基于语义增强的改进混合特征选择的文本分类

高洁云,赵逢禹,刘 亚
(上海理工大学 光电信息与计算机工程学院,上海 200093)

摘 要:如何从文本中抽取出能够体现文本特点的关键特征,抓取特征到类别之间的映射是文本分类核心问题之一。传统的词袋模型的优点是将每个词视为一个特征,而缺点是计算成本会随特征数量和文本与特征之间的关系的增加而增加,并且没有考虑文本特征自身的语义关系,语义关系的优势是获取文本和特征之间的相关性。针对这个问题,提出一种增强混合特征选择方法,该方法使用混合特征选择进行降维,然后再使用词向量对低频词进行语义增强。为了验证增强的混合特征选择对文本分类的作用,构建了两个实验,使用LSTM算法进行分类模型训练与测试。对爬取的71 825个新闻文本数据进行实验表明,基于语义的增强混合特征选择方法在文本分类时既提高了分类效率又能保证分类精度。

关键词:混合特征选择;语义分析;词向量;文本分类;LSTM
中图分类号:TP391.41 **文献标识码**:A **文章编号**:1673-629X(2021)01-0024-06
doi:10.3969/j.issn.1673-629X.2021.01.005

Text Classification of Modified Hybrid Feature Selection Based on Semantic Enhancement

GAO Jie-yun, ZHAO Feng-yu, LIU Ya
(School of Optoelectronic Information and Computer Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China)

Abstract: One of the core problems of text classification is how to extract the key features that can reflect the characteristics of the text from the text and capture the mapping between features and categories. The advantage of the traditional bag-of-words model is to treat each word as a feature, while the disadvantage is that the calculation cost increases with the increase in the number of features and the relationship between text and features, and the semantic relationship of the text features themselves is not considered. The advantage of semantic relationships is to get the correlation between text and features. Aiming at this problem, we propose an enhanced hybrid feature selection method which uses hybrid feature selection to reduce the dimension, and then uses word embedding to semantically enhance low-frequency words. In order to verify the effect of enhanced hybrid feature selection on text classification, two experiments are constructed, using the LSTM algorithm to train and test the classification model. Experiments on 71 825 news text data crawled show that the semantic-based enhanced hybrid feature selection method not only improves the classification efficiency but also ensures the classification accuracy in text classification.

Key words: hybrid feature selection; semantic analysis; word-embedding; text classification; LSTM

0 引 言

文本分类是基于文本的内容将文本分配给一个或多个预定义类别[1]。传统上,文本分类是基于词的向量空间模型,每个文本被表示为高维空间中的向量,文本之间的相似性是基于词匹配计算的,即取决于文本之间词特征的共现率。因此,Salton等人[2]提出的矢量空间模型,也称为特征向量模型,有助于在二维空间中用词频建模来表示文本。通常,文本表示包括特征提取和特征加权两个步骤,特征提取可捕获文本上下文的重要特征,特征加权是各特征的赋予不同的权重值,表明其在该特定文本以及整个语料库中的重要性。词袋(BOW)[2]模型是把文本中的词与词频抽取出来,构成词袋,形成文本的特征空间,但是文本特征空间的大小随文本大小增加而变得极度稀疏,并且没有考虑词间的语义。

基于语义分析的技术可用于文本分类过程以提高

收稿日期:2020-03-20 **修回日期**:2020-07-22
基金项目:国家自然科学基金(61803264)
作者简介:高洁云(1993-),女,硕士,研究方向为文本分类;赵逢禹,博士,教授,CCF会员(E20-20 15341M),研究方向为软件工程与软件质量控制;刘 亚,博士,副教授,研究方向为信息安全、密码学。

性能和准确性。但是,大多数现有的文本分类方法都使用统计加权方法来计算特征加权。Deerwester 等[3]提出了一种称为潜在语义分析(LSA)的纯统计技术,通过合并与具有相似含义的词相关联的维度来很好地解决同义词问题,但是多义性问题仍没有很好地解决。Gabrilovich 和 Markovitch[4]提出了一种显式语义分析(ESA)技术,利用维基百科中的概念将自然语言文本表示成细粒度的语义概念,对自然语言处理的研究发现,ESA 在文本分类方面是成功的。然而,ESA 算法在很大程度上依赖于维基百科的现有知识,这非常耗时。Banik 等人[5]已经提出了类似的方法,其目的在于开发概念本体,其中从维基百科提取的背景知识用作语义核以改进文档表示。但是,基于词特征空间的文档表示有时不能很好地反映文档之间的语义相关性。

在 Google 推出词向量 word2vec[6]后,出现了一种新的文档表示方法,陈磊等人[7]通过 word2vec 词向量特征选择的方法来创建分类特征。词向量是一种分布式表示,其中词以低维和实值向量表示,在向量空间中,语义上相似的词往往具有相似的向量。最近的研究已经应用词向量来提高文本分类任务的性能[8]。

为了更进一步优化文本分类性能并提高准确性,针对海量文本数据,该文提出了一种新的混合特征选择技术(hybrid feature selection,HFS),对海量文本使用 HFS 技术删除不相关且冗余的文本特征,通过去除不必要的特征来减少数据维度。由于 word2vec 会忽略词内部的形态特征这一问题,在应用混合特征选择之后,提出使用预训练的 fastText 词向量技术用于发现语义上相似的特征,以增强原始特征集,并应用这种增强的特征集进行分类。该文把增强特征方法(enhanced hybrid feature selection,EHFS)与两种特征选择方法(AC 和 MAD)以及著名的文本分类算法 LSTM 一起使用,并通过实验验证了 EHFS 对文本分类的有效性。

1　相关技术

word2vec[9-10]从输入中构建词汇表,然后学习词的向量表示为每个词生成一个向量。但它忽略了词内部的形态特征,比如:“文本数据”和“文本”,这两个词有公共字符“文本”,即它们的内部形态类似,但是在传统 word2vec 中,这种词内部形态信息因为被转换成不同的词向量而被忽略。为了克服这个问题,该文引入了 fastText,它可以计算两个向量在语义上的相似度,对相似特征进行语义上的增强。

1.1　fastText

Facebook 于 2016 年推出的 fastText 是一个开源的词向量计算和文本分类工具[11-12]。在 fastText 模型中,字符级别的 n-gram 信息和词内部顺序的隐藏信息可以用于词表示,对于“文本数据”这个词,假设 n 的取值为 3,则它的 trigram 有:“<文本”,“文本数”,“本数据”,“数据>”。其中,<表示前缀,>表示后缀,于是,可以用这些 trigram 来表示“文本数据”这个词,进一步,“文本数据”的词向量可以用这 4 个 trigram 的向量叠加来表示。对于未登录词的词向量可以使用词典中相应的子词向量之和的平均来进行表示。

fastText 还可以计算两个向量在语义上的相似度。首先从输入中构建词汇表,然后学习词的向量表示。对于两个词向量,使用余弦相似度值确定它们的相似度,该值越大,两个向量在语义上就越接近,这些词向量可以用作分类问题中的特征。

1.2　LSTM 算法

LSTM 是由 Hochreiter&Schmidhuber 在 1997 年提出的[13],它是 RNN 的一种特殊类型,可以学习长期依赖关系[14]。LSTM 结构如图 1 所示。

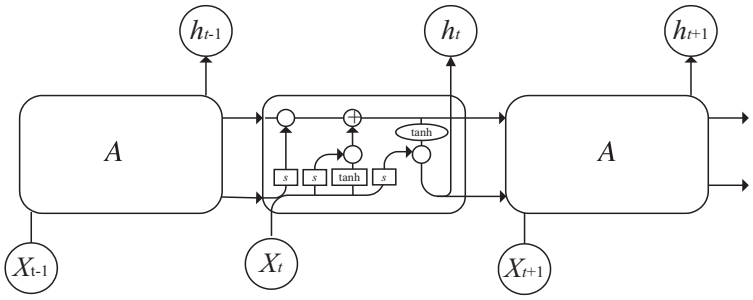

图 1　LSTM 结构

RNN 是具有内部存储器的网络,可高效地预测时间序列[15]。在 RNN 中,信息从每个神经元流到其同层中的其他每个神经元。LSTM 是 RNN 单元的扩展[16],它克服了 RNN 单元的缺点。与传统的循环神经网络相比,LSTM 仍然是基于 x_t 和 h_{t-1} 来计算 h_t,只不过内部的结构加入了输入门、遗忘门以及输出门三个门和一个内部记忆单元 c_t[17]。输入门确定何时将当前计算的新状态更新到记忆单元中;遗忘门确定何时将前一步记忆单元中的信息遗忘掉;输出门确定何时将当前的记忆单元输出。

2 增强的混合特征选择 EHFS

文本经过预处理后一般含有数以万计个不同的词组,这些词组所构成的向量规模同样也很庞大,计算机运算成本就比较高,因此进行特征选择,对文本分类具有重要的意义。该文采用改进的混合特征选择方法,从全局特征中提取最具区分度和较多文档中出现的特征[18]。由于相似特征对分类价值不大,该文采用基于特征向量的绝对余弦(AC)相似度,去除部分相似的冗余特征。然后采用基于词频(term frequency,TF)的平均绝对差值(MAD)与基于词文档频率(document frequency,DF)的平均绝对差值(MAD)相结合的方法选择特征。

2.1 绝对余弦(AC)

绝对余弦(AC)基于相似性得分去除冗余特征,将两个词经过预训练好的 fastText 模型转成词向量,计算余弦相似度,如果相似度得分过高,则删除其中之一。特征的绝对余弦相似度可以通过式(1)计算,其中 w_i 和 w_t 是词,$v(w_i)$ 与 $v(w_t)$ 是对应的词向量。

$$\mathrm{sim}(w_i,w_t)=\left|\frac{v(w_i)*v(w_t)}{\|v(w_i)\|\ \|v(w_t)\|}\right| \tag{1}$$

2.2 基于词频的平均绝对差值

该方法将每个词作为一个特征,词频作为该词的特征值 x_{ij}(x_{ij} 是具有第 j 个文本的第 i 个词的词频),MAD 值是由特征值 x_{ij} 和所有其他文本特征的均值 $\overline{x_i}$ 之差来分配,MAD 值越大,这些词对文本分类越有价值。式(2)给出第 i 个特征的词频 MAD 值,其中词频平均值 $\overline{x_i}$ 按照式(3)计算得到。

$$\mathrm{MAD}_i=\frac{1}{n}\sum_{j=1}^{n}\left|x_{ij}-\overline{x_i}\right| \tag{2}$$

$$\overline{x_i}=\frac{1}{n}\sum_{j=1}^{n}(x_{ij}) \tag{3}$$

2.3 基于词文档频率的平均绝对差值

该方法将每个词作为一个特征,该词文档频率(DF,包含该词的文档数目)作为它的特征值,MAD 值是由特征值的 DF 和所有其他文本特征的 $\overline{\mathrm{DF}}$ 之差来分配,MAD 值越大,这些词对文本分类越有价值。式(4)给出第 i 个特征的词文档频率 MAD 值,其中文档频率均值 $\overline{\mathrm{DF}_i}$ 按照式(5)计算得到。

$$\mathrm{MAD}_i=\frac{1}{n}\sum_{i=1}^{n}\left|\mathrm{DF}_i-\overline{\mathrm{DF}_i}\right| \tag{4}$$

$$\overline{\mathrm{DF}_i}=\frac{1}{n}\sum_{i=1}^{n}(\mathrm{DF}_i) \tag{5}$$

2.4 混合特征选择 HFS

为了得到对文本分类重要的特征,该文提出了一个混合特征选择方法 HFS,获取最具区分度和较多文档中出现的特征。算法 1 给出了 HFS 算法的描述。

算法 1:混合特征选择 HFS。

输入:经过数据预处理后的文本特征集 $T=\{t_1,t_2,\cdots,t_f\}$。

输出:特征子集 $\mathrm{features}_{\mathrm{HFS}}$。

(1)在文本特征集 T 中,用绝对余弦 AC 进行特征选择去除冗余特征获取特征子集 FS_1。

(2)在 FS_1 中,用基于词频的平均绝对差值 MAD 进行特征选择获得 MAD 值较大的特征子集 FS_2。

(3)在 FS_1 中,用基于词文档频率的平均绝对差值 MAD 进行特征选择获得 MAD 值较大的特征子集 FS_3。

(4)选取 $\mathrm{FS}_2\cup\mathrm{FS}_3$ 特征子集作为 $\mathrm{features}_{\mathrm{HFS}}$。

2.5 增强混合特征选择 EHFS

通过混合特征选择方法 HFS 获取的特征子集 $\mathrm{features}_{\mathrm{HFS}}$ 并未考虑低频且对分类有重要价值的特征,由于这类特征出现的频率低,在分类训练算法中的作用常被忽略。为了解决这一问题,该文使用特征增强方法对这类特征在语义上进行增强。算法 2 给出了 EHFS 算法的描述。

算法 2:特征增强方法 EHFS。

输入:经过数据预处理后的文本特征集 $T=\{t_1,t_2,\cdots,t_f\}$。特征子集 $\mathrm{features}_{\mathrm{HFS}}$。

输出:特征子集 $\mathrm{features}_{\mathrm{enhanced}}$。

(1)在混合特征选择方法 HFS 的基础上,选择 MAD 值较高但是词频较低的部分特征。

(2)使用预训练好的 fastText 模型计算每个特征 f 与未标记数据集的其他所有特征之间的余弦相似度得分。

(3)选择每个特征语义上最相似的前 k 个特征,如果这 k 个特征在 T 中,且不在特征子集 $\mathrm{features}_{\mathrm{HFS}}$ 中,则将其添加到最终的特征集 $\mathrm{features}_{\mathrm{enhanced}}$ 中。

3 文本分类器训练

为了评估针对文本分类提出的增强混合特征选择方法 EHFS 的性能,该文将分类算法 LSTM 应用于使用该算法生成的最终文本向量上得到分类模型,然后验证模型的正确性。图 2 描述了文本分类流程。

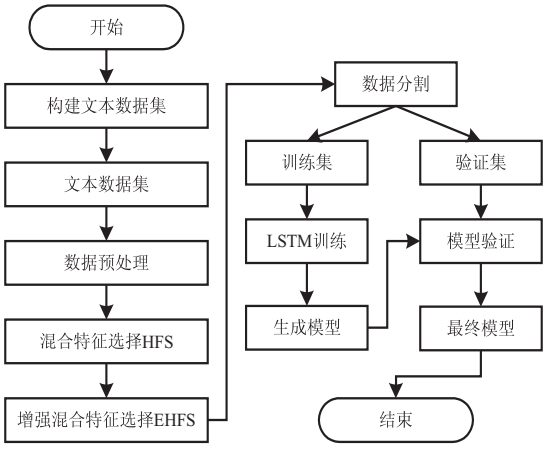

图 2 基于语义分析的文本分类流程

4　实验过程

为了更好地验证增强的混合特征选择 EHFS 算法对文本分类的作用,构建了一个实验。该实验首先抓取不同类型的文本数据,构建文本数据集,然后进行数据预处理并采用增强混合特征选择方法进行特征选取,最后采用 LSTM 模型对文本分类,并基于分类结果验证该方法的性能与效率。

4.1　语料集数据

使用 python 抓取工具 scrapy 收集了来自新浪、今日头条、腾讯、百度、人民网五个热门站点共 71 825 个文本数据,数据收集自过去 3 年到 2019 年 12 月之间。类别和文本数量如下:文化(6 024),经济(13 167),科技(7 580),法律(4 122),教育(9 809),军事(2 675),旅游(8 352),娱乐(1 954),历史(5 124),体育(13 018)。由于来自所有网站的合并数据集并未将数据公平地分布在所有类别中,因此该研究仅限于十个类别。

4.2　文本数据预处理

4.2.1　文本词频

该文用 python 编写了对各类别中每个文本处理程序,提取词干,统计每个类别对应的词干和词频,然后在数据库中构建文本词频表,表的字段有文本文档、词干、词频,以法律文本 law0. txt 为例,law0. txt、诉讼、13。

4.2.2　文本的词文档频率

根据文本词频的统计,统计每个词的词文档频率 DF 。以每个类别的文本特征为例,针对每个词干计算被多少个文本所覆盖,将结果插入到词文档频率表中,表的字段有词干、词文档频率,以法律词干“诉讼”为例,诉讼、217。

4.3　增强混合特征选择 EHFS

在获取每个类别每个文本的特征后(包括词干与词频),通过 merge_stem 函数可以将每个类别的文本进行词干合并,最后得到该类别所有词干集合 all_features。以法律为例,将 law0. txt、law1. txt … law4122. txt 的特征合并到 law. txt 文本中。

4.3.1　词干去冗余

根据类别得到合并后的文本,然后计算任意两个词之间的的绝对余弦 AC 值,这里词的特征向量通过 fastText 计算。如果特征相似度得分超过设置的 0.8,就认为这两个特征是冗余关系,可去掉其中 DF 值低的词。本文通过调用 removed _ redundancy (all _ features)程序获得去冗余后的数据集 FS_1。

4.3.2　提取重要特征 FS_2

本实验提供 mad_tf 程序,该程序基于每个类别特征子集 FS_1 中各类别词频的平均绝对差值获取每个类别的特征子集 FS_2。该程序的处理方法是:(1)计算 FS_1 中所有词的词频均值(合并文本时,将相同的词干,所对应的词频相加);(2)根据每个词干 T_i 的词频,计算 MAD_{tf_i} ;(3)将词干的 MAD_{tf} 由大到小排序,取前 60% 的词干加入到特征子集 FS_2 中,获得的特征子集 FS_2 的信息(文本文档、词干、词频、词频的平均绝对差值 MAD_{tf})放在数据库的表 1 中。

表 1　词频的平均绝对差值

文本文档	词干	词频	词频的平均绝对差值
law. txt	诉讼	6 329	11.85
…	…	…	…
law. txt	仲裁被诉人	1	1.417

4.3.3　提取重要特征 FS_3

本实验提供 mad_df 程序,该程序基于每个类别特征子集 FS_1 中各类别词文档频率 DF 的平均绝对差值获取每个类别的特征子集 FS_3。该程序的处理方法是:(1)计算所有词文档频率的均值;(2)根据每个词干 T_i 的词文档频率,计算 MAD_{df_i} ;(3)将词干的 MAD_{df} 由大到小排序,取前 60% 的词干加入到特征子集 FS_3 中,获得的特征子集 FS_3 的信息(词干、词文档频率、词文档频率的平均绝对差值 MAD_{df})放在数据库的表 2 中。

表 2　词文档频率的平均绝对差值

词干	词文档频率	词文档频率的平均绝对差值
诉讼	217	1.386
…	…	…

4.3.4　获取增强特征子集 $features_{enhanced}$

根据算法 2,该文提供了 enhanced_features 程序对低频特征在语义上进行增强,来获取每个类别增强后的特征子集 $features_{enhanced}$ 。

4.4　训练 LSTM 分类模型

获得每个类别的特征子集后,使用 TensorFlow 库中的 LSTM 的实现类 BasicLSTMCell 用于构建和训练深度学习模型。LSTM 的建模工作是:(1)拆分数据集为 80% 训练集和 20% 验证集;(2)定义一个 LSTM 的序列模型,模型的第一层是嵌入层(Embedding),它将上面获得的每个类别的特征子集 features 经过 fastText 模型构建的词向量而形成的矢量矩阵作为模型的输入;(3)输出层则为包含 10 个分类的全连接层。因为是多分类问题,所以激活函数设置为“softmax”,损失函数为分类交叉熵。

5　实验结果分析

基于上面的实验过程,构建了两个实验。实验 1

采用混合特征选择 HFS 选择的特征训练分类模型,实验 2 采用增强混合特征选择方法 EHFS,在混合特征选择后对低频特征做语义上的增强,然后用增强后的特征训练分类模型。针对两个实验,从精确率(Precision)、召回率(Recall)、F 值(F1-score)[19-21]的得分来评判分类的效果。

实验 1:混合特征选择 HFS 和 LSTM 模型的评估。

表 3 描述了混合特征选择 HFS 进行特征选择获取特征子集,然后将其作为 LSTM 模型的输入数据,进行文本分类,分类的实验结果中,军事文本的分类精度为 96%,经济文本的分类精度为 68%。

表 3 使用 HFS 和 LSTM 的分类结果

类别	Precision	Recall	F1-score
文化	0.89	0.60	0.72
经济	0.68	0.92	0.78
科技	0.82	0.76	0.79
法律	0.82	0.35	0.49
教育	0.83	0.91	0.87
军事	0.96	0.25	0.40
旅游	0.81	0.84	0.82
娱乐	0.83	0.86	0.84
历史	0.77	0.81	0.79
体育	0.93	0.93	0.93
平均值	0.83	0.72	0.74

表 4 使用经典的算法 TF-IDF 来做一组对比实验。实验表明 HFS 方法的分类效果优于 TF-IDF 算法。

表 4 HFS 和经典算法 TF-IDF 的平均实验结果对比

性能指标	HFS	TF-IDF
Precision	0.83	0.80
Recall	0.72	0.67
F1-score	0.74	0.73

实验 2:增强混合特征选择方法 EHFS 和与 HFS 对比。

首先使用该文提出的基于语义的增强混合特征选择方法 EHFS 获取特征子集,然后将其作为 LSTM 模型的输入数据,进行文本分类,得到的分类结果见表 5。

表 5 使用 EHFS 和 LSTM 的分类结果

类别	Precision	Recall	F1-score
文化	0.90	0.75	0.82
经济	0.86	0.86	0.86
科技	0.85	0.79	0.82
法律	0.82	0.75	0.78

续表 5

类别	Precision	Recall	F1-score
教育	0.92	0.87	0.89
军事	0.87	0.90	0.88
旅游	0.88	0.85	0.86
娱乐	0.89	0.86	0.87
历史	0.91	0.86	0.88
体育	0.96	0.93	0.95
平均值	0.89	0.84	0.86

从实验结果可以得知,考虑了语义的特征选择实验 2 的分类效果较实验 1 有很大改善。并且实验 2 在准确度上也比实验 1 更高一些。实验的分类准确性如图 3 所示。

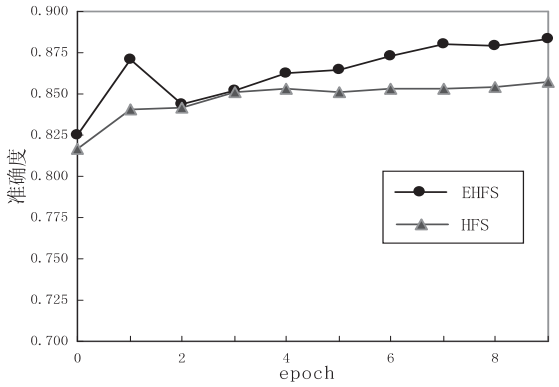

图 3 实验 1 和实验 2 的分类准确性与 epoch 的关系图

6 结束语

在混合特征选择上,提出一种新的增强混合特征选择方法 EHFS。该方法先使用改进的混合特征选择,再使用特征相似分析对低频且对分类有重要价值的特征进行语义上的增强。然后将得到的特征向量矩阵作为 LSTM 模型的输入进行模型的训练。数据集来自五个热门网站的数据,并将其效果与指标进行了比较。实验结果表明,提出的增强混合特征选择方法 EHFS 比不考虑语义只进行混合特征选择 HFS,分类效果有很好的改善。文本数据集被分为十个不同的类别,其中体育的准确度最高为 96%,法律类的准确度为 82%。实验基于相对较小的数据集,还可以获取更多文本数据集来进行改进。另外,该文只是基于 10 类文本进行分类,还需要增加数据类别验证文本分类方法的实用性。

参考文献:

[1] YANG Y M,LIU X. A re-examination of text categorization on methods[C]//Proceedings of the 22nd annual international ACM SIGIR conference research and development in information retrieval. Berkeley, California, USA: ACM,

1999:42-49.

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

[3] DEERWESTER S, DUMAIS S T, FURNAS G W, et al. Indexing by latent semantic analysis[J]. Journal of the American Society For Information Science, 1990,41(6):391-407.

[4] GABRILOVICH E, MARKOVITCH S. Wikipedia-based semantic interpretation for natural language processing[J]. Journal of Artificial Intelligence Research, 2009,34(1):443-498.

[5] BANIK P, GAIKWAD S, AWATE A, et al. Semantic analysis of Wikipedia documents using ontology[C]//2018 IEEE international conference on system, computation, automation and networking (ICSCA). Pondicherry: IEEE, 2018:1-6.

[6] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient estimation of word representations in vector space[J]. Computer Science, 2013,25(5):213-219.

[7] 陈　磊,李　俊. 基于词向量的文本特征选择方法研究[J]. 小型微型计算机系统,2018,39(5):991-994.

[8] 马　力,李沙沙. 基于词向量的文本分类研究[J]. 计算机与数字工程,2019,47(2):281-284.

[9] 熊富林,邓怡豪,唐晓晟. Word2vec 的核心架构及其应用[J]. 南京师范大学学报:工程技术版,2015,15(1):43-48.

[10] 周　练. Word2vec 的工作原理及应用探究[J]. 科技情报开发与经济,2015,25(2):145-148.

[11] 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: EACL, 2017:427-431.

[12] BOJANOWSKI P, GRAVE E, JOULIN A, et al. Enriching word vectors with subword information[J]. Transactions of the Association for Computational Linguistics, 2017,5(1):135-146.

[13] 孙运淼,林　锋,周激流. 长短时记忆网络在移动场景中的应用研究进展[J]. 现代计算机,2017(35):10-15.

[14] DOĞAN E, KAYA B, MÜNGEN A, et al. Generation of original text with text mining and deep learning methods for turkish and other languages[C]//2018 international conference on artificial intelligence and data processing (IDAP). Malatya, Turkey: IEEE, 2018:1-9.

[15] DU C, HUANG L. Text classification research with attention-based recurrent neural networks[J]. International Journal of Computers Communications & Control, 2018,13(1):50-61.

[16] 朱小燕,王　昱,徐　伟. 基于循环神经网络的语音识别模型[J]. 计算机学报,2001,24(3):213-218.

[17] AL-SMADI M, TALAFHA B, AL-AYYOUB M, et al. Using long short-term memory deep neural networks for aspect-based sentiment analysis of Arabic reviews[J]. International Journal of Machine Learning and Cybernetics, 2018,10(8):2163-2175.

[18] BHOPALE A P, KAMATH S S. Novel hybrid feature selection models for unsupervised document categorization[C]//2017 international conference on advances in computing, communications and informatics (ICACCI). Udupi: IEEE, 2017:1471-1477.

[19] 汪　静,罗　浪,王德强. 基于 Word2Vec 的中文短文本分类问题研究[J]. 计算机系统应用,2018,27(5):209-215.

[20] 王　盛,樊兴华,陈现麟. 利用上下位关系的中文短文本分类[J]. 计算机应用,2010,30(3):603-606.

[21] 王义真,郑　啸,后　盾,等. 基于 SVM 的高维混合特征短文本情感分类[J]. 计算机技术与发展,2018,28(2):88-93.