情报学报 ISSN 1000-0135
第27卷第2期175-179，2008年4月

JOURNAL OF THE CHINA SOCIETY FOR SCIENTIFIC AND TECHNICAL INFORMATION ISSN 1000-0135
Vol.27 No.2, 175-179 Apirl 2008

停用词表对中文文本情感分类的影响[1)]

王素格[1,2] 魏英杰[1]

（1.山西大学数学科学学院，太原 030006；2.上海大学计算机工程与科学学院，上海 200436）

摘要 本文利用三种特征选择方法、两种权重计算方法、五种停用词表以及支持向量机分类器对汽车语料的文本情感类别进行了研究。实验结果表明，不同特征选择方法、权重计算以及停用词表，对文本情感分类的影响也不尽相同；除形容词、动词和副词外的其余词语作为停用词表以及不使用停用词表对情感分类作用较大，得到的分类结果比较好；总体上，采用信息增益和布尔型权重进行中文文本情感分类的效果较好。

关键词 停用词 文本情感分类 特征选择 支持向量机

The Influence of Stoplist on the Chinese Text Sentiment Categorization

Wang Suge[1,2] and Wei Yingjie[1]

(1. *School of Mathematics Science, Shanxi University, Taiyuan* 030006;
2. *School of Computer Engineering and Science, Shanghai University, Shanghai* 200072)

Abstract In this paper, using three kinds of feature selection methods, two kinds weighing assignment methods, the five kinds of Stoplist and SVM on text sentiment classification are studied. The experiment results indicate that the greater text sentiment classification impact depends on other corpus, excluded adjective, verb, adverb as stop words and none stop words. As a whole, for text sentiment classification, information gain is superior to other feature selection methods and Boolean type weighting is superior to frequency type weighing.

Keywords stop word, text sentiment classification, feature selection, support vector machine

1 引 言

随着网络的快速发展，信息量的剧增，人们对信息处理提出了更高的要求。网上大量的文本信息不仅包含了主题信息，而且还包含了个人主观信息。对于后者，这些信息的主要特点是带有个人主观情感、情绪以及态度。情感是一个非常广泛的概念，它涉及人们的观点、看法和评价，包括人类行为相对于社会标准的评价，产品相对于审美观的评价。相对于情感概念的特征至少包括情感、情绪或态度、情感倾向（判断语义倾向为正面还是反面）以及其强度。本文针对汽车评论语料的情感进行分类。可以将其看成正面、反面的二分问题：假设预定义的文本类型集为 $S=\{P,N\}$，其中 P 表示对相关汽车持积极态度的评论，也可称正面的（positive），N 表示对相关汽车持消极态度的评论，也可称反面的（negative）。待分类的文本集为 $D=(d_1,d_2,\cdots,d_n)$，本文的任务就是将文本集 D 中的文档 $d_i(i=1,2,\cdots,n)$ 自动判断为正面或者反面。

传统的中文文本分类通常是基于主题分类，停

收稿日期：2007年2月9日

作者简介：**王素格**，副教授，博士研究生，研究方向：文本挖掘，自然语言处理与机器学习。E-mail：wsg@sxu.edu.cn。**魏英杰**，硕士研究生，研究方向：文本挖掘与自然语言处理。

1） 国家自然基金项目（60573074）；山西省自然科学基金（20041040）；山西省科技攻关项目（051129）；山西高校科技研究开发项目（200611002）。

用词表对其分类结果有较大的影响[1]，G:W.Hart发现[2]，在典型英文段落中所用词的50%可以包含在一个具有135个词的普通词表中，应在文本分析预处理中去除；Yang和Pedersen认为[3]，若对停用词按照其出现的文本频数降序排列，仅用前10个停用词降低特征向量空间维数，不会产生负面影响；用前100个停用词降低特征向量空间维数，所产生的负面影响非常小，但再大一些，效果会有明显的影响。与此同时，Silva验证了应用停用词表降低特征空间的维数，对提高文本分类器的准确率会产生积极的作用[4]；顾益军等提出联合熵自动获取停用词表[1]。

对于情感分类来说，将什么样的词作为停用词以及停用词对其的影响还没有看到相关的报道。本文选用了5种停用词表，采用常用的3种特征提取方法[5]：信息增益(IG)、互信息(MI)和 χ^2 统计(X2)，两种权重的计算方法：基于文档和基于词频，利用支持向量机分类方法[5]，分别考察了5种停用词表对汽车评论的情感类别判断的影响，为后续情感分类问题的进一步研究提供参考和依据。

2 特征选择方法与权重计算方法

2.1 特征选择方法

本文采用信息增益(IG)、互信息(MI)、χ^2 统计(X2)[5]等常见的特征选择方法。首先对候选特征计算其度量值，然后根据预先设定的阈值 T，将度量值大于 T 的特征选为有效特征。

设 f 为特征，c 为文档，C 为文档类。

(1) 信息增益(IG)

信息增益表示文档中包含某一特征时文档类的平均信息增量，它被定义为某一特征在文档中出现前后的信息熵之差。对于特征 f，其信息增益 $IG(f)$ 被定义为：

$$IG(f) = H(C) - H(C \mid f) = \sum_{c \in C}\left(P(c,f)\log\left(\frac{P(c,f)}{P(c)P(f)}\right) + P(c,\bar{f})\log\left(\frac{P(c,\bar{f})}{P(c)P(\bar{f})}\right)\right) \quad (1)$$

(2) 互信息(MI)

在统计学中，互信息用于表征两个随机变量间的相关性。对于特征 f，其互信息 $MI(f)$ 被定义为：

$$MI(c,f) = \log\left(\frac{P(c,f)}{P(c)P(f)}\right) \quad (2)$$

应用时取平均值：

$$MI_{avg}(f) = \sum_{c \in C} P(c)MI(c,f) \quad (3)$$

(3) χ^2 统计(X2)

χ^2 统计也是用于表征两个随机变量间的相关性的统计量，但它比互信息更强，因为它同时考虑了特征出现与不出现两种情况。对于特征 f，其 $\chi^2(f)$ 统计值被定义为：

$$\chi^2(c,f) = \frac{(P(c,f)P(\bar{c},\bar{f}) - P(c,\bar{f})P(\bar{c},f))}{P(c)P(f)P(c)P(f)} \quad (4)$$

类似于互信息，取平均值：

$$\chi^2_{avg}(f) = \sum_{c \in C} P(c)\chi^2(c,f) \quad (5)$$

在实际应用中，上述公式中的概率均用频率代替，详见文献[5]。

2.2 权重计算方法

根据2.1介绍的3种特征选择方法和预先设定的阈值，可以选择出对文本分类起作用的特征。要将待处理的文本表示为向量的形式，还需要计算特征在此文本中的权重(特征的取值)，向量的维数即为特征的个数。本文中，我们采用布尔型和词频型两种权重形式：

(1)布尔型特征权重：如果特征 f_i 出现在文档 c_i 中，则其权重 $w_{ij}=1$，否则其权重 $w_{ij}=0$。

(2)词频型特征权重：以特征 f_i 出现在文档 c_i 中的频次 t_{ij} 作为该特征的权重，即 $w_{ij}=t_{ij}$。

3 语料库选取与评价

3.1 语料库的选取

本文实验语料全部来自汽车点评网，评论时间集中于2006年6～8月。该网收集了对国内外11种品牌的轿车的评论，总计400篇约41万字。为了反映网站评论的真实情况，其中正面、反面的语料约占网站评论总量的10%，这样导致正面、反面语料数量不一致，比例约为5:3。评论人群主要是车主和即将购车的人，他们大多是从非专业的角度进行评论，另外还有少量媒体评论。有别于其他类别的评论，汽车类的评论一般篇幅较长而且综合影响因素较多。总体来看评论基本上可以分为两大类：一是显性的，即无论是从题目、内容用语，还是结论都明确地表明作者的情感倾向和态度，相反另一类是隐性的，这对问题的研究有一定的影响。从评论的内容上来看可以分为以下几种：

媒体评论文章。这类评论一般比较专业,用词恰当、准确,处处都突出鲜明的观点。如"高档与高性价比一肩挑","作为国内中高档车市场的主力干将,别克君威一直口碑素著"。

专业的评论。特点是汽车各项参数居多,对判断情感有一定的作用,但很难从参数上直接得出情感倾向性信息。如"固特异 205/55R16 的宽扁平比"。

日志式的评论。一般以时间为序,篇幅很长,且内容很散乱,褒贬之语反复交替出现。如"每月统计分析:8 月……,9 月……"。

大众化的评论。这类评论基本上是申明观点、比较、下结论的格式,但大多数情况是观点结论与中间评论内容不符,例如虽然对该车有很多不满,但结论仍为正面的,这对研究情感的判定是个较大的影响。如:"我是根据我开我的小 Q500 多公里初步感觉实事求是的说的这些,总体上我是喜欢 QQ 的!我也希望奇瑞能够都听听车主们的意见,不断改进,我支持奇瑞!"。

基于以上的考虑,语料中正面、反面类别的判定是由 3 个判断者人工进行判别,以减少人为的误差。

3.2 评价指标

评价指标是在测试过程中所使用的一些用来评价分类性能的量化指标,通常采用的分类评价指标有查全率(Recall,简记为 R)、查准率(Precision,简记为 P),其定义如下:

$$\text{反面查全率 } RN = \frac{a_1}{c_1} \quad \text{反面查准率 } PN = \frac{a_1}{b_1}$$

$$\text{正面查全率 } RP = \frac{a_2}{c_2} \quad \text{正面查准率 } PP = \frac{a_2}{b_2}$$

$$\text{总体查全率 } F1 = \frac{a_1 + a_2}{c_1 + c_2}$$

$$\text{总体查准率 } F2 = \frac{a_1 + a_2}{b_1 + b_2}$$

其中,a_1 表示系统判断为反面的文档与实际应为反面的文档相等的数量,a_2 表示系统判断为正面的文档与实际应为正面的文档相等的数量;c_1 表示实际应为反面的文档数量,c_2 表示实际应为正面的文档数量;b_1 表示系统判断为反面的文档数量,b_2 表示系统判断为正面的文档的数量;由于本文判断的是正反两类问题,因此 $c_1 + c_2 = b_1 + b_2$,即 $F1 = F2$,这样混合两种类别的查全率和查准率应为相等。因此在实验中,我们仅仅考虑 $F1$。

4 实验结果与分析

4.1 中文文本情感分类的步骤

对于中文文本情感分类,其步骤如下:

(1) 将文本集通过分词预处理;

(2) 去掉一些与文本情感分类无关的词即停用词,得到候选特征;

(3) 对每一个候选特征,通过特征抽取方法,计算其度量值,然后根据设定的阈值 T,将度量值大于 T 的候选特征选为文本的情感特征;

(4) 对文本情感特征利用文本分类器(支持向量机)得到情感分类结果。

4.2 停用词表的选择

传统的停用词表中是对主题没有描述能力和区别能力的词,是一些噪声词。而对情感分类来说,特征的选取就是要选择既要带有情感色彩又要有区分能力的词。在英文中,人们首先选择名词(n)、动词(v)、形容词(a)、副词(d)确定为候选特征[6-8],而中文[9]具有情感色彩的确定为名词(n)、动词(v)、形容词(a)、副词(d)、区别词(f)、叹词(e)、拟声词(o)、代词(r)、成语、简称等作为具有情感色彩的词,为了测试选用不同的候选特征对于情感分类的影响,我们构造不同的停用词表作为选择候选特征的依据。

(1)停用词表 1:不含动词(v)、形容词(a)、副词(d)的停用词表。即将动词(v)、形容词(a)、副词(d)作为候选特征。

(2)停用词表 2:不含名词(n)、动词(v)、形容词(a)、副词(d)的停用词表。即将名词(n)、动词(v)、形容词(a)、副词(d)作为候选特征。

(3)停用词表 3:不含名词(n)、动词(v)、形容词(a)、副词(d)、区别词(f)、叹词(e)、拟声词(o)、连词(c)的停用词表。由于汉语的句式对情感分类有很大的影响,比如转折句中情感描述主要在于后半分句,因此候选特征将连词加入。将名词(n)、动词(v)、形容词(a)、副词(d)、区别词(f)、叹词(e)、拟声词(o)、连词(c)作为候选特征。

(4)停用词表 4:主题停用词表:采用李荣陆[5]的停用词表。

(5)停用词表 5:无停用词表,即将所有的词作为候选特征。

4.3 中文文本情感分类实验

为了减少训练语料与测试语料对测试结果的影

响，本实验采用五次交叉检验，考察五种停用词表对汽车评论情感类别（正面、负面）判断的影响。实验采用特征维数为4 000维，其结果见表1，为了反映其结果的趋势，将各分类结果的 F 值绘成图1。

表1 五种停用词表、三种特征抽取方法以及两种权重计算的情感分类正反面、综合的 F 值

停用词表	评价 F	基于文档			基于词频		
		IG	*MI*	*X2*	*IG*	*MI*	*X2*
1	正面	83.90%	83.33%	82.24%	81.12%	84.50%	81.43%
	反面	68.32%	65.97%	63.41%	62.18%	68.64%	63.42%
	综合	78.48%	77.11%	77.70%	74.44%	78.94%	76.12%
2	正面	82.35%	82.22%	82.86%	83.25%	81.52%	82.73%
	反面	64.63%	66.03%	64.56%	65.91%	65.57%	63.95%
	综合	76.73%	74.95%	76.95%	76.40%	76.45%	75.18%
3	正面	82.62%	81.19%	83.46%	82.45%	82.27%	80.90%
	反面	65.85%	63.28%	66.62%	65.85%	65.64%	59.64%
	综合	76.19%	76.19%	76.69%	77.11%	75.40%	76.43%
4	正面	81.63%	80.39%	83.32%	80.60%	80.51%	81.77%
	反面	64.20%	60.51%	66.88%	60.36%	61.98%	62.32%
	综合	75.40%	73.68%	77.70%	73.57%	74.18%	73.94%
5	正面	85.36%	78.31%	83.80%	83.23%	77.97%	84.11%
	反面	71.29%	52.86%	68.77%	66.00%	51.98%	65.92%
	综合	80.31%	70.11%	78.45%	77.23%	69.43%	78.21%

由表1和图1可以看出：

(1)从语料的正反查全率、查准率来看，正面查全率、查准率比较高，主要原因一是正面的规模比反面语料的规模大。我们已做了测试，当反面评论数量增加时，反面的查全率和查准率、总体评价性能都会得到相应的提高；二是语料自身的一些特点（如评论本身正面、反面用语交替出现）也是导致这一结果不可忽视的原因。

(2)从特征抽取方法看，基于信息增益(IG)在各种停用词表的情况下，分类效果最好，χ^2 统计(X2)次之，互信息(MI)方法最差。

(3)从权重计算看，利用停用词表1得到的分类结果，词频型与布尔型一致，其余的停用词表，词频型权重比词频型权重的整体分类效果要好，这与英文的认识是一致的[6]。

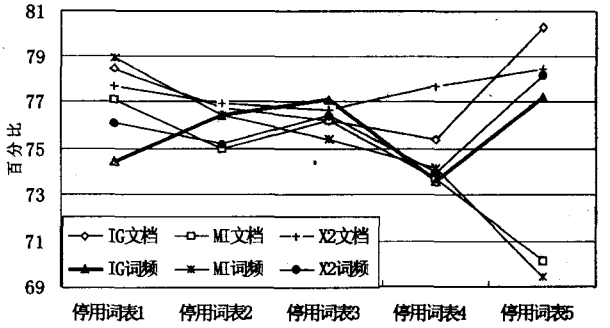

图1 不同停用词表、不同特征抽取方法以及不同权重的情感分类结果 F 值

(4)以停用词表2和停用词表3作为停用词的筛选，各种特征选取方法得到的分类结果相差不大，大部分集中于74%～77%，说明去掉这两种停用词后对特征抽取方法依赖不大；对于不使用停用词表(停用词表5)时各个特征抽取方法得到的分类结果波动最大，最好的性能达到了80.3%，最差的为69.43%，最大差超过了10%；对于选用停用词表4和停用词表1时各个特征抽取方法得到的结果波动低于无停用词表(停用词表5)的情况，但使用停用词表1时各个特征抽取方法得到的性能比选用停用词表4的整体效果好。总的来说，不使用停用词表(停用词表5)与停用词表1(不含avd)时各个特征抽取方法得到的结果较好，这与人的直觉认识是一致。

利用五种停用词表得到候选特征后，再使用信息增益进行特征选择，得到了524个特征的交集，部分特征为“盈利、异常、严重、先进、心疼、喜悦、偷工减料、热爱、强烈、轻微、耐用、勉强、辉煌、华丽、脆弱、高档、刺激、满意、安静、沉稳、安全”，表明这524个特征不论采用什么样的停用词表均被选为特征，即它们对情感分类的作用是比较大的。

5 结束语

文本情感分类在信息检索、文本过虑、产品的在线跟踪评价，民情民意调查分析以及聊天系统具有广泛的应用。本文针对三种特征抽取方法信息增益(IG)、互信息(MI)与 χ^2 统计(X2)，两种权重计算方法文档频率与词频统计，在五种不同的停用词表中利用支持向量机分类器对文本情感分类进行了实验，结果表明信息增益(IG)和布尔型对情感分类的效果整体比较好。

当选用不同的停用词表时，它们文本情感分类

的影响也不尽相同,停用词表5(不去掉停用词)、停用词表1(仅选用形容词、动词和副词)对情感分类作用较大,整体性能效果较好。但同时发现对于情感倾向性的判定中,特征词应该是形容词、副词、动词、助词、感叹词等与情感表达相关的带有较强情感色彩的短语或词汇[9]。但实验的结果是"路程、外方、GPS、眼球、桑塔纳"等名词居多,它们与主题相关。另一个现象是很多正面的形容词、副词、动词,如"巨大、完好、有效、风范、热销"同时也出现在反面评论中,使得文本情感分类的各项性能指标均低于已有相关主题文本分类报道的结果[1],说明了中文文本情感分类比主题分类更复杂。通过对于汽车评论的情感分类,帮助用户对所关注的汽车表现有一个整体了解,而且还可以对汽车产品起到推荐作用。

参 考 文 献

[1] 顾益军,樊孝忠,王建华,汪涛,黄维金.中文停用词表的自动选取[J].北京理工大学学报,2005,25(4):337-340.

[2] Hart G W. To decode short cryptograms[A]. Communications of the ACM [C]. New York Association for Computing Machinery,1994:102-108.

[3] Yang Y, Pedersen J O. Acomparative study on feature selection in text categorization// Proceedings of ICML-97,14th Internationa Conference on Machine Learning [C]. San Francisco Morgan Kaufmann Publishers Inc,1997:412-420.

[4] Silva C, RibeiroB. The importance of stop word removal on recall values in text categorization [J]. Neural Networks, 2003,3:20-24.

[5] 李荣陆.文本分类若干关键技术研究.上海:复旦大学博士论文,2005.

[6] Pang B, Lee L, Vaithyanathan S. Thumbs up? Sentiment classification using machine learning techniques [C]. The Conference on Empirical Methods in Natural Language Processing,2002:79-86.

[7] Peter D Turney, Michael L Littman. Measuring Praise and Criticism: Inference of Semantic Orientation from Association. ACM Transaction on information systems,2003, 21(4):315-346.

[8] Hatzivassilolou Kathleen V, Mckeown R. Predicting the semantic orientation of adjectives. Proceeding of the 35th Annual meeting of the association for computational linguistics and the 8th conference of the European Chapter of the ACL. Association for Computational Linguistics, New Brunswick, 1997:174-181.

[9] 王治敏,朱学峰,俞士汶.基于现代汉语语法信息词典的词语情感评价研究. Computational Linguistics and Chinese Language Processing,2005,10(4):581-592.

(责任编辑 王建平)