

TDQE:一种面向深度学习的文本数据质量评估方法

罗春旭^{1,2}, 熊海旭^{1,2}, 叶雅珍^{1,2}, 丁滢³, 宗世泽³, 熊贇^{1,2}, 朱扬勇^{1,2}

1. 复旦大学计算机科学技术学院, 上海 200438;
2. 上海市数据科学重点实验室, 上海 200438;
3. 中国人民解放军国防科技大学计算机学院, 湖南 长沙 410073

摘要

文本数据质量作为语言模型性能的重要影响因素, 其评估方法对模型训练效果具有决定性作用。针对现有文本数据质量评估方法存在的计算成本过高与评估指标不全面的问题, 提出了一种面向深度学习的文本数据质量评估方法 TDQE。具体而言, TDQE 利用文本摘要生成模型的 Dropout 生成一系列随机子网络, 得到数据样本嵌入表示, 捕获数据样本的语义一致性, 用于评估样本的鲁棒性; 利用文本相似匹配模型计算数据样本与其文本摘要之间的匹配度, 用于评估样本的准确性; 设计加权鲁棒性与准确性指标, 用于衡量文本数据质量。在公开数据集上, 将 TDQE 与主流训练数据质量评估算法进行对比实验, 实验结果表明 TDQE 的性能优于对比算法。

关键词

深度学习; 文本数据; 数据质量; 质量评估

中图分类号: TP311.1

文献标志码: A

doi: 10.11959/j.issn.2096-0271.2025073

TDQE: a quality evaluation method for text data in deep learning

LUO Chunxu^{1,2}, XIONG Haixu^{1,2}, YE Yazhen^{1,2}, DING Yan³, ZONG Shize³, XIONG Yun^{1,2}, ZHU Yangyong^{1,2}

1. School of Computer Science, Fudan University, Shanghai 200438, China
2. Shanghai Key Laboratory of Data Science, Shanghai 200438, China
3. Department of Computer Science, National University of Defense Technology, Changsha 410073, China

Abstract

Text data quality is an important factor affecting the performance of language models, and its evaluation methodology is considered decisive for model training effectiveness. To address the issues of high computational costs and incomplete evaluation metrics in existing text data quality assessment methods, a deep learning-oriented text data quality evaluation (TDQE) method was proposed. Specifically, (1) the Dropout module of a text summarization model was utilized to generate multiple stochastic sub-networks, producing embedded representations of data samples to capture semantic consistency, thereby evaluating sample robustness; (2) a text similarity matching model was employed to compute the alignment between data samples and their summaries, assessing sample accuracy; (3) weighted robustness and accuracy metrics were designed to quantify overall text data quality. Comparative experiments were conducted on

public datasets between TDQE and state-of-the-art methods, and the results demonstrated that TDQE outperformed existing mainstream algorithms.

Key words

deep learning, text data, data evaluation, quality evaluation

0 引言

语言模型具有强大的表达能力及泛化能力,并在文本生成、知识推理等任务上都表现出了优异的性能。研究表明,使用更高质量的训练数据能够有效提升语言模型在多个下游任务中的性能表现^[1]。然而,庞大的数据量与复杂的数据来源导致高质量的文本数据难以获取。并且,大多数语言模型的训练数据是由部分标注数据与大量未标注数据构成的,这两方面原因导致语言模型的训练数据中包含大量的低质量数据样本。研究表明,删除部分低质量数据样本可以在减少模型训练量的同时获得同样甚至更高的模型性能^[2-3],而移除低质量数据样本的前提是对训练数据集中的数据样本质量进行评估。

在图像质量评估方法中,研究者指出图像的质量应当能够反映其预期的模型性能^[4]。类比于文本数据,文本数据的质量应当能够反映其预期的语言模型的性能,即语言模型在下游任务的性能越强,则用于模型训练的数据质量越高。基于上述思想,研究者提出了基于沙普利值(Shapley value)的方法^[5-6]与基于强化学习(reinforcement learning, RL)的方法^[7-8]。然而,上述方法的时间复杂度与训练数据的数量、语言模型的参数量呈正相关性。对于训练数据规模不断扩大、模型结构日益复杂、参数量快速增加的语

言模型而言,这样的训练成本是难以接受的。并且,上述质量评估方法普遍依赖于模型下游任务的性能指标,这就导致其质量评估结果更加倾向于能够完成下游任务的数据样本,而忽略数据样本质量的整体性。例如,在文本分类任务中,数据样本含有类别相关关键词的数量是影响分类结果的关键因素之一,这将导致现有方法倾向于为相关关键词数量多的数据样本赋予更高的质量分数,而忽略数据样本本身的其他质量因素,如文本的流畅度、可阅读性等。

基于上述不足,从文本数据质量的整体性角度而言,能够反映数据样本质量的并非语言模型的性能指标,而是数据样本被语言模型理解的程度。对此,本文提出了一种全新的文本数据质量评估方法TDQE(text data quality evaluation)。TDQE利用文本摘要生成模型,设计语义一致性检测模块与匹配度检测模块,从鲁棒性与准确性两个角度评估数据样本被语言模型理解的程度,即数据样本的质量。选择文本摘要生成模型的原因是该任务需要语言模型对训练数据的词汇、语法、语义等多方面进行理解与处理,文本摘要生成的好坏能够直观反映数据样本是否被语言模型理解。具体而言,在语义一致性检测模块中,TDQE利用文本摘要生成模型中的Dropout^[9]生成多个随机子网络。在文本摘要生成模型的每次前向传播中,Dropout会以不同的方式随机丢弃一部分神经元,进而生成随机子网络。对于同一

数据样本，每个随机子网络均会生成一个嵌入表示，并通过嵌入表示之间的平均距离评估数据样本的鲁棒性。其中，不同随机子网络的嵌入表示代表文本摘要生成模型在不同情况下对数据样本的理解结果。通过评估数据样本的鲁棒性，TDQE 能够反映不同情况下模型对数据样本的处理结果的稳定程度，即模型对同一数据样本是否会产生不同的理解结果。但单一的鲁棒性评估无法反映语言模型对数据样本理解的准确程度，可能导致 TDQE 为模型理解错误的数据样本赋予较高的质量分数。因此，本文进一步设计了匹配度检测模块，利用文本相似匹配模型衡量数据样本与其摘要之间的匹配度，评估数据样本的准确性。准确性越高，则表示模型对数据样本的理解越准确。最终，设计加权鲁棒性与准确性，评估文本数据的质量分数。本文主要贡献如下。

- 本文针对现有方法依赖模型性能指标导致的整体性问题，设计了一种新的文本数据质量评估方法 TDQE。该方法同时利用鲁棒性指标与准确性指标，分别从文本数据被语言模型理解的稳定程度与准确程度评估数据质量。

- 本文提出基于 Dropout 的语义一致性检测方法。利用 Dropout 生成的不同随机子网络的嵌入表示之间的平均距离评估数据样本的鲁棒性，以反映语言模型对数据样本的多次理解结果的稳定程度；提出样本与生成摘要间匹配度检测方法，利用数据样本和对应摘要间的匹配度评估数据样本的准确性，以反映语言模型对数据样本理解的准确程度。

- 本文在公开数据集上与主流算法进行了性能对比。结果表明，本文提出的 TDQE 方法的效果优于目前的主流算法，并且 TDQE 方法并不需要额外训练模型。

1 研究现状

训练数据的质量评估通常是通过量化数据样本对模型性能的整体影响来计算的。针对训练数据的质量评估，一种常用的评估方法是留一法（leave-one-out, LOO）^[10]，它通过移除一个特定的数据本来量化模型前后性能的差异，并将此作为该数据样本的数据质量分数。LOO 的计算复杂度与数据样本的数量成线性相关，对于大规模文本数据集和复杂的大语言模型来说，计算成本过高。研究者通过改进 LOO 方法提出了影响函数法^[11-12]。影响函数法是一种以高效计算方式近似 LOO 的方法，它利用损失函数的梯度来估计数据样本的质量。但值得注意的是，影响函数法需要计算（或近似）Hessian 矩阵的逆，这对于参数量庞大的深度学习模型而言，计算成本仍然很高。

基于沙普利值的方法^[5-6,13-14]是另一类数据质量评估的方法。这类方法通常需要考虑模型在所有可能子集上的性能，并将模型的边际性能提升作为数据质量评估的结果。然而，精确计算数据样本的沙普利值需要穷举地训练和评估每个数据点在每个可能的数据子集上的边际性能。为避免穷举带来的高成本问题，研究者引入近似求解沙普利值方法，例如采用截断的蒙特卡罗采样^[6]和采用联邦学习的方法^[13]。Stephanie Schoch 等^[5]进一步改进了基于沙普利值的方法，并针对分类问题提出了一个用于区分数据样本对类内和类外的贡献度的评估函数 CS-SHAPLEY，以提升针对分类算法的数据质量评估。尽管采用了高效计算的方法近似地求解沙普利值，但基于沙普利值的方法还是受其高昂的计

算成本限制, 仅能应用于训练数据量较少、结构相对简单的模型。

Yoon 等^[7]提出的 DVRL (data valuation using reinforcement learning) 模型为解决基于沙普利值的方法计算复杂度高的问题, 将强化学习引入了数据质量评估。DVRL 使用一个数据选择器用于输出每个数据样本被选择的概率, 并以此概率进行数据采样以用于深度学习模型训练。将深度学习模型的损失函数及模型性能作为反馈更新优化数据选择器。最终, 将数据选择器输出的概率作为数据样本的质量评估结果。DVRL 方法使模型训练和数据质量评估能够同时进行, 减少了数据质量评估的计算成本。随着数据量及模型参数的不断增大, 该方法仍然面临较大的计算量。并且, 上述质量评估方法普遍依赖于模型下游任务的性能指标, 导致其质量评估结果更加侧重于性能指标而忽略了数据样本的整体质量。

针对现有方法计算量大的问题, 本文提出了一种新的文本数据质量评估方法。该方法从文本在语言模型嵌入空间的鲁棒性以及语言模型对文本理解的准确性两方

面衡量文本数据的质量。其中, 鲁棒性指标用于衡量数据样本被语言模型理解的稳定性, 准确性指标则用于评价数据样本被语言模型理解的程度。

2 方法框架

TDQE 由语义一致性检测模块与匹配度检测模块组成, 其模型框架如图 1 所示。首先, 对于给定的数据样本 T , 语义一致性检测模块利用一个文本摘要生成模型生成 T 的文本摘要 S 。该文本摘要生成模型需要经过 Dropout 训练, 确保其能够通过随机子网络提取 T 的嵌入表示 $V(T) = \{v_i, i \in \{1, 2, \dots, n\}\}$, 并计算 T 的鲁棒性指标 $R(V(T))$, 其中 v_i 为不同随机子网络生成的 T 的嵌入表示。其次, 匹配度检测模块使用一个文本相似匹配模型计算 T 与 S 的匹配度, 并将此作为 T 的准确性指标 $M(T, S)$ 。最终, 综合计算鲁棒性指标与准确性指标, 得到 T 的质量分数 $Q(T)$ 。

具体而言, 将待评估的数据样本 T 输入文本摘要生成模型。由于该模型使用了 Dropout 技术, 在语义一致性检测模块中, 通过不同的 Dropout 模式生成多个不同的随机子网络。每个随机子网络生成对应嵌入向量 v_i , 通过计算不同嵌入向量的平均距离来衡量 T 的鲁棒性指标。同时, 利用文本摘要生成模型生成 T 对应的文本摘要 S 。在匹配度检测模块中, 利用文本相似匹配模型计算 T 与 S 间的匹配度, 并将此作为 T 的准确性指标。 T 与 S 的匹配度越高, 语言模型对 T 的理解程度越高, 表明 T 更容易被语言模型准确理解。最终, 通过鲁棒性指标和准确性指标加权得到 T 的质量分数。TDQE 计算的文本质量分数反映的是数据样本被语言模型理解的程度,

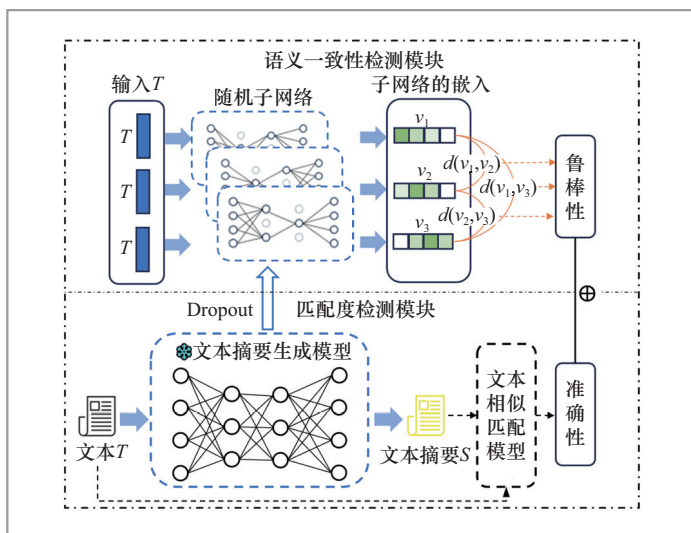


图1 TDQE 模型框架

分数越高表示数据样本越容易被语言模型理解，对语言模型训练越有利，反之亦然。

2.1 语义一致性检测模块

语义一致性检测模块通过一个文本摘要生成模型中的不同随机子网络计算数据样本的鲁棒性指标。绝大多数文本摘要生成模型训练中都会使用 Dropout 技术，目的是防止模型过拟合并提高模型泛化能力。为了实现鲁棒性指标计算，语义一致性检测模块使用不同的 Dropout 模式对模型的编码器进行随机前向传播，从而生成 m 个不同的随机子网络。每个随机子网络生成不同的嵌入向量 v_i ，将这些嵌入向量收集形成集合 $V(T) = \{v_i\}, i \in \{1, 2, \dots, n\}$ 。可以将 T 的鲁棒性指标 $R(V(T))$ 定义为式 (1)。

$$R(V(T)) = \sigma \left(-\frac{2}{m(m-1)} \sum_{i < j} d(v_i, v_j) \right) \quad (1)$$

其中， m 表示随机子网络的个数， v_i 与 v_j 表示随机子网络输出的嵌入向量， $d(v_i, v_j)$ 则表示嵌入向量对 (v_i, v_j) 的欧氏距离， $\sigma(\cdot)$ 表示 Sigmoid 函数，以保证 $R(V(T)) \in [0, 0.5]$ 。

Gal 的理论证明在神经网络上重复使用 Dropout 可以近似高斯过程的不确定性^[15]，通过多次 Dropout 生成多个嵌入向量，再利用嵌入向量间的平均欧氏距离衡量 T 的鲁棒性。 $d(v_i, v_j)$ 越大，表示嵌入向量间的变化越大，语言模型对 T 的理解的不确定性越大，反之亦然。

2.2 匹配度检测模块

匹配度检测模块通过一个文本相似匹配模型^[16]将 T 与文本摘要 S 作为输入，计算 T 与 S 的文本匹配度分数 $M(T, S) \in [0, 0.5]$ ，

并将 $M(T, S)$ 定义为 T 的准确性指标。准确性指标越高，意味着模型对文本的理解越准确，反之亦然。尽管目前已经存在多种衡量生成文本摘要质量的量化评估指标，如 ROUGE-L^[17]、BLUE^[18] 等，但上述量化指标需要提前得到数据样本的摘要，这在实际应用中是难以实现的。因此，匹配度检测模块选择利用文本相似匹配算法进行准确性指标的评估。

2.3 数据样本质量评估

本文通过设计加权鲁棒性与准确性，将数据样本 T 的质量分数 $Q(T)$ 定义如下。

$$Q(T) = R(V(T)) + M(T, S) = \sigma \left(-\frac{2}{m(m-1)} \sum_{i < j} d(v_i, v_j) \right) + M(T, S) \quad (2)$$

值得注意的是，本文提出的方法使用的文本摘要生成模型与文本相似匹配模型在模型结构上并无特殊不同。在实际实验中，也尝试过使用其他模型。不同结构的模型并不会对本文提出的方法输出的数据质量评分排名产生明显影响，但模型原本的性能差异会使鲁棒性指标与准确性指标的评估区分度有所不同。使用性能较好的模型会使质量评估的区分度较高，反之亦然。并且，本文的方法只有一个超参数 m ，整体时间复杂度为 $t = O(n^2)$ ，相较于现有方法，本文提出的方法无须训练模型，在计算开销上具备优势。

3 实验与结果分析

3.1 实验设置

实验中，首先通过训练数据微调一个预训练模型，并利用多标签文本分类任务

来验证提出的方法的有效性。具体微调参数设置见表1。

TDQE使用20NG数据集^[19]进行实验。20NG数据集是文本摘要生成任务中的常用数据集，由20个不同类别的19 000篇新闻文本组成，文本平均长度为266个词，最长文本长度为10 334个词。本文按照8：2的比例划分训练集和测试集。

3.2 对比算法及评价指标

此次研究采用目前主流的3种训练数据质量评估算法，分别为DVRL算法、Data Shapley算法和LOO算法，与本文提出的方法进行对比实验。

DVRL算法通过设计一个数据选择器输出每个数据样本被选择的概率，并以此概率进行数据采样，从而挑选数据样本用于深度学习模型训练。DVRL算法通过多轮迭代优化数据选择器，并最终将数据选择器的预测概率作为训练数据的质量分数。

Data Shapley算法通过沙普利值衡量单个数据样本对模型的性能影响，并将模型的边际性能提升作为该数据样本的质量评估结果。精确计算所有数据样本的沙普利值计算量巨大，因此Data Shapley算法采用截断的蒙特卡罗采样对该过程进行

近似。

LOO算法则通过移除一个特定的数据样本并计算模型前后性能的差异，利用模型前后性能的差异量化该数据样本的数据质量。

Data Shapley算法和LOO算法计算复杂度较高，因此对比实验分为两部分：全数据集样本对比实验和部分数据集样本对比实验。在全数据集样本对比实验中，采用20NG数据集中的全部数据样本进行实验，并将TDQE与DVRL算法进行对比。在部分数据集样本对比实验中，随机抽取4 000条20NG数据集中的数据样本进行实验，并将本文方法与DVRL算法、Data Shapley算法及LOO算法进行对比，以达到减少计算量并验证本文方法有效性的目的。

两部分对比实验均为文本的多标签分类任务实验，并通过准确率（*P*）和召回率（*R*）衡量方法有效性。

3.3 全数据集对比实验结果

本文通过TDQE计算20NG数据集中训练数据的质量评分，并从高到低进行排序；通过删除排序靠后的低质量训练数据和删除排序靠前的高质量训练数据两个实验，验证方法有效性。删除排序靠后的低质量数据能够缩短模型的训练时间，并在一定程度上提升模型的性能。而删除排序靠前的高质量数据则会使模型性能迅速下降。总之，删除低质量训练数据和删除高质量训练数据都是表现数据质量评估性能的有力指标。删除低质量数据实验的结果如图2所示。其中，纵坐标分别表示准确率和召回率，横坐标表示删除低质量数据的比例，0则是微调后的基础预训练模型的性能。

结果显示，TDQE全面优于对比算法

表1 参数设置

参数	规模
输出文本特征维度	768
输入最大长度	512
训练轮次/次	10
批处理大小/条	8
学习率	0.01
激活函数	Leaky-ReLU

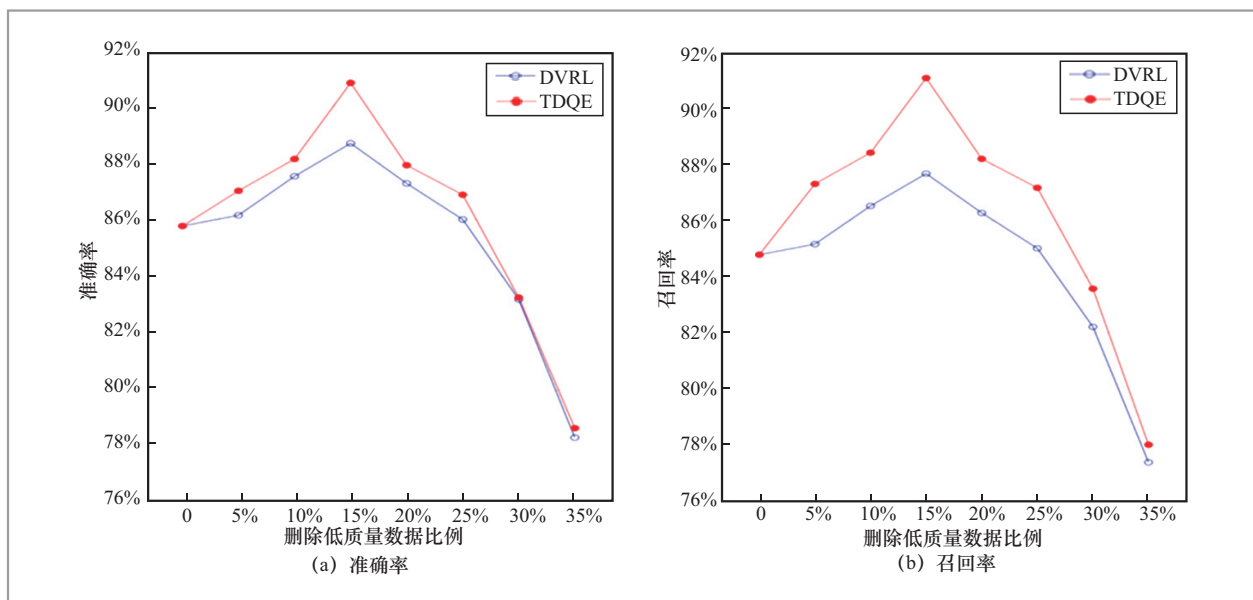


图2 全数据集对比实验中删除低质量数据的实验结果

DVRL, 并且在删除一定比例的低质量数据后, 模型性能得到明显提升 (在 15% 时性能提升达到最大)。这说明 TDQE 有效识别出了训练数据中的低质量数据。当删

除约 25% 的低质量数据后, 模型性能回落至未删除数据前的性能。

删除高质量数据的结果如图 3 所示。其中, 横坐标表示删除高质量数据的

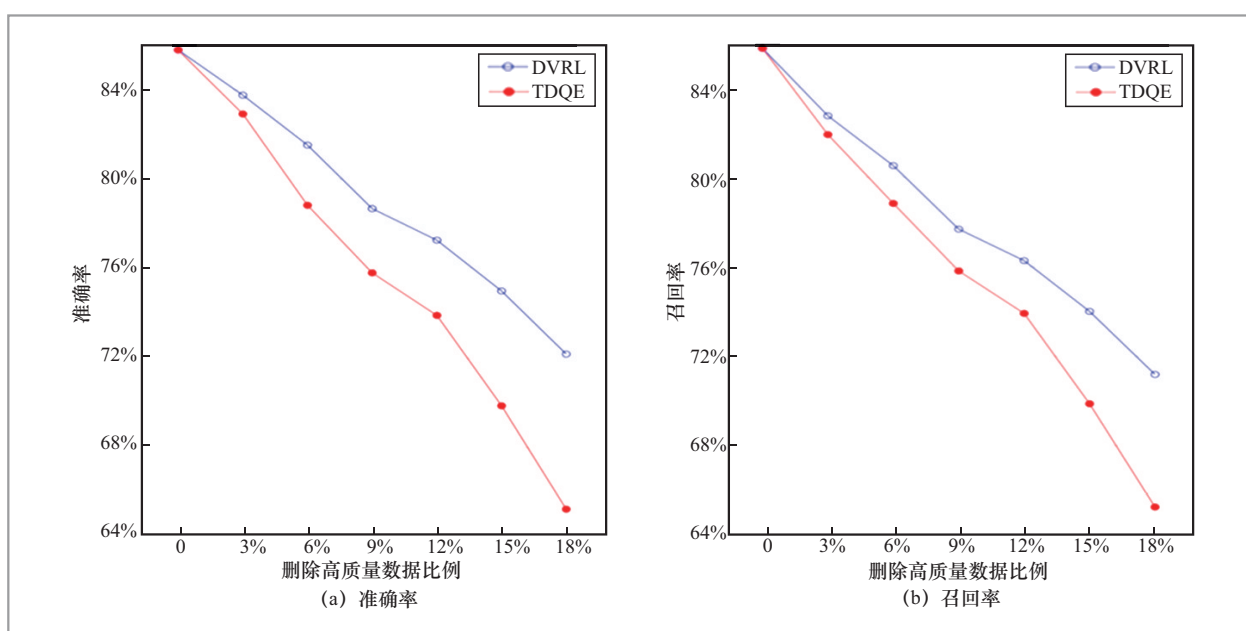


图3 全数据集对比实验中删除高质量数据的实验结果

比例。

结果表明，相较于 DVRL，TDQE 删除一定比例高质量数据后的模型性能下降更显著。这说明 TDQE 能够更有效识别训练数据中的高质量数据，并赋予较高的数据质量评分。

3.4 部分数据集对比实验结果

进一步，随机抽取 20NG 数据集的 4 000 个数据样本，并将 TDQE 与 DVRL 算法、Data Shapley 算法及 LOO 算法进行对比实验。实验同样采用删除低质量数据与删除高质量数据两种方式进行，实验结果如图 4、图 5 所示

在删除低质量数据的实验中，4 种算法性能变化趋势相似，在删除 5%~10% 低质量数据后，模型性能有一定提升；删除比例达到 15% 及以上后，模型性能下降较为严重。而在上述实验中，TDQE 均优于对比算法。

在删除高质量数据实验中，模型性能下降率越高，则表明模型识别高质量数据

能力越强。在该实验中，TDQE 大幅优于对比算法。

值得注意的是，DVRL 算法、Data Shapley 算法及 LOO 算法的计算耗时均与模型的训练耗时呈正相关，而 TDQE 则避免了模型训练，因此计算效率高于需要进行模型训练的方法。

3.5 模型泛化性实验及消融实验

本节在基于 LSTM 结构的文本分类模型^[20]上删除低质量数据进行实验，以验证 TDQE 的泛化性，实验结果如图 6 所示。

本次实验中使用 DVRL 方法对基于 LSTM 结构的文本分类模型进行重新训练，以得到针对该模型数据质量评估的排序结果。而 TDQE 则无须重复数据质量评估过程。上述实验结果显示，相较于 DVRL 算法，TDQE 对语言模型的结构不敏感。在不同的模型结构下，TDQE 仍优于 DVRL 方法，具有较好的泛化性。

在上述实验的基础上，进一步探究 TDQE 的不同组件对数据质量评估结果的

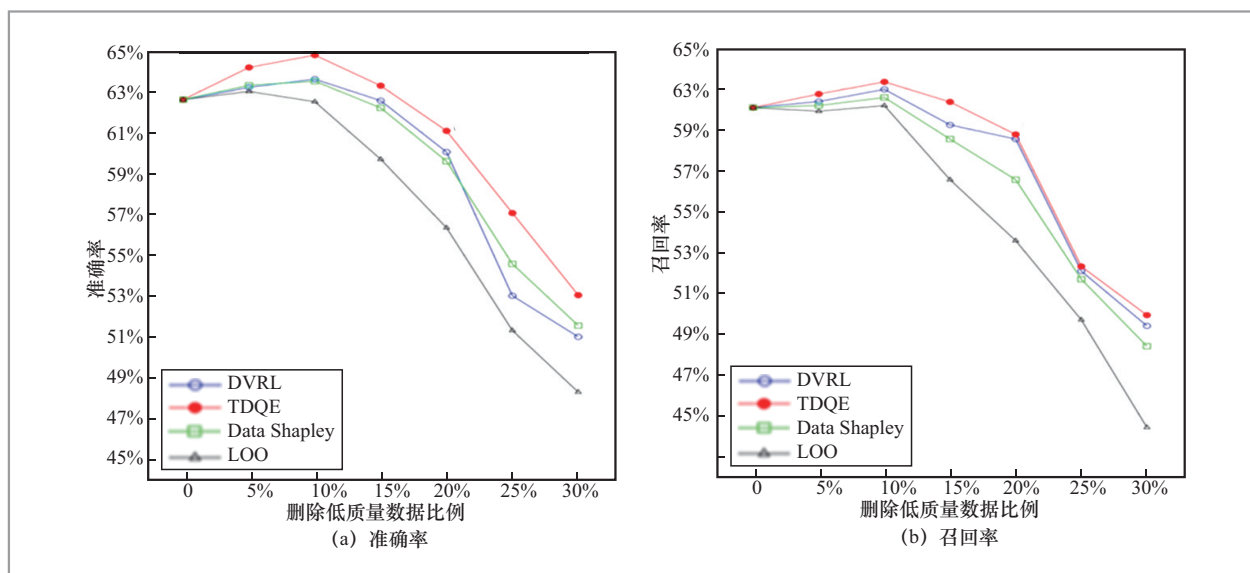


图4 部分数据集对比实验中删除低质量数据的实验结果

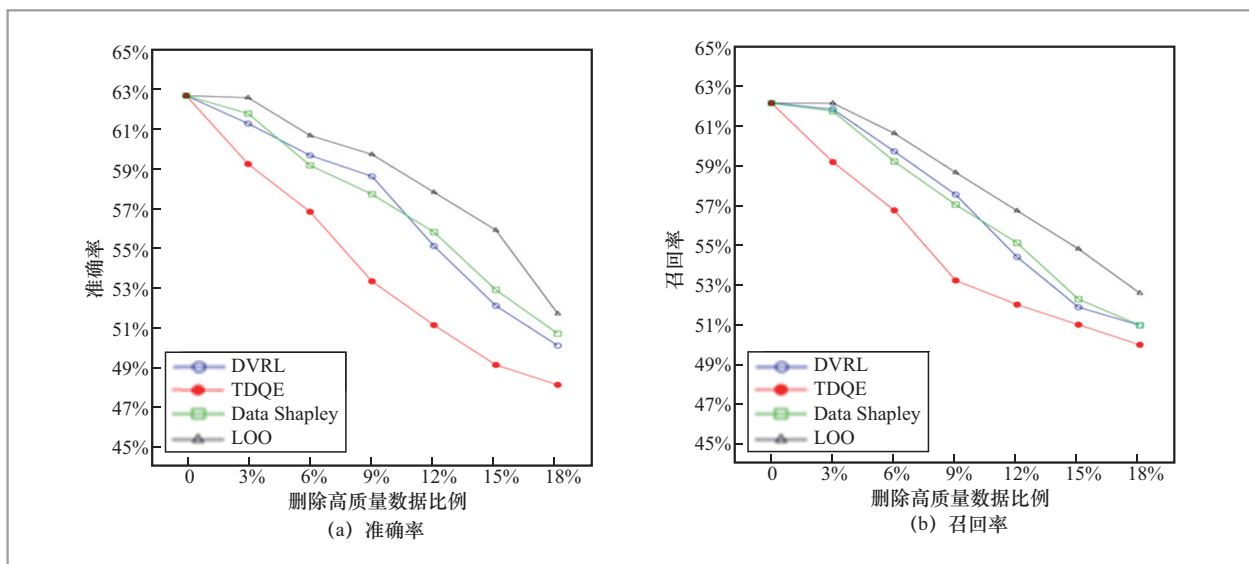


图5 部分数据集对比实验中删除高质量数据的实验结果

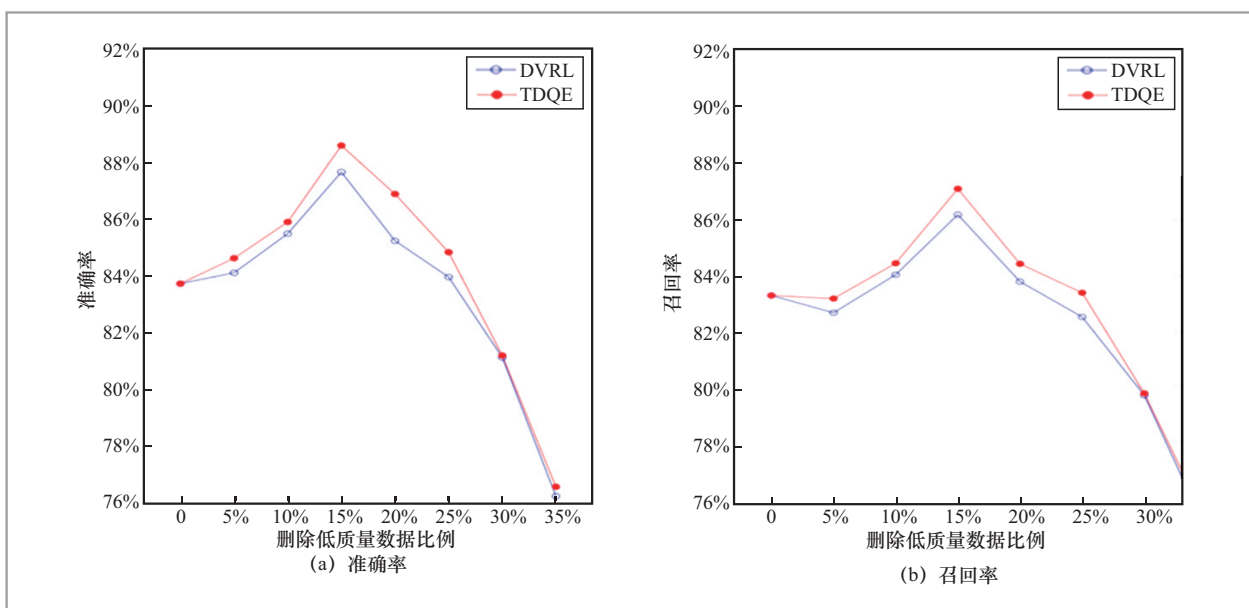


图6 在基于LSTM结构的文本分类模型上删除低质量数据的实验结果

影响，以确定鲁棒性指标与准确性指标的有效性。鲁棒性指标或准确性指标为仅用单个指标对文本数据进行质量评估，并采用删除低质量数据的实验方法所得到的实验结果，实验结果如图7所示。

由图7可知，鲁棒性指标与准确性

指标均能在一定程度上筛选出低质量数据，但综合这两个指标进行数据质量评估结果最优。因此，本文提出将鲁棒性指标与准确性指标综合进行计算的方法是有效的。

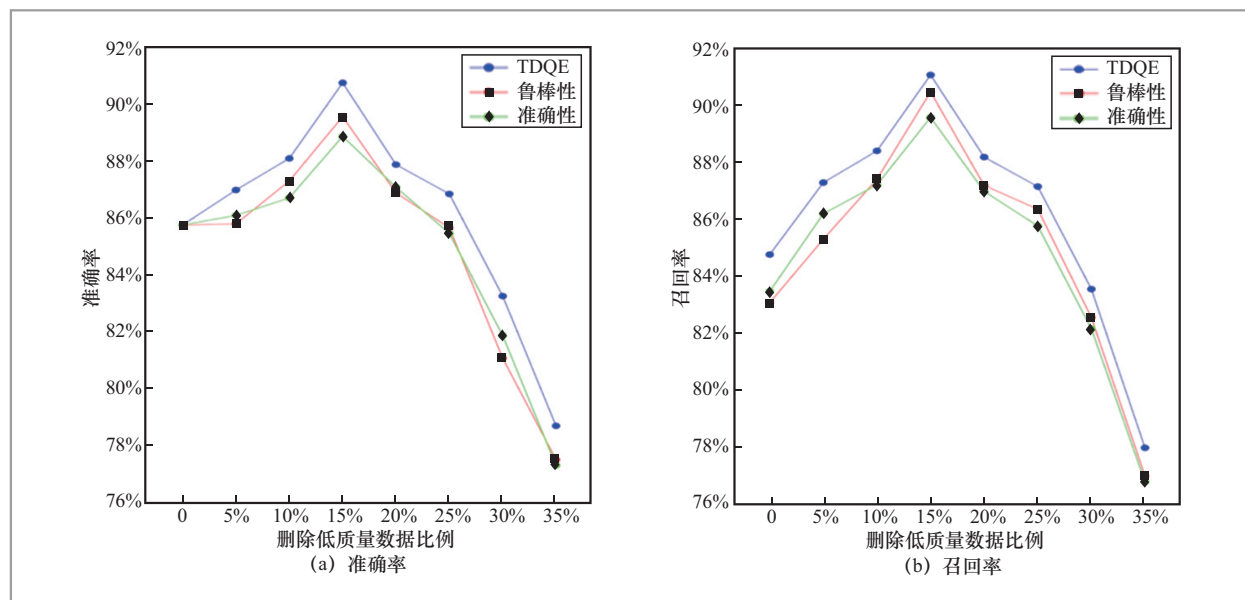


图7 消融实验结果

4 结束语

文本数据质量评估旨在预测文本数据是否适合用于语言模型。以往的研究大多基于模型性能指标评估，忽略了文本数据质量评估的整体性。本文提出了一种新的文本数据质量评估方法TDQE，该方法使用Dropout训练的文本摘要生成模型，通过测量文本摘要生成模型的随机子网络产生的嵌入变化，评估数据样本的鲁棒性；通过文本相似匹配模型计算数据样本与其摘要间的匹配度，评估数据样本的准确性。实验表明，TDQE方法的性能优于主流算法，并能够通过删除部分低质量数据实现在减少模型训练量的同时提升模型性能。

参考文献：

- [1] LONGPRE S, YAUNEY G, REIF E, et al. A pretrainer's guide to training data: measuring the effects of data age, domain coverage, quality, & toxicity [C]//Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers). Stroudsburg: ACL, 2024: 3245-3276.
- [2] FERDOWSI H, JAGANNATHAN S, ZAWODNIOK M. An online outlier identification and removal scheme for improving fault detection performance[J]. IEEE Transactions on Neural Networks and Learning Systems, 2014, 25(5): 908-919.
- [3] FRENAY B, VERLEYSEN M. Classification in the presence of label noise: a survey[J]. IEEE Transactions on Neural Networks and Learning Systems, 2014, 25(5): 845-869.
- [4] TERHÖRST P, KOLF J N, DAMER N, et al. SER-FIQ: unsupervised estimation of face image quality based on sto-

- chastic embedding robustness[C]//Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2020: 5650–5659.
- [5] SCHOCH S, XU H F, JI Y F. CS-SHAPLEY: class-wise Shapley values for data valuation in classification [C]//Proceedings of the 36th International Conference on Neural Information Processing Systems. New York: Curran Associates Inc., 2022: 34574–34585.
- [6] GHORBANI A, ZOU J. Data Shapley: equitable valuation of data for machine learning[C]//Proceedings of the 36th International Conference on Machine Learning, Long Beach: PMLR 97, 2019: 2242–2251.
- [7] YOON J S, ARIK S, PFISTER T. Data valuation using reinforcement learning [C]//Proceedings of the 37th International Conference on Machine. Stroudsburg: PMLR, 2020: 10842–10851.
- [8] ABOLFAZLI A, PALMER G, KUDENKO D. Data valuation for offline reinforcement learning[EB]. arXiv preprint, 2022, arXiv: 2205.09550.
- [9] SRIVASTAVA N, HINTON G, KRIZHEVSKY A, et al. Dropout: a simple way to prevent neural networks from overfitting[J]. The Journal of Machine Learning Research, 2014, 15(1): 1929–1958.
- [10] CHI H L, JIN W, AGGARWAL C, et al. Precedence-constrained winter value for effective graph data valuation[EB]. arXiv preprint, 2024, arXiv: 2402.01943.
- [11] KOH P W, LIANG P. Understanding black-box predictions via influence functions[C]//Proceedings of the 34th International Conference on Machine Learning – Volume 70, Sydney: JMLR, 2017: 1885–1894.
- [12] WANG Z F, ZHU H, DONG Z H, et al. Less is better: unweighted data subsampling via influence function[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(4): 6340–6347.
- [13] WANG T H, RAUSCH J, ZHANG C, et al. A principled approach to data valuation for federated learning[M]//Federated Learning: Privacy and Incentive. Cham: Springer International Publishing, 2020: 153–167.
- [14] JIA R X, DAO D, WANG B X, et al. Towards efficient data valuation based on the shapley value[C]//Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics. Naha: DBLP, 2019: 1167–1176.
- [15] SEEGER M. Gaussian processes for machine learning[J]. International Journal of Neural Systems, 2004, 14(2): 69–106.
- [16] 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: Human Language Technologies, Volume 1 (Long and Short Papers). Stroudsburg: ACL, 2019: 4171–4186.
- [17] LIN C Y. ROUGE: a package for automatic evaluation of summaries[C]//Proceedings of the Text Summarization Branches Out, Barcelona: ACL, 2004: 74–81.
- [18] PAPINENI K, ROUKOS S, WARD T, et al. BLEU: a method for automatic evaluation of machine translation[C]//Proceedings of the Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 2018: 311–318.
- [19] PEDREGOSA F, VAROQUAUX G, GRA-

MFORT A, et al. Scikit-learn: Machine Learning in Python[J]. The Journal of Machine Learning Research, 2011, 12(1/2): 2825-2830.

[20] 孙刘成, 黄润才. 融合LSTM和注意力机制的新闻文本分类模型[J]. 传感器与微系统, 2022, 41(9): 38-41.

SUN L C, HUANG R C. News text classification model fuses LSTM and attention mechanism[J]. Transducer and Microsystem Technologies, 2022, 41(9): 38-41.

作者简介



罗春旭（1974-），男，复旦大学计算机科学技术学院、上海市数据科学重点实验室博士生，主要研究方向为垂域模型、知识工程。



熊海旭（1996-），男，复旦大学计算机科学技术学院、上海市数据科学重点实验室博士生，主要研究方向为数据质量、深度学习。



叶雅珍（1985-），女，博士，复旦大学计算机科学技术学院教师，上海市数据科学重点实验室数据资产研究室主任，主要研究方向为数据科学和数字经济。



丁滢（1977-），女，中国人民解放军国防科技大学计算机学院副研究员，主要研究方向为操作系统、系统安全、机密计算。



宗世泽（1999-），男，中国人民解放军国防科技大学计算机学院工程师，主要研究方向为操作系统、系统安全、机密计算。



熊贇（1980-），女，博士，复旦大学计算机科学技术学院教授，上海市数据科学重点实验室副主任，主要研究方向为数据科学和数字经济。



朱扬勇（1963-），男，博士，复旦大学计算机科学技术学院教授、博士生导师，主要研究方向为数据科学和数字经济。

收稿日期: 2023-12-31

通信作者: 熊海旭, 22110124113@m.fudan.edu.cn