基于多特征融合的评论文本情感分析

龚 安,费 凡
(中国石油大学(华东) 计算机与通信工程学院,山东 青岛 266580)

摘 要:评论文本情感分析现已成为自然语言处理的重要研究领域。针对评论文本语法不规则、特征稀疏的问题,设计了一种针对评论文本的多特征融合的情感分类算法。首先提出一种改进的情感规则方法;然后从规则方法中提取出有效信息,将每一个情感信息量扩展为多维向量,再融合一元词特征、句法特征以及依存词语搭配特征构成向量空间,形成更有效的融合特征模板;最后利用信息增益理论进行特征选择,作为支持向量机的输入对评论文本进行识别和分类,实现了机器学习方法与规则方法相融合。以中文酒店评论数据集作为语料进行实验,结果表明该方法能让机器学习算法更加充分地利用规则特征,相比单纯地使用规则方法或机器学习方法,能够达到更好的分类性能,进一步提高分类精度。

关键词:文本情感分析;多特征融合;机器学习;情感规则
中图分类号:TP391.9 **文献标识码**:A **文章编号**:1673-629X(2018)08-0091-05
doi:10.3969/j.issn.1673-629X.2018.08.019

Comment Text Sentiment Analysis Based on Multi-feature Fusion

GONG An,FEI Fan
(School of Computer & Communication Engineering,China University of Petroleum,Qingdao 266580,China)

Abstract:The analysis on text emotional inclination has received much attention from natural language processing filed in recent years. In order to solve the problem of grammatical irregularity and feature sparsity,we design an emotional classification approach based on multi-feature fusion for text sentiment. At first,an improved method based on emotional rules is proposed. Then the effective information extracted from the ruled-based method is extended to a multidimensional vector and an effective integration feature set is obtained by adding various rule-based features to the basic feature set after expanding and converting them. Finally,the information gain theory is used to select features as the input of SVM. Thus,a method via a combination of rule-based and machine learning method is realized. We use the Chinese hotel reviews data set as the corpus for the experiment which shows that this method can make machine learning algorithm more full use of the rule features and it works better than simply using rule-based method or machine learning method.

Key words:text sentiment analysis;multi-feature fusion;machine learning;emotional rules

0 引 言

随着互联网的蓬勃发展,各类网络评论也相应激增。大量用户通过新闻网站、购物网站及微博等载体表达自己对时政、商品、电影及各类社会现象的观点及看法,这其中隐含着大量的高价值信息,而合理地分析和利用这些评论文本能够为个人消费决策、商家营销策略规划、政府舆情检测等方面提供帮助,因此有效地挖掘评论文本中蕴含的情感具有重要的社会价值与商业价值[1]。

目前,文本情感分析[2-3]的主流方法一般分为两种:一种是基于情感词典的规则方法[4-5];另一种是基于机器学习的方法[6-8]。基于情感词典的方法主要是根据情感词典的先验信息进行计算来判断文本所蕴含的情感,但情感词典的大小是有限的,且因为忽视语义往往不能得到准确的分类。机器学习方法是以模式分类的思想来处理这个问题,通过人工设计特征,将文本进行特征向量化输入到各种分类器中进行分类。从整体来看,机器学习方法的表现好于规则方法。然而对于复杂的汉语来说,传统的机器学习的建模方法不能取得令人满意的结果。对此,充分利用规则情感分析的结果,提出了一种机器学习与情感规则相融合的中文文本情感分析方法。

收稿日期:2017-08-26 **修回日期**:2018-01-10 **网络出版时间**:2018-03-08
基金项目:国家油气重大专项(2017ZX05013-001)
作者简介:龚 安(1971-),男,副教授,CCF 会员(62929M),研究方向为大数据智能处理;费 凡(1993-),男,硕士研究生,研究方向为自然语言处理、行为识别。
网络出版地址:http://kns.cnki.net/kcms/detail/61.1450.tp.20180307.1432.072.html

文中主要内容如下：

(1)对现有中文情感词典进行了扩充整合，整理建立了网络情感词典库，形成比较全面的情感词典。

(2)针对评论文本特点，提出一种改进的基于词典的情感规则分类方法，在处理指代问题和特殊语言结构时分类结果更加精确。

(3)充分利用情感规则方法分析的结果，将经过情感语义规则方法中提取出的有效信息与人工设计的多种特征进行融合映射到 N 维特征空间中，使模型可以学习到更多的情感知识。

(4)将建立的情感分类算法在酒店评论分析任务上进行实验验证。

1 机器学习与情感规则融合的文本情感分析方法

对于 NLP 问题，由于汉语情感分析资源有限，且因其绝无仅有的复杂度，从而使得识别中文评论的情感成为一项具有挑战性的任务。文中提出了一种机器学习与情感规则相结合的多特征融合的中文文本情感分析方法，目标是对现有评论文本进行情感分类，从而发现用户对产品、主题的评价信息。将情绪结果映射到极性，并将其分为三类：正向情感、负向情感和中性情感。

1.1 情感词典构建

情感词典是构建的带有情感极性色彩标记的一个集合，是文本情感分析任务中不可或缺的重要组成部分，通常情况下情感词典越完备，得到的识别结果越精准。为了得到更好的识别结果，对目前使用广泛的各大情感词典(如 HowNet、Ntusd、Tsing 等)进行整合与扩展，建立了包含基础情感词、表情情感词、程度副词、否定词及转折连词的综合情感词典。

除此之外，还建立了网络情感词词典。对于网络新词的出现，有很多文献研究了基于机器学习的扩充情感词典的方法，取得了一定的效果[9]。但是针对现在各种层出不穷的网络用语，如“惊不惊喜”、“2333”等词，由于分词及候选词抽选等问题不能用算法得到很好的处理效果。故以知乎爬取的网络用语词典为基础，对其他网络情感词进行了整理和扩充，构建了情感词数量为 726 的网络情感词典。

1.2 文本预处理

中文评论文本通常包含了极强的个人风格和个人感情色彩，表达内容丰富，除了具有不规范性、语法基本都是偏向生活化和口语化之外，还包含大量不规范用语、错别字、链接以及表情符号等，所以在进行文本情感分析任务之前，需要对其进行预处理。

为了提高文本情感分析的效率，首先进行滤除网址、标签、不规则用语以及去除停用词的处理。在文本预处理阶段，分词是非常重要的组成部分之一。由于评论文本口语化特点明显，且包含大量网络新词，使用一般的分词工具效果不是非常理想，所以采用中科院开发的可加入用户自定义词典的中文分词系统 ICTCLAS[10]对评论文本进行分词处理，以达到更好的分词效果。

1.3 基于词典的情感规则分类方法

基于情感词典的分类方法是以情感词为中心，根据情感词典的先验知识来判断文本的情感倾向，最经典的是对情感词进行累加得到文本的情感倾向值，公式如下：

$$S = \sum_{i=1}^{n} \mathrm{Sw}_i \tag{1}$$

其中，Sw_i 为第 i 个情感词的极性；n 为情感词的总数。

根据式 1 将所有情感词的极性进行叠加，根据最后得到的数值来判断文本情感倾向值。但是在文本中决定情感极性的不仅只是情感词，其他如否定词、程度副词以及语言结构等都会对情感倾向造成一定影响。

针对经典方法存在的缺陷，提出基于词典的情感规则分类方法。由于评论文本一般较短，首先将文本中每个子句作为一个单元，通过以情感词典为基础设立的情感规则方法得到的情感计算公式 2 对每个单元进行情感倾向计算，最后将所有的单元得分值进行叠加，得到整个评论文本的情感倾向性。

$$S_{\mathrm{unit}} = \sum_{i=1}^{n} (K * \mathrm{Pw}_i * \prod_{j=1}^{m} \mathrm{mod}_j) \tag{2}$$

其中，n 表示文本中情感词的总数；Pw_i 表示第 i 个情感词的极值；m 表示修饰第 i 个情感词的词数；mod_j 表示其对应的修饰词的权值；k 表示强化削弱系数，是为了避免主语混淆所导致的情感分析偏差。

在文本情感分析任务各种算法中，往往由于缺少指代判定，所得出的情感极性并不是对主语的判定，结果存在偏差。

情感规则如表 1 所示。

表 1 情感规则

规则名称	规则内容
否定规则	检测到否定修饰词，对情感极性和强度作反转
表情规则	检测到表情词，匹配表情词典，直接赋予相应得分
转折连词规则	前转折词(如“虽然”、“尽管”)予以削弱，后转折词(如“但是”、“但”)予以加强；对“如果”、“假如”、“要是”等连词作反转处理
指代规则	对指代进行加强和削弱，对于”这个”、“这些”一类代词予以加强，对“那个”、“那些”一类代词予以削弱

1.4　机器学习方法

基于机器学习的分类方法是将情感分析看作一个模式分类问题,建立分类模型来判断情感极性。首先,机器学习方法需要对文本进行标注工作,将其作为训练集,然后提取特征对分类器进行训练,最后对测试语料进行测试得到分类结果。

文本特征选择是机器学习的关键步骤,决定着情感分类的精度。文中选择三大类特征:一元词(unigram)特征、句法特征以及依存词语搭配特征。其中句法特征是研究组成部分和排列顺序的特征,考虑到短语结构可以减少句子歧义,将二元词(bigram)及其组合词性标注作为其特征添加到特征集中;依存关系特征是从依存解析树中得到的依存关系标识,它对情绪类别信息的标注有着重要的作用,可以保存情感词与情感词直接相关联的信息及其他隐藏信息。

以"华为手机确实不错,我很喜欢!"为例句进行特征提取。首先采用中科院 ICTCLAS 分词工具进行处理,得到的词性标注以及分词结果如下所示:

华为/nz 手机/n 确实/ad 不错/a/,/wd 我/rr 很/d 喜欢/vi!!/wd

其中,/nz 代表专有名词;/n 代表名词;/ad 代表副形词;/a 代表形容词;/wd 代表标点符号;/rr 代表代词;/d 代表副词;/vi 代表动词。

从上述结果中可以得到例句的一元词特征及句法特征。然后在分词的基础上,调用斯坦福大学的 StanfordNlp 工具包,获得文本的依存关系及其词语搭配特征。例句的依存关系及词语搭配表如图 1 所示。

依存关系	
assmod(手机-2,华为-1)	punct(不错-4,-5)
nsubj(不错-4,手机-2)	nsubj(喜欢-8,我-6)
advmod(不错-4,确实-3)	advmod(喜欢-8,很-7)
root(ROOT-0,不错-4)	conj(不错-4,喜欢-8)

图 1　例句依存关系及词语搭配

从图中可以发现文本的根节点及其蕴含的 4 种依存关系:关联修饰(assmod)、名词性主语修饰(nsubj)、副词修饰(advmod)、并列词连接(conj)。

由上述分析步骤可以得到机器学习方法的 3 种基本特征模板。为了避免由于原始特征空间维数较大导致的分类器效果下降的问题,采用信息增益(IG)[11]的特征选择方法对原始特征空间进行维数约简以选择相应的特征,其公式如下所示:

$$IG(T)=H(C)-H(C\mid T)=-\sum_{i=1}^{n}P(C_i)\log_2P(C_i)+P(t)\sum_{i=1}^{n}P(C_i\mid t)\log_2P(C_i\mid t)+P(\bar{t})\sum_{i=1}^{n}P(C_i\mid \bar{t})\log_2P(C_i\mid \bar{t}) \quad (3)$$

其中,$P(C_i)$ 表示类别 C_i 的概率;$P(t)$ 表示特征 t 出现的概率;$P(C_i\mid t)$ 表示特征 t 出现时,类别 C_i 出现的概率;$P(C_i\mid \bar{t})$ 表示特征 t 不出现时,类别 C_i 出现的概率。

1.5　机器学习与情感规则结合的多特征融合方法

机器学习方法和规则方法相融合的算法受到了很多研究者的关注,如 Qiu 等[12]将词典分类结果作为分类模型的训练语料,形成一个层级迭代的分类框架;Mohammad 等[13]将情感词累加和和收尾词的极性作为特征。受前人的启发,文中提出一种机器学习与情感规则相结合的多特征融合的分类算法,其流程如图 2 所示。

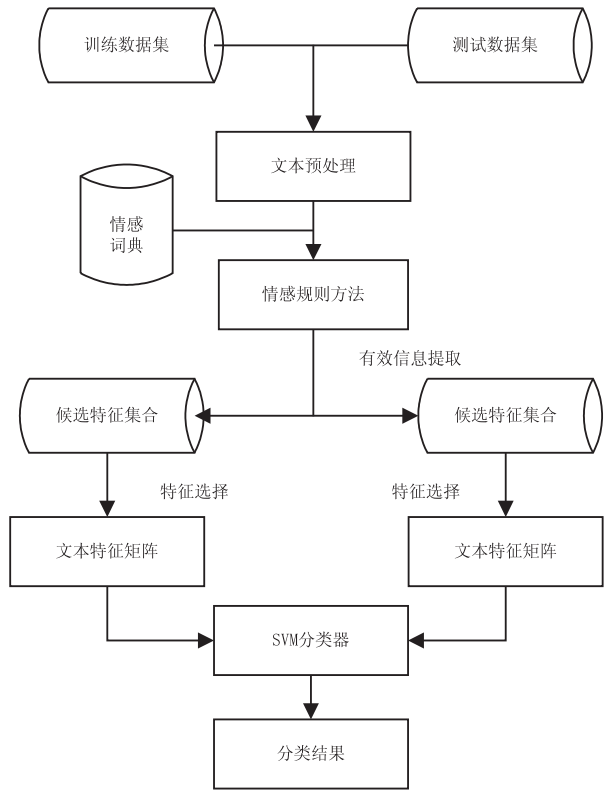

图 2　情感分类流程

作为机器学习和情感规则融合方法的必要步骤,在根据改进的情感规则方法计算出情感得分后,对其有效信息进行提取和扩展,用以与机器学习特征相融合。文中提取了情感词得分、正/负向情感词数量之比、加强次数与削弱次数之比、褒/贬情感句数量之比四种特征,对其归一化处理后扩展到机器学习特征模板中,训练 SVM 分类器,再用测试语料进行测试。通过上述流程,实现了机器学习方法与基于词典的情感规则方法相结合的多特征融合的文本分类方法,将从规则算法中提取出的多个有效情感信息扩展到向量空间,使得机器学习算法能更充分地利用规则特征,学习

到更多的情感知识。

2 实 验

2.1 实验准备

实验具体的配置如下:处理器为 Intel(F) Core(TM)i5-6500 CPU @3.2 GHz;内存 8 GB;编程平台为 Eclipse;开发语言为 Java;数据库为 SqlSever2008。

实验数据来自(学者谭松波)从携程网上收集整理的酒店评论语料,随机抽取正向类别和负向类别样本各 4 000 条。其中 70% 的语料作为训练数据,其余 30% 的语料作为测试数据。

为对实验效果进行评价,采用情感分类准确率(accuracy),即分类正确的样本数占所有样本数的比例,作为评价指标:

$$accuracy = num(correct)/num(all) \quad (4)$$

2.2 实验结果与分析

当前较为著名的分类器有支持向量机(SVM)、朴素贝叶斯(NB)、K 近邻分类器(KNN)等,文中选择在文本分类领域中性能较好的 SVM 算法来测试分类效果[14]。目前应用最为广泛的 SVM 分类器主要有 LibSVM 和 SVMLight 两种,采用由台湾大学林智仁教授开发的 LibSVM[15] 进行分类测试,将所获得的文本特征矩阵转化成 LibSVM 所对应的格式,最终获得情感分类类别。

表 2 比较了两种基于词典的情感规则方法的效果,结果表明,经过改进的情感规则方法的准确率得到了有效提升,但由于设定的情感规则仍较为粗糙,需要继续改进。

表 2 基于词典的情感规则方法分类性能

情感分类方法	准确率/%
基于词典情感分类的方法	70.12
改进的基于词典的情感规则的方法	74.20

为了更好地验证情感规则方法与机器学习方法融合的有效性,将机器学习的基本特征模板作为基准,加入情感规则方法提取的有效信息特征作对比。其中,Ft_1 是一元词特征,Ft_2 是依存关系特则,Ft_3 是句法特征。为了避免特征冗余现象造成的向量空间维数过大对分类器效果的影响,根据信息增益公式计算每个特征的信息增益分数,选择分数靠前的 1 000、2 000、4 000项特征构成文本向量。SVM 的核函数选取径向核函数。结果如表 3 和表 4 所示。

表 3 机器学习与情感规则融合的方法

信息增益值	情感规则特征融合	Ft_1	Ft_1+Ft_2	Ft_1+Ft_3	$Ft_1+Ft_2+Ft_3$
Top1000	否	81.02	81.98	81.78	81.54
	是	82.80	82.91	82.81	82.97
Top2000	否	81.67	82.33	82.08	82.01
	是	82.96	83.34	83.10	83.66
Top4000	否	81.50	82.16	82.10	81.95
	是	82.84	83.24	83.17	83.12

表 4 对比实验

情感分析方法	准确率/%
情感词典方法	70.12
传统 SVM 方法[16]	79.48
Qiu[12]	81.99
Mohammad[13]	82.84
文中	83.66

从表 3 可以得知,在不加入从情感规则方法提取转化的有效特征情况下,最好的分类精度在一元词特征与依存特征取信息增益值前 2 000 项时达到了最好的分类效果,识别率为 82.33%。并且一元词特征与依存特征相结合取得的识别率高于一元词特征与句法特征相结合取得的识别率,说明在这种短文本的语料中,依存关系特征带来的性能提升大于句法特征。在融合从情感规则方法提取的有效特征后,识别效果均有较大提升,并且在三种基本特征信息增益分值前 2 000项时与情感规则特征相融合得到了最好的识别效果,识别率为 83.66%。

从表 4 可以得知,文中提出的方法相比单一的情感词典方法、机器学习方法在识别准确率上有较大提升,且高于 Qiu[12]、Mohammad[13] 提出的机器学习与规则方法相融合的算法,更加适合中文评论文本情感分类,证明了该算法的有效性。

从以上结果可知,在文本情感分析任务中,提出的改进情感规则方法的准确率得到了有效提升,在提取其有效信息进行多特征融合后达到了最好的分类正确率。

3 结束语

对基于词典的情感规则方法进行改进,提出一种

基于多特征融合的文本情感分类算法,将从改进的规则方法中提取有效信息进行转化扩展,融合基本特征模板形成了更为有效的特征模板,实现了机器学习方法与情感规则方法的融合。通过酒店评论语料测试,实验结果表明,该方法在文本情感分类任务中取得了较好的效果。

参考文献:

[1] 董强柱. 微博中的意见领袖和新闻伦理[J]. 长安大学学报:社会科学版,2013,15(2):118-121.

[2] 赵妍妍,秦　兵,刘　挺. 文本情感分析[J]. 软件学报,2010,21(8):1834-1848.

[3] 朱圣代. 细颗粒度情感分析若干技术研究[D]. 杭州:杭州电子科技大学,2013.

[4] JANSEN B J,ZHANG Mimi,SOBEL K,et al. Twitter power:tweets as electronic word of mouth[J]. Journal of the American Society for Information Science & Technology,2009,60(11):2169-2188.

[5] PANDARACHALIL R,SENDHILKUMAR S,MAHALAKSHMI G S. Twitter sentiment analysis for large-scale data:an unsupervised approach[J]. Cognitive Computation,2015,7(2):254-262.

[6] LIU Bing. Sentiment analysis and opinion mining[M]. [s. l.]:Morgan & Claypool Publishers,2012.

[7] BARBOSA L,FENG Junlan. Robust sentiment detection on Twitter from biased and noisy data[C]//Proceedings of the 23rd international conference on computational linguistics:posters. Beijing,China:[s. n.],2010:36-44.

[8] TAN Songbo,ZHANG Jin. An empirical study of sentiment analysis for chinese documents[J]. Expert Systems with Applications,2008,34(4):2622-2629.

[9] PENG Wei,PARK D H. Generate adjective sentiment dictionary for social media sentiment analysis using constrained nonnegative matrix factorization[C]//Proceedings of the fifth international conference on weblogs and social media. Barcelona,Catalonia,Spain:[s. n.],2011:273-280.

[10] 刘　群,张华平,俞鸿魁,等. 基于层叠隐马模型的汉语词法分析[J]. 计算机研究与发展,2004,41(8):1421-1429.

[11] YANG Yiming,PEDERSEN J O. A comparative study on feature selection in text categorization[C]//Fourteenth international conference on machine learning. [s. l.]:Morgan Kaufmann Publishers Inc.,1997:412-420.

[12] QIU Likun,ZHANG Weishi,HU Changjian,et al. SELC:a self-supervised model for sentiment classification[C]//Proceedings of the 18th ACM conference on information and knowledge management. Hong Kong:ACM,2009:929-936.

[13] MOHAMMAD S M,KIRITCHENKO S,ZHU Xiaodan. NRCCanada:building the state-of-the-art in sentiment analysis of tweets[C]//Proceedings of the 7th international workshop on semantic evaluation. Atlanta:[s. n.],2013:321-327.

[14] JOACHIMS T. Text categorization with support vector machines:learning with many relevant features[C]//European conference on machine learning. Berlin:Springer,1998:137-142.

[15] CHANG C C,LIN C J. LIBSVM:a library for support vector machines[J]. ACM Transactions on Intelligent Systems and Technology,2011,2(3):27.

[16] 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(1):1-4.

(上接第 74 页)

international conference on computer and information science. Okayama,Japan:IEEE,2016:1-5.

[7] 李　倩. 深度网络模型构建及学习算法研究[D]. 西安:西安电子科技大学,2014.

[8] 余　凯,贾　磊,陈雨强. 深度学习:深度学习:推进人工智能的梦想[J]. 程序员,2013(6):22-27.

[9] 周　超. 基于深度学习混合模型的文本分类研究[D]. 兰州:兰州大学,2016.

[10] 陈翠平. 基于深度信念网络的文本分类算法[J]. 计算机系统应用,2015,24(2):121-126.

[11] HINTON G E,SALAKHUTDINOV R R. Reducing the dimensionality of data with neural networks[J]. Science,2006,313(5786):504-505.

[12] HINTON G E,OSINDERO S,TEH Y W. A fast learning algorithm for deep belief nets[J]. Neural Computation,2006,18(7):1527-1554.

[13] 张　祥. 一个网页分类系统的研究与实现[D]. 北京:北京邮电大学,2013.

[14] 胡　燕,吴虎子,钟　珞. 基于改进的 kNN 算法的中文网页自动分类方法研究[J]. 武汉大学学报:工学版,2007,40(4):141-144.

[15] 梁宏胜,徐建民,成岳鹏. 一种改进的朴素贝叶斯文本分类方法[J]. 河北大学学报:自然科学版,2007,27(3):327-331.