多层次中文微博情绪分析

刘宝芹,牛 耘
(南京航空航天大学 计算机科学与技术学院,江苏 南京 210016)

摘 要:文中旨在对中文微博文本中表达的情绪进行自动分析。目前,微博情绪分析的方法主要是平面型分类方法。该方法认为各个情绪类之间相互独立,相互并列,它们处在同一个平面层次上,只需要一次性构建一个分类器就可以完成情绪分类任务。事实上,Ekman六类情绪之间的关系并不完全独立。文中将Ekman六类情绪按照情感极性及情绪间的相互关系组织成三层树状结构,在此基础上提出了一种基于朴素贝叶斯模型的多层次中文微博情绪分析方法。实验结果表明,与传统的平面型朴素贝叶斯分类方法相比,文中提出的多层次微博情绪分析方法降低了各情绪类微博分布不平衡对分类结果造成的影响,提高了微博情绪识别的精度。

关键词:微博情绪;朴素贝叶斯;平面型分类方法;层次型分类方法
中图分类号:TP391 **文献标识码**:A **文章编号**:1673-629X(2015)11-0023-04
doi:10.3969/j.issn.1673-629X.2015.11.005

Multi-hierarchy Emotion Analysis of Chinese Microblog

LIU Bao-qin, NIU Yun
(School of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China)

Abstract: Focus on automatic classifying emotion expressed in Chinese microblog. At present, most of the research on emotion analysis of microblog has focused on flat classification approach. The approach considers emotion categories are independent and coordinate of each other. They are in the same flat level and need only one classifier to achieve the emotion classification task. In fact, the relationship between Ekman's six emotion classes is not completely independent. Ekman's six emotion classes are arranged into a tree structure according to the relationship between emotional polarity and mutual relations. Based on this tree structure, an approach of hierarchical emotion analysis of Chinese microblog based on Naïve Bayesian Model is proposed. Compared with the flat classification, the experimental results show that hierarchical classification method reduces the effect of the highly imbalanced distribution of the emotional categories on the classification results, and improves the accuracy of microblog emotion analysis.

Key words: micro-blog emotion; Naïve Bayesian; flat classification approach; hierarchical classification approach

0 引 言

随着互联网的蓬勃发展,用户通过微博、博客、论坛等社交媒体主动发布的文本越来越多。微博以内容简短、即时分享、快速传播的特色成为用户分享、传播、获取信息以及抒发个人情绪的重要社交网络平台。分析微博中所包含的情绪,可以帮助用户及时了解自身情绪的波动情况,帮助企业理解用户的消费习惯,制定营销策略,还可以帮助政府分析热点事件的舆情,从而为政府制定决策提供重要依据。

文中的研究目的是自动分析判别微博文本中表达的情绪。情绪是人基于个体本能的需要而产生的身体与心理状态。Ekman[1]通过研究人的面部表情,将情绪划分为六种基本状态:喜(joy)、哀(sad)、怒(anger)、惧(fear)、惊(surprise)、恶(disgust)。文中将以这六类情绪作为情绪类别对微博文本中表达的情绪进行自动分析。

目前,微博情绪分析的方法主要以平面型分类(flat classification)方法为主[2-7]。该方法认为各个情绪类之间相互独立,相互并列,它们处在同一个平面层次上,只需要一次性构建一个分类器就可以完成情绪分类任务。事实上,Ekman的六类情绪之间的关系并不完全独立。如joy类可以被划分为positive类,其他

收稿日期:2015-02-02 **修回日期**:2015-05-06 **网络出版时间**:2015-11-04
基金项目:国家自然科学基金资助项目(61202132,61170043)
作者简介:刘宝芹(1989-),女,硕士研究生,研究方向为自然语言处理;牛 耘,博士,副教授,研究方向为自然语言处理。
网络出版地址:http://www.cnki.net/kcms/detail/61.1450.TP.20151104.0948.010.html

五类可以被划分为 negative 类;而 anger 类和 disgust 类非常接近,不易区分。文中根据六类情绪的情感极性及情绪间的相互关系将六类情绪组织成三个层次,然后利用基于朴素贝叶斯模型的多层次分类(hierarchical classification)方法对中文微博进行情绪分析。实验结果表明,文中提出的多层次微博情绪分析方法降低了各情绪类微博分布不平衡对分类结果造成的影响,提高了微博情绪识别的精度。

1 相关工作

1.1 微博情绪分析

目前,微博情绪分析主要有监督机器学习的方法[2-4]和基于规则的方法[5-7],而这些方法主要是平面型方法。

以 Twitter 为代表的英文微博的情绪分析已取得了一定的研究成果。Purver 等[2]利用远程监督(distant supervision)的方法进行 Ekman 情绪分类。他们利用人工选取的标签(hashtag)和表情符(emoticon)来自动标注微博情绪,省去了人工标注语料的过程。实验结果表明,fear,surprise,disgust 三类情绪不易区分。Paltoglou 和 Thelwall[5]利用基于情绪词典 LIWC 的无监督方法对 Twitter、MySpace、Digg 三个社交媒体的微博进行主、客观分类和正负向情感分类。实验结果表明,多数情况下该方法对不同领域的适应性较强,性能优于有监督机器学习的方法。

中文微博情绪分析的研究起步较晚。刘志明和刘鲁[3]采用三种机器学习算法、三种特征选择方法、三种特征项权重计算方法对微博进行正负向情感分类。实验结果表明,使用 SVM,信息增益和 TF-IDF 三者结合的方法对微博情感的分类效果最好。张晶等[6]先根据常用情绪词和情绪短语构建情绪词典,再结合情绪表达方式、标点符号、表情符号建立情绪规则来识别微博情绪。该方法虽然具有可行性,但是建立情绪规则比较困难。

1.2 微博情绪分析中的层次型分类方法

在很多分类问题中,类别之间并不是相互独立的,而是具有一定的层次关系,层次型分类方法即针对这类问题。层次型分类方法将类别组织成某种类型的层次结构(一般组织成树状结构),然后利用类别层次结构提供的信息帮助提高分类性能。

目前,微博情绪分析主要以平面型分类方法为主。层次型分类方法主要用于文本分类任务中,而在情感分类任务中主要用于对微博情感进行正负情感极性的判别[8-10]。Jiang 等[8]将 Tweets 情感划分为两个层次,通过考虑主题相关特征以及 Tweets 间的转发关系,采用二步分类法对 Tweets 的情感进行分类。该方法首先对 Tweets 进行主、客观分类,然后对被分为主观的 Tweets 进行正、负向分类,准确率达到 66%。Jiang 等[9]建立了与文献[8]类似的情感层次结构,通过提取微博结构特征、句子结构特征、情感词典特征、表情符特征,采用二步二分类法对中文微博情感进行分类。该方法第一步对微博进行有情感、无情感分类,第二步对有情感的微博进行正、负向情感分类,准确率达到 60.1%。

Keshtkar 和 Inkpen[11]利用层次型分类方法将博客的心情分为 132 类,然而情绪与心情不同,情绪的持续时间比心情的持续时间短,两者的层次结构和分类任务也不同。因此,文中的研究工作与 Keshtkar 和 Inkpen 的研究工作有所不同。文中根据六类情绪间的情感极性及情绪间的相互关系为六类情绪建立树状层次结构,并且利用该结构对微博情绪进行自动分析。

2 多层次微博情绪分析方法

2.1 微博情绪的层次结构

文中根据情绪的正负极性及情绪间的关系,将六类情绪分为三个层次,微博情绪的层次结构如图 1 所示。

其中,第一个层次是对情感正负极性的划分,文中将 joy 类的情感极性划分为正向(positive),其他五类情绪的情感极性划分为负向(negative)。第二个层次是对负向情感的划分,其中 anger 类和 disgust 类合并为 anger-disgust 类。Ekman 在文献[12]中指出 anger 类和 disgust 类的面部表情最易混淆;Alm[13]在对儿童故事情绪进行分类时也将 anger 类和 disgust 类进行了合并。第三个层次是对 anger-disgust 类的划分。

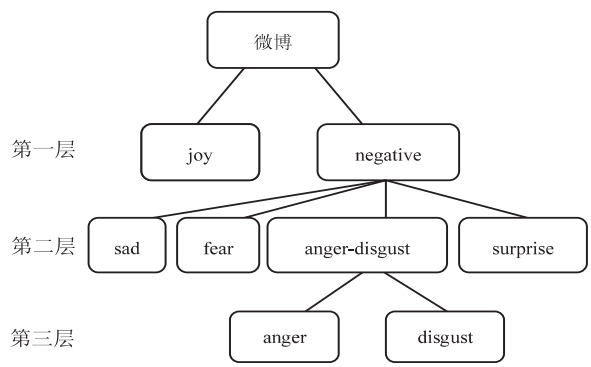

图 1 微博情绪的层次结构

2.2 分析方法

文中提出了基于朴素贝叶斯模型的多层次微博情绪分析方法。使用的分类算法是朴素贝叶斯(NB)算法。该算法是一种基于概率统计的算法,它实现简单,分类准确率高,速度快。

层次型分类方法一般将类别组织成树状结构,每个类别相当于树中的一个节点。根据分类过程中分类

器的数目,可以将层次型分类方法分为如下两种类型:局部分类器方法(local classifier approach)和全局分类器方法(global classifier approach)。下面以树状层次结构为例,简单介绍一下这两种类型以及根据图 1 中的层次结构如何利用它们进行微博情绪分析。

(1)局部分类器方法:又称为自顶向下方法(top-down approach)。该方法需要建立多个分类器。Silla 和 Freitas[14] 根据建立分类器方式的不同将该方法细分为三类:每个节点一个分类器(a Local Classifier per Node,LCN);每个父节点一个分类器(a Local Classifier per Parent Node,LCPN);每层一个分类器(a Local Classifier per Level,LCL)。

LCN 方法为除根节点以外的每个节点建立一个二元分类器。该方法的缺点是对微博进行情绪分类时,会导致父类-子类间的预测结果不一致。如:为 negative 类建立的二元分类器将情绪判断为非 negative 类。为 fear 类建立的二元分类器将情绪判断为 fear 类。在图 1 中,fear 类属于 negative 类,即如果该条微博的情绪是 fear 类,那么该条微博的情绪也一定是 negative 类,但为 negative 建立的二元分类器却将情绪判断为非 negative 类,这就出现了父类-子类间预测结果不一致的情况。

LCPN 方法为每个父节点建立一个多类分类器,将情绪判断为其子类中的某一类。

LCL 方法为在每层建立一个多类分类器,将情绪判断为该层中的某类。该方法不同层次的分类结果也可能导致父类-子类间的预测结果不一致。

(2)全局分类器方法:又称为爆炸式(big-bang)方法。该方法从层次结构的全局出发,考虑层次结构中的所有类别,为它们建立一个分类器,该分类器包含了整个类别层次结构的信息。

考虑到 LCN 方法和 LCL 方法的缺点,文中选取 LCPN 方法和全局分类器方法来分析微博情绪。LCPN 方法为图 1 中的每个父节点建立一个分类器,共建立三个分类器。首先利用根节点分类器将情绪分为图 1 第一层中的某类。若该分类器将情绪判断为 joy 类,则分类结束。若判断为 negative 类,则利用为 negative 节点建立的分类器将情绪判断为图 1 第二层中的某一类。若该分类器将情绪判断为 sad,fear,surprise 中的某类,则分类结束。若判断为 anger-disgust 类,则利用为 anger-disgust 节点建立的分类器将情绪判断为 anger 类或 disgust 类。

全局分类器方法为图 1 中包括父节点在内的八个情绪类只建立一个分类器,而文中的分类任务是将微博情绪判断为六个叶节点中的某类。所以在测试阶段,若分类器将情绪判断为层次结构中的叶节点类,则分类任务结束;若分类器将情绪判断为 negative 类,则需进一步将情绪判断为层次结构第二层的某类或第三层的某类。若该分类器将情绪判断为第二层 anger-disgust 类,则进一步将情绪判断为 anger 类或 disgust 类。

3　实　验

3.1　实验数据

文中利用新浪提供的 API 抓取不同话题的微博文本。选取若干话题的微博,由两名标注人员各自独立对文本进行情绪标注。每条微博标注为喜、哀、怒、惧、恶、惊和其他共七类中的一类。将两名标注员标注结果一致的微博文本提取出来作为实验数据集以保证数据的可靠性。六类情绪的微博分布情况如表 1 所示。

表 1　六类情绪的微博分布情况

情绪类别	joy	sad	anger	fear	disgust	surprise	合计
数量	473	173	276	147	145	121	1 335

3.2　实验设置

文中评测方法采用五折交叉验证(five-fold cross-validation),评测指标选用精确率 P(Precision = TP/(TP+FP))、召回率 R(Recall = TP/(TP+FN))和 f-score(f-score = $2P \times R/(P+R)$)。

3.3　实验结果分析

将第 2 节提出的多层次情绪分析方法和平面型分类方法进行比较,结果如表 2 所示。

表 2　三种方法的分类结果比较

	精确率			召回率			f-score		
	GC	LCPN	NB	GC	LCPN	NB	GC	LCPN	NB
negative	82.8	90.1	65.8	79.6	90.8	64.6	81.2	90.4	65.2
joy	87.6	83.4	81.3	77.8	81.8	83.3	82.4	82.6	82.3
sad	54.7	58.1	55.4	47.4	47.4	47.4	50.8	52.2	51.1
anger	60.3	67.3	65.8	82.4	73.6	79.5	69.7	70.3	72.0
fear	84.1	81.2	84.0	60.5	64.6	62.5	70.4	72.0	71.7
disgust	55.9	54.0	61.2	84.0	84.0	82.6	67.1	65.7	70.3
surprise	74.2	72.2	75.7	37.2	47.0	36.3	49.5	56.9	49.1
六类情绪平均值	69.5	69.4	70.6	64.9	66.4	65.3	67.1	67.9	67.8

注:LCPN 代表每个父节点一个分类器方法,GC 代表全局分类器方法,NB 代表平面型分类方法,即文中使用的朴素贝叶斯方法。

由表 2 可以看出,文中提出的两种方法在 negative 类上的分类效果都优于平面型方法。其中,LCPN 方法的召回率和 f-score 比平面型 NB 方法分别高 26.2%、25.2%。LCPN 方法直接利用根节点分类器将微博情绪判断为 negative 类或 joy 类,根节点分类器实现的是二类分类任务。而平面型 NB 方法将微博情绪一次性分为六类,然后根据除 joy 类外的其它五类负向情绪的分类结果计算 negative 类的分类效果。观察

平面型方法在六类情绪上的分类结果发现,该方法受数据集中各类别数据分布不平衡的影响较大,所以 joy 类和其它五类负向情绪的分类效果差异很大。因此平面型方法根据五类负向情绪的分类结果计算 negative 类的分类效果时,negative 类的分类效果不如文中方法。

LCPN 方法在 Ekman 的六类情绪上召回率的平均值和 f-score 的平均值都高于平面型方法。NB 方法在 joy 类的召回率高于 LCPN 方法和 GC 方法,但在 surprise 类的召回率低于 LCPN 方法和 GC 方法。这是因为在实验数据中,joy 类微博占的比例最大,而 surprise 类微博占的比例最小。当采用平面型分类方法时,分类器将很多微博情绪判断为 joy 类,而将微博数量仅占语料库 9% 的 surprise 类微博情绪判断为了其他五类情绪中的一类。当采用文中提出的两种方法时,弱化了 joy 类微博在数量上的优势,从而在 joy 类上取得了更高的精确率和 f-score,在 surprise 类上取得了更高的召回率和 f-score。相比平面型方法,GC 方法在 anger 类的召回率提高了 2.9%。LCPN 方法和 GC 方法在 disgust 类的召回率都提高了 1.4%。

LCPN 方法明显优于 GC 方法,LCPN 方法的召回率和 f-score 都普遍高于 GC 方法。但在 anger 类的召回率却比 GC 方法低 8.8%。这是因为 GC 方法为层次结构中的所有类别建立一个分类器,这类似于平面型分类方法,只是 GC 方法考虑到层次结构,将父节点算作一种类别,建立的分类器含有整个类别的层次结构信息。GC 方法也会受数据集中各类别数据分布不平衡的影响,在实验数据中 anger 类的微博数量占 anger-disgust 类微博数量的 65%,这导致了在 GC 方法将许多 disgust 类判断为了 anger 类,所以 GC 方法在 anger 类的召回率高于 LCPN 方法。

4 结束语

文中提出了基于朴素贝叶斯模型的多层次微博情绪分析方法。将文中提出的方法与传统的平面型朴素贝叶斯方法对比发现,文中方法降低了各情绪类微博分布不平衡对分类结果造成的影响,提高了微博情感识别的精度。但文中提出的方法并没有考虑类别层次结构中各类别之间的联系,如类别间的相似性。下一步将考虑如何充分利用层次结构中各类别之间的关系来帮助识别微博情绪。

参考文献:

[1] Ekman. Facial expression and emotion[J]. American Psychologist,1993,48(4):384-392.

[2] Purver M, Battersby S. Experimenting with distant supervision for emotion classification[C]//Proceedings of the 13th conference of the European chapter of the association for computational linguistics. [s. l.]: Association for Computational Linguistics,2012:482-491.

[3] 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.

[4] 欧阳纯萍,阳小华,雷龙艳,等. 多策略中文微博细粒度情绪分析研究[J]. 北京大学学报:自然科学版,2014,50(1):67-72.

[5] Paltoglou G, Thelwall M. Twitter, Myspace, Digg: unsupervised sentiment analysis in social media[J]. ACM Transactions on Intelligent Systems & Technology,2012,3(4):67-83.

[6] 张 晶,朱 波,梁琳琳,等. 基于情绪因子的中文微博情绪识别与分类[J]. 北京大学学报:自然科学版,2014,50(1):79-84.

[7] 牛 耘,潘明慧,魏 欧,等. 基于词典的中文微博情绪识别[J]. 计算机科学,2014,41(9):253-258.

[8] Jiang L, Yu M, Zhou M, et al. Target-dependent twitter sentiment classification[C]//Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies - volume 1. [s. l.]: Association for Computational Linguistics,2011:151-160.

[9] Jiang F, Cui A, Liu Y, et al. Every term has sentiment: learning from emoticon evidences for chinese microblog sentiment analysis[M]//Natural Language Processing and Chinese Computing. Berlin: Springer,2013:224-235.

[10] 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.

[11] Keshtkar F, Inkpen D. A hierarchical approach to mood classification in blogs[J]. Natural Language Engineering,2012,18(1):61-81.

[12] 保罗·艾克曼. 情绪的解析[M]. 海口:南海出版公司,2008.

[13] Alm E C O. Affect in text and speech[M]. [s. l.]: ProQuest,2008.

[14] Silla Jr C N, Freitas A A. A survey of hierarchical classification across different application domains[J]. Data Mining and Knowledge Discovery,2011,22(1-2):31-72.

多层次中文微博情绪分析

作者： 刘宝芹， 牛耘， LIU Bao-qin， NIU Yun
作者单位： 南京航空航天大学 计算机科学与技术学院,江苏 南京,210016
刊名： 计算机技术与发展 ISTIC
英文刊名： Computer Technology and Development
年，卷(期)： 2015,25(11)