性的训练过程[23,17]。比如,文献[24]提出 COSINE,加入比较正则化和基于置信度的权重机制。文献[25]提出了一种协同训练框架 MixTex,采用 TMix 去增强训练样本,同时计算有监督的损失和一致性损失。

尽管现有的半监督学习模型有效利用了未标注数据,主要局限于这些数据的使用,并未解决训练数据集与测试数据集特征分布不对齐问题。大都基于独立同分布(i.i.d)的假设,忽略了训练集和测试集在特征分布上的不一致性。直接使用训练集和未标注数据集来训练模型参数,并用其预测测试数据集时,并没有采取措施来缩小训练集和测试集之间的特征偏差。此外,由于引入了带伪标签的数据参与训练过程会带来错误累积问题,虽然已经尝试了各种策略来减少错误累积,如使用信息熵过滤掉可能预测错误的样本,但这些策略基本上都是在假设深度神经网络(DNN)能正确预测的前提下进行。DNN 在预测错误的样本上也给出了较高的置信度,这说明需要更深入地解决这个问题。笔者提出的方案提出一种新的方法,不仅能有效利用未标注数据,而且能处理训练集和测试集之间的特征分布不一致性,减轻伪标签引入的错误累积效应,提高半监督学习的整体性能。

2　句子级情感分析任务定义

研究考虑句子情感二分类问题,即分类器只需要标注每个句子是正情感还是负情感。定义如下:

考虑 1 个拥有大量评论的语料库 $\{r_0,r_1,\cdots,r_n\}$,每个评论数据集 r_j 由一系列句子 $\{s_{j1},s_{j1},\cdots,s_{jn}\}$ 组成。句子级别情感分类的目标是为每个句子预测 1 个情感标签,这个标签指示该句子是表达正面情感(标签=1 表示正面情感)还是负面标签(标签=0 表示负面情感)。

3　基于 K-近邻的 LOSS 权重机制

尽管深度神经网络(DNN)的预测概率越极端通常意味着对该样本的预测置信度越高,现实情况却常常并非如此。DNN 可能对其预测结果过于自信,导致对错误标注的样本给出了过高的置信度,很难对这些错误的预测进行纠正。研究提出了一种基于 K-近邻的加权损失机制,促使 DNN 模型能根据邻近样本的预测标签重新评估和调整自己对当前样本的预测准确性。现有研究通常通过计算 DNN 预测概率的信息熵过滤掉置信度低的未标注数据,避免错误地进一步传播。其他方法可能包括直接过滤出预测概率低于特定阈值的样本,或者根据样本在 2 个类别上的概率差异进行排序,选择置信度较高的部分样本进行下 1 轮训练。然而,这些方法主要基于 DNN 对当前样本预测概率准确的前提下进行计算的,并没有充分考虑当前样本的预测是否准确。笔者通过将目标样本的邻近样本纳入置信度计算过程中,提出一种新的方法,基于对当前样本预测准确性的全面考量。因此,所提出的基于 K-近邻加权的损失机制能够从 1 个新的角度选择置信度较高的样本参与训练,展示了方法的创新性。

具体来说,该方法通过使用 K-近邻算法,基于样本的嵌入向量(embedding)计算余弦距离(cosine)(或相似度),识别每个样本在其所在 batch 内的最近邻居。接下来,统计这些近邻中与目标样本预测标签相同样本所占的比例。这一比例反映了在所有最接近的邻居中,有多少比例的样本与目标样本具有相同标签。可以用以下公式表示

$$w(i)=1+\frac{a_i\ln(a_i)}{\log(B)}, \tag{1}$$

$$D_{\mathrm{KL}}=\sum_{i=1}^{B}\tilde{\boldsymbol{y}}\log\frac{\tilde{\boldsymbol{y}}}{f(i;\boldsymbol{W})}, \tag{2}$$

$$L_{\mathrm{KL}}=\frac{1}{|B|}\sum_{i=1}^{B}w(i)\,D_{\mathrm{KL}}\ , \tag{3}$$

其中：a_i是同1个batch内部相同标签占所有最近邻居的比例；$w_i(i)$是本文通过对K-近邻计算得到的样本i的权重系数，B是batch size用于正则化；D_{KL}是Kullback-Leibler(KL)散度，表示度量预测标签$f(i;\boldsymbol{W})$分布和伪标签$\tilde{\boldsymbol{y}}$分布之间的分布差异，$\boldsymbol{W}$表示模型参数；L_{KL}表示提出的基于K-近邻的加权重系数后更精准的损失值。实验中设置K为16近邻。

4 半监督训练过程

尽管大多数现有的半监督情感分类研究依赖于教师—学生(Teacher-Student)训练或协同训练来降低未标注数据中的噪声，往往没有充分利用测试数据集中的特征，而仅集中挖掘未标注数据的特征。笔者提出了一种2阶段的优化模式。在模型训练过程中，采用K-近邻加权的方法，对可能预测错误的样本赋予较小权重，对可能预测正确的样本赋予较大权重，最大限度减轻错误累积的问题。在优化的第2阶段，采用自训练(self-training)方法，其中Teacher模型和Student模型交替对测试集数据进行标注，利用已标注的测试数据集特征作为训练数据，用于下1轮的特征学习。这种策略不仅增强了模型对未标注数据的利用效率，还通过直接引入测试数据集进一步提高模型的泛化能力和准确性。

在本文提出的2阶段优化模式中，第1阶段的自训练(self-training)主要依靠有限的带标签数据和初始未标注的数据共同参与训练，以学习未标注数据的特征，最终为这些未标注数据分配预测标签。第2阶段的自训练(self-training)将训练集与第1阶段标注好的未标注数据合并，形成新的扩展训练集。随后，在迭代过程中，每1轮都会从测试数据中选取部分预测准确率较高的样本，加入到训练集中，持续优化模型。通过这种方式，模型能够在迭代中不断提升性能，最终对测试集中的所有样本进行1次性预测。这个过程不仅增强了模型对未标注数据特征的学习能力，还通过逐步引入测试数据进一步优化模型，提高对新数据的适应性和预测准确性。

4.1 第1阶段self-training

图2表示所提半监督方案的第1阶段self-training，主要流程是使用带标签据初始化RoBERTa模型，使用训练好的模型预测未标注数据上的伪标签，合并未标注数据和训练数据一起联合训练RoBERTa模型，此时使用所提出的基于K-近邻的损失权重机制最小化模型中的损失。

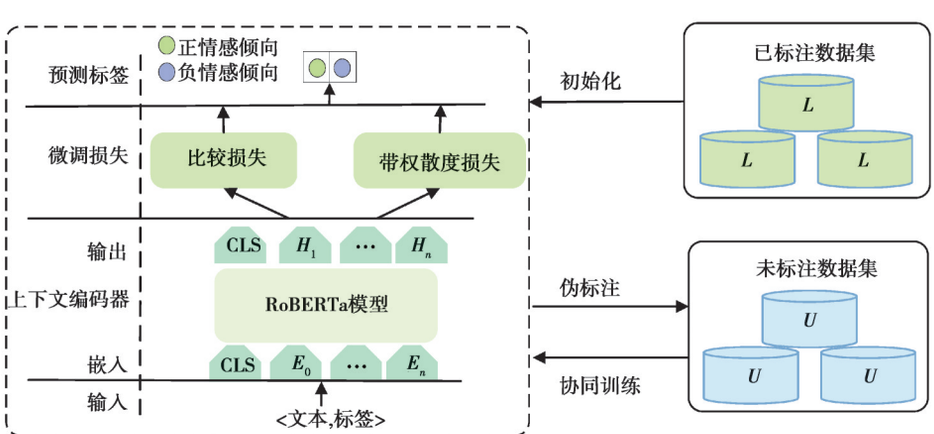

图2 第1阶段self-training框架图

Fig. 2 The frame diagram of the first stage self-training

4.1.1 使用带标签的数据初始训练模型

研究所使用的分类器$f=g\circledcirc$RoBERTa包含2部分，RoBERTa是1个预训练的模型，输出隐藏层向量表示，g是任务相关的分类器头，输出2维的向量，每个维度相应于指定类的预测概率，◎表示连接符号。第1阶段self-training框架如图1所示，使用带标签的数据初始化预训练模型。此外，类似之前研究，采取early stopping方法在半监督中比较广泛，减轻模型对标签中噪声的过拟合问题。

4.1.2　联合训练过程

联合训练过程主要通过同时利用未标注(unlabeled)数据集和已标注(labeled)数据集进行微调(fine-tuning),旨在减轻可能由于错误标注的数据在训练集中引起的误差传播问题。首先,利用已标注数据集对模型进行初始训练,预测未标注数据集的伪标签,将这些伪标注的数据与已标注数据集合并。为了缓解误差传播,采用了基于K-近邻加权的损失函数及当前被广泛认为有效的对比损失函数(contrastive loss,CL)。通过这种方法,模型能不断更新伪标签和模型参数,提高整体训练过程的准确性和鲁棒性。所提方法还结合了比较损失[6]。比较损失是用于指导DNN学习更加清晰的分类边界,主要通过引导模型学习同1个类别的数据具有相似的表示,不同类别的数据具有不同表示,否则如果相同类别具有较大距离,或不同类别如果具有较小的距离值则通过在损失函数中加入距离值作为惩罚。

$$D_{\mathrm{KL}}=\sum_{i=1}^{N_{\mathrm{b}}}p_i\log\frac{p_i}{q_i}\ , \tag{4}$$

$$L_{\mathrm{KL}}=\frac{1}{|N_{\mathrm{b}}|}\sum_{i=1}^{N_{\mathrm{b}}}w\ (i)\,D_{\mathrm{KL}}\ , \tag{5}$$

$$L_{\mathrm{SCL}}=\sum_{i=1}^{N_{\mathrm{b}}}-\frac{1}{N_{\mathrm{b},y_i}-1}\sum_{j=1}^{N_{\mathrm{b}}}\prod_{i\neq j}\prod_{y_i=y_j}\log\frac{\exp(\boldsymbol{\varphi}(x_i)\cdot\boldsymbol{\varphi}(x_j)/\tau)}{\sum_{k=1}^{N_{\mathrm{b}}}\prod_{i\neq T}\exp\ (\boldsymbol{\varphi}(x_i)\cdot\boldsymbol{\varphi}(x_R)/\tau)}\ , \tag{6}$$

$$L=\beta_1 L_{\mathrm{KL}}+\beta_2 L_{\mathrm{SCL}}\ , \tag{7}$$

公式(4)是Kullback-Leibler(KL)散度计算公式,主要度量当前模型预测分布和伪标签分布之间的差异,p_i是第i个样本的伪标签,q_i表示样本i的预测概率输出,公式(5)中L_{KL}是笔者定义的加入K-近邻权重后的损失函数,$w(i)$是基于K-近邻的权重系数。公式(6)中L_{SCL}是比较损失函数,其中,$\boldsymbol{\varphi}(x_i)$是深度网络中编码层的输出,表示对任意句子$x_i$的高维向量表示。$N_{\mathrm{b}}$是batch size,$N_{\mathrm{b},yi}$是同个batch内部与第$i$个实例相同标签的实例个数。$\prod$表示指示器函数,用于检测是否满足指定条件,值为1表示符合指定条件,值为0表示不符合指定条件时。公式(7)是对基于K-近邻的损失和比较损失加权后的损失进行联合训练,β_1和β_2是权衡这2部分损失占比的超参数,在具体的实验中设置$\beta_1=0.7$,$\beta_2=0.3$。

4.2　第2阶段self-training

第2阶段主要针对在Test数据上抽取部分准确率较高的数据合并到训练集中训练模型,如图3所示。该过程同样在前1阶段已经fine-tuning好模型的基础上再次训练该模型。training数据和test数据特征分布不对齐,通过接着学习test数据上所拥有的特征,能打破现有半监督方法存在的独立同分布特性(i.i.d假设)。

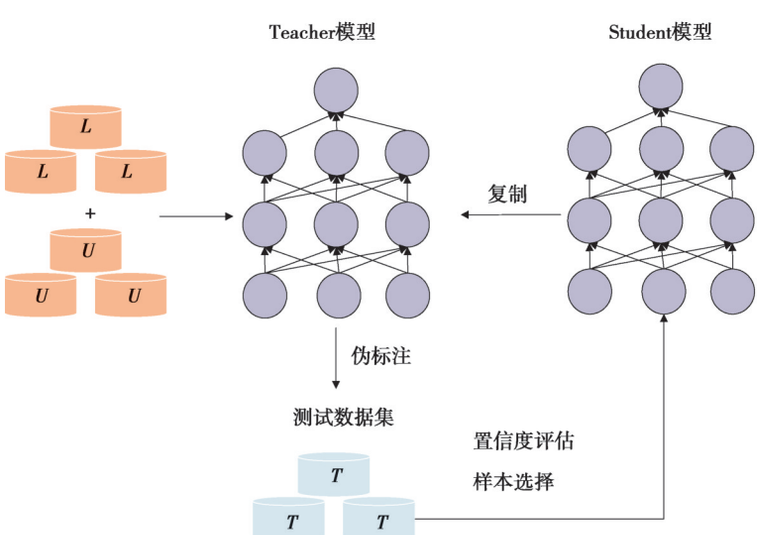

图3　第2阶段self-training框架图

Fig. 3　Diagram of the second stage self-training framework.

4.2.1 基于BALD的未标注样本选择

所提方案采用BALD[16]从测试数据中选择未标注样本。BALD(bayesian active learning by disagreement)方法的目标是选择最大化模型参数信息熵样本,或最大化预测和模型后验之间的信息增益。利用模型对数据的不确定性指导数据选择,从最有信息量的样本中学习,提高学习效率和模型的性能。对具有较低信息增益的样本,模型更加确定,由于具有较低信息增益,模型从样本中学习到的信息较少,直接使用较低信息增益的样本训练模型会导致过拟合,相反,具有较高信息增益的样本对模型学习贡献较多,但也容易受到错误伪标签的破坏。为了权衡这2个场景,采用不同的权重抽取样本,较低熵的样本抽取得更多一些,较高熵的样本抽取更少一些。具体计算过程如下所示公式,对于1个数据样本 $x_i \in \text{Test}$,

x_i的采样权重s_i为

$$s_i = \frac{1-\hat{B}(y_i, \boldsymbol{W}|\boldsymbol{\varphi}(x_i), \text{Test})}{\sum_{i=1}^{|\text{Test}|}[1-\hat{B}(y_i, \boldsymbol{W}|\boldsymbol{\varphi}(x_i), \text{Test})]}, \tag{8}$$

$$\hat{B}(y_i, \boldsymbol{W}|\boldsymbol{\varphi}(x_i), \text{Test}) = -\sum_{c=1}^{C}\left(\frac{1}{T}\sum_{t=1}^{T}\hat{p}_c^t\right)\log\left(\frac{1}{T}\sum_{t=1}^{T}\hat{p}_c^t\right) + \frac{1}{T}\sum_{t=1}^{T}\sum_{c=1}^{C}\hat{p}_c^t\log(\hat{p}_c^t), \tag{9}$$

$$\hat{p}_c^t = \text{soft}\max(\boldsymbol{f}^{W_t}(\boldsymbol{\varphi}(x_i))), \tag{10}$$

通过公式(8)可以得到样本x_i的采样权重s_i。$\hat{p}_c^t$是在MC dropout中模型在第t轮对类c的预测概率,具体计算过程如公式(10)所示,$\boldsymbol{f}^{W_t}(\boldsymbol{\varphi}(x_i))$表示第$t$轮模型$\boldsymbol{f}^{W}$得到的logits,$\boldsymbol{W}$是模型的参数,Test是测试数据集,$y_i$是预测样本$x_i$的预测概率输出。$\hat{B}$是根据公式(9)计算得到的BALD数值,$C$表示类别总数,$T$是迭代次数,实验中设置$T=30$。

4.2.2 训练过程

首先,Teacher模型使用带标签数据和第1阶段获取的带伪标签的已标注数据作为训练集参与训练,得到训练好的模型$\boldsymbol{f}^{W}$,$\boldsymbol{W}$表示模型参数。接着采用基于BALD的方法从这些测试数据集上选择部分数据作为新增的训练集合并到原有的训练集中训练Student模型。Student模型训练好后copy模型参数给Teacher模型,Teacher模型再次预测测试数据集,得到最新预测概率,再次应用BALD方法最新选择部分测试数据添加到原始的训练数据集中训练Student模型。Teacher模型及其模型使用的基础模型都是RoBERTa模型。因为RoBERTa模型在文本分类领域性能比较稳定,且擅长做语义理解类相关任务[2]。

算法1.第2阶段self-training训练过程

$\boldsymbol{W}$表示模型参数,W_t表示第t轮的模型参数,T表示模型迭代的总共轮数,Test表示测试数据集,$\boldsymbol{x}$为未标注数据D_u中的句子向量,$\boldsymbol{f}^{W_t}(\boldsymbol{x})$表示第$t$轮模型得到的logits,$\hat{p}_c^t$表示模型在$t$轮的预测概率输出

输入:训练数据集D_L和测试数据集Test

输出:更新后的模型$\boldsymbol{f}^{W}$

1.在带标签的数据集上D_L微调模型$\boldsymbol{f}^{W}$

2.While $n <$ Itr do

3. 从Test中随机选择一部分未标注的数据D_u

4. For $\boldsymbol{x} \in D_u$

5. For $t \in T$

6. W_t~Droupout($\boldsymbol{W}$)

7. $\hat{p}_c^t$ = softmax($\boldsymbol{f}^{W_t}(\boldsymbol{x})$)

8. End For

9. 根据公式(9)计算样本x_i的BALD值

10. 根据公式(8)计算样本x_i的采样权重s_i

11. End For

12. 在D_u中根据每个样本采样权重s_i抽取$|R_u|$个实例

13. 使用模型f^W伪标注R_u

14. 使用新的训练集$D_L -> D_L + R_u$重新训练模型f^W

15. End While

16. 输出优化后的模型f^W

5　实验

为了实验评估,使用了4个句子情感分类任务公开使用的标准数据集,分别是MR,CR,Twitter2013和Twitter2016。其中:MR是电影评论集合;CR是电子商品评论集合;Twitter2013和Twitter2016是微博评论,内容长度上更加精简。表1列出所有数据集的统计信息。

表1　CR,MR,Twitter2013,Twitter2016数据集的统计信息

Table 1　Statistical information for the CR, MR, Twitter 2013, and Twitter 2016 datasets

数据集名称	训练集	验证集	测试集
MR	8 534	1 066	1 050
CR	2 262	754	754
Twitter2016	6 920	872	1 821
Twitter2013	5 098	915	2034

5.1　对比实验

后续的实验不仅与目前最好的有监督情感分类模型做对比,且与目前最好的半监督情感分类模型做对比。由于研究主要是句子级情感二分类问题,使用的度量标准是准确率和Macro-F_1(文章中简写为F_1)。目前性能最好的有监督情感分类模型主要有以下:

1)RoBERTa模型[2]。文本分类主要采用RoBERTa模型,性能比较稳定,且擅长执行语义理解相关类任务。

2)XLNet模型[3]。XLNet是对BERT模型的优化改进,是通用的自回归预训练模型,能够学习双向文本语义。

3)EFL[12]。该模型通过把类标签转化为辅助句子,使更多的任务能够统一转化为文本蕴含任务。

4)DualCL[26]。最近提出用于情感分类的模型,能同时学习输入句子的特征和分类器的参数特征。

目前性能最好的基于半监督的情感分类任务模型主要有以下:

1)UST[16]。该模型是一种Teacher-Student半监督方案,主要用于文本分类,使用不确定度对unlabeled数据进行采样的方法选取置信度高的伪标注数据。

2)COSINE[24]。也是一种文本分类的半监督方案,使用比较loss且结合了信息熵的置信度权重机制以减少错误累积。

3)MTGT[19]。研究提出了一种半监督文本分类方案,采用2个Teacher训练,1个Teacher在Labeled数据上训练,另1个Teacher在增强后的数据中训练,然后加权这2种伪标签后得到新的unlabeled数据集上的伪标签作为Student模型使用的训练集。

4)DisCo[27]。采用一种新颖的协同训练技术,通过促进不同视图下的Student模型之间的知识共享来优化多个Student模型。

5)RNT[28]。为了缓解噪声,采用基于来自标签文本的证据支持度计算不确定性排序unlabeled文本,同时使用负训练方式训练RNT。

表2展示了在CR和Twitter2016数据集上,所提方法(Ours)与当前流行的最佳方法之间的性能比较。可以看出提出的方法不仅优于最佳的有监督模型,也超过了最佳的半监督模型。特别是在仅有0.25%训练数据的情况下,CR和Twitter2016数据集上,相比于有监督模型,所提方案在准确率上分别提高了10.34%和16.06%;与最佳的半监督模型相比,分别提高了3.56%和3.12%。当训练数据增至1%时,CR和Twitter2016数据集上,所提方法相比现有最佳半监督模型分别提高了1.01%和0.96%。在训练数据为3%的情况下,提升分别为0.82%和1.54%。此外,还在MR和Twitter2013数据集上进行了类似的对比实验。表3展示了这2个数据集在不同训练数据比例下的性能表现,证明了方法在不同数据集和不同数据规模下的普适性和有效性。

表2 CR和Twitter2016在不同比例训练集的性能对比

Table 2 Performance comparison of CR and Twitter 2016 on different training set proportions %

训练数据比例	模型	CR		Twitter2016	
		准确率	F_1	准确率	F_1
0.25	RoBRERTa	56.88	47.25	31.89	2.09
	XLNet	57.08	67.00	32.35	3.01
	EFL	64.37	78.20	33.94	7.68
	DualCL	62.91	63.73	54.16	46.63
	UST	76.55	81.77	60.68	62.87
	COSINE	71.05	77.31	61.40	63.21
	MTGT	75.63	81.71	56.79	59.91
	DisCo	80.15	83.46	65.32	68.51
	RNT	79.25	82.69	67.10	69.92
	Ours	83.71	86.53	70.22	74.31
0.50	RoBERTa	50.00	39.06	35.24	12.89
	XLNet	62.32	74.03	37.44	16.99
	EFL	65.17	77.46	51.10	47.30
	DualCL	64.37	65.47	69.20	73.70
	UST	81.67	83.59	73.32	76.22
	COSINE	80.01	82.11	73.00	75.92
	MTGT	79.87	82.38	71.20	73.53
	DisCo	82.20	84.96	74.22	76.15
	RNT	83.71	85.30	75.18	79.31
	Ours	85.43	88.93	76.24	83.72
1.00	RoBERTa	64.10	79.85	36.12	16.72
	XLNet	64.11	78.13	37.66	18.04
	EFL	66.49	79.10	79.83	84.74
	DualCL	71.19	73.60	83.22	82.66
	UST	84.15	87.26	84.12	86.19
	COSINE	83.57	82.35	76.14	83.15
	MTGT	81.59	86.41	76.03	84.56
	DisCo	84.22	86.05	84.20	86.92
	RNT	84.82	87.30	84.39	87.38
	Ours	85.83	89.80	85.35	81.61

续表 2

训练数据比例	模型	CR		Twitter2016	
		准确率	F_1	准确率	F_1
3.00	RoBERTa	87.95	90.09	83.61	88.53
	XLNet	82.65	86.83	77.96	85.38
	EFL	90.93	93.18	84.84	89.81
	DualCL	85.30	88.33	87.38	86.66
	UST	88.07	90.62	87.03	89.21
	COSINE	89.30	88.33	85.57	87.65
	MTGT	82.02	87.13	84.85	86.85
	DisCo	90.13	92.30	86.30	89.21
	RNT	91.76	93.59	89.03	91.79
	Ours	92.58	94.26	90.57	93.30

从表 3 可以看出,在 1% 的 MR 和 Twitter2013 数据集上时,本文所提方法在准确率上分别可以达到 86.22%, 90.02%,macro-F_1 分别可以达到 86.45%,92.93%,比目前最好的方法半监督方法在准确率上分别超出 2.5%,1.18%,在 macro-F_1 上分别超出 2.75%,4.09%。在 0.5% 训练集的时候,MR 和 Twitter2013 在准确率上分别比最好的方法超出 1.55% 和 1.11%。当数据负载为 3% 的训练集的 MR 和 Twitter2013 时,所提方法在准确率上可以超出目前最好模型 0.33% 和 0.15%。综上可以看出,半监督模型普遍比有监督模型性能好,因为半监督模型利用了 unlabeled 数据上的特征信息,所提方法也同样利用了 unlabeled 数据上的特征,不仅止步于如何充分利用 unlabeled 数据上的特征,同时也利用了 test 数据上的特征,试图缩减训练集和测试集之间特征分布差异。

表 3　MR 和 Twitter2013 在不同比例训练集的性能对比

Table 3　Performance comparison of MR and Twitter 2013 on different training set proportions　%

训练数据	模型	MR		Twitter2013	
		ACC	F_1	ACC	F_1
0.25	RoBERTa	49.77	66.46	72.52	84.07
	XLNet	49.76	66.45	68.68	80.42
	EFL	64.94	70.87	65.92	77.04
	DualCL	62.32	54.32	65.63	62.11
	UST	65.13	70.69	76.19	82.70
	COSINE	65.42	71.15	75.91	70.20
	MTGT	66.35	72.19	79.15	85.31
	DisCo	80.30	82.05	79.69	81.63
	RNT	81.00	82.69	78.30	80.17
	Ours	82.10	83.41	81.02	86.28

续表3

训练数据	模型	MR		Twitter2013	
		ACC	F_1	ACC	F_1
0.50	RoBERTa	62.14	71.51	72.81	84.00
	XLNet	58.07	36.58	69.02	79.61
	EFL	80.97	82.20	70.50	82.26
	DualCL	76.38	75.86	73.65	74.57
	UST	82.15	82.10	82.12	85.51
	COSINE	82.76	82.74	82.65	85.42
	MTGT	82.99	81.82	82.22	85.63
	DisCo	81.30	81.45	83.03	85.20
	RNT	82.09	82.28	84.19	87.32
	Ours	84.54	83.84	85.30	89.51
1.00	RoBERTa	81.82	80.57	78.66	86.38
	XLNet	82.09	82.31	72.57	84.08
	EFL	82.94	82.87	87.89	89.63
	DualCL	81.81	81.43	88.84	88.84
	UST	82.57	82.31	83.43	88.69
	COSINE	83.72	83.70	87.71	83.91
	MTGT	83.13	83.68	88.25	88.67
	DisCo	83.09	85.37	88.09	90.17
	RNT	83.00	86.30	87.38	89.20
	Ours	86.22	86.45	90.02	92.93
3.00	RoBERTa	84.91	84.62	87.66	91.41
	XLNet	84.37	84.98	85.35	90.11
	EFL	83.97	84.16	89.45	92.81
	DualCL	86.46	86.21	90.28	90.75
	UST	84.53	83.99	89.87	92.01
	COSINE	85.66	85.68	90.17	93.12
	MTGT	86.12	86.68	90.81	93.58
	DisCo	85.97	85.91	90.19	93.27
	RNT	86.06	86.05	90.21	93.15
	Ours	86.79	86.84	90.96	93.61

5.2 敏感性测试

5.2.1 self-training

为了证明所提方法中第2阶段self-training的重要性，接着展示了只执行第1阶段self-training和同时执

行第 1 阶段和第 2 阶段 self-training 在 CR 和 twitter2016 2 个数据集上的表现结果。从图 4 和图 5 中可以看出,在 CR 和 Twitter2016 2 个数据集上 2 阶段 self-training 比 1 阶段 self-training 性能明显好一些。说明了相比 1 阶段 self-training, 2 阶段的 self-training 可缓解训练集和测试集之间的特征分布偏差问题。

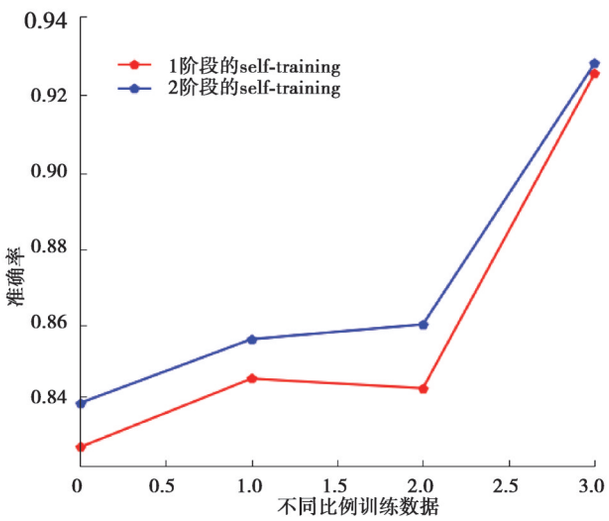

图 4　在 CR 数据集 1~2 阶段 self-training

Fig. 4　1~2 self-training on the CR dataset.

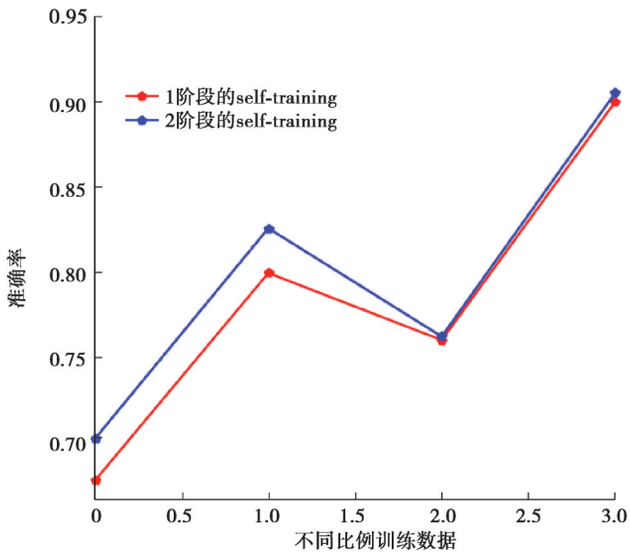

图 5　在 Twitter2016 数据集 1~2 阶段 self-training

Fig. 5　1~2 self-training on the Twitter 2016 dataset.

5.2.2　加入 K-近邻权重的 loss VS 不加入 K-近邻权重的 loss

通过表 4 中加入 K 近邻 loss 前后在 CR 和 Twitter2016 这 2 个数据集上的效果可看出,在没有加入 K-近邻 loss 前,3%CR 数据集和 3%Twitter2016 数据集的准确率分别是 0.896 7 和 0.841 3,加入 K-近邻 loss 后的准确率分别是 0.928 5,0.905 7,分别增加了 3.18% 和 6.44%。说明所提方法采用 K-近邻 loss 机制后在模型训练过程中提供不同视角检查出有可能标注错误的伪标签,同时给予可能标注错误的伪标签较低的学习权重,这种方法从一定程度上降低错误伪标签造成的影响。

表 4　CR 和 Twitter 上 K-近邻 loss

Table 4　CR with K-nearest neighbors loss on Twitter　　%

数据集名称	比例	加入 K-近邻损失		去掉 K-近邻损失	
		准确率	F_1	准确率	F_1
CR	0.25	0.837 1	0.865 3	0.735 1	0.795 9
	0.50	0.854 3	0.889 3	0.826 5	0.861 1
	1.00	0.858 3	0.898 0	0.845 0	0.889 9
	3.00	0.928 5	0.942 6	0.896 7	0.922 3
Twitter2016	0.25	0.702 2	0.743 1	0.611 3	0.629 6
	0.50	0.762 4	0.837 2	0.755 2	0.796 9
	1.00	0.853 5	0.816 1	0.832 4	.0.878 0
	3.00	0.905 7	0.933 0	0.841 3	0.886 8

6 结 论

1)研究在现有半监督方案的基础上提出一种基于K-近邻正则化样本权重机制的loss,有效降低预测错误样本产生的误差累积问题,通过敏感性实验观测到该方法在一定程度上提升准确率;

2)提出一种小样本上基于2阶段的半监督情感分类模型,学习Test数据上的特征信息且对Test数据上已经标注的样本标签进行再修改,结果显示所提出方案的有效性。

参考文献

[1] Devlin J, Chang M, Lee K,et al. BERT: pre-training of deep bidirectional transformers for language understanding[C]//2019 Conference of the North American Chapter of the Association for Computational Linguistics. Minnesota: Association for Computational Linguistics, 2019: 4171-4186.

[2] Liu Y, Ott M, Goyal N, et al. Roberta: A robustly optimized bert pretraining approach[EB/OL].(2019-7-27)[2024-11-6]. https://arxiv.org/abs/1907.11692.

[3] Yang Z, Dai Z, Yang Y, et al. Xlnet generalized autoregressive pretraining for language understanding[C]//33rd International Conference on Neural Information Processing Systems. Red Hook, USA:Curran Associates Inc, 2019:5753-5763.

[4] Zhao Z, Zhou L, Duan Y, et al. DC-SSL: Addressing mismatched class distribution in semi-supervised learning[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE, 2022:9747-9755.

[5] Wang P, Li J, Hou J. S2SAN: A sentence-to-sentence attention network for sentiment analysis of online reviews[J]. Decision Support Systems, 2021,149:113603.

[6] Zhang Y, Zhang Z, Miao D, et al. Three-way enhanced convolutional neural networks for sentence-level sentiment classification [J].Information Sciences, 2019, 477:55-64.

[7] Rao G, Huang W, Feng Z, et al. LSTM with sentence representations for document-level sentiment classification[J]. Neurocomputing, 2018, 308(35):49-57.

[8] Fu X, Liu W, Xu Y, et al.Combine HowNet lexicon to train phrase recursive autoencoder for sentence-level sentiment analysis [J].Neurocomputing, 2017, 241(7):18-27.

[9] Lan Z, Chen M, Goodman S, et al. ALBERT: a Lite BERT for self-supervised learning of language representations[EB/OL]. (2019-9-26)[2024-11-6]. https://arxiv.org/abs/1909. 11942.

[10] He P, Liu X, Gao J, et al. DeBERTa: decoding-enhanced BERT with disentangled attention[J]. (2020-6-5)[2024-11-6].https://arxiv.org/abs/2006.03654.

[11] Pasquier C, Da Costa Pereira C, Tettamanzi A G B. Extending a fuzzy polarity propagation method for multi-domain sentiment analysis with word embedding and pos tagging[C]//ECAI 2020-24th European Conference on Artificial Intelligence. Spain:IOS Press, 2020: 2140-2147.

[12] Tian H, Gao C, Xiao X, et al. SKEP: sentiment knowledge enhanced pre-training for sentiment analysis[C]//58th Annual Meeting of the Association for Computational Linguistics. Pennsylvania, United States: Association for Computational Linguistics, 2020:4067-4076.

[13] Zhao Q, Ma S, Ren S. KESA: a knowledge enhanced approach for sentiment analysis[C]// 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing. Pennsylvania,United States：Association for Computational Linguistics, 2022:766-776.

[14] Ke P , Ji H, Liu S, et al. SentiLARE: sentiment-aware language representation learning with linguistic knowledge[C]// 2020 Conference on Empirical Methods in Natural Language Processing. Pennsylvania, United States：Association for Computational Linguistics, 2020:6975-6988.

[15] Li W L, Qian T Y. From consensus to disagreement: multi-teacher distillation for semi-supervised relation extraction[EB/OL]. (2021-12-2)[2024-11-6]. https://arxiv.org/abs/2112.01048.

［16］Mukherjee S, Awadallah A H. Uncertainty-aware self-training for few-shot text classification[C]//34th International Conference on Neural Information Processing Systems. Canada:Curran Associates Inc, 2020:21199-21212.

［17］Tsai A C Y, Lin S Y, Fu L C. Contrast-enhanced semi-supervised text classification with few labels[C]//AAAI Conference on Artificial Intelligence.Vancouver, Canada:AAAI Press, 2022:11394-11402.

［18］Zuo S, Yu Y, Liang C, et al. Self-training with differentiable teacher[C]//Findings of the Association for Computational Linguistics. Pennsylvania,United States：Association for Computational Linguistics, 2022:933-949.

［19］Lin Q, Ng H T. A semi-supervised learning approach with two teachers to improve breakdown identification in dialogues[C]//、AAAI Conference on Artificial Intelligence.Vancouver, Canada: AAAI Press, 2022:11011-11019.

［20］Karamanolakis G, Mukherjee S, Zheng G, et al. Self-training with weak supervision[C]//2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Pennsylvania, United States: Association for Computational Linguistics, 2021:845-863.

［21］陈 珂, 黎树俊, 谢 博 . 基于半监督学习的微博情感分析[D]. 茂名：广东石油化工学院, 2018.
Chen K, Li S J, Xie B. Sentiment analysis of Chinese micro-blog based on semi-supervised[D]. Maoming: University of Petrochemical Technology,2018.

［22］郭卡, 王芳 . TS-Aug架构的半监督自训练情感分类算法[D]. 合肥：安徽外国语学院，2024.
Guo K, Wang F. Semi-supervised self-training sentiment classification algorithm based on TS-Aug architecture[D]. Hefei: Anhui University of Foreign Languages, 2024.

［23］Li C, Li X, Ouyang, J. Semi-supervised text classification with balanced deep representation distributions[C]//59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing. Pennsylvania,United States: Association for Computational Linguistics, 2021:5044-5053.

［24］Yu Y, Zuo S, Jiang H, et al. Fine-tuning pre-trained language model with weak supervision: a contrastive-regularized self-training[C]//2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Pennsylvania,United States: Association for Computational Linguistics,2020:1063-1077.

［25］Chen J, Yang Z, Yang D. Mixtext: linguistically-informed interpolation of hidden space for semi-supervised text classification [C]//Association for Computational Linguistics. Pennsylvania,United States: Association for Computational Linguistics, 2020: 2147-2157.

［26］Chen Q, Zhang R, Zheng Y, et al. Dual contrastive learning: text classification via label-aware data augmentation[EB/OL]. (2022-1-21)[2024-11-6]. https://arxiv.org/abs/2201.08702.

［27］Jiang W, Mao Q, Lin C, et al. DisCo: distilled student models co-training for semi-supervised text mining[C]//2023 Conference on Empirical Methods in Natural Language Processing. Pennsylvania, United States: Association for Computational Linguistics, 2023:4015-4030.

［28］Murtadha A, Pan S, Wen B, et al. Rank-Aware negative training for semi-supervised text classification[J]. Transactions of the Association for Computational Linguistics. 2023, 11:771-786.

（编辑　侯　湘）