Journal of Computer Applications ISSN 1001-9081 2018-09-10
计算机应用,2018,38(9):2469-2476 CODEN JYIIDU http://www.joca.cn

文章编号:1001-9081(2018)09-2469-08 DOI:10.11772/j.issn.1001-9081.2018030643

# 基于词向量和卷积神经网络的垃圾短信识别方法

赖文辉*,乔宇鹏

(华南理工大学 自动化科学与工程学院,广州 510640)

(* 通信作者电子邮箱 lwhscut1994@163.com)

**摘 要:** 对垃圾短信进行过滤识别研究具有重要的社会价值和时代背景意义。针对传统的人工设计短信特征选择方法中存在数据稀疏、特征信息共现不足和特征提取困难的问题,提出一种基于词向量和卷积神经网络(CNN)的垃圾短信识别方法。首先,使用 word2vec 的 skip-gram 模型根据维基中文语料库训练出短信数据集中每个词的词向量,并将每条短信中各个词组所对应的词向量组成表示短信的二维特征矩阵;然后,把特征矩阵作为卷积神经网络的输入,通过卷积层的不同尺度卷积核提取多尺度短信特征,以及利用 1-max pooling 池化策略得到局部最优特征;最后,将局部最优特征组成融合特征向量放入 softmax 分类器中得出分类结果。在 10 万条短信数据上进行的实验结果表明,在特征提取方式相同的情况下,基于卷积神经网络模型的识别准确率能够达到 99.5%,比传统的机器学习模型提高了2.4%~5.1%,且各模型的识别准确率均保持在 94% 以上。

**关键词:** 垃圾短信;识别;word2vec;skip-gram;词向量;卷积神经网络

**中图分类号:** TP181;TP391.1 **文献标志码:** A

## Spam messages recognizing method based on word embedding and convolutional neural network

LAI Wenhui*, QIAO Yupeng

(*School of Automation Science and Engineering, South China University of Technology, Guangzhou Guangdong* 510640, *China*)

**Abstract:** It is of great social value and times background significance to filter and recognize spam messages. Traditional artificially designed feature selection methods may lead to data sparseness, insufficient co-occurrence of feature information and difficulty in feature extraction. To solve above problems, a spam messages recognizing method based on word embedding and convolutional neural network was proposed. Firstly, word2vec's skip-gram model was used to train the word embedding of each word in the short message dataset according to the Wiki Chinese corpus, and the two-dimensional feature matrix representing short message was composed of word embedding of each word in a short message. Then, the feature matrix was used as the input to the convolutional neural network. The multi-scale short message features were extracted by using different scale convolution kernels of the convolution layer, and the 1-max pooling strategy was used to obtain the local optimal features. Finally, the fusion feature vector, composed of the local optimal features, was put into the softmax classifier to get the classification results. Experiments were performed on 100 000 short messages. The experimental results show that the recognition accuracy based on the convolutional neural network model can reach 99.5%, which is 2.4% to 5.1% higher than that of the traditional machine learning models with the same feature extraction method, and the recognition accuracy of each model maintains above 94%.

**Key words:** spam message; recognizing; word2vec; skip-gram; word embedding; Convolutional Neural Network (CNN)

## 0 引言

近年来,随着通信技术的不断进步,我国使用手机的用户数量日益增多,短信成为一种方便的信息传递渠道。然而,短信在使人们的日常生活变得更加方便的同时,垃圾短信的泛滥也越来越严重,严重干扰了人们的生活,成为了危害社会公共安全的一大公害。360 互联网安全中心于 2017 年 2 月发布的《2016 年中国互联网安全报告》显示,360 手机卫士在 2016 年内为全国手机用户共拦截约 173.5 亿条垃圾短信。绝大多数垃圾短信的内容都是广告推销或者影响社会稳定团结的谣言、诈骗等。工业和信息化部于 2015 年 6 月颁布施行的《通信短信息服务管理规定》,其中规定任何短信服务提供商和短信内容发送者在未经用户同意的情况下,都不得向用户发送商业性的信息。因此垃圾短信成为当前一个重要的社会问题,有效识别垃圾短信对维护国家安全、社会稳定和人们正常生活具有重要的时代背景意义。

对垃圾短信进行有效识别的方法主要有三种[1-4],分别是基于黑白名单的方法、基于规则的方法和基于短信内容的方法。基于黑名单和规则的识别方法比较简单,缺点是需要手动添加号码名单和关键词,能够添加的数量相对有限且难以全面,导致识别的效果较差。鉴于前两种方法的局限性,目前对垃圾短信识别技术的研究主要集中在短信的内容上,即利用文本分类技术将垃圾短信识别问题转化为一个有监督的学习问题。文本分类技术是以机器学习算法为基础,先对已

收稿日期:2018-03-29;修回日期:2018-04-23;录用日期:2018-04-25。

作者简介:赖文辉(1994—),男,江西赣州人,硕士研究生,主要研究方向:机器学习、自然语言处理;乔宇鹏(1981—),女,黑龙江海林人,副研究员,博士,主要研究方向:布尔网络、博弈论、自然语言处理。

经过人工标注的文本进行特征提取，然后利用算法对文本进行自动分类。文献[5]提出基于多特征融合的方法来向量化表示短信文本，并分别比较了朴素贝叶斯(Naïve Bayes，NB)、逻辑回归(Logistic Regression，LR)、支持向量机(Support Vector Machine，SVM)和随机森林(Random Forest，RF)等分类器的性能差别，各分类器的识别效果较好，但是特征提取较为复杂。文献[6]提出一种文本加权K最近邻(K Nearest Neighbor，KNN)算法，通过特征词在短信中出现的频率赋予合适的权重，同时对垃圾短信数据集进行频繁词挖掘，并以此提高垃圾短信文本的权重，虽然在性能上有所提升，但仍然没有解决垃圾短信语法和句法格式干扰的问题。文献[7]用信息增益矩阵作为提取短信特征的方法，并在朴素贝叶斯和随机森林这两种分类器上进行垃圾短信的识别。文献[8]针对基于对抗环境下的垃圾短信短信检测技术进行研究，提出了基于特征长度与权重相结合的好词攻击和feature reweighting防御方法来识别垃圾短信。文献[9]提出了消息主题模型(Message Topic Model，MTM)方法提取短信特征，并利用k-means算法将垃圾短信训练成不规则的类，然后把所有的垃圾短信聚合为单个文件以捕获单词的共现模式。文献[10]通过深入研究垃圾短信的特征后发现了10个特征可以有效地过滤垃圾短信，并在随机森林分类器上实现了96.5%的真阳性率和1.02%的假阳性率。这些分类算法使用的前提都是先利用人工设计的文本特征选择方法来提取短信文本的特征，但这些特征提取方法忽略了短信文本长度比较短且上下文的关联性比较强的特点，导致数据特征稀疏，无法体现短信上下文语义之间的联系，并且丢弃了词序、语法等文本结构信息，阻碍了垃圾短信识别效果的提升。相比人工设计文本特征的方式，深度学习能够高效便捷地完成文本特征提取。文献[11]将词向量与卷积神经网络(Convolutional Neural Network，CNN)相结合用于癌症特征文本分类，引入最新的癌症领域数据集进行评估，并获得了良好的效果。文献[12]将CNN用于情感分析和主题模型的分类，模型在测试数据上的准确率相比传统的情感分析模型有显著的提升。

为了提高垃圾短信识别的准确率，针对人工设计特征选择方法提取短信特征时产生的数据稀疏、文本特征信息共现不足和文本特征提取困难等问题，本文尝试将基于Word2Vec模型的词向量特征提取方法和CNN模型相结合，进行垃圾短信识别。词向量方法的优势在于可以通过控制向量的维数来解决“维数灾难”的问题，并且词向量在训练过程中注重词与词之间的位置关系，保留词组在语义之间的联系。CNN是一种出色的深度学习算法，尤其在特征提取方面有更加优异的性能。CNN起初主要应用于图像处理领域，以图像像素矩阵作为模型的输入，因此需要将短信数据转化为图像像素矩阵的形式。首先利用Word2Vec中的Skip-Gram模型根据维基中文语料训练出每个词的词向量；然后由训练好的词向量按照短信的词序组合成表示每条短信的特征矩阵；最后将特征矩阵作为CNN的输入参与到模型的训练测试过程中。为了验证词向量和CNN在垃圾短信识别中的性能优势，本文用同样的数据集和特征提取方法在文献[5]、文献[7]和文献[10]中所采用的分类器即朴素贝叶斯、逻辑回归、支持向量机和随机森林四种机器学习模型上进行对比实验。实验结果表明本文所设计的基于深度学习的词向量和CNN模型相比机器学习模型具有更强的特征提取能力，能有效提高识别的准确率。

## 1 相关理论

### 1.1 中文分词方法

本文是以词向量作为组成文本特征矩阵的基本单位，因此应该以分开的具有独立语言意义的词组作为文本的最小组成要素。

在以单词作为文本基本组成元素的拉丁语言中，单词之间都会以空格分开，因此对这些语言进行分词并不困难。

然而这种天然优势在汉语里并不存在，主要原因是中文文本是由中文词组无空格紧密地连接在一起所组成的，因此中文文本预处理的首要任务是先对文本进行中文分词，将词与词分隔开来。有研究表明，分词质量高低与最终的文本分类效果息息相关，因此，快速准确的分词算法是非常重要的。

目前，研究人员不断开发出一些中文分词器，中文分词技术不断地趋向成熟。比较流行的有NLPIR(Natural Language Processing & Information Retrieval)、THULAC(THU Lexical Analyzer for Chinese)、jieba分词和SnowNLP等分词器，使用者可以根据应用场景选用适合的中文分词器。由于jieba分词器是一个用python语言开发的免费开源的分词工具，并且用户可以根据自己的任务环境自定义词典和词库，所以本文采用jieba分词器中比较适合于文本分析的精确分词模式对短信文本进行分词处理。

### 1.2 文本的传统特征提取方法

目前文本表示通常采用向量空间模型(Vector Space Model，VSM)。VSM是20世纪70年代由Salton等提出的，并在SMART文本检索系统中成功应用[13]。简要概括，VSM用统计的方法提取文本特征的向量表示，然后计算向量之间的距离从而判断文本之间在语义上是否相似[14-16]。VSM的建模过程如下：文本集$D$中某个文本$d \in D$含相互独立的$q$个不同的关键词，即$d=(x_1,x_2,\cdots,x_q)$；以这$q$个关键词在文本中的权重组成的向量$\boldsymbol{V}_d=(w_1,w_2,\cdots,w_q)$作为文本$d$的特征表示。在VSM的建模思想中，如何给各个关键词赋予合适的权重值是最为关键的问题。权重的计算方法一般是利用文本的统计信息，主要是词频，给关键词赋予一定的权重。常用的权重计算方法有布尔权重、词频率-逆文档频率(Term Frequency-Inverse Document Frequency，TF-IDF)和熵权重等。

VSM在表示文本的特征向量时可以降低向量计算的复杂度，具有简单易行的优点，但同时也存在很多缺陷和不足：1)当样本数据集比较大，含有较多的关键词时，文本的特征向量维度较高，可能导致维数灾难。2)VSM只是单纯地将每个词的词频或者权重值作为统计量，割裂了前后文之间的语义联系，造成语义信息的丢失。对于维数灾难问题，可以考虑用文档频率(Document Frequency，DF)、信息增益(Information Gain，IG)、$\chi^2$统计量(Chi-square，CHI)和互信息(Mutual Information，MI)等特征选择方法进行降维。这些特征选择方法虽然可以使“维数灾难”问题在某种程度上得到缓解，但是也加剧了信息的流失。

进而，当面对通常只有几个到几十个词且整个词汇空间非常大的短信文本时，用传统特征提取方法提取短信的文本特征时将产生数据稀疏、文本特征信息共现不足的问题。因此针对传统的特征提取方法存在的不足，为更好地提取短信文本特征，应该尝试新的特征提取方法。

### 1.3　分布式特征提取方法

为解决文本数据特征稀疏的问题，Hinton [17] 提出了一种叫作 word embedding 的词向量方法，其核心理论是将词从高维度的向量空间分布式地投影到低维度空间，不同词之间的语义关联性可以由它们所分别对应的词向量之间的位置关系反映，这种方法保留了词序、语法等文本结构信息，有助于提升文本分类效果 [18-19]。

Mikolov 等 [20] 提出了针对语料库能够快速高效训练词向量的 Word2Vec 模型。Word2Vec 中的 Skip-Gram 模型可以根据所给语料库快速地训练出每个词的词向量。

Skip-Gram 的主要思想是由当前词 $W_t$ 的概率来预测前后文词 $W_i$ 的概率，即预测 $P(W(i) \mid W(t))$，其中 $t-c \leqslant i \leqslant t+c$，每个词向量反映了前后文词的位置情况，并且其训练方式如图 1 所示。

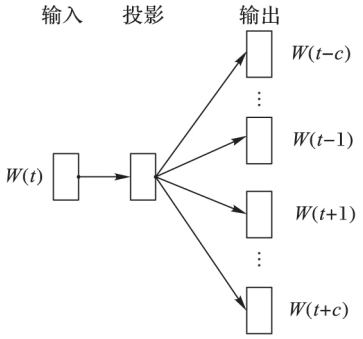


图 1　Skip-Gram 模型结构

Fig. 1　Skip-Gram model structure

假设有一条短信文本经过分词、去除停用词等预处理步骤后，产生一系列词为 $W(1), W(2), \cdots, W(n)$。Skip-Gram 模型的目的是使式(1) 的值最大化：

$$E = \frac{1}{n}\sum_{t=1}^{n}\sum_{-c\leqslant i\leqslant c, i\neq 0} \mathrm{lb}\, P(W(t+i) \mid W(t)) \tag{1}$$

其中：$c$ 表示窗口的长度，即当前词 $W(t)$ 的前面的 $c$ 个词和后面的 $c$ 个词。

综上所述，相较于传统的人工提取特征的方法，词向量方法的优势在于可以控制特征向量维数、在解决维数灾难问题的同时，不会忽略词组在文本中的相对位置关系，而且保留了词组在语义之间的关联。

### 1.4　短信文本分类算法

在完成提取短信文本特征的任务后，后续步骤就是用分类器过滤识别。目前可以完成文本分类的算法主要有两类：一类是当前在研究应用上已经十分成熟的机器学习算法；另一类是时下热点之一的基于深度学习理论的算法。

本文涉及的 CNN 模型是深度学习理论中应用最为广泛的一种结构，具有很多显著的特性：特有的卷积层和池化层使得模型对微小的局部特征十分敏感；模型在训练数据时既可以提取出更加抽象完备的特征信息，也可以完成分类任务。CNN 的诸多优势使其在图像处理和语音识别中大获成功。然而文本数据与图像语音有很大的不同点，文本数据不单是由词语所组成，它更包含了属于人类特有的语义信息，因此文本处理过程中需要更完备可靠的特征信息。鉴于 CNN 的优异的特征提取能力，本文尝试将其用于垃圾短信识别。

## 2　短信文本中文分词及特征提取

垃圾短信识别总体流程如图 2 所示。

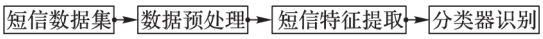


图 2　垃圾短信识别流程

Fig. 2　Spam message recognition flow chart

短信作为日常的通信交流的工具，行文比较随意，结构也不规范；因此在获得短信样本后，首先必须对原始的短信作预处理以得到较为纯净的文本。进一步纯文本数据仍无法被计算机所识别，需要将短信文本转化为计算机可以处理的形式，即用特征提取的结果来表示短信文本，然后使用已经提取到的短信文本特征和类别已知的样本将垃圾短信识别任务转化为有监督的学习问题，设计算法完成最终的识别任务。

### 2.1　数据预处理

由于原始的短信文本数据的格式没有统一的规范，包含许多标点符号和表情图形或颜文字等特殊元素，无法直接处理，需要先对其进行数据清洗。

原始的短信文本数据里的标点符号、表情图和颜文字等与短信的上下文语境之间没有语义关联，首先需要将这些特殊元素过滤掉，只保留具有语言信息意义的中文词组和一些专有的外文词汇。

全半角转换。有一些外文字符是在全角状态下输入的，有些是在半角状态输入的，这导致文本格式都不尽相同，为使下一步的分词操作顺利进行，需要将这些不规范的地方进行格式统一。

原始的短信文本经过数据预处理步骤后，格式更为统一规范，为后续的分词处理奠定了良好的基础。

### 2.2　中文分词处理

由于中文不像英文以空格作为单词之间的分隔符，中文词汇之间没有明确的界限，因此需要先对短信进行中文分词处理，以词作为短信的组成要素。本文采用 python 第三方库中的 jieba 分词器里的精确分词模式作为分词工具。

以短信文本内容是“如果天气逐渐变凉，记得要添加衣服，而且注意防寒保暖☺☺……”为例，首先去除短信中所有的标点符号和表情图，短信内容变为“如果天气逐渐变凉记得要添加衣服而且注意防寒保暖”；然后进行分词处理，结果为“如果 天气 逐渐 变凉 记得 要 添加 衣服 而且 注意 防寒 保暖”。

分词完毕后，还要去除短信文本中的停用词，减少冗余，使文本分类更准确。常见的停用词有“的”“如果”“可以”“要”和“而且”等对垃圾短信识别不重要的词。上述短信在去除停用词后最终的分词结果为“天气 逐渐 变凉 记得 添加 衣服 注意 防寒 保暖”，取得了较好的分词效果。

### 2.3　分布式特征提取

短信文本在经过中文分词处理并滤去停用词后，可以使用 Word2Vec 工具并结合 Skip-Gram 模型训练数据集中各词的词向量。由于 python 语言在自然语言处理任务中的优势，本文使用 python 的第三方开源库 Gensim 作为训练词向量的工具。

若要得到质量较高的词向量，需要具备较大规模的语料库，目的是为了充分地反映出词组在语义空间的位置关系。而维基中文语料是公认的大型中文语料，本文拟用维基中文语料训练词向量。

Skip-Gram 模型有两个重要的参数需要设置，分别是窗口长度参数 $c$ 和词向量维数 $k$。

原则上 $c$ 值越大，考虑前后文的关系就更全面，一般能使预测的结果更加精确，但也会使训练时间更长，因此需要不断尝试来确定 $c$ 值的大小。由于短信文本的长度较短，故 $c$ 的值不宜太大，否则容易引起关联到很多语义不相关的词汇。

词向量维数 $k$ 可以根据所解决问题的要求和语料库的大小确定。为更能体现各个词组在语义空间上的分布情况，应该使词向量的维数尽量大一些，但前提是需要有大而均匀的语料库作为支撑，为避免发生过拟合现象，也需要更高要求的模型表达能力和硬件计算能力。

在确定了合适的 $c$ 和 $k$ 以后，模型可根据数据集训练出各个词组的词向量。

在短信文本中，词与词、句子与句子之间在语境上更依赖前后文的关系，而不是割裂孤立起来的，这种语义关联直接影响到最终的识别结果[21]。与用向量空间模型作为特征提取方法相比，Word2Vec 更重视词组前后文之间的语序和语法上的联系；而且所有词组的词向量各个维度上都有数值，这解决了传统特征提取方法中数据稀疏和维数灾难问题。

## 3 卷积神经网络模型研究

CNN 是深度学习中一种具有代表性的结构。CNN 由多层神经网络组成，本质上是神经网络的一种拓展，一个典型的 CNN 由输入层、卷积层、池化层和全连接层四部分构成。

CNN 区别于其他神经网络之处在于 CNN 采用了局部连接和权值共享技术，对局部微小的特征更加敏感，这更有利于提取短信文本的特征信息。通过对短信文本进行卷积和池化操作，可以在词和词的位置信息之间提取出更多的抽象特征值和相关语义信息。CNN 模型起初用在图像处理领域中，以每张图片的像素矩阵作为模型的输入。灰度图像以一个二维矩阵表示，由于每个像素点只能有一个值表示颜色，因此灰度图像也称为单通道图像；彩色图像也叫 RGB 通道图像，以一个三维矩阵表示，每一维矩阵分别代表红色(R)通道、绿色(G)通道和蓝色(B)通道。本文用于垃圾短信识别的 CNN 模型如图 3 所示，模型的输入层是类似于灰度图像的表示各个短信文本的二维特征矩阵，因此每条短信都只有一个通道。

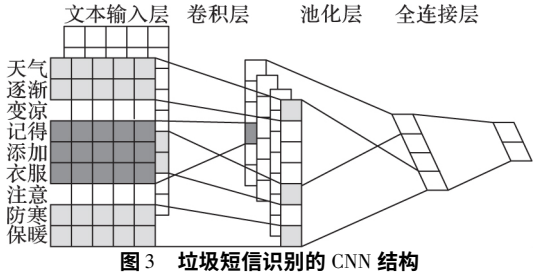


图 3 垃圾短信识别的 CNN 结构

Fig. 3 CNN structure of spam messages recognition

### 3.1 输入层设计

输入层的作用是把提取到的表示短信的特征矩阵作为输入数据传送到 CNN 模型中，并和下一层的卷积层连接起来。为了在垃圾短信识别任务中应用卷积神经网络模型，模型的输入应该是类似于图像像素的特征矩阵。

要实现以词向量来表示整个短信文本，可以采取一种较为简单的组织词向量的方式：假设一条短信文本经过分词处理后由“天气 逐渐 变凉 记得 添加 衣服 注意 防寒 保暖”组成。这 9 个词语的词向量依次为 $\boldsymbol{\Omega}_1,\boldsymbol{\Omega}_2,\cdots,\boldsymbol{\Omega}_9$，按照词组的顺序作纵向排列，就得到一个表示该短信的特征矩阵 $\boldsymbol{\Omega}$，可表示为：

$$\boldsymbol{\Omega}=\begin{pmatrix}\boldsymbol{\Omega}_1\\ \boldsymbol{\Omega}_2\\ \vdots\\ \boldsymbol{\Omega}_9\end{pmatrix}=\begin{bmatrix}\omega_{11} & \omega_{12} & \cdots & \omega_{1k}\\ \omega_{21} & \omega_{22} & \cdots & \omega_{2k}\\ \vdots & \vdots & & \vdots\\ \omega_{91} & \omega_{92} & \cdots & \omega_{9k}\end{bmatrix} \tag{2}$$

式(2)的排列方式可以通过图像形象化地表示为图 3 的文本输入层所示的格式。

为便于后续的卷积层和池化层提取出更加抽象的高层次文本特征，需要将各条短信的特征矩阵设置为同一大小。取特征矩阵的宽度为各个词组的词向量的维数 $k$。然而由于短信长短不一，特征矩阵的高度应该由短信数据集 $D$ 中长度最长的短信决定。若短信数据集 $D$ 中各条短信在经过中文分词处理后，词组数目最多的一条短信含有 $m$ 个词，则该短信由 $m$ 个 $k$ 维向量按照词组的顺序进行纵向排列成 $m\times k$ 的特征矩阵表示。

因此，$m$ 只是整个短信数据集 $D$ 中含词汇数目最多的一条短信的长度，而对于含词汇数目小于 $m$ 的短信则需要作补零处理，即对于含有 $q(q<m)$ 个词的任意一条短信 $d$，其特征矩阵中的前 $q$ 行由这 $q$ 个词的词向量表示，后面 $m-q$ 行用 0 行填补。短信数据集 $D$ 中的任意一条短信 $d$ 补 0 后的特征矩阵 $\tilde{\boldsymbol{\Omega}}$ 可表示为：

$$\tilde{\boldsymbol{\Omega}}=\begin{pmatrix}\boldsymbol{\Omega}_1\\ \boldsymbol{\Omega}_2\\ \vdots\\ \boldsymbol{\Omega}_q\\ \boldsymbol{\Omega}_{q+1}\\ \vdots\\ \boldsymbol{\Omega}_m\end{pmatrix}=\begin{bmatrix}\omega_{11} & \omega_{12} & \cdots & \omega_{1k}\\ \omega_{21} & \omega_{22} & \cdots & \omega_{2k}\\ \vdots & \vdots & & \vdots\\ \omega_{q1} & \omega_{q2} & \cdots & \omega_{qk}\\ 0 & 0 & \cdots & 0\\ \vdots & \vdots & & \vdots\\ 0 & 0 & \cdots & 0\end{bmatrix} \tag{3}$$

### 3.2 卷积层设计

卷积层作为 CNN 的核心组成部分，其主要功能是用卷积核对输入层的各个特征矩阵进行卷积操作，得到更加抽象的高层文本特征。卷积核是一个 $h\times k$ 的权重矩阵，可表示为：

$$\boldsymbol{W}=\begin{bmatrix}w_{11} & w_{12} & \cdots & w_{1k}\\ w_{21} & w_{22} & \cdots & w_{2k}\\ \vdots & \vdots & & \vdots\\ w_{h1} & w_{h2} & \cdots & w_{hk}\end{bmatrix} \tag{4}$$

高度 $h$ 可根据文本的长度设置合适的值，本文设计的卷积核的宽度等于词向量的维数 $k$，用符号 $\boldsymbol{W}\in\mathbf{R}^{hk}$ 代表整个卷积核。卷积核以大小为 1 的步长从文本特征矩阵的顶部由上至下开始扫描并和卷积核窗口内的矩阵进行卷积计算，每一步都提取出一个新的特征值。用 $\boldsymbol{X}_i$ 表示一条短信中每一个词的词向量，也是卷积核窗口中第一个词的词向量。$\boldsymbol{X}_i,\boldsymbol{X}_{i+1},\cdots,\boldsymbol{X}_{i+h-1}$ 依次表示卷积核窗口中 $h$ 个词的词向量。卷积核窗口中的部分特征矩阵可表示为：

$$\tilde{\boldsymbol{X}}_i=\begin{pmatrix}\boldsymbol{X}_i\\ \boldsymbol{X}_{i+1}\\ \vdots\\ \boldsymbol{X}_{i+h-1}\end{pmatrix}=\begin{bmatrix}x_{i1} & x_{i2} & \cdots & x_{ik}\\ x_{(i+1)1} & x_{(i+1)2} & \cdots & x_{(i+1)k}\\ \vdots & \vdots & & \vdots\\ x_{(i+h-1)1} & x_{(i+h-1)2} & \cdots & x_{(i+h-1)k}\end{bmatrix} \tag{5}$$

每一步的特征值可以通过式(6)得到：

$$V_i=f(\boldsymbol{W}*\tilde{\boldsymbol{X}}_i+b)=f\left(\sum_{r=1}^{h}\sum_{s=1}^{k}w_{rs}x_{(i+r-1)s}+b\right) \tag{6}$$

其中：$b$ 为偏置量，符号 $*$ 为卷积运算符，$f(\cdot)$ 为激活函数。常见的激活函数有逻辑函数、正切函数和线性整流函数等。本文采用线性整流函数即 Relu 函数[22]作为激活函数，可以使模型较快地收敛，其计算公式为：

$$f(x)=\max(0,x) \tag{7}$$

对每一步的卷积核窗口中的部分特征矩阵 $\tilde{\boldsymbol{X}}_1,\tilde{\boldsymbol{X}}_2,\cdots,$

$\widetilde{X}_{m-h+1}$ 均用式(6) 进行卷积操作可以得到一个特征图:

$$V = (V_1, V_2, \cdots, V_{m-h+1})^{\mathrm{T}} \tag{8}$$

$V$ 表示卷积核对整个特征矩阵扫描完毕后提取出的新的特征图。由于卷积核的窗口高度为 $h$,特征矩阵为 $m$ 行,所以扫描完一个特征矩阵需要 $m-h+1$ 次,即特征图 $V$ 的高度也为 $m-h+1$;卷积核的窗口宽度等于词向量的维数,所以特征图 $V$ 的宽度为 1。窗口高度不同的卷积核可以提取出高度也不同的特征图 $V$,意味着可以从不同的角度提取短信特征,得到更为完善的特征信息。

根据上述分析可知,CNN 的强大之处就在于其卷积块强大的特征提取能力。为了从不同角度提取特征而不增加计算的复杂度,Kim 所提出的模型[12] 使用了三种不同的卷积窗口。由于不同的短信,其前后文的语义关联不一样,为从短信文本中提取较为完备的特征,因此本文也使用三种窗口高度不同的卷积核提取相应的局部语义特征。因短信文本的长度一般在几个到几十个词之间,可将卷积核的高度分别设置为 3、4、5,最终得到 3 种不同粒度的特征。CNN 提取特征的具体流程如图 4 所示。

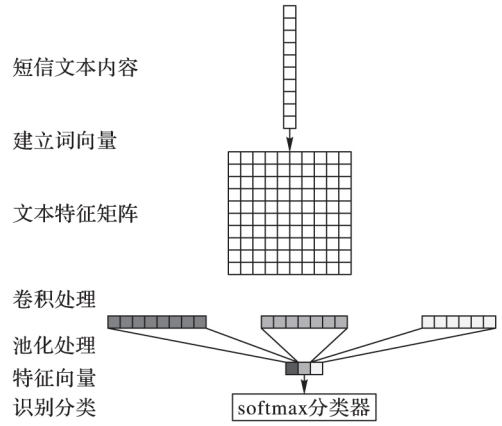


图 4　CNN 提取特征流程

Fig. 4　Flow chart for extracting features using CNN

三种卷积核的大小分别设置为 $3\times k$、$4\times k$、$5\times k$,通过卷积核扫描运算后,获得 3 种粒度大小分别为 $(m-2)\times 1$、$(m-3)\times 1$ 和 $(m-4)\times 1$ 的特征图,其过程如图 4 的卷积处理所示。本文设置每种尺寸卷积核各 128 个,即对每条短信文本的特征矩阵输入,一种卷积核可得到 128 个特征图,卷积层输出 384 个特征图。

### 3.3　池化层设计

针对自然语言处理过程中卷积操作的结果,CNN 的池化层的功能是对局部信息再总结,使卷积层提取到的文本向量维数减小,防止出现过拟合。池化操作包括常用的 Max-Pooling 和 Average-Pooling。Max-Pooling 是输出所接收到特征图 $V$ 中的最大值,可以认为这个最大值是短信的最显著的特征。Average-Pooling 是输出特征图 $V$ 中所有值的平均值。常见的自然语言处理任务中通常使用 Max-Pooling 方法,而且文献[23] 也认为 Max-Pooling 更适合于文本分类,因此本文以 Max-Pooling 为基础进行研究。

考虑到短信文本的长度差异,本文采取 1-max pooling 的池化策略,即对每一个特征图只取一个最大值 $\hat{V}$,$\hat{V}$ 可表示为:

$$\hat{V} = \max\{V_1, V_2, \cdots, V_{m-h+1}\} \tag{9}$$

$\hat{V}$ 即为短信的最优局部特征,过程如图 4 的池化处理所示。卷积层输出的 384 个特征图经过池化处理后得到 384 个特征值。

### 3.4　全连接层设计

池化操作得到的 384 个特征值是对垃圾短信识别最具影响的局部特征。全连接的操作是对这 384 个特征值进行融合,本文以串接的方式把这 384 个特征值串联起来,形成一个固定长度的融合特征向量来表示短信的语义特征,过程如式(10)所示:

$$\boldsymbol{V}_{\mathrm{message}} = \hat{V}^{1} \oplus \hat{V}^{2} \oplus \cdots \oplus \hat{V}^{384} \tag{10}$$

其中⊕表示串接操作。从式(10) 可以看出,特征向量的长度不再与短信的长度相关联,每条短信的有效特征都是一个 384 维的向量。把表示短信的语义特征向量 $\boldsymbol{V}_{\mathrm{message}}$ 输入最后的 softmax Regression 分类器[24] ,从全局的角度对特征进行分析,进而完成垃圾短信的识别。

## 4　实验及结果分析

### 4.1　实验环境

本文的实验环境:编程语言为 python3.6,深度学习框架为 Tensorflow1.0,内存 8 GB,操作系统为 Windows 10,处理器为 Intel Core i5。

### 4.2　实验数据

本文训练词向量所使用的维基中文语料库包含 232 894 个中文文本。所有文本经过分词处理后,共有 1 亿多个中文词组和少量的英语单词。

参与实验的短信数据集分为垃圾短信(negative)和正常短信(positive)两大类,短信样本总数量为 10 万条,其中正常短信数量为 5 万条,垃圾短信数量为 5 万条。这些数据在初始时即具有特征标签。

### 4.3　实验设计

维基中文语料库的规模足以训练高质量和高维度的词向量,而且每个词组所附带的语义信息也可以被尽量地保存下来。高维度的词向量虽然可以更充分地表达词组的语义特征,但它也会增加 CNN 模型参数的数量,同时增加过拟合的风险。因此经过综合权衡之后,本文将词向量的维度设置为 100,即 $k=100$。在训练词向量时为避免关联到更多语义不相关的词汇和缩短训练的时间,窗口大小设置为 5,即 $c=5$。本文用到的短信数据集在通过中文分词、去除停止词等操作后,长度最长的一条短信含有 100 个词,即 $m=100$;因此,每条短信都须表示为 $100\times 100$ 的特征矩阵。

在以词向量组成表示短信文本特征矩阵的基础上,结合 CNN 模型来完成垃圾短信的识别,并与传统的机器学习模型进行比较以验证 CNN 模型的性能优势,具体的实验设计方案如下:

1)词向量 + CNN 模型。将表示短信文本的 $100\times 100$ 特征矩阵作为 CNN 的输入参与到模型的训练测试过程中。

2)词向量 + 传统的机器学习模型。在同一数据上,同样用分布式特征提取方法获取每个词组的词向量,以每条短信中各个词所对应的词向量的各维度上的均值组合成表示短信的特征向量。该组实验采用文献[5]、文献[7] 和文献[10] 所用到的分类器模型,包括:朴素贝叶斯、逻辑回归、支持向量机

和随机森林。朴素贝叶斯是基于贝叶斯定理与特征条件独立假设的分类方法，而且模型比较简单，在处理大量数据时效率较高。逻辑回归是用于二分类最基本的算法之一，该模型处理数据时计算量非常小，效率高，所需的存储资源较低。支持向量机可以解决高维的大型特征空间问题，能够处理非线性特征。随机森林是通过集成学习的思想将多棵决策树集成的一种算法，具有模型简单、容易实现和计算开销小的特点，适合用于大型数据集。

### 4.4 模型训练测试和CNN参数设置

目前在训练CNN时所采用的方法仍然是传统的梯度下降法。欲使收敛效果最佳，可以采用批量梯度下降法，但是这种方法需要所有的短信文本数据参与每一次的迭代过程，这使收敛速度受到严重限制。欲使收敛速度较高，亦可采用随机梯度下降法，每次只需要一个短信文本数据参与迭代过程，但这种方法在每次迭代中没有用到全部信息，收敛效果不佳，很可能最终只是收敛到局部最优解。为使训练过程既可以获得较好的收敛效果也可以达到较高的收敛速度，本文使用mini-batch梯度下降法进行CNN模型的训练，即用一部分短信参与迭代过程。短信样本容量为100 000，为实现收敛效果和收敛速度之间的平衡，将每批样本大小设置为1 000。

在模型的测试过程中，用交叉熵(Cross Entropy)损失函数来衡量模型的损失并采用L2范数对参数进行约束，这样做的优点是防止隐藏层参数自适应，从而减轻过拟合的程度。假设短信样本为$(\tilde{\Omega},Y)$，其中$\tilde{\Omega}$是表示文本内容的特征矩阵，$Y$表示类别标签并可表示为：

$$Y=\begin{cases}(1,0), & \text{正常短信}\\(0,1), & \text{垃圾短信}\end{cases} \tag{11}$$

交叉熵损失约束表示如式(12)所示：

$$L(Y,G(\tilde{\Omega}))=-\left[(1-G(\tilde{\Omega}))\ln(1-G(\tilde{\Omega}))+Y\ln G(\tilde{\Omega})\right]+\frac{\lambda}{2}\|\theta\|_2 \tag{12}$$

其中：$G$表示CNN模型，$G(\tilde{\Omega})$表示模型的输出向量，$\lambda$为正则项系数，$\theta$为参数向量。

在训练全连接层参数时，为避免发生过拟合，采取dropout策略使部分神经节点失效，即一些已经训练过的参数在每一次更新时将被随机选择丢弃[25]。本文参照Kim所提出的模型[12]，在训练神经网络模型的参数时将dropout策略的概率值设置为0.5，选择随机抛弃一半的参数。本文选择应用最广泛、一般而言效果最好的AdamOptimizer优化器。

CNN模型训练的参数设置如表1所示。

表1 CNN算法参数设置

Tab. 1 CNN algorithm parameter settings

| 参数名称 | 参数值 | 参数名称 | 参数值 |
|---|---|---|---|
| 卷积核的个数 | 128 | dropout | 0.5 |
| 卷积核窗口高度 | 3,4,5 | 训练次数 | 40 |
| 词向量维度 | 100 | 学习率 | 0.1,0.5,0.8 |
| Mini-batch | 1 000 | | |

### 4.5 评估标准

本文以准确率作为评估模型在垃圾短信识别任务中的有效性的指标，假设短信文本数据集$D$中包含$N$条短信，短信样本$\tilde{\Omega}_i$的类别标签和分类结果可分别表示为$Y_i$和$\hat{Y}_i$，准确率[26]为可表示为：

$$\text{识别准确率}=\frac{1}{N}\sum_{i=1}^{N}|Y_i=\hat{Y}_i| \tag{13}$$

其中$|\cdot|$是指示函数，当里面的内容为真时取值为1，当内容为假时取值为0。

本文为验证CNN模型在垃圾短信识别中的可靠性，采取十折交叉验证法来评判模型在短信测试集上的准确率。具体措施为：将短信数据集中的所有100 000个样本分为10等份，每份包含5 000条正常短信和5 000条垃圾短信。每次实验以其中1份样本进行测试，另外9份样本进行训练。10份样本轮流作为测试集，共进行10次实验，以这10次实验所测得结果的平均值作为评估模型的指标。

### 4.6 实验结果分析对比

#### 4.6.1 CNN模型实验结果

CNN的收敛速度与学习率的大小有关，当学习率过小时，无法快速找到好的下降方向，导致训练时间较长，收敛较慢；当学习率太大时，会造成神经网络出现超调或剧烈振荡；因此需设置合理的学习率以获得预期的结果。实验结果如图5所示，可以观测出当学习率分别为0.1和0.8时，整个神经网络收敛相对稳定，但是收敛的速度相对较低；当学习率为0.5时模型收敛得最快，准确率最高。从图6可以观测出，随着学习的不断深入，损失在不断减小，当学习率为0.5时损失收敛得最快。最终准确率收敛为99.5%，损失收敛为0.03。

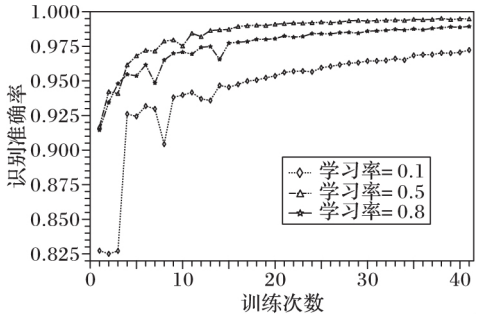


图5 不同学习率下的准确率变化

Fig. 5 Change of accuracy under different learning rates

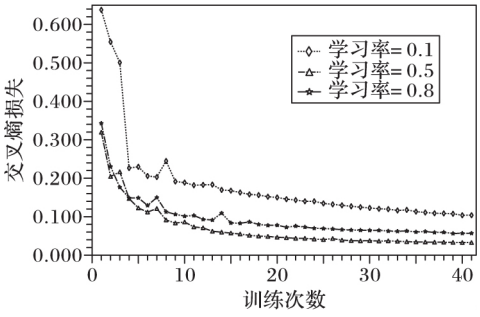


图6 不同学习率下的损失变化

Fig. 6 Change of cost under different learning rates

#### 4.6.2 模型对比分析

本文使用朴素贝叶斯(NB)、逻辑回归(LR)、支持向量机(SVM)和随机森林(RF)四种机器学习模型与CNN模型进行对比实验，不同模型下的准确率对比如图7所示。使用CNN模型的识别准确率为99.5%，而使用朴素贝叶斯、逻辑回归、支持向量机和随机森林四种机器学习模型的准确率分别为94.4%、97.1%、96.3%和95.8%。可以看出，基于深度学习的CNN模型对垃圾短信的识别准确率高于传统的机器学习模型，因此CNN模型在垃圾短信识别任务中能有效提高识别

的准确率。

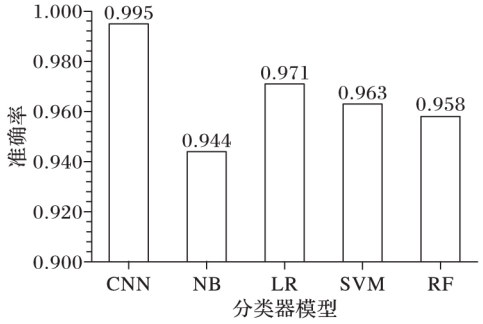


图 7　不同模型下的准确率比较

Fig. 7　Comparison of accuracy under different models

朴素贝叶斯、逻辑回归、支持向量机和随机森林对垃圾短信识别率偏低的主要原因分别是:1) 朴素贝叶斯模型比较简单,在处理大量数据时效率较高,但由于朴素贝叶斯模型假设样本各个特征之间是相互独立的,而这个假设在垃圾短信识别中往往不成立。主要因为训练的词向量不仅仅表示单个词组,而且还保留了词组在语义空间上的联系。当特征个数较多或各特征之间关联性较大时,分类效果不好。朴素贝叶斯模型在预测时需要知道先验概率,而先验概率的计算取决于模型的假设,很多时候因为模型假设的原因导致预测效果不佳。2) 逻辑回归模型处理数据时计算量非常小,效率高,所需的存储资源较低;但是当特征空间较大时,容易出现欠拟合,阻碍了分类效果的提升。3) 支持向量机可以解决高维的大型特征空间问题,能够处理非线性特征,但是当样本容量比较大时,数据处理的速度较低、耗时长;对非线性特征的处理没有通用的方法,选取合适的核函数也较困难。4) 随机森林模型的训练和预测速度较高,能有效处理大型数据集;但是当数据集中的样本数据的噪声较大时,随机森林模型容易引起过拟合。究其原因,传统的机器学习模型在训练的过程中对输入的特征没有更进一步的分析;而 CNN 首先利用卷积层和池化层对输入的特征矩阵提取出更高层次的文本特征,然后再通过分类器进行模式分类。

因此,通过对以上结果的分析可知,在同一短信样本集和特征提取方法下,基于深度学习的 CNN 模型相比传统的机器学习算法,在垃圾短信识别任务中有更出色的性能优势,能有效提高识别的准确率,同时也证明 CNN 模型在短信数据有噪声的情况下有更好的健壮性。

## 5　结语

垃圾短信泛滥一直是一个使人们非常困扰的社会问题。为有效识别垃圾短信并建立可靠的模型,本文首先深入研究了常用的文本特征提取方法,提出了一组更适合于垃圾短信识别的文本数据特征提取方法和识别分类方法:采用 Skip-Gram模型根据维基中文语料库自动训练出短信样本集中所有中文词组的词向量,这在一定程度上解决了短信表示面临的数据稀疏、维数过高和词间语义关系建模困难等问题,并且将词向量按照词序纵向排列成每条短信的分布式特征矩阵。然后在充分了解了深度学习的理论和应用后,尝试用深度学习中的 CNN 模型的来解决垃圾短信识别问题。将表示文本的特征矩阵作为 CNN 的输入层,然后利用 3 个窗口大小不同的卷积核提取 3 种相应粒度的更高层的文本特征。为使所提取到的特征向量的维数进一步减小,防止出现过拟合,采用 1-max pooling 策略进一步筛选特征,再将筛选出来的特征在全连接层重新组合成特征向量输入分类器中完成垃圾短信识别任务。模型对比实验表明,CNN 模型在垃圾短信识别任务中相比传统的机器学习模型有更高的准确率,证明了 CNN 模型在垃圾短信识别任务中的有效性和性能优势。

然而,基于 CNN 模型的垃圾短信识别分类任务中也存在着一些不足需要加以改进,如网络结构中参数过多、训练时间过长。因此为了提高短信文本识别分类的效率,缩短训练的时间,未来将尝试在分布式平台上进行卷积神经网络的训练测试。为了不使识别的准确率因样本数据类别分布的差异而产生偏差,本文使用的短信据集在类别分布上保持均衡,但是现实中往往是正常短信多于垃圾短信,而现有的分类器在设计时都假设数据集中的样本类别分布均衡,如果用这些分类器对样本类别分布不均衡的数据集进行分类,将导致分类器性能的下降并且也会因数据分布不均衡引入额外的误差而对最终的分类结果产生影响。对类别分布不均衡的短信数据集进行垃圾短信识别将是下一步的研究重点。

参考文献 (References)

[1] 陈功平,沈明玉,王红,等. 基于内容的短信分类技术[J]. 华东理工大学学报(自然科学版),2011,37(6):770 - 774. (CHEN G P, SHEN M Y, WANG H. et al. SMS classification technology based on content [J]. Journal of East China University of Science and Technology (Natural Science Edition), 2011, 37(6):770 - 774.)

[2] ZHANG L, MA J, WANG Y. Content based spam text classification: an empirical comparison between English and Chinese [C] // INCOS '13: Proceedings of the 2013 5th International Conference on Intelligent Networking and Collaborative Systems. Washington, DC: IEEE Computer Society, 2013: 69 - 76.

[3] SHARMA N, GAGANPREETKAUR, VERMA A. Survey on text classification (spam) using machine learning [J]. International Journal of Computer Science and Information Technologies, 2014, 5(4): 5098 - 5102.

[4] SHAHI T B, YADAV A. Mobile SMS spam filtering for nepali text using Naïve Bayesian and support vector machine [J]. International Journal of Intelligence Science, 2014, 4(1): 24 - 28.

[5] 李润川,酋红英,申圣亚,等. 基于多特征融合的垃圾短信识别[J]. 山东大学学报(理学版),2017,52(7):73 - 79. (LI R C, ZAN H Y, SHEN S Y, et al. Spam messages identification based on multi-feature fusion [J]. Journal of Shandong University (Natural Science), 2017,52(7):73 - 79.)

[6] 黄文明,莫阳. 基于文本加权 KNN 算法的中文垃圾短信过滤[J]. 计算机工程,2017,43(3):193 - 199. (HUANG W M, MO Y. Chinese spam message filtering based on text weighted KNN algorithm [J]. Computer Engineering, 2017, 43(3): 193 - 199.)

[7] SETHI P, BHANDARI V, KOHLI B. SMS spam detection and comparison of various machine learning algorithms [C] // Proceedings of the 2017 International Conference on Computing and Communication Technologies for Smart Nation. Piscataway, NJ: IEEE, 2017, 28 - 31.

[8] CHAN P P K, YANG C, YEUNG D S, et al. Spam filtering for short messages in adversarial environment [J]. Neurocomputing, 2015, 155(C): 167 - 176.

[9] MA J, ZHANG Y, LIU J, et al. Intelligent SMS spam filtering using topic model [C] // Proceedings of the 2016 International Conference on Intelligent Networking and Collaborative Systems. Piscataway, NJ: IEEE, 2016: 380 - 383.

[10] CHOUDHARY N, JAIN A K. Towards filtering of SMS spam mes-

sages using machine learning based technique [M] // Advanced Informatics for Computing Research. Berlin: Springer, 2017: 18 - 30.

[11] BAKER S, KROHONEN A, PYYSALO S. Cancer hallmark text classification using convolutional neural networks [EB/OL]. [2018-01-11]. https://www.repository.cam.ac.uk/bitstream/handle/1810/270037/BIOTXTM2016.pdf; jsessionid=FAE7EA1B196FA600CC643D798DD04A0D? sequence=1.

[12] KIM Y. Convolutional neural networks for sentence classification [EB/OL]. [2018-01-11]. http://www.anthology.aclweb.org/D/D14/D14-1181.pdf.

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

[14] KESORN K, POSLAD S. An enhanced bag-of-visual word vector space model to represent visual content in athletics images [J]. IEEE Transactions on Multimedia, 2012, 14(1): 211 - 222.

[15] CASTELLS P, FERNANDEZ M, VALLET D. An adaptation of the vector-space model for ontology-based information retrieval [J]. IEEE Transactions on Knowledge and Data Engineering, 2007, 19(2): 261 - 272.

[16] TURNEY P D, PANTEL P. From frequency to meaning: vector space models of semantics [J]. Journal of Artificial Intelligence Research, 2010, 37(1): 141 - 188.

[17] HINTON G E. Learning distributed representations of concepts [C] // Proceedings of the 8th Annual Conference of the Cognitive Science Society. New York: Clarendon Press, 1986: 1 - 12.

[18] HU B, TANG B, CHEN Q, et al. A novel word embedding learning model using the dissociation between nouns and verbs [J]. Neurocomputing, 2016, 171: 1108 - 1117.

[19] BIAN J, GAO B, LIU T Y. Knowledge-powered deep learning for word embedding [C] // Proceedings of the 2014 Machine Learning and Knowledge Discovery in Databases, LNCS 8724. Berlin: Springer, 2014: 132 - 148.

[20] MIKOLOV T, SUTSKEVER I, CHEN K, et al. Distributed representations of words and phrases and their compositionality [EB/OL]. [2018-01-12]. http://www.eecs.wsu.edu/~sji/classes/DL16/CNN-text/word2vec2.pdf.

[21] 郑世卓,崔晓燕. 基于半监督LDA的文本分类应用研究[J]. 软件,2014,35(1):46 - 48. (ZHEN S Z, CUI X Y. Research on text classification based on semi-supervised LDA [J]. Computer Engineering and Software, 2014, 35(1): 46 - 48.)

[22] GLOROT X, BORDES A, BENGIO Y. Deep sparse rectifier neural networks [EB/OL]. [2018-01-12]. http://proceedings.mlr.press/v15/glorot11a/glorot11a.pdf.

[23] ZHANG X, ZHAO J, LeCUN Y. Character-level convolutional networks for text classification [EB/OL]. [2018-01-12]. http://www.eecs.wsu.edu/~sji/classes/DL16/CNN-text/5782-character-level-convolutional-networks-for-text-classification.pdf.

[24] HINTON G E, SALAKHUTDINOV R R. Replicated softmax: an undirected topic model [C] // Proceedings of the 2009 International Conference on Neural Information Processing Systems. [S.l.]: Curran Associates Inc., 2009: 1607 - 1614.

[25] SRIVASTAVA N, HINTON G, KRIZHEVSKY A, et al. Dropout: a simple way to prevent neural networks from overfitting [J]. Journal of Machine Learning Research, 2014, 15(1): 1929 - 1958.

[26] 宗成庆. 统计自然语言处理[M]. 北京:清华大学出版社,2008:352 - 363. (ZONG C Q. Statistical Natural Language Processing [M]. Beijing: Tsinghua University Press, 2008: 352 - 363.)

**LAI Wenhui**, born in 1994, M. S. candidate. His research interests include machine learning, natural language processing.

**QIAO Yupeng**, born in 1981, Ph. D., associate professor. Her research interests include Boolean network, game theory, natural language processing.

(上接第2468页)

[15] 袁丽. 基于文本的情绪自动归因方法研究[D]. 哈尔滨:哈尔滨工业大学,2014:49 - 60. (YUAN L. The Study on Text-based Emotion Cause Detection [D]. Harbin: Harbin Institute of Technology, 2014: 49 - 60.)

[16] GHAZI D, INKPEN D, SZPAKOWICZ S. Detecting emotion stimuli in emotion-bearing sentences [C] // Proceedings of the 2015 International Conference on Intelligent Text Processing and Computational Linguistics, LNCS 9042. Berlin: Springer, 2015:152 - 165.

[17] GUI L, HU J, HE Y, et al. A question answering approach for emotion cause extraction [C] // Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Copenhagen: EMNLP, 2017: 1593 - 1602.

[18] 慕永利,李旸,王素格. 基于E-CNN的情绪原因识别方法[J]. 中文信息学报,2018,32(2):120 - 128. (MU Y L, LI Y, WANG S G. Emotion cause detection based on ensemble convolution neural networks [J]. Journal of Chinese Information Processing, 2018, 32(2): 120 - 128.)

[19] RABINER L R. A tutorial on hidden Markov models and selected applications in speech recognition [J]. Readings in Speech Recognition, 1990, 77(2): 267 - 296.

[20] 张子睿,刘云清. 基于BI-LSTM-CRF模型的中文分词法[J]. 长春理工大学学报(自然科学版),2017,40(4):87 - 92. (ZHANG Z R, LIU Y Q. Chinese word segmentation based on bi-directional LSTM-CRF model [J]. Journal of Changchun University of Science and Technology, 2017, 40(4): 87 - 92.)

[21] GUI L, WU D, XU R, et al. Event-driven emotion cause extraction with corpus construction [C] // Proceedings of the 2016 Conference on Empirical Methods on Natural Language Processing. Austin: EMNLP, 2016: 1639 - 1649.

[22] HUANG Z, XU W, YU K. Bidirectional LSTM-CRF models for sequence tagging [EB/OL]. [2017-12-27]. https://arxiv.org/pdf/1508.01991.pdf.

This work is partially supported by the Surface Project of National Natural Science Foundation of China (61772378), the Surface Project of Hubei Natural Science Foundation (2018CFB690).

**ZHANG Chen**, born in 1992, M. S. candidate. His research interests include nature language processing, deep learning.

**QIAN Tao**, born in 1975, Ph. D., associate professor. His research interests include nature language processing, machine learning.

**JI Donghong**, born in 1967, Ph. D., professor. His research interests include nature language processing, data mining.