知识点文本与新闻标题文本语义相似度的计算方法

王星月，松　云

(内蒙古师范大学 计算机科学技术学院,内蒙古 呼和浩特 010022)

摘　要：针对构建融入课程思政内容的知识图谱过程中,知识点文本与思政内容文本相关性难以判定和样本分布失衡问题,提出了基于 BERT 微调的知识点-新闻相关性判断方案。首先,将文本语义相似度判断视为二分类任务,选用微调后的 BERT 模型完成分类任务;其次,引入多权重策略进行对比试验,以有效缓解正负样本量差距导致的预测偏移现象,发现 PR-AUC 的值提高了 2.85%;最后,通过对比实验证实该方法能较好地处理相关性判断问题,$F1$ 值达到 90.87%,相较于其他算法有较大优势且能有效应对小样例数目下的样本分布不平衡情况。

关键词：语义相似度；BERT 微调；类别权重；文本分类

中图分类号：TP391.1　　**文献标志码**：A　　**文章编号**：1001-8735(2026)02-0214-07

DOI:10.3969/j.issn.1001-8735.2026.02.012

知识图谱(knowledge graph, KG)是基于节点和边所形成的图结构在大数据背景下被广泛使用的知识表征形式,它能有效地将一个领域内所有事物的概念及彼此的关系进行组织,从而形成一种大规模的数据组织方式[1]。近年来,随着人工智能的发展,知识图谱被应用于各个领域。在教育领域,对提高学生的认知水平具有较大作用。知识图谱能通过可视化的呈现方式让学生更好地了解并掌握每个知识点的内容及相互之间的关系,具有传统授课方式无法达到的效果。因此,可以利用知识图谱可视化的方式让学生更直观地理解每一个知识点。万海鹏等[2]利用学科知识图谱和学习元平台技术,构建高中信息技术在线课程体系,实现个性化诊断与实时反馈。许智宏等[3]运用多模态信息融合与生成对抗网络(GAN)技术,提出多模态课程学习知识图谱实体预测模型(MMCL),提升知识图谱实体预测的准确性和训练效果。

随着教育改革的持续深入,如何培养全面发展的高素质人才,已成为当前教育实践中亟须解决的关键问题。其中,新闻作为意识形态传播的重要载体,承担着塑造社会共识的使命。将知识图谱应用于课程思政建设,能更好地实现“显性知识与隐性知识”的连接。已有研究中,学者通过搭建基于知识图谱的课程思政资源库[4],采用相关度计算方法进行知识实体与思政内容之间关系的确定,但该方法需要人工设定阈值来判断二者之间是否相关。同时,在判断新闻和知识点是否相关时,大多数的新闻都与知识点无关。针对以上问题,本文提出在课程思政内容融入知识图谱的过程中,开展基于 BERT 微调的知识点-新闻相关性判断研究。首先将文本相关度计算转变为二分类任务;其次提出多种权重方案来缓解正负样本不平衡的问题;最后微调 BERT 模型提高识别的准确率。本文旨在为构建高质量、高关联度的课程思政知识图谱提供核心技术支持,进而助力教师高效备课,实现专业知识传授与价值引领的有机融合。

1　相关工作

1.1　文本语义相似度判断研究

根据场景和领域的不同,关于文本相似度的定义也并不相同,有的偏重表面字符相似,有的更加注重语

收稿日期：2025-07-19

基金项目：内蒙古自治区自然科学基金资助项目“基于标签迁移和神经网络的蒙古文领域术语识别研究”(2025MS06015)。

作者简介：王星月(2001－),女,在读硕士研究生。

通信作者：松　云(1973－),女,教授,博士,主要从事蒙古文信息处理、多语种教育智能化研究,E-mail:1348012800@qq.com。

义的相似性[5]。目前最典型的方法有三类，即基于字符串的方法、基于统计的方法和基于深度学习的方法。

基于字符串的方法是通过对原始文本进行多维度比对，以文本之间的匹配程度作为衡量相似性的依据，具有直观、易理解且实现门槛低的特点。主要包括编辑距离(levenshtein distance，LD)[6]、最长公共子序列(longest common sequence，LCS)[7]、余弦相似度(cosine similarity)和 Jaccard 相似度[8]等方法。邵清等[9]提出一种基于改进编辑距离和相似度的汉字字符串近似匹配算法，解决中文字符串匹配精度低的问题。LD 算法计算虽然比较准确，但计算时间长、普适性不足。LCS 算法则通过求解两个文本中最长公共子序列的长度，并结合原始文本长度进行归一化，以此评估文本间的相似性。张胜楠[10]在 LD 算法的基础上，融合 LCS 算法与最长公共子串算法(longest common substring，LCCS)，提出一种改进的字符串相似度计算方法。该研究从数据结构层面优化了 LD 和 LCS 的求解算法，显著降低了空间复杂度，提升计算的准确性和普适性。LCS 算法在长文本相似度计算上表现欠佳，其更适用于短文本相似度计算。余弦相似度通过将文本转换为向量空间的点，然后计算两点之间的余弦值，越接近 1 则表示相似度越高。许浩等[11]提出基于余弦相似度的英语作文评分算法，可以更高效地判断学生的作文是否跑题。然而余弦相似度无法捕捉词序信息对语义的影响。鲜翠琼等[12]提出图文组合的相似度算法，该算法引入 Jaccard 相似系数对余弦相似度进行改进，提升文本语义匹配的准确性，同时利用感知哈希算法计算图片相似度，最后将二者相似度加权结合得到整体图文相似度。

基于统计的方法通过分析文档的词汇评论衡量文本间的相似性，能够有效捕捉上下文语境信息，更准确反映文本的语义内容，因此，该方法被认为是对字符串方法的改进和扩展[13]。其代表主要包括向量空间模型(vector space model，VSM)[14]和主题模型(topic model)。VSM 将每个文档表示为一个向量，每个维度对应词汇表中的一个词，文档向量的每个元素通常表示该词在文档中的某种统计值，如词频(term frequency，TF)或逆文档(inverse document frequency，IDF)权重。杨宏伟等[15]提出了一种基于 TF-IDF 加权的文本语义相似度算法，应用于变电站一键顺控测试中，显著提升测试效率和准确率。VSM 充分利用了词语在文本中的频率和权重信息，直观易懂、灵活调节。但在处理大规模数据时，VSM 会生成高维稀疏矩阵，导致计算成本增加。同时，忽略词语间的依赖信息，可能导致相似度计算不准。主题模型是一种用于发现文档集合或语料库中隐藏主题结构的统计模型，在 VSM 的基础上引入对文档内容的深层次理解和处理，提供了更加丰富和灵活的方法。最常用的主题模型是潜在狄利克雷分配[16](latent dirichlet allocation，LDA)。LDA 将每篇文档表示为若干主题的概率分布，并且每个主题被表示为词项的概率分布，进而计算文本相似度。程蔚等[17]提出了一种基于双语 LDA 模型的跨语言文本相似度计算方法，使用双语平行语料库构建双语 LDA 模型预测新语料的主题分布，并将文档映射至同一主题向量空间；同时采用改进的主题频率-逆文档频率(ITFIDF)方法计算特征主题权重，结合余弦相似度计算双语文档间的相似度。

基于深度学习的方法利用神经网络模型强大的表达能力和自动特征学习能力，捕捉文本的深层次语义信息[13]，是现代自然语言处理(natural language processing，NLP)领域的重要研究方向。主要包括 Word2Vec(word to vec)和 GloVe(global vectors)[18]等词向量模型。Word2Vec 是最早广泛使用的词嵌入模型之一，主要包括跳元模型(skip-gram，SG)[19]和连续词袋模型(continuous bag of words，CBOW)[20]两种。其中，SG 模型基于当前中心词预测其上下文词汇；CBOW 模型则相反，通过上下文词汇预测中心词。崔洁[21]提出结合 TF-IDF 加权的 Word2vec 文本相似度算法，解决传统算法无法反映词性和词项权重的问题。孙志远等[22]提出一种结合词向量与聚类的加权 Word2Vec 算法，用于改进移动营销领域短文本相似度计算效果，通过构建语义簇特征空间，提升文本分类性能。随着深度学习技术的不断发展，基于预训练模型的方法逐渐成为计算文本相似度的主流。常见的预训练模型包括语言模型嵌入表示方法(embeddings from language models，ELMo)[23]、生成式预训练模型(generative pre-training，GPT)[24]和 BERT(bidirectional encoder representations from transformers)预训练模型[25]。ELMo 通过为每个词生成包含前后向信息的动态向量表示，对这些向量进行加权平均，得到文本的整体语义表示，进而完成文本语义相似度计算。虽然 ELMo 能够捕捉词语的多义性和复杂的语义关系，但计算复杂度较高。GPT

在 Transformer 解码器的基础上,预训练一个用于文本表示的序列语言模型,应用于下游任务。BERT 使用双向 Transformer 编码器预训练大规模语料库,生成上下文敏感的词嵌入,通过将一对文本拼接后输入全连接层,计算文本语义相似概率。李伊仝等[26]提出在融合新闻标题信息基础上,结合 TextRank 算法、LDA 模型与 BERT 预训练模型构建综合计算模型,有效解决了文本间长度差异较大、信息缺失导致的相似度计算不准确问题。

1.2 BERT 模型

ELMo 选用双层双向 LSTM 对上下文进行编码,生成上下文相关的词向量,但其表示依赖于特定任务的模型架构。GPT 则基于单向 Transformer 结构,从左到右进行上下文建模,具备任务无关的预训练能力,适用于多种下游任务。BERT 结合了 ELMo 的双向上下文建模能力和 GPT 的任务无关特性,采用多层双向 Transformer 编码器,能够同时捕捉上下文中左右两侧的语义信息,在大多数自然语言处理任务中仅需微调即可取得优异性能。

BERT 是一种基于 Transformer 架构的深度双向预训练语言模型,预训练阶段包括两个任务:掩蔽语言模型(masked language modeling,MLM)和下一句预测(next sentence prediction,NSP)。这两个任务协同作用,帮助模型学习上下文感知的词向量表示和跨句子的语义关系。其中,MLM 通过随机掩蔽文本中的词元,利用双向上下文信息以自监督方式预测掩蔽词元;但该任务无法显式对文本对之间的逻辑关系进行建模,因此引入 NSP 任务以理解两个文本序列之间的关联。

在下游任务中,BERT 通过对预训练模型进行微调以匹配特定任务。以文本语义相似度计算为例,BERT 将两个待比较的文本拼接后输入模型,经过多层双向 Transformer 编码器对文本进行深层次语义建模,利用[CLS]标记的隐藏状态或句子嵌入来表征两个文本的语义相似度,这一过程主要依赖于 Transformer 编码器的核心架构。编码器由多个相同的层叠加构成,每个层包含两个子层(sublayer):第一个子层是多头自注意力,第二个子层是基于位置的前馈网络。

为使用相同的注意力机制学习不同的行为,将不同的行为作为知识组合,并采用多头注意力机制,以捕获序列内不同依赖范围的语义关联。其核心思想是将查询向量(query,Q)、键向量(key,K)和值向量(value,V)分别通过 h 组不同的线性变换矩阵 $W_i^{(q)}$,$W_i^{(k)}$,$W_i^{(v)}$,映射到低维空间,形成 h 个独立的注意力头。每个注意力头 $h_i(i=1,2,\cdots,h)$的计算方式为 $h_i=f(W_i^{(q)}q, W_i^{(k)}k, W_i^{(v)}v)$,其中 f 表示注意力函数,通常定义为

$$f(Q,K,V)=\mathrm{softmax}(\frac{QK^T}{\sqrt{d_k}})V。$$

随后,将所有头的输出拼接,并通过一个可学习的线性变换矩阵整合,得到最终的多头注意力输出 $\mathrm{MultiHead}(Q,K,V)=\mathrm{Concat}(h_1,h_2,\cdots,h_h)W^O$。

基于位置的前馈网络对所有位置的序列表示进行独立变换。该网络引入了多层感知机与非线性变换,旨在增强模型的整体表达能力。用 Z 表示多头注意力的输出,则两层全连接神经网络的通常形式为 $\mathrm{FFN}(Z)=\max(0,ZW_1+b_1)W_2+b_2$,其中,$W_1$、$W_2$ 是权重矩阵,b_1、b_2 是偏置项。由于 Transformer 本身不包含递归或卷积结构,因此无法捕捉序列中词与词之间的位置信息。为弥补这一缺陷,引入位置编码为模型提供序列顺序信息,将词嵌入与位置编码相加作为输入。在每个子层周围添加一个残差连接,缓解梯度消失问题,加速模型收敛。同时在残差连接之后进行层归一化,使模型更稳定。

2 知识点-新闻二分类模型

本研究对新闻与知识点的相关性进行分析,旨在通过计算知识点文本与新闻标题文本的语义相似度,从大规模新闻标题数据中筛选出相关新闻条目,将任务形式化为二分类问题。最终应用目标是为“计算机网络”等专业课程,精准匹配蕴含思政内容的新闻内容,采用微调 BERT 模型实现端到端分类。实际场景中存在大量与知识点无关的新闻,导致数据集呈现类别不平衡现象。为此,本研究提出系统性解决方案,设计多权重实验框架,通过实验确定最优的类别权重方案,以提升模型对正样本的识别能力。

模型主要由五部分构成，分别为文本输入层、嵌入层、加和输出层、BERT 模型层和输出分类层。首先，将知识点文本与新闻标题文本输入，为确保后续数据的完整性对数据进行预处理，删除缺失值。随后，将输入的文本拆分为更小的单元(token)，在分词过程中，添加表示序列开始的特殊标记[CLS]和表示序列结束或不同文本段的分隔符[SEP]，并通过[SEP]拼接知识点文本和新闻标题文本。

为将离散的文本符号转换为连续的向量表示，BERT 模型使用嵌入层将数据转化为结构化的表示。首先，文本经分词处理后，通过词嵌入(token embedding)记录词汇的语义信息，每个子词映射为固定维度的向量；然后，利用片段嵌入(segment embedding)将输入文本的不同段落标记。在句子对任务中，将知识点文本标记为 0、新闻标题文本标记为 1，不同的句子使用[SEP]分隔，[CLS]标记用于分类任务；最后，利用位置嵌入(position embedding)的绝对位置编码捕捉序列顺序，使模型感知词语在序列中的位置信息。将 Token Embedding、Segment Embedding 和 Position Embedding 三种嵌入结果相加，作为 BERT 模型层的输入。

输入 BERT 模型的数据，经过 12 层 Transformer 编码器输出富含上下文信息的隐藏状态序列。位于输入序列起始位置的特殊标记[CLS]对应的向量，会逐步学习到整个输入序列的综合语义表示，具有良好的全局表征能力，适用于各类分类任务。为适应下游任务，Dense 全连接层通过线性变换将高维特征转化为目标维数的特征向量，实现降维和下游任务匹配。通过 Dropout 层随机丢弃一些神经元，降低过拟合风险。利用 softmax 函数进行归一化，转化为分类概率，其计算公式为 $P=\mathrm{softmax}(Wh_{\mathrm{d}}+b)$，其中，$W$ 为权重，h_{d}为 Dropout 层输出，b 为偏置。最后利用交叉熵损失函数更新模型参数。

为解决类别不平衡问题，设计多权重对比实验，探究不同权重策略对模型性能的影响。在知识点文本与新闻标题文本关联二分类任务中，通常存在正类样本远少于负类样本的现象，导致未经调整的模型倾向于预测负类样本，进而忽视正类样本，造成正类样本 $F1$ 值等关键指标降低。本文设计多组权重方案，系统化测试不同权重策略对模型的影响，量化评估各类权重对多维度指标的提升效果。

3　结果分析

3.1　数据来源

鉴于目前针对知识点文本和新闻标题文本相关性分析的公共数据集较为缺乏，本文首先选用领域专家定义的“计算机网络”课程知识点共计 56 个；随后利用爬虫技术从新华网、人民网等权威网站以知识点为关键词获取包含新闻标题、发布时间及 URL 在内的有效样本 1 477 条；最后对数据集进行人工标注，用标签 0 表示知识点和新闻不相关，标签 1 表示知识点和新闻相关，只要新闻中直接出现知识点或与知识点相关内容均可视为与新闻相关；完成标注后，将数据集划分为训练集和测试集，数据集分布情况见表 1。

表 1　数据集分布

Tab. 1　The distribution of datasets

类别及总数	训练集	测试集
1	393	98
0	809	213
总数	1 202	301

3.2　参数设置

实验处理器为 AMD Ryzen 7 7840HS 和 Radeon 780M Graphic，采用 pytorch2.4.1 深度学习框架，编程语言采用 Python3.8。实验超参数设置见表 2。

3.3　评估指标

本文选用精确率(precision，P)、召回率(recall，R)和 $F1$ 值作为模型的评价指标。在评估分类任务中通常选用 ROC-AUC，绘制假阳性率作为横轴，真阳性率作为纵轴的 ROC 曲线，通过计算曲线下的面积衡量模型整体判别能力，值越接近 1，代表模型性能越好。然而在研究知识点文本和新闻标题文本分类任务中更加关注正类样本，因此本文进一步将精确率-召回率下面积(PR-AUC)作为主要评价指标，PR-AUC 以 R 为横轴、P 为纵轴，能够更直观反映模型在不同召回水平下的精确性表现。该指标在数据不平衡情况下更具敏感性，值越接近 1，说明模型在识别正类样本方面的性能越强。具体计算公

式为

$$P=\frac{TP}{TP+FP},R=\frac{TP}{TP+FN},$$

$$F1=\frac{2PR}{P+R},\mathrm{PR-AUC}=\int_0^1 P(R^{-1}(r))\mathrm{d}r。$$

其中,TP 表示被正确识别的正类样本数,FN 和 FP 分别表示被错误识别的负类样本数和正类样本数。

3.4 结果分析

3.4.1 *多权重实验* 进行多权重实验,目的在于系统化解决类别不平衡问题对模型性能的影响。在真实场景中,新闻标题与知识点的关联数据普遍存在类别不均衡现象,由表 1 可知,正类样本共计 491 条,负类样本共计 1 022 条,直接训练会导致模型偏向负类样本而忽视正类样本。通过设计无权重、基础权重、1.5 倍加权、2 倍加权多组对比方案,将不同的权重方案直接作用于损失函数,改变模型对不同类别样本的关注程度,验证类别加权策略对缓解失衡问题的有效性。同时探索正负样本权重比的阈值效应,为实际应用提供可复用的调参策略,最终实现模型在保持整体准确率的同时,提升对正类样本的识别能力。实验结果见表 3。

表 2 配置信息

Tab. 2 The configuration information

超参数名称	内容
MAX_LENGTH	128
Batch_size	16
Epoch	10
Learning_rate	0.000 02
Weight_decay	0.01
Warmup_step	100

表 3 多权重实验结果对比

Tab. 3 The comparison of results of multi-weighting experiments

权重方案	P-0	R-0	$F1$-0	P-1	R-1	$F1$-1	P	R	$F1$	PR-AUC
无权重	92.68	93.60	93.14	86.46	84.69	85.57	89.57	89.15	89.36	86.38
基础权重	93.66	94.58	94.12	88.54	86.73	87.63	91.10	90.65	90.87	89.23
1.5倍权重	90.09	94.09	92.05	86.52	78.57	82.35	88.30	86.33	87.20	86.61
2倍权重	90.87	93.10	91.97	84.95	80.61	82.72	87.91	86.85	87.34	85.72

分析表 3 实验结果可知,与无权重方案相比,基础权重方案不仅有效提升了对正类样本的识别能力,同时还保持了对负类样本良好的识别能力。表明模型在处理类别不平衡问题时,能够较好地平衡两类样本的识别效果,避免出现过度偏向某一类样本的情况。综合对比精确率(P)、召回率(R)、$F1$ 值及 PR-AUC 四大评价指标,基础权重方案表现最优,尤其在衡量不平衡样本场景下模型性能的 PR-AUC 指标上,基础权重方案取得 89.23% 的最高值,进一步证实了其在缓解类别不平衡问题上的有效性与优越性。1.5 倍权重和 2 倍权重实验结果表明,正类样本识别能力下降的同时负类样本的识别也受到影响,说明过高的正类样本权重会干扰模型对负类样本的学习。尽管 1.5 倍权重的 PR-AUC 值较无权重方案有所提升,但 P、R、$F1$ 值均有所下降。因此选取基础权重实现类别平衡,既避免了无权重时模型对负类样本的过度偏向,也防止了高权重设置下模型对正类样本的过度拟合,使得模型在保持分类边界清晰度的同时,有效提升正类样本的识别覆盖面。在实际应用中,该方案能尽可能地筛选出更多真正相关的思政新闻内容,有效降低教师在思政教学案例库构建过程中的人工筛选成本,提高了知识图谱构建的效率。

3.4.2 *对比实验* 为验证模型的有效性,将其与三类不同维度的模型进行对比实验,三类模型涵盖传统方法、特征融合及序列建模方法;实验过程中采用相同的训练集、测试集与验证集,以确保实验的可比性与公平性,不同模型的实验结果见表 4。

表 4 不同模型的对比实验结果

Tab. 4 The comparative experimental results of different models

模型	P	R	$F1$
SVM+TF-IDF	65.24	63.32	63.87
BERT+TF-IDF	87.34	87.91	87.61
BERT-BiLSTM	90.18	88.35	90.09
BERT	91.10	90.65	90.87

(1)SVM+TF-IDF 模型:采用 TF-IDF

算法从“知识点”和“新闻标题”拼接的文本中提取特征，结合 SVM 分类器进行关联判断，通过设置 class_weight=‘balanced’参数缓解数据不平衡问题。

(2)BERT＋TF-IDF 模型：采用特征融合架构，旨在综合深度语义理解与传统统计特征的优势。使用相同参数的 BERT 模型，同时采用 TF-IDF 方法提取文本的关键词统计特征。将两类特征拼接形成融合特征向量，最后通过全连接神经网络进行分类。

(3)BERT-BiLSTM 模型：采用与本文模型相同参数的预训练 BERT 模型提取文本嵌入特征，后续接入 BiLSTM 网络，进一步捕捉序列中的长距离依赖关系和上下文信息，独立处理“知识点”与“新闻标题”双序列输入，通过序列建模增强上下文关联能力。

从表 4 可知，传统 SVM＋TF-IDF 方法因受限于数据不平衡和特征表达能力，$F1$ 值仅 63.87%，效果不理想。融合预训练 BERT 语义特征与 TF-IDF 统计特征的模型将 $F1$ 值提升至 87.61%，验证了多模态特征融合的有效性，但仍未挖掘二者互补性。进一步采用双向 LSTM 捕捉序列依赖的 BERT-BiLSTM 模型，凭借对双输入序列知识点文本与新闻标题文本的上下文关联建模能力，$F1$ 值达 90.09%。然而，通过对比可知，直接使用微调后的 BERT 模型在 P、R 和 $F1$ 值上均展现出更优的性能，这表明微调后的 BERT 模型在数据不平衡场景下具有突出的鲁棒性。说明额外增加特征或复杂网络结构，可能会给模型引入噪声、增加模型复杂性，反而导致分类性能下降。

4　结语

本文针对教育领域知识图谱构建中，新闻与知识点相关性判断的关键难题，提出一种基于 BERT 微调的二分类模型，并通过多权重实验，系统化解决样本不平衡问题。实验结果表明，基础权重方案在保持负类样本识别能力的同时，提升正类样本的识别能力，在 PR-AUC 指标上表现优异，验证了类别加权策略对失衡数据的优化效果。此外，通过对比试验可知，相较于传统 SVM＋TF-IDF 方法及特征融合模型 BERT＋TF-IDF、序列建模模型 BERT-BiLSTM，BERT 微调模型在精确率、召回率和 $F1$ 值方面均表现最优，体现了其强大的上下文语义建模能力和鲁棒性。

参考文献：

[1] 王启飞，刘昊霖，王俊龙，等．安全政策知识图谱的构建与推理技术[J]．安全与环境学报，2025，25(11)：4346-4356.

[2] 万海鹏，成玲娜，程玉梅．基于学科知识图谱的信息技术在线课程设计研究[J]．中国教育信息化，2023，29(8)：121-128.

[3] 许智宏，郝雪梅，王利琴，等．多模态课程学习知识图谱实体预测方法研究[J]．计算机科学与探索，2024，18(6)：1590-1599.

[4] 李梦楠．基于知识图谱的课程思政资源库建设研究[D]．秦皇岛：燕山大学，2024.

[5] 魏嵬，丁香香，郭梦星，等．文本相似度计算方法综述[J]．计算机工程，2024，50(9)：18-32.

[6] LEVENSHTEIN V I. Binary codes capable of correcting deletions, insertions and reversals[J]. Soviet Physics Doklady, 1966, 10: 707-710.

[7] SANKOFF D. Matching sequences under deletion-insertion constraints[J]. Proceedings of the National Academy of Sciences of the United States of America, 1972, 69(1): 4-6.

[8] JACCARD P. The distribution of the flora of the alpine zone[J]. New Phytologist, 1912, 11(2): 37-50.

[9] 邵清，叶琨．基于编辑距离和相似度改进的汉字字符串匹配[J]．电子科技，2016，29(9)：7-11.

[10] 张胜楠．基于编辑距离的字符串相似度算法研究[J]．现代计算机，2023，29(14)：23-26.

[11] 许浩，周亚萍，赵亚慧．基于余弦文本相似度计算的英语作文评分算法的应用研究[J]．教育教学论坛，2018(6)：255-256.

[12] 鲜翠琼，秦学，朱道恒，等．一种图文组合相似度算法的设计与优化[J]．软件工程，2020，23(8)：9-12.

[13] 李莹，伍胜，徐聪，等．语义文本相似度计算方法研究综述[J]．软件导刊，2024，23(11)：1-11.

[14] SALTON G, WONG A, YANG C S. A vector space model for automatic indexing[J]. Communications of the ACM, 1975, 18(11): 613-620.

[15] 杨宏伟，张红梅，张骥，等．基于 TF-IDF 加权文本语义相似度算法的变电站一键顺控测试方法研究[J]．电力科学与技术学报，2023，38(5)：269-278.

[16] BLEI D M, NG A, JORDAN M I. Latent dirichlet allocation[J]. The Journal of Machine Learning Research, 2003(3): 993-1022.
[17] 程蔚,线岩团,周兰江,等. 基于双语 LDA 的跨语言文本相似度计算方法研究[J]. 计算机工程与科学,2017,39(5):978-983.
[18] 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(EMNLP). Doha:Association for Computational Linguistics,2014:1532-1543.
[19] MIKOLOV T, SUTSKEVER I, CHEN K, et al. Distributed representations of words and phrases and their compositionality[J]. Advances in Neural Information Processing Systems,2013:arXiv. 1310. 4546.
[20] MIKOLOV T,CORRADO G,CHEN K,et al. Efficient estimation of word representations in vector space[J]. Computer Science,2013: arXiv. 1301. 3781.
[21] 崔洁. 基于加权 word2vec 算法的文本相似度研究[J]. 电子测试,2021,(21):53-55.
[22] 孙志远,王伟,马迪,等. 移动营销领域的文本相似度计算方法[J]. 计算机应用,2017,37(S1): 292-294.
[23] PETERS M, NEUMANN M, IYYER M, et al. Deep contextualized word representations[C]//Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. New Orleans: Association for Computational Linguistics, 2018: 2227-2237.
[24] RADFORD A, NARASIMHAN K, SALIMANS T, et al. Improving language understanding by generative pre-training [J]. Computer Science,2018,3(2):324-336.
[25] 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. Minneapolis: Association for Computational Linguistics, 2019: 4171-4186.
[26] 李伊仝,王红斌,程良. 融入新闻标题信息的新闻文本与评论的语义相似度计算方法[J]. 吉林大学学报(理学版),2022,60(6):1399-1406.

A Method for Calculating Semantic Similarity Between Knowledge Texts and News Headlines

WANG Xingyue, Songyun

(*College of Computer Science and Technology*, *Inner Mongolia Normal University*, *Hohhot* 010022, *China*)

Abstract: To address the difficulty in determining the relevance between knowledge texts and ideological education content during the construction of knowledge graphs that incorporate curriculum-based ideological education, as well as the issue of imbalanced sample distributions, this paper proposed a BERT fine-tuning-based approach for judging the relevance between knowledge points and news. First, the semantic similarity between texts was treated as a binary classification task, and a fine-tuned BERT model was employed to perform the classification. Second, a multi-weight strategy was introduced for comparative experiments to effectively mitigate prediction bias caused by differences in positive and negative sample sizes, resulting in a 2.85% increase in PR-AUC. Finally, comparative experiments demonstrated that the proposed method performed well in relevance judgment, achieving an $F1$ score of 90.87%, which significantly outperformed other algorithms and can effectively handle situations with a small number of samples and imbalanced distributions.

Key words: semantic similarity; BERT fine-tuning; class weight; text classification

【责任编辑 闫立华】