

基于多部情感词典和规则集的中文微博情感分析研究

吴杰胜 陆奎

(安徽理工大学计算机科学与工程学院 安徽 淮南 232001)

摘要 微博情感分析是对微博文本情感极性的判断并实现微博消息分类,可以对网络舆情进行及时有效的决策。做好微博情感分析的关键点是在原有的基础上更加准确地分析出每条微博文本的情感极性,因此以此为目标对微博进行情感分析。对情感词典进行改进与扩充,主要包括构造程度副词、否定词词典、微博领域词典等相关词典。同时将文本之间的语义规则集考虑到情感分析中,主要涵盖了句间分析规则和句型分析规则。多部情感词典和规则集相结合的方式实现了对微博的情感分析。实验结果证明了该方法对微博情感分析有一定的作用。

关键词 微博 情感词典 规则集 情感分析

中图分类号 TP391

文献标识码 A

DOI:10.3969/j.issn.1000-386x.2019.09.017

CHINESE WEIBO SENTIMENT ANALYSIS BASED ON MULTIPLE SENTIMENT LEXICONS AND RULE SETS

Wu Jiasheng Lu Kui

(College of Computer Science and Engineering, Anhui University of Science and Technology, Huainan 232001, Anhui, China)

Abstract Weibo sentiment analysis is the judgment of the emotional polarity of Weibo text and realizes the classification of Weibo messages. This research can make timely and effective decision on network public opinion. The key point of doing a good analysis of Weibo sentiment is to analyze the emotional polarity of each Weibo text more accurately on the basis of the original. Therefore, this paper aimed to analyze the sentiment of Weibo. The Emotional Dictionary was improved and expanded. It mainly included related dictionaries such as constructive degree adverbs, negative word dictionaries and microblog domain dictionaries. The semantic rule sets between texts were considered in sentiment analysis, mainly covering inter-sentence analysis rules and sentence analysis rules. The combination of multiple emotional dictionaries and rule sets finally realized the sentiment analysis of Weibo. The experimental results show that this method has a certain effect on the analysis of Weibo sentiment.

Keywords Weibo Sentiment dictionary Rule set Sentiment analysis

0 引言

微博是近些年来一个新生的适用于大众的社交媒体平台,随着移动互联网的普及,大众对微博的使用率越来越高,微博也得以快速发展。广大的用户群体都可以通过微博来发表自己对当前的一些热点话题的看法,所以他们每天都在提供海量且丰富的观点文本数据,而这些数据中包含着很多情感信息。如何充分挖

掘情感信息并进行分析就是情感分析。情感分析在当今的研究很广泛,提取情感信息对社会发展起到一定的作用,而微博除了作为一个社交媒体平台之外,还具有其他特性,因此对微博的情感分析研究至关重要。

目前国内外都在对微博进行研究,但中文微博和英文微博的研究进展差距很大,英文微博的研究成熟度高于中文微博,而且中文微博与英文微博的特性几乎不同,因此如何能利用中文微博情感信息来进行研究分析是我们现在要做的工作。本文利用多部情感词

典和中文语义规则集相结合的方式判断中文微博的情感极性。

1 相关工作

文献[1]中指出情感即文本作者的意见和观点,因此对情感的分析也可以理解为对意见的挖掘,文本意见挖掘属于数据挖掘的子类,主要是利用现有的计算机技术挖掘出蕴含在文本间的观点、情绪等元素。在当今可以通过构造相应的情感词典和利用机器学习算法来对微博文本进行情感分析、极性分类。构造情感词典来对微博进行情感分析出现比较早,而且它对微博文本这种细粒度的情感分析效果极佳。文献[2]就是在基础情感词典的基础上,构造了两种计算词汇语义的情感权值方法。文献[3]也在基础情感词典的基础上,构造了一种分类器,可以对文本语义之间的歧义进行消除,从而提高情感分析准确率。

基于机器学习的方法来进行情感分析,主要是通过选取一些特征来标注训练集和测试集,接着利用朴素贝叶斯、支持向量机等分类器进行情感分类。文献[5]利用支持向量机或朴素贝叶斯与支持向量机相结合的方法对微博进行情感分析。文献[7]首先构造微博语料库,再用朴素贝叶斯算法进行分类。

总之,微博情感分析常用的两种方法都有一定的作用,但谁也不能做到更高的准确率,只能在这个基础上不断地加以改进方法提高准确性。基于情感词典的方法擅长处理细粒度的文本情感分析,因此本文主要也是利用情感词典,在此基础上加以改进,并结合文本之间的语义规则集来对微博进行情感分析,最后通过各个部分的情感权值加权求和得到微博的情感极性。微博的整体情感分析流程图如图1所示。

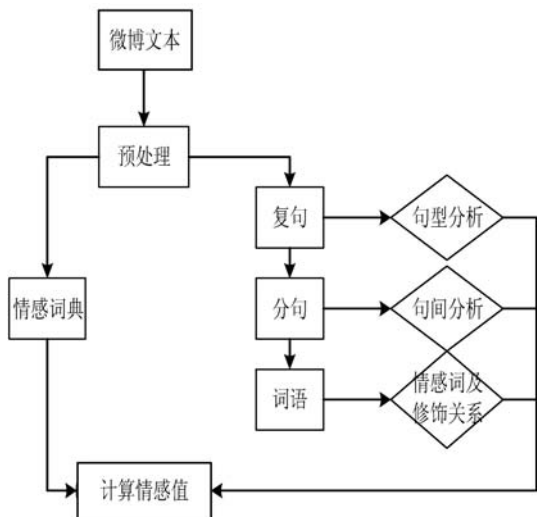


图1 微博整体情感分析流程

2 情感词典的构建

目前国外的情感词典《General Inquirer》完善度很高,但在国内还没有一部这样比较完善的词典,所以对微博来说,有一部完善的情感词典是很有必要的。现在国内使用常见的代表性情感词典有知网 HowNet 情感词典,台湾大学的正、负面情感词典和大连理工大学中文情感词典库等等。所以本文在此基础词典的基础上进行整合和优化,构建一个扩展的多部情感词典,同时还需要单独构建一个微博特定领域的情感词典来一起组成微博情感词典,从而进行微博情感分析。

2.1 微博文本的预处理

微博文本具有元素多样性、随意性、口语化等特点,所以需要进行预处理。预处理步骤如下:

1) 将网页中的链接、图片、视频、动画删除;将“@+用户名”删除;将“#话题#”删除。这些内容虽对微博情感分析有一定作用,但是影响不大,可以删除。

2) 将文本中的繁体字、英文等其他语言都翻译成中文,这是为了后续工作的方便,可使用特定的工具来进行翻译。

3) 保留微博文本中的表情符号。因为表情是情感状态的外在表现,与情感有关,可以参与情感权值计算。

4) 分词,本文使用中科院 ICTCLAS 软件进行分词与词性标注。

5) 删除停用词,比如助词“的”,代词“她”、“他”等之类的词。

在预处理完成之后,微博文本就是词语连接成串的形式,比如“我国运动员武大靖在短道速滑男子500米决赛中夺冠。”就会变为{我国,运动员,武,大靖,在,短道速滑,男子,500,米,决赛,中,夺冠}。

2.2 构建多部情感词典

目前中文情感词典还没有完整成熟的情感词典,所以除了构造基础情感词典外,还有否定词词典和双重否定词词典、程度副词词典、关系连词词典、表情符号词典。

2.2.1 基础情感词典

基础情感词典是取自大连理工大学的中文情感词典库。这个词典库将情感词分成了五个强度和三类词。本文用数字1表示正面词,数字2表示反面词,0表示中性词且它的权值为0。示例如表1所示。

表1 基础情感词典示例

情感词	词性种类	权值	极性
绝望	形容词	9	2
瑞雪	名词	5	1
开心	形容词	5	1
数落	动词	0	0

2.2.2 否定词词典和双重否定词词典

否定词词典包括否定副词和反问词这两部分。文献[10]中指出否定副词和反问词修饰情感词时,都会改变词的情感极性,但反问词语气更强,而双重否定不会改变词的情感极性,但是语气会更加强烈。通过人工筛选共获取25个否定词,示例如表2所示。

表2 否定词词典和双重否定词词典示例

词语类型	词语	权值
否定词	不、没、无、否、……	-1
反问词	难道、难不成、岂、……	-2
双重否定词	不是不、绝非不、……	1

2.2.3 程度副词词典

程度副词词典来自于知网词典库。将这些词一共分为6个等级。等级分别是超、最、很、较、稍、欠。分别对这6个等级给予一定的权值,对所修饰的情感词的情感强度扩大一定的倍数。示例如表3所示。

表3 程度副词词典示例

等级	副词	权重倍数	个数
超	超、过度、忒、……	3	30
最	百分百、极度、过于、……	2.5	69
很	何等、不过、太、……	2	42
较	那么、大不了、更、……	1.5	37
稍	稍微、略微、稍稍、……	1	29
欠	不那么、弱、不甚、……	0.5	12

2.2.4 关系连词词典

关系连词主要有转折、让步、递进、因果、假设等关系,它们在句子与句子之间的连接起到作用。本文收集整理常用的一些词构建了一个关系连词词典,并赋予一定的权值,示例如表4所示。

表4 关系连词词典示例

词性	词语	权值	个数
转折	但是、然而、而、……	0.5	10
让步	虽然、尽管、即使、……	1.5	10
递进	甚至、并且、况且、……	2	9
因果	因此、所以、以便、……	1.5	14
假设	如果、倘若、若、……	1	12

2.2.5 表情符号词典

微博表情在微博文本中具有很强的情感倾向性,可以通过它去判断微博情感极性有一定的作用。本文通过微博抓取了一些频率使用比较高的部分表情构造表情词典,共计217个表情。示例如表5所示。

表5 表情符号词典示例

表情符号	权值	个数
 ……	2	33
 ……	1	42
 ……	0	77
 ……	-1	33
 ……	-2	32

2.3 微博领域情感词典的构建

由于基础的情感词典还不完整,对情感词的概括是有限的,所以还需要针对微博上一些特有的情感新词进行识别,从而对这些新词集合构建一个词典。首先要基于统计信息来识别新词,然后在新词中进行情感识别。

2.3.1 基于统计信息的新词识别

文献[6]中给出三个定义,分别称作字串频数、内部耦合度、邻字集信息熵,一个字串能否成词与这三个定义有关。微博文本是由一连串词语组成的文本,首先我们用一个长字串来表示微博文本,同时将一个新词的成词长度设定为一个值,本文设定为7。同时再考虑上面三个定义,它们每个都要设定一个参数阈值,如果有任何一个条件不满足,即超过阈值范围,则这个字串不是一个词。最后剩下的能构成的词语集合中,仍需要比对情感词典中的词语,若该词在已有的词典中找不到,即成为新词。

2.3.2 新词情感分析与 PMI 算法改进

通过以上方法能识别并挖掘出新词,但是对这些词的情感极性还需要继续识别,从而构建出一个微博特定领域的情感词典。首先根据以上方法识别出新词,按照词频进行统计并排序,按照从上到下的方式来筛选,筛选出情感极性较强而且词频比较高的词语作为种子词。然后对这些词的情感极性作出判断,紧接着利用 PMI 算法计算其他未知词与它们之间的语义相似度,最后计算未知新词的情感极性,方法如下:

点互信息主要是可以计算词与词之间的相似度。

两个词 w_1 和 w_2 之间的相似度计算公式为:

$$PMI(w_1, w_2) = \log_2 \frac{p(w_1, w_2)}{p(w_1) \times p(w_2)} \quad (1)$$

式中: $P(w_1, w_2)$ 表示 w_1, w_2 共同出现的概率, $p(w_1)$ 、 $p(w_2)$ 分别表示 w_1, w_2 单独出现的概率。

w_1 表示未知词, w_2 表示种子词, 若式(1)的计算结果较大即相似度高, 则可知两个词情感极性相同, 否则就不同。但仅仅计算一对词的语义相似度在微博情感分析中不具有说服力, 所以本文在考虑这个的基础上, 在词阈的范围内选取了 30 对正负面情感极性的种子词, 同时考虑到使用频率高的表情元素, 选取了 5 对正负面情感极性表情符号作为种子词, 一起构成正面的情感词集合 W_p 和负面情感词集合 W_n , 用来考察多词之间的语义相似度。同时对 PMI 公式进行改进, 得出新词 w 的情感极性判断的新公式:

$$Sen_PMI(w) = \sum_{w_p \in W_p} PMI(w, w_p) - \sum_{w_n \in W_n} PMI(w, w_n) \quad (2)$$

式(2)的值如果大于 0, 则新词 w 的情感极性为正面; 等于 0, 新词 w 的情感极性为中性; 小于 0, 新词 w 的情感极性为负面。

最后一起构建成微博特定领域的情感词典, 本文识别并挖掘出 2018 年微博新词共计 164 个, 将这些词分为 4 个级别, 并赋予一定权值, 示例如表 6 所示。

表 6 微博新词词典示例

微博新词	权值	个数
真香、skr、锦鲤、……	2	18
官宣、佛系、确认过眼神、……	1	40
安排、凉凉、大猪蹄子、……	-1	65
坑爹、尼玛、中年油腻男、……	-2	41

3 微博文本规则集的情感分析

微博文本也是普通文本, 都是由汉字等其他元素构成的表达文本, 而文本之间肯定存在着一些语法关系和语义规则, 它们对文本的情感分析也有一定作用。

3.1 句间分析规则

一条微博文本可以通过标点符号划分成若干个复句, 一条复句可以分成若干个分句, 句间分析规则就是考虑分句与分句之间的关系, 而句间关系主要有三类: 转折、递进、假设。这里用 S 表示整个复句, S_i 表示复句的各个分句。定义集合 $\{S_1, S_2, \dots, S_i\}$ 为复句的分句集合, R_i 表示句间规则对分句 S_i 的情感权值。

3.1.1 转折关系规则

转折关系中, 基本都会实现前后的情感翻转作用, 转折之前的分句情感会变弱, 而主要突出后面分句的情感, 后面分句与前面分句的情感极性相反。规则定义如下:

1) 若复句 S 中只有单一的转折后接词出现(如“但”, “可是”, “却”等)在分句 S_i 中, 则 S_i 之前的分句权值 R_i 都设为 0, S_i 之后的分句权值 R_i 都设为 1。

2) 若复句 S 中只有单一的转折前接词出现(如“虽然”, “如”, “尽管”等)在分句 S_i 中, 则 S_i 之前的分句权值 R_i 都设为 1, S_i 之后的分句权值 R_i 都设为 0。

3) 若复句 S 中出现成对的转折连接词(如“虽然…但是…”等), 且转折后接词出现在分句 S_i 中, 则 S_i 之前的分句权值 R_i 都设为 0, S_i 之后的分句权值都 R_i 设为 1。

3.1.2 递进关系规则

递进关系, 顾名思义, 在这个关系规则中, 复句的每个分句根据从前到后的顺序逐渐增强情感。规则定义如下:

若复句 S 中出现递进关系的连接词(如“更”, “更加”, “更重要的是”等), 则分句的权值为:

$$R_i = 1 \quad R_{i+1} = 1.5 \quad \dots \quad R_j = 1 + 0.5 \times (j - i)$$

3.1.3 假设关系规则

假设关系建立在现实情况中的一种设想, 它表达的情感主要在假设复句的前半分句, 而对后半分句的情感相对弱化一些。比如: 如果 A, 那么 B。则句子强调的是内容 A。

1) 若复句 S 中未出现否定的假设连接词, 但是出现假设关系的后接词(如“那么”), 且假设后接词出现在分句 S_i 中, 则 S_i 之前的分句权值 R_i 都设为 1, S_i 之后的分句权值 R_i 都设为 0.5。

2) 若复句 S 中出现否定的假设连接词, 而且假设后接词(如“那么”)出现在分句 S_i 中, 则 S_i 之前的分句权值 R_i 都设为 -1, S_i 之后的分句权值 R_i 都设为 -0.5。

上面描述的这三种句间关系都能影响到整个微博文本的情感极性, 所以情感分析中要考虑到它们。至于其他的句间关系如因果、并列等, 对情感分析的影响可以忽略不计。

3.2 句型分析规则

上一节所说的是复句的分句之间的关系, 这一节说明的是复句的句型对整个文本的情感极性的影响。本文主要讨论陈述句、疑问句、反问句和感叹句这四类常见句型。它们常以“?”、“!”、“。”等标点符号结尾。一个文本用 D 来表示, 则文本分割成各个分句即复句,

用集合定义为 $\{D_1, D_2, \dots, D_i, \dots, D_n\}$ 。复句用 D_i 来表示,定义 T_i 为句型规则对复句 D_i 的情感权值。具体的规则定义如下:

1) 如果微博文本中有复句 D_i 以感叹号“!”结尾,则表示此复句为感叹句,它的权值 T_i 设为 1.5。

2) 如果微博文本中有复句 D_i 以反问号“?”结尾且结尾处有反问标志词或者没有以反问号“?”结尾但有反问标志词,则表示此复句为反问句,它的权值 T_i 设为 -1。

3) 如果微博文本中有复句 D_i 以反问号“?”结尾且结尾处无反问标志词,则表示此复句为疑问句,它的权值 T_i 设为 0。

4) 如果微博文本中有复句 D_i 以句号“。”等其他标点符号结尾,则表示此复句为陈述句,它的权值 T_i 设为 1。

4 微博综合情感计算

本文基于多部情感词典和规则集的微博情感分析,对微博从词到句进行整体综合情感计算。用 D 表示整个文本,文本中各个复句用 D_i 表示; S 对应一个复句 S_i 表示复句中的各个分句; E 表示情感权值, R_i 表示分句的句间关系规则情感权值, T_i 表示复句的句型关系规则情感权值, sen_i 表示词典匹配得到的权值。

1) 词语情感值 $E(W_i)$ 计算公式为:

$$E(W_i) = N \times A \times sen_i \quad (3)$$

式中: N 表示情感词前对应的否定词或者双重否定词, A 表示情感词前对应的程度副词, sen_i 表示情感词与词典匹配得到的权值, W_i 表示情感词语。

词语的情感权值计算不仅与它自身的权值有关,还与在其前面修饰的程度副词、否定词有关,所以在情感权值计算时要将它们考虑进去。

2) 分句情感值 $E(S_i)$ 计算公式为:

$$E(S_i) = \sum_{i=1}^n E(W_i) \times R_i \quad (4)$$

3) 复句情感值 $E(D_i)$ 计算公式为:

$$E(D_i) = \sum_{i=1}^n E(S_i) \times T_i \quad (5)$$

4) 文本情感值 E 的计算公式为:

$$E = \sum_{i=1}^n E(D_i) \quad (6)$$

5) 表情情感值 E_m 计算公式为:

$$E_m = \frac{1}{n} \sum_{i=1}^n sen_i \quad (7)$$

6) 微博情感值 E_{last} 计算公式为:

$$E_{last} = m \times E + n \times E_m \quad (8)$$

式(8)表示微博的最终情感值计算, m 和 n 表示文本情感值和表情情感值在微博情感权值计算中所占分量的大小,本文根据文献[9]中分析分别设置为 0.6 和 0.4,计算得出 E_{last} 的大小。如果 E_{last} 大于 0,则表示此微博的情感倾向为正面的,如果 E_{last} 小于 0,则表示此微博的情感倾向为负面的,如果 E_{last} 等于 0,则表示此微博情感为中性的。

5 微博情感分析实验

5.1 实验方法

首先通过爬虫工具爬取了微博上两个相关的微博话题,然后对这些数据进行情感分析,具体的实验步骤如下:

1) 获取实验数据。利用爬虫软件爬取微博上比较两个热门话题“#短视频整顿#”和“#《我不是药神》爆红引社会热议#”的文本数据。

2) 情感极性的人工标注。获取数据的情感极性没有进行标注,采用人工方法对这两个话题进行标注。人工标注主要是通过统计抽取随机选择三名实验同学对这两个话题进行主观判断,标注情感极性,最后统计结果。

3) 预处理。根据上述对应的方法构建六部情感词典。

4) 话题情感分析。分别在一部基础情感词典、六部情感词典和基于六部情感词典与规则集的基础之上对这两个话题进行三组实验,得出微博的情感分析结果。

5.2 实验数据

本文通过爬虫软件爬取到关于两个微博话题的数据集,接着利用人工标注的方法,将这些文本进行情感极性标注,给出每条微博的情感权值并进行分类。共筛选出话题“#短视频整顿#”共计 25 720 条,其中正面数据 18 634 条,负面数据 1 385 条,中性数据 5 701 条;话题“#《我不是药神》爆红引社会热议#”共计 17 695 条,其中正面数据 10 672 条,负面数据 2 856 条,中性数据 4 167 条。判断标准是:微博情感权值大于 0 为正面,小于 0 为负面,等于 0 为中性。从筛选结果可知正面微博数据所占比例较大,负面微博数据和中性微博数据所占比例较小,且数据较少。

5.3 实验性能评估指标

本实验根据本文提出的微博情感分析方法对每一

条微博文本进行情感分析,然后将在此方法下自动分析得出的结果与我们人工分类得出的结果进行对比,看情感分析的效果如何。采用以下三个指标进行分析,分别是正确率 P 、召回率 R 和综合度量 F 指标值,具体公式如下:

$$P = \frac{\text{判断正确的该类别微博数}}{\text{判断为该类别的微博数}} \quad (9)$$

$$R = \frac{\text{判断正确的该类别微博数}}{\text{应判断为该类别的数目}} \quad (10)$$

$$F = \frac{2 \times P \times R}{P + R} \quad (11)$$

5.4 实验分析与结果

为了验证本文提出的方法具有更好的作用,还另外做了只基于一部情感词典和只基于六部情感词典的实验。将本文提出的方法实验结果与这两种方法得出的实验结果进行对比,利用性能评估指标对结果进行分析。

对两个话题分别做如下三组实验:

第一组实验:分别对话题“#短视频整顿#”和“#《我不是药神》爆红引社会热议#”采用基于一部基础情感词典的微博情感分析,并进行微博分类。

第二组实验:分别对话题“#短视频整顿#”和“#《我不是药神》爆红引社会热议#”采用基于六部基础情感词典的微博情感分析,并进行微博分类。

第三组实验:分别对话题“#短视频整顿#”和“#《我不是药神》爆红引社会热议#”采用基于六部基础情感词典和规则集的微博情感分析,并进行微博分类。

实验结果如表 7 和表 8 所示。

表 7 #短视频整顿#实验结果

实验方法	类别	P	R	F
一部基础情感词典	正面	0.648	0.643	0.645
	负面	0.625	0.598	0.611
	中性	0.607	0.667	0.636
	平均	0.627	0.636	0.631
六部情感词典	正面	0.782	0.778	0.780
	负面	0.754	0.703	0.728
	中性	0.736	0.795	0.764
	平均	0.757	0.759	0.757
六部情感词典 + 规则集	正面	0.849	0.843	0.846
	负面	0.818	0.785	0.801
	中性	0.797	0.853	0.824
	平均	0.821	0.827	0.834

表 8 #《我不是药神》爆红引社会热议# 实验结果

实验方法	类别	P	R	F
一部基础情感词典	正面	0.657	0.651	0.654
	负面	0.586	0.545	0.565
	中性	0.546	0.594	0.569
	平均	0.596	0.597	0.596
六部情感词典	正面	0.745	0.736	0.741
	负面	0.671	0.619	0.645
	中性	0.639	0.715	0.675
	平均	0.685	0.690	0.687
六部情感词典 + 规则集	正面	0.806	0.785	0.795
	负面	0.733	0.653	0.691
	中性	0.701	0.756	0.727
	平均	0.747	0.731	0.738

通过表 7 和表 8 的数据,对实验结果进行如下分析:

1) 实验结果表明本文提出的方法提高了微博的情感分析的正确率。若只单纯靠一部基础情感词典,那么正确率是较低的,因为微博的特殊文本包含了很多普通文本不具有的特性,所以要在原来的基础上扩建多部情感词典,提高词典的覆盖面,同时将文本语义规则集考虑进去,更有利于微博的情感分析。

2) 通过两个话题的实验结果可以看出,话题“#短视频整顿#”的正确率高于话题“#《我不是药神》爆红引发社会热议#”的正确率。这是因为前者所获取的正面数据居多,而且对后者话题中一些判断失误的微博文本进行分析发现这是一部关于电影反讽刺的话题,有网友发表微博就使用了一些反讽刺的表达。比如“电影中的药商真的好棒啊,竟然可以把药卖给病人,真的是好样的!”,这其中“好棒”“好样”都是正面情感词,但实际上是起到讽刺作用,是负面的微博,因此在后续对微博的情感分析中还可以继续对语义规则进行完善分析。

3) 通过表格中数据发现正确率和 F 值都是正面微博偏高,通过微博分析得知是由于正面、负面、中性数据分布不平衡造成的,因为这两个微博都是社会热点话题,众多网友支持支持态度。

4) 通过对比 F 值可以发现在引入六部情感词典之后, F 值相对于一部情感词典下有很大提高,这是因为在六部情感词典下,匹配微博文本的面更广,尤其加入了微博特定领域的情感词典,而且在加入规则集以后, F 值又有了一定的提升。虽然 F 值总体上提高了,但还可以继续提高,因为实验预处理过程中有个分词

过程,还有语义规则的分析过程,这两个过程的优劣程度都会影响最后结果。当然还有一些其他因素,比如一词多义现象等。

实验表明,本文提出的方法利用多部情感词典,并考虑文本语义规则集,对微博的情感分析效果有明显的提升,且在三个指标下,都验证了此方法对微博情感分析有效果。

6 结 语

基于词典的情感分析是已有的研究方法,本文在基于词典的基础上,构建了除基础情感词典之外的其他五部词典,这些词典范围更广,其中微博特定领域的情感词典构造至关重要,未来还需要继续不断完善这部词典。最后在六部词典的基础上,考虑文本之间的语义规则,因此提出一种基于多部情感词典和规则集的中文微博情感分析方法,通过实验验证了此方法具有很好的作用。

微博的情感分析研究还有很多可以改进之处,比如要考虑微博的点赞数、转发数和阅读数等。我们将继续改进方法,力争使中文微博情感分析更上一个台阶。

参 考 文 献

- [1] 杨立公,朱俭,汤世平. 文本情感分析综述[J]. 计算机应用,2013,33(6):1574-1578,1607.
- [2] 朱嫒岚,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.
- [3] Jose R, Chooralil V S. Prediction of election result by enhanced sentiment analysis on Twitter data using Word Sense Disambiguation [C]//International Conference on Control Communication&Computing India, 2015: 638-641.
- [4] Park S, Kim Y. Building thesaurus lexicon using dictionary-based approach for sentiment classification[C]//2016 IEEE 14th International Conference on Software Engineering Research, Management and Applications(SERA). IEEE,2016: 39-44.
- [5] 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.
- [6] 王志涛,於志文,郭斌,等. 基于词典和规则集的中文微博情感分析[J]. 计算机工程与应用,2015,51(8):218-225.
- [7] Shahheidari S, Dong H, Bin Daud M N. Twitter sentiment mining: A multi domain analysis [C]//Proceedings of the 2013 Seventh International Conference on Complex, Intelligent, and Software Intensive Systems. IEEE, 2013: 144-149.
- [8] 姜杰,夏睿. 机器学习与语义规则融合的微博情感分类方法[J]. 北京大学学报(自然科学版),2017,53(2):247-254.
- [9] 赵天奇,姚海鹏,方超,等. 语义规则与表情加权融合的微博情感分析方法[J]. 重庆邮电大学学报(自然科学版),2016,28(4):503-510.
- [10] 陈国兰. 基于情感词典与语义规则的微博情感分析[J]. 情报探索,2016,18(2):1-6.
- [11] Liu B. Sentiment Analysis and Opinion Mining[J]. Synthesis Lectures on Human Language Technologies, 2016, 30(1): 152-153.
- [12] Zou X, Yang J, Zhang J, et al. Microblog Sentiment Analysis with Weak Dependency Connections [J]. Knowledge-Based Systems, 2017, 142:170-180.
- [13] Zhang S X, Wang Y, Zhang S Y, et al. Building associated semantic representation model for the ultra-short microblog text jumping in big data[J]. Cluster Computing, 2016, 19(3): 1399-1410.
- [14] Ouyang Y, Guo B, Zhang J, et al. SentiStory: multi-grained sentiment analysis and event summarization with crowdsourced social media data[J]. Personal and Ubiquitous Computing, 2017, 21(1):97-111.
- [15] Araque O, Corcuera-Platas I, Sánchez-Rada, J. Fernando, et al. Enhancing deep learning sentiment analysis with ensemble techniques in social applications[J]. Expert Systems with Applications, 2017, 77:236-246.
- [16] Zhang S X, Wei Z L, Wang Y, et al. Sentiment analysis of Chinese micro-blog text based on extended sentiment dictionary[J]. Future Generation Computer Systems, 2018, 81:395-403.

(上接第57页)

- [8] Dimovski T, Mitrevski P. On the Performance Potential of Connection Fault-Tolerant Commit Processing in Mobile Environment[J]. International Journal of Wireless & Mobile Networks, 2012, 4(5):29-44.
- [9] Stuedi P, Mohamed I, Terry D. WhereStore: Location-Based Data Storage for Mobile Devices Interacting with the Cloud [C]//Proceedings of the 1st ACM Workshop on Mobile Cloud Computing & Services: Social Networks and Beyond. ACM, 2010: 137-143.
- [10] 严蔚敏,李冬梅,吴伟民. 数据结构(C语言版)[M]. 北京:清华大学出版社,2011.
- [11] 施伯乐,丁宝康,汪卫. 数据库系统教程[M]. 北京:高等教育出版社,2008.
- [12] 杜金莲. 高级数据库技术[M]. 北京:清华大学出版社,2010.
- [13] 王元元. 离散数学教程[M]. 北京:高等教育出版社,2010.