基于 ERNIE-RCNN 模型的中文短文本分类

王浩畅,孙铭泽
(东北石油大学 计算机与信息技术学院,黑龙江 大庆 163318)

摘 要:由于中文短文本存在特征词少、规范性差、数据规模量大等难点,ERNIE 预训练模型占用内存大,进行短文本分类时会造成向量空间稀疏、文本预训练不准确、时间复杂度高等问题。针对以上短文本分类存在的问题,提出基于 ERNIE-RCNN 模型的中文短文本分类。模型运用 ERNIE 模型作为词向量,对实体和词语义单元掩码,后连接 Transformer 的编码层,对 ERNIE 层输出的词嵌入向量进行编码,优化模型过拟合问题,增强泛化能力,RCNN 模型对 ERNIE 输入的词向量进行特征提取,卷积层利用大小不同的卷积核提取大小不同的特征值,池化层进行映射处理,最后通过 softmax 进行分类。将该模型与七种深度学习文本分类模型在中文新闻数据集上进行训练实验,得到了模型在准确率、精准率、召回率、F1 值、迭代次数、运行时间上的对比结果,表明 ERNIE-RCNN 模型能够很好地提取文本中的特征信息,减少了训练时间,有效解决了中文短文本分类的难点,具有很好的分类效果。

关键词:中文短文本分类;ERNIE 模型;ERNIE-RCNN 模型;词向量;特征提取;深度学习
中图分类号:TP391 **文献标识码**:A **文章编号**:1673-629X(2022)06-0028-06
doi:10.3969/j.issn.1673-629X.2022.06.005

Chinese Short Text Classification Based on ERNIE-RCNN Model

WANG Hao-chang, SUN Ming-ze
(School of Computer and Information Technology, Northeast Petroleum University, Daqing 163318, China)

Abstract: Due to the difficulties in short Chinese texts such as fewer feature words, poor standardization and large data size, the ERNIE pre-training model occupies a large amount of memory, which causes problems such as sparse vector space, inaccurate text pre-training and high time complexity when classifying short texts. In response to the above short text classification problems, we propose a Chinese short text classification based on the ERNIE-RCNN model. The model uses the ERNIE model as a word vector, masks entities and word sense units, and then connects to the encoding layer of Transformer and outputs to the ERNIE layer. The word embedding vector is encoded to optimize the model over-fitting problem and enhance the generalization ability. The RCNN model performs feature extraction on the word vector input by ERNIE. The convolution layer uses convolution kernels of different sizes to extract feature values of different sizes. The pooling layer is mapped and finally classified by softmax. The proposed model is trained on the Chinese news data set with seven deep learning text classification models, and the comparison results of accuracy, precision, recall, F1 value, number of iterations and running time are obtained. It is showed that ERNIE-RCNN can extract the feature information in the text well, reduce the training time, effectively solve the difficulties in the classification of Chinese short texts with excellent classification effect.

Key words: Chinese short text classification; ERNIE; ERNIE-RCNN; word vector; feature extraction; deep learning

0 引 言

文本分类(text classification)是自然语言处理(natural language processing, NLP)学科的一门重要研究方向,而短文本分类则是文本分类的一个重要分支。随着自然语言处理技术研究的不断深入,文本分类算法研究也获得了巨大突破[1]。文本分类由最初依靠人工进行规则提取方式,转向基于机器学习的自动分类方式,通过机器学习方法提取文本分类规则进行自动分类,将机器学习算法应用到文本分类领域[2]。

文本分类过程一般包括:文本预处理、特征提取、模型选择、损失函数计算、测试评估[3]。文本分类方法经过长期的研究在很多场景下已经得到了应用,但是短文本分类研究起步比较晚,且一直没有什么通用的、效果良好的方法。短文文本分类一般存在两个问题,其一是短文本提供的词语少,提供的有效信息有限;其二是根据分词结果构建的词频或者特征矩阵稀疏,大

收稿日期:2021-07-12 **修回日期**:2021-11-16
基金项目:国家自然科学基金(61402099,61702093)
作者简介:王浩畅(1974-),女,教授,研究方向为自然语言处理;孙铭泽(1997-),男,硕士研究生,研究方向为自然语言处理。

多数算法重点放在处理稀疏矩阵,效果都不好。因此短文本分类重心放在特征处理和分类算法环节上,就需要加深对深度学习模型的研究。

深度学习模型在文本分类上表现出了相对较好的分类效果,得益于模型在复杂特征提取和文本表示方面有着更强大的能力[4]。例如,快速文本分类 FastText[5]模型,具有模型结构简单、训练速度快的特点,能够处理样本数量大、类别标签多的任务,文本分类任务中将整篇文档的词叠加得到文档向量。卷积神经网络(convolutional neural network,CNN)文本分类模型 TextCNN[6],简化了卷积层,具有参数数目少、计算量少、训练速度快等优势。在循环神经网络(recurrent neural network,RNN[7])和 CNN 基础上,Lai S[8]等提出了一种循环卷积神经网络(recurrent convolutional neural network,RCNN)分类方法,汲取了 RNN 和 CNN 共同的优势,具有训练时间更短、训练速度更快、处理样本更多等优势,采用双向循环结构,能最大程度捕捉下文信息,极大提高了分类的准确率,分类效果更明显。

近年来,随着深度学习的发展,使得预训练语言模型受到广泛关注,预训练模型是利用训练好的词向量初始化网络文本表征问题。当前最好的预训练模型有 Bert[9]、ERNIE[10]模型。李可悦等[11]提出基于 BERT 的社交电商文本分类算法,采用 BERT 预训练语言模型,完成社交电商文本句子层面的特征向量表示,有针对性地将获取特征向量输入分类器进行分类,能够高效准确地判断文本所描述商品的类别。邢照野等[12]提出基于改进 ERNIE 模型的中文文本分类,通过利用知识增强的语义表示预训练模型生成基于上下文信息的词向量,有效提高了中文文本分类性能。

伴随文本分类技术的成熟,对于短文本分类技术的需求日益突显,一些研究者陆续开始着重短文本分类技术研究。王玉燕等[13]针对短文本存在篇幅短、特征稀疏、主题多变等问题,提出基于深度学习的短文本分类技术,采用 CNN 和 RNN 系列模型,结合场景需要,应用到文本分类方案中,通过实验验证,各个模型都表现出了更好的性能。吕飞亚等[14]针对短文本分类中存在特征表示高维稀疏、语义分布不明显、上下文语意联系不强等问题,会对信息抽取造成困扰,提出了基于 BiLSTM 与 Bert 的短文本分类方法,其中 BiLSTM 层获取更多上下文不同距离的语义信息,注意力机制层对经过的编码数据进行转变加权提升序列的学习任务。

段丹丹等[15]针对短文本分类算法存在的特征稀疏、用词不规范和数据海量问题,提出一种基于 Transformer 的双向编码器表示 BERT 的中文短文本分类算法,使用 BERT 预训练语言模型对短文本进行句子层面的特征向量表示,将获得的特征向量输入到 Softmax 回归模型进行训练分类。齐佳琪等[16]针对短文本分类中存在的长度短、数据海量、文本噪音大等问题,提出了 ERNIE 词向量与深金字塔卷积神经网络模型的短文本分类研究,运用 ERNIE 实体掩码方式捕获词汇和语义信息,使用卷积神经网络进行特征提取,取得了很好的分类效果。

根据以上研究,该文针对短文本存在的特征词少、规范性少、数据规模量大的难点,提出将 ERNIE 预训练模型与 RCNN 模型进行融合的短文本分类方法。该模型以 ERNIE 作为词向量,对实体和词语义单元掩码,后连接 Transformer[17]的 Encoder 层,对 ERNIE 层输出词嵌入向量进行编码,优化模型过拟合问题,增强泛化能力,RCNN 模型对 ERNIE 输入的词向量进行特征提取,卷积层利用大小不同的卷积核提取大小不同的特征值,池化层进行映射处理,最后通过 softmax 进行分类,提高了中文短文本分类性能。

1　基于 ERNIE-RCNN 模型的中文短文本分类

1.1　ERNIE 模型

ERNIE 模型是一种基于知识增强策略的持续学习语义理解模型,通过不断吸收大量文本数据里的结构、语义、词汇等知识,实现模型效果不断进化。与 BERT 相比,ERNIE 也是由微调和预训练两部分组成,不同的是其预训练过程利用了更丰富的语义知识和语义任务,在多个 NLP 任务上效果显著。ERNIE 模型结构是经过多层、双向 Transformer 编码和 ERNIE 词向量构成,如图 1 所示。

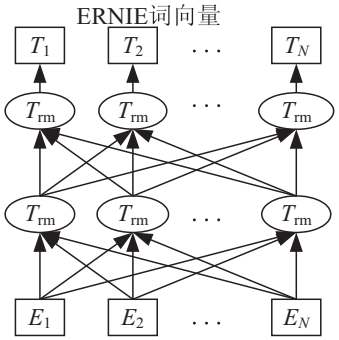

图 1　ERNIE 模型

由图 1 模型看出,E_1,E_2,…,E_N 表示文本输入,经过 Transformer 编码后,输出 ERNIE 词向量。在整个预训练过程中,ERNIE 使用的数据是对整个词语进行屏蔽,从而学习到词与实体表达。

1.1.1　ERNIE 结构详解

ERNIE 结构是由 12 个 Encoder 组成,从输入上看

第一个输入是一个特殊的 CLS,CLS 表示的是分类任务。底层是单词输入,其中共有 768 个隐藏单位,对输入的单词通过 Mask 机制进行中文实体掩码,然后把结果传输到下一个 Encoder 层,最后输出结果。

1.1.2 ERNIE Encoder 模型结构

ERNIE Encoder 基本上是 Transformer 的 Encoder 部分,并且 Encoder 在结构上全部是一样的,但并不共享权重。ERNIE Encoder 结构如图 2 所示。

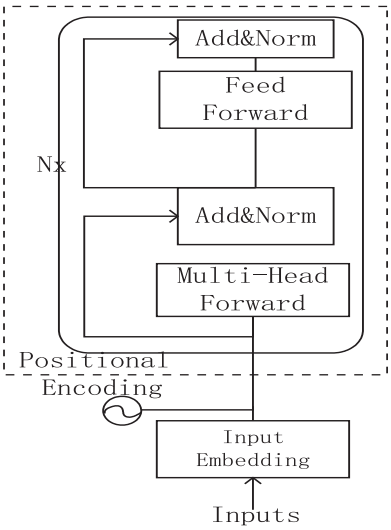

图 2 ERNIE Encoder 结构

由图 2 可以看出,最下层输入的是 embedding 的向量,然后经过一个位置信息的嵌入,输出到多头自注意力机制层,进行多头自注意力计算。接下来 ERNIE Encoder 的输出会经过一层 Add&Norm 层,Add 表示对来自多头自注意力机制层的输入和输出进行残差连接,Norm 表示对输入和输出进行归一化处理,归一化处理后的结果会传入前馈神经网络层。然后再经过一层 Add&Norm 层,通过同样的处理后会输出归一化的词向量列表。

1.2 构建 RCNN 模型

RCNN 是卷积神经网络用于目标检测的模型,其中 CNN 具有良好的特征提取和分类回归性能。算法步骤如下:(1)候选区域选择;(2)CNN 特征提取;(3)分类与边界回归。RCNN 模型如图 3 所示。

1.2.1 候选区域选择

候选区域选择是一种传统的区域提取方法,方法用的是选择性搜索(selective search,SS[18])方法,SS 用来查看现有的小区域,合并两个最有可能的区域,然后重复操作,最后输出候选区域。候选区域一般为 1k ~ 2k 左右,可理解为将信息划分为 1k ~ 2k 个网格,之后再对网络进行特征提取或卷积操作。

1.2.2 CNN 特征提取

CNN 特征提取可以再次提取文本中的关键信息及深层结构信息,且 CNN 可以并行运行,能够加快训练速度。如图 3 所示,CNN 由若干个卷积层、池化层、全连接层组成。卷积层会将 Encoder 层输出的向量提取出不同长度词语的信息和结构信息。如输入一个句子,卷积层首先会对这个句子进行切分,假设 $C_1 \sim C_n$ 为 1 到 n 个单词,对每个单词进行词嵌入,可以得到 $X_1 \sim X_n$ 词向量。假设词向量共 d 维,将 $X_1 \sim X_n$ 词向量拼接($X_1, X_2, \cdots, X_n$),那么对于这个句子便可以得到 n 行 d 列的矩阵 $\boldsymbol{X}$ 。

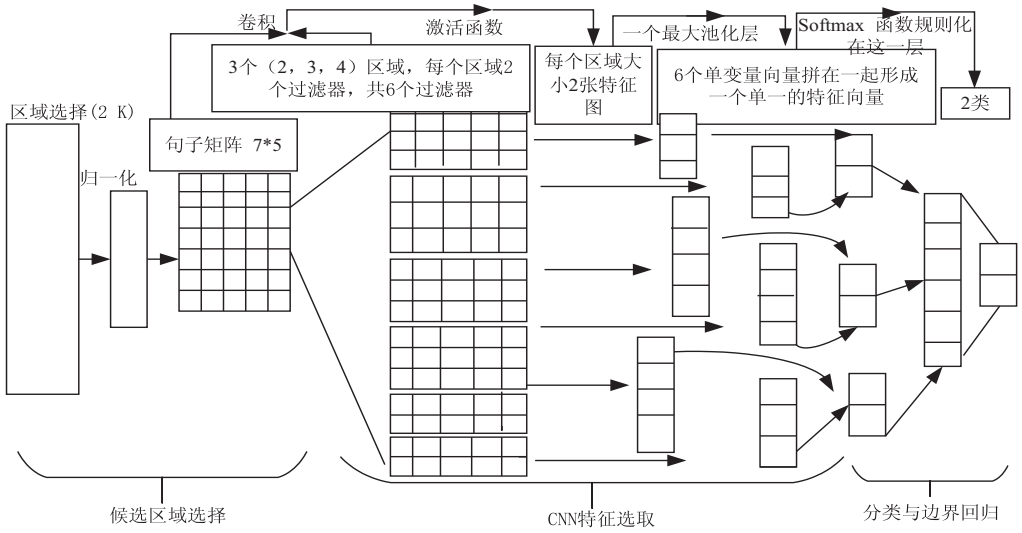

图 3 RCNN 模型

文本生成的词向量通过拼接构建成的句子向量是二维向量,因此卷积过程可由如下公式表示:

$$s(i,j) = (X,W)(i,j) = \sum_m \sum_n X(i-m, j-n) W(m,n)$$

式中,X 为卷积核,$\boldsymbol{W}$ 为被卷积矩阵,m 为对应矩阵的词向量行数,n 为矩阵的维数,i 和 j 为映射后的行和列,$s(i,j)$ 为卷积和 $\boldsymbol{W}$ 对应的输出矩阵的对应位置元素的值。

池化层会对卷积层获得的特征值进行特征映射处理,由于不同尺寸的卷积核得到的特征值大小是不一样的,因此池化层会对每个特征图使用池化函数,使得它们的维度相同,最常用的就是最大池化层,提取出特征图句子的最大值,这样卷积核得到的特征就是一个值,然后对所有的卷积核使用最大池化层,最后经过全连接层把所有卷积核连接起来,就可以得到最终的特征向量。为了防止过拟合,全连接层还引入了 drop out 机制。

1.2.3　分类与边界回归

分类与边界回归共有两个子步骤:第一个是对前一步的输出向量进行分类;第二个是通过边界回归框回归获得精确的区域的信息。目的是准确定位和合并完成分类的预期目标,并避免多重检测。

1.3　ERNIE-RCNN 模型建立的具体步骤

步骤 1:对输入的数据集进行预处理,得到输入文本,记为 $E=(E_1,E_2,\cdots,E_i,\cdots,E_n)$,其中 $E_i(i=1,2,\cdots,n)$ 表示文本的第 i 个字。

步骤 2:将每个 E_i 输入到 ERNIE 预训练层,进行 Mask 掩码,然后经过 Transformer 编码器编码后,将文本 E 进行序列特征化,输出文本 $W_i=(W_{1i},W_{2i},\cdots,W_{ni})$,其中 W_{ni} 表示文本中第 i 句中的第 n 个词的词向量,将 $W_1 \sim W_n$ 词向量拼接($W_1,W_2,\cdots,W_n$),得到矩阵 $\boldsymbol{W}$,即 ERNIE 词向量。

步骤 3:将 ERNIE 词向量输入到 RCNN 模型,RCNN 模型经过再次特征提取,将 ERNIE 输入的词向量 $\boldsymbol{W}$ 经过卷积层操作,输出 $X_i=(X_{1i},X_{2i},\cdots,X_{ni})$,将 $X_1 \sim X_n$ 词向量拼接($X_1,X_2,\cdots,X_n$)得到的矩阵 $\boldsymbol{X}$,然后经过池化层映射处理,得到统一的特征值,经过全连接层连接和 softmax 回归,生成新的特征向量,最后输出。ERNIE-RCNN 模型如图 4 所示。

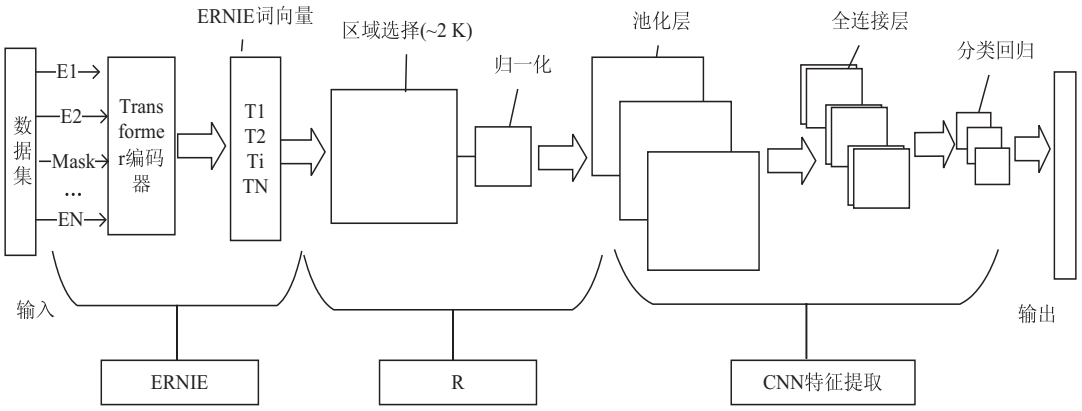

图 4　ERNIE-RCNN 模型

2　实验及结果分析

2.1　实验数据集

实验选用的是 THUCNews 中文新闻数据集[19],THUCNews 是根据新浪新闻 RSS 订阅频道 05 到 11 年间的历史数据筛选过滤生成,包括 74 万篇新闻文档,均为 UTF-8 纯文本格式。从 THUCNews 中抽取了 20 万条新闻标题,其中 18 万条作为训练集,1 万条作为测试集,1 万条作为验证集,文本长度在 20 到 30 之间。一共 10 个类别,分别为财经、家居、房产、教育、科技、时尚、时政、体育、游戏、军事,每个类别数据共 2 万条,数据分类均衡。

2.2　实验环境

实验采用的硬件 GPU 为 NVIDIA-SMI,内存容量为 8 G,CPU 为 Intel(R) Core(TM) i7-9700KCPU @ 3.60 GHz,Python 版本为 3.7.7。

2.3　评价指标

实验用了文本分类中常用的评价指标:精确率、召回率、F1 值,计算公式如下:

$$\text{Precision}=\frac{\text{TP}}{\text{TP}+\text{FP}}$$

$$\text{Recall}=\frac{\text{TP}}{\text{TP}+\text{FN}}$$

$$\text{F1}=\frac{2PR}{P+R}$$

式中,TP 为实际值和预测值均为正值时数据的数量,FP 为实际值为负值、预测值为正值时的数据数量,FN 为实际值为正值、预测值为负值时数据的数量,P 为精确率,R 为召回率。

2.4　实验结果

为对比模型分类性能,选择 TextRNN、Transformer、TextCNN、TextRCNN、Bert、ERNIE、Bert-RCNN[20] 7 种模型与 ERNIE-RCNN 模型进行三组对比实验,为进一步比较模型的性能,查阅了两篇使用同样数据集文章,进行一组简单对比。

对比 1:从图 5 可以看出,ERNIE-RCNN 模型准确率最高,说明分类效果最好。其中相对 ERNIE 和 Bert

-RCNN 差别不大，但相对 TextRNN、Transformer、TextCNN、TextRCNN 模型，分类效果有明显差别。

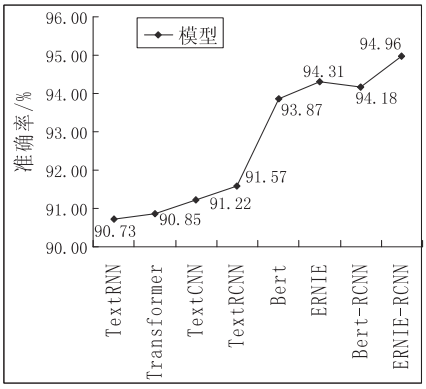

图 5 模型准确率的变化

对比 2：从表 1 可以看出，ERNIE-RCNN 模型在精准率、召回率、F1 值这三个指标上均优于其他模型，其中在精确率上，比 TextRNN、Transformer、TextCNN、TextRCNN、Bert、ENRIE、Bert-RCNN 分别提高了 4.23、4.11、3.74、3.39、1.09、0.65、0.78 个百分点。TextRCNN、Transformer、TextCNN、TextRCNN模型分类结果差距不大，说明 embedding 部分没有很好地提取文本特征，下游模型的变化对分类结果影响不大，而对于 Transformer 模型分类也并不理想，说明 embedding 后的下游模型是决定分类结果的重要部分。

对比 3：从表 2 可以看到，ERNIE-RCNN 模型比 Bert、ENRIE、Bert-RCNN 模型训练时间更短一些，ERNIE-RCNN 模型随着数据的增加，模型训练效果时间成本更低。

表 1 不同模型的测试结果

模型	精确率/%	召回率/%	F1 值
TextRNN	90.73	90.73	0.907
Transformer	90.85	90.79	0.907 9
TextCNN	91.22	91.20	0.912
TextRCNN	91.57	91.52	0.915 2
Bert	93.87	93.87	0.938 6
ERNIE	94.31	94.27	0.942 7
Bert-RCNN	94.18	94.15	0.941 6
ERNIE-RCNN	94.96	94.95	0.949 2

表 2 不同模型的训练时间

epoch	训练时间/s							
	TextRNN	Transformer	TextCNN	TextRCNN	Bert	ERNIE	Bert-RCNN	ERNIE-RCNN
1	15	16	18	11	664	663	663	661
2	29	30	37	23	1 318	1 315	1 334	1 314
3	43	46	57	34	1 972	1 964	1 995	1 919

对比 4：齐佳琪等[16]和雷景生等[21]使用共同 THUCNews 中文新闻数据集，针对其论文的实验结果与本实验结果进行了精确率、召回率、F1 值上的对比，如表 3 所示。可以看出，ERNIE-RCNN 模型在精确率、召回率上比 ERNIE-BiGRU 模型分别高出了 0.64、0.83 个百分点，比 ERNIE-CNN 模型分别高出了 1.02、1 个百分点，表明 ERNIE-RCNN 模型具有更好的分类效果。

表 3 使用 THUCNews 数据集模型比较

模型	精确率/%	召回率/%	F1 值
ERNIE-CNN	93.94	93.95	0.939 4
ERNIE-BiGRU	94.32	94.12	0.942 2
ERNIE-RCNN	94.96	94.95	0.949 2

3 结束语

在解决中文短文本存在难点时，提出一种基于 ERNIE-RCNN 模型的中文短文本分类方法，利用 ERNIE 预训练模型提取文本特征信息，输出对应的词向量，将输出结果作为 RCNN 模型的输入进行训练。从实验结果表明，ERNIE-RCNN 模型在测试集上具有更高的准确率，分类性能更强。不足之处是训练时间效果并不明显，还需进一步提高模型训练性能，缩短训练时间。

参考文献：

[1] 何 锴. 基于自然语言处理的文本分类研究与应用[D]. 南京：南京邮电大学，2020.

[2] 赵博轩，房 宁，赵群飞，等. 利用拼音特征的深度学习文本分类模型[J]. 高技术通讯，2017，27(7)：596-603.

[3] 孙昭颖，刘功申. 面向短文本的神经网络聚类算法研究[J]. 计算机科学，2018，45(S1)：392-395.

[4] 万家山，吴云志. 基于深度学习的文本分类方法研究综述[J]. 天津理工大学学报，2021，37(2)：41-47.

[5] JOULIN A，GRAVE E，BOJANOWSKI P，et al. Bag of tricks for efficient text classification[C]//Proceedings of the 15th conference of the European chapter of the association for computational linguistics. Valencia，Spain：EACI，2016：427-431.

[6] KIM Y. Convolutional neural networks for sentence classification[C]//Proceedings of the 2014 conference on empirical methods in natural language processing. Doha，Qatar：EMNLP，2014：1746-1754.

[7] LIU P，QIU X，HUANG X. Recurrent neural network for text classification with multi-task learning[C]//Proceeding of the

twenty-fifth international joint conference on artificial intelligence. New York, USA: IJCAI, 2016: 2873-2879.

[8] LAI S, XU L, LIU K, et al. Recurrent convolutional neural networks for text classification[C]//Twenty-ninth AAAI conference on artificial intelligence. Austin, TX, USA: AAAI, 2015: 2267-2273.

[9] DEVLIN J, CHANG M W, LEE K, et al. BERT: pretraining of deep bidirectional transformers for language understanding [EB/OL]. [2019-09-16]. https://arxiv.org/abs/1810.04805.

[10] SUN Y, WANG S H, LI Y K, et al. ERNIE: enhanced representation through knowledge integration[EB/OL]. (2019-04-19) [2019-12-23]. https://arxiv.org/abs/1904.09223v1.

[11] 李可悦,陈　轶,牛少彰. 基于 BERT 的社交电商文本分类算法[J]. 计算机科学,2021,48(2):87-92.

[12] 邢照野,刘晓群. 基于改进 ERNIE 模型的中文文本分类方法[J]. 信息与电脑,2021,33(8):87-89.

[13] 王玉燕. 基于深度学习的短文本分类技术研究[D]. 西安:西安电子科技大学,2020.

[14] 吕飞亚. 基于 BiLSTM 与 Bert 的短文本分类方法研究[D]. 太原:太原科技大学,2020.

[15] 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.

[16] 齐佳琪,迟呈英,战学刚. ERNIE-CNN 文本分类模型[J]. 辽宁科技大学学报,2021,44(1):56-61.

[17] DAI Z, YANG Z, YANG Y, et al. Transformer-xl: attentive language models beyond a fixed-length context[C]//Proceedings of the 57th annual meeting of the association for computational linguistics. Florence, Italy: ACL, 2019: 2978-2988.

[18] UIJLINGS J R R, SANDE K E A, GEVERS T, et al. Selective search for object recognition[J]. International Journal of Computer Vision, 2013, 104(2): 154-171.

[19] 清华大学自然语言处理实验室. THUCNews 中文新闻数据集[DB/OL]. 2006. http://thuctc.thunlp.org/message.

[20] 李悦晨,钱玲飞,马　静. 基于 BERT-RCNN 模型的微博谣言早期检测研究[J]. 情报理论与实践,2021,44(7):173-177.

[21] 雷景生,钱　叶. 基于 ERNIE-BiGRU 模型的中文文本分类方法[J]. 上海电力大学学报,2020,36(4):329-335.

(上接第 27 页)

[13] ZHU J, QIAO J, DAI X, et al. Relation classification via target-concentrated attention CNNs[C]//International conference on neural information processing. Guangzhou: Springer, 2017: 137-146.

[14] 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, MN, USA: NAACL, 2019: 4171-4186.

[15] 张东东,彭敦陆. ENT-BERT:结合 BERT 和实体信息的实体关系分类模型[J]. 小型微型计算机系统,2020,41(12):2557-2562.

[16] REIMERS N, GUREVYCH I. Sentence-BERT: sentence embeddings using siamese BERT-networks[C]//Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP). [s.l.]: [s.n.], 2019.

[17] 潘理虎,张佳宇,张英俊,等. 煤矿领域知识图谱构建[J]. 计算机应用与软件,2019,36(8):47-54.