



引用格式:朱颢东,李雯琦. 基于语义规则和表情加权的中文微博情感分析方法[J]. 轻工学报,2020,35(2):74-82.

中图分类号:TP301 文献标识码:A

DOI:10.12187/2020.02.010

文章编号:2096-1553(2020)02-0074-09

基于语义规则和表情加权的 中文微博情感分析方法

Chinese micro-blog emotional analysis method based on semantic rules and expression weighting

朱颢东,李雯琦

ZHU Haodong, LI Wenqi

郑州轻工业大学 计算机与通信工程学院 河南 郑州 450001

School of Computer and Communication Engineering, Zhengzhou University of Light Industry, Zhengzhou 450001, China

关键词:

微博情感;表情符;
情感词;语义规则

Key words:

micro-blog emotion;
emoticon; emotional
word; semantic rules

摘要:针对目前中文微博情感分析方法考虑因素不全面,从而导致情感分析结果欠佳的问题,提出一种基于语义规则和表情加权的中文微博情感分析方法。该方法在使用传统情感词典分析中文微博情感倾向的基础上,在普通情感词典中融入否定词、程度副词和网络新词,根据中文微博文本独有的语言特点和句式特点,采用从词语到分句再到复句的方式对整个中文微博进行情感分析,进而使用表情加权和语义规则进行权值求和,以确定情感倾向。实验结果表明,较另外3种中文微博情感分析方法,该方法效果更显著,其平均准确率为78.4%,平均查全率为75.2%,平均F值为76.7%。

收稿日期:2019-10-08

基金项目:河南省高等学校重点科研项目(19A520009)

作者简介:朱颢东(1980—),男,河南省虞城县人,郑州轻工业大学副教授,博士,主要研究方向为智能信息处理与智能计算。

Abstract: Aiming at the problem that the current Chinese micro-blog emotional analysis methods were not comprehensive, which led to poor sentiment analysis results, a Chinese micro-blog emotional analysis method based on semantic rules and expression weighting was proposed. On the basis of using traditional emotion dictionary to analyze the emotion tendency of Chinese micro-blog, negative words, degree adverbs and network neologisms were incorporated into the general emotion dictionary. According to the unique language characteristics and sentence pattern characteristics of Chinese micro-blog text, the method of emotional analysis from words to clauses and then to complex sentences was adopted to analyze the whole Chinese micro-blog. Expression weighting and semantic rules were used to perform weight summation to determine emotional tendency. The experimental results showed that compared with the other three Chinese micro-blog emotional analysis methods, the proposed method was more effective. It had an average precision rate of 78.4%, an average recall rate of 75.2%, and an average F value of 76.7%.

0 引言

伴随着互联网和社交媒体的不断发展及广泛应用,其所承载和产生的数据信息迅速增长.微博中承载着海量有价值的信息,如何对其进行合理的分析和利用,受到了国内外学者的广泛关注.用户通过微博平台可以随时随地获取信息、表达情感,海量的微博信息包含着用户对事件、产品、政策等的态度、意见和评价,对这些信息进行情感趋向分析,能够很好地挖掘网络群体的行为规律,获得潜在信息.因而,如何快速准确地分析微博情感倾向是当前研究的热点之一.

当前微博情感分析的主流方法为基于语义的分析方法,该方法通过情感词典统计出微博文本中情感词的情感权值,对其进行相关运算,计算出整个语句或文本的情感值,由此判断微博的情感倾向.

国外的网络社交媒体文本情感分析研究主要针对 Twitter 数据. Y. Yu^[1]提出了针对 Twitter 的情感分析方法,将从 Twitter 中提取的词语与情感词典中被标注为正向或负向的情感词语进行比较,而后计算出文本的情感值. A. Pak 等^[2]在标注 Twitter 文本情感极性数据集的基础上,设计实现了基于朴素贝叶斯算法、支持向量机和条件随机场的情感分类器. E. Riloff 等^[3]利

用人工定制模板,选取种子情感词语,运用迭代法获取了名词词性的情感词语,以此对微博进行情感倾向分析.

由于国内外语言结构和表达方式不同,中文微博的情感分析比英文微博的情感分析要复杂得多.国内的网络社交媒体文本情感分析研究主要针对新浪微博数据展开.朱嫣岚等^[4]基于 HowNet 中情感词汇的情感信息,通过语义相似度和语义相关场来计算词语情感倾向,以提高情感分析的准确率.谭皓等^[5]引进表情符和注意力机制,将表情符和词语同时训练,使用表情加权与普通情感词相结合的方式进行情感分类.李继东^[6]结合普通情感词典和语义规则计算感情值,对中文微博进行情感分析.王文等^[7]基于普通情感词典计算语义相似度,考虑正面情感增强因素,综合表情和语义规则获得情感倾向结果.刘志明等^[8]通过对比 3 种机器学习算法、3 种特征权重计算方法和 3 种特征选择算法后发现,使用支持向量机、词频-逆文本频率(TF-IDF)和信息增益(IG)选择特征项权重时,情感倾向分析效果最佳.但是上述研究对于影响中文微博情感倾向分析结果的因素考虑不够全面,导致情感倾向分析结果欠佳.

鉴于此,本文拟提出一种基于语义规则和表情加权的中文微博情感分析方法,根据中文微博特有的语言特点,在情感词典的基础上,融

入语义规则和表情符号,计算文本情感值以确定微博情感倾向,以期有效改善中文微博情感分析结果.

1 中文微博情感分析方法与流程

中文微博情感分析方法的流程一般分为以下4个步骤:

- 1) 通过微博爬虫和 API 接口获取微博语料信息;
- 2) 对获取的微博数据进行预处理,通过中文分词将微博文本中的句子划分为词语,并进行词性标注;
- 3) 对微博文本情感词和常用的表情符号进行抽取和极性标注;
- 4) 借助情感词典和语义规则计算情感值,对其进行加权处理,根据最终情感值将微博文本划分为正向、中性和负向3类.

中文微博情感分析流程如图1所示.

1.1 中文微博语料的收集和文本预处理

中文微博数据的获取方法主要有两种:一种是通过爬虫获取微博数据,将解析到的数据流内容转换成文本存入数据库,该方法可以收集大量的微博数据,但受限于访问速度,且需要下载大量的页面,因而效率不高;另一种是利用微博官

方 API 获取数据^[9],但如果仅使用 API 接口,所采集到的数据范围不够广泛,并且微博 API 开放接口会限制每段时间的请求次数,而过度或者恶意调用又会对微博平台性能产生不良影响.

因此,本文使用爬虫与 API 接口相结合的方式来获取微博数据,而后对其进行预处理.对没有价值的信息进行降噪处理,如删除微博中的广告、具有转发性质的微博、只含有表情符号的微博等.本文运用当前使用频率较高、性能较稳定、准确率较高的 jieba 组件(Python 版)对获取的微博文本进行分词处理,并分析词法,标注词性.

1.2 微博文本表情符号标注

表情符号直接影响着微博这种短文本的最终情感分析结果.中文微博中常常存在大量的表情符号,其中使用最为广泛的的就是微博中的默认表情符和输入法中自带的 emoji 表情符^[10].这些表情符大多具有明显的情感倾向,在输入表情符号时,会出现相对应的标注,可以利用它们进行明确的情感倾向划分.

对于表情符号的情感分析研究主要有3种方式:将文本中表情符的情感极性直接作为句子的情感极性,忽略文本信息对情感极性的影响;将表情符和情感词视为相同级别的元素,赋予相同的权值进行情感极性计算;将表情符和文本视为相同级别的元素,但使用不同的权值对微博情感极性进行计算^[11].本文选择第3种方式对微博的情感值进行计算.通过人工方式选择微博常用的表情符号,删除其中不具有情感倾向的表情符号.常用表情符情感倾向标注如表1所示.

除了对表情符号进行情感极性标注,本文对表情符号的情感强度也进行了相应的标注.利用王文等^[7]对正向和负向表情情感权值的定义,正向表情情感权值在(0,1]区间内取值,负向表情情感权值在[-1,0)内取值,部分表情符号及其权值如表2所示.

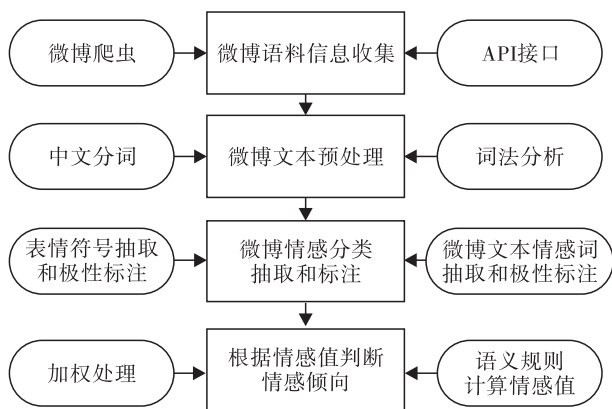


图1 中文微博情感分析流程

Fig.1 The process of Chinese micro-blog emotional analysis

表1 常用表情符情感倾向标注

Table 1 Emotionality tagging with common emoticons

















情感倾向	表情符
正向	
负向	

表2 部分表情符号及其权值

Table 2 Partial emoticons and their weights

情感倾向	表情符号	权值	情感倾向	表情符号	权值
正向		1.0	负向		-1.0
		0.9			-0.9
		0.8			-0.8
		0.7			-0.7
		0.6			-0.6
		0.5			-0.5
		0.4			-0.4

1.3 微博文本情感词的抽取和标注

微博文本情感词的抽取和标注离不开情感词典. 在情感分析领域, 目前使用较为广泛的情感词典有知网的 HowNet 情感词典、大连理工大学词汇本体库和台湾大学的中文情感词典^[12]. 三者各有优缺点, HowNet 情感词典和中文情感词典对正向和负向情感词进行了细致的划分, 但未标注情感词词性及其情感权值, 而词汇本体库虽然不如 HowNet 情感词典词汇丰富, 却对词性和情感权值都进行了标注. 因此, 本文将融合 HowNet 情感词典和词汇本体库来确定情感词的极性及其情感权值, 并在此基础上, 针对微博文本特点, 加入否定词词库、网络新词词库和程度副词词库, 以提高情感分析的

准确性.

在对网络新词进行词性标注时, 引入 PMI 算法, 通过计算基准词与网络新词之间的相似度来判断其情感极性, 具体公式为

$$PMI(word_1, word_2) = \log_2 \left(\frac{P(word_1 \& word_2)}{P(word_1) \times P(word_2)} \right)$$

式中, $P(word_1)$ 和 $P(word_2)$ 分别为词语 $word_1$ 和 $word_2$ 在微博文本中出现的概率, $P(word_1 \& word_2)$ 为词语 $word_1$ 和 $word_2$ 共同出现在微博文本中的概率. 以 PMI 为基础, 通过正向基准词集 P_{words} 和负向基准词集 N_{words} 来计算网络新词 $word_1$ 的情感倾向值:

$$SO-PMI(word_1) = \sum_{word_2 \in P_{words}} PMI(word_1, word_2) - \sum_{word_3 \in N_{words}} PMI(word_1, word_3)$$

若 $SO-PMI(word_1) > 0$, 则网络新词 $word_1$ 为正向情感词汇, 否则为负向情感词汇.

在对微博进行情感分析时, 应充分考虑到微博文本中程度副词和否定词对情感极性判定的影响. 程度副词会加重或减轻微博情感倾向的程度, 根据蒯璜等^[13]对程度副词的特点、范围和分类的研究, 将程度副词分为 4 个级别, 即极度、高度、中度和微度, 如果某个情感词语前出现多个程度副词, 只对其前面出现的 3 个以内程度副词进行权值求积, 超出 3 个的程度副

词按无效词处理. 否定词在情感分析中的作用更是不容小觑,若一个分句中存在两个否定词,双重否定则表示为否定之否定即肯定. 因此,本文将否定词的权值设置为负值,用累乘的方式来计算其情感权值^[14]. 部分网络新词、程度副词和否定词示例见表3—表5.

1.4 语义规则

中文的微博情感分析不同于英文微博,不仅要考虑情感词(表情符、否定词、网络新词和程度副词)的权值,还要考虑语义规则对整体情感分析的影响^[15]. 本文的语义规则主要包括句型关系语义规则和句间关系语义规则.

1.4.1 句型关系语义规则 在进行情感分析时,先根据标点符号将复句分割成若干个小的分句,再对每个分句进行分析. 用户在发布微博

表3 部分网络新词示例表

Table 3 Example table of partial network new words

网络新词	情感倾向	权值
给力	正向	1.6
不明觉厉	正向	1.3
正能量	正向	1.4
香菇	负向	-0.8
蓝瘦	负向	-0.6
亚历山大	负向	-0.9

表4 部分程度副词示例表

Table 4 Example table of partial degree adverbs

程度副词	级别	权值
极其、极端、完全、绝对、非常、万分、异常、相当、无比、极、过度、……	极度	2.0
格外、尤其、特别、很、太、十分、越来越、何等、颇为、分外、……	高度	1.5
怪、挺、更加、愈发、较为、还、更、较、足足、……	中度	0.9
几乎、稍微、略微、有点儿、稍稍、丝毫、不怎么、弱、……	微度	0.5

表5 部分否定词示例表

Table 5 Example table of partial negative words

否定词	权值
不、没有、没、别、崩、不用、不必、不要、非、……	-1
不得不、无非、未必不是、无不、……	1

时,常用的句型主要有陈述句、问句和感叹句. 陈述句对整体的情感影响不大;问句又分为反问句和疑问句,反问句会使整个句子的情感倾向从正向变为负向或者从负向变为正向,疑问句通常只是传递疑惑的态度,句子本身无情感变化;感叹句具有加强情感倾向的作用^[16]. 如果用 s_i 来表示整个句子的影响权值,那么这3种句型的影响权值如下:陈述句 $s_i = 1$;疑问句 $s_i = 0$,反问句 $s_i = -1.5$;赵天奇等^[17]研究发现,当感叹句 $s_i = 2$ 时合格率最大,本文根据目前中文微博的语言习惯,对其研究成果进行改进,如果感叹句的尾部只有1个或2个感叹号,则 $s_i = 2$,如果有3个或3个以上感叹号,则 $s_i = 3$.

1.4.2 句间关系语义规则 除句型关系外,句间关系也会对情感分析产生重要的影响. 在中文的日常表达中,复句存在众多句间关系,如并列关系、递进关系、选择关系、转折关系、假设关系、条件关系、因果关系等,本文仅对中文微博中出现频率较高的转折、递进和假设3种句间关系进行权值分析. 通常使用集合 $\{C_1, C_2, \dots, C_i, C_j, \dots, C_n\}$ 来表示整个复句,其中 C_i 用于表示复句中的分句, $S(C_i)$ 用于表示影响权值.

1) 转折关系影响权值:对于存在转折词“但是”“然而”“却”的句型,其情感表达的重点在后半句,因此其 $S(C_i) = 0, S(C_j) = 1$;对于只有一个转折前接词(如“尽管……”)的句型,其情感表达的重点在含有转折词的分句中,因此其 $S(C_i) = 1, S(C_j) = 0$.

2) 递进关系影响权值:递进关系常用的关联词有“不但……而且”“甚至”“更”等,带有关联词的句子所表达的情感更加强烈,所以其 $S(C_i) = 1, S(C_j) = 1.5$.

3) 假设关系影响权值:如果是形如“如果……那么”的肯定句型时,前面分句的情感倾向更强一些,则 $S(C_i) = 1, S(C_j) = 0.5$;如果是形如“如果不……就”的否定形式时,因为有

否定词,则 $S(C_i) = -1, S(C_j) = -0.5$.

1.5 本文方法流程

根据上述分析,在对中文微博文本进行情感分析时,首先应对整个中文微博文本进行词语划分,统计情感词,然后对网络新词、程度副词和否定词进行权值分析,将这些权值进行累乘计算,再融入表情符号和语义规则进行权值求和,以确定其情感倾向.本文提出的中文微博情感分析方法流程如图2所示.

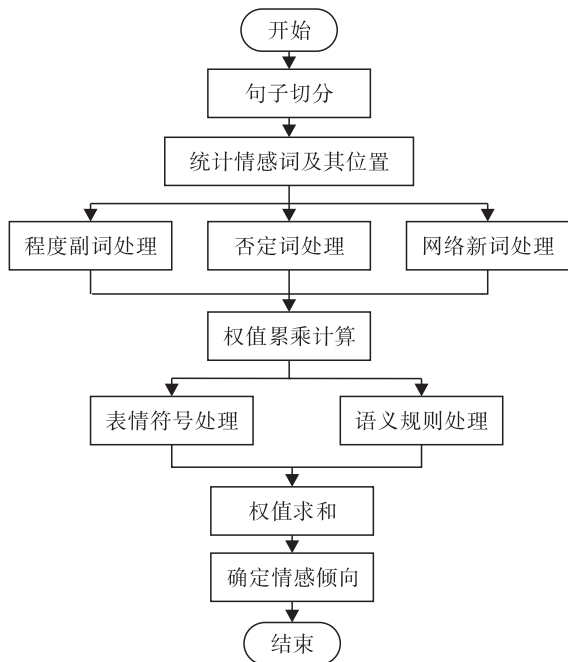


图2 中文微博情感分析方法流程图

Fig. 2 Flow chart of Chinese micro-blog emotional analysis method

在词语级别上采用权值累乘的方法,对情感倾向进行计算,用 $E(W_i)$ 来表示词语的情感权值,其计算公式为

$$E(W_i) = adv \times Neg \times new \times sen$$

其中, adv 表示用于修饰情感词的情感副词权值, Neg 表示否定词权值, new 表示网络新词权值, sen 表示情感词在情感词典中的情感权值.

整个句子的情感值需要考虑到语义规则对情感倾向的影响,首先是对于分句而言的句型关系,根据上文对句型关系的分析,以 $E(t_i)$ 来

表示融入句型关系的分句情感权值,其计算公式为

$$E(t_i) = \sum E(W_i) \times s_i$$

转折、递进和假设关系在复句中对情感分析有着重要的影响,对其情感权值进行求值时,要先对各个分句的情感权值进行求和,再乘以句型关系权值,以最终确定复句的情感权值,其计算公式为

$$E(T_i) = \sum_{j=1}^n E(t_j) \times S_i$$

其中, $\sum_{j=1}^n E(t_j)$ 用于计算各个分句的情感权值之和.

整个微博文本的情感权值用 E_{text} 来表示,其计算公式为

$$E_{text} = \sum E(T_i)$$

表情符号情感权值 E_{emo} 为文本中出现的所有表情符号权值的算术平均值,其计算公式为

$$E_{emo} = \frac{1}{n} \sum_{j=1}^n sen_{emo_j}$$

其中, sen_{emo_j} 表示微博表情符号的情感权值, n 为某条微博中表情的数量.

微博情感值 $E = \beta \times$ 微博表情情感权值 $+ (1 - \beta) \times$ 微博文本情感权值,本文选用准确率最高的 $\beta = 0.4$ ^[18],则微博情感值公式为

$$E = 0.4 \times E_{emo} + 0.6 \times E_{text}$$

其中, $E > 0.1$ 表示微博情感倾向为正向,值越大则正向情感越强烈; $E \in [-0.1, 1]$ 表示微博情感倾向为中性感; $E < -0.1$ 表示微博情感倾向为负向,值越小则负向情感越强烈.

2 实验结果与分析

2.1 实验数据与评价指标

为验证本文提出的情感分析方法的有效性,将其与文献[5-7]的3种方法进行对比实验.采用将网络爬虫与API接口相结合的方式从新浪微博中获取8000条微博数据,剔除无效

数据,经过筛选,有效微博为6834条,通过人工的方式将这6834条微博进行情感标注后得到具有正向情感倾向的微博1358条,具有中性情感倾向的微博2349条,具有负向情感倾向的微博3127条。

在信息检索、模式识别等领域,准确率(P)和查全率(R)是最基本的评价指标,可以从不同侧面对结果进行质量评估,而 F 值是综合这两项指标的评估指标,用于综合反映实验结果的整体质量。在对微博情感进行分析时,选择这3项指标作为判别分析结果的标准,其计算公式分别为

$$P = \frac{TP}{TP + FP}$$

$$R = \frac{TP}{TP + FN}$$

$$F = \frac{2P \times R}{P + R}$$

其中, TP 表示微博情感分析中正确分类的微博文本数量, FP 表示检索到但分类错误的微博文本数量, FN 表示未检索到但属于此类别的微博文本数量。本文使用Matlab分析计算结果。

2.2 结果与分析

4种情感倾向分析方法对6834条有效微博数据的测试结果如图3—图5所示。

由图3可知,文献[7]方法对中文微博分析平均准确率高于文献[5]方法和文献[6]方法,说明在中文微博情感倾向分析中,表情加权和语义规则对情感分析具有一定的积极作用。本文方法针对中文微博的语言特点和独有的表情特点,在结合了表情加权和语义规则的基础上,又融入一些特殊的影响情感倾向的情感词,从多个方面对情感倾向进行分析,对情感倾向的判别具有一定的修正作用,相比较其他3种方法,本文提出的微博情感倾向分析的平均准确率达78.4%,整体上提高了情感倾向分析的

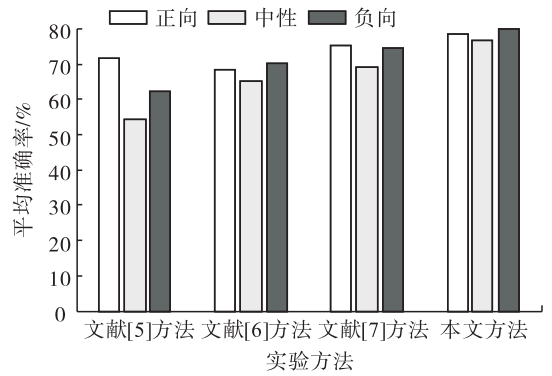


图3 4种情感倾向分析方法平均准确率比较
Fig. 3 Comparison of the average precision rate of four methods of emotional tendency analysis

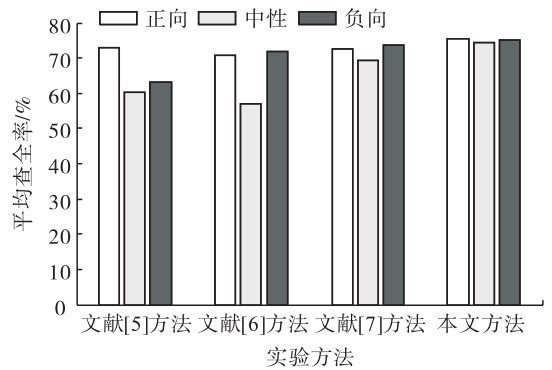


图4 4种情感倾向分析方法平均查全率比较
Fig. 4 Comparison of the average recall rate of four methods of emotional tendency analysis

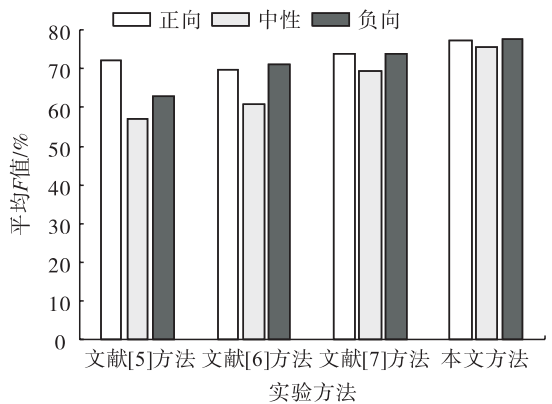


图5 4种情感倾向分析方法平均F值比较
Fig. 5 Comparison of the average F value of four methods of emotional tendency analysis

准确率。

由图4可知,4种情感倾向分析方法的平

均查全率都大于 70%,而本文方法的平均查全率最高,为 75.2%。这是因为中文微博比较口语化且时效性强,经常出现一些新兴的网络词汇,而普通情感词并不包含这些。因此,本文方法在普通情感词中同时融合否定词、程度副词和网络新词的方式,对情感倾向的分析更加全面、准确。

由图 5 可知,本文方法的平均 F 值最高,为 76.7%,其中中性倾向的中文微博 F 值整体上都低于正向和负向情感倾向的 F 值。这主要是因为正向和负向情感倾向的中文微博通常都带有明显的情感倾向标志,而中性情感倾向的中文微博容易因为中文语言特点造成误判,例如“垃圾分类是一个严肃的问题”,“严肃”一词,在情感词典中被划分为负向情感词,但此文本是在阐述事实,是一种中性情感,这就影响了中性情感分析的结果。

整体而言,随着实验方法中衡量微博情感倾向的因素增多,正向、中性和负向倾向的中文微博情感分析指标都有所提高,而通过融入表情符、语义规则、否定词、程度副词和网络新词众多特征后,中文微博情感倾向分析的平均准确率、平均查全率和平均 F 值都有一定程度的提高。但即使从多方面考虑影响微博情感倾向的因素并对其偏差进行修正,本文方法的平均准确率仍未突破 80%。造成这种状况的原因可能是,用户在发布微博时,存在语言、符号和表情滥用的情况,或者文本中存在着讽刺意味,这对微博情感倾向的分析带来较大的干扰。

3 结语

针对目前中文微博情感分析方法考虑因素不全面,从而导致情感分析结果不佳的问题,本文根据中文独特的语言方式和微博文本信息量大、时效性强、随意性强等特点,充分考虑影响情感分析结果的各方面因素,提出了一种基于

语义规则和表情加权的中文微博情感分析方法。该方法从微博文本的多方面特征考虑,在传统情感词的基础上进行拓展,融入了否定词、程度副词、网络新词,并使用表情加权和语义规则进行权值求和确定情感倾向。实验结果表明,本文提出的中文微博情感分析法效果显著,相较于其他 3 种分析方法,本文方法的平均准确率(78.4%)、平均查全率(75.2%)和平均 F 值(76.7%)相对较高。但该方法对于歧义句的分析效果还不令人满意,在对具有讽刺意味的语句进行情感分析时,发生误判的概率较高,后续将会结合具体领域和情境对中文微博情感分析做更深入的研究。

参考文献:

- [1] YU Y. Text emotional analysis based on Twitter data [J]. Information and Computer, 2018, 46 (19):151.
- [2] PAK A, PAROUBEK P. Twitter as a corpus for sentiment analysis and opinion mining [C] // European Language Resource Association. Proceedings of International Conference on Language Resource and Evaluation. Valletta: LREC, 2010:1320.
- [3] RILOFF E, WIEBE J. Learning extraction patterns for subjective expressions [C] // Association for Computational Linguistics. Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing (EMNLP-03). Sapporo: EMNLP, 2013:105.
- [4] 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 (1):14.
- [5] 谭皓, 邓树文, 钱涛, 等. 基于表情符注意力机制的微博情感分析模型 [J]. 计算机应用研究, 2019, 36(9):2647.
- [6] 李继东. 基于扩展词典和规则的中文微博情

- 感分析[D].北京:北京交通大学,2018.
- [7] 王文,王树锋,李洪华.基于文本语义和表情倾向的微博情感分析方法[J].南京理工大学学报(自然科学版),2014(6):733.
- [8] 刘志明,刘鲁.基于机器学习的中文微博情感分类实证研究[J].计算机工程与应用,2012,48(1):1.
- [9] 于韬,李伟,代丽伟.基于Python的新浪新闻爬虫系统的设计与实现[J].电子技术与软件工程,2018(9):188.
- [10] 林江豪,顾也力,周咏梅,等.基于表情符号的情感词典的构建研究[J].计算机技术与发展,2019,29(6):181.
- [11] 梁亚伟.基于表情词典的中文微博情感分析模型研究[J].现代计算机(专业版),2015(21):7.
- [12] 宋沛玉.面向中文微博情感分析的多特征融合方法研究[D].广州:广东工业大学,2018.
- [13] 蔺璜,郭姝慧.程度副词的特点范围与分类[J].山西大学学报(哲学社会科学版),2003(2):71.
- [14] 杨立月,王移芝.微博情感分析的情感词典构造及分析方法研究[J].计算机技术与发展,2019,29(2):13.
- [15] 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.
- [16] 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(2):1.
- [17] 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.
- [18] 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.