◎博士论坛◎

基于机器学习的中文微博情感分类实证研究

刘志明,刘 鲁
LIU Zhiming, LIU Lu
北京航空航天大学 经济管理学院,北京 100191
School of Economics and Management, Beihang University, Beijing 100191, China

LIU Zhiming, LIU Lu. Empirical study of sentiment classification for Chinese microblog based on machine learning. Computer Engineering and Applications, 2012, 48(1):1-4.

Abstract: With the development of microblog, it is more convenient to comment on the Web. Up to now, there are very few studies on the sentiment classification for Chinese microblog, therefore this paper uses three machine learning algorithms, three kinds of feature selection methods and three feature weight methods to study the sentiment classification for Chinese microblog. The experimental results indicate that the performance of SVM is best in three machine learning algorithms, IG is the better feature selection method compared to the other methods, and TF-IDF is best fit for the sentiment classification in Chinese microblog. Combining the three factors the conclusion can be drawn that the performance of combination of SVM, IG and TF-IDF is best. For the movie domain it is found that the sentiment classification depends on the review style.

Key words: microblog; sentiment classification; machine learning; feature selection; term weight

摘 要: 使用三种机器学习算法、三种特征选取算法以及三种特征项权重计算方法对微博进行了情感分类的实证研究。实验结果表明,针对不同的特征权重计算方法,支持向量机(SVM)和贝叶斯分类算法(Naïve Bayes)各有优势,信息增益(IG)特征选取方法相比于其他的方法效果明显要好。综合考虑三种因素,采用SVM和IG,以及TF-IDF(Term Frequency-Inverse Document Frequency)作为特征项权重,三者结合对微博的情感分类效果最好。针对电影领域,比较了微博评论和普通评论之间分类模型的通用性,实验结果表明情感分类性能依赖于评论的风格。

关键词: 微博;情感分类;机器学习;特征选取;特征项权重
DOI:10.3778/j.issn.1002-8331.2012.01.001 文章编号:1002-8331(2012)01-0001-04 文献标识码:A 中图分类号:TP39

1 引言

互联网的兴起,特别是近几年随着Web2.0应用的增多,网民对各种产品以及热点事件的评论变得更加方便。针对产品的评论,不管对商家还是买家都是十分有价值的;对于热点事件的评论,对于政府了解网民对特定事件的观点也是十分重要的。情感分类作为一项新兴技术已经在这些领域得到了大量的研究[1-3]。情感分类技术就是将人们的情感分为正面情感和负面情感,当前研究使用的主要方法分为两种:基于机器学习的方法[1,3]和基于语义的方法[4-5]。基于机器学习的方法将情感分析问题看作是一个分类问题,标注好的训练集通过机器学习算法训练得到分类模型,用于以后的情感分类。基于语义的方法将表示情感的词语分为正面情感词语和负面情感词语,构造一个情感词典,然后通过计算一个句子中的正负情感词语的相对数量决定句子的情感倾向。当前很多研究结果表明[1-2],基于机器学习的方法性能表现比基于语义的方法好。

微博作为近几年发展起来的一种应用,正在受到研究者的关注。微博相比于传统的评论有五种特有的性质:

(1)长度:微博的长度限制在140个字符,相比于传统的评论,长度相差很大,根据收集到的语料统计,平均长度为40个字符;正是因为长度有限制,所以微博中网民的观点更容易理解。

(2)数据易获取性:数据获取相对更加容易,当前大部分微博都提供API,可以很方便地获取大量的数据。

(3)特有的语言风格:微博信息的来源是多样的,网民可以通过手机、客户端、插件多种形式发布信息,所以相比于传统的博客以及产品评论来说,微博的语言更多地会出现一些新兴的词语,或者是错误的拼写。

(4)信息多样性:微博中的信息来自不同领域,网民可以针对产品发表评论,也可以针对当前热点事件发表评论,所以从微博中可以获取不同领域的信息。当前大部分微博都提供关键词搜索功能,可以根据相关领域关键词搜索相关的信息。

(5)实时性:发布微博的渠道多种多样,网民随时随地都可以将自己的观点发布到微博,所以微博的实时性相比于传统的评论更加及时,这对于那些对时间要求更高的应用无疑是一个更加合适的信息来源。

从上面分析的特点来看,将微博作为评论来源进行情感分类的研究是十分有意义的。目前,国内外相关的研究相对较少,国外一些学者对twitter进行了情感分类的相关研究[6-7];针对中文微博的研究当前十分缺乏,文献[8]针对微博提出了一种基于语义的方法,通过定义态度词典、权重词典、否定词典、程度词典以及连接词词典来计算每条微博的情感指数,他

基金项目:国家自然科学基金(No.90924020);教育部博士点基金(No.200800060005)。
作者简介:刘志明(1979—),男,博士研究生,研究方向:管理信息系统,数据挖掘;刘鲁(1947—),通讯作者,女,教授,博士生导师。
E-mail:liulu@buaa.edu.cn
收稿日期:2011-06-17;修回日期:2011-09-19;CNKI出版:2011-10-24;http://www.cnki.net/kcms/detail/11.2127.TP.20111024.1013.068.html

们使用的数据来自中文微博饭否。然而当前没有发现使用机器学习方法进行微博情感分类的相关研究。为了填补这个空白,本文使用三种机器学习方法、三种特征选取算法以及三种特征项权重计算方法对中文微博进行了实证研究,并且比较了微博和普通评论之间分类模型的通用性。

2 相关知识介绍

2.1 机器学习方法

2.1.1 支持向量机分类方法

支持向量机分类方法(SVMs)是基于结构风险最小化原理的一种新颖的机器学习算法[9],是一种具有很好泛化能力的预测工具,已经被广泛应用于文本分类以及人脸识别等领域。在文本分类领域,SVM被证明是非常高效的,与传统的方法相比SVM鲁棒性更好[10]。

在样本可分情况下的支持向量机称为线性支持向量机,由于大部分文本数据是线性可分的,所以本文只考虑线性支持向量机。本文使用LIBLINEAR用于分类模型的训练及测试,LIBLINEAR是Rong-En Fan[11]提出的用于大规模线性文本分类的一种SVMs算法,对于高维稀疏数据集特别有效。

2.1.2 贝叶斯分类算法

贝叶斯文本分类算法(Naïve Bayes)是经常被使用的一种文本分类方法,Naïve Bayes利用贝叶斯定理来预测一个未知类别的样本的可能属性,选择其可能性最大的类别作为样本的类别。模型虽然简单,但是在文本分类领域应用非常广泛[12]。针对文本分类存在两种不同的贝叶斯模型:多项式模型(multinomial model)和多变量贝努利模型(multi-variate bernoulli model)。

当前有大量的学者使用多项式模型进行文本分类的研究[2,13-14],所以本文也选取多项式贝叶斯分类算法进行实验。

多项式贝叶斯分类模型通过公式(1)计算给定类 c_j 词 W_t 出现的概率:

$$p(w_t|c_j)=\frac{\sum_{i=1}^{N_j} n_{it}}{\sum_{s=1}^{W}\sum_{i=1}^{N_j} n_{is}} \tag{1}$$

其中,n_{it} 表示文档 i 中词 t 出现的次数,N_j 表示类别 c_j 的训练集的大小,W 表示词典大小。

后验概率通过公式(2)计算:

$$p(c_j|d_i)=\frac{p(c_j)p(d_i|c_j)}{p(d_i)} \tag{2}$$

2.1.3 n元语言模型

使用 n 元语言模型进行文本分类是自然语言处理中的一个新的模型[15],与传统的向量空间模型不同,n 元语言模型把文档看作是词的序列,这样词的出现与否可以看成是一种语言结合模式,使用这些结合模式可以用来对文档进行分类。

对于一个字符串 $s=c_1c_2\cdots c_{n-1}c_n$,$n$ 元语言模型假设认为第 n 个字符出现的概率只与前 $n-1$ 个字符有关,即

$$p(c_n|s_{c_1c_2\cdots c_{n-1}})=p(c_n|c_1c_2\cdots c_{n-1}) \tag{3}$$

2.2 特征选取方法介绍

2.2.1 信息增益

信息增益(IG)方法是文本分类中经常使用的一种特征选取方法[16],对于特征 t,通过测量加入特征 t 相对于去掉特征 t 对分类性能的影响来衡量特征 t 的分类能力。信息增益公式如下:

$$IG(t)=-\sum_{i=1}^{|C|}P(c_i)\lg P(c_i)+P(t)\sum_{i=1}^{|C|}P(c_i|t)P(c_i|t)+P(\bar{t})\sum_{i=1}^{|C|}P(c_i|\bar{t})\lg P(c_i|\bar{t}) \tag{4}$$

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;$P(\bar{t})$ 表示特征 t 不出现的概率。

2.2.2 CHI统计

CHI统计的方法通过测量特征与类别之间的依赖性来进行特征的选取,CHI值越大意味着特征与类别之间的依赖性更强,相反CHI值越小意味特征与类别之间相对独立。CHI值的计算公式如下:

$$CHI(t,c_i)=\frac{N(N_{11}N_{00}-N_{10}N_{01})^2}{(N_{11}+N_{01})(N_{11}+N_{10})(N_{10}+N_{00})(N_{01}+N_{00})} \tag{5}$$

$$CHI(t)=\max_i(CHI(t,C_i)) \tag{6}$$

其中,N 表示训练集中文档总数;N_{11} 表示特征 t 与类别 c_i 同时出现的次数;N_{10} 表示包含特征 t 并且类别不是 c_i 的文档数;N_{01} 表示类别为 c_i 但不包含特征 t 的文档数;N_{00} 表示类别不是 c_i 并且不包含特征 t 的文档数。

2.2.3 文档频率

文档频率(DF)是一种最简单的特征选择方法,通过设置文档频率阈值来进行特征的选取。文档频率是指包含某个特征的文档数量,文档频率特征选取方法认为,文档频率太低或者太高的特征对文本的分类作用都不大,因此可以删去这些特征。虽然简单,但是DF方法在中文分类以及英文分类中都有较好的性能[17-18]。

3 数据集的收集

3.1 微博数据集(datasetA)

因为国内现在没有通用的微博数据集,所以通过爬虫程序从新浪微博上抓取了部分数据,新浪微博按照主题对微博进行了分类,为了避免实验结果过于依赖特定领域,抓取的数据来自四个主题:甲流疫苗、王家岭矿难、电影评论以及春游活动。首先小组三个成员分别对语料进行情感倾向标注,然后从三个标注中选取最多的情感倾向赋予每条评论,最后共得到2 134条评论,其中包括正面评论1 002条,负面评价1 132条。

3.2 微博影评与普通影评数据集(datasetB)

为了测试微博与传统评论之间的情感分类模型通用性,从新浪微博和豆瓣分别采集了电影领域的评论集。新浪微博的电影评论共4 000条,其中正面评论2 000条,负面评论2 000条,标注的方式与数据集datasetA相同;豆瓣评论共1 000条,因为豆瓣评论有1~5星的评级,将4星和5星评级的评论标注为正面评论,将1星和2星评级的评论标注为负面评论,其他有没有评级的评论过滤掉,共得到正面评论和负面评论分别为500条。

通过对收集到的影评进行统计,微博评论平均长度为40个字符,普通影评平均长度为1 155个字符。

4 实验

4.1 实验设计

实验中首先采用ICTCLAS[19]对每条评论进行中文分词,

根据实验的需要选取某种特征项权重计算方法构建向量空间模型,然后采用相应的特征选择方法进行特征的选择,最后使用三种机器学习算法训练分类模型。

用WEKA实验环境(http://www.cs.waikato.ac.nz/ml/weka/)进行SVM、Naïve Bayes算法实验,使用Lingpipe(http://alias-i.com/lingpipe/index.html)进行n元语言模型的实验。

实验使用10折交叉验证方法,选取F-SCORE作为性能评测指标,公式(7)为F-SCORE的公式。

$$F=\frac{2\times Recall\times Precision}{(Recall+Precision)} \tag{7}$$

其中,$Recall$表示算法召回率,$Precision$表示算法准确率。

4.2　实验结果及分析

4.2.1　不同特征项权重表示的性能比较

本实验比较下列三种特征项权重计算方法:

(1)布尔型特征权重(Presence):如果特征出现在文档中,权重为1,否则权重为0。

(2)词频型特征权重(TF):将特征出现在文档中的次数作为该特征的权重。

(3)TF-IDF(Term Frequency-Inverse Document Frequency)特征权重:修正词频型特征权重,将包含此特征的文档数作为一个考虑因素,认为包含特征的文档数越多,特征的区分能力越差,使用如下的公式计算:

$$W(t,d)=tf(t,d)\times\lg(\frac{N}{n_t}) \tag{8}$$

其中,N表示总的训练文档集中文档数量,n_t表示包含词t的文档数量。

当前大多数研究都是直接采用某种特定的特征表示方法[2,14],文献[1]对英文情感分析中Presence和TF进行了比较,结果表明Presence表现更好。文献[20]对中文新闻的情感分析比较了Presence和TF的性能,结果表明Presence性能更好。而对于微博领域,这方面的比较研究还是空白的,所以本文设计实验来比较三种不同权重计算方法的性能。

实验中特征选取算法选用信息增益算法,分类算法采用SVM和Naïve Bayes算法,图1、图2是采用三种权重计算方法的性能比较。

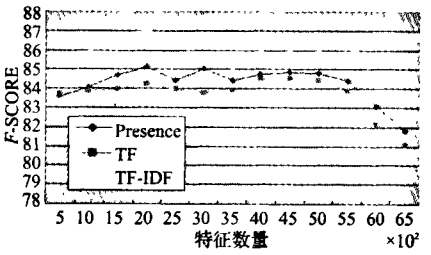

图1　三种权重在SVM中的性能比较图

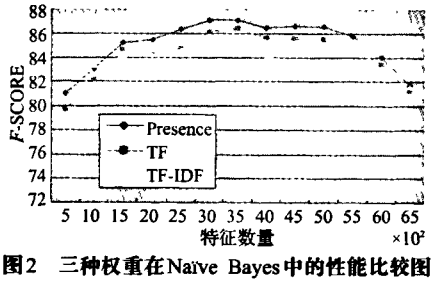

图2　三种权重在Naïve Bayes中的性能比较图

从图1可以看出,三种权重对于不同的机器学习方法各有优势。从图1可以看出,当采用SVM分类算法时,TF-IDF的性能是最好的,而Presence和TF的性能比较接近。从图2可以看出,当采用Naïve Bayes算法时,Prensence的性能是最好的,TF的性能也比较接近Presence,但是TF-IDF的性能表现不好,特别是特征数量取3 000~4 000时性能出现明显的下降。

将分类算法和权重表示综合考虑,可以发现当权重采用TF-IDF时,SVM在特征数为2 000的时候性能达到最优,F-SCORE值为87.07;当权重采用Presence时,BAYES在特征数为3 000的时候性能达到最优,F-SCORE值为87.07。所以针对IG特征选取方法而言,选取SVM结合TF-IDF权重表示的组合性能最佳。

4.2.2　不同特征选取方法的比较

实验采用SVM作为分类算法,采用TF-IDF权重计算方法,比较不同特征选取方法的性能。实验结果如图3。

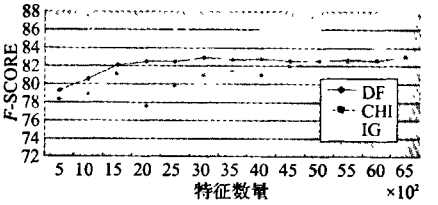

图3　特征选取方法的比较图

从图3可以看出IG相比于CHI统计、DF优势十分明显,当特征数选择在2 000时,IG性能就得到了最优,准确率高达87.07,而CHI统计、DF两种特征选取方法性能差别不大,但是CHI统计表现不稳定。当特征数选取在2 500以上时,三种方法的性能基本稳定。

4.2.3　三种机器学习算法的比较

实验的目的是比较三种机器学习算法的性能,由于在实验1中发现SVM、Naïve Bayes的性能是依赖于不同的权重计算方法的,所以实验中,考虑三种不同的权重计算方法来进行比较。n元语言模型不存在权重计算问题,实验中依次设置n从2到8,取其中最优的值作为结果。实验结果如表1。

表1　三种机器学习算法性能比较表

分类算法	权重表示		
	Presence	TF	TF-IDF
SVM	85.10	84.54	87.07
Naïve Bayes	87.07	86.41	84.91
N-GRAM		82.32	

由表1可知,相比于SVM和Naïve Bayes算法,n元模型性能最差。而SVM和Naïve Bayes方法的性能是依赖于不同权重计算方法的,当采用TF-IDF时,SVM性能更好,而采用Presence时,Naïve Bayes性能更好。

从前面的实验可得到,权重计算方法采用TF-IDF,分类算法使用SVM,特征选取算法使用IG的组合表现是最好的。后面的实验如无特殊说明,都采用这种组合。

4.2.4　微博与普通评论之间的比较

文献[2]通过实验证明情感分类器严重依赖于不同的领域或者主题。由于微博和普通评论有很多不同的特点,同一领域两种不同风格的评论分类器能否通用值得去研究。

实验的目的是通过对两种不同风格的影评进行情感分类的比较研究,分析同一领域评论的情感分类器是否依赖于评论的风格。

将datasetB中微博评论集分为训练集和测试集，训练集3 000条，测试集1 000条；同样将豆瓣评论集也分为训练集和测试集，训练集700条，测试集300条。然后分别对两个训练集进行训练，得到相应的分类模型，然后分别用两个测试集进行测试。得到的分类性能比较如表2。

表2 微博与普通评论模型通用性比较表

训练集	测试集	SVM	Bayes	*N*-GRAM
微博训练集	普通测试集	75.77	78.50	75.23
微博训练集	微博测试集	86.68	84.32	85.12
普通训练集	微博测试集	63.00	73.25	70.45
普通训练集	普通测试集	76.08	79.89	79.61

从表2可以发现，相比于同一种类型的评论分类性能来说，不同评论建立的模型之间的通用性表现并不好。这很可能是因为两种评论所表达情感的方式不同造成的，微博更倾向于直接表达情感，句子中更多地包含了情感词语，而普通的评论情感词语掺杂在一些事实语句中。

5 结语

针对微博进行了情感分析的研究，通过实验发现3种机器学习的方法对于微博的情感分析都是有效的，其中采用TF-IDF权重计算方法，分类算法采用SVM，通过IG进行特征的选取得到的分类效果是最佳的。进而针对电影评论研究了微博和普通评论之间情感分类模型的通用性，实验数据说明两种不同风格的评论在通用性方面相对比较差，建立对不同风格的评论通用的情感分类算法也是一个需要研究的问题。本文只是对机器学习算法在微博的情感分析领域进行了初步的研究，一些更深入的研究还需要进一步进行，比如基于机器学习的算法和基于语义的算法之间性能的比较，以及针对某些具体领域微博情感分析的应用是否有效。未来的工作是具体到生物医药领域，通过微博研究大众对突发事件所表达的情感演化问题。

参考文献：

[1] Pang Bo, Lee L, Vaithyanathan S.Thumbs up?Sentiment classification using machine learning techniques[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing(EMNLP), 2002:79-86.

[2] Tan Songbo, Zhang Jin.An empirical study of sentiment analysis for Chinese documents[J].Expert Systems with Applications, 2008:2622-2629.

[3] Mullen T, Collier N.Sentiment analysis using support vector machines with diverse information sources[C]//Proceedings of Methods in Natural Language Processing, Barcelona, Spain, 2004: 412-418.

[4] Hatzivassiloglou V, McKeown K.Predicting the semantic orientation of adjectives[C]//Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics(ACL), 1997: 174-181.

[5] 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(ACL), Philadelphia, 2002:417-424.

[6] O'Connor B, Balasubramanyan R.From tweets to polls: linking text sentiment to public opinion time series[C]//Proceedings of the International AAAI Conference on Weblogs and Social Media, Washington DC, 2010:1-8.

[7] Jansen B J, Zhang Mimi.Twitter power: tweets as electronic word of mouth[J].Journal of the American Society for Information Science and Technology, 2009:2169-2188.

[8] Shen Yang, Li Shuchen.Emotion mining research on micro-blog[C]// 2009 1st IEEE Symposium on Web Society, 2009:71-75.

[9] Vapnik V N.The nature of statistical learning theory[M].New York: Springer, 1995.

[10] Joachims T.Text categorization with support vector machines: learning with many relevant features[C]//Proceedings of 10th European Conference on Machine Learning, 1998:137-142.

[11] Fan Rongen, Chang Kaiwei.LIBLINEAR:a library for large linear classification[J].Journal of Machine Learning Research, 2008: 1871-1874.

[12] Domingos P, Pazzani M J.On the optimality of the simple bayesian classifier under zero-one loss[J].Machine Learning, 1997: 103-130.

[13] McCallum A, Nigam K.A comparison of event models for naive bayes text classification[C]//AAAI-98 Workshop on Learning for Text Categorization, 1998:41-48.

[14] Ye Qiang, Zhang Ziqiong, Law R.Sentiment classification of online reviews to travel destinations by supervised machine learning approaches[J].Expert Systems with Applications, 2009, 36: 6527-6535.

[15] Carpenter B.Scaling high-order character language models to gigabytes[C]//Proceedings of the 2005 Association for Computational Linguistics Software Workshop, 2005:1-14.

[16] Yang Y, Pedersen, Jan O.A comparative study on feature selection in text categorization[C]//ICML, 1997:412-420.

[17] Dai Liuling, Huang Heyan, Chen Zhaoxiong.A comparative study on feature selection in Chinese text categorization[J]. Journal of Chinese Information Processing, 2004, 18(1):26-32.

[18] 单松巍，冯是聪，李晓明.几种典型特征选取方法在中文网页分类上的效果比较[J].计算机工程与应用，2003，39(22)：146-148.

[19] 张华平.基于多层隐马尔科夫模型的中文词法分析[C]//第41届ACL会议暨第二届SIGHAN研讨会，札幌，日本，2003：63-70.

[20] 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.