

# 基于类别主题词集的加权相似度短文本分类

王小楠, 黄卫东

(南京邮电大学 管理学院, 江苏 南京 210003)

**摘要:** 由于短文本存在特征稀疏的问题, 在分类问题上效果不佳, 该文充分利用词向量模型, 在词层面提出一种基于类别主题词集的加权相似度的短文本分类算法。首先训练词向量模型, 其次使用 TF-IDF 选择出最能代表各类别的主题词形成类别主题词集, 将短文本的关键词与各类别主题词分别进行相似度计算, 将类别主题词对主题的贡献度表示在权重中, 选择相似度最高的结果作为该短文本的类别。实验结果表明, 基于类别主题词集的加权相似度短文本分类方法在精确率上相较 KNN 算法、Logistic 回归算法、决策树分类算法分别提高了 2.9%、1.8%、10.2%; 在召回率上分别提升了 3.0%、1.7%、10.4%。但是类别主题词对类别的贡献度量维度简单。基于主题词集的加权相似度短文本分类算法在词的层面解决了短文本分类中的特征不足的问题, 提高了短文本分类的性能。

**关键词:** Word2Vec; 短文本分类; 相似度; 类别主题; 加权

中图分类号: TP391.1

文献标识码: A

文章编号: 1673-629X(2022)09-0095-05

doi:10.3969/j.issn.1673-629X.2022.09.015

## Short Text Classification with Weighted Similarity Based on Category Topic Word Set

WANG Xiao-nan, HUANG Wei-dong

(School of Management, Nanjing University of Posts and Telecommunications, Nanjing 210003, China)

**Abstract:** Due to the problem of sparse features of short text, it is not effective in classification. We make full use of the word vector model and propose a short text classification algorithm based on the weighted similarity of the category topic word set at the word level. Firstly the word vector model is trained. TF-IDF is used to select the subject words that can best represent each category to form the category subject word set. The similarity between the keywords of the short text and the subject words of each category is calculated respectively. The contribution degree of the category subject words to the topic is expressed in the weight, and the result with the highest similarity is selected as the category of the short text. The experiment shows that the precision of the short text classification method based on the weighted similarity of the category topic word set is 2.9%, 1.8%, and 10.2% higher than that of the KNN algorithm, the Logistic regression algorithm, and the decision tree classification algorithm respectively. The recall rate increased by 3.0%, 1.7%, and 10.4% respectively. The metric dimension of the contribution of topic words to category is simple. The short text classification algorithm based on the weighted similarity of the topic word set solves the problem of insufficient features in short text classification at the word level, and improves the performance of short text classification.

**Key words:** Word2Vec; short text classification; similarity; category topic; weighting

## 0 引言

在互联网快速发展的信息时代, 各主流平台每天都会产生数以万计的信息, 其中短文本的数量更是数不胜数。因此对短文本的研究有非常长远的意义和广阔的前景。对于文本处理的技术也越来越先进。

短文本分类是自然语言处理的一部分, 广泛应用于数据挖掘、知识检索、情感分类等领域。针对短文本分类的方法, 有统计的方法和机器学习的方法, 深度学

习近年来在自然语言处理领域也发挥了强大的作用。但是短文本分类最大的问题在于长度短, 特征数量少, 特征稀疏, 提取短文本有用的特征才是对短文本分类最大的挑战<sup>[1]</sup>。针对这个问题, 有很多的学者进行探索, 都试图去扩展短文本的特征来进行短文本分类。该文在词层面上, 没有对短文本进行扩展, 而是充分利用词向量来计算词语间的语义信息, 对短文本进行分类。

收稿日期: 2021-10-15

修回日期: 2022-02-16

基金项目: 国家自然科学基金项目(7217011293); 国家社会科学基金重大项目(16ZDA054); 江苏省研究生科研创新计划(KYCX21\_0836)

作者简介: 王小楠(1997-), 女, 研究生, 研究方向为网络舆情; 黄卫东, 教授, 博士, 研究方向为应急管理、网络舆情。

## 1 相关研究

对于短文本的分类,包括传统意义上的统计学习方法和深度学习方法。

赵晓平用 TF-IDF 提取短文本中频率为前  $N$  的词语进行 Word2Vec 向量表示,再计算文本空间距离进行分类<sup>[2]</sup>;TF-IDF 算法解决了短文本分类中外部语料依赖的问题,但在计算文本特征时存在权重集中和文本区分度低的问题。因此,Duan 提出了一种基于卡方统计和 TF-IWF 算法的短文本分类方法,在准确率、召回率、F 值上均有提高<sup>[3]</sup>。Zhou 提出了一种基于语义扩展的短文本算法,通过涉及 Word2Vec 和 LDA 模型,以提高经常因语义依赖和特征稀缺而恶化的分类性能<sup>[4]</sup>。盖璇计算分词权重,提出构建邮件的特征空间,将邮件特征量化<sup>[5]</sup>;霍光煜用 LDA 主题模型和 K-means 聚类算法构建模型,对于新的短文本则采用 fast-text 深度学习进行档案数据的智能分类<sup>[6]</sup>。余本功提出一种结合主题模型和词向量的方法构建 SVM 的输入空间向量,并融合集成学习的方式提出的 nBD-SVM 文本分类模型<sup>[7]</sup>。

Zhang 针对短文本分类数据不足的问题,提出了一种基于 TextCNN 的中文短文本分类模型,利用回译实现数据增广,弥补了训练数据的不足<sup>[8]</sup>。段丹丹利用 BERT 模型表示短文本的特征向量,再输入 softmax 模型进行回归训练和分类<sup>[9]</sup>。付静提出改进的 BERT 模型,把词向量和位置向量作为模型的输入,通过多头注意力机制获取长距离的语义关系来提取短文本特征,其次利用 Word2Vec 融合主题模型来拓展短文本的特征表示<sup>[10]</sup>。张斌艳提出基于半监督图的神经网络模型,在模型构建中引入了词项和文档之间的关系来增强短文本的表示<sup>[11]</sup>。雷明珠在 resLCCNN 模型的基础上,引入神经主题模型,将信息存储在记忆网络中,加入序列因素,最后,将其输入具有残差结构的卷积神经网络以及双向 GRU 中,提取局部以及全局的语义特征进行分类<sup>[12]</sup>。王渤茹在对短文本的特征提取阶段,对比了三种方法,其中基于字词向量的双路卷积神经网络比单一的卷积神经网络效果更好,在此基础上,提出了深度神经决策森林的分类算法<sup>[13]</sup>。

尽管深度学习在自然语言处理方面效果惊人,但是现有的传统方法利用外部知识来处理短文本的稀疏性和歧义性,由于忽略了上下文相关的特征,准确率仍有待提高。Liu 针对这个问题将上下文相关特征与基于时间卷积网络(TCN)和 CNN 的多阶段注意力模型相结合,并证实了方法的有效性<sup>[14]</sup>。Cheng 针对卷积神经网络(CNN)和双向长短期记忆(BiLSTM)无法区分重要性词的问题,提出一种改进的基于 ERNIE\_BiGRU 模型的分类方法,提高了计算速度和分类

效果<sup>[15]</sup>。

针对短文本特征稀疏,分类困难的问题,该文提出一种基于类别主题词集的加权相似度的短文本分类。选择出最能代表各类别的词语组成类别主题词集,通过计算关键词到主题词的加权相似度来选择短文本的类别。解决了短文本特征稀疏、特征抽取难度大的问题。

## 2 基于类别主题词集的加权相似度分类

针对短文本存在的数据稀疏和特征选择难度大的问题,提出的模型和传统的特征拓展不同,而是计算短文本的关键词和类别主题词之间的加权相似度来对短文本进行分类。该文提出的基于类别主题词集的加权相似度算法,其核心思想是通过 TF-IDF 选取各类别下的类别主题词,保留各词语的 TF-IDF 值,使用 Word2Vec 训练出词向量模型,将短文本预处理之后的关键词与各类别下的主题词的相似度进行加权求和,选择相似度最大的类别作为短文本的类别。

基于主题词集的加权相似度短文本分类算法主要分为四个模块:关键词提取模块,对短文本关键词进行分词、去停用词处理;类别主题词模块,选择最能代表本类别的词语构成类别主题词集;词向量训练模块,基于内部数据语料使用 Word2Vec 训练词向量,得到词向量模型;算法分类模块,将短文本关键词和类别主题词相似度进行计算,融合主题词的权重,以进行分类。框架设计如图 1 所示。

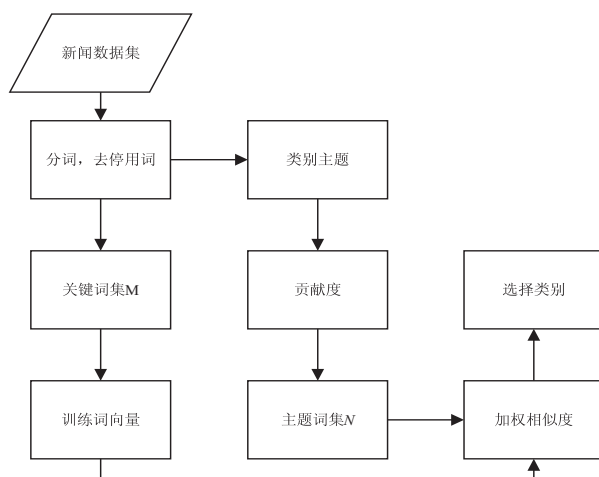


图 1 框架设计

### 2.1 word2vector 模型

word2vector 是词语向量化表示的升级。从传统的独热编码发展到根据上下文语义更好地表示词语。word2vector 也叫词嵌入,词向量是神经网络算法进行 N-gram 语言模型训练过程中的一个副产品,并能够在训练过程中得到词的向量化表示。语言模型训练时的目标函数为:

$$L = \prod_{i=1}^m \prod_{j=1}^n p((w_{ij} | \text{context}_{ij})) \quad (1)$$

其中,  $m$  表示文档的数量,  $n$  表示每篇文档的单词数,  $p((w_{ij} | \text{context}_{ij}))$  表示在上下文为  $\text{context}_{ij}$  的条件下  $w_{ij}$  出现的概率。词向量就是最大化上述目标函数时的产物。word2vector 提供了两种经典的语言模型进行训练, 分别是 CBOW 和 Skip-gram 模型。CBOW 是根据上下文词语来预测中间词语。Skip-gram 模型与 CBOW 模型不同, 是利用当前词推测上下文中的相关词汇。在训练过程中, 两种架构又各有侧重: CBOW 在词向量的训练速度方面表现出色; Skip-gram 虽然在训练速度上较慢, 但是其训练低频词的效果较好。在该文的模型中需要训练全部的特征词, 所以选择 Skip-gram 模型。

## 2.2 类别主题词集和贡献度

主题词要能最大程度地反映类别信息。将各个类别下的词语按照 TF-IDF 值来对词语进行降序排序。选取前 TOP-N 个词语作为类别的主题词集  $N_i$ 。主题词反映类别的不同程度用贡献度来表示, 并将主题词的 TF-IDF 值作为主题词对类别的贡献度。在对新闻标题文本进行分词, 去停用词之后, 将所有文本用作语料库。TF 表示短文本中词语出现的频率, IDF 表示出现这个词语的类别数。则 TF-IDF 的计算方法如公式 2:

$$W(\text{TF-IDF}) = \frac{w_{ij}}{\sum w_j} \log \frac{n+1}{n(w_i+1)} + 1 \quad (2)$$

其中,  $w_{ij}$  表示某一类别中的特定词语出现的次数,  $\sum w_j$  表示特定类别的词语总数,  $n$  表示类别总数,  $w_i$  表示含有这一词语的类别数。为了防止对数的真数和分式的分母为零, 用上述公式进行修正。

## 2.3 关键词到类别的相似度

短文本的关键词为  $M_i$ , 主题词为  $N_i$ , 短文本中的关键词到主题词的相似度用余弦公式(公式 3)来计算。并且考虑到各类别下的主题词的 TF-IDF 值差距过大会对结果产生影响, 所以每个类别下相同顺序的词语权重值取平均值作为第 TOP-N 词的权重。用主题词的 TF-IDF 值来代表主题词对类别的贡献度。每个关键词与类别的相似度用模型  $f(x_i)$  表示(公式 4),  $x_{ij}$  表示的是短文本的第  $i$  个关键词与第  $j$  个主题词的相似度。贡献度体现在模型的权重  $w$  中。

$$x_{ij} = \cos\theta = \frac{\sum_1^n (x_i * y_i)}{\sqrt{\sum_1^n (x_i)^2} * \sqrt{\sum_1^n (y_i)^2}} \quad (3)$$

$$f(x_i) = w_1 x_{i1} + w_2 x_{i2} + \dots + w_i x_{ij} + \dots + w_n x_{in} \quad (4)$$

因此, 短文本到类别的相似度为  $g(x)$ , 如公式 5

所示。

$$g(x) = \frac{f(x_1) + f(x_2) + \dots + f(x_i) + \dots + f(x_m)}{m} \quad (5)$$

选择短文本相似度最大的类别作为短文本的类别。

## 3 实验结果与分析

### 3.1 数据集以及数据预处理

实验在内存为 16G 的 windows10 系统上进行, 使用的编程语言为 python3.6, 编译器为 jupyter notebook。

实验目的是为了测试基于类别主题词集的加权相似度算法的分类效果。该文使用公开的 THUCNews 语料库。THUCNews 语料库是新浪新闻 RSS 订阅频道 2005 年—2011 年的数据, 共有 74 万篇新闻文档, 14 个类别。选取其中房产、股票、教育、社会、时政、体育、游戏 7 个类别的文本进行实验。其中每个类别训练集为 18 000 条数据, 测试集为 1 000 和 2 000 条, 并将 1 000 条和 2 000 条数据结果进行对比。具体实验数据如表 1 所示。

表 1 新闻数据集

文本类别	Train	Test
Realty	18 000	1 000, 2 000
Stock	18 000	1 000, 2 000
Education	18 000	1 000, 2 000
Society	18 000	1 000, 2 000
Politics	18 000	1 000, 2 000
Sports	18 000	1 000, 2 000
Game	18 000	1 000, 2 000

首先将所有数据进行分词和去停用词处理。选取每个类别下 TF-IDF 值为前 TOP-50 的特征词作为类别主题词, 如教育类的 TOP-30 主题词和 TF-IDF 值, 如图 2 所示, ‘考研’一词对教育类别的贡献度最高。

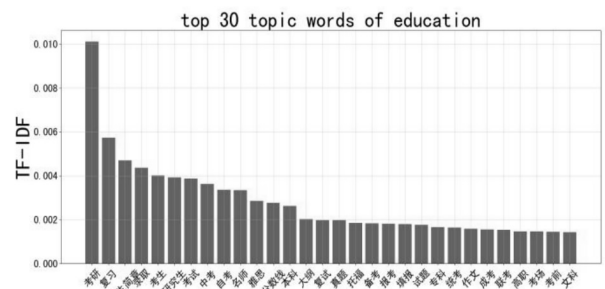


图 2 教育类别主题词集

考虑到各类别下的主题词的 TF-IDF 值差距过大会对结果产生影响, 所以每个类别下相同顺序的词语权重值取平均值作为第 TOP-N 词的权重, 因此选取

各类别下 TOP 顺序在同一位置的特征词的 TF-IDF 值进行平均,得到 TOP-50 个主题词的权重,如图 3 所示。

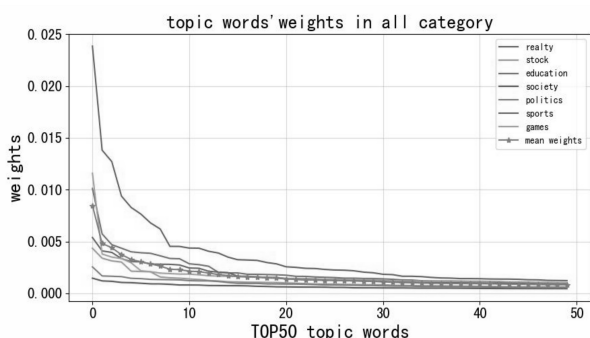


图 3 各类别平均权重

### 3.2 评价指标

实验中采用精确率(PR)、召回率(RC)、调和平均值(F1)来评价模型的分类效果,其计算方法如下。三个指标分别来评估测试集为 1 000 和 2 000 时的精确率、召回率和调和平均值。

$$PR = \frac{TP}{TP + FP} \quad (6)$$

$$RC = \frac{TP}{TP + FN} \quad (7)$$

$$F1 = \frac{2PR \times RC}{PR + RC} \quad (8)$$

其中,TP 是正确地预测为正例,FP 是错误地预测为正例,FN 是错误地预测为反例。精确率(公式 6)是正确地被预测为正例(TP)占有所有实际被预测为正例(TP+FP)的比例,召回率(公式 7)是正确地被预测为正例(TP)占有应该被预测为正例(TP+FN)的比例,F1 是 PR 和 RC 的调和平均值(公式 8)。

### 3.3 词向量模型对召回率的影响

使用 Python 环境下的 Gensim 库训练词向量模型,Skip-gram 模型中>window 表示窗口大小,size 表示词向量的维度。通过不断增加 size 的大小,分类召回率在不断变化。当 window 设置为 8,size 大小为 15 时,达到曲线的拐点,此时的召回率最高,达到了 88.9%,在此基础上通过调节参数 window 的大小,当 window 为 16 时,达到最高召回率 91%,如图 4 所示,选用此时的 Skip-gram 模型训练并计算关键词与各类别词的相似度。

### 3.4 基于类别主题词集的加权相似度分类算法

如第一条测试集数据为[词汇 阅读 关键 考研 暑期 英语 复习 指南],类别标记为教育类别,标签数字为 2,短文本到各类别的相似度分别为[0.319 846 16, 0.287 555 1, 0.475 932 26, 0.334 259 93, 0.296 793 82, 0.294 998 77, 0.323 647 86],由此判断此条新闻标题属于教育类别,分类正确。图 5 展示了社会类别数

据的分类结果,社会类别标签为 3,预测正确的是类别 3,预测错误的是 3 以外的其他标签数字。

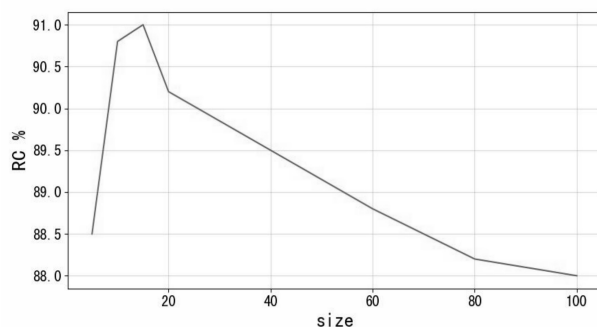


图 4 召回率随 size 的变化情况

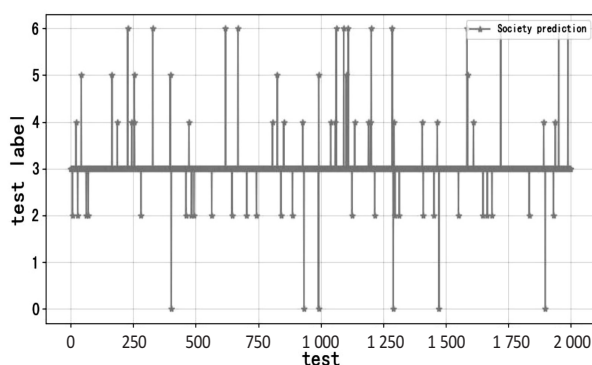


图 5 社会类别分类结果

### 3.5 实验结果分析

在测试集为 1 000 条和 2 000 时,测试文中方法在分类任务上的分类效果,各类别的精确率、召回率以及调和平均值如表 2 所示。

表 2 各类别分类指标

	Label	PR	RC	F1	Support
0	Reality	97.0	85.2	90.7	1 000
		96.9	84.2	90.1	2 000
1	Stock	90.6	85.3	87.9	1 000
		90.3	85.9	88.0	2 000
2	Education	95.5	92.8	94.1	1 000
		95.9	92.4	94.1	2 000
3	Society	85.6	96.3	90.6	1 000
		84.9	96.4	90.2	2 000
4	Politics	84.2	90.4	87.2	1 000
		85.1	90.1	87.5	2 000
5	Sports	95.3	94.2	94.8	1 000
		95.2	94.5	94.8	2 000
6	Game	92.5	94.6	93.5	1 000
		92.1	94.1	93.1	2 000
	Mean	91.5	91.2	91.3	1 000
		91.4	91.0	91.1	2 000

表 2 显示文中分类方法在数据集各个领域类别均



能获得满意的分类效果,是一种有效的分类算法。其中房产领域效果尤其明显,在时政类别效果略逊色于其他类别。可能时政类别新闻标题的内容较短,而这里基于所有标题同样长度来训练 Word2Vec 所导致的。

将文中方法与三种基于单一模型的分类方法(KNN、Logistic 分类、决策树分类)进行比较,表3展示了测试集为1 000时各种算法的精确率(PR)、召回率(RC)和调和平均值(F1)。

表3 算法对比结果 %

分类算法	PR	RC	F1
KNN 算法	88.6	88.2	88.2
Logistic 回归	89.7	89.5	89.5
决策树算法	81.3	80.8	80.9
文中方法	91.5	91.2	91.3

表3显示,前三种基于单一模型的分类方法中,基于决策树的分类算法效果最差,表明决策树分类模型并不适用于文本分类,决策树需要足够多的特征支持,要想取得一个较好的效果,须从数据中构建非常多的特征,做大量的特征工程相关工作,但是短文本特征稀疏,因此决策树并不适合处理高维稀疏矩阵数据;与三种基于单一模型的分类方法相比,文中方法相较 KNN 算法、Logistic 回归算法、决策树分类算法在精确率上分别提高了2.9%、1.8%、10.2%;在召回率上分别提升了3.0%、1.7%、10.4%;在调和平均值上分别提高了3.1%、1.8%、10.4%。用加权相似度算法融合词向量与类别主题词集对短文本进行建模,能够更精细在词层面表示文本的语义信息,从而提高短文本的分类效果。

#### 4 结束语

提出了一种基于类别主题词集的加权相似度算法,在词的层面充分利用词向量和词语之间的相似性来进行文本分类,并且还探索了词向量维度的大小对结果的影响。与其他分类算法相比具有一定的优势,例如和机器学习与深度学习的算法相比,适合数据量不多的情况,无监督学习不需要过多数据进行训练和学习,算法简单,分类速度快。但是该模型中的权重选取过于简单,缺乏依据。后续将重点研究如何通过训练得出最优的权重组合。

#### 参考文献:

- [1] WANG H, TIAN K, WU Z, et al. A short text classification method based on convolutional neural network and semantic extension[J]. International Journal of Computational Intelligence Systems, 2020, 14(1): 367-375.
- [2] 赵晓平, 黄祖源, 黄世锋, 等. 一种结合 TF-IDF 方法和词向量的短文本聚类算法[J]. 电子设计工程, 2020, 28(21): 5-9.
- [3] QIJUN D. Method of short text classification based on TF-IWF feature selection[J]. International Journal of Social Science and Education Research, 2021, 4(4): 367-375.
- [4] ZHOU Y, DENG D, CHI J. A short text classification algorithm based on semantic extension[J]. Chinese Journal of Electronics, 2021, 30(1): 153-159.
- [5] 盖璇. 基于聚类分析算法的垃圾邮件识别[J]. 计算机与现代化, 2020(10): 17-22.
- [6] 霍光煜, 张勇, 孙艳丰, 等. 基于语义的档案数据智能分类方法研究[J]. 计算机工程与应用, 2021, 57(6): 247-253.
- [7] 余本功, 陈杨楠, 杨颖. 基于 nBD-SVM 模型的投诉短文本分类[J]. 数据分析与知识发现, 2019, 3(5): 77-85.
- [8] ZHANG T, YOU F. Research on short text classification based on TextCNN[J]. Journal of Physics: Conference Series, 2021, 1757(1): 012092.
- [9] 段丹丹, 唐加山, 温勇, 等. 基于 BERT 模型的中文短文本分类算法[J]. 计算机工程, 2021, 47(1): 79-86.
- [10] 付静, 龚永罡, 廉小亲, 等. 基于 BERT-LDA 的新闻短文本分类方法[J]. 信息技术与信息化, 2021(2): 127-129.
- [11] 张斌艳, 朱小飞, 肖朝晖, 等. 基于半监督图神经网络的短文本分类[J]. 山东大学学报: 理学版, 2021, 56(5): 57-65.
- [12] 雷明珠, 邵新慧. 短文本分类模型的优化及应用[J]. 计算机应用研究, 2021, 38(6): 1775-1779.
- [13] 王渤茹, 范菁, 张王策, 等. 基于深度神经决策森林的新闻标题分类[J]. 云南民族大学学报: 自然科学版, 2020, 29(5): 472-479.
- [14] LIU Y, LI P, HU X. Combining context-relevant features with multi-stage attention network for short text classification[J]. Computer Speech & Language, 2022, 71: 101268.
- [15] CHENG X, ZHANG C, LI Q. Improved Chinese short text classification method based on ERNIE\_BiGRU model[J]. Journal of Physics: Conference Series, 2021, 1993(1): 012038.