DOI:10.11992/tis.201605019
网络出版地址:http://www.cnki.net/kcms/detail/23.1538.TP.20160808.0830.002.html

基于多情绪源关联模型的中文微博情感分析

李凌霄[1,2], 李绍滋[1,2],曹冬林[1,2]
(1.厦门大学 智能科学与技术系,福建 厦门 361005; 2. 厦门大学 福建省仿脑智能系统重点实验室,福建 厦门 361005)

摘 要:社交媒体信息的爆炸式增长,使得依据其对公众舆论情感的分析受到越来越多的关注。与传统文本不同,新浪微博中存在包括情感词、表情、图片和视频等特征在内的多情绪源,本文针对中文社交短文本情感分析中情感词典时效性问题和多情绪源间的关联性问题,提出了一种多情绪源关联模型。该模型考虑微博中的情感词和表情特征及其之间的关联关系,在经典的词典规则投票方法基础上,引入多情绪源以及关联概率,通过概率建模的方式对情感词和表情两类情绪源建立关联模型,实现对微博情感的判别。实验表明,在 6 171 条微博数据集中,多情绪源关联模型分类准确率达到了 85.3%,强于包含情感词和表情的传统投票模型(83.4%)以及包含同类多特征的 SVM 方法(82.9%)。

关键词:多模态情感分析;多情绪源;社交媒体;关联性
中图分类号: TP391 **文献标志码**:A **文章编号**:1673-4785(2016)04-0546-08

中文引用格式:李凌霄,李绍滋,曹冬林. 基于多情绪源关联模型的中文微博情感分析[J]. 智能系统学报, 2016, 11(4): 546-553.
英文引用格式:LI Lingxiao, LI Shaozi, CAO Donglin. Emotional multi-source correlation model for chinese micro-blog sentiment analysis[J]. CAAI Transactions on Intelligent Systems, 2016, 11(4): 546-553.

Emotional multi-source correlation model for chinese micro-blog sentiment analysis

LI Lingxiao[1,2], LI Shaozi[1,2], CAO Donglin[1,2]
(1. Cognitive Science Department, Xiamen University, Xiamen 361005, China; 2. Fujian Key Laboratory of the Brain-like Intelligent Systems, Xiamen 361005, China)

Abstract: With the explosion of social media information, sentiment analysis of public opinion is attracting more and more attention. Compared with traditional text, the Sina micro-blog contains a variety of emotional sources, including sentiment words, emoticons, pictures, etc. To solve the problem of the poor timeliness of lexicons in Chinese social short messages and to utilize the correlation between different emotional sources, an emotional multi-source correlation model (EMCM) is proposed to carry out sentiment analysis on a micro-blog. In particular, it takes advantage of the correlation between sentiment words and emoticons. It imports the multi-sources and correlation probabilities, and then builds a correlation model between the two emotional sources, emotional words and emoticons, based on a voting model using sentimental words. Experimental results show that this model achieved an accuracy of 85.3% in 6 171 micro-blogs, higher than either the traditional method based on voting (83.4%) or the SVM method based on similar multi-features (82.9%).

Keywords: multi-modal sentiment analysis; emotional multi-sources; social media; correlation

时下,社交媒体正成为人们生活中不可或缺的一部分,通过微博、微信等工具,人们可以随意发表对电影、商品的喜恶,对社会事件的个人观点,甚至对国家政策的看法。如何从包含这些信息的大规模

收稿日期:2016-05-19. 网络出版日期:2016-08-08.
基金项目:国家自然科学基金项目(61202143, 61305061, 61402386, 61572409);福建省自然科学基金项目(2013J05100).
通信作者:曹冬林. E-mail:another@xmu.edu.cn.

数据中获取诸如情感倾向在内的潜在信息,对于产品导向、广告精确投放、国家舆情控制等领域都具有重要意义,社交信息的数据挖掘与分析正成为研究者们关注的热门课题。

对英文社交媒体(如 Tweet)的情感分析已经有很多进展,分析的方法主要分为有监督方法[1-5]和基于词典或逐点互信息(PMI)[7]的无监督方法。而类似针对中文社交媒体的情感分析工作则仍处于起步阶段,所使用的方法大都源于英文情感分析方法,但由于社交媒体表现形式的多样化和中文网络语境多变性等原因,传统分类方法仍存在很大改进空间,本文针对目前存在的两个问题进行建模:

1)情感词典时效性差,中文新词的出现更为频繁,基于统计的方法在短周期内难以判断其情感;

2)传统方法未考虑多情绪源之间的关联。

这里的多情绪源是指微博中可能出现的能够体现其情感的多种异构特征,如情感词、表情符号、图片和视频等。并且这些情绪源之间存在以下在情感分析上可以进行互补利用的关联关系:

1)不同情绪源表达的情感强度可能不同,强情绪源可以对弱情绪源进行极性加强;

2)同一情绪下不同情绪源之间存在较强的关联性,例如在"哈哈"表情下出现正情感词的概率较大。

根据以上分析,我们提出了一种多情绪源关联模型,该模型对微博中的情感词和表情符号两种情绪源及其之间的关联进行建模。我们的实验结果显示,该模型在微博数据上优于经典分类算法,并且该模型具有拓展性,可以继续加入诸如图片和视频在内的其他情绪源。

1　情感分析相关工作

文本情感分析近几年逐渐成为热门研究课题,其内容主要包括情感极性分析和主客观分析等,本文主要关注情感极性分析。目前情感极性分析的方法主要分为两类:有监督的分类器学习方法和无监督的基于情感词典或者 PMI 的方法。

1.1　有监督方法

有监督方法大多通过机器学习技术从文本中选取合适的特征构建分类器,包括朴素贝叶斯、最大熵和支持向量机等,进而对不同情感进行分类。

分类器选择上,Pang 等[1]用以上 3 种分类器将影评分为正、负两类极性,引入了一元语法特征、二元语法特征、词性特征和词位置特征等 8 种组合特征,最终使用基于出现与否的一元语法特征 SVM 分类器效果最好,在其语料集中达到 83%的准确率。

特征选择上,D.Kushal 等[2]对语法规则、n-gram 特征进行了分析;Hatzivassiloglou 等[3]使用了情感词作为特征,对句子级别的情感倾向进行了分析;J.C.Na 等[4]对指定词语和否定短语特征进行了分析。

这类机器学习方法,例如多特征 SVM 情感分类方法,并未考虑到不同特征之间的关联关系。

1.2　无监督方法

无监督方法利用文本中带有情感的词汇的情感倾向,综合考虑文本的语法规则、句法构成等要素对文本进行情感极性的判别,通常采用投票的方法。在该类方法中,主要依靠文本分析,并未关注社交媒体信息中情绪源多并且不同情绪源之间存在关联性的特点。

基于情感词方法的基础是判断词的情感,对词汇的情感判断方法包括:基于情感词典、基于监督学习[5]和基于种子词[7-9]的方法等。

常用的中文情感词典有知网情感分析用词语集、台湾大学中文情感极性词典(NTUSD)和大连理工大学中文情感词汇本体库等。基于情感词典的方法主要缺陷在于覆盖面窄、无法包含网络新词。

Wilson 等[5]提出了一种二步分类的有监督方法判断短语的极性:1)判断将短语分类为有极性和中性;2)将第 1 步中得出的有极性短语进一步划分为 4 类极性,每一步使用不同的特征进行分类,分类器相同(BoosTexter AdaBoost.HM[6])。最终在其数据集上准确率达到 75.9%。

Turney[7]提出了一种判断单词情感的方法,通过在大规模语料集中分别计算目标单词与正负极性种子词(正种子词:excellent;负种子词:pool)的逐点互信息,将两个结果进行对比得出目标单词的情感,最终在其数据集中达到 82.8%的准确率,缺点是需要大规模语料集,运算量大。

此外,Xia H.等[9]研究了英文社交媒体中出现的情感标记信号在无监督情感分析中的应用,取得了良好的效果。

1.3　中文微博情感极性分析研究现状

中文微博情感极性分析主要方法来源于上文提及的英文文本情感分析相关方法[10]。

目前,由中国中文信息学会(CIPS)主办的中文倾向性分析评测(The Fifth Chinese Opinion Analysis Evaluation, COAE)聚集了该领域大量研究成果。COAE 评测由 2008 年开始每年举办一次,发布中文倾向性分析的相关任务,包括情感识别、新词发现、观点句提取和评价对象识别等。表 1 给出了 COAE2013 http://ccir2013.sxu.edu.cn/COAE.aspx 任务 1(基于否定句的句子级倾向性分析)的最佳评测结果。

表 1 COAE2013 任务 1 最佳评测宏平均结果

Table 1 COAE2013 Task1 best evaluation results

参数	褒义	中性	贬义
准确率	0.741	0.445	0.836
召回率	0.619	0.725	0.464
F_1	0.674	0.551	0.597
精度		0.615	

最佳结果[11]使用了集成学习的方法,通过多次欠采样训练 NB、ME、SVM 基分类器,通过 product rule 融合多个基分类器。该方法针对标注数据集较少的情况,提高了分类器的鲁棒性和泛化能力。

在中文微博情感分析的多种方法中,SVM 方法虽然引入了不同特征,但是认为特征之间相互独立;基于规则投票的方法主要依赖情感词典和语法规则,也有引入表情符号等情绪源的方法,但未考虑不同情绪源之间的关联。

此外,谢丽星等[12]提出了基于层次结构的 SVM 分类方法,选取主题相关特征构建分类器对微博情感进行三分类。通过分句考虑了 3 类极性的句子数目以及首尾句情感极性,并且依据主题选取了多种特征训练分类器,在其数据集上达到 67.283% 的准确率。但通过对我们的 6171 条微博进行分析发现,句子数目大于 2 的微博仅占 12%,因此分句对情感分析效果不大。此外由于本文针对没有主题标签的微博,因此最终在实验中选择文献[12]中与主题无关的不分句最佳特征 SVM 以及无关联多情绪源模型作为对比方法。

2 算法实现

多情绪源关联模型受基于词典投票的情感分析方法启发,对包括情感词在内的多情绪源及其间的关联进行建模(本文只考虑情感词和表情两种情绪源)。因此本章从基于词典投票的分类模型,到加入表情特征进行改进,近而引入后验概率联合建模 3 个过程来介绍模型的产生原理,最后介绍多情绪源关联模型的构建方法(算法将微博分为负面、中性和正面 3 种情感)。

2.1 原理框图

图 1~3 分别展示了 3 种情感分类模型的组成原理,可以看出相比其他两类模型只考虑单一或者相互独立的情绪源特征,本文提出的多情绪源关联模型综合考虑了不同情绪源及其之间的关联进行建模,并且在第 2.4 节的实验中证明了这种关联对于情感分析的作用。

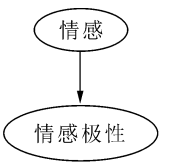

图 1 情感词投票模型

Fig.1 Word voting model

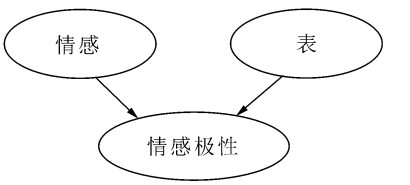

图 2 无关联模型

Fig.2 Uncorrelated model

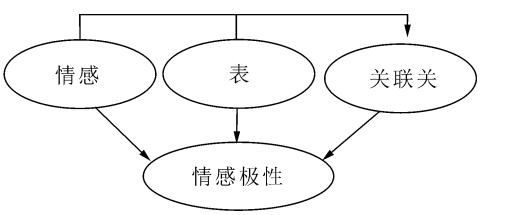

图 3 多情绪源关联模型

Fig.3 Emotional multi-source correlation model

2.2 基于词典投票的情感分类模型

本节介绍了传统方法中基于情感词典投票的情感分类模型,并对其进行了概率转换,再依据否定词和感叹句对情感词极性进行了修正。

2.2.1 情感词典概率模型

基于情感词典的分析方法将情感词典中标注为正负极性的情感词作为特征,先对文本进行分词(本文中涉及的分词工具使用了中科院计算所开发的 ICTCLAS50 分词系统 http://www.ictclas.org/),将正负情感词在文本中出现次数的差值作为文本正负情感判断的依据。根据式(1)进行极性投票判断。

$$\text{文本极性} = \begin{cases} \text{正}, \text{正情感词数} - \text{负情感词数} > 0 \\ \text{中}, \text{正情感词数} - \text{负情感词数} = 0 \\ \text{负}, \text{正情感词数} - \text{负情感词数} < 0 \end{cases} \tag{1}$$

如果将以上判断方法用概率模型进行表示,可以得到式(2)。

$$\text{微博极性} = \begin{cases} \text{正}, P_w(p) - P_w(n) > 0 \\ \text{中}, P_w(p) - P_w(n) = 0 \\ \text{负}, P_w(p) - P_w(n) < 0 \end{cases} \tag{2}$$

式中:$P_w(p) = \dfrac{\text{正情感词数}}{\text{总情感词数}}$,$P_w(n) = \dfrac{\text{负情感词数}}{\text{总情感词数}}$。

2.2.2 否定词和感叹句分析

针对中文微博里存在否定词、感叹句等语法结

构的特点,本文对情感词的极性权值进行了修正。

与文献[12]中类似,模型对否定词的出现进行了处理,自定义了 24 个常用否定词,如表 2 所示,将以否定词为中心,大小为 3 窗口的中出现的情感词极性反转。

表 2　自定义否定词表

Table 2　Custom privative words list

自定义否定词
不、不会、不可能、不是、不应该、并非、并不、 不、不会、没、无、非、莫、勿、未、否、别、无、不曾、 未必、没有、不要、难以、未曾、毫无、毫不

感叹句通常起到的是加强语义的作用,而对于语句的情感影响也会起到类似的加强效果。我们认为出现感叹句的句子中,情感词表达效果翻倍,因此使用了最为直接的处理方法,将感叹句中的情感词个数在原基础上乘以 2。

2.3　无关联的情感词和表情模型

很多情况下,单独使用情感词难以判断微博所表达的极性,因此可以通过引入其他情绪源来综合判断极性,我们考虑了表情符号作为联合特征,因为表情和情感词在微博情感分析中具有如下优势互补的特性。

1)微博中情感词分布广泛,一条微博中往往包含多个情感词。但仅利用情感词进行情感判别的缺点在于情感词典时效性差:情感新词出现较频繁,但刚出现时数量少,使用基于统计的新词极性判别方法在新词出现初始周期内难以对新词进行识别和判断。

2)微博上表情符号的使用相对固定,但利用表情进行情感判别的缺点在于一条微博中表情个数不多,同时并非所有微博都包含表情。

此外,经过试验表明,微博表情特征的以下特点也能够提升情感分类效果:

1)微博表情对情感的表达比文本更为直接和显著;例如微博“终于通关了😁”,文本中并未出现情感词,仅通过词典将其判断为中性情感,加入表情特征后判断为正面情感。

2)微博表情可能直接作为句子成分出现在句子当中。例如“今天下雨了,不过😊😊😊”,这条微博将表情符号“太开心”作为句子成分加入转折句当中,最终表示了正极性情感。

因此我们对情感词和表情符号联合建模,以综合利用二者在微博情感判断中的互补优势,和表情特征的自身判别优点,具体模型如式(3)~(5)所示:

$$S_p^0 = \operatorname{argmax}_{\omega_w,\omega_f}(\omega_w P_w(p) + \omega_f P_f(p)) \quad (3)$$

$$S_n^0 = \operatorname{argmax}_{\omega_w,\omega_f}(\omega_w P_w(n) + \omega_f P_f(n)) \quad (4)$$

$$\text{微博极性} = \begin{cases} \text{正}, S_p^0 - S_n^0 > 0 \\ \text{中}, S_p^0 - S_n^0 = 0 \\ \text{负}, S_p^0 - S_n^0 < 0 \end{cases} \quad (5)$$

式中:$P_f(p) = \dfrac{\text{正表情数}}{\text{总表情数}}$,$P_f(n) = \dfrac{\text{负表情数}}{\text{总表情数}}$,$\omega_w$ 和 ω_f 为情感词和表情的权重系数,本文通过遍历系数空间选取准确率最高的系数值。

2.4　多情绪源关联模型

2.3 节模型认为情感词与表情之间是相互独立的,没有考虑情感词和表情之间的关联关系,以及这种关系对情感极性判断的影响,因此这里引入了后验概率对其进行修正。

表 3 给出了一个例子,在该例中,虽然出现的情感词都为正极性,但表情符号却只有负面表情,通过 2.3 模型进行判断,将这条微博错分成负极性。

表 3　无转折词的转折句实例

Table 3　Examples of transitional sentences without transitional words

类型	实例
正面情感微博	天兔遇上给力的海航, 终于跟坐快艇似的回到广州。杭州 之行说起来还算圆满吧,多年 未见的大学死党、越来越漂亮的 老妹鱼头阿奋来平,还有闺蜜 菁菁茜女人想念大家了。表情:“泪”
正情感词	给力;圆满;漂亮
负表情符号	泪

通过 2.3 中的方法,对这条微博的情感极性判断为负,但实际极性为正面情感。我们引入了概率模型 $P(w,f \mid p)$,$P(w,f \mid n)$ 来增强类似的情感极性判断,构建了关联模型(6)~(8):

$$S_p = \operatorname{argmax}_{\omega_w,\omega_f} \frac{\omega_w P_w(p) + \omega_f P_f(p) + P(p \mid w,f)}{\text{normal}} \quad (6)$$

$$S_n = \operatorname{argmax}_{\omega_w,\omega_f} \frac{\omega_w P_w(n) + \omega_f P_f(n) + P(n \mid w,f)}{\text{normal}} \quad (7)$$

$$\text{微博极性} = \begin{cases} \text{正}, S_p - S_n > 0 \\ \text{中}, S_p - S_n = 0 \\ \text{负}, S_p - S_n < 0 \end{cases} \quad (8)$$

式中：normal 为归一化因子。

$$\text{normal} = [\omega_w P_w(p) + \omega_f P_f(p) + P(p \mid w,f)] + [\omega_w P_w(n) + \omega_f P_f(n) + P(n \mid w,f)]$$

$P(p \mid w,f)$ 和 $P(n \mid w,f)$ 计算如下（默认 $P(p) = P(n) = 0.5$）：

$$P(p \mid w,f) = \frac{P(w,f \mid p)P(p)}{P(w,f)} \cong P(w,f \mid p)P(p) = \frac{P(w,f,p)P(p)}{P(p)} = P(w \mid f,p)P(f \mid p)P(p) \cong 0.5\prod_{i=1}^{a}\prod_{j=1}^{b}P(w_i \mid f_j,p)\prod_{j=1}^{b}P(f_j \mid p) \quad (9)$$

类似地

$$P(n \mid w,f) = \frac{P(w,f \mid n)P(n)}{P(w,f)} \cong P(w,f \mid n)P(n) = \frac{P(w,f,n)P(n)}{P(n)} = P(w \mid f,n)P(f \mid n)P(n) \cong 0.5\prod_{i=1}^{a}\prod_{j=1}^{b}P(w_i \mid f_j,n)\prod_{j=1}^{b}P(f_j \mid n) \quad (10)$$

式中：a 和 b 分别表示一条微博中情感词和表情符号的个数。而 $P(w_i \mid f_j,p)$、$P(f_j \mid p)$、$P(w_i \mid f_j,n)$、$P(f_j \mid n)$ 是对数据集进行统计后得出的结果。该模型认为词与词（表情与表情）之间相互独立；但是词与表情、词与微博极性、表情与微博极性之间存在关联，用情感词与表情之间的关联得出的结果来改善原始结果。

此外，为了消除 $P(p \mid w,f)$ 与 $P(n \mid w,f)$ 中多小数相乘使值过小的问题，实际计算时，取

$$P(p \mid w,f) = \frac{P(p \mid w,f)}{P(p \mid w,f) + P(n \mid w,f)}$$

$$P(n \mid w,f) = \frac{P(n \mid w,f)}{P(p \mid w,f) + P(n \mid w,f)}$$

在表 3 所示的示例中，使用 2.3 节中的方法进行极性判断，结果如下：

$$S_p^0 = \omega_w P_w(p) + \omega_f P_f(p) = 1.0 \times \frac{3}{3} + 0 = 1.0$$

$$S_n^0 = \omega_w P_w(n) + \omega_f P_f(n) = 0 + 1.5 \times \frac{1}{1} = 1.5$$

因此，$S_p^0 - S_n^0 < 0$，3.3 判断为负极性，而在关联模型中：$P(p \mid w,f) = 1$，$P(n \mid w,f) = 0$，$S_p - S_n = 0$

$$S_p = \frac{\omega_w P_w(p) + \omega_f P_f(p) + P(w,f \mid p)}{\text{normal}} = \frac{1 + 0 + 1}{(1 + 0 + 1) + (0 + 1.5 + 0)} = 0.57$$

$$S_n = \frac{\omega_w P_w(n) + \omega_f P_f(n) + P(w,f \mid n)}{\text{normal}} = \frac{0 + 1.5 + 0}{(1 + 0 + 1) + (0 + 1.5 + 0)} = 0.43$$

$S_p - S_n = 0.14 > 0$，最终结果为正性（本数据集下，取 $\omega_w = 1$，$\omega_f = 1.5$）。分类正确的原因是通过“泪”与上述情感词之间的关联性，考虑了“泪”与上述情感词出现情况下，分类为正极性的概率。

多情绪源关联模型不限于情感词和表情符号两个情绪源，可以通过加入更多的情绪源，例如图片、视频等，来拓展关联模型。

3 实验结果与分析

3.1 实验数据及验证方法

上文中提及的 COAE 评测给出了公共数据集，但由于其数据集中所包含的有表情微博数量十分稀少，不适合测试本方法，因此本文通过新浪微博 API 爬取微博信息，并对爬取的 6 171 条微博进行了人工标注，经过统计，微博数据来自社会、电影、电视剧、美食、娱乐八卦、科技等多个领域。

所选择数据集中正极性微博所占比例偏大，中极性比例偏小，并且含有表情的微博较多（主要分布于电影、电视剧、娱乐八卦和美食等领域），但用于比较的各个分类方法所用数据集相同，不会对结果比较造成影响。

我们所使用的情感词典为大连理工大学中文情感词汇本体库 http://ir.dlut.edu.cn/EmotionOntologyDownload.aspx? 以及自定义的少量新词（如坑爹、给力等），一共 27 488 个（正极性词 13 556 个，负极性词 13 932 个）。在分词时，使用 ICTCLAS50 自定义词典接口，调用了情感词典和否定词典。

模型中，使用表情符号和情感词进行了联合建模，表情符号选择微博常用表情中默认的 50 个表情符号，如😁（正性）和😡（负性）。如表 4。

表 4 实验数据极性分布

Table 4 Dataset sentiment polarity distribution

极性	微博数目	比例/%
正极性	4 196	67.9
中性	621	10.0
负极性	1 354	21.9
含表情微博	5 182	84

3.2 对比实验说明

对比实验 1 采用文献[12]中一步三分类最佳特

征组合(去除了情感短语和中文是否出现这两个特征),此外因为本文数据集中的微博包含的多句子情况少,因而不考虑分句的情况进行第 2 次分类;同时本文的情感极性分析针对无主题标签的微博,因此不考虑主题特征。在文献[12]所做的实验中,url 特征与主客观分类对最终效果有负面影响,因此也不将这两个因素考虑在内。此外,选用的情感词典和表情符号、标点符号也与之不同。最终使用的特征表示如表 5 所示,用词袋模型(BOW)表示。其中否定词采用与 3.2.2 中相同处理方法。对比实验二采用 3.3 节中方法。实验采用五折交叉验证。

表 5　对比实验特征表示

Table 5　Sentiment features of baseline

序号	类型	特征描述	维度	备注
1	表情	正、负向表情符号个数(50 个表情)	2	微博常用表情中的默认表情
2	情感词	正、负向情感词个数(2 461 个情感词)	2	使用大连理工大学词典
3	形容词	形容词个数	1	
4	动词	动词个数	1	
5	感叹号	是否出现! 或!	1	
6	问号	是否出现? 或?	1	

3.3　实验结果及分析

分类器说明:

1)关联模型:多情绪源关联模型(情感词、表情关联建模);

2)NB:朴素贝叶斯模型,所使用的特征与对比实验一的 SVM 方法相同,使用 BOW 表示特征;

3)传统词典:传统的基于情感词典以及规则进行投票的方法(2.2 中的方法);

4)词典+表情:传统基于情感词典及规则进行投票的方法,辅以表情特征(2.3 中的方法)。

5)SVM:文献[12]中一步三分类方法。

从表 6 的实验结果可以看出,本文提出的多情绪源关联模型分类效果最佳,达到 85.3%,比传统基于情感词加表情投票的方法高出了 1.9%,比同类多特征 SVM 高出了 2.4%。说明了对情绪源进行关联性建模,能够有效提高情感分类效果,表明不同情绪源之间的关联关系与情感极性也是相关的。缺点在于对情绪源单一的微博(例如无表情的微博)则主要依赖于传统情感词典分类方法。

表 6　总体结果

Table 6　Experimental results

方法	正极性		中极性		负极性		准确率
	P	R	P	R	P	R	
关联模型	0.906	0.945	0.506	0.395	0.806	0.779	0.853
NB	0.945	0.753	0.270	0.747	0.657	0.537	0.705
传统词典	0.833	0.617	0.162	0.617	0.608	0.312	0.550
词典+表情	0.894	0.925	0.507	0.369	0.750	0.763	0.834
SVM	0.870	0.945	0.538	0.330	0.769	0.702	0.829

注:P、R 分别表示准确率(Precision)和召回率(Recall)。

3.4　错误分析

本节中对混合概率模型的错误分类样本进行了分析,研究了造成分类错误的原因,如表 7 所示。

表 7　错误类别及相关示例

Table 7　Misclassified examples

序号	错误类别描述	示例
1	情感词未包含在词典中	各种消失。信号标识消失时间消失电池电量各种消失苹果系统 ios7 真心是坑爹啊。
2	无表情符号特征	有时候突然不能输入中文,关机重启后正常。已发生两次。
3	负面表情或情感词表达正面或中性情感	略凶残的相机效果,自恋狂可以点赞。
4	反问句式加强了负面情感	坑不坑!
5	反讽句式	3 天内出现 4 次这种情况,还能不能一起愉快地玩耍了?
6	转折句式	其实我还是挺喜欢 ios7 的如果他不卡的话
7	正面表情表达中性	同志们不好意思,我刚才发错了,那个是草稿箱里的表情:“嘻嘻”“哈哈”

实验结果表明,在缺乏表情符号特征的微博中分类效果较差,主要原因还是由于当没有表情特征时,分类器只依赖于情感词以及简单规则进行分类。此外,对转折句、反讽句等句式的判断存在不足,原因是微博中很多反讽句式的出现往往是伴随着网络

新词出现的,并且没有明显的句式标识词(例如,“这小偷真是太机智了”),使得对反讽句和转折句的判断比较困难。

4.4 对比分析

通过在同一数据集上对不同模型的实验表明,多情绪源关联模型能够很好地解决基于情感词判别方法时效性差的问题,并且在分类时综合考虑了不同情绪源之间的关联性,提高了分类效果。相对于对比实验 2 的普通情感词和表情建模的方法,多情绪源关联模型通过引入后验概率,利用情感词与表情符号之间的关联性,加强情感判断性能。另外,使用对比实验 1 中的 SVM 分类器时,虽然加入了包括表情、否定词在内的多特征,但认为不同特征之间相互独立。多情绪源关联模型所能解决的一些错分类问题如表 8 所示。

表 8 关联模型分类正确样本

Table 8 Experimental examples

序号	错误类别描述	示例	关联模型	SVM 模型	3.3 模型
1	情感词与表情间的关联关系主导分类结果	趴在墙上不能更萌 表情:“可怜”	正确	错误	错误
2	微博包含否定词	这啥,太不稳定了,又抽风般地自己好了	正确	正确	正确
3	SVM 误判	祈福。我叫不生气!! 表情:“蜡烛”“生病”“抓狂”	正确	错误	正确

4 结论及展望

新浪微博作为时下最为流行的社交网站之一,不仅是民众钟爱的社交工具,更是研究者挖掘数据的天堂,其商业价值和学术价值都不断升温。本文对微博数据挖掘领域的情感分析进行了研究,提出多情绪源关联模型,针对传统基于词典的方法重新进行了关联性建模,使得分类准确率相比传统模型(3.3 节模型)提高了 1.9%;相比多特征 SVM 提高了 2.4%。但该方法仍是较为简单的情感分析方法,就方法本身而言,也存在很大的提升空间,可以对以下几个方面进行改进:

1)拓展模型,引入更多情绪源,包括图片和视频等,使模型更适合于微博语境。

2)在概率模型中引入更加复杂的语法规则分析,例如祈使句式、多重否定、反讽句等;

3)挖掘微博用户之间的社交网络关系对情感分析的影响,通过有关联用户来参与判断情感。

参考文献:

[1] PANG Bo, LEE L, VAITHYANATHAN S. Thumbs up?: sentiment classification using machine learning techniques [C]//Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, PA, USA: ACM, 2002, 10: 79-86.

[2] DAVE K, LAWRENCE S, PENNOCK D M. Mining the Peanut gallery: opinion extraction and semantic classification of product reviews[C]//Proceedings of the 12th International Conference on World Wide Web. Budapest, HU: ACM, 2003: 519-528.

[3] YU HONG, HATZIVASSILOGLOU V. Towards answering opinion questions: separating facts from opinions and identifying the polarity of opinion sentences[C]//Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, PA, USA: ACM, 2003: 129-136.

[4] NA J C, SUI H, KHOO C, et al. Effectiveness of simple linguistic processing in automatic sentiment classification of product reviews[C]//MCILWAINE I C. Knowledge Organization and the Global Information Society: Proceedings of the Eighth International ISKO Conference. Wurzburg, Germany: Ergon Verlag, 2004: 49-54.

[5] WILSON T, WIEBE J, HOFFMANN P. Recognizing contextual polarity in phrase-level sentiment analysis[C]//Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing. Stroudsburg, PA, USA: ACM, 2005: 347-354.

[6] SCHAPIRE R E, SINGER Y. BoosTexter: a boosting-based system for text categorization[J]. Machine Learning, 2000, 39(2/3): 135-168.

[7] TURNEY P D. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews [C]//Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. Stroudsburg, PA, USA: ACM, 2002: 417-424.

[8] 朱嫣岚, 闵锦, 周雅倩, 等. 基于 HowNet 的词汇语义倾向计算[J]. 中文信息学报, 2006, 20(1): 14-20.
ZHU Yanlan, MIN Jin, ZHOU Yaqian, et al. Semantic orientation computing based on HowNet[J]. Journal of Chinese information processing, 2006, 20(1): 14-20.

[9] HU Xia, TANG Jiliang, GAO Huiji, et al. Unsupervised sentiment analysis with emotional signals[C]//Proceedings

of the 22nd international conference on World Wide Web. Rio de Janeiro, Brazil: ACM, 2013: 607-618.

[10]赵妍妍, 秦兵, 刘挺. 文本情感分析[J]. 软件学报, 2010, 21(8): 1834-1848.
ZHAO Yanyan, QIN Bing, LIU Ting. Sentiment analysis [J]. Journal of software, 2010, 21(8): 1834-1848.

[11] 魏现辉, 任巨伟, 何文译, 等. DUTIR: 中文短文本倾向性分析及要素抽取方法研究[C]//第五届中文倾向性分析评测研讨会论文集. 太原, 2013: 116-129.
WEI Xianhui, REN Juwei, HE Wenyi, et al. DUTIR: method research of sentiment analysis and elements extraction of Chinese short text [C]//Proceedings of the Fifth Chinese Opinion Analysis Evaluation. Taiyuan, 2013: 116-129.

[12]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.
XIE Lixing, ZHOU Ming, SUN Maosong. Hierarchical structure based hybrid approach to sentiment analysis of Chinese micro blog and its feature extraction[J]. Journal of Chinese information processing, 2012, 26(1): 73-83.

作者简介:

李凌霄,男,1990 年生,硕士研究生,主要研究方向为跨媒体舆情分析。

曹冬林,男,1977 年生,博士,厦门大学智能科学与技术系助理教授,主要研究方向为自然语言处理、信息检索、跨媒体舆情分析、计算机视觉、模式识别。

李绍滋 ,男,1963 年生,博士,教授,博士生导师,主要研究方向为人工智能及其应用、计算机视觉与机器学习、运动目标检测与识别、跨媒体舆情分析等。主持过多项国家、省市级项目研究,获得省科学技术三等奖两项,发表学术论文 200 余篇,其中:27 篇被 SCI 检索、171 篇 EI 检索。

2016 年国际云和可信计算研讨会
International Symposium on Cloud and Trusted Computing 2016

Current and future software needs to remain focused towards the development and deployment of large and complex intelligent and networked information systems, required for internet-based and intranet-based systems in organizations. Today software covers a very wide range of application domains as well as technology and research issues. This has found realization through Cloud Computing, Big Data, and data intensive applications. Vital element in such networked information systems are the notions of trust, security, privacy and risk management.

Cloud and Trusted Computing (C&TC 2015) is the 6th International Symposium on Cloud Computing, Trusted Computing and Secure Virtual Infrastructures, organized as a component conference of the OnTheMove Federated Conferences & Workshops. C&TC 2016 will be held in Rhodes, Greece.

The conference solicits submissions from both academia and industry presenting novel research in the context of Cloud Computing, Big Data, and data intensive applications, presenting theoretical and practical approaches to cloud and big data trust, security, privacy and risk management. The conference will provide a special focus on the intersection between cloud, big data, and trust bringing together experts from the three communities to discuss on the vital issues of trust, security, privacy and risk management in Cloud Computing, shedding the light on novel issues and requirements in big data domains. Potential contributions could cover new approaches, methodologies, protocols, tools, or verification and validation techniques. We also welcome review papers that analyze critically the current status of trust, security, privacy and risk management in the cloud. Papers from practitioners who encounter trust, security, privacy and risk management problems and seek understanding are also welcome.

Website: http://www.otmconferences.org/index.php/conferences/ctc-2016