文章编号:1001-9081(2010)04-1011-04

基于机器学习的网络新闻评论情感分类研究

周 杰,林 琛,李弼程
(信息工程大学 信息工程学院,郑州 450002)
(zhoujie-0001@163.com)

摘 要:首先对网络新闻评论数据的特点进行归纳总结,选取不同的特征集、特征维度、权重计算方法和词性等因素进行分类测试,并对实验结果进行分析比较。对比结果表明:情感词和论据词语搭配效果优于仅使用情感词作为评论特征;另外该类数据中特征维度对分类准确率的影响减小,且TF-IDF权重计算方法仍优于布尔型权重;在词性选择上,名词和动词词性比形容词和副词取得更好的分类效果。

关键词:网络新闻评论;中文信息处理;情感分类;机器学习;口语化评论
中图分类号:TP181 **文献标志码**:A

Research of sentiment classification for netnews comments by machine learning

ZHOU Jie, LIN Chen, LI Bi-cheng
(*Institute of Information Engineering, Information Engineering University, Zhengzhou Henan 450002, China*)

Abstract: Netnews comments has become an important channel to express personal opinions for the common people, and sentiment analysis can find out the whole attitude of the common people for the news events. This paper summarized the characteristics of netnews comments firstly, and selected different sets of feature, different feature dimensions, different feature-weight methods and parts of speech to construct classifiers; then made the comparison and analysis to the experimental results. The results of comparison show that the features combining sentiment words and argument words perform well to those only employing sentiment words; otherwise, feature dimension has less influence on the accuracy of classification for this kind of data, and the feature-weight method of TF-IDF is still better than boolean method. As for part of speech selection, nouns and verbs as features obtain better performance than adjectives and adverbs.

Key words: netnews comments; Chinese information processing; sentiment analysis; machine learning; oral comments

0 引言

网络新闻以及时、全面的特点越来越受广大网民关注,人们从网络新闻中实时了解社会新闻,并针对新闻事件发表观点性评论。目前,针对产品评论的观点挖掘技术逐步发展,该技术对给定对象(如产品)的相关信息进行搜索,生成属性列表(如质量、外观等),并概括对各属性的观点(好、中等、坏)。在部分文献中,也称观点挖掘为情感分析,文献[1]对两种称呼的起源和发展做出考查,认为两个领域已相互融合,可以相互替代。考虑到本文讨论的新闻评论数据不具有产品的属性列表,这里一致表述为情感分析。

网络互动能力的迅速增强,促使网络舆情的影响力不容忽略。合理地利用评论数据,提取并整合评论者观点,就能及时掌握广大民众对新闻事件(如新政策、社会热点事件等)的总体观点,为管理者提供决策支持。

网络新闻评论不同于专业评论,它更接近于口语化语言,形式上与论坛跟帖相似,因此也称为网络新闻跟帖。其特点主要包括:

1)长度很短,特征非常稀疏:一般只有几十个字符,仅包含几个词典词语;

2)表达简洁,错误噪声多,用语不规范,较多引入网络用语;

3)指代不明确,发散性思维强。

本文主要比较和分析了不同特征集、特征维度、权重计算方法和词性等因素选择对基于机器学习的情感分类的影响,为研究这种口语化的短文本数据提供参考。

1 相关研究

目前,情感分析的研究多针对产品评论信息,这些评论信息的评论对象较为固定,网络中发表有大量相关评论,可以容易获取充足数据,并且部分评论已有发表者评定等级,节省大量语料标注时间。而汉语方面的研究相对较少,特别是针对口语评论的研究。

1.1 无监督词典建立和情感分析

一篇文档表现为一个由文字和标点符号组成的字符串,字或字符组成词,词组成短语,然后再形成句子、段落和篇章。因此,对文档进行情感分析往往从判定词语的情感极性开始。

文献[2]的作者早在1997年就利用语言学上的连词对情感词倾向的约束性,由其中一个已知情感倾向的形容词推测另一个形容词的倾向。文献[3]中利用词语与一些具有明显语义倾向的种子词之间的关联特性,用统计的方法识别词语的倾向性,并分别对SO-PMI和SO-LSA两种方法,以及"and"和"near"两种关联程度进行分析,但由于使用了网络搜索引擎或大型语料库,实际应用较难实现。另一种思路是利用WordNet中的同义、反义和上下位等关系计算词语情感极性[4-5]。文献[5]中发现词语情感倾向间的一致性随着搜索距离的增加而逐渐减弱,为了克服此缺陷,提出一种基于路径的分析方法。情感词典的建立往往只是前期工作,最终的目

收稿日期:2009-09-08;**修回日期**:2009-11-27。 **基金项目**:国家863计划项目(2007AA01Z439)。

作者简介:周杰(1984-),男,湖北武汉人,硕士研究生,主要研究方向:文本情感分析; 林琛(1981-),女,山东威海人,博士研究生,主要研究方向:网络数据挖掘; 李弼程(1970-),男,湖南衡阳人,教授,博士生导师,主要研究方向:智能信息处理。

标是实现对评论者情感的判断。实验发现,仅利用情感词极性加权值判断情感倾向效果较差,文献[6]中先使用语法分析器对句子进行语法分析,再利用情感词典和情感模式库对句子的语义关系进行分析,实现对产品评论的观点识别。

1.2 机器学习方法进行情感分析

传统的机器学习方法很好地应用于基于主题的文本分类,主要依据是文本的主题往往可以由关键的主题词独立确定,这点不同于基于情感的文本分类。

文献[7]最早将机器学习方法用于基于情感的文本分类,分别采用 Naïve Bayes、最大熵和支持向量机(Support Vector Machines,,SVM)分类器以及多种特征选择方法等进行测试,后期研究[8]对更多方法进行比较分析,其中:

1)权重计算方法。不同于基于主题的分类,情感分类中布尔型权重优于词频权重。

2)位置信息。如处于文档前端、中间或靠近尾部,常作为分类特征[9]。

3)词性标注信息。情感分析系统中常引入词性信息,它是一种词语含义消歧的方法[10]。另一方面,形容词常被认为对情感具有很强的表征能力。

此外,还有许多情感特征的选择和处理方面的工作,如句法关系约束[11]、否定词处理[12]等。

实际环境中包含大量客观性文本,它们并不包含具体的情感倾向信息,并降低情感分析的性能。如句子:"'高尚'一词意思是指道德水平高,含赞扬的感情色彩",词语"高尚"、"赞扬"都含有明显的情感色彩,常会被误判为积极情感类别。相关研究逐渐重视主观性识别,文献[13]中指出,区分主客观实例要难于其后的情感倾向分类。由此可以体会出情感分类之前进行主观性识别的必要性。

1.3 汉语情感分析现状

汉语情感分析研究起步较晚,所做的研究工作相对较少。在汉语词汇语义倾向研究方面,文献[14]中提出了基于HowNet两种词语语义倾向性计算方法:基于语义相似度的方法和基于语义相关场的方法。在主观性语言方面,文献[15]中对汉语网络非正规语言(Network Informal Language,NIL)进行了研究,利用BBS文本建立NIL语料库。文献[16]中采用手工分类和自动获取相结合的方法填充词汇本体的框架,完成大规模情感词汇本体的构造。

在中文情感分类上,文献[17]中比较了常用的机器学习和特征选择算法的性能,使用五个主题的网络评论数据,得出结论:采用布尔权重、信息增益特征选择方法和SVM分类器,在大训练集和特定特征维度情况下,情感分类效果最优。文献[18]中针对口语化的中日论坛发帖进行情感分类,并简单分析了这类文档的特点。

2 分类方法

文本中包含着人类所使用的自然语言,计算机难以直接进行处理,需要先将文本转化成计算机能够处理的表示方式。文本特征是指反映文本的元数据,分为描述性特征和语义性特征。向量空间模型(Vector Space Model,VSM)是近年来应用最多且效果较好的方法之一。在该模型中,文本空间被看作由一组正交特征向量所组成的向量空间,每个文本表示成特征向量:

$$V(d) = (t_1, w_1(d); \cdots; t_i, w_i(d); \cdots; t_n, w_n(d)) \quad (1)$$

其中:t_i 为特征项,$w_i(d)$ 为 t_i 在文档 d 中的权重,n 为特征总数。

SVM是一种新型的通用知识发现方法,在分类方面具有良好的性能,在本文测试中,SVM的性能也远远好于KNN、RBF网络等分类方法。假设存在训练样本 $\{(x_i, y_i) \mid i = 1, \cdots, l, x \in \mathbf{R}^n, y \in \{-1, +1\}\}$,且可被一个超平面 $(w \cdot x) + b = 0$ 线性分割。如果训练数据可以无误差地被划分,且与该超平面最近的向量之间距离最大,则称这个超平面为最优超平面,其中距离超平面最近的向量称为支持向量。

在线性可分情况下,求解最优超平面可以看成解二次型规划问题,即在 $y_i(w \cdot x_i + b) - 1 \geqslant 0 (i = 1, 2, \cdots, l)$ 约束下,找到权值 w 和偏移 b 的最优值,使得权值代价函数最小化:

$$\min \Phi(w) = \frac{1}{2} \| w \|^2 \quad (2)$$

线性不可分情况下,先通过非线性变换将输入空间变换到一个高维空间,使样本线性可分;再在线性可分的情况下求取最优超平面。超平面是对两类的划分,对于大于两类的多类文本分类,就对每个类构造一个超平面,将这一个类与其余的类分开。

3 情感分类的因素比较与分析

3.1 语料数据

实验语料是从互联网中采集的新闻评论数据,由于网络的开放特性致使评论的发表过程更加随意。使用这类语料之前,需要经过数据去重、数据规范化处理和情感标注过程。

1)数据去重。相同评论者在较短时间间隔内发表的相同评论认为是重复评论,这部分评论所含信息量非常小,需要在采集以后去除。

2)数据规范化处理。网络新闻评论的书写格式多样,需要进行规范化处理,提高分类的准确性。常见步骤有:简繁转换、全角半角转换、字母大小写转换,以及不规范分隔号识别去除和特殊符号去除等。

3)情感语料标注。分别从文档、句子和词语三种粒度对评论语料进行标注。先确定整篇文档的情感倾向,再分析各个句子,提取句子的评价对象和情感词,针对每个情感词给出情感倾向及其强度。

实验使用"超市禁止免费提供塑料袋"(塑料袋政策)、"央视停播NBA"(停播NBA)和"刘翔退赛"三个事件的新闻评论数据,经相关处理和情感标注后,得到数据样本,如表1所示。

表1 实验数据样本信息

事件名称	评论总数	样本数量		样本平均长度	
		负面样本	正面样本	负面样本	正面样本
塑料袋政策	1675	854	821	30	54
停播NBA	1570	830	740	32	37
刘翔退赛	1003	603	400	50	59

从评论平均长度可以看出,这种网络新闻评论一般长度很短。其中,大部分评论观点明确,表达直截了当(如例1)。另一方面,这种评论的思维具有发散性,常常给出其他对象表达自身观点并对观点进行论证,如例2、3中使用"森林"、"垃圾袋"说明自己持反对观点及其原因。

例1 "坚决支持!"

例2 "那岂不是要提供纸袋了,又要有多少森林被砍了。"

例3 "现在不提供了,要去买专门的垃圾袋。这不一样吗,并没有减少对塑料袋的使用,没看出怎么环保了?"

此外,口语化的表述和一些不规范的格式如下面例子所示,部分会在规范化处理中矫正,但仍给分类准确性造成影响。

例 4　“不播就不播有什么大不了的。电视台有权选择放什么节目。全国都不播一定有理由的,电视台是靠广告赚钱的。不播他们损失更大,所以坚决支持。”

例 5　“支 - 持 - 封 - 杀 - NBA　其实本人又是 NBA 粉丝,但是在 - 这 - 样 - 的 - 情 - 况 - 下 我 - 可 - 以 - 不 - 看”

3.2　实验结果和分析

3.2.1　人工确定特征表示的分类实验。

文献[7]中对电影评论数据进行测试,人工挑选最能表征文本情感的特征词,然后利用这些特征对文本分类,精度仅为 60% 左右,差于根据统计方法选取的特征对文本分类所得的结果。此外,人工挑选的特征词主要是具有情感倾向信息的形容词或动词,词语之间具有较强关联性。

本实验从标注的情感语料中提取所有的情感词和评价对象信息,然后从情感词集合中人工挑选最具有情感表征能力的词语,形成测试特征集一;从评价对象中挑选能够支持评价者观点的论据词语(也可认为成潜情感词),与已挑选的情感词集合并,形成测试特征集二。分别使用 50% 的语料作为训练数据,剩余 50% 作为测试数据,选择 TF-IDF 权重计算方法和 SVM 分类器,得到实验结果如表 2 所示。

表 2　人工确定特征表示的分类实验结果比较

事件名称	测试特征集一			测试特征集二		
	特征数	准确率/%	无特征样本/%	特征数	准确率/%	无特征样本/%
塑料袋政策	40	40.21	62.9	90	80.79	5.7
停播 NBA	35	54.39	57.6	90	75.03	8.8
刘翔退赛	39	77.64	41.9	92	84.23	9.8

由前两个事件的结果可以看出,情感词独立作为特征时分类准确率低于 60%,并且特征覆盖范围具有很大局限性,其中无特征样本占全部语料的半数,导致分类准确率受到很大影响。引起这种大量无特征样本的原因主要包括特征维度过低和人工选取的情感词特征之间的强关联性两个方面。当加入与情感词关联性较弱的论据特征时,无特征样本数量急剧减少,同时分类准确率上升,说明在网络新闻评论中,除了情感词以外,部分论据词语对情感倾向性同样具有较好表征能力,如“塑料袋事件”中的“森林”、“垃圾袋”等,“停播 NBA”事件中的“奥运”等。

在“刘翔退赛”事件中,加入论据词语后准确率也有较大提升。该事件中部分情感词对情感倾向性分类具有巨大表征能力,致使在特征维度非常低时仍有较好分类准确率,如“加油”一词,删除该词后准确率由 77.64% 下降到 53.89%。

当去除数据中无特征样本后,对三个事件重新分类,得到结果均超过 75%,说明人工挑选的特征也能较好区分情感类别。然而,人工挑选时往往选择频繁使用的情感词或论据特征,特征覆盖范围有限。

3.2.2　不同特征维度的分类实验。

文本主题分类和专业评论情感分类过程中,往往会遇到高维特征问题。而网络新闻评论由于长度限制,致使特征非常稀疏,特征的最高维度也远远低于其他数据。本实验选择不同的特征维度,使用 50% 数据进行训练,剩余 50% 用于测试,选择 TF-IDF 权重计算方法和 SVM 分类器,得到实验结果如表 3 所示。

表 3　不同特征维度的分类准确率

特征维度	准确率/%		
	塑料袋政策	停播 NBA	刘翔退赛
100	79.00	75.29	83.23
300	83.77	80.00	85.43
500	84.49	81.91	87.43
700	86.40	81.66	86.03
1 000	85.20	80.51	86.03
1 400	84.96	80.51	85.63
1 800	85.44	80.76	85.83
2 000	85.92	—	84.63

由表 2、3 可以看出,采用机器学习方法得到的分类结果普遍高于人工选择的特征。与基于主题的分类过程相同,随着特征维度的增加分类准确率也逐渐增大,当超过一定值时又呈现下降趋势。但不同的是,这种趋势不明显,并且当维度降到足够低时,仍然保持较好的准确率。说明网络新闻评论中使用词语相对集中,并且这部分词语对情感类别具有良好表征能力,这种现象在“刘翔退赛”事件中表现更加明显。

当特征维度减少时,评论中的无特征样本数量增加,它们主要是单个词或短句构成的评论,即使选取最高维度,同样存在少量无特征样本。

3.2.3　不同权重计算方法的分类实验。

文本主题分类中,TF-IDF 权重计算方法性能明显优于布尔型权重方法。文献[7,19]中得出结论:不同于主题分类,在文本情感分类中,布尔权重能够得到更高的准确度。本实验使用 TF-IDF、TF 和布尔型权重对网络新闻评论数据进行测试,训练和测试语料仍各占一半,得到实验结果如表 4 所示。

表 4　不同权重计算方法的分类准确率

权重	准确率/%		
	塑料袋政策	停播 NBA	刘翔退赛
TF	84.57	78.77	83.43
TF-IDF	85.36	81.74	86.49
布尔型	84.73	79.96	84.90

对于新闻等数据的情感分类,人们一般认为具有正面或负面情感倾向性的词语往往不会重复出现,并且出现一次就决定文档的情感倾向,实验结果也显示布尔型权重准确率更高[19]。本实验结果表明,网络新闻评论情感分类中 TF-IDF 权重计算方法获得较优准确率,但三种权重计算方法之间的差距表现不明显,主要原因有两个方面:

1)由于评论语料的长度很短,同一条评论中具有情感倾向性的词语出现的可能性大大降低,导致三种方法的权重值较接近;

2)一些出现次数较多的词语(主要是评价对象和论据词语)也能对情感倾向性具有较好的表征能力,采用 TF-IDF 权重计算能提高分类准确率。

3.2.4　不同词性特征的分类实验。

情感分析系统中常引入词性信息,人们在分析哪类词语对文本情感分类更具表征能力的过程中,逐渐提升形容词(A)的地位。在汉语中,其他词性的词语同样具有较强的情感倾向,如名词(N)、动词(V)、副词(D)等。以下分别选择不同词性进行实验,训练和测试语料各占一半,选择 TF-IDF 权

重计算方法,采用中国科学院分词系统ICTCLAS30,得到实验结果如表5所示。

表5 不同词性特征的分类准确率 %

词性	塑料袋政策	停播NBA	刘翔退赛	词语所占比例
A, N, V, D	88.19	81.53	86.83	67.77
A	71.24	60.13	72.26	4.78
N	84.96	74.39	81.04	20.63
V	83.65	76.56	82.63	30.44
D	78.76	63.06	72.06	11.92

除了以上四种词性词语在评论中占有较大比例以外,还有代词和助词比例都接近10%。从实验结果可以看出,名词和动词作为特征的性能明显好于形容词和副词。一方面,因为形容词和副词数量太少,导致特征覆盖范围受限,无特征文档数量远远多于名词和动词特征;另一方面,网络新闻评论中与论据相关的词语词性一般为名词和动词,促使这两种词性有良好的性能。

4 结语

文中针对网络新闻评论进行情感分析,此类语料最大的特点是口语化和短文本。实验中可以发现,单纯地使用情感词并不能达到最佳的分类性能,评论中较多地运用支撑评价者观点的论据词语,加入此类词语作为特征能较好地提升性能。另外,由于新闻评论的短文本特性,致使特征稀疏,特征的最高维度仅2000左右,远远低于正规评论语料,评论中部分词语对情感类别具有良好表征能力,促使在较低特征维度时也能取得较高准确率,同时导致无特征样本数量急剧增加。不同于其他实验结果[7,19],该评论数据受到论据词语的影响,TF-IDF权重计算方法仍优于布尔型,词性选择上也偏向于特征覆盖范围更广的名词和动词。

研究口语评论为了解广大民众真实想法提供了途径,本文对该类语料进行初步的探讨,目的是为后面研究口语化评论数据提供参考。从实验数据可以看出,要真正做到实用仍有许多工作需要研究:

1)无关评论的过滤问题,网络新闻评论中存在各种无关评论,去除该类评论能得到更加可靠的数据来源。

2)网络中新词汇的出现和评论的不规范致使分词的性能远远低于书面语数据,特别是网络中人名、机构名以及对应别称的分析需要改善。

3)语料数据中存在单个词或短句评论,当选择较优特征时往往无法达到较广的数据覆盖范围。另外,网络新闻评论中存在部分具有很强情感表征能力的特征,需要协调好利用这些特征和解决数据稀疏两者之间的问题。

及时准确地了解广大民众对新闻事件的整体观点,以及它随着各种措施制定发生变化的情况,将为相关部门提供必要的参考。本文对该类语料的特点进行初步讨论,后续将针对已提出的工作进行研究,建立起一个实用的口语化评论的情感分析系统。

参考文献:

[1] PANG B, LEE L. Opinion mining and sentiment analysis [M]. Boston: Now Publishers Inc, 2008: 8 - 10.

[2] HATZIVASSILOGLOU V, MCKEOWN K R. Predicting the semantic orientation of adjectives [C]// Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics. Madrid: ACL, 1997: 174 - 181.

[3] TURNEY P D. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews [C]// Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. Philadelphia: ACL, 2002: 417 - 424.

[4] KAMPS J, MARX M, MOKKEN R J, *et al.* Using WordNet to measure semantic orientation of adjectives [C]// Proceedings of the 4th International Conference on Language Resources and Evalvation. Lisbon: LREC, 2004: 1115 - 1118.

[5] GODBOLE N, SRINIVASAIAH M, SKIENA S. Large-scale sentiment analysis for news and blogs [C]// Proceedings of the International Conference on Weblogs and Social Media. Colorado: [s. n.], 2007: 219 - 222.

[6] YI J, NASUKAWA T, BUNESCU R C, *et al.* Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques [C]// Proceedings of the 3rd IEEE International Conference on Data Mining. Florida: IEEE, 2003: 427 - 434.

[7] PANG B, LEE L, VAITHYANATHAN S. Thumbs up? Sentiment classification using machine learning techniques [C]// Proceedings of the Conference on Empirical Methods in Natural Language Processing. Philadelphia: [s. n.], 2002: 79 - 86.

[8] MULLEN T, COLLIER N. Sentiment analysis using support vector machines with diverse information sources [C]// Proceedings of the Conference on Empirical Methods in Natural Language Processing. Barcelona: NLP, 2004: 412 - 418.

[9] KIM S, HOVY E H. Automatic identification of pro and con reasons in online reviews [C]// Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics. Sydney: ACL, 2006: 483 - 490.

[10] WILKS Y, STEVENSON M. The grammar of sense: Using part-of-speech tags as a first step in semantic disambiguation [J]. Journal of Natural Language Engineering. 1998, 4(2): 135 - 144.

[11] KUDO T, MATSUMOTO Y. A boosting algorithm for classification of semi-structured text [C]// Proceedings of the Conference on Empirical Methods in Natural Language Processing. Barcelona: NLP, 2004: 301 - 308.

[12] DAS S, CHEN M. Yahoo! for Amazon: Extracting market sentiment from stock message boards [C]// Proceedings of the 8th Asia Pacific Finance Association Annual Conference. Bangkok: [s. n.], 2001: 22 - 25.

[13] MIHALCEA R, BANEA C, WIEBE J. Learning multilingual subjective language via cross-lingual projections [C]// Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics. Prague Czech Republic: ACL, 2007: 976 - 983.

[14] 朱嫣岚, 闵锦, 周雅倩, 等. 基于HowNet的词汇语义倾向计算[J]. 中文信息学报, 2006, 20(1): 14 - 20.

[15] XIA Y, LI W. A Phonetic-based approach to Chinese chat text normalization [C]// Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics. Sydney: ACL, 2006: 993 - 1000.

[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.

[17] 唐慧丰, 谭松波, 程学旗. 基于监督学习的中文情感分类技术比较研究[J]. 中文信息学报, 2007, 21(6): 88 - 108.

[18] 王素格, 李伟. 面向中日关系论坛的情感分类问题研究[J]. 计算机工程与应用, 2007, 43(32): 174 - 177.

[19] 徐军, 丁宇新, 王晓龙. 使用机器学习方法进行新闻的情感自动分类[J]. 中文信息学报, 2007, 21(6): 95 - 100.