文章编号：2096-1472(2016)-12-33-03

基于微博文本的情感倾向分析

宋继红，葛达明
（沈阳工业大学信息科学与工程学院，辽宁 沈阳 110870）

摘　要：微博作为一种用户发表看法和观点的载体已成为互联网上一个重要的情感交流平台，博文搜索为这种交流提供了方便快捷的途径。基于HowNet等中文情感词典的微博情感词的抽取和分类，计算词语语义相似度和倾向性。对文本情感倾向的加权值、表情、和情感词增强因素等进行综合考虑。实验结果表明表情情感倾向对微博情感倾向起着重要作用；在表情和文本情感倾向比值固定的情况下，调整因素和中性区间的选择会对情感倾向判断准确率产生影响；通过与基于HowNet语义相似度的计算模型比较，该文方法使得情感倾向判断准确率有所提高。

关键词：情感提取；情感分析；微博文本
中图分类号：TP399　　**文献标识码**：A

An Analysis of the Emotional Tendency Based on Micro-Blog Text

SONG Jihong,GE Daming
(*School of Information Science and Engineering,Shenyang University of Technology,Shenyang* 110870,*China*)

Abstract: As a carrier for users to express opinions and viewpoints,micro-blog has become an important platform for emotional communication.The function of blog search brings more convenience and efficiency to the communication. The paper proposes an algorithm to extract and categorize the emotional words in micro-blog based on HowNet and other Chinese dictionaries of emotional words.Comprehensively considering weight values of text emotional tendency,emoticons and enhancement factors of emotional words,the algorithm computes semantic similarity and tendency among words and expressions.The experiment result shows that:the emotional tendency of emoticons plays an important role in the emotional tendency in micro-blog,the selection of adjustment factors and neutral ranges affects the judgment accuracy of emotional tendency under the circumstance that the emotional tendency ratio between the emoticon and the text is fixed,compared with the algorithm model based on HowNet semantic similarity,the method proposed in this paper effectively improves the judgment accuracy of emotional tendency.

Keywords:sentiment extraction;emotion analysis;micro-blog text

1 引言(Introduction)

微博文本中往往包含了大量的文本作者对于某事件的情感，例如对微博文本、时事的态度、意见、评价等，研究如何高效的对舆论信息进行情感挖掘与趋势分析，从而更好地分析网民群体的行为规律。通过分析，能够实现对网络流行事件或突发事件的快速分析，对于政府机构舆情分析、企业市场决策、消费行为分析等方面具有重要意义。当前，主要有两大类针对情感分析的方法，分别是基于语义的方法与基于机器学习的方法[1]。一个词汇的语义倾向是指通过对微博文本个体词汇褒贬度进行分析得到的度量值，取值区间为±1。微博文本的情感倾向值最终通过汇总组合个体词汇的情感倾向度量值得到[2]。基于机器学习的情感分析方法的思路是构造一个分类器，并使用已分类的训练集来训练这一分类器，研究重点在于如何提高训练效果[3]与获得高质量的训练集[4]。

中文微博的情感分析一般可以分为三个步骤。第一步为微博语料的收集和预处理；第二步根据给定的规则从微博文本中抽取出情感词并且标注情感词极性；第三步依据情感倾向值计算方法，对微博文本进行倾向性计算，得出整体情感倾向值。微博情感分析工作的主要流程如图1所示。

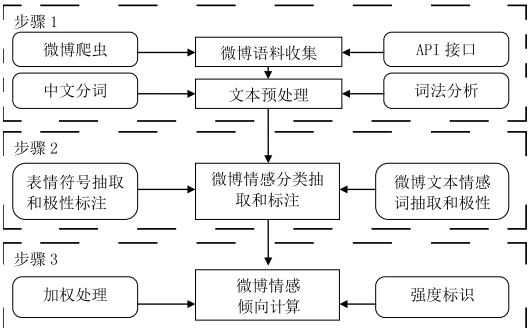

图1 中文微博情感分析工作流程图

Fig.1 Flow chart of Chinese micro-blog emotion analysis

2 微博语料的收集和文本预处理(Collection and text pretreatment of micro-blogging corpus)

微博语料的获取途径主要有通过互联网上提供的语料库获得和通过微博爬虫程序或网站应用程序编程接口（API）函

数获取三种方式[5]。互联网上提供的微博语料数量大质量优，但因需要经过收集整理再发布的过程，往往难以获得最新的数据。基于微博的爬虫程序不同于传统的爬虫程序依靠超链接关系而是通过节点之间的社交关系来获取整个微博的网络数据。单独采用微博爬虫程序往往会在局部陷入深度优先搜索，难以采集到大范围的微博数据[6]。大多数网站都加入了反机器人机制如验证码、验证滑块等，进一步增加了爬虫程序获取数据的难度；多数微博平台都提供了使用API接口访问的功能，但在实际使用过程中常常受到诸多限制，因此本文采用三者结合的方法进行语料收集。

文本预处理主要使用自然语言处理技术对微博文本进行分词与词性标注[7]。国内对于自然语言处理技术的研究由来已久，目前已比较成熟。本文采用中国科学院计算技术研究所的汉语词法分析系统(ICTCLAS)作为词法分析器，该系统历经多次内核升级，分词速度与精度均处于国内一流水平。

3 微博情感分类抽取和标注(The extraction and labeling of micro-blogging emotional classification)

3.1 文本表情处理

中文微博网站提供了丰富的表情符号，借助这些符号，可以比较容易地将微博分为主观性文本和客观性文本，从而使得分析结果更加精确。主观性文本的情感倾向又分为积极和消极两类，客观性文本的情感倾向为中性。本文对表情强度采取了人工标注强度的方法。强度值为(0,1)之间代表正向情感表情，消强度值为(−1,0)之间代表负向情感表情，强度值为0表示中性表情。下表列举了一些具有代表性的正向、负向和中性表情，如表1所示。

表1 常用表情符号及强度

Tab.1 Common expression symbol and intensity

积极表情符号		强度	消极表情符号		强度
	哈哈	1.0		怒	−1.0
	微笑	0.9		泪	−0.9
	嘻嘻	0.8		悲伤	−0.8
	可爱	0.6		失望	−0.5
	挤眼	0.5		怒骂	−0.8

3.2 微博文本情感词典构建

现有的中文情感词典较少，比较成熟的有台湾大学简体情感词典(NTUSD)[8]和知网(HowNet)中文情感词典[9]。本文对NTUSD、HowNet、情感词汇本体库和中文褒贬意词典等词典进行整理、去重，同时加入搜狗实验室提供的互联网词库，扩充网络流行语。加入了程度副词和否定词，整理后的情感词典包括正向情感词4800个、负向情感词6200个。

与基于句子的情感分析不同，微博文本段落的情感分析与判断对情感词典提出了更高的要求。对情感词和程度副词赋予权重，从而定量地度量文本倾向性可以提升文本情感的准确性。情感词的强度划分在(−1,1)，分别用(0,1)和(−1,0)代表正面情感词语的权重和负面情感词语的权重；程度副词的权重划分在（0.2,1.2），按照强度由强到弱划分为5级，如表2所示。若出现多个程度副词修饰一个词语的情况，则多个程度副词的综合权重为所有程度副词权重之乘积。

表2 常用程度副词及权重

Tab.2 Adverbs of degree and weight

程度级别	示例词语	权重
5级	完全、非常、最为	1.2
4级	尤其、多、格外	0.9
3级	比较、愈发、更加	0.6
2级	有点、稍微、略微	0.3
1级	不大、不很、丝毫	0.2

4 基于文本和表情的情感计算方法(Emotional computing method based on text and emotion)

基于表情的微博情感分析可以使用五元组Q(A,S,F,E,T)表示，其中A、S、F、E、T分别表示程度副词、情感倾向、表情情感倾向、增强因子与发表的时间。其中发表时间T对于微博情感分析结果影响可以忽略不计，特将五元组简化为四元组Q(A,S,F,E)。微博情感值的计算过程从而可以转换为从微博文本中抽取出程度副词、情感倾向、表情情感倾向、增强因子并对其进行处理的过程。

微博文本的情感倾向由表情和文本的情感两部分组成，微博文本的情感倾向值可以通过对这两部分的情感倾向值加权处理来得出。

$$Q(P)=\lambda Q(P_S)+(1-\lambda)Q(P_T) \tag{1}$$

其中，$Q(P)$、$Q(P_S)$、$Q(P_T)$分别为微博总体的情感倾向值、微博表情的情感倾向值，以及微博文s本的情感倾向值。其中λ为变量，取值区间为(0,1)，代表总体情感倾向值中表情与文本情感倾向所占的比重。

微博表情的情感倾向值可根据如公式(2)得到：

$$Q(P_S)=\frac{1}{n}\sum_{n=1}^{n}Q(p_{si}) \tag{2}$$

其中，$Q(p_{si})$为微博文本中第i个表情的情感强度。

使用HowNet提供的词汇语义相似度计算工具计算义原之间的相似度，可以得到词语之间的相似程度。进而计算出词语的情感倾向，最终计算出微博文本的情感倾向值$Q(P_T)$。对于两个汉语词语W_1和W_2，如果W_1有n个义项：$x_1,x_2,\cdots,x_n$；W_2有m个义项：$y_1,y_2,\cdots,y_m$，则规定W_1和W_2的相似度为各义项相似度之最大值，即

$$S(W_1,W_2)=\max_{i\in[1,n],j\in[1,j]}(S(x_i,y_j)) \tag{3}$$

义原相似度的计算公式为

$$S(x_i,y_j)=\frac{\alpha}{\alpha+d(x_i,y_j)} \tag{4}$$

其中，α为变量，取值区间为（0,+∞）；$d(x_i,y_j)$表示义原x_i和义原y_j的义原距离，由词汇语义相似度计算工具得出。一般地对于一个不在情感词典中的词语，其情感倾向值可以通过对比其与情感词典中的词之间的距离得到。具体计算方法为：将词语W分别与正面和负面情感词典中的每个种子词进行比较得到其正、负面情感倾向值，再通过比较其与正负向情感值之间的均差，得出其情感倾向值。某个词语W的情感倾向值可以通过下式计算得出

$$Q(W)=\frac{1}{n}\sum_{i=1}^{n}S(W,P_i)-\frac{1}{m}\sum_{j=1}^{m}S(W,N_j) \tag{5}$$

其中，P_i、N_j分别表示情感词典中的一个正向情感种子词与一个负向情感种子词。

对于得到的情感倾向值，可以应用程度副词和否定词对其进行修正，经过修正后$Q(W)$的计算公式为

$$Q(W)=M_n\times M_a\times[\frac{1}{n}\sum_{i=1}^{n}S(W,P_i)-\frac{1}{m}\sum_{j=1}^{m}S(W,N_j)] \tag{6}$$

其中，M_n与M_a分别表示否定词权重与程度副词权重，$M_n=\prod_{i=1}^{n}N_i$，N_i为第i个否定词的极性权重，$M_a=\prod_{i=1}^{n}A_i$，A_i代表情感词典中第i个程度副词的权重。

对一个语句中多个情感词倾向值进行累加可以得到整个语句的情感倾向值$Q(W)$，而对构成微博文本的多条语句的情感倾向值求和可以得出微博文本的总体情感倾向值$Q(P_T)$，计算公式如下

$$Q(P_T)=\sum_{i=1}^{k}Q(W_i) \tag{7}$$

微博情感倾向$Q(P)$的最终计算公式由表情和文本的情感两部分组成，公式为

$$Q(P)=\lambda\times\frac{1}{n}\times\sum_{i=1}^{n}Q(p_{fi})+(1-\lambda)\times\sum_{k=1}^{n}[M_n\times M_a\times(\frac{1}{n}\sum_{i=1}^{n}S(W_k,P_i)-\frac{1}{m}\sum_{j=1}^{m}S(W_k,N_j))]$$

5　实验结果与分析(Experimental results and analysis)

实验目的是对测试集中的每条文本赋予一个情感倾向值来代表文本的褒贬意程度，文本的情感倾向值由其中包含的情感词的情感值相加得到。情感倾向值判断准确率=判断正确的文本数与测试集总文本数之比。实验数据来源于微博搜索与搜狗实验室提供的互联网语料库数据，样本集中共计含有微博文本5000余条，其中正向、负向、中性文本数量分别为1500条、2000条、1200条。对测试数据进行比对分析，同时考虑微博表情符号、程度副词和反向词的影响因素对文本进行加权处理，得到的实验结果如图2所示，P代表分析准确率。

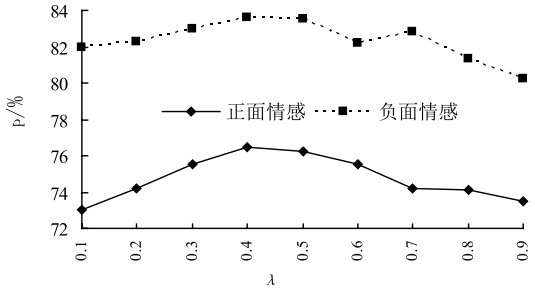

图2 正负面情感判断准确率实验结果

Fig.2 Experimental results of positive and negative emotion judgment accuracy

分析图中的折线数据中可以得出，λ使正面情感与负面情感的分析准确率表现均较好的区间为（0.4,0.5）。表3给出了当λ=0.4时程度副词与否定词等增强因素加权前后对正面情感，以及负面情感的分析准确率对照情况。可以看出，当λ取值区间为（±0.2,±0.8）时，加权后的判断准确率与加权之前均有提升，无论是正向情感还是负向情感，当λ=±0.4准确率达到最大值。超过±0.4后判断准确虽也有提升，但是幅度不及之前。同时，负面情感倾向的判断准确率要明显高于正面情感倾向的判断准确率，其主要原因可能是受情感字典中正向与负向词语数量不同和文本样本空间中正向与负向文本的比例不同的影响。其中Pqz、Phz分别为加权修正前后正面情感判断准确率，Pqf、Phf分别为判断准确率以及加权修正后负面情感判断准确率。

表3 加权前后微博情感分析准确率对照表

Tab.3 The contrast table of the emotion judgment accuracy of micro-blog before and after the neutral section

中性区间	Pqz/%	Phz/%	Pqf/%	Phf/%
±0.2	73.1	75.6	78.8	82.5
±0.3	76.2	78.3	81.7	83.2
±0.4	75.3	79.6	80.6	85.2
±0.5	71.8	74.7	79.2	82.1
±0.6	71.3	72.7	76.9	77.9
±0.7	69.6	72.2	72.8	74.3
±0.8	67.9	71.6	71.3	73.7

6　结论(Conclusion)

本文方法通过对NTUSD、HowNet、情感词汇本体库和中文褒贬意词典进行整理，基于HowNet的义原情感判别，加入程度副词和表情，以及否定词对文本情感的影响。实验结果显示程度副词与表情倾向对微博文本情感倾向起着至关重要的作用，当λ参数取值一定时，情感值倾向判断的准确率会

（下转第29页）