基于增强语言表示模型的网络新闻长文本分类的研究

许楠桸，柯圆圆，胡晓莉*
（江汉大学 人工智能学院，湖北 武汉 430056）

摘 要：基于网络实时新闻内容数据，对一份具有时效的中文长文本数据集进行了新闻主题分类。利用年度关键词增强的分词方案提升分词精度，采用一种长文本压缩方法处理中文长文本的特殊数据，具体方法为选择关键句并利用TF－IDF算法提取长文本中关键词，再将组合的新文本进行词向量训练。最后，采用增强的语言表示模型进行新闻主题分类，并与6种机器学习和深度学习模型进行对比评估，评价指标为召回率、准确率、精度和F1分数等。实验结果表明：本文的模型可通过提取16个重要词对实时新闻长文本进行有效的分类。

关键词：ERNIE模型；预训练模型；新闻分类；长文本处理；中文文本
中图分类号：TP391.1 **文献标志码**：A **文章编号**：1673－0143(2024)04－0037－08
DOI：10.16389/j.cnki.cn42－1737/n.2024.04.004

Long Text Classification for Web News Based on Enhanced Language Representation Model

XU Nanxi, KE Yuanyuan, HU Xiaoli*
(School of Artificial Intelligence, Jianghan University, Wuhan 430056, Hubei, China)

Abstract: Based on the real－time news content data of the Internet, the author classified the news topic of a time－limited Chinese long text data set. The segmentation scheme of annual keyword enhancement was used to improve the segmentation accuracy. In addition, the author adopted a long text compression method to process the special data of Chinese long text. The specific method was to select key sentences, and extract the keywords in long text using the TF－IDF algorithm, then carry out word vector training on the combined new text. Finally, the author used an enhanced language representation model to classify news topics and compared them with six machine learning and deep learning models, including recall rate, accuracy, precision, and F1 score. The experimental results show that the model can effectively classify long text in real－time news by extracting 16 important words.

收稿日期：2023 － 10 － 10
基金项目：江汉大学研究生科研创新基金项目(KYCXJJ202350)
作者简介：许楠桸(2001—)，女，硕士生，研究方向：数学教育。
*通信作者：胡晓莉(1984—)，女，教授，博士，研究方向：数学与大数据、学科教学。E－mail：xxds1234@qq.com

Key words: ERNIE model; pretraining model; news classification; long text processing; Chinese text

随着大数据时代的到来,网络媒体迅速发展壮大,网络新闻已成为人们获取信息的重要途径之一。相较于传统媒体,网络新闻形式更加丰富多样,并且具有更高的时效性。然而,随之而来的信息爆炸问题让人们在海量信息中获取目标信息变得困难。为解决这一问题,文本分类技术提供了一种进行文本信息管理的新思路,其通过将文档中的信息按照一定的规则划分为不同的类别,能够快速而准确地对海量信息进行高效的分类和归纳。这不仅极大地提高了信息处理的效率而且还可以大大缩小信息检索的范围,更加迅速准确地获取所需信息,极大程度上解决信息杂乱的问题[1]。

文本分类问题的研究可追溯到20世纪50年代,Luhn[2]在1958年提出基于词频加权的自动摘录和标引技术,这一技术的提出标志着文本分类的自动化发展迈出重要的一步。在20世纪90年代,机器学习文本分类技术开始兴起。Lewis于1992年建立用于实验测试的数据集[3]。然而,由于传统特征工程对语义表达能力较弱,当面对大规模数据量时,机器学习的分类效果并不令人满意,因而有着处理海量数据优势的深度学习模型开始走向历史舞台[4]。2006年,Hinton等[5]通过逐层的无监督学习方法优化深度神经网络,从而奠定了深度学习的基本框架和理论基础。同年,他们还提出了深度学习(Deep Learning)的概念[6]。近年来,关于文本分类的研究取得了较大进展。2015年,陈翠平[7]利用深度信念网络进行特征提取,并结合Softmax回归分类器实现高维特征的降维并提升了训练速度。谷歌的Vaswani等[8]提出Transformer框架,该框架被广泛应用到各种预训练模型和文本分类任务中。2020年,张曼等[9]将全卷积思路应用于字符级文本分类任务,成功地解决了传统神经网络中的过拟合问题并提高了模型的收敛速度。

与机器学习和深度学习模型不同,预训练大语言模型目前已成为文本分类中有效的建模方法。这些模型从文本中捕获语言信息,并利用特定自然语言处理(NLP)任务的信息实现语言分类任务。预训练方法可以分为两类:基于特征的方法和微调方法。与上述基于特征的语言方法仅使用预训练的语言表示作为输入特征不同,Dai等[10]通过在未标记的文本上训练自动编码器,然后使用该预训练模型作为其他特定NLP模型的起点,提出微调方法。在该研究的基础之上,又有学者提出了更多用于微调的预训练语言表示模型。Howard等[11]提出了改进的AWD-LSTM[12]来构建通用语言模型(ULMFiT)。而Rodford等[13]则提出生成式预训练模型(GPT)来学习语言表示。Devlin等[14]提出了具有多层Transformer的深度双向模型(BERT),该模型实现各种NLP任务的最新结果。为了应对中文文本的挑战,Zhang等[15]提出了一种中文语境分析模型(ERNIE)。该模型通过使用大规模文本语料库和知识图谱进行训练,增强了语言表示模型,被广泛用于处理中文文本数据。

本文通过优化和改进文本分类算法,可以更好地分析和理解大量的网络新闻信息,发现其中的规律和趋势,进而为用户提供准确的个性化信息推荐和搜索服务。这不仅有助于提升用户的信息获取体验,还对新闻行业的发展和媒体内容的传播起到积极的推动作用。因此,研究新闻文本分类方法具有重要的现实意义和广阔的研究前景。

1 基于增强的语言表示模型的文本分类

1.1 数据集介绍

本文的数据来源于文献[16]提供的中文网络新闻数据,其数据集涵盖各个不同领域和主题,

包括科技、财经、游戏、体育、社会、房产、娱乐、教育、家居和时政共10个类别。这些新闻数据的时间跨度为2022－2023年，具有实时性，并且与时事紧密相关。此外，在中文新闻文本中通常包含大量的文本内容，其中还包含各种命名实体，如人名、地名和组织名等。这使得实体识别和命名实体词汇表的构建具有挑战性。

1.2 中文文本预处理

中文新闻文本通常是长文本，并包含繁体字和标点符号等。本文对这些文本进行了预处理，将繁体字转换为简体字，并去除数字、字母和特殊符号，得到清理后的文本。与英文文本不同，中文语句包含丰富的语境、语气和大量的各种专业名词。由于网络新闻具有时效性，许多专业名词来源于突发新闻事件，这些名词与新闻主题密切相关，往往难以被分词方法进行区分。

因此，本文提出使用年度关键词增强的jieba方案，以实现更加高级的停用词方法。本文对比了4种分词方法：NLTK、jieba、pkuseg和关键词增强的jieba。通过对中文文本进行分词处理，样例的分词效果如表1所示。

表1　样例分词效果

Tab. 1　Sample segmentation effect

方法	效果
样例	俄乌冲突不仅给冲突双方带来巨大损失，也对欧洲安全格局形成重大冲击
NLTK	俄/乌/冲/突/不/仅/给/冲/突/双/方/带/来/巨/大/损/失/，/也/对/欧/洲/安/全/格/局/形/成/重/大/冲/击
jieba	俄乌/冲突/不仅/给/冲突/双方/带来/巨大损失/，/也/对/欧洲/安全/格局/形成/重大/冲击/
pkuseg	俄/乌/冲突/不仅/给/冲突/双方/带来/巨大/损失/，/也/对/欧洲/安全/格局/形成/重大/冲击/
关键词增强的jieba	俄乌冲突/不仅/给/冲突/双方/带来/巨大/损失/，/也/对/欧洲/安全/格局/形成/重大/冲击/

4种停用词处理方法在专业名词和形容词的识别方面存在差异。其中英文分词方法NLTK不适合中文文本的分词任务；jieba能够识别机构名词，其词库更加丰富；而pkuseg则能更加详细地划分动词与名词。本文采用的年度关键词来源于国家语言资源监测与研究网络媒体中心的统计数据[17]，这些年度关键词在新闻文本中频繁出现，对于提升专业名称分词的准确性至关重要。因此，为了更好地提取长句中的关键名词，本文选择使用关键词增强的jieba作为更高效的分词方法。

本文提出一种有效的方法来对长文本进行精简处理，以压缩长文本信息并加速生成可用于主题分类的词向量数据。该过程可以分为以下步骤：

步骤1　将新闻长文本划分为句子，得到[Text_1, Text_2 ,…, Text_n]。提取新闻文本的第一句Text_1，并使用TF－IDF法统计后续文本[Text_2, Text_3, …, Text_n]中的关键词，并从中选择前n个词语，构成语句[Words]。最后，将长文本压缩表示为[Text_1, Words]。

步骤2　对压缩后的文本进行分词，并建立新闻的总词汇表。根据词汇表中词语的频率，为每个词语分配唯一的整数编号。

步骤3　使用词嵌入方法Word2Vec[18]，Word2vec通过一个简单神经网络训练每个词语的分布式表示，即词向量。训练后的词向量既包含了原始文本的信息，又便于分类模型的训练使用。

通过以上步骤能够将长文本精简化，并生成可用于主题分类的词向量数据，这既能够压缩文本信息，又能够提升计算速度。

1.3 增强语言表示的文本分类模型

增强语言表示模型（ERNIE）采用与BERT类似的掩码语言模型（MLM）和下一个句子预测（NSP）作为预训练任务，因此ERNIE能够从文本标记中捕获词汇和句法信息。其中每一个token

对应于一个特殊的[CLS]令牌,并生成其对应的嵌入向量,作为分类任务的输入表示。此外,ERNIE还具有针对知识驱动任务(如关系分类和实体类型)的特殊微调过程。

两层多头自注意力层被用于建模分词嵌入计算和实体嵌入计算。$\{w_1^{i-1},w_2^{i-1},\cdots,w_n^{i-1}\}$和$\{e_1^{i-1},e_2^{i-1},\cdots,e_n^{i-1}\}$分别表示分词和实体的输入,表达式如下:

$$\{\bar{w}_1^{i-1},\bar{w}_2^{i-1},\cdots,\bar{w}_n^{i-1}\}=MH-att(\{w_1^{i-1},w_2^{i-1},\cdots,w_n^{i-1}\}),$$

$$\{\bar{e}_1^{i-1},\bar{e}_2^{i-1},\cdots,\bar{e}_n^{i-1}\}=MH-att(\{e_1^{i-1},e_2^{i-1},\cdots,e_n^{i-1}\})。$$

信息融合层用于分词和整体的互相集成计算,其中分词w_j和整体$e_k=f(w_j)$。信息融合的计算公式为

$$h_j=\sigma(\bar{W}_t^{(i)}\bar{w}_j^{(i)}+\bar{W}_e^{(i)}\bar{w}_t^{(i)}+\bar{b}_t^{(i)}),$$

$$w_j^{(i)}=\sigma(\bar{W}_t^{(i)}h_j+\bar{b}_t^{(i)}),$$

$$e_k^{(i)}=\sigma(\bar{W}_e^{(i)}h_j+\bar{b}_e^{(i)})。$$

完整的ERNIE模型的分类过程可以简化为流程图1。

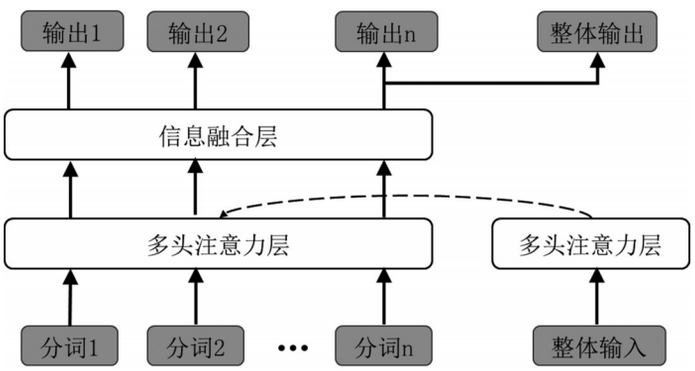

图1 ERNIE模型流程图

Fig. 1 ERNIE model flow chart

综合来看,ERNIE的高效性源于多个因素的结合,包括其预训练模型、知识融合、多任务学习和微调等。这些特点使得ERNIE能够更好地理解中文文本,捕捉上下文信息,并根据不同任务的需求进行适应性调整。因而ERNIE模型在中文数据分类任务中展现出色的性能,目前已成为处理中文数据分类任务的强大工具。

2 实验方法

2.1 实验环境及配置

本文实验环境采用Windows10操作系统,并使用Python作为编程语言,深度学习框架为Pytorch。本文实验数据为中文网络新闻数据集,具体为文献[16]提供的98 762条新闻内容数据,分别对应10个类别,如表2所示。在实验条件允许的情况下,将使用80%的数据用于模型的训练;剩下10%的数据作为验证集用于评估模型的泛化性能;另外10%的数据作为测试集,用于评估模型在实际分类任务中的性能。

2.2 评价指标

在模型训练过程中,验证集和测试集的数据不参加训练。因此,可以使用验证集和测试集在模型上的表现来评估模型的效果。对于评估分类模型,常见的指标有准确率、精度(P)、召回率(R)以及F1值。此外,混淆矩阵也是评估分类模型的重要工具,其展示见表3。

表2　新闻文本数据类别分布

Tab. 2　Data category distribution of news text

新闻类别	训练集＋验证集	测试集数量	总计数量
科技	9 908	1 052	10 960
财经	13 040	873	13 913
游戏	7 874	833	8 707
体育	9 330	1 022	10 352
社会	8 501	996	9 497
房产	9 618	859	10 477
娱乐	8 015	922	8 937
教育	8 601	953	9 554
家居	6 876	1 139	8 015
时政	6 999	1 351	8 350
总计	88 762	10 000	98 762

表3　混淆矩阵

Tab. 3　Confusion matrix

混淆矩阵		真实值	
		正向	负向
预测值	正向	*TP*	*FN*
	负向	*FP*	*TN*

F1值是精度P和召回率R的调和均值，其计算公式为

$$\mathrm{F1}=\frac{2TP}{(2TP+FP+FN)},\ P=\frac{TP}{(TP+FP)},\ R=\frac{TP}{(TP+FN)}, \tag{3}$$

其中TP、FN、FP、TN的值来源于混淆矩阵的计算结果。

2.3　模型构建

本文使用Python语言实现文本分类，其中机器学习模型通过Scikit－Learn库构建，深度学习模型通过Keras和TensorFlow来实现，深度学习模型参数见表4。

表4　深度学习模型参数

Tab. 4　Parameters of deep learning model

参数名称	CNN模型	RNN模型	LSTM模型
序列长度	600	600	600
向量维度	64	64	64
dropout	0.5	0.5	0.5
学习率	0.001	0.001	0.001
批处理大小	64	64	64
隐藏层神经元	—	128	128
隐藏层个数	—	2	2
卷积核数量	256	—	—
卷积核大小	5	—	—
全连接层神经元	128	—	—

注：“—”表示模型没有该参数。

ERNIE模型参数见表5，使用Pytorch来实现，其预训练模型可在http://image.nghuyong.top/ERNIE.zip下载。

表5 ERNIE模型参数

Tab. 5 ERNIE model parameters

参数名称	参数值	参数名称	参数值
类别数	10	样本数	98 762
测试集	78 762	验证集	10 000
测试集	10 000	Epochs	10
学习率	5e−5	隐藏层	768
序列长度	80	Dropout	0.2

3 实验结果及分析

按照2.2节中的步骤对中文长文本数据进行处理,然后输入训练集的词向量数据训练ERNIE模型,获得训练过程的损失值下降和准确率上升的训练迭代结果见图2。

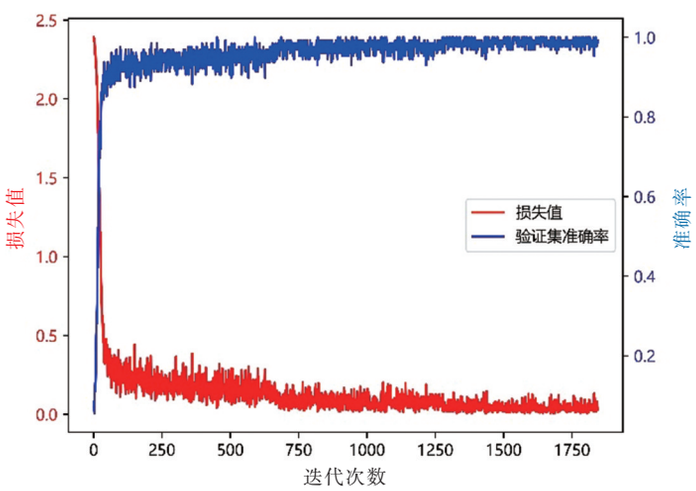

图2 ERNIE模型训练迭代图

Fig. 2 ERNIE model training iteration graph

完成ERNIE机器学习模型和深度学习模型的训练后,得到了最终的实验结果。为了评估其在测试集上的分类效果,采用准确率、召回率、精度以及F1分数作为评价指标。同时,设定ERNIE(n)来提取n个重要词进行计算,以探究最佳的文本分类模型。表6为在测试集上6种模型的新闻文本分类结果。

表6 文本分类结果

Tab. 6 Text categorization results

分类模型	准确率/%	召回率/%	精度/%	F1值/%
朴素贝叶斯	91.18	91.18	91.20	91.18
逻辑回归	92.28	92.28	92.32	92.29
支持向量机	92.65	92.65	92.66	92.65
CNN	93.06	93.06	93.09	93.06
RNN	93.31	92.31	92.30	92.29
LSTM	93.72	93.72	93.86	93.76
ERNIE(8)	95.82	95.89	95.82	95.89
ERNIE(16)	96.53	96.51	96.53	96.51
ERNIE(32)	96.04	96.06	96.04	96.05

由表6可见，在本次实验中，机器学习模型和深度学习模型均取得了良好的分类效果，所有模型在测试集上的准确率、召回率、精度以及F1分数均超过了90%。从表6中各模型在测试集上的分类表现可以看出预训练模型ERNIE在多分类任务中精度整体高于其他模型，并且3个ERNIE模型的多分类精度差距不大，其中ERNIE(16)模型在所有模型中均取得了最优的4项指标。因此，根据模型在具体的新闻类别分类情况，选择新闻文本分类精度最高的ERNIE(16)进行详细分析，其在测试集的详细分类结果如表7所示。

表7　ERNIE(16)的文本测试集分类结果

Tab. 7　Text test aggregation classification of ERNIE (16)

类别	准确率/%	召回率/%	F1值/%	样本数
体育	98.76	99.14	98.95	1 044
娱乐	95.94	98.82	97.36	932
家居	98.48	97.98	98.23	791
房产	93.51	95.95	94.71	1036
教育	97.20	94.85	96.01	951
时政	95.41	96.20	95.80	842
游戏	99.18	98.48	98.47	856
社会	94.99	94.80	98.83	1 019
科技	95.00	95.87	95.43	1 089
财经	97.28	94.51	95.88	1 440
平均值	96.53	96.51	96.51	10 000(总值)

在表7中各文本的分类结果中，10个类别的分类结果模型准确率均达到90%以上。ERNIE(16)模型在体育、家居和游戏等领域的分类准确率更高，基本达到98.5%以上，本文认为预处理文本能帮助ERNIE模型接受足够的信息；而在社会、时政和房产领域的分类准确率相对较低，通常来说这些分类领域容易受突发事件的影响，其新闻主题通常比较新颖并且相对具有时效性，需要采用更完整区分新词的停词方法来提升对应领域的分类精度。

综上，ERNIE模型对于输入数据长度的编码存在影响。这可能与TF－IDF模型选取重要词的强度有关，设置更多的无效重要词会稀释输入信息的有效性，从而影响后续ERNIE模型对新闻文本的分类精度。例如，在社会领域中增加重要词的数量会导致重要性下降，从而使得分类准确率下降。

4　结语

本文通过建立一种长文本处理办法，探究在处理长文本数据时不同方法的文本分类效果。实验结果表明，不同模型均展现出了较好的分类效果，特别是ERNIE模型在对于中文长文本的处理上表现突出，能够对新闻分类进行精准预测。同时，笔者认为在处理中文文本数据上，中文分词是提取关键词的重要基础，本文根据时效性采用了年度关键词来对新闻文本进行处理；为提升精度，可以从词语重要度入手，挖掘更多的重要词；也可以从重要语句入手，提取新闻中的关键句。本研究对于长文本的分类是一项具有实用价值的技术方法，可以帮助新闻行业精准投放到相应平台，为新闻爱好者和阅读者提供便利，同时也帮助新闻平台获取更多的浏览量。

参考文献(References)

[1] 栗征征. 基于CNN和RNN的文本分类方法研究[D]. 荆州:长江大学,2022.

[2] LUHN H P. The automatic creation of literature abstracts [J]. IBM Journal of Research and Development, 1958, 2(2):159-165.

[3] LEWIS D D. An evaluation of phrasal and clustered representations on a text categorization task [C]. Research and development in information retrieval,1992.

[4] 张欣. 基于多尺度CNN与LSTM混合模型的中文新闻分类研究[D]. 青岛:青岛理工大学,2022.

[5] HINTON G E, OSINDERO S, TEH Y W, et al. A fast learning algorithm for deep belief nets [J]. Neural computation, 2006, 18(7):1527-1554.

[6] HINTON G E, SALAKHUTDINOV R. Reducing the dimensionality of data with neural networks [J]. Science,2006,313(5786):504-507.

[7] 陈翠平. 基于深度信念网络的文本分类算法[J]. 计算机系统应用,2015,24(2):121-126.

[8] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need [C]. Proceedings of the 31st international conference on neural information processing systems, 2017: 6000-6010.

[9] 张曼,夏战国,刘兵,等. 全卷积神经网络的字符级文本分类方法[J]. 计算机工程与应用,2020,56(5):166-172.

[10] DAI M A, LE Q V. Semi-supervised sequence learning [M]. Cambridge:MIT Press, 2015.

[11] HOWARD J, RUDER S. Universal language model fine-tuning for text classification [C]. Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: long papers), 2018:328-339.

[12] MERITY S, KESKAR N S, SOCHER R. Regularizing and optimizing LSTM language models [J]. arXiv: 1708.02182, 2017.

[13] RADFORD A, NARASIMHAN K, SALIMANS T, et al. Improving language understanding by generative pre-training [EB/OL]. [2023-09-20]. https://www.mikecaptain.com/resources/pdf/GPT-1.pdf.

[14] DEVLIN J, CHANG M W, LEE K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding [C]. Proceedings of NAACL-HLT, 2019: 4171-4186.

[15] ZHANG Z, HAN X, LIU Z, et al. ERNIE: Enhanced language representation with informative entities [C]. Proceedings of the 57th annual meeting of the association for computational linguistics, 2019: 1441-1451.

[16] CSDN. 10万条新闻数据的数据集[EB/OL]. (2020-6-30)[2023-09-20]. https://download.csdn.net/download/qq_40395874/12563669.

[17] 国家语言资源监测与研究网络媒体中心. 2020年度十大流行语[EB/OL]. (2022-12-20)[2023-09-20]. http://nlp.ccnu.edu.cn/conference/15.

[18] Word2vec-api[EB/OL]. (2021-1-22)[2023-09-20]. https://github.com/3Top/word2vec-api.

(责任编辑:胡燕梅)