·人工智能及识别技术· 文章编号：1000—3428(2012)13—0156—03 文献标识码：A 中图分类号：TP391

基于情绪知识的中文微博情感分类方法

庞 磊[1,2]，李寿山[1,2]，周国栋[1,2]

(1. 苏州大学计算机科学与技术学院，江苏 苏州 215006；2. 江苏省计算机信息处理技术重点实验室，江苏 苏州 215006)

摘 要：通过对新浪微博文本进行情感信息方面的分析与研究，提出一种基于情绪知识的非监督情感分类方法。利用情绪词和表情图片2种情绪知识对大规模微博非标注语料进行筛选并自动标注，用自动标注好的语料作为训练集构建微博情感文本分类器，对微博文本进行情感极性自动分类。实验结果表明，该方法对微博文本的情感极性分类达到较好的效果。

关键词：中文信息处理；无监督学习；情绪知识；微博；情感分类

Sentiment Classification Method of Chinese Micro-blog Based on Emotional Knowledge

PANG Lei[1,2], LI Shou-shan[1,2], ZHOU Guo-dong[1,2]

(1. School of Computer Science and Technology, Soochow University, Suzhou 215006, China;
2. Provincial Key Lab of Computer Information Processing Technology of Jiangsu, Suzhou 215006, China)

【Abstract】 This paper proposes an unsupervised method of sentiment classification and applies it to perform sentiment classification on Sina micro-blog. The approach employs emotional images and emotional words as the emotional knowledge to extract pseudo-labeled samples, and uses them to train a classifier for automatically classification on polarities of the miro-blog. Experimental results show that the method achieves a decent performance on sentiment classification for Chinese micro-blog.

【Key words】 Chinese information processing; unsupervised learning; emotional knowledge; micro-blog; sentiment classification

DOI: 10.3969/j.issn.1000-3428.2012.13.046

1 概述

微博是Web2.0时代兴起的一种集成化、开放化的互联网社交服务。它打通了移动通信网和互联网的界限，用户可以通过手机、IM软件和外部API接口等途径，即时向外发布140字以内的文本，越来越受到互联网用户的青睐。统计显示，截止到2011年5月底，仅在Twitter网上的微博注册用户就已达3亿。在发展相对较晚的中国，微博也呈爆炸性增长，从2009年8月新浪微博开始发布到2011年4月，仅20个月的时间，新浪微博注册用户便达到1.42亿，用户每天要发布近5 000万条微博内容。规模庞大的微博文本的自动处理给自然语言处理研究提出了新的挑战和机遇。在这些海量的文本信息中，有很大一部分是带有情感的文本信息。这些情感文本信息是非常宝贵的意见资源，可以利用这些文本信息进行情感文本分类研究。

本文通过对中文微博的研究与分析，提出一种利用情绪知识实现微博上非监督的情感文本分类方法，通过情绪知识从大规模未标注语料中自动获取伪标注语料，然后利用这些伪标注语料作为训练集训练分类器，实现微博文本情感的自动分类。根据微博的特点，利用2种情绪知识(即情绪词和表情图片)自动标注语料。通过该方法可以很容易地获得训练集，从而省去了人工标注的过程。

本文利用情绪词作为情绪知识自动标注语料主要考虑到以下3个原因：

(1)情绪是人在受到外界事物的影响后发出的，如果在一篇微博中同时出现了情绪词和与某一话题相关的评论，那么，该评论的情感倾向性应与作者的情绪倾向性保持一致。所以，可以通过作者的情绪好坏来判断其发表评论的极性。

(2)情绪的表达具有一般性，没有领域界限，在任何话题中，都可以通过作者情绪来判断评论的极性，而且情绪词都是固定的。

(3)一般来讲，微博评论是作者情绪产生的原因，它不会影响作者对评价对象发表评论的极性，这一点和表情图片类似，可以将它看成微博评论的极性标识。

2 相关工作

2.1 情感分类方法

情感分类是按照文本表达的情感倾向性对文本进行分类[1]。例如，判断文本对某个事物的评论是“好”还是“坏”。情感分类的研究历史虽然不长，但是已成为自然语言处理方向里面的一个研究热点，近年来文本情感分类技术已比较成熟。目前，情感分类的研究方法主要可以分为2种研究思路：非监督的分类方法和监督的分类方法。前者主要依靠一些已有的情感知识(情感词典或领域词典)以及一些规则获取情感文本的极性，例如，文献[2-4]首先分析文本中评价词语的极性，然后进行极性加权求和。这种方法的重点一般都放在评价词语的抽取和极性判断方法的研究上；后者主要是使用机

基金项目：国家自然科学基金资助项目(61003155, 60873150)

作者简介：庞 磊(1985－)，男，硕士研究生，主研方向：自然语言处理；李寿山，副教授；周国栋，教授、博士生导师

收稿日期：2011-08-31 **E-mail**：panglei.nlp@gmail.com

器学习的方法，选取大量有意义的特征来完成分类任务。在监督的分类方法中有很多代表性的研究工作，这类方法一般都是基于特征的。文献[1]首次将机器学习的方法应用于篇章级的情感分类任务中，并指出这种方法比非监督的分类方法在分类性能上有明显的优势。文献[5-7]在有效特征的发现以及特征选择和特征融合等方面做了相应研究。文献[8]在分类器的选择上和分类器融合等方面做了相应研究。近年来，中文情感文本分类方面的研究也得到了迅速的发展，例如，文献[8-10]基于特征的监督分类方法是目前主流的情感分类方法。

2.2 微博情感分类方法

微博是一种新兴的社交网络服务，目前针对微博文本的情感分类研究还相对较少。文献[11]提出了利用距离监督的方法对 Twitter 上的微博进行情感分类，他们的主要贡献在于利用了 Twitter 上的表情符号作为标签收集英文语料作为训练集，省去了人工标注语料的过程。文献[12]提出对英文微博语料进行收集与自动标注，进行情感分析与意见挖掘的研究。文献[13]提出在微博上通过加入评价对象相关的特征来提高情感分类的效果。文献[14]利用 Twitter 上的标签和笑脸表情符对 Twitter 上的微博语料进行强化学习研究。

不同于以上文献的工作，本文采用表情图片，并收集使用了情绪关键词对微博语料进行收集与自动标注，使得标注样本在规模和性能上都有进一步的提升。本文是首次对中文微博进行情感分类方法研究。

3 中文微博语料收集与标注

3.1 语料收集

从新浪微博上收集了 2 个话题(电影、手机)的语料。然后以情绪词和表情图片作为情绪知识来过滤未标注样本，得到只含有情绪词或表情图片的样本，通过情绪词以及表情图片所反映的用户情绪(积极与消极)对含有这些信息的文本进行自动标注。

3.2 语料自动标注方法

文献[11]通过收集含有表情符号的微博，以这些表情符号作为微博情感倾向的标识。这样就省去了人工标注的环节。例如，将含有:)、:-)的样本标注为正例样本；将含有:(、:-(的样本标注为负例样本。与文献[11]不同的是，本文利用 2 种情绪知识(表情图片和情绪词)对大规模未标注样本进行自动标注，以获取训练样本。

3.2.1 表情图片自动标注方法

在中文微博中经常出现一些表情图片，这些表情图片要比表情符号反映的情绪更为明确。选取了正面表情与负面表情各 8 种情绪倾向比较明确的表情图片，如图 1 所示。

图 1 表情图片

将微博中含有正面表情图片的文本归为正面评论，含有负面表情图片的文本归为负面评论。

3.2.2 情绪词自动标注方法

由于仅以表情图片作为情绪知识自动标注的样本数量相对较少，用来训练一个好的分类器还远远不够。于是，提出一种结合情绪词作为情绪知识对未标注样本进行自动标注的方法。

情绪是相对于人来说的，它所反映的是一个人的心情好坏，可以从一个人谈论某个话题时所流露出的情绪，来判断其在该话题上的评论是正面的还是负面的，例如，“×××这电影看得，太郁闷了，死了半天都没死过去，导演差，编剧烂，从没看过这么差的电影。”在这条微博中，根据“郁闷”这个词将其自动标为负面评论。在该例中，“郁闷”是人的情绪，它并不是微博作者对电影的评价，但是，和表情图片类似，它可以作为评论的情感倾向标识。

构建一个情绪词表库，分为正面情绪词表和负面情绪词表，其中，正面情绪词有 50 个，负面情绪词有 67 个。表 1 给出了部分情绪词。

表 1 情绪词

正面情绪词	负面情绪词
开心	悲催
快乐	痛心
兴奋	郁闷
…	…
放松	心寒
感动	懊悔

利用情绪词作为情绪知识进行自动标注，共分为以下 3 步来完成：

(1)将含有情绪词的文本粗分为 2 类，含有正负情绪词的评论归为正面评论，含有负面情绪词的文本归为负面评论。

(2)对含有否定词的文本进行处理，收集了 17 个否定词：避免，并不是，不，不会，没有，不可能，很难，不太，减少，没，不再，没能，一改，怎么能，怎么会，不可以，很少。

针对第(1)步的粗分结果，将情绪词前面带有否定词的文本放入到相反的类别中去，例如，在“今天去看了×××，×××果然没有让我失望。”这句话中，“失望”前面有否定词“没有”，那么第(1)步粗分的结果与实际情况相反，所以要将情绪词前带有否定形式的文本放入粗分结果相反的类别中去。

(3)针对第(2)步处理结果，将在文本中同时含有 2 种冲突情绪的文本删除。在处理这些有情绪冲突的语料时，并没有进一步判断这些语料到底是什么极性，为了保证自动标注语料的精确性，直接将有冲突情绪的语料删除，例如，在“从电影开场，紧张的神经就一直没有放松过，团队合作，领导力，坚持永不撤退的信念，这道主旋律一直贯穿始终，让人感动。×××，值得一看。”这句话中，“放松”是正面情绪词，通过第(1)步处理将它归入正面评论类；在第(2)步处理时，因为“放松”前有否定词“没有”，所以要将它放入负面评论类；最后，又在这条论评里面发现了正面情绪词“感动”，所以，在这条评论中情绪有冲突的现象，要将这条语料删除。

通过以上 3 步处理，就能获取比仅用表情图片作为情绪知识自动标注语料规模大的标注语料。

4 实验结果及分析

4.1 自动标注样本的结果分析

表 2 给出在手机和电影这 2 个话题的微博(各 10 000 条)中，分别以表情图片和情绪词作为情绪知识自动标注语料的规模。

表 2 自动标注语料数量

话题	表情图片		情绪词	
	正例	负例	正例	负例
电影	682	148	1 303	405
手机	418	132	693	378

通过人工校对发现，这些自动标注的语料都达到了很高的准确率，表 3 给出了自动标注语料的准确率。

表 3 自动标注语料准确率 (%)

话题	表情图片	情绪词
电影	89.2	86.3
手机	87.1	83.2

从表 3、表 4 可以看出，用表情图片自动标注的语料在准确率上要比用情绪词自动标注的语料高，这是因为情绪词相对于表情图片来说，反映微博作者的情绪要弱一些，而且在用情绪词自动标注语料时，要对含有否定词的语料进行处理，会有误差存在。例如，在“今天看了×××，3D 的画面很美，音乐也很给力，好久没有这样开心过了!”这句话中，“开心”前面有否定词“没有”，仅根据情绪词自动标注的规则就会将它错误标注为负面评论。但是，发现利用情绪词自动标注的语料比利用表情图片自动标注的语料规模上大很多，所以这也是结合表情图片和情绪词作为情绪知识来自动标注语料的原因。

4.2 基于自动标注样本的分类器性能分析

在获得标注样本后，使用机器学习的方法构建分类器，分类任务是将评论分为正面和负面。本实验使用了从新浪微博上收集并自动标注的 2 个话题(手机和电影)的语料作为训练集。从表 3 可以看出，自动标注的语料在 2 个话题内是不平衡的，这里仅简单采用随机抽样的方法使训练集达到平衡，选取了这 2 个话题的训练集样本正负例各 500 个(其中，用表情图片自动标注的样本 130 个，用情绪词自动标注的样本 370 个)。测试集是手工标注的，其中 2 个话题中正负例各 150 个。本文采用分类正确率评价分类的效果，其定义如下：

$$Acc=\frac{\text{正确分类文本数}}{\text{测试集中文本总数}}\times 100\%$$

本实验使用了 3 种分类方法作为比较，分别是支持向量机(SVM)、朴素贝叶斯(NB)和最大熵(ME)。其中，SVM 使用的是标准工具 light-SVM，NB 和 ME 使用的是 Mallet 机器学习工具包。在使用这些工具的时候，所有参数都设置为它们的默认值。

在分类之前，首先采用中国科学院计算机研究所的分词软件 ICTCLAS 对中文文本进行分词操作，然后选取词的 Unigram 和 Bigram 作为特征进行实验。

表 4、表 5 分别给出只用表情图片自动标注语料作为训练集(正、负例各 130 个)和只用情绪词自动标注语料作为训练集(正、负例各 370 个)在 2 个话题上的分类结果。从表 4、表 5 可以看出，用这 2 种情绪知识自动标注的语料作为训练集得到的 3 种分类器在分类性能上都得到了比较好的效果。

表 4 以表情图片自动标注语料作为训练集的分类正确率 (%)

话题	SVM 分类正确率	ME 分类正确率	NB 分类正确率
电影	73.1	73.8	74.6
手机	70.7	72.4	72.8

表 5 以情绪词自动标注语料作为训练集的分类正确率 (%)

话题	SVM 分类正确率	ME 分类正确率	NB 分类正确率
电影	79.8	79.4	80.3
手机	77.0	77.2	76.8

表 6 给出了将表情图片自动标注语料和情绪词自动标注语料合并作为训练集在 2 个话题上的分类结果。与表 4 及表 5 的分类结果进行对比，可以发现将 2 组训练集合并时，分类效果能够有更进一步的提高，正确率都超过了 80%。从表 4~表 6 可以看出，在这 3 种分类方法中，NB 在分类性能上要略好于其他 2 种分类方法。

表 6 以情绪词与表情图片自动标注语料的分类正确率 (%)

话题	SVM 分类正确率	ME 分类正确率	NB 分类正确率
电影	83.6	83.3	84.2
手机	79.2	79.8	80.3

5 结束语

面向微博的情感分析研究迫切需要解决的问题就是微博文本的情感语料收集和标注问题。本文提出了一种基于情绪词和表情图片作为情绪知识的语料自动标注方法，能够在微博上获取一定规模的自动标注样本。实验结果表明，本文的方法能够收集高精度的自动标注样本，并发现使用这些样本训练的分类器能够在微博文本的情感分类中取得非常不错的效果，正确率超过了 80%。

微博的情感分析研究刚刚起步，还有许多相关的研究需要进一步开展。例如，本文仅对话题评论的情感极性分类做相关研究，对于主客观分类的研究还有待进一步探讨。下一步工作将针对微博的主客观分类展开研究，寻找一种较好的方法，自动获得主客观分类研究中所用到的训练语料。

参考文献

[1] 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.

[2] Kim Soo-Min, Hovy E. Automatic Detection of Opinion Bearing Words and Sentences[C]//Proc. of International Joint Conference on Natural Language Processing. Jeju Island, Korea: [s. n.], 2005.

[3] Yu Hong, Hatzivassiloglou V. Towards Answering Opinion Questions: Separating Facts from Opinions and Identifying the Polarity of Opinion Sentences[C]//Proc. of Conference on Empirical Methods in Natural Language Processing. Sapporo, Japan: [s. n.], 2003.

[4] Hu Minqing, Liu Bing. Mining and Summarizing Customer Reviews[C]//Proc. of Conference on Knowledge Discovery and Data. [S. l.]: ACM Press, 2004.

[5] Cui Hang, Mittal V, Datar M. Comparative Experiments on Sentiment Classification for Online Product Reviews[C]//Proc. of the 21st National Conference on Artificial Intelligence. [S. l.]: ACM Press, 2006.

[6] Kim Soo-Min, Hovy E. Automatic Identification of Pro and Con Reasons in Online Reviews[C]//Proc. of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics. Sydney, Australia: [s. n.], 2006.

[7] Zhao Jun, Liu Kang, Wang Gen. Adding Redundant Features for CRFs-based Sentence Sentiment Classification[C]//Proc. of Conference on Empirical Methods in Natural Language Processing. [S. l.]: ACM Press, 2008.

[8] 李寿山, 黄居仁. 基于 Stacking 组合分类方法的中文情感分类研究[J]. 中文信息学报, 2010, 24(5): 56-61. (下转第 162 页)

的数据应剔除；为消除实验过程中中央偏向性的影响，剔除每张图像第 1 次注视数据。最后得到 19 位有效被试的数据，对应每幅测试图像有 19 份眼动数据。

4.2.2 相似度计算

在计算相似度时，有 2 种模式：一种是从细节(像素点)的角度考虑，称为点对点相似度，它主要考虑每个点的差异，没有整体上的概念；另一种是从整体(区域)的角度考虑，称为区域相似度，它考虑了区域信息却忽略了区域的细节。点对点相似度与区域相似度从 2 个不同的角度反映 2 个显著图之间的相似度，缺一不可。在计算图像相似度时，先计算出 2 幅图像的颜色直方图，然后综合考虑点对点的相似度和区域相似度。通过式(3)计算出点对点区域相似度：

$$Sim(G,S)_{\text{point}}=\frac{1}{N}\sum_{i=1}^{N}\left(1-\frac{|g_i-s_i|}{\max(g_i-s_i)}\right) \tag{3}$$

其中，G、S 分别为由眼动数据和底层特征得到的二值图像；N 为图像的像素个数；g_i、s_i 分别是 2 幅二值图像第 i 个像素点对应的像素值。通过式(4)计算出区域相似度：

$$Sim(G,S)_{\text{regional}}=1-\frac{\sqrt{(X_g-X_s)^2+(Y_g-Y_s)^2}}{\sqrt{w^2+h^2}} \tag{4}$$

其中，(X_g, Y_g)和(X_s, Y_s)为两待评价区域质心坐标；w 和 h 是图像的宽度和高度。通过式(5)整合点对点相似度和区域相似度，得到最终相似度，作为评价指标。

$$Sim(G,S)=Sim(G,S)_{\text{point}}\times Sim(G,S)_{\text{regional}}=\left(\frac{1}{N}\sum_{i=1}^{N}\left(1-\frac{|g_i-s_i|}{\max(g_i-s_i)}\right)\right)\times\left(1-\frac{\sqrt{(X_g-X_s)^2+(Y_g-Y_s)^2}}{\sqrt{w^2+h^2}}\right) \tag{5}$$

其中，$Sim(G,S)\in[0.0,1.0]$，其值越接近 1.0 说明两者越相似，相似度为 1 时说明两者完全相同。

将眼动数据得到的显著图作为比较标准，计算由 4 个模型得到的显著图之间的相似度，表 2 的平均相似度结果从另一个角度验证了上文中各种模型适用和不适用情况的分析。

表 2 4 个模型平均相似度

模型	海滩组	鲜花组	大象组
Itti-Koch 模型	0.82	0.85	0.74
Stentiford 模型	0.89	0.86	0.73
光谱剩余假说模型	0.81	0.83	0.76
Hu-Rajan-Chia 模型	0.76	0.78	0.71

5 结束语

本文选择 4 种典型的视觉注意模型，在编程实现各自显著图的基础上，通过大量实验，分析生成显著图时不同的图像该选择哪种视觉注意模型。实验结果表明：Itti-Koch 模型适用于显著物体颜色变化大的图像，或者显著物体细小且与背景差别明显的图像；Stentiford 模型适用于背景颜色均匀的图像；光谱剩余假说适用于显著物体较小或者是人造物体的图像；Hu-Rajan-Chia 主要适用于细小物体。同时本文设计了眼动实验，并以眼动数据得到的显著图作为比较标准，分别计算其与 4 种模型生成显著图的相似度，进一步验证上述 4 种相关模型的适用性。

参考文献

[1] Itti L, Koch C, Niebur E. A Model of Saliency Based Visual Attention for Rapid Scene Analysis[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1998, 20(11): 1254-1259.

[2] Stentiford F W M. An Attention Based Similarity Measure with Application to Content Based Information Retrieval[C]// Proceedings of Storage and Retrieval for Media Databases Conference. Bellingham, USA: [s. n.], 2003.

[3] 曾志宏，李建洋，郑汉垣. 融合深度信息的视觉注意计算模型[J]. 计算机工程, 2010, 36(20): 200-202.

[4] Baluch F, Itti L. Training Top-down Attention Improves Performance on a Triple Conjunction Search Task[J]. PLoS ONE, 2010, 5(2): 1-10.

[5] Hou Xiaodi, Zhang Liqing. Saliency Detection: a Spectral Residual Approach[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Minneapolis, USA: IEEE Computer Society, 2007.

[6] Hu Y, Rajan D, Chia L T. Adaptive Local Context Suppression of Multiple Cues for Salient Visual Attention Detection[C]// Proceedings of IEEE International Conference on Multimedia and Expo. Amsterdam, the Netherlands: IEEE Computer Society, 2005.

[7] 陈再良，邹北骥，李海冰，等. 方向特征融合 ROI 提取算法[J]. 华中科技大学学报：自然科学版, 2011, 39(12): 102-106.

编辑 索书志

(上接第 158 页)

[9] 潘 宇，林鸿飞. 基于语义极性分析的矮馆评论挖掘[J]. 计算机工程, 2008, 34(17): 208-210.

[10] 宋光鹏. 文本的情感倾向分析研究[D]. 北京：北京邮电大学, 2008.

[11] Go A, Bhayani R, Huang Lei. Twitter Sentiment Classification Using Distant Supervision[Z]. 2009.

[12] Pak A, Paroubek P. Twitter as a Corpus for Sentiment Analysis and Opinion Mining[C]//Proc. of Language Resources and Evaluation Conference. Lisbon, Portugal: [s. n.], 2010.

[13] Jiang Long, Yu Mo, Zhou Ming, et al. Target-dependent Twitter Sentiment Classification[C]//Proc. of the 49th Annual Meeting of the Association for Computational Linguistics. Portland, USA: [s. n.], 2011: 151-160.

[14] Davidov D, Tsur O, Rappoport A. Enhanced Sentiment Learning Using Twitter Hashtags and Smileys[C]//Proc. of the 23rd International Conference on Computational Linguistics. Beijing, China: [s. n.], 2010: 241-249.

编辑 顾逸斐