浙江理工大学学报(社会科学版),第40卷,第1期,2018年2月
Journal of Zhejiang Sci-Tech University (Social Sciences)
Vol. 40, No. 1, Feb. 2018

DOI:10.3969/j.issn.1673-3851(s).2018.01.005

基于词向量和情感本体的短文本情感分类

王正成,李丹丹
(浙江理工大学经济管理学院,杭州 310018)

摘　要:目前短文本情感分类主要采取统计自然语言处理、情感语义特性两种方式,而将这两种方式相结合进行情感分类的研究较少,故提出将这两种方式进行结合,设计基于词向量与情感本体相融合的短文本情感分类方法。首先利用Word2Vec模型训练词向量,以相加平均法合成短文本向量;在此基础上结合基于情感本体所得出的每条短文本的情感值,构建词向量与情感本体相融合的短文本表示模型;最后采用K最近邻分类算法完成短文本情感分类。相比传统的基于词向量、基于情感本体或其他单一技术路线的分类方法,词向量与情感本体相融合的分类方法在准确率、召回率、$F1$值均有明显的提升。

关键词:短文本情感分类;词向量;情感本体
中图分类号:G350　　**文献标志码**:A　　**文章编号**:1673-3851 (2018) 02-0033-06

目前以微博等众多平台为代表的自媒体不断涌现,用户成为信息传播的主动者,在网络社区中分享知识、经验和感受等,于是大量主观性的评论数据爆发性增长[1]。从大规模的文本数据中挖掘用户情感价值信息具有重要意义,而挖掘用户情感价值信息前提在于判别用户情感倾向。情感倾向的判别是在给定的文本分类模型下,依据文本内容所体现的情感特征,自动地对文本进行分类,从而帮助人们组织文本、挖掘文本信息。

一、文献综述

本文研究的短文本情感分类是根据文本内容所体现的用户意见的情感极性,将带有相同特定情感倾向的短文本归为一类,即文本情感分类[2]。目前短文本情感分类主要采取统计自然语言处理、情感语义特性两种方式[3]。统计自然语言处理是指利用文本中情感词的权重等特性对分类器进行训练来识别文本。Pang等[4]的研究表明,文本分类中若采用布尔值作为权重的Unigram,分类效果最好。Isidoros等[5]所设计的集成分类器架构是利用统计机器学习分类方法,确定情感的极性。杨锋等[6]根据词语顺序共现随机网络和情绪词表对短文本进行情感分类。杨小平等[7]对用户评论数据进行结构化处理和分析,通过构建网络节点和拓扑连接关系的知识图谱进行情感分析。

根据情感语义特性进行分类的方法是指利用情感词极性来判别文本情感倾向。Philipp[8]根据情感词极性将人类的情感划分成6种基本类型,包括愤怒、厌恶、恐惧、欢乐、悲伤和惊喜。Gamon[9]采用了NLPWin自然语言处理系统,利用情感文本的句法结构特征来进行文本分类。Tong[10]建立了电影评论情感词典,对每一个情感词汇的情感极性进行人工标记,以评论中情感词的极性判别情感倾向。桂斌等[11]根据情感词在正负微博文本中出现的概率对文本进行情感分类。史伟等[12]在划分情感本体(分为评价词本体和情感词本体)的基础上,构建了模糊情感本体作为分类依据。唐晓波等[13]对微博表情符号进行情感分析,构建了某微博产品的领域情感词典,进行微博产品评论的文本分类。

此外,张群等[14]提出了词向量与LDA相融合

收稿日期:2017-09-11　　网络出版日期:2018-01-11
基金项目:国家自然科学基金项目(71271192)
作者简介:王正成(1974-),男,浙江武义人,教授,博士,主要从事先进制造模式与技术、企业信息化方面的研究。

的短文本分类方法,将表示文本情感分类的两种方式相结合,解决了短文本特征稀疏问题及主题聚焦性差的问题。这表明情感分类可将情感词与统计处理两者不同的技术路线相结合进行。目前将这两种文本特征表示方法相结合的研究较少,大多数情感分类方法都是采用单一技术路线进行。因此本文以微博评论文本为研究对象,提出一种结合两种方式的短文本特征表示方法,即将情感词语义特性进行量化,与基于统计自然语言处理的分类方法相结合,利用情感本体词库的语义特性,融合基于Word2Vec的词向量,形成新的文本特征向量,之后通过KNN分类器进行短文本情感分类,查看分类效果评估该方法。

二、基于词向量和情感本体的短文本特征表示方法

Word2Vec采用神经网络语言模型(Neural network language model,NNLM)和N-gram语言模型,将每个词都表示成一个实数向量[15]。而情感本体则是对某个中文词汇或短语予以特性的描述。本文的情感本体主要借鉴大连理工大学信息检索研究室整理和标注的中文情感词汇本体库,其词汇情感共分为7大类21小类,并标注各个词汇的词性、情感类别及极性[16]。

本文将以上两种方法进行结合,其方法过程描述如下:

(一)基于Word2Vec的词向量合成

采用向量相加平均法得到文本数据集D的基于词向量的短文本特征,可以用公式表示为:

$$d_m' = \sum_{w_j \in d'_m} C(w_j)/N_m$$
$$= \{C(w_{m1}),C(w_{m2}),C(w_{m3}),\cdots,C(w_{mj})\},$$

其中:d_m'表示D中第m篇短文本的基于词向量合成的短文本表示,N_m为词数,w_j为当前短文本第j个词($j=1,2,3,\cdots,N_m$),$C(w_j)$是经过Word2Vec模型计算得到的词w_j的词向量,$C(w_{mj})$是经过相加平均后得到的第m篇短文本中词w_j的词向量。

(二)情感值计算

每条评论的情感值是由其情感词的数量和在情感本体中指定的强度来决定,定义公式如下:

$$d_m'' = \sum_{i=1}^{n} N_i \times Q_i,$$

其中:d_m''表示数据集D中第m篇短文本的情感值,N_i表示短文本中第i个词的情感极性,N_i的取值范围为{-1,0,1}(-1表示贬义,0表示中性,1表示褒义),Q_i表示第i个情感词的情感强度。

由于中文情感本体库中情感词数量有限,并没有包括所有的情感词,特别是微博表情库。微博表情是在微博评论文本中常见的一种抒发用户情感的途径,而对于某些表情,无法在情感本体库中找出对应的情感词进行计算。因此,对于无法进行情感值计算的表情,本文采用与这些表情相近的且是包含在情感本体库中的情感词进行代替,例如微博表情库中的"[亲亲]",在情感本体库中找不到与之直接对应的情感词,但可以用"亲热"这一相近的词进行代替,该词在情感本体库中属于"乐"这一大类,"快乐(PA)"这一小类,情感极性为褒义,情感强度为5。此外,针对网络词汇,很难能够权威地肯定该词的情感值,本文通过Word2Vec词向量模型,根据词向量之间的余弦距离,查找与之相关的近义词来判断该新词的情感极性和强度,如表1所示。

表1 网络词汇情感值示例

网络词汇	最近词	余弦距离	情感极性	情感强度
坑爹	郁闷	0.94390	贬义	5
萌萌哒	可爱	0.91689	褒义	5
腹黑	黑心	0.96465	贬义	7
心塞	难受	0.93084	贬义	5
吐槽	抱怨	0.90823	贬义	9

即便如此,仍然有一部分微博表情无法确定其情感倾向,例如"熊猫"、"咖啡"、"话筒"等表示静物或动物的微博表情,很难确定其情感极性以及情感强度。故本文将这一类微博表情的情感极性定为中性,即不计入评论文本的情感值计算。

(三)词向量和情感本体相融合的文本特征表示

将所得到的每条评论文本的情感值,记为d_m'',作为该文本的除词向量以外的一大特征,与基于词向量合成的模型d_m'进行顺序拼接,得到词向量与情感本体结合的短文本特征,公式定义如下:

$$d_m = \{d_m';d_m''\}$$
$$= \{C(w_{m1}),C(w_{m2}),C(w_{m3}),\cdots,C(w_{mj});d_m''\},$$

其中:";"表示向量顺序拼接操作,d_m为文本数据集D中第m篇短文本的词向量与情感本体结合的向量表示。

三、基于词向量和情感本体的短文本情感分类过程

根据本文所提出的短文本特征表示方法,现对如何应用该方法提出对应的短文本情感分类过程,其方法流程如图1所示。

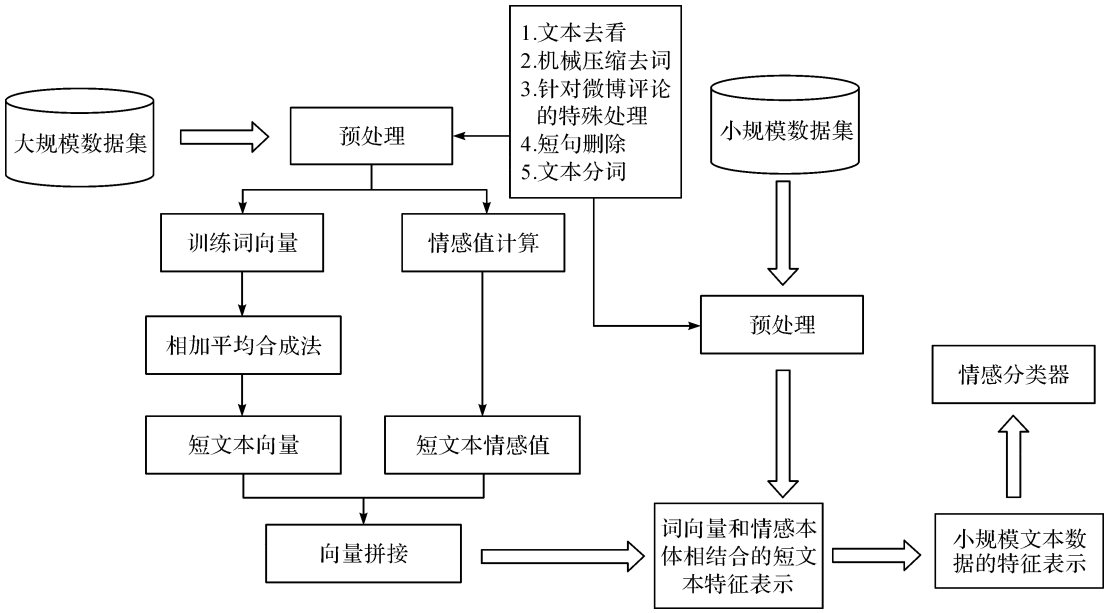

图 1　基于词向量和情感本体的短文本情感分类流程

(一)文本数据集构建

本文的小规模数据集用于训练文本分类器,需要进行标注。而大规模数据集用于训练生成词向量和情感值相结合的模型,所需数据并不需要标注。因此本文分类流程属于半监督学习。

由于本文是以微博评论文本为例进行研究,故大规模无标注数据集和小规模有标注数据集都应来自微博的评论文本,其所包含的领域与分类任务一致,且大规模无标记数据集应包含足够的领域(包括科学技术、社会、金融、互联网等)。

(二)评论文本预处理

1. 文本去重

文本去重即是去除文本评论数据中重复的部分。针对微博平台,有些用户在转发微博时系统可能会自动进行评论,评论内容为“转发微博”;其次同一用户由于想要多次表达自己的观点可能会进行重复评论;再者用户在评论时存在复制他人评论的可能性,导致出现不同人的评论内容相同。因此需要删除重复的文本数据,但为了存留更多的有用语料,文本去重只对完全重复的语料进行处理。

2. 机械压缩去词

机械压缩去词的目的就是去掉一些连续重复累赘的表达,将多个连续重复的词语压缩至一个。由于微博的评论文本数据质量参差不齐,没有意义的文本数据很多。例如,“非常好非常好非常好”以及“好呀好呀好呀”。这类语料的特点是在于将某些词语连续地重复地进行表达,这对基于本文方法的短文本特征结果产生较大的干扰。

3. 针对微博评论文本的其他处理

删除评论中存在的大量网页链接,这些对于评论文本的情感倾向挖掘没有意义。除此之外,用户在评论微博时会@其他用户,或是用户在回复其他用户的评论时,该用户所抓取的评论内容的形式为“回复@其他用户名:+‘该用户的评论内容’”。并且,用户在评论某微博或回复他人评论时,该用户的评论会多次出现,意味着用户多次进行情感表达,这会对情感分类产生影响。故在进行文本数据预处理时,不仅需要把每条评论的用户名删除。另对同一用户的多次不重复评论进行处理,只保留该用户评论中最长的一条,避免用户的评论重复次数对分类结果产生影响(在此视同一用户多次对同一微博评论的情感倾向是一致的)。

4. 短句删除

虽然精简的辞藻在很多时候是一种比较良好的习惯,但是由语言的特点知道,从根本上说,字数越少所能够表达的意思就越少,要想表达一些相关的意思就一定要有相应量的字数。因此,要删除掉过短的评论文本数据,以去除掉没有意义的评论。根据实验经验,4～8 个国际字符都是较为合理的下限。故此,经过前三步预处理后得到的短文本评论若小于等于 4 个国际字符,则将该语料删去。

5. 分词

进行预处理以后,主要对短文本数据进行分词。本文选用 Python 的中文分词包“jieba”进行中文分词。经过实验测试,“jieba”的分词精度高达 97% 以上。

(三)短文本特征表示

根据本文所提出的基于词向量和情感本体的短文本特征表示方法,计算预处理后的文本数据相对应的短文本特征,后将短文本特征数据输入情感分类器进行情感分类。

(四)分类器选择

情感分类器主要有支持向量机、K 最近邻和朴素贝叶斯。支持向量机对处理样本量少的数据效果较好,但大样本时优势并不明显,且这种非线性算法的计算复杂度较高,也不适合在大样本数据上做训练。贝叶斯分类是根据某对象的先验概率,利用贝叶斯公式得出后验概率,以最大后验概率的类作为该对象所属的类。而 KNN 分类器是根据距离度量个体间的差异性,将距离相近的归为一类。而本文所提出的对某条短文本数据的特征表示的本质就是 n 维向量,可依据向量之间的距离来进行短文本分类。故本文将选用 KNN 分类器进行分类。

四、实验结果及分析

(一)实验设置

通过 Python 对微博平台数据进行网络爬虫,爬取不同领域(包括科技、社会、财经、互联网等)下的热门微博评论,本文总计收集 50 万条微博评论样本作为训练生成词向量和情感值计算的大规模样本,另收集 3 万条微博评论作为情感分类的样本数据。使用 Python 进行短文本预处理,包括文本去重、机械压缩、短句删除、包括针对微博文本的特殊处理以及评论文本分割。

经过以上处理,最后总计得到大规模数据集 201245 条,小规模数据集 16832 条。

本文将利用 Python 进行 Word2Vec 的词向量训练,根据经验设置词向量维数为 50、100 和 150,用以测试不同维数下在 KNN 分类器下的分类效果。

(二)评价指标

1. Precision、Recall、$F1$ 分数评价指标

本文采用的评价指标主要是准确率 P(Precision),召回率 R(Recall)和 $F1$ 值。为了方便描述三种评价指标的计算公式,建立分类结果表 2。

表 2 分类结果

归属类别	属于此类	不属于此类
实际属于此类	TP	FN
实际不属于此类	FP	TN

表格中 TP、FP 表示分类系统与实际分类结果一致的文本数,FN、TN 则表示分类系统与实际分类结果不一致的文本数。

准确率 P 表示分类的正确率,计算公式为:

$$P/\%=\frac{TP}{TP+FP}\times 100.$$

召回率 R 表示分类的完整性,计算公式为:

$$R/\%=\frac{TP}{TP+FN}\times 100.$$

$F1$ 值综合考虑准确率和召回率,计算公式为:

$$F1/\%=\frac{2PR}{P+R}\times 100.$$

2. AUC(Area under curve)评价指标

AUC 是指 ROC 曲线下的面积,取值范围一般在[0.5,1.0],且其面积越大,分类效果越好。而 ROC 曲线是指对于某分类器在不同的阈值下,会得到一组(FPR,TPR),以 FPR 作为横坐标,TPR 作为纵坐标,作出曲线图。其中对 FPR 和 TPR 定义如下:

$$FPR/\%=\frac{FP}{FP+TN}\times 100,$$

$$TPR/\%=\frac{TP}{TP+FN}\times 100.$$

为消除本文小规模数据的样本数目对本文分类结果的影响,本文采用 AUC 评估不同样本量下的分类结果。

(三)结果分析

针对不同维度下(50 维、100 维、150 维),测试本文方法的 AUC 值随样本量大小变化的情况,并比较其他两种分别基于情感本体和词向量模型的分类方法。其中样本量从已预处理好的 16832 条小规模数据中随机抽取 2000,4000,…,14000,16832 条进行实验。训练集与测试集的比例依照 K-Fold Cross-Validation(K 取常用值,即 $K=10$)方法进行 10 次交叉验证,最后得到结果如图 2 所示。

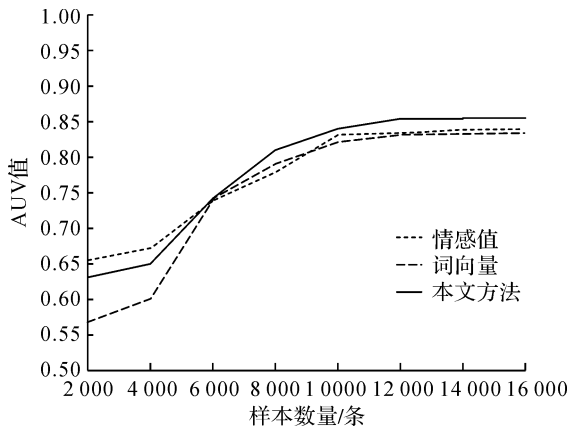

图 2 不同样本数量的分类效果(50 维)

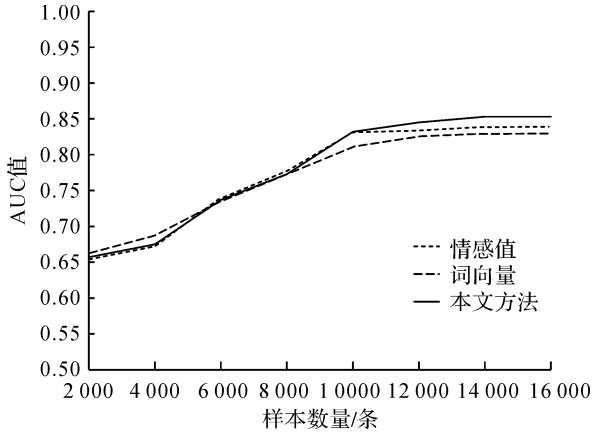

图 3　不同样本数量的分类效果(100 维)

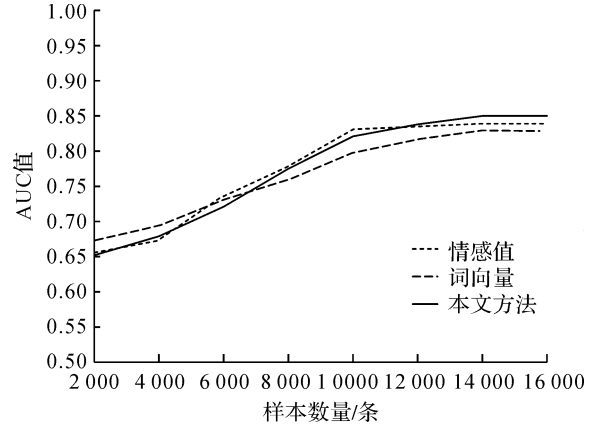

图 4　不同样本数量的分类效果(150 维)

由图 2—图 4 可知,当样本数目大于 12000 时,三种分类方法的分类效果达到较好的水平,并且此时维数对分类效果的影响较小。因此从计算效率的角度考虑将维数设置为 50,样本量设置为 12000,对三种分类方法进行测试,取 10 次 10 折交叉验证结果的均值,如表 3 所示。

表 3　分类结果比较(50 维)

分类方法	*F*1 值/%	*P*/%	*R*/%
本文方法	76.6	76.7	76.4
词向量	73.3	73.2	73.4
情感本体	75.1	74.9	75.3

从表 3 可以看出,相比其他两种分类效果,本文的分类效果较佳,*F*1 值至少提升 3.3%,准确率 *P* 至少提升 3.5%,召回率 *R* 至少提升 3.0%。并且从图 2—图 4 可看出,本文方法的 AUC 值在样本量大于 12000 时,均较优于其他两种方法。综上,本文所提出的基于词向量和情感本体的文本特征表示方法可有助于提升短文本情感分类效果。

五、结　语

本文提出一种新的短文本特征表示方法,即综合情感词特征和词向量的文本特征表示。实验部分探讨了不同词向量维数以及分别基于词向量和基于情感本体在 KNN 分类器上的分类效果,综合得出本文方法的优势。然而本文方法不足之处在于,一是仅简单地将情感值与词向量拼接作为文本的特征项,二是缺乏对网络词汇情感倾向判定是否恰当的评估。后续将对某一领域和网络词汇的情感强度和极性判定,以及词向量和情感值的其他结合方式等情况展开研究。

参考文献:

[1] 夏火松,刘建,朱慧毅. 中文情感分类挖掘预处理关键技术比较研究[J]. 情报杂志,2011,30(9):160-163.

[2] Yu N. Exploring co-training strategies for opinion detection[J]. Journal of the Association for Information Science&Technology,2014,65(10):2098-2110.

[3] 唐晓波,朱娟,杨丰华. 基于情感本体和 KNN 算法的在线评论情感分类研究[J]. 情报理论与实践,2016,39(6):110-114.

[4] Pang B, Lee L. Seeing stars: exploiting class relationships for sentiment categorization with respect to rating scales [C]//43rd Annual Meeting of the Association for Computational Linguistics. Michigan: The Association for Computational Linguistics,2005.

[5] Isidoros P, Ioannis H. Recognizing emotions in text using ensemble of classifiers[J]. Engineering Applications of Artificial Intelligence,2016,51(C):191-201.

[6] 杨锋,彭勤科,徐涛. 基于随机网络的在线评论情绪倾向性分类[J]. 自动化学报,2010,36(6):837-844.

[7] 杨小平,马奇凤,余力,等. 评论簇在网络舆论中的情感倾向代表性研究[J]. 现代图书情报技术,2016,32(z1):51-59.

[8] Philipp M. Support vector machines in automated emotion classification[D]. Cambridge: Churchill College,2003.

[9] Gamon M. Sentiment classification on customer feedback data: noisy data, large feature vectors, and the role of linguistic analysis[C]//20th International Conference on Computational Linguistics. Geneva: International Committee on Computational Linguistics,2004.

[10] Tong R M. An operational system for detecting and tracking opinions in on-line discussion[C]//the 24th Annual International ACM SIGIR Conference. New Orleans: ACM,2001.

[11] 桂斌,杨小平,张中夏,等. 基于微博表情符号的情感词典构建研究[J]. 北京理工大学学报,2014,34(5):537-541.

[12] 史伟,王洪伟,何绍义. 基于知网的模糊情感本体的构

建研究[J].情报学报,2012,31(6):595-602.
[13] 唐晓波,兰玉婷.基于特征本体的微博产品评论情感分析[J].图书情报工作,2016(16):121-127.
[14] 张群,王红军,王伦文.词向量与LDA相融合的短文本分类方法[J].现代图书情报技术,2016,32(12):27-35.
[15] Mikolov T, Chen K, Corrado G, et al. Efficient Estimation of Word Representations in Vector Space [C]// International Conference on Learning Representations. Scottsdale: the Computational and Biological Learning Society,2013.
[16] 徐琳宏,林鸿飞,潘宇,等.情感词汇本体的构造[J].情报学报,2008,27(2):180-185.

Short text sentiment classification: Based on the word vector and emotional ontology

WANG Zhengcheng, LI Dandan

(School of Economics and Management, Zhejiang Sci-Tech University, Hangzhou 310018, China)

Abstract: At present, two ways are mainly adopted for short text sentiment classification: statistical Natural Language Processing and emotional semantic characteristics, while the researches on the combination of the two methods is few. Thus, this paper will design a classification method that is based on the combination of word vector and emotional ontology is designed in this paper. Firstly, the word vector was trained by Word2Vec model and short text vector was synthesized by adding average method. On this basis, short text expression model which integrates word vector and emotional ontology was constructed by combining emotional value of each short text. Eventually, short text sentiment classification was completed by using KNN algorithm. Compared with the traditional classification methods which are based on the word vector, emotional ontology, or some other single technical route, the classification method combining word vector and emotional ontology gets an obvious improvement on precision, recall rate and $F1$ value.

Key words: short text sentiment classification; word vector; emotional ontology

(责任编辑:任中峰)