[文章编号] 1007－7405(2020)05－0392－09 DOI:10.19715/j.jmuzr.2020.05.12

基于关键词策略和CNN的中文文本有害信息分类

陈德意[1]，张宏怡[1]，刘彩玲[1]，张光斌[2]

（1. 厦门理工学院光电与通信工程学院，福建 厦门 361024；
2. 厦门市美亚柏科信息股份有限公司，福建 厦门 361005）

[摘要] 提出一种新颖的中文文本分类框架。在该框架中，首先基于Word2Vec构建词向量模型，然后采用分词频文档频率（segmentation term frequency-document frequency,STF-DF）筛选出类别区分能力强的关键词，同时构建一种适合于中文文本分类的卷积神经网络（convolution neural network，CNN）进行分类。实验结果表明，采用该框架使THUCNews和复旦大学中文文本数据集中的准确率分别达到了94.51%和95.04%，同时在真实的有害信息数据集中取得了99.70%的召回率，这验证了所提出框架的有效性和实用价值。

[关键词] 词向量；分词频文档频率；特征词集合；Word2Vec模型；卷积神经网络
[中图分类号] TP 312

Classification of Chinese Text Harmful Information Based on Keywords Strategy and Convolutional Neural Network

CHEN Deyi[1], ZHANG Hongyi[1], LIU Cailing[1], ZHANG Guangbin[2]

(1. College of Optoelectronics and Communication Engineering, Xiamen University of Technology, Xiamen 361024, China;
2. Xiamen Meiya Pico Information Co., Ltd., Xiamen 361005, China)

Abstract: The rapid development of internet and big data technology has greatly facilitated people's access to various Chinese text information, but also greatly increased the risk of dissemination of harmful information in Chinese text. The traditional text processing method based on vector representation is mainly used to process English text. To deal with these problems, a novel Chinese text classification framework was proposed. In this framework, a word vector model based on Word2Vec was constructed firstly. Then the keywords with distinguishing category ability were selected by using word document frequency (segmentation term frequency-document frequency, STF-DF). Meanwhile, a suitable convolution neural network (CNN) was build for Chinese text classification. The experimental results show that the accuracy of this framework in THUCNews and Fudan University Chinese text data set is 94.51% and 95.04% respectively, and the recall rate is 99.70% in the real harmful information data set, which verifies the effectiveness and good practical value of the proposed framework.

Keywords: word embedding; STF-DF; feature word set; Word2Vec model; convolution neural network (CNN)

[收稿日期] 2019－06－05

[作者简介] 陈德意（1994—），男，硕士生，从事智能信息处理与应用研究。通信作者：张宏怡（1973—），女，教授，从事智能生物信息处理、模式识别方向研究。E-mail:zhanghongyi@xmut.edu.cn

0　引言

随着网络的快速发展，互联网每天都产生数以万计的文本数据。文本作为信息的重要载体[1]，相对于图像和语音等载体而言，具有容量占有率少，更方便存储和管理的特点[2-3]。但是，如果只凭借传统人工方法对这些大量的文本数据进行管理，需耗费大量人力和财力[4]；并且，文本的表达没有图像和语音那么直观[5]，这也提高了分类的难度。因此，对文本数据进行系统的分类变成一项极具挑战的任务[6]。

文本分类是文本信息挖掘的基本功能，其目的是在预定义的分类体系下，根据文本的特征（内容或属性），将给定的文本与一个或多个类别相关联的过程[7-8]。研究文本表示和分类模型是文本分类的核心问题[9]。传统的文本表示主要是运用词袋模型（bag of words），将文本中的词映射到高维向量空间，这种做法存在严重的特征稀疏、语义特征不明显等缺点，无法表达文本之间词序的相对关系[10]。如：韩琪恒等[11]在论文中提到使用 One-Hot 编码方式来进行文本词映射，存在严重的维数灾害；Nugaliyadde 等[12]运用词向量模型，较好地实现了文本分类任务，有效地缓解了自然语言处理中的数据稀疏问题。

在分类模型研究方面，当前有很多深度学习模型应用在英文新闻文本分类、垃圾邮件检测等场景中[13]。如：Kim 等[14]提出了一种新的文本分类深度神经网络模型 Seq2CNN，可以训练不同的文本长度，且其在短文本上的分类效果要好于长文本；Helmy 等[15]提出一种基于卷积神经网络，独立于语言的文本编码方法的模型，实验准确率达到了 91.99%，但该方法是基于字符级训练的词向量，存在丢失词序信息的缺点。目前，深度学习在中文文本分类的应用和研究也逐渐流行起来。例如：侯小培等[16]采用卷积神经网络进行中文文本分类训练，与传统方法相比，分类效果明显提高，但该方法是利用空间向量模型从文本信息中得到中文词的词向量，并不能很好地表达文本词序之间的关系；蓝雯飞等[17]在经典的 LSTM（long short-term memory）分类模型基础上加入了注意力机制（attention），相比传统方法分类效果更好，但存在计算成本高的缺点；黄贤英等[18]针对社交网络文本传统情感分类模型存在先验知识依赖以及语义理解不足问题，提出一种基于 Word2Vec 和双向 LSTM 的情感分类模型，减小了词向量间的稀疏度，但是其未考虑到文本的局部特征；代令令等[19]将 fastText 应用到中文文本分类中，验证了其可行性，由于其在文本表示时，是将所有的词向量取平均作为文本向量，存在文本上下文语义丢失的问题；Shen 等[20]利用深度学习网络具有良好的特征提取能力，将低层次的特征提取出来，形成更适合分类的高级抽象表示，较好地解决了中文文本分类的降维问题；Chung 等[21]在中文数据集中应用了字符级神经网络，是以单个汉字为基础进行文本表示，但没考虑词序之间的关系，分类准确率还有待提高。

中文与其他语言有很大的不同。对英文而言，一个单词就是一个词，而中文是以字为基本的书写单位，词语之间没有明显的区分标记，需要人为切分，所以中文关键词的选取就显得尤为重要。因此，本文从关键词提取角度出发，同时针对中文文本表示时存在噪声多、特征稀疏的问题，提出一种新的框架。

1　中文文本分类相关算法及流程

本文提出的适于中文文本分类的卷积神经网络算法（convolution neural network，CNN）的框架流程如图 1 所示，主要从数据预处理、文本表示、分类器等几方面进行研究：

1）首先针对中文文本数据集进行分词、去除停用词等预处理，使用 Word2Vec 模型来训练特定分类场景下的文本数据集，能够较好地解决维度过高及文本前后词的关联性低等问题。

2）本文采用基于类别关键词提取的 STF-DF 算法来得到特征词集合。通过该算法得到的类别关键词个数远小于文档的总词数，并以此组成一个特征词集合。在对样本进行文本表示时，把未在特征词集合中的词去除，这样不仅保留了相应词语的上下文关系，也去除了对文本分类没有作用的特征词，经过词向量模型映射为文本特征矩阵，作为分类器模型的输入。

3）在分类器模型上，将采用当前热门的卷积神经网络来作为研究方向，训练分类器，使每个文本尽可能得到正确的分类结果。

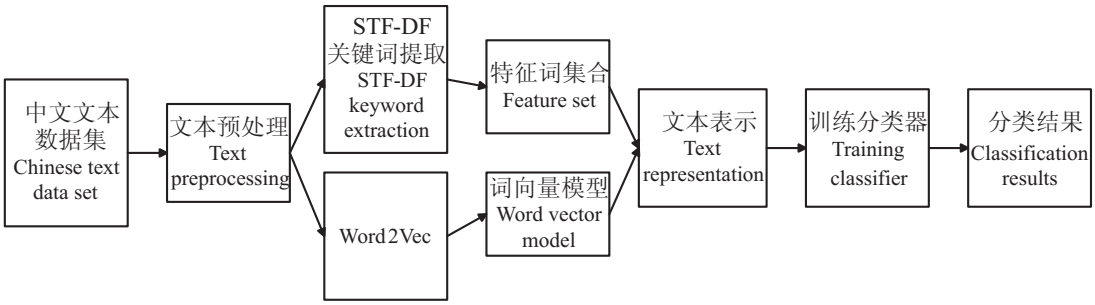

图 1 中文文本分类框架流程图

Fig.1 The flow chart of Chinese text classification framework

1.1 构建文本表示模型

针对中文文本分类中获取到的词向量模型，是通过 Word2Vec 从大规模中文语料库训练中获得的，分类效果还可以提升。因此，本文通过特定分类场景下的中文文本数据集来构建的词向量模型能够有效提高分类准确率。

Word2Vec 是根据上下文之间的出现关系来训练词向量。主要分为 CBOW（continuous bag-of-words）与 Skip-gram 两种模型。CBOW 根据上下文预测目标单词，Skip-gram 根据目标单词预测上下文，最后使用模型的部分参数作为词向量。本文将采用 CBOW 模型来进行词向量的训练。

本文采用 Python 中的结巴（Jieba）分词器来对数据集做分词。首先针对数据集构建一个专有名词字典，将其作为结巴分词器的用户字典，这在一定程度上能提高分词准确率，同时利用中文停用词去除文本中介词、代词、虚词等词以及特殊符号。然后，把分好词的数据集作为 CBOW 模型的输入，经过训练得到该数据集下的词向量模型，方法流程如图 2 所示。

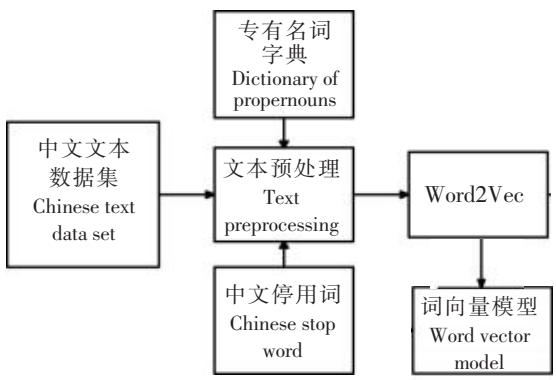

图 2 词向量模型训练

Fig.2 Word2Vec model training

1.2 STF-DF 关键词特征提取算法

STF-DF 算法包含了两种概念，即分词频和分词频文档频率。该算法从分词频的角度计算文档频率，充分考虑相同特征词在不同词频下对分类的影响。

1）分词频是指将一个特征词 t_w 切分成 F 个词频为 f 的特征词 $t(w,f)$，该过程可形式化表示为：

$$t_w \rightarrow \{t(w,1),\cdots,t(w,f),\cdots,t(w,F)\}, \tag{1}$$

其中：F 为特征词在语料库文档中的最大词频；$f \in (0,F]$。

2）STF-DF 的核心思想是将特征词 t_w 的文档频率 d_w 划分为不同类别下的文档频率 $d_{kw}(0 < k < K)$，然后根据分词频的概念，将每个类别中的文档频率 d_{kw} 分成在不同词频下的文档频率 $d_k(w,f)$。该过程表示为：

$$d_w \rightarrow \{d_{1w},\cdots,d_{kw},\cdots,d_{Kw}\}, d_{kw} \rightarrow \{d_k(w,1),\cdots,d_k(w,f),\cdots,d_k(w,F)\}。 \tag{2}$$

首先，对于特征词 t_w 而言，通过上述步骤得到的分词频文档频率可通过一个 $F \times K$ 阶的矩阵 $\boldsymbol{D}_w$ 来描述：

$$\boldsymbol{D}_w = \begin{bmatrix} d_1(w,1) & \cdots & d_k(w,1) & \cdots & d_K(w,1) \\ \vdots & & \vdots & & \vdots \\ d_1(w,q) & \cdots & d_k(w,q) & \cdots & d_K(w,q) \\ \vdots & & \vdots & & \vdots \\ d_1(w,F) & \cdots & d_k(w,F) & \cdots & d_K(w,F) \end{bmatrix}, \tag{3}$$

其中，矩阵的行表示不同词频下的文档频率分布，矩阵的列表示不同类别下的文档频率分布。对于每一个特征词来说，经分词频处理后都可以得到相应的矩阵 $\boldsymbol{D}$，称为分词频文档频率矩阵。

接着，分别计算出特征词 $t(w,f)$ 的 STF-DF 的类间方差 $S^2[t(w,f)]$，计算公式为：

$$S^2[t(w,f)] = \sum_{k=1}^{K}(d_k(w,f)/K - \overline{d(w,f)}), \tag{4}$$

以及不考虑分词频时特征词 t_w 的文档频率类间方差 $S^2[t_w]$，计算公式为：

$$S^2(t_w) = \sum_{k=1}^{K}(d_{kw} - \overline{d_w})/K。 \tag{5}$$

其中：$d_k(w,f)$ 表示矩阵 $\boldsymbol{D}_w$ 中第 w 行第 f 列的 STF-DF 值，即在该类 C_k 中词频为 f 的特征词 t_w 的文档频率；$\overline{d(w,f)}$ 表示 $\boldsymbol{D}_w$ 中第 w 行的分词频文档频率的平均值，即在所有类中词频为 f 的特征词 $t(w,f)$ 文档频率的平均值；d_{kw} 为不考虑分词频时特征词 t_w 在类 C_k 中的文档频率；$\overline{d_w}$ 为不考虑分词频时特征词 t_w 在所有类中文档频率的平均值；K 为类别总数。同时考虑到特征词词频的高、低对分类效果的影响，应加大高词频带来的影响，减小低词频的影响，于是本文引入词频权重因子 A，其值在 0 ~ 1 之间，$A(w,f) = f/F$，由此构造出的特征词重要程度度量公式为：

$$T(t_w) = \sum_{q=1}^{F}A(w,f) * S^2[t(w,f)] + S^2(t_w)。 \tag{6}$$

最后，根据权重值以降序排列所有特征词，并选择前 n 个最佳特征词作为数据集样本的特征词集合。

1.3　文本特征向量表示

在文本特征向量化时，需要对词语进行数字化编码，这是将文本预处理后生成的词语列表转换为数字列表的过程。在本次实验中加入了特征词集合，目的在于去除文本分词后不在特征词集合中的词语。经过这样处理，能够更加准确地提取出每个文本的有用信息，将剩下的文本特征词，用前面训练得到的词向量模型，将文本特征词映射为文本特征向量矩阵。图 3 为文本特征向量表示流程图。

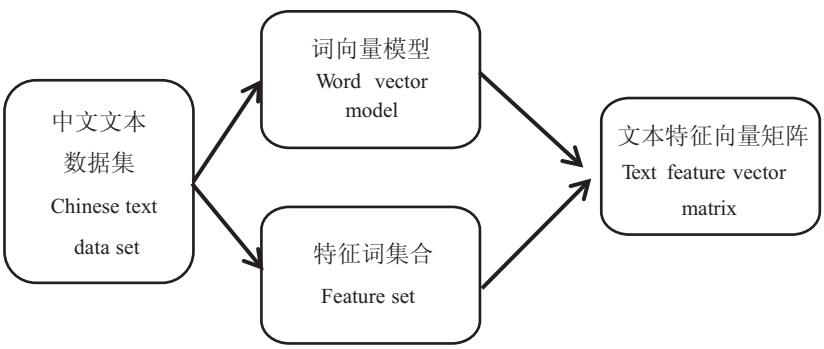

图 3　文本特征向量表示过程

Fig.3　The process of text feature vector representation process

1.4　构建卷积神经网络模型

如图 4 所示，在输入卷积神经网络之前，根据词向量映射原理，用前述的文本表示模型将文本提取出的特征词映射到词向量中。考虑由 n 个特征词组成的文本特征，可以得到维数为 $n \times d$ 的文本矩阵，矩阵中的每一行是一个特征词向量，d 是特征词向量的长度。卷积核的长度通常等于词向量的长度，卷积核的高度 h 称为“感受域”的大小。令 $O_p = q \cdot A[p:p+h-1]$，这里 $p = 1,2,\cdots,n-h+1$，“·”表示子矩阵和滤波器 q 之间的点积（元素乘法的总和）。考虑到偏置项 $\boldsymbol{b} \in R^{n-h+1}$ 和激活函数 f，卷积核的特征图可表示为：$\boldsymbol{z} = f(\boldsymbol{o} + \boldsymbol{b})$。

训练目标是尽量减少交叉损失。在这里，要估计的参数包括卷积核的权重向量、激活函数中的偏差项和 softmax 函数的权重向量。使用随机梯度下降法（SGD）来进行优化，算法具体步骤为：

1）输入层　输入图 4 中最左边 $n \times d$ 的文本特征向量矩阵，所有样本都表示成 $n \times d$ 形式的矩阵

作为卷积层的输入；

2）卷积层　图 4 中卷积核尺寸大小设为［1，2］的一维卷积核各 200 个，对输入的特征矩阵进行卷积，生成相应个数的特征图；

3）池化层　该层是一个 1-max pooling 层，对每个特征图进行池化，这样不同长度的特征图经过 pooling 层后都变成定长的表示；

4）全连接层　该层将各个经过卷积和池化后的特征列向量进行全连接，得到最终文本表示的特征矩阵，作为 softmax 层的输入；

5）softmax 层　输出每个类别的概率，选取其中概率最大的类别作为最终的预测类别。

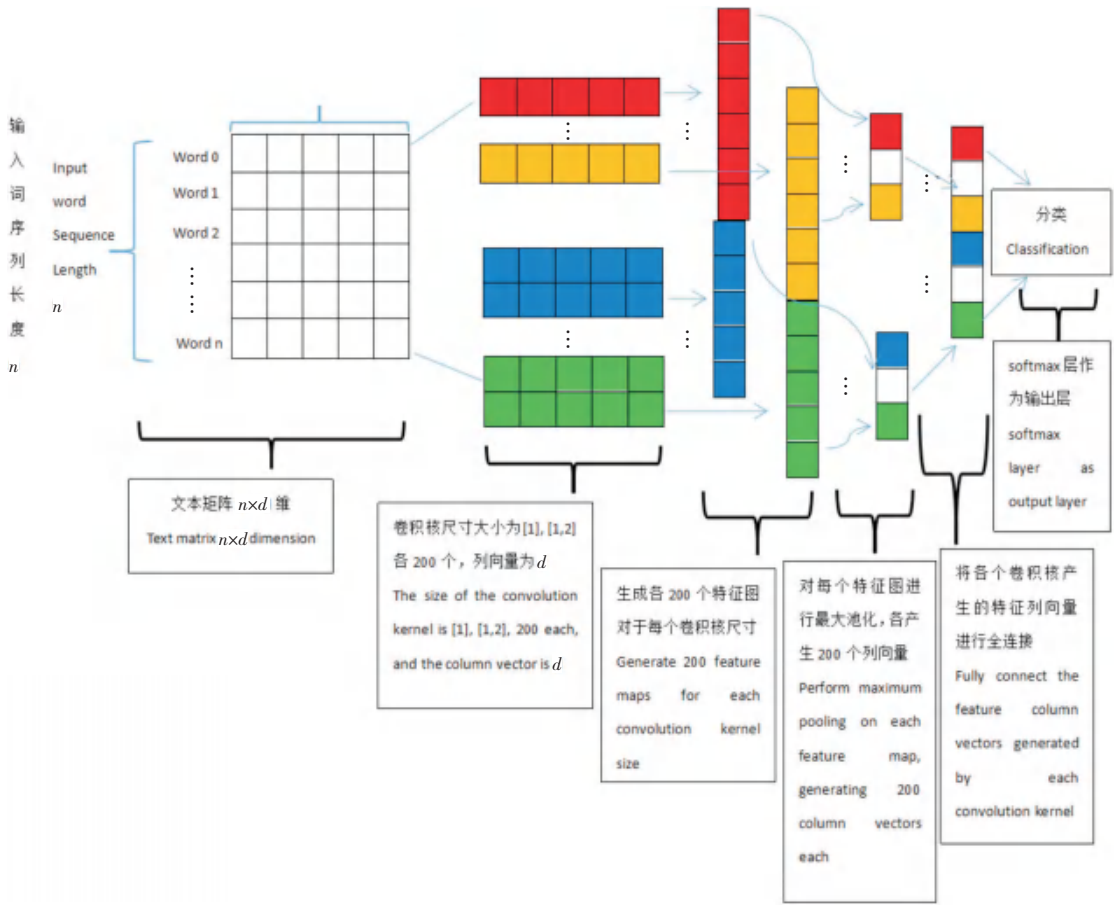

图 4　基于中文文本分类的卷积神经网络原理结构

Fig.4　The principle structure of convolutional neural network based on Chinese text classification

2　实验结果与分析

本文实验基于 Windows10 操作系统，配备 GTX1070 显卡，编程语言为 Python3. 5，使用 Tensorflow1. 7 GPU 版深度学习框架进行开发。

2. 1　实验数据

本次实验数据共分两部分，第一部分数据集是用来验证提出的框架，第二部分是实际应用场景下的有害信息数据集。

第一部分数据集共两种。第一种来自清华大学 THUCNews，共 83 万多条新闻文本，包括彩票、教育、财经、娱乐等 14 个类，实验中将数据随机分成比例为 9∶1 的训练集和测试集。第二种是复旦大学计算机信息与技术系国际数据库中心自然语言处理小组整理的中文语料库，其中训练语料 9804

篇（20 个类），测试语料 9833 篇（20 个类）。表 1 所示为部分语料介绍。

表 1　部分复旦大学数据集

Tab. 1　Part of Fudan University data set

类别名称 Category name	对应中文类别 Corresponding Chinese category	文本数 The article number/条
C3-Art	艺术类 Art	740
C4-Literature	文学类 Literature	33
C5-Education	教育类 Education	59
C6-Philosophy	哲学类 Philosophy	44
C7-History	历史类 History	466
C11-Space	太空类 Space	640
C15-Energy	能源类 Energy	32
C16-Electronics	电子类 Electronics	27
C17-Communication	通信类 Communication	25
C19-Computer	计算机类 Computer	1537

第二部分数据集是有害信息文本和含有政治、经济、科技、教育等类型的中文新闻文本，共两类，其中有害信息文本指的是淫秽色情、低俗庸俗、暴力血腥、恐怖惊悚、赌博诈骗、网络谣言等文本。本实验将采取淫秽色情文本作为实验数据集，数据集是通过相关网站爬取得到，并对两类数据集进行预处理，分别得到 10 000 条文本，其中每条文本的字数都大于 100，实验中将数据随机分成 4∶1 的训练集和测试集。

2.2　实验内容

本次所有实验中，经过 STF-DF 算法提取特征词后，3 种数据集统一保留 20 000 个关键词，通用词向量模型由大量中文语料库训练得到（维基百科语料库，文件大小为 6.5 GB），维数 128。

1）实验一　卷积神经网络卷积核尺寸大小对分类模型的影响

对两种数据集分别采用尺寸大小为［1,2］,［2,3,4］,［3,4,5,6］的卷积核来进行实验，两种数据集训练得到的词向量维数为 128，且加入了特征词集合，实验结果如表 2 所示。

通过实验可以得到本文提出的新型中文文本分类框架在卷积核尺寸大小为［1,2］时具有较好的分类效果，两种数据集的准确率 A 分别达到了 94.51% 和 95.04%。但在复旦大学数据集中，其宏精确率、宏召回率、宏 F_1 值相对较低，原因在于该数据集类别间样本数相差较大，导致某些类别的评价指标值较低，从而宏平均值就变低。接下来的实验将采用卷积核尺寸为［1,2］的卷积核进行实验。

表 2　THUCNews 和复旦大学数据集实验结果

Tab. 2　The experiment results of THUCNews data set

数据集 Data set	卷积核尺寸 Convolution kernel size	准确率 A	宏精确率 Macro_P	宏召回率 Macro_R	宏 F_1 值 Macro_F_1
THUCNews	［1,2］	0.9451	0.9431	0.9419	0.9423
	［2,3,4］	0.9445	0.9424	0.9415	0.9418
	［3,4,5,6］	0.9432	0.9420	0.9412	0.9415
复旦大学 Fudan University	［1,2］	0.9504	0.8877	0.8037	0.8355
	［2,3,4］	0.9488	0.8843	0.8016	0.8329
	［3,4,5,6］	0.9484	0.8831	0.8012	0.8317

2）实验二　基于特定分类场景下数据集的词向量模型与通用词向量模型的分类效果比较

在实验中使用的词向量维数为 128，同时加入特征词集合，得到的实验结果如表 3 所示。其中 ***A*** 为THUCNews 数据集（基于数据集训练的词向量模型），***B*** 为 THUCNews 数据集（通用词向量模型），***C*** 为复旦大学数据集（基于数据集训练的词向量模型），***D*** 为复旦大学数据集（通用词向量模

型）。由实验结果可得，使用基于数据集训练的词向量模型比使用通用词向量模型分类效果好，两种数据集的准确率分别提升了 0.73% 和 0.91%。

表 3 基于数据集训练的词向量模型与通用词向量模型比较

Tab.3 The comparison between the Word2Vec model based on data set training and the general Word2Vec model

类型 Type	准确率 A	宏精确率 Macro_P	宏召回率 Macro_R	宏 F_1 值 Macro_F_1
A	0.9451	0.9431	0.9419	0.9423
B	0.9378	0.9344	0.9336	0.9339
C	0.9504	0.8877	0.8037	0.8355
D	0.9413	0.8791	0.7952	0.8264

3）实验三 是否加入特征词集合对模型分类的影响

使用特定分类场景下数据集训练的词向量模型，词向量维数为 128，分别对两种数据集进行有无加入特征词集合的实验。实验结果如表 4 所示，其中 ***E*** 为 THUCNews 数据集（含特征词集合），***F*** 为 THUCNews 数据集（不含特征词集合），***G*** 为复旦大学数据集（含特征词集合），***H*** 为复旦大学数据集（不含特征词集合）。由实验结果可得，使用特征词集合的模型比没有使用特征词集合的模型在两种数据集上的准确率分别提高了 1.05% 和 1.07%，从而验证了加入特征词集合的模型具有更好的分类效果。

表 4 是否加入特征词集合对模型分类的影响

Tab.4 The influence of whether to add feature word set on model classification

类型 Type	准确率 A	宏精确率 Macro_P	宏召回率 Macro_R	宏 F_1 值 Macro_F_1
E	0.9451	0.9431	0.9419	0.9423
F	0.9346	0.9322	0.9315	0.9318
G	0.9504	0.8877	0.8037	0.8355
H	0.9397	0.8761	0.7939	0.8231

4）实验四 不同深度学习模型下的分类效果比较

目前，已有很多深度学习模型应用在文本分类上。为了验证本文提出的 CNN 分类算法优于其他算法，选取典型的长短时记忆网络（LSTM）分类算法来做对比，同时加入特征词集合。从实验结果（见表 5）可以得出，本文提出的 CNN 分类算法在两种数据集中的评价指标的表现都优于 LSTM，其准确率分别提升了 1.46% 和 1.70%。

表 5 不同深度学习模型下的分类效果比较

Tab.5 Comparison of classification effects under different deep learning models

类型 Type	准确率 A	宏精确率 Macro_P	宏召回率 Macro_R	宏 F_1 值 Macro_F_1
THUCNews 数据集 THUCNews Dataset（CNN）	0.9451	0.9431	0.9419	0.9423
THUCNews 数据集 THUCNews Datase（LSTM）	0.9305	0.9298	0.9229	0.9256
复旦大学数据集 Fudan University dataset（CNN）	0.9504	0.8877	0.8037	0.8355
复旦大学数据集 Fudan University dataset（LSTM）	0.9334	0.8722	0.7825	0.8168

5）实验五 基于提出的中文文本分类模型下的有害信息文本分类效果

使用有害信息数据集训练的词向量模型，词向量维数为 128，同时进行有无加入特征词集合的实验。实验结果如表 6 所示，在不加入特征词集合的情况下，准确率达到了 99.12%，同时在加入特征词集合后，准确率提高了 0.28%。

表 6　中文文本有害信息分类效果

Tab. 6　Classification effect of harmful information in Chinese text

类型 Type	准确率 A	宏精确率 Macro_P	宏召回率 Macro_R	宏 F_1 值 Macro_F_1
有害信息数据集(含特征词集合) Harmful information data set(including feature word set)	0.9940	0.9940	0.9940	0.9940
有害信息数据集(不含特征词集合) Harmful information data set(excluding feature word set)	0.9912	0.9912	0.9913	0.9912

6）实验六　基于不同分类模型下的有害信息文本分类效果比较

为了说明本文框架的优势，采用传统的机器学习分类模型和深度学习模型 LSTM，对有害信息文本进行分类，并对实验结果进行对比。使用的传统方法包括朴素贝叶斯（NB）、最近邻（KNN）、支持向量机（SVM）、极端梯度提升算法（XGBoost）。同时，为了消除不同的特征构造方式导致的实验结果没有可比性，本次传统方法的特征构造方式同样是基于词向量的，是将文本的所有词的词向量取平均值，作为文本的特征表示。实验结果如图 5 所示，采用准确率 A、精确率 P、召回率 R、F_1 值为实验评价指标。

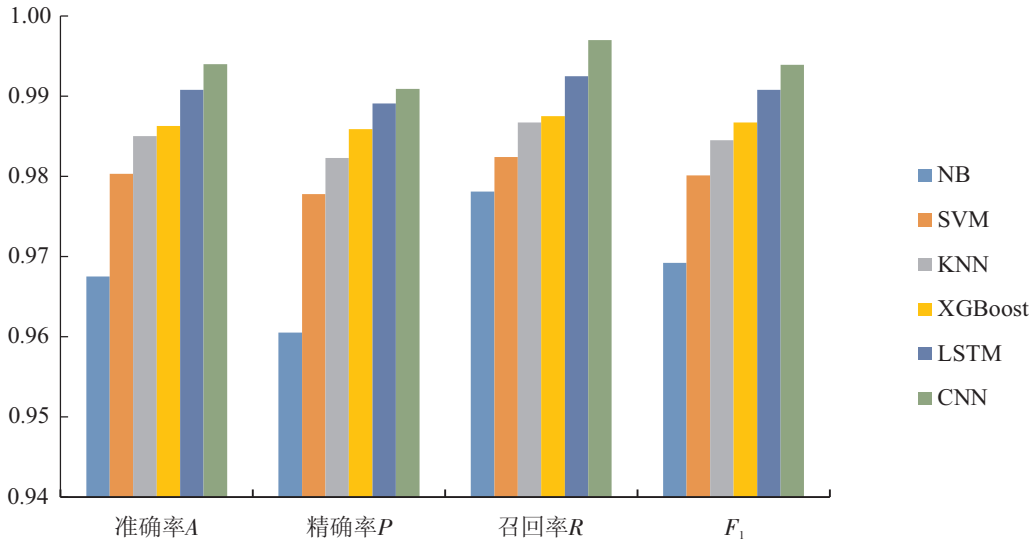

图 5　不同分类模型下的有害信息文本分类效果比较

Fig.5　Comparison of text classification effects of harmful information under different classification models

从实验结果可以得出，在相同的有害信息文本数据集下，以 Word2Vec 训练的词向量来作为文本的特征表示，各种分类模型都取得了较好的分类效果，其中本文方法（CNN）在各类评价指标上都取得了最好结果，特别是色情文本这类的召回率达到了 99.70%。因此可以证明本文方法的分类效果要优于传统方法，究其原因在于 CNN 模型具有自动提取相关语义特征的优势，具有更好的分类效果。

3　结论

本文提出了一种新的中文文本分类框架，通过 Word2Vec 模型来训练特定分类场景下的中文文本数据集，能得到对分类文本表达更准确的词向量模型。同时使用 STF-DF 算法提取关键词，组成特征词集合，从分词频的角度计算文档频率，充分考虑了相同特征词在不同词频下对分类带来的影响。还把特征词集合加入到卷积神经网络的输入文本的特征向量表示中，对卷积神经网络的卷积核尺寸进行选择，最终的分类效果是由以上各个环节共同作用的结果，在两种标准数据集实验中验证了该框架的有效性。并通过使用该模型对中文有害信息文本进行判别，与当前主流的深度学习模型和传统分类模型作比较，证明了本文方法分类效果显著。在下一步工作中，将考虑不平衡数据集在该框架下的分类

效果，并加入其他类别的有害信息数据，进行多分类的中文文本有害信息分类实验。

[参 考 文 献]

[1] RAHMAN A，QAMAR U. A bayesian classifiers based combination model for automatic text classification [C] //7th IEEE International Conference on Software Engineering and Service Science (ICSESS). Beijing：IEEE，2016：63-67.

[2] ABDUR R，KASHIF J，BABRI H A，et al. Selection of the most relevant terms based on a max-min ratio metric for text classification [J]. Expert Systems with Applications，2018，114：78-96.

[3] LIU P，ZHAO H H，TENG J Y，et al. Parallel naive Bayes algorithm for large-scale Chinese text classification based on spark [J]. J Cent South Univ，2019，26(1)：1-12. DOI:10.1007/S11771-019-3978-x.

[4] 马成龙，颜永红. 基于概率语义分布的短文本分类 [J]. 自动化学报，2016，42(11)：1711-1717.

[5] JOHNSON R，ZHANG T. Convolutional neural networks for text categorization：shallow word-level vs. deep character-level [J]. arxiv:1609.00718，2016：402-408.

[6] 林奕欧，雷航，李晓瑜，等. 自然语言处理中的深度学习：方法及应用 [J]. 电子科技大学学报，2017，46(6)：913-919.

[7] HSIEH Y L，LIU S H，CHANG Y C，et al. Neural network-based vector representation of documents for reader-emotion categorization [C] //IEEE International Conference on Information Reuse and Integration. Califorina：IEEE，2015：569-573.

[8] AYINDE B O，INANC T，ZURADA J M . Regularizing deep neural networks by enhancing diversity in feature extraction [J]. IEEE Transactions on Neural Networks and Learning Systems，2019(3)：1-12.

[9] BURKHARDT S，KRAMER S. Online multi-label dependency topic models for text classification [J]. Machine Learning，2018，107(5)：859-886.

[10] SINGH J，SINGH G，SINGH R. Optimization of sentiment analysis using machine learning classifiers [J]. Human-centric Computing and Information Sciences，2017，7(1)：32.

[11] 韩琪恒. 机器学习方法在文本分类中的应用 [J]. 电子制作，2018，359(18)：63-64，66.

[12] NUGALIYADDE A，WONG K W，SOHEL F，et al. Enhancing semantic word representations by embedding deep word relationships [C] //Proceedings of the 2019 11th International Conference on Computer and Automation Engineering. Perth：ACM，2019：82-87.

[13] WANG Q，XU J，HE B，et al. An improved convolutional neural network for sentence classification based on term frequency and segmentation [C] //International Conference on Artificial Neural Networks. Alghero，surdinia，Italy：Springer，2017：56-63.

[14] KIM T，YANG J . Abstractive text classification using sequence-to-convolution neural networks [J]. arxiv：1805.07745，2018.

[15] HELMY A A，OMAR Y M K，HODHOD R. An innovative word encoding method for text classification using convolutional neural network [C] //14th International Computer Engineering Conference (ICENCO). Ciza，Egypt：IEEE，2018：42-47.

[16] 侯小培，高迎. 卷积神经网络 CNN 算法在文本分类上的应用研究 [J]. 科技与创新，2019(4)：74.

[17] 蓝雯飞，徐蔚，汪敦志，等. 基于 LSTM-Attention 的中文新闻文本分类 [J]. 中南民族大学学报（自然科学版），2018，37(3)：133-137.

[18] 黄贤英，刘广峰，刘小洋，等. 基于 word2vec 和双向 LSTM 的情感分类深度模型 [J]. 计算机应用研究，2019(12)：13.

[19] 代令令，蒋侃. 基于 fastText 的中文文本分类 [J]. 计算机与现代化，2018(5)：35.

[20] LUO X，CHEN Y，SHEN F. Text classification dimension reduction algorithm for chinese web page based on deep learning [C] //International Conference on Cyberspace Technology (CCT 2013). Beijing：IET，2013：451-456. DIO：10.1049/cp.2013.2171.

[21] CHUNG T，XU B，LIU Y，et al. Empirical study on character level neural network classifier for Chinese text [J]. Engineering Applications of Artificial Intelligence，2019，80：1-7.

（责任编辑 朱雪莲 英文审校 黄振坤）