基于表情图片与情感词的中文微博情感分析

张 珊 于留宝 胡长军
(北京科技大学计算机与通信工程学院 北京 100083)

摘 要 微博是 Web 2.0 时代新生的社会化媒体平台,网民通过微博抒发自己的情感,表达自己的喜怒哀乐与爱恶,从而产生了海量的情感文本信息。通过对情感信息的分析,可以得到网民的情绪状况、对某个社会现象的观点、某个产品的喜好等信息,其不仅有一定的商业价值,还对社会的稳定有所帮助。利用微博中的表情图片,并结合情感词语的方法来构建中文微博情感语料库,既保证了语料库的规模与准确性,又省去了人工的负担;在情感语料库的基础上,构建贝叶斯分类器;最后利用熵的概念对语料库进行优化,提高了分类的准确性,并比较了使用不同 n-gram 特征项的性能。最终发现,使用 UniGram 特征项并用熵进行优化之后,分类的效果最好,召回率和准确率都可以达到 85% 以上,F 值甚至可以达到 89% 以上。

关键词 情感分析,表情图片,情感词,微博
中图法分类号 TP391 **文献标识码** A

Sentiment Analysis of Chinese Micro-blogs Based on Emoticons and Emotional Words

ZHANG Shan YU Liu-bao HU Chang-jun
(School of Computer and Communication Engineering, University of Science and Technology Beijing, Beijing 100083, China)

Abstract Micro-blog is a new social media platform based on Web 2.0. Internet users express their feelings, emotions, favorites and disgust through micro-blogs, resulting in a large number of emotional text information. We can know the emotional state of the Internet users, the point of a social phenomenon and preference of a product, through analysis of the emotional text information, which not only has a certain kind of commercial value, and is helpful to the stability of the society. In this paper, we use the emoticons form micro-blogs, combined with emotional words to build the Chinese emotional corpus, ensuring the scale and accuracy of the corpus, eliminating the need for artificial burden. Based on the corpus, we construct Bayes classifier and use the entropy to improve the performance. We compare different performance while changing the type of n-gram. Finally, we get the best classification results using unigrams as features and optimizing with entropy. Recall rate and accuracy can be achieved above 85%, the F measure can even reach more than 89%.

Keywords Sentiment analysis, Emoticons, Emotional words, Micro-blog

1 引言

微博是 Web 2.0 时代新生的社会化媒体平台,是一个基于用户关系的信息分享、传播以及获取的平台。用户可通过 Web、WAP 以及各种客户端,以不超过 140 个字的文字更新信息,实现即时分享。微博变革了传统的信息传播模式,开创了新的社区互动模式,同时又是一个自发传播的平台。微博的使用人群数量极大,截至 2012 年 7 月 1 日,Twitter 的用户数已经达到 5.17 亿。在国内,最新开发微博平台的新浪微博截止到 2012 年 5 月 15 日,其用户已达到 3.24 亿,日活跃用户比例超过 9%。在微博平台中,信息以前所未有的速度增长,并快速传播,这些信息涉及生活、社会热点、人际交往等各种信息。网民通过微博抒发自己的情感,表达自己的喜怒哀乐与爱恶,形成了海量的情感文本信息。这些情感文本信息是非常宝贵的资源,通过情感分析技术来分析这些情感文本,得出文本的情感倾向性,可以得到网民的情绪状况、对某个社会现象的观点、某个产品的喜好等信息,其不仅有一定的商业价值,还对社会的稳定有所帮助。

国内外学者已经在情感分析领域做了大量的研究,并取得了很多成果。目前的研究工作可以分为两种思路:基于情感知识的方法与基于特征分类的方法。基于情感知识的方法主要是建立情感词典或领域情感词库,利用主观文本中情感词语或词语的组合来判断文本的极性;基于特征分类的方法主要是使用机器学习的方法,将情感分析看作传统的分类,抽取特征并进行判断。文献[1-3]利用情感词库中的词作为种子词语,来确定与种子词语有一定语法关系(解释性关系、同义词关系、反义词关系)的词语的极性,然后进行极性的加权求和。这类方法通常的重点是情感评价词语或其组合的极性判断以及极性求和的方法。文献[4]首次将机器学习的方法用在情感分类中,面向的是篇章级的文本,使用了 n-gram 词

本文受国家十二五科技支撑计划课题"面向群体的网络热点传播分析与监测技术研究"(2011BAK08B04)资助。
张 珊(1987—),女,硕士;于留宝(1987—),男,硕士;胡长军 教授,E-mail:huchangjun@gmail.com。

语特征与词性特征，同时对比了 SVM(Support Vector Machine)、NB、和 ME 3 种分类模型，最后发现 Unigram 特征效果最好。文献[5]在特征的提取中进行了研究，提出了一种将多种语法特征(词性、词组内部构成模式、词语上下文语境)集成的框架。文献[6]在分类方法方面进行了探索，提出了一种组合的方法，即通过将多种分类器进行混合叠加来提高情感分类的性能。

虽然微博是新兴的网络媒体，但由于其对社会的影响巨大，也吸引了很多学者进行研究。文献[7]使用 twitter 中的表情符号来区分文本的情感极性，从而实现了自动建立情感语料库，减少了人工的标注。文献[8]提出了利用距离监督的方法，对 twitter 中的文本进行情感分类。文献[9]使用了词性与词语两种特征对 twitter 中的文本进行分类，并研究了提高分类准确性的优化方法。

在国内，对微博情感分析的研究才刚起步，针对微博的语料库很少。本文利用微博中的表情图片，并结合情感词语的方法来构建中文微博情感语料库，既保证了语料库的规模与准确性，又省去了人工的负担；在情感语料库的基础上，使用机器学习的方法构建贝叶斯分类器；最后利用熵的概念对语料库进行优化，提高了分类的准确性，并比较了不同 n-gram 类型的性能。最终发现，使用 UniGram 特征项并用熵进行优化之后，其分类的效果最好，召回率和准确率都可以达到 85%以上，F 值甚至可以达到 89%以上。本文第 2 节介绍基于情感图片与情感词的语料库收集与标注；第 3 节介绍微博情感分类器与性能的优化；第 4 节展示实验数据；最后进行总结，并提出下一步的工作设想。

2 微博语料库的收集与标注

2.1 语料库的收集

微博中常常含有一定量的表情图片，表情图片代表了微博作者的情感，表示此微博是主观的，因此，利用表情图片便可很容易地得到主观微博。本文的语料资源来自新浪微博，利用新浪微博提供的 API，可以很容易地得到新浪微博的实时数据，从中筛选带有表情的微博作为语料库的候选。主观微博一般来自于微博个体用户，微博的作者有着不同背景、工作、性别等，并来自不同的地区。客观微博来自于媒体的官方微博，例如新浪财经、新浪体育、新京报等。

2.2 语料库的标注

文献[7,9]都利用了 twitter 中的表情对语料库进行标注。积极表情有：":-)"，":)"，"=)"，":D"等；消极表情有：":-("，":("，"=("，":("。然而 twitter 中限制的是 140 个英文字母，一条英文微博基本上就是一两句话，因此，可以用表情直接表示整条微博的情感。对于中文来说，一条微博的字数限制是 140 个汉字，比 140 个英文字母表示的内容要多很多，如表 1 中的比较。

表 1 中英文微博对比

英文微博：120	中文微博：134
Hi my name is Kendyl ！i love music and dancing singing u know that kind of stuff i also Love Everyone !!^_^ also im 14.	近日，湖南卫视@平民英雄节目，在武汉一辆583路双层公交车上，分别让演员扮演小偷与失主。观察乘客在目睹小偷行窃过程会做何种反应。61位被"试验"乘客中，35位用不同方式提醒被偷者，9位挺身而出直接指出小偷，武汉人不冷漠！http://t.cn/zO8IW5E [怒]。

可见，中文微博虽然有 140 个字的限制，但明显比英文 140 个字母所表达的内容多很多。而本条中文微博中虽然有表情符号"[怒]"(新浪微博中的表情符号字符化格式)，但是由于有"武汉人不冷漠！"这句话，其整体情感应该是积极的，因此，中文微博情感语料库标注不能单纯依赖于表情符号。

文献[10]利用情绪词语来进行中文微博语料库标注，然而情绪词没有表情图片反映的情绪强，并且在对含有否定词的微博语料进行处理时，会存在误差。于是提出了结合情感图片与情感词的语料库标注方法。其步骤如下：

(1)利用微博中的情感图片，将微博分成积极和消极两类。由于可能会出现两种情感图片共存的现象，这可能是由微博前后的情感变化导致的，因此语料库要筛选两种表情图片只存其一的微博。本文筛选了一些代表性的积极与消极表情，如图 1 所示。

图 1 情感表情图片

(2)对分好类的语料，再利用情感词进行判定。此时要注意否定词，比如上文中的"武汉人不冷漠！"中的否定词"不"。如果出现否定，便将情绪翻转。如果利用情感词判定得出的情感与利用情感图片得出的情感不相同，则将此条冲突语料删除。

3 情感分类器

3.1 特征抽取

在表达情感的文本中，某个关键词出现一次或者多次，所表达的情感是一致的，因此使用布尔权值来进行文本表示时，出现为 1，不出现为 0。本文使用 n-gram 进行特征抽取，并分别采用 Unigram、Bigram 和 Trigram 进行实验，并比较性能。

在训练分类器之前，必须对语料库中的微博文本进行预处理，其步骤如下：

(1)删除 URL 链接、用户名(@XXX，@后面跟的是用户 ID)、表情图片(新浪微博中的表情图片表现在文本中是以"[XX]"格式表示)。

(2)分词，本文使用中科院 ICTCLAS 软件进行分词与词性标注。

(3)去除停用词，包括代词、助词等，如"的"、"他/她"之类。

完成上述的处理之后，一条微博文本就会表示成一串词

语的组合，比如“我国游泳小将叶诗文继 400 米混合泳夺冠后，又在 200 米混合泳比赛中勇夺冠军，兴奋剂的质疑不攻自破。祝贺她又打破奥运会纪录！”就会变为{我国，游泳，小将，叶，诗文，继，400，米，混合泳，夺冠，后，又，在，200，米，混合泳，比赛，中，勇夺，冠军，兴奋剂，质疑，不攻自破，祝贺，又，打破，奥运会，纪录}。在此基础上就可以构建 n-gram。

3.2 分类器

本文使用了朴素贝叶斯(NB)分类器，下面介绍其原理。

朴素贝叶斯(NB)分类器的依据是贝叶斯定理。

$$P(q|W)=\frac{P(W|q)P(q)}{P(W)}$$

式中，q 表示情感(积极、消极、客观)，W 表示一条微博。对于不同的情感来说，$P(W)$ 是相同的，而在本文的训练语料库中，3 种情感的微博数量相等，因此 $P(q|W)$ 的大小就只取决于 $P(W|q)$。于是，得出下面的公式。

$$P(q|W)\sim P(W|q)$$

而一条微博最终使用 n-gram 特征项表示，即 $W=\{w_1,w_2,\cdots,w_n\}$，于是便得出最终的公式。

$$P(q|W)\sim\sum_{k=1}^{n}\log P(w_k|q)$$

3.3 性能优化

在分类任务中，特征项的选择对分类的准确性起着至关重要的作用。为了提高分类的性能，要选择那些具有较高类区分度的特征，在本文中便是要选择在某个情感类中出现概率很高的 n-gram，而删除那些在所有类中出现概率都差不多的 n-gram。计算特征项 n-gram 在不同情感类中出现概率的熵，即

$$Entropy(q)=-\sum_{i=1}^{n}p(q_i)\log p(q_i)$$

式中，$p(q_i)$ 为 n-gram 特征项在某个情感语料出现的概率，本文共有 3 类，即为 3。由于 $p(q_i)$ 之间越接近，熵越大，因此，若 n-gram 特征项在 3 类语料库中分布得越平均，熵也越大。实际部分特征项的熵大小如表 2 所列。只需要选取熵小于某个值 θ 的特征项。于是贝叶斯分类器计算公式变为

$$P(q|W)\sim\sum_{k=1}^{n}\log P(w_k|q)f(w_k)$$

$$f(w_k)=\begin{cases}1, & \text{Entropy}(w_k)\leqslant\theta\\ 0, & \text{Entropy}(w_k)>\theta\end{cases}$$

表 2 部分特征项的熵

熵比较小的特征项		熵比较大的特征项	
恶心	0.201881	有	1.584342
讨厌	0.243612	学	1.584367
恭喜	0.342979	人生	1.582105
哈哈	0.361509	会	1.582337
变态	0.371194	多	1.578883
气愤	0.39979	最	1.578254

4 实验数据

采用本文基于情感图片与情感词的方法，获得了积极、消极、客观语料库各 5000 条，人工标注的测试数据如表 3 所列。

表 3 测试数据

积极	消极	客观	总量
998	1005	989	2992

本文采用的评价指标有召回率 R、准确率 P 和 F 值，使用 UniGram、BiGram 和 TriGram 3 种特征项。

$$\text{召回率 } R=\frac{\text{系统检索到的相关文件数}}{\text{系统所有相关的文件总数}}$$

$$\text{准确率 } P=\frac{\text{系统检索到的相关文件数}}{\text{系统所有检索到的文件总数}}$$

$$F\text{ 值}=\frac{2PR}{P+R}$$

使用熵优化前、后的结果如表 4、表 5 所列，熵优化之后的 F 值整体情况柱状图如图 2 所示。

表 4 优化前的测试结果

	召回率	准确率	F 值
积极—UniGram	0.833667	0.968568	0.896069
消极—UniGram	0.927363	0.884250	0.905294
客观—UniGram	0.966633	0.886006	0.924565
积极—BiGram	0.766533	0.897887	0.827027
消极—BiGram	0.820896	0.825826	0.823353
客观—BiGram	0.939333	0.814198	0.872300
积极—TriGram	0.309619	0.804688	0.447178
消极—TriGram	0.433831	0.768959	0.554707
客观—TriGram	0.939333	0.455169	0.613201

表 5 优化后的测试结果

	召回率	准确率	F 值
积极—UniGram	0.855711	0.969353	0.908994
消极—UniGram	0.919403	0.914851	0.917122
客观—UniGram	0.976744	0.877384	0.924402
积极—BiGram	0.775551	0.896871	0.831811
消极—BiGram	0.811940	0.835210	0.823411
客观—BiGram	0.942366	0.809028	0.870621
积极—TriGram	0.301603	0.811321	0.439737
消极—TriGram	0.430846	0.775986	0.554063
客观—TriGram	0.942366	0.451769	0.610747

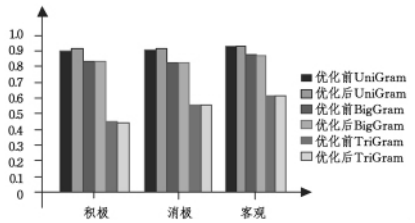

图 2 优化前后 F 值对比

由上述数据可以看出，UniGram 作为特征项的分类效果最好，其 F 值可以达到 89% 以上，同时准确率和召回率也都很高；使用熵优化之后，UniGram 特征项的效果的提高比 BiGram 和 TriGram 要明显，对于 TriGram 特征项来说反而下降。总体来看，使用 UniGram 特征项并进行熵优化后的分类效果最好。

结束语 微博在网络舆论中扮演着越来越重要的角色，网民通过微博平台抒发自己的情感，同时也影响着整个社会。微博的情感分析最重要的是语料库的建立和特征的选择抽取，本文提出了表情图片与情感词相结合的语料库标注方法，提高了语料库标注的速度与准确性，在此基础上训练贝叶斯分类器，并利用熵进行筛选特征项，提高了分类器的性能，使得情感分类结果的 F 值达到 90% 以上，准确率和召回率都达到 85% 以上。

中文微博的情感分析起步比较晚，还有很多研究需要进一步开展。下一步将针对情感的强烈程度展开研究，以探索一种较好的判断情感程度的方法。

(下转第 176 页)

环，直到每一个 $P(t)$ 都为空为止。

3.4 算法应用

应用以上算法对表 1 中的 $t1$ 和 $t2$ 进行清洗，过程如下：

(1)$\Sigma=\{\phi1,\phi2,\phi3\}$；

(2)$P[t1]=\{(\phi1,s1),(\phi2,s1)\}$，$P[t2]=\{(\phi3,s1)\}$；

(3)将$(\phi1,s1)$应用在 $t1$ 中，得到 $t1[AC,str]=[131,51$ Elm Row]，将$(\phi2,s1)$应用在 $t1$ 中，得到 $t1[FN]=$ Robert

(4)$Q[t1]=\{(\phi1,s1),(\phi2,s1)$，$P[t1]=\emptyset$

(5)将$(\phi3,s1)$应用在 $t2$ 中，得到 $t2[str,city,zip]=[$ 51 Elm Row，Edi，EH7 4AH]

(6) $Q[t2]=\{(\phi3,s1)\}$

(7)更新后得到 $P[t2]=\{(\phi1,s1)\}$

(8)将$(\phi3,s1)$应用在 $t2$ 中，没有任何更新。

(9)$P[t2]=\emptyset$，$Q[t2]=\{(\phi3,s1),(\phi1,s1)\}$

(10)最终 $P[t2]$和 $P[t1]$都为空，修复结束。同理可得 $t3$，$t4$ 的修复。

虽然算法可以准确地修复数据，但是算法并不能完全修复所有数据，因为一些元组找不到与其相适用的 Editing Rule，即前文所定义的 Region 无法完全覆盖所有规则，所以算法可以在如何定义新的 Region 方面进行改进，此外还需根据条件函数依赖构造另外一种算法来修复数据，参考文献[9]，这种算法得到的数据修复无法保证数据的精确性。

结束语　本文提出了一种全新的概念，即应用 Editing rule 并结合 Master Data 来更新元组，修复数据，而不是简单地依据传统的完整性约束，框架可以得到唯一的、确定性的修复而且不会引入新的错误。以前提出的有些算法是基于约束规则的，例如 CFD 和 MD，虽然有些可以得到确定性的修复，但是算法基本上存在一定的前提条件，例如文献[6]中的算法需要在每一个属性上放置一个标识可信度的数值，根据这个数值来进行修复，但是在所有属性上放置这样的数值显然是不太现实的。本文算法考虑到这种情况，将 Eidting rule 和 Master Data 相结合来清洗数据，因此大大提高了效率和精确性。但是本文尚有纰漏，算法不可能完全修复存在的“脏”数据，因为定义的 Editing Rules 无法完全覆盖所有的元组，总会存在一些元组找不到与之对应的修复法则，因此算法有待改进。

参 考 文 献

[1] Redman T. The impact of poor data quality on the typical enterprise[J]. Communication，Communication. ACM，1998，2：79-82

[2] Eckerson W W. Data Quality and the Bottom Line：Achieving Business Success through a Commitment to HighQuality Data [Z]. The Data Warehousing Institute，2002

[3] Gartner. Forecast：Data quality tools，worldwide，2006-2011. Technical report[Z]. Gartner，2007

[4] Herzog T N，Scheuren F J，Winkler W E. Data Quality and Record Linkage Techniques[C]// Springer. 2009

[5] Arenas M，Bertossi L E，Chomicki J. Consistent query answers in inconsistent databases[J]. TPLP，2003，3(4/5)

[6] Fan W，Geerts F，Jia X，et al. Conditional functional dependencies for capturing data inconsistencies[J]. TODS. 2008，33(1)

[7] Fan W，Jia X，Li J，et al. Reasoning about record matching rules [C]// VLDB. 2009

[8] Chiang F，Miller R. Discovering data quality rules[C]// VLDB. 2008

[9] Cong G，Fan W，Geerts F，et al. Improving data quality：Consistency and accuracy[C]// VLDB. 2007

[10] Fan W，Tang N，Li J，et al. Towards Certain Fixes with Editing Rules and Master Data[C]// VLDB. 2011

(上接第 148 页)

参 考 文 献

[1] Andreevskaia A，Bergler S. Mining WordNet for a fuzzy sentiment：Sentiment tag extraction from WordNet glosses[C]// Proceedings of the European Chapter of the Association for Computational Linguistics (EACL). 2006

[2] Kanayama H，Nasukawa T. Fully automatic lexicon expansion for domain-oriented sentiment analysis[C]// Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics，Sydney，Australia，July 2006：355-363

[3] Turney P. Thumbs up or thumbs down？Semantic orientation applied to unsupervised classification of reviews[C]// Proceedings of the Association for Computational Linguistics (ACL). 2002：417-424

[4] Pang Bo，Lee L，Vaithyanathan S. Thumbs up？Sentiment Classification Using Machine Learning Techniques[C]// Proc. of Conference on Empirical Methods in Natural Language Processing. [S. l.]：ACM Press，2002

[5] Wei Jin，Ho H H，SrihariR K. OpinionMiner：A novel machine learning system for web opinion mining and extraction[C]// Proceedings of the 15th ACM SIGKDD

[6] Prabowo R，Thelwall M. Sentiment analysis：A combined approach[J]. Journal of Informetrics，2009，3(2)：143-157

[7] Read J. Using emoticons to reduce dependency in machine learning techniques for sentiment classification[C]// Proceedings of ACL-05，43nd Meeting of the Association for Computational Linguistics. Association for Computational Linguistics，2005

[8] Go A，Huang L，Bhayani R. Twitter sentiment classification using distant supervision[C]// CS224N Project Report. Stanford，2009

[9] Pak A，Paroubek P. Twitter as a corpus for sentiment analysis and opinion mining[C]// Proceedings of the Seventh Conference on International Language Resources and Evaluation. 2010：1320-1326

[10] 庞磊，李寿山，周国栋. 基于情绪知识的中文微博情感分类方法[J]. 计算机工程，2012，38(13)：156-158