Advances in Applied Mathematics 应用数学进展, 2022, 11(1), 269-277
Published Online January 2022 in Hans. http://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2022.111034

Hans汉斯

基于Doc2vec的微博评论情感倾向研究

李荟珍

南京信息工程大学数学与统计学院，江苏 南京

收稿日期：2021年12月21日；录用日期：2022年1月11日；发布日期：2022年1月24日

摘　要

该文针对疫苗接种的相关微博评论进行情感倾向分析，首先利用基于神经网络的Doc2vec模型训练文本向量，继而使用支持向量机(SVM)、随机森林(RF)、逻辑回归(LR)三种机器学习的算法完成情感分类任务，且分别讨论了三种算法在四种不同的Doc2vec模型设定方案下的分类表现。其中Distributed Memory version of Paragraph Vector (PV-DM)算法训练的文本向量中，RF表现最优，在方案一与方案二上其F1分数值均为最高，分别为87.24%、87.50%。基于Distributed Bag of Words version of Paragraph Vector (PV-DBOW)算法训练的文本向量中，SVM表现最优，在方案三与方案四上其F1分数值达到最高，分别为84.11%、83.91%。

关键词

情感倾向，Doc2vec模型，文本分类，机器学习，微博评论

Research on Emotional Tendency of Microblog Comments Based on Doc2vec

Huizhen Li

School of Mathematics and Statistics, Nanjing University of Information Science & Technology, Nanjing Jiangsu

Received: Dec. 21st, 2021; accepted: Jan. 11th, 2022; published: Jan. 24th, 2022

Abstract

Firstly, Doc2vec model based on neural network was used to train the text vector, and then three machine learning algorithms including Support Vector Machine (SVM), Random Forest (RF) and Logistic Regression (LR) were used to complete the emotion classification task. The classification performance of the three algorithms under four different Doc2vec model setting schemes is discussed respectively. Among the text vectors trained by the Distributed Memory version of Paragraph Vector (PV-DM) algorithm, RF performs best, and its F1 score is the highest in plan 1 and

文章引用: 李荟珍. 基于 Doc2vec 的微博评论情感倾向研究[J]. 应用数学进展, 2022, 11(1): 269-277.
DOI: 10.12677/aam.2022.111034

plan 2, which are 87.24% and 87.50%, respectively. Among the text vectors trained by the Distributed Bag of Words Version of Paragraph Vector (PV-DBOW) algorithm, SVM has the best performance, and its F1 score is the highest in scheme 3 and scheme 4, which are 84.11% and 83.91% respectively.

Keywords

Emotional Tendency, Doc2vec Model, Text Classification, Machine Learning, Weibo Comments

1. 引言

随着互联网日新月异的发展，我们的日常生活方式产生了巨大的变化，移动智能设备为人们的家居出行提供了极大便利，各类平台也随之迅速崛起，例如电商平台上商品的用户评价或者是微博等社交平台的网友发言，这些文本数据中都潜藏着巨大的价值。如果能够有效挖掘出文本的情感倾向，我们可以据此了解消费者的消费偏好或者掌握大众的舆论动向等等，但是由于文本信息具有数据庞大、非结构化或者半结构化等特质，仅仅依靠人工识别语义信息是不可取的，而情感倾向研究正可以为此提供具体解决方向与技术手段。

情感分析(Sentiment Analysis)也叫情感倾向分析，作为时下自然语言处理领域热点之一，主要是根据文本所提供的语义信息将文本分为积极(褒义)、消极(贬义)两类或者更多类别。虽然情感分析是自然语言处理领域的一个新的研究方向，但其应用范围已经十分广泛。电商平台上的商家更希望通过消费者对商品的好差评来制定相应的策略；政府也可以通过从网友的评论挖掘出代表性的观点看法以便监督舆论动向，例如随着新冠疫苗在全国全面推广，通过收集网友在微博平台所发表的意见看法，从中了解群众的情感倾向，以便更好地推进后续疫苗接种工作。

2. 国内外研究现状

关于文本情感分析的方法目前主要分为三大类：基于情感词典的情感分析方法、基于机器学习算法的情感分析方法、基于深度学习的情感分析方法。

基于情感词典的情感分析法主要依赖于情感词典的制定。国外最早出现的英文情感词典是 SentiWordNet，后来 Hu (2004)等[1]建立的 Sentiment lexicon 词典，已经包含了 6800 多个情感词，基于此来判断文本的情感极性。Taboada (2011)等[2]提出了一种基于词典的文本情感提取方法，利用带有情感极性和强度注释的字典为每个情感词计算得分并求和，从而实现文本的情感极性分类。对于中文情感词典，比较具有代表性的主要有知网(HowNet)中文情感词典、台湾大学 TUSD 中文情感词典和大连理工大学中文情感词汇文本库，由于微博存在海量语料，极具分析价值。赵妍妍等(2017) [3]基于文本统计算法，根据微博海量文本数据构建了一个超大情感词典，能够大幅提高微博领域情感分类性能。在这之后为了克服单一的情感词典覆盖面不够广泛，吴杰胜(2020) [4]提出了可以同时使用多部情感词典并结合语义规则，能够有效提高微博情感分析的准确性。

基于机器学习的情感分析属于有监督学习方法，需要事先将原始语料分为训练集和测试集两部分，在划分的训练数据中训练分类器，继而对划分的测试集进行类别预测。Pang (2002) [5]提出将机器学习算

法应用到文本情感分析，将支持向量机 SVM (Support Vector Machine)、最大熵模型、朴素贝叶斯分类算法运用在电影评论的情感分类进行比较并得出 SVM 分类性能最佳的结果；刘志明等人(2012) [6]利用不同机器学习算法、不同的特征选取算法以及不同的特征项权重计算相结合面对微博文本情感分类，得出在电影方面，模型对微博评论和普通评论是通用的。孙建旺等人(2014) [7]是利用情感词典对微博文本动词形容词计算极性值并基于此提出位置权重方法进行特征提取继而采用支持向量机分类。李明等人(2019) [8]通过机器学习方法对商品评论分类后进一步利用了互信息扩充商品属性，完成细粒度的情感分析。

利用机器学习算法进行情感分类需要提取特征、人工标注，随着互联网数据规模与日俱增，上述方法无法满足分类需求。深度学习是多层神经网络在学习中的应用，解决了以往机器学习难以解决的大量问题。基于深度学习的情感分析主要是将深度学习的模型应用到情感分析的研究之中，通常可以取得不错的效果[9]。Mikolov 等人(2013) [10] [11]首次提出了 Word2vec 模型来训练词向量，该模型主要利用特定单词的上下文信息将一个词转化成低维的稠密向量，且语义相近的词会被映射到向量空间相近的位置；考虑到需要生成句子或者段落向量，提出 Doc2vec 模型，该模型克服了利用 Word2vec 训练的词向量简单加权生成段落向量从而忽略了语句语序的缺陷。秦胜君等人(2013) [12]提出基于限制玻尔兹曼机的无词汇标注情感分类算法，克服手工标注的方式适应能力较低的缺陷，一定程度上提升了泛化能力。梁军等人(2014) [13]提出结合递归神经网络与情感极性转移模型强化了获取文本关联特征的能力，该方法在处理微博语料时不仅分类效果好且节省了人工标注的工作量。

该文基于 Doc2vec 模型训练文本向量，得到向量矩阵，后续结合三种不同的机器学习算法进行情感分类，根本几种常用的分类评估指标来评估分类器性能。

3. 关于微博评论的情感倾向研究

拟选取疫苗接种阶段的微博在线评论做情感倾向研究，框架大致如下图 1 所示：

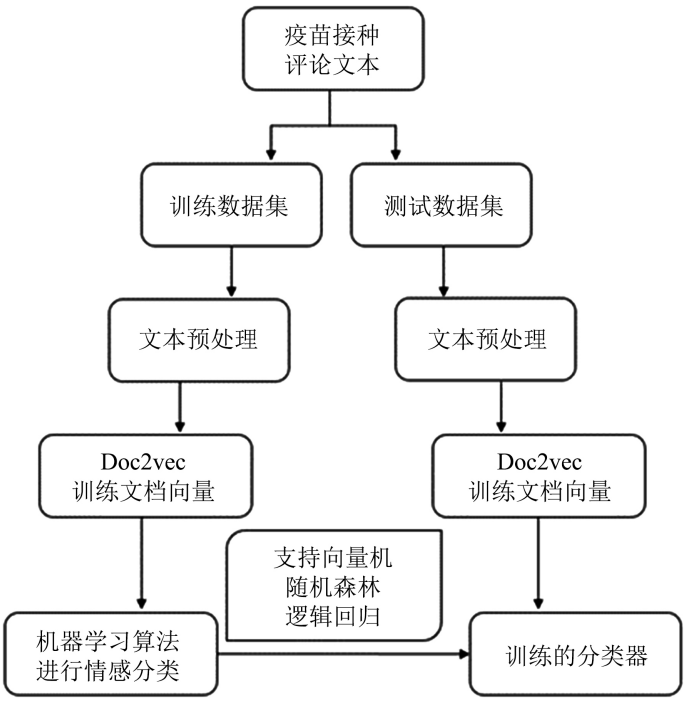

Figure 1. Emotion classification framework based on Doc2vec model
图 1. 基于 Doc2vec 模型的情感分类框架

训练 Doc2vec 模型

文本需要进一步转化才能为计算机所识别，即需要对这些文本信息进行结构化处理。Mikolov 等人于 2013 年[10]首次提出了 Word2vec 模型来训练词向量。能有效避免造成向量空间维数灾难。其训练过程中根据训练向量方法不同可分为两种模型，即 Continuous Bag-of-Word Model (CBOW)模型和 Skip-gram 模型。

为了克服 Word2vec 训练的词向量表示句子或者文本向量容易忽略语序的缺陷，Mikolov [11]同年提出 Doc2vec 方法，该方法同 Word2vec 一样也分有两种模型结构，分别为 Distributed Memory version of Paragraph Vector (PV-DM)和 Distributed Bag of Words version of Paragraph Vector (PV-DBOW)，DM 算法原理图如图 2 所示：模型主要建立在单层神经网络基础上，将段落向量与词向量取平均值或者相连接作为输入项输入网络，然后通过梯度下降法进行优化，继而实现对上下文中下一个将要出现的单词进行预测。DBOW 算法该模型是在随机梯度下降过程中每一次迭代采样一个文本窗口，再从中随机采样一个单词，实现输入段落向量进行单词的预测，这种方式也可以获得文档的向量表示。原理图如图 3 所示。

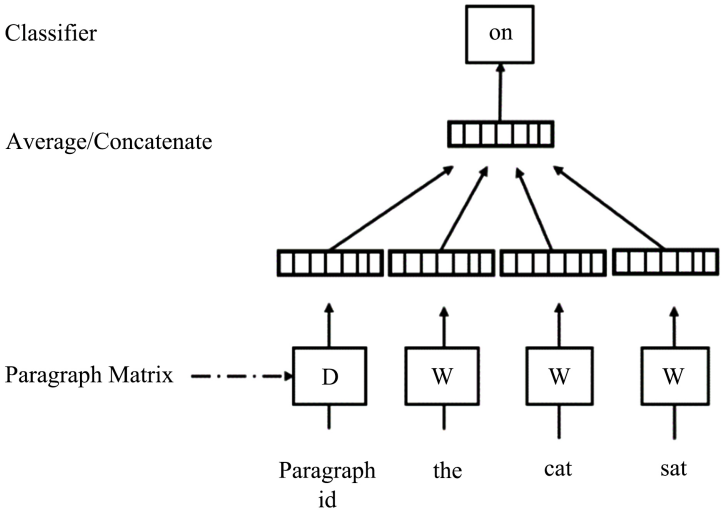

Figure 2. Principle of PV-DM algorithm
图 2. PV-DM 算法原理

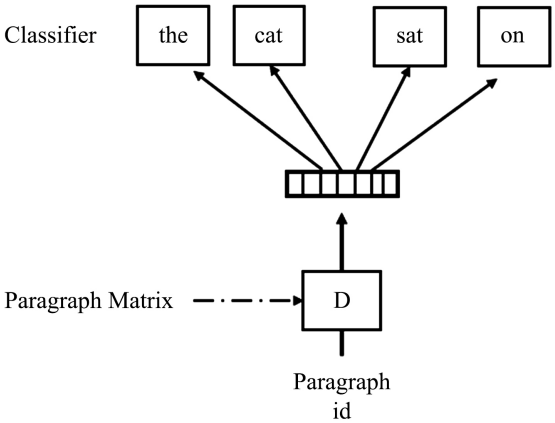

Figure 3. Principle of PV-DBOW algorithm
图 3. PV-DBOW 算法原理

4. 实验及分析

4.1. 微博语料的收集以及预处理准备

本文通过爬虫工具八爪鱼软件对人民日报微博官方平台关于疫苗的重要报道博文爬取相关留言评论，在完成语料的收集后，需要对冗杂的文本数据做初步的清洗与处理，文本数据的预处理流程大致为：文本去重、中文分词、去停用词等。由于本文所涉及的情感分析的机器学习方法是有监督学习，需要对训练数据进行标注，并且划分训练集、测试集，故对爬取的数据经过整理如下表，总评论数 900 条，经过去重、去除无关评论后剩 810 条。进行人工标注后，积极评论为 555 条，消极情感评论 255 条。训练集 567 条，测试集 243 条。具体概况及部分人工标注原始数据见表 1 与表 2。

Table 1. Experimental data structure
表 1. 实验数据结构

数据集类型	积极评论	消极评论	总计
训练集	389	178	567
测试集	166	77	243
总计	555	255	810

Table 2. Manual annotation of original document (part)
表 2. 原始文档人工标注(部分)

编号	评论文本	评论类别
24	一定会去积极接种的	1
48	中国做得好棒[好喜歡] 👍	1
206	没有第三期临床试验报告，说再多没用	0
318	不是应该用数据说话？只是简单的没有出现严重不良反应。搞笑呢	0
454	棒棒棒！一到位马上去接种👏👏 👏👏	1
577	中国疫苗，科学严谨，接种安全，让人放心，其他疫苗，不良反应，接二连三，可喜可贺。💉 💉	1
689	未有严重不良反应，这句话咋感觉听着那么不舒服呢	0
806	我想自愿打，但是现在单位一刀切，只要不是马上死，就都得打，不然就通报开除	0

由于爬取的微博语料含有大量的非中文字符例如表情符号、英文、标点符号，这些均需在预处理过程中进行剔除，经过处理之后，文档只保留中文字符。随后，本文主要使用 Jieba 分词系统完成分词操作，由于分词后的语料还存在一些无意义的停用词，例如“啊，呀，的，……”等等，可利用哈工大停用词库去除大量出现却无实义的词。经过预处理后的部分文档如下表 3：

Table 3. Pre-processed comment text (part)
表 3. 预处理后的评论文本(部分)

编号	类别	原始评论文本	预处理后评论文本
754	1	我们全家符合条件的该打的都打咯！	[“全家”，“符合条件”]
241	1	太棒啦！！	[“太棒”]
51	1	我们这里各单位也在统计疫苗接种人数	[“单位”，“统计”，“疫苗”，“接种”，“人数”]
167	1	支持他们最需要	[“支持”]
129	1	太棒了	[“太棒了”]
298	0	打了疫苗有什么副作用吗？	[“疫苗”，“副作用”]
168	0	医院里的医护人员强制接种，没有正式上市的疫苗算哪门子事？	[“医院”，“医护人员”，“强制”，“接种”，“正式”，“上市”，“疫苗”，“算”，“哪门子”，“事”]
544	1	我爱中国	[“爱”，“中国”]
458	1	辛苦了！谢谢。	[“辛苦”，“谢谢”]
501	1	中国免费治疗接种疫苗太强了吧哈哈哈！	[“中国”，“治疗”，“免费”，“接种”，“疫苗”，“太强”]

4.2. 基于 Doc2vec 模型训练文本向量

经过对文本的一系列预处理之后，原始的文本信息变成了单独的词语个体，但计算机仍然无法识别出这些词语中蕴涵的语义和情感。因此需要利用数学模型继续进行结构化处理步骤。

结构化处理主要使用 Doc2vec 将文本向量化表示，具体是由 python 里的 gensim 库实现模型训练、输出文本向量矩阵。gensim 库中 Doc2vec 函数中向量维数(vector_size)默认为 100 维，一般来说语料越多，该值越大。函数参数 dm 默认取 1 即采用 PV-DM 算法，后续将分别使用两种算法进行训练，向量维度分别设置 100 维和 50 维，分别观察各向量表示方案下不同分类器的分类效果。

4.3. 基于机器学习算法进行情感分类

该部分拟采用三种机器学习的算法分别对不同模型训练的结果进行分类：支持向量机 SVM (Support Vector Machine)、随机森林 RF (Random Forest)、逻辑回归 LR (Logistic Regression)，经过分类后需要评估分类效果，主要用到的指标有精确度(Accuracy)、查全率(Recall)、查准率(Precision)、*F*1 分数(*F*1 score)，分类结果的混淆矩阵如下表 4 所示，因此各指标具体表达公式如下：

Table 4. Confusion matrix for classification results
表 4. 分类结果的混淆矩阵

真实 / 预测	属于类别 y 的文本	不属于类别 y 的文本
判断属于类别 y 的文本	*TP*	*FP*
判断不属于类别 y 的文本	*FN*	*TN*

$$\text{准确率：accuracy} = \frac{TP+TN}{TP+FN+FP+TN} \tag{1}$$

$$\text{查全率：recall} = \frac{TP}{TP+FN}. \tag{2}$$

$$\text{查准率：precision} = \frac{TP}{TP+FP}. \tag{3}$$

$$F1\text{ 分数：}F1 = 2\times\frac{\text{precision}\cdot\text{recall}}{\text{precision}+\text{recall}}. \tag{4}$$

表 5 为特定的 Doc2vec 模型设定下三种机器学习算法的分类性能评估结果，图 4 也直观上对各个分类器的分类指标值进行对比。其中方案一表示模型参数设定为 dm = 1，size = 100，方案二表示模型参数设定为 dm = 1，size = 50，方案三表示模型参数设定为 dm = 0，size = 100，方案四表示模型参数设定为 dm = 0，size = 50。

Table 5. Comparison of classification results of machine learning algorithms under different Doc2vec model schemes
表 5. 各 Doc2vec 模型方案下机器学习算法分类结果比较

方案	评估指标	机器学习算法		
		SVM	RF	LR
方案一	Time	45 ms	304 ms	53 ms
	Accuracy	71.79%	81.68%	80.25%
	Recall	95.38%	91.28%	87.66%
	Precision	72.38%	83.67%	84.39%
	*F*1	82.27%	87.24%	85.93%
方案二	Time	30 ms	367 ms	68 ms
	Accuracy	72.50%	82.19%	81.30%
	Recall	94.10%	90.74%	89.96%
	Precision	73.41%	84.62%	83.98%
	*F*1	82.44%	87.50%	86.81%
方案三	Time	41 ms	348 ms	52 ms
	Accuracy	74.77%	76.17%	66.67%
	Recall	97.17%	89.71%	80.47%
	Precision	74.16%	78.82%	73.76%
	*F*1	84.11%	83.82%	76.85%
方案四	Time	21 ms	271 ms	61 ms
	Accuracy	74.44%	75.69%	72.15%
	Recall	97.18%	91.02%	87.40%
	Precision	73.84%	77.57%	75.78%
	*F*1	83.91%	83.71%	81.12%

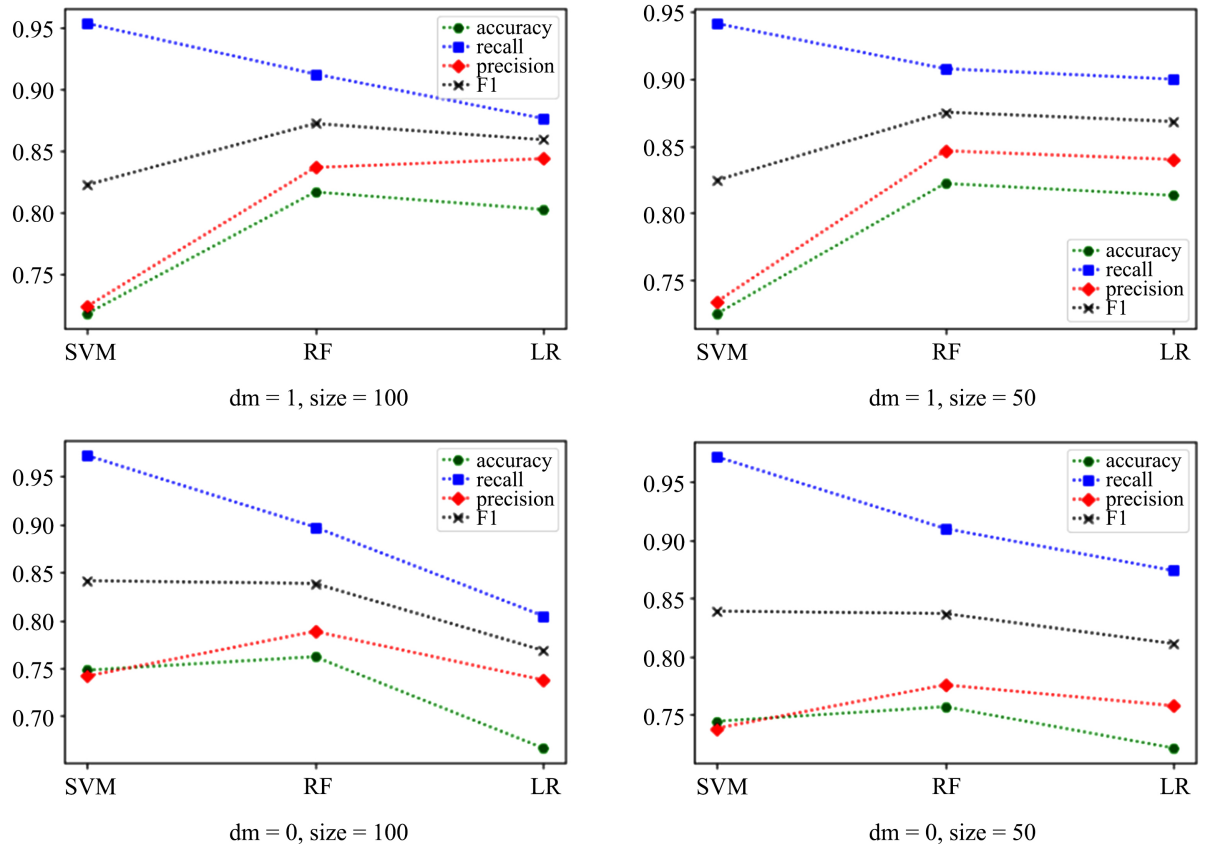

Figure 4. Comparison of machine learning algorithm classification performance under different Doc2vec model schemes
图 4. 各 Doc2vec 模型方案下机器学习算法分类性能对比

从上述结果来看，各个方案下，SVM 在查全率指标方面表现最好，而在评估分类性能时，查全率和查准率均各有所长与不足，故在关注上述指标同时还需关注 F1 分数等指标；而从各个方案的综合情况来看，RF 整体表现要比 SVM 和 LR 好。从具体各个指标来看，SVM 在方案三、四表现较方案一、二更优，即 SVM 在 PV-DBOW 算法下表现更优；LR 与 SVM 相反，在方案一、二上取得更佳的表现，即在 PV-DM 算法下表现更好。

5. 结论

通过使用 Doc2vec 模型对微博评论文本进行向量化，且分别使用 PV-DM 算法和 PV-DBOW 算法并设定不同的向量维度训练，对输出的向量矩阵进一步利用 SVM、RF、LR 完成情感分类任务，并根据正确率、查全率、查准率、F1 分数四个方面来评估分类性能。通过比较可以看出，对经过 PV-DM 算法训练的文本分类，随机森林表现最优，维数为 100 时 F1 分数值达到 87.24%，当维度为 50 时，F1 分数值达到 87.50%；对经过 PV-DBOW 算法训练的文本分类，支持向量机表现更优，维数为 100 时 F1 值达到 84.11%，维数为 50 时 F1 值达到 83.91%，实际操作中，支持向量机分类用时最短，能够节约训练成本。

参考文献

[1] Hu, M. and Liu, B. (2004) Mining and Summarizing Customer Reviews. *Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, New York, 168-177. https://doi.org/10.1145/1014052.1014073

[2] Taboada, M., Brooke, J., Tofiloski, M., *et al.* (2011) Lexicon-Based Methods for Sentiment Analysis. *Computational Linguistics*, **37**, 267-307. https://doi.org/10.1162/COLI_a_00049

[3] 赵妍妍, 秦兵, 石秋慧, 等. 大规模情感词典的构建及其在情感分类中的应用[J]. 中文信息学报, 2017, 31(2). 187-193.

[4] 吴杰胜. 基于多部情感词典和深度学习的中文微博情感分析研究[D]: [硕士学位论文]. 淮南: 安徽理工大学, 2020.

[5] Pang, B., Lee, L., Vaithyanathan, S., *et al.* (2002) Thumbs Up?: Sentiment Classification Using Machine Learning Techniques. *Conference on Empirical Methods in Natural Language Processing* (*EMNLP* 2002), **10**, 79-86. https://doi.org/10.3115/1118693.1118704

[6] 刘志明, 刘鲁. 基于机器学习的中文微博情感分类实证研究[J]. 计算机工程与应用, 2012, 48(01): 1-4.

[7] 孙建旺, 吕学强, 张雷瀚. 基于词典与机器学习的中文微博情感分析研究[J]. 计算机应用与软件, 2014, 31(07): 177-181.

[8] 李明, 胡吉霞, 侯琳娜, 等. 商品评论情感倾向性分析[J]. 计算机应用, 2019, 39(S02): 15-19.

[9] 王颖洁, 朱久祺, 汪祖民, 等. 自然语言处理在情感分析领域应用综述[J/OL]. 计算机应用. https://kns.cnki.net/kcms/detail/51.1307.TP.20210928.1611.014.html, 2021-09-29.

[10] Mikolov, T., Sutskever, I., Chen, K., *et al.* (2013) Distributed Representations of Words and Phrases and Their Compositionality. *Advances in Neural Information Processing Systems* 26 (*NIPS* 2013), Lake Tahoe, 5-10 December 2013.

[11] Mikolov, T., Chen, K., Corrado, G., *et al.* (2013) Efficient Estimation of Word Representations in Vector Space. arXiv:1301.3781 [cs.CL]

[12] 秦胜君, 卢志平. 基于限制玻尔兹曼机的无极性标注情感分类研究[J]. 科学技术与工程, 2013, 13(35): 10703-10707.

[13] 梁军, 柴玉梅, 原慧斌, 昝红英, 刘铭. 基于深度学习的微博情感分析[J]. 中文信息学报, 2014, 28(5): 155-161.