面向情感倾向性识别的特征分析研究

李妍坊,许歆艺,刘功申
(上海交通大学 信息安全工程学院,上海 200240)

摘 要:随着互联网和信息技术的迅速发展,网络上用户的评论信息越来越多。利用计算机技术分析网络中大规模文本的情感倾向,在政府的舆情分析和企业的产品评价智能回馈等应用中有着非常巨大的发展前景。文中着重研究了选取不同的文本特征对文本情感倾向性分类精度的影响。实验中所研究的不同文本特征主要包括情感词、形容词、副词、语气词和标点符号等。实验结果表明,选取情感词、形容词、副词作为特征项对情感分类具有较好的效果,在此基础上添加语气词和标点特征可以有效地提高情感分类的精度。该研究成果可用于社会舆情分析、垃圾博客过滤、商品评论与推荐、影视评价等领域。

关键词:情感分析;文本分类;特征选取;支持向量机
中图分类号:TP31 **文献标识码**:A **文章编号**:1673-629X(2014)09-0033-04
doi:10.3969/j.issn.1673-629X.2014.09.007

Research on Feature Analysis Oriented Text Sentiment Identification

LI Yan-fang, XU Xin-yi, LIU Gong-shen
(School of Information Security Engineering, Shanghai Jiaotong University, Shanghai 200240, China)

Abstract: With the rapid development of the Internet and information technology, the online comments of users are also increasing. Using computer technology to analyze emotional tendencies of large-scale network texts in the government's public opinion analysis and evaluation of the company's product applications such as intelligent feedback has enormous development prospects. Mainly study the influence of selecting different text features on the final classification accuracy of sentiment classification in this paper. Different text features studied in the experiment include emotional words, adjectives, adverbs, modal and punctuation. The experimental results show that selecting emotional words, adjectives, adverbs as feature items on sentiment classification can achieve good classification performance, and adding modal and punctuation features can effectively improve the sentiment classification accuracy. The research findings can be applied to social public opinion analysis, filtering spam blog, commodity reviews and recommendations, film evaluation and so on.

Key words: sentiment analysis; text classification; feature selection; SVM

0 引 言

当今时代,随着互联网和信息技术的迅速发展,越来越多的人把网络作为社交平台,网络上的评论文章等用户生成内容的数量不断增长,对海量文本进行批量的情感倾向性识别已经日益成为一种极其迫切的需要。通过对带有情感的主观性信息进行分析处理,可以挖掘出人们的态度和见解,从而为政府、企业或用户提供重要的决策支持,也可将分析结果应用于商品的评估与推荐、影视评价、社会舆情监控、网络内容管理、垃圾信息过滤等领域。

文本情感倾向性识别可大致分成以下几个层次:词语、句子、篇章情感倾向性识别[1]。在国外,Kim 等人将工作重点放在情感词汇的倾向性分析上,在基准词集的基础上使用 WordNet 计算未知词汇的情感倾向性[2];Pang 等人对电影评论的数据按照倾向性分成两类,利用人工标注了文本倾向性的训练语料,基于 unigram 和 bigram 等特征,学习分类器[3]。近年来,跨领域的情感倾向性识别也成为一大研究热点[4]。目前文本情感倾向性识别的主要研究方法是基于统计的文本分类法,即以情感倾向为标准对文本进行分类。康乃尔大学的 Salton 教授提出的向量空间模型[5](Vector Space Model)是目前最流行的文本表示模

收稿日期:2013-11-01 **修回日期**:2014-02-16 **网络出版时间**:2014-07-17
基金项目:国家自然科学基金资助项目(61272441,61171173);国家科技支撑计划项目(2011BAK05B03)
作者简介:李妍坊(1991-),女,硕士生,研究方向为自然语言处理、信息安全;刘功申,副教授,研究方向为自然语言处理、社交网络、信息安全。
网络出版地址:http://www.cnki.net/kcms/detail/61.1450.TP.20140717.1233.042.html

型[6],使得各种文本分类算法有效地应用于情感倾向性识别。其中,支持向量机(SVM)算法是文本分类领域非常高效的算法,具有相对较好的分类效果[7]。要进行情感分类,首先需要对文本进行特征提取。通过对中文文本进行分词、词性标注、停用词过滤[8]等预处理,以突出能够代表文本特征的内容。一般情况下在特征提取前,文本中的语气词“啊”“呢”“吗”等和标点符号往往会被当成冗余信息被停用词表过滤掉,然而在情感的表达方面,语气词和标点符号却有着画龙点睛的作用,针对此,文中选择的文本特征也包括了语气词和标点符号。此外,文中从不同词性特征出发,对比研究了选取不同的文本特征项及其组合对文本情感分类精度的影响。

1 文本特征提取

特征提取将一个无结构的原始文本转化为结构化的计算机可以识别处理的信息,即对文本进行科学的抽象。特征提取一般分为特征表示、特征项选择和特征构造。

1.1 文本的特征表示

特征表示即选取哪种形式的信息作为文本的特征项。特征项是表示文本的基本单位。特征项必须具备一定的特性:能够确实代表文本内容,可以区分目标文本与其他文本,个数不能太多,分离要方便实现。在中文文本中可以采用字、词或短语等作为表示文本的特征项。此外,N-gram[9]、词性、句型等也可作为文本的特征项。

1.2 文本的特征项选择

特征项选择要考虑的是特征项数量多少的问题。特征选择是文本分类中最重要的环节之一,准确、有代表性的特征项才能更有效地区分文本的类别。特征选择也可称为特征降维,特征选择的好坏直接关系到后期分类的准确度。

目前通常采用 VSM 来描述文本,但是如果直接用分词算法和词频统计方法得到的特征项来表示文本向量中的各个维,那么这个向量的维度将非常得大。这种未经处理的文本矢量不仅给后续工作带来巨大的计算开销,使整个处理过程的效率非常低下,而且会损害分类、聚类算法的精确性,从而使所得到的结果很难令人满意。为了解决这个问题,最有效的办法就是通过特征选择来降维。传统的特征选择方法包括文档频率法、信息增益法、互信息法等[10]。姜鹤等人描述了一种基于法矢量权重的特征选取方法[11],并通过实验证明该特征选取方法相对于传统的特征选取方法可以产生更优的分类性能。张希娟等人提出了一种基于最小冗余的特征选取方法[12],也可以改善特征选取的效果。

1.3 文本的特征权重计算

在 VSM 中,每个特征项成为空间中的一个维度,各个维度的权重表示该特征项在该文档中的重要程度。

(1)基于词频的特征权重:词频(Term Frequency)是指某一词语出现的频率。即把某一特征项在文档 d 中出现的次数作为该特征项的权重。

(2)基于词频和反文档频率的特征权重:即 TF * IDF(Term Frequency - Inverse Document Frequency),Salton 将 IDF 定义为式(1):

$$\mathrm{IDF}_t = \log\left(\frac{N}{n_t}\right) \tag{1}$$

其中,N 表示文本集中的文档总数;n_t 表示文本集中出现过特征项 t 的文本数;IDF 表示该特征项在文本集中的分布情况,也表示该特征项的区分能力,过高的 IDF 表示该特征项区分度很低,不能体现该文本的特点。

结合 TF 与 IDF,该特征项的最终权重为式(2):

$$w_{t,d} = \mathrm{TF}_{t,d} * \mathrm{IDF}_t \tag{2}$$

(3)基于文本语义信息的特征权重:即在基于词频的基础上,根据特征项前后的语义信息来调整该特征项的权重[13]。

2 词性特征分析与选取

针对文本情感倾向性识别,需要考虑的是哪些特征对文本情感的表达具有重要的影响。在文本情感倾向性分类出现之前,传统的文本分类大多是基于主题的[14],因此特征抽取只需筛选出具有主题信息的特征,继而通过分类器便能得到分类结果。“主题”,顾名思义,通常都是名词,所以传统分类的关键就是找到能代表特征的名词,难度不大。然而,文本中所传达的情感并不是仅仅存在于某些名词之中,情感的表达方式也是多种多样的。因此,比起传统的基于主题的文本分类,情感分类要更复杂一些。

目前大多数的情感分析研究常利用一些现有的情感资源,其中比较有权威性的是知网发布的《知网情感分析用词语集》。知网情感分析用词语集中总结出了英文和中文的正面情感词语和负面情感词语、正面评价词语和负面评价词语、主张词语等等。情感词语,顾名思义,即能够表达正面或负面情绪的词语,例如“快活”“欢天喜地”“悲痛欲绝”等。情感词语是大多数传统的文本情感分类研究中最重要最常用的特征项。

然而除了情感词语,从词性的角度来看,形容词、副词、动词、语气词等都可能对情感倾向的表达产生影

响。例如“酒店的装饰非常漂亮”,这是一句明显的带有正面情感的句子,其中形容词“漂亮”对本句的情感表达起着至关重要的作用。再如“服务员的态度极其差”和“服务员的态度稍微有点儿差”相比,前者中的副词“极其”更能强烈地表达出对酒店服务员负面的情感。“酒店的地理位置很好,值得推荐”中动词“推荐”很明显地表达了作者对酒店的积极评价,带有正面的情感。

在大多数情况下对原始文本进行预处理时,语气词“啊”“呢”“吗”等和标点符号往往会被当成冗余信息通过停用词表过滤掉,然而在情感的表达方面,语气词和标点符号却有着画龙点睛的作用,使情感表达更鲜明、细腻、生动。比如“大床房中的电视能收到中央三台的综艺栏目,哈哈”,这句酒店评论里的语气词“哈哈”更加生动地强调了对酒店的正面情感;再如“出差这么久了,住在这里很有家的感觉!”其中句末的感叹号传达出了作者较为强烈的正面情感。因此,语气词和标点符号看似是文本中的冗余信息,但是在情感表达中所起到的作用不可忽视。针对这两点,文中的实验也将语气词和标点符号作为文本特征项。

3　实验及结果分析

3.1　语料选取及词表构建

实验用到的原始语料库为携程网酒店评价的平衡语料库,首先从已标注过的原始语料中随机选取了 800 篇酒店评价的文本(其中正面情感和负面情感的分别 400 篇)作为原始待分析文本。从 800 篇文本中随机选取 600 篇作为训练集(正面和负面的各 300 篇),其余 200 篇作为测试集,训练集与测试集的比例为 3:1。

文中的研究将分别提取情感词、形容词、副词、动词、语气词、标点符号以及它们的组合作为特征项,因此需要构建好各种词表。情感词表用了《知网情感分析用词语集》提供的正面情感词语和负面情感词语,形容词表、副词表、动词表是以现代汉语形容词表、现代汉语副词表、现代汉语动词表为基础,通过去重复、排序等处理构成。语气词表和标点符号是以汉语语气词的研究文章为基础,通过总结得来,也经过去重复等处理最终形成。为了后续实验方便,将要提取的不同文本特征进行编号,如表 1 所示。

3.2　实验步骤

实验中采用了中科院 ICTCLAS 系统对文本分词,在已分词的文本中,识别并提取出情感词、形容词、副词、动词、语气词等,在后续实验中分别或组合作为特征项。

考虑到实验所用的语料库为酒店评价语料,其中包含了许多能反映情感倾向性但是区分度并不大的词,如:“好”“大”“干净”等等,所以为了避免此类词汇的权重受到影响,在特征权重计算时采用词频(TF)法而未采用 TF * IDF 法。

表 1　特征编号与特征对应表

特征编号	提取的特征
0	情感词
1	形容词
2	副词
3	动词
4	语气词
5	标点符号

分类结果评估的标准包括查准率(Precision)、查全率(Recall)和 F 值(F-Measure)三部分,这三个评价指标都是针对某一类别的分类指标。其计算公式分别如式(3)~式(5)。

$$\text{Precision} = \frac{\text{正确判为类别 } C \text{ 的文本数}}{\text{所有被判为类别 } C \text{ 的文本数}} \times 100\% \tag{3}$$

$$\text{Recall} = \frac{\text{正确判为类别 } C \text{ 的文本数}}{\text{所有属于类别 } C \text{ 的文本数}} \times 100\% \tag{4}$$

$$F_\beta(P,R) = \frac{(\beta^2+1)PR}{\beta^2 P + R} \tag{5}$$

其中,β 是一个调整参数,引入 β 的目的是以不同权重综合查准率和查全率。通常取 $\beta = 1$。此外,为了准确地评价某一分类器整体上的有效性,需要将各类别的有效性指标综合起来,可以用分类精度(Accuracy)来衡量。精度计算公式如式(6):

$$\text{Accuracy} = \frac{\text{分类正确的文本数}}{\text{测试文本总数}} \tag{6}$$

上述的这几个评价指标的值越高,说明文本分类的有效性越好。

3.3　实验结果

选取不同文本特征的最终分类精度如表 2 所示。

3.4　实验数据分析

对实验数据进行对比分析,可以得出以下结论。

(1)选取单一文本特征时,情感词、形容词、副词表现出了相对较好的分类效果,相比之下,选取动词、语气词和标点作为文本特征时分类效果则不尽人意。其原因可能在情感表达方面,形容词和副词具有更好的表现力,单纯的动词和语气词还不足以代表文本的情感意图。

(2)选取多项文本特征进行组合实验时,由于情感词、形容词、副词作为特征时情感分类效果较好,因此在这三项特征的基础上分别添加了语气词、标点符号这两项特征。由表 2 数据得知,特征编号为 0 时情感分类精度为 87%,特征编号为 0+4、0+5、0+4+5 时情感分类的精度有非常明显的提升,全部都已达到

90% 以上。实验结果表明,添加语气词和标点特征可以从不同程度上提高情感分类精度。

表 2 选取不同文本特征的分类精度

特征编号	提取的特征	正确分类文本数/测试文本总数	分类精度(Accuracy)/%
0	情感词	174/200	87
1	形容词	166/200	83
2	副词	146/200	73
3	动词	100/200	50
4	语气词	100/200	50
5	标点	104/200	57
0+4	情感词+语气词	197/200	98.5
0+5	情感词+标点	185/200	92.5
0+4+5	情感词+语气词+标点	199/200	99.5
1+4	形容词+语气词	172/200	86
1+5	形容词+标点	172/200	86
1+4+5	形容词+语气词+标点	169/200	84.5
2+4	副词+语气词	148/200	74
2+5	副词+标点	143/200	71.5
2+4+5	副词+语气词+标点	148/200	74
4+5	语气词+标点	128/200	64

传统的文本情感分类大多只提取出了情感词作为基本的特征项,且在使用停用词表时把类似于“呢”“啊”的语气词和类似于“!”“~”的标点符号过滤掉,忽略了语气词和标点对情感表达的重要作用。文中的实验研究表明了虽然语气词和标点单独作为特征项时效果并不理想,但与其他文本特征组合使用时,则可以提高情感分类的精度。

4 结束语

由于人类情感的复杂性,再加上存在着文本的语义歧义等难题,文本情感倾向性识别具有较高的难度。文中主要研究了选取不同的文本特征对文本情感分类精度的影响。研究的文本特征包括了传统实验中忽略掉的语气词和标点特征。实验结果显示,选取情感词、形容词、副词作为特征项对情感分类具有较好的效果,在此基础上添加语气词和标点特征可以有效地提高情感分类的精度。文中提取的文本特征没有考虑句型、修辞等语义信息,这些以后还需要进一步深入研究。

参考文献:

[1] 黄萱菁,赵 军. 中文文本情感倾向性分析[J]. 中国计算机学会通讯,2008,4(2):39-46.

[2] Kim S M, Hovy E. Automatic detection of opinion bearing words and sentences[C]//Proceedings of IJCNLP-05. [s. l.]:[s. n.],2005:61-66.

[3] Pang B, Lee L, Vaithyanathan S. Thumbs up? sentiment classification using machine learning techniques[C]//Proceedings of the 2002 conference on empirical methods in natural language processing. New Jersey: ACL, 2002:79-86.

[4] 吴 琼,谭松波,张 刚,等. 跨领域倾向性分析相关技术研究[J]. 中文信息学报,2010,24(1):77-83.

[5] Salton G, Wong A, Yang C S. A vector space model for automatic indexing[J]. Communications of the ACM, 1975, 18(11):613-620.

[6] Lewis D D. An evaluation of phrasal and clustered representations on a text categorization task[C]//Proceedings of the fifteenth annual international ACM SIGIR conference on research and development in information retrieval. [s. l.]:[s. n.],1992:37-50.

[7] Sharma A, Dey S. A comparative study of feature selection and machine learning techniques for sentiment analysis[C]//Proceedings of the 2012 ACM research in applied computation symposium. San Antonio, Texas: ACM, 2012:1-7.

[8] 王素格,魏英杰. 停用词表对中文文本情感分类的影响[J]. 情报学报,2008,27(2):175-179.

[9] 于津凯,王映雪,陈怀楚. 一种基于 N-gram 改进的文本特征提取算法[J]. 图书情报工作,2004,48(8):48-50.

[10] 刘丽珍,宋瀚涛. 文本分类中的特征选取[J]. 计算机工程,2004,30(4):14-15.

[11] 姜 鹤,陈丽亚. SVM 文本分类中一种新的特征提取方法[J]. 计算机技术与发展,2010,20(3):17-19.

[12] 张希娟,王会珍,朱靖波. 面向文本分类的基于最小冗余原则的特征选取[J]. 中文信息学报,2007,21(5):56-60.

[13] 李媛媛,马永强. 基于潜在语义索引的文本特征词权重计算方法[J]. 计算机应用,2008,28(6):1460-1462.

[14] 刘金岭. 基于主题的中文短信文本分类研究[J]. 计算机工程,2010,36(4):30-32.

(上接第 32 页)

2012:587-591.

[9] Golder S A, Macy M W. Diurnal and seasonal mood vary with work, sleep, and day length across diverse cultures[J]. Science, 2011, 333(6051):1878-1881.

[10] Paltoglou G, Twitter T M. MySpace, Digg: unsupervised sentiment analysis in social media[J]. ACM Transactions on Intelligent Systems and Technology, 2012, 3(4):66-66.

[11] 谢丽星,周 明,孙茂松. 基于层次结构的多策略中文微博情感分析和特征抽取[J]. 中文信息学报,2012,26(1):73-83.

[12] 杨 亮,林 原,林鸿飞. 基于情感分布的微博热点事件发现[J]. 中文信息学报,2012,26(1):84-90.

[13] 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.

[14] NLPIR[EB/OL]. 2013. http://ictclas.nlpir.org/.

面向情感倾向性识别的特征分析研究

作者： 李妍坊， 许歆艺， 刘功申， LI Yan-fang， XU Xin-yi， LIU Gong-shen
作者单位： 上海交通大学 信息安全工程学院,上海,200240
刊名： 计算机技术与发展 ISTIC
英文刊名： Computer Technology and Development
年，卷(期)： 2014(9)