山西大学学报(自然科学版)37(4):580－587,2014
Journal of Shanxi University(Nat. Sci. Ed.)

文章编号:0253-2395(2014)04-0580-08

# 基于情感语义词典与PAD模型的中文微博情感分析

孙晓[1],叶嘉麒[1],龙润田[2],任福继[1]

(1.合肥工业大学 计算机与信息学院情感计算与先进智能机器安徽省重点实验室,安徽 合肥 230009;
2.加州理工学院 工程与应用科学学院计算机与数学科学系,美国 加利福尼亚州 91125)

**摘　要**:随着社交网络影响的不断增加,微博作为人类社会交流、发布观点信息的重要载体,其所包含的情感状态具有重要的研究意义。文章通过对微博文本及其包含的情感词汇的分析研究,引入神经网络语言模型和语义向量,结合心理学、情感计算领域相关知识,采用心理学PAD连续维度情感描述模型作为文本情感分析量化的基础,对微博文本所蕴含的情感状态进行分析研究,以获得更加精确的情感分析结果,达到情感分析的目的。同时实现了从个性角度的微博文本情感的可计算性。实验表明,所述方法能较好地提高微博文本情感分析的准确性和精确度,在不同主题不同情感特征中均能够得到很好的应用。

**关键词**:情感词;PAD情感模型;情感量化;中文微博;情感分析

**中图分类号**:TP391　　**文献标志码**:A　　**DOI**:10.13451/j.cnki.shanxi.univ(nat.sci.).2014.04.017

# Sentiment Analysis of Chinese Microblog based on Emotional Semantic Words Dictionary and PAD Model

SUN Xiao[1], YE Jiaqi[1], LONG Runtian[2], REN Fuji[1]

(1. *Key Laboratory of Affective Computing and Advanced Intelligent Machines, Hefei University of Technology, Hefei* 230009, *China*;
2. *School of Engineering and Applied Science, Computer and Mathematical Sciences, California Institute of Technology, Pasadena* 91125, *USA*)

**Abstract**: With the increasing impact of social networks, Microblog becomes important carrier of information and social interaction for human beings, which contains emotional states that have important research significance. Based on the analysis of microblog text that contains the emotional vocabulary, combining domain knowledge of psychology and affective computing, continuous dimension of emotion psychology PAD model is adopted as basis of sentiment analysis quantified for text sentiment analysis, anel emotional state inherent in the text is analyzed to obtain a more accurate result and achieve purposes of emotional analysis. At same time, emotional Microblog text computability is achievel from the aspect of personal characteristics. Experiment results show that the method can improve the Microblog text sentiment analysis accuracy and precision. In the different themes and different emotional features, the method is able to get a good application.

**Key words**: emotional word; PAD emotional model; emotional quantification; Chinese microblog; sentiment

**收稿日期**:2014-08-27;**修回日期**:2014-09-17

**基金项目**:国家自然科学基金(61203315);国家高新科技发展计划(863,No2012AA011103);安徽省科技攻关项目(1206c0805039)

**作者简介**:孙晓(1980－),山东龙口人,工学博士,副教授,研究领域为自然语言处理,机器学习,人机交互,E-mail:suntian@gmail.com

analysis

## 0 引言

微博(Microblog)作为数字信息时代的产物,是一种新兴的依靠社交网络的信息传播平台。用户在微博中可公开发布 140 字以内的实时文本信息,并允许任何人阅读或者只能由用户选择的群组阅读。目前来看,这些信息可以通过包括短信、IM 客户端、手机、API 接口等多种途径进行传递。同时大多数的微博平台也可以发布多媒体信息。以新浪微博平台为例,截止到 2012 年 12 月底,新浪微博的注册用户数目就超过 5 亿。微博日发送量超过 1 亿条。微博已成为人们在互联网领域中的重要信息来源和交流媒介。

日发送量巨大的微博,以其传播迅速、更新快的特点吸引了一大批的专家学者对其进行研究。同时,在每天发送的微博文本中,发现其中非常大的一部分带有很重的感情色彩。因此,作为自然语言处理的一个重要方面,情感分析逐渐成为微博文本研究的重要领域。文本情感分析的最主要任务就是信息观点的挖掘和情感分析分类,例如,对于用户评价信息的情感挖掘技术可以帮助生产商获取产品的反馈信息[1]应用于新闻文本领域的文本自动应答系统[2]和文本摘要自动生成技术[3],可以帮助用户迅速准确地获取需要的信息。就目前的文本情感分析技术而言,主要包括基于词典和规则的分析方法和基于机器学习和统计思想的情感分析方法。其中,第一类方法根据文本中所包含的情感词数目和种类进行分析,另一种方法则是采用机器学习技术选择文本情感特征进行训练标注,并使用最大熵、支持向量机等分类器进行情感的分析分类。

有研究表明在自然语言和文本的表达过程中,人们的言语和表述中常常并不只含有一种情感,很多语句有可能是几种情感融合后的表达。在微博文本中,也可以看到,很多情况下,文本信息中含有很多极性不同的情感词。单一的通过离散情感标签来标记这些语句的情感状态,并不能很恰当地表达出该文本所表达的情感状态。情感维度空间描述情感的方法也称为情感维度论。维度论把不同的情感看作是逐渐的、平稳的转变,通过不同情感在维度空间中的距离来衡量彼此的相似性和差异性[4]。这使得通过研究不同情感词的混合表达关系就可以获得较为精确的文本情感,以实现文本情感的可计算。

在各种情感维度模型中,由 Mehrabian 提出的 PAD 三维情感模型[5]是其中较为成熟的维度空间情感描述模型,它将人类的情感投影到由 P(*Pleasure-displeasure*)、A(*Arousal-nonarousal*)和 D(*Dominance-submissiveness*)组成的三维空间中。经心理学、情感计算的研究表明[6],利用 P、A、D 可有效地描述和解释人类的情感组成,有效地区分不同的情感状态,并可以采用量化计算的方法获取各种情感间的关联关系。利用该情感描述模型对情感词进行建模分析,就有可能实现文本情感的定量分析。

目前,对中文微博情感分析的研究较少,大多数的研究者都是借鉴普通文本情感分析的方法进行研究。本文研究基于情感词和情感模型的中文微博情感分析方法,结合 PAD 情感建模方法和基于词典的文本情感分析方法,提出了一种基于 PAD 距离测度的文本情感分析方法。通过对情感进行建模,将文本中的情感词投影到情感空间中,进行距离测度和聚类,以实现文本情感的定量分析。实验表明,本文提出的方法有效地分析了文本情感的组成,首次实现了从个性化角度的微博文本情感的可计算性,提高了文本情感分类分析的精确程度和正确率。

### 0.1 中文情感分析

中文微博的情感分析主要可分为三个部分:文本预处理、特征信息提取以及情感分类分析[7]。其中文本预处理包括文本的分词、标注、句法分析等自然语言处理技术。特征信息提取则是针对微博文本抽取其中有价值的情感信息。包括情感词的抽取和判别、主题抽取、关系抽取等。在情感信息提取判别中,朱嫣岚等[8]提出的基于 HowNet 的语义词汇倾向性相似度计算方法,大大提高了文本判别的准确率。王素格等[9]考虑了中文情感同义词间的关系,提出基于同义词的文本情感倾向判别方法,获得了不错的效果。谢丽星等[10]对微博文本中的表情符号进行分析研究。王岩[11]将微博文本解析为文档链,利用文本模型对文本主题判别进行了分析。在本文中采用基于情感词典的情感词抽取方法,对文本中的情感特征进行提取。

微博情感的分类分析则是针对所提取的文本情感特征进行分类分析,以获取微博文本中所蕴含、所表达的情感状态。总的来说,中文的情感分析方法与英文类似,大致有两种:第一种是有监督的机器学习方

法[12-14]。这类方法主要采用机器学习中的最大熵、支持向量机以及朴素贝叶斯等分类器对文本的情感进行分析。Zhao 等[12]采用 CRF 模型并引入"冗余特征"研究了情感分类的问题,Li 等[13]提出了基于 DS-LDA 模型用于评论数据的情感二分类。第二种是基于规则和组合的方法[15-18]。李寿山等[15]研究了四种不同的分类方法在中文情感分类上的应用。谢丽星等[16]则对于文本的层次结构进行了分析,针对主题无关特征和主题有关特征均得到了较高的情感分析准确率。对于微博情感分析工作,大多从文本角度,将微博情感分类问题视为文本分类问题,采用机器学习模型解决。

### 0.2 文本情感建模

在情感建模领域,目前对于汉语词汇或文本的情感建模报道较少。这一定程度上制约了汉语文本情感分析识别研究的发展。文献[6]首次针对汉语情感词汇进行了情感建模,将 88 个情感词汇投影到 PAD 情感模型中进行距离度量,最终获得 14 个情感类别及其情感距离关系。

PAD 三维情感模型是一种较常用的利用维度空间标识情感状态的模型。在音视频语音合成、情感计算等领域均有较为广泛的应用。PAD 情感模型由 Mehrabian 等于 1974 年首次提出。PAD 模型由三个维度组成,愉悦度(Pleasure)反映的是情感的本质,表现情感的极化程度。愉悦度越高,情感的正面积极程度越高。激活度(Arousal)表现情绪生理的激活水平和个体警觉性,反映了个体对所处环境的活跃程度。优势度(Dominance)表示不同情感下的主观控制水平,主要用于区分所表现的情感状态是由外界环境引起的,还是由自身主观激发出的。与其他模糊情感描述方法相比,三维情感模型主要有以下特点:在 PAD 模型中,每一种情感都唯一对应一个 PAD 空间坐标位置。当 PAD 参数归一化后,情感可以用唯一的三维坐标来标识,具有高置信度的评价。在模型中,通过一组标准情感量表完成 PAD 参数坐标的确定,PAD 各维度间的独立性能够更容易的区分位于不同情感维度的文本情感。基于 PAD 模型,可以对个性化的情感因素建模,即区分了外界环境和个性等内外因素对于情感的影响,可以更全面地描述微博情感。PAD 的建模形式,比通常的将情感简单分为几类要更加全面地反映出影响情感的因素和情感的真实状态。

## 1 基于情感模型的微博情感分析方法

### 1.1 微博语料典型情感词表的建设

情感是人主观的心理、生理现象,人的情感表达是通过表情、语义、语音以及姿态等多通道作用的结果[19],是一种由内而外通过主观冲动引起的心理、生理状态。在微博、博客等文本中,情感的主要表达方式是情感词、情绪词的使用。研究者通过对于微博文本内作者情绪词汇的使用情况,就可分析判断出在该条文本中,作者的情感表达是积极的还是消极的。举个例子来讲,微博文本"录了一个晚上的歌! 终于又让我吃到你了!!! 除了涮肉我在北京的最爱! 吃完就可以安心地睡觉喽!"中,可以看出,在文本中所含有的情感词包括"最爱"、"安心",根据这些情感词就可以判断出,该条文本表现的是正面、积极的情感。

由此看来情感词是情感极性判别中较为重要的考量标准和依据。情感词典中的情感词汇极性的正确与否及其精确程度都将对情感的判断产生影响。本文通过分析整理微博中常见的情感表征词汇,首先选取其中 60 个建立微博语料典型情感种子词典。这些典型情感词汇在任何情况下,其情感表现都应具有绝对性。这些词汇作为种子词汇,通过引入 Google 的 word2vec 模型工具[20-22],对这些种子词汇进行扩展,在 2 亿词的微博语料上进行训练,获取了与该 60 条词汇最相关的 300 个词汇(每个词选取 Top5 相关词汇),通过不断的迭代可以不断扩充该情感词汇表,获得更广泛更有代表性的情感词表。部分情感种子词汇如表 1 所示。

**表 1 典型情感种子词汇表**

**Table 1 Typical emotional seed vocabulary**

| 情感词汇 |
| --- |
| 喜欢,高兴,愉悦,乐意,关怀,感激,同情,兴奋,崇拜,后悔 |
| 悲愤,失望,不平,心寒,自卑,痛苦,无聊,郁闷,懊悔,放松 |
| 悲催,讨厌,得意,欣慰,尊崇,快乐,舒畅,炫耀,自满,感动 |

利用以上情感词汇,进行 PAD 评价标注建模,就可获得各个典型情感词汇的 PAD 模型位置。因为中文微博的口语化较为普遍,传统的固定词表的方法无法进行扩展,有一定的局限性,而且处理口语化的词汇

尤其是新的词汇需要有一个可以不断进行扩展的词典。通过引入语义词典,基于种子词汇表和词义向量模型(word2vec),可以不断扩展 PAD 的词汇表,而该词汇表是在海量的语料上训练得到的,例如表 2 是通过 word2vec 学习到的新的具有情感倾向的词。通过引入 word2vec 和语义词典,使得该模型在中文微博上可以不断地学习新的语言现象,更适合处理中文微博中的特殊语言现象。

**表 2 基于种子词典和 word2vec 得到新情感词**

**Table 2 New emotional words based on seed dictionary and word2vec**

| 新情感词 | 近义词 | 情感倾向 |
|---|---|---|
| 善意 | 玩笑话、执掌、敬爱、世人、约翰·布拉格 | 正面情感 |
| 道听途说 | 胡编、耳光、何故、难于、超然物外 | 中性情感 |
| 听话 | 家伙、豆汁儿、心领神会、随口、大人 | 负面情感 |
| 吃喝嫖赌 | 菜叶、澳门、防护兵、夜生活、浴场 | 负面情感 |
| 盗窃案 | 厂房、郊区、站台、辱骂、刑讯逼供 | 负面情感 |
| 老龄化 | 估量、全球化、产物、必然性、基本矛盾 | 负面情感 |
| 感染力 | 文学、应和、歌唱、魅力、评论家 | 正面情感 |
| 打击报复 | 控告人、检举、堵塞、泄漏、举报人 | 负面情感 |

### 1.2 基于 PAD 的情感词建模

在心理学的研究中,PAD 情感坐标的评定是通过一套设计完整的量表体系来完成的。在 PAD 情感模型中,由于人类语言的多样化、差异化和情感表达的不同,典型情感在 PAD 情感模型中的映射关系没有一个统一的标准。

国外的研究者大多采用 Mehrabian 编制的完整量化表,共包含 34 个测试项目,其中测量 $P$ 值 16 项,$A$ 值和 $D$ 值各 9 项。由于完整量表的测试较为复杂,研究者又进一步提出了简化量表,对 3 个维度各用 4 个项目进行测量。中国的 PAD 模型情感评价量化表由中国科学院心理研究所修订[19],是语义差异量表,共 3 个维度 12 项,每项划分为 9 段。如表 3 所示。

**表 3 PAD 标准项目表**

**Table 3 PAD standard table**

| 标号 | 项目 | 标号 | 项目 |
|---|---|---|---|
| S1 | 愤怒的——感兴趣的 | S7 | 痛苦的——高兴的 |
| S2 | 清醒的——困倦的 | S8 | 感兴趣的——放松的 |
| S3 | 受控的——主控的 | S9 | 谦卑的——高傲的 |
| S4 | 友好的——轻蔑的 | S10 | 兴奋的——激怒的 |
| S5 | 平静的——兴奋的 | S11 | 拘谨的——惊讶的 |
| S6 | 支配的——顺从的 | S12 | 有影响力的——被影响的 |

当获得一个情感词汇后,评价者根据表 3 中的项目依次对情感词汇进行判别打分,打分范围为"−4—+4"之间,最后维度分数由测量该维度的 4 个项目得分经参数计算公式计算后得出。其中,归一化后的 PAD 情感计算公式如表 4 所示。

**表 4 PAD 参数归一化计算公式**

**Table 4 PAD parameter normalization formula**

| 维度 | 归一化公式 |
|---|---|
| $P$ | $P=(S1-S4+S7-S10)/16$ |
| $A$ | $A=(-S2+S5-S8+S11)/16$ |
| $D$ | $D=(S3-S6+S9-S12)/16$ |

在归一化的 PAD 情感空间中对所有典型词汇进行 PAD 判别计算,将其投影至情感空间中进行分析聚类就可获得该类型情感点的 PAD 位置。

在中国,中科院心理所对无差别文本下的情感语音进行了 PAD 建模分析对照,共获得 11 类典型情感点的位置及其 PAD 参考值。由文本与语音及相同情感在各表达通道间的相互关系,本文对文本分类引入语音情感的分类方法,同样将文本情感划分为 11 个类别,依次为中性、放松、温顺、惊奇、喜悦、轻蔑、厌恶、恐

惧、悲伤、焦虑和愤怒。使用 K-means 对情感词表中所有的情感词进行 11 种情感类别的聚类,获得各类情感点的情感空间位置。因为引入了 word2vec 神经网络语言模型,所有的情感词都已经表示为一个语义向量,因此可以通过语义向量的空间距离来进行聚类。

### 1.3 文本情感特征提取

在微博文本中,由于其口语化特点,文字较为简练、简洁,与传统中文写作差异较大。常常会出现否定前置、双重否定以及文本口语化和表情使用等。这些都将会对文本情感特征的提取和判断产生较大的影响。因此在进行文本情感特征提取时,需要对文本及其对应的上下文关系、环境关系等进行分析。

首先,对所获取的微博文本提取其中的情感表征词汇。

其次,针对所提取出来的每一个情感词汇,依次分析其所在句子的语法关系以及上下文关系,依存分析采用的是哈尔滨工业大学的开源工具 HIT_IRLAB 的 LTP。例如"对于产品质量不怎么满意",该句子的语法分析树如图 1 所示。

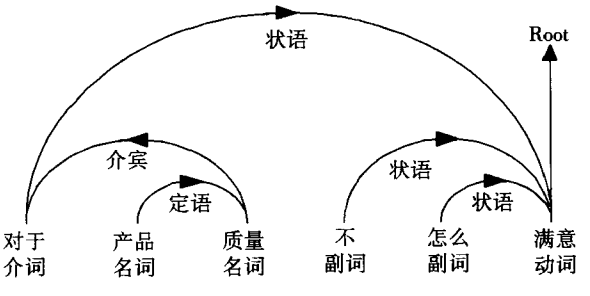


Fig. 1 Syntax analysis tree

图 1 语法分析树

对含有否定前缀和否定后缀的情感词进行处理。将其划分或表征成为相反情感的情感词汇。在研究中分析发现,在各种情感词汇中,表现为喜悦、悲伤和愤怒的三类情感词最容易受到否定的影响。

最后,针对已经进行处理的情感词汇,将这些词汇放入连续维度的情感描述空间中,进行距离分析,当出现情感词表中没有的情感词汇时,将对新的情感词汇进行 PAD 标注评分(空间距离),以使得各个情感词都能够在情感空间中找到对应点或近似点。

### 1.4 基于情感聚类的文本情感判别

将所提取的情感词汇投影至 PAD 连续维度空间内后,对所提取的情感词汇及其 PAD 空间内位置进行主成分分析获取该微博在 PAD 空间内的情感位置,设为待测情感点。而后,计算各典型情感类别到待测情感点的距离和权重比例,以此对微博情感进行分析。

在 PAD 空间中,本文采用欧氏距离来计算待测情感点到各个典型情感类别间的距离。计算公式如下:

$$S(p_1 - p_2) = || p_1^2 - p_2^2 || \tag{1}$$

其中 $p_1$,$p_2$ 为两情感点在 PAD 空间内的 $p$ 参数观察值。由于 PAD 为三维情感空间,则两情感间的最终距离为:

$$S = \sqrt{S_P^2 + S_A^2 + S_D^2} \tag{2}$$

由空间距离测度聚类可知两种情感之间的距离越小,则这两种情感的相似程度越高,采用此方法就可进行文本情感的分析判别。由此可以计算出待测情感点到各情感类别间的距离,获得待测微博文本的基本情感组成。

在文中,设文本待测点到各情感类别的欧式距离分别为:$S_1$,$S_2$,…,$S_n$。则待测文本的情感组成权重为:

$$M_{\max} = \frac{S_{\max}}{\sum_{i=1}^{n} S_i} \tag{3}$$

其中 $M_{\max}$ 为各类情感中距离待测点最近的情感类别所占权重。利用排列组合的方式,就可获得该待测点的所有情感组成,得到中文微博文本的情感。

## 2 实验及结果分析

### 2.1 实验设置

本文使用新浪微博 API 的 OAuth2.0 接口,采用以下方式准备实验数据集。

1)在微博数据中以"索尼"、"iphone5s"、"人人网"、"小时代"、"致青春"、"毕业季"、"郭敬明"、"科比"、"川菜"、"必胜客"等为关键字分别抓取这 10 种主题的微博文本数据各 200 条,共 2 000 条微博文本数据。

在抓取的文本中应该至少包含有 2 个以上的情感词(情感关键词或表情符号)。这 2000 条微博文本作为实验数据 1。

2)以除中性外的 10 类情感词集合为关键词,抓取剔除广告、回复的微博情感文本各 100 条作为实验数据 2。

为了验证本文所提出方法的准确性,采用对比实验的方法进行结果比较,将自动分析的结果和人们主观分析的结果进行相互比较。同时实验中的参数设置均采用交叉验证得到最优参数。在本实验中,实验 1 对实验数据进行 11 种情感状态的归属分析。实验 2 仅考虑本文方法所获得的微博情感与抓取所使用的情感关键词是否对应。

在实验 1 中,先对各条微博进行情感特征提取,再通过 PAD 情感空间进行分析聚类,判断其情感组成。同时使用主观评测的方法,让 5 位评测者(多数原则)直接对所抓取的微博文本进行情感分析判断,获取该文本的主观情感表现,作为文本情感的标准答案。若主观评测与机器情感分析的结果相似,则判断该微博文本的情感分析正确。实验 2 中,则只考虑文本情感属性分析是否正确。本实验的评价指标采用目前应用较为广泛的正确率、召回率、精确度。实验 3 中的数据集源于 COAE 2013 任务 3 的数据,从任务 3 中随机抽取 2000 条数据进行人工标记使用 DataSet2 表示,在该数据集上进行比较。因为 COAE2013 的数据任务只有两种情感倾向,因此也将 11 种情感划分为两类,保持与 COAE2013 的标准一致,其实验结果如表 5 所示。

**表 5 实验 3 COAE2013 语料上开式分析结果**

**Table 5 Analysis of experiment on COAE2013 corpus result**

| | *P* | *R* | *F* |
|---|---|---|---|
| 正向 | 81.2% | 93.2% | 86.8% |
| 负向 | 80.6% | 89.0% | 84.6% |

## 2.2 实验结果分析

实验 1 和实验 2 结果如表 6 和表 7 所示。

**表 6 基于 PAD 模型的实验 1 结果**

**Table 6 Experiment results based on PAD model**

| | *P* | *R* | *F* |
|---|---|---|---|
| 放松 | 78.10% | 67.23% | 72.26% |
| 温顺 | 75.23% | 70.22% | 72.64% |
| 惊奇 | 67.97% | 71.78% | 69.82% |
| 喜悦 | 81.54% | 71.34% | 76.10% |
| 轻蔑 | 64.23% | 75.14% | 69.25% |
| 厌恶 | 81.56% | 76.34% | 78.86% |
| 恐惧 | 69.56% | 89.32% | 78.21% |
| 悲伤 | 79.00% | 70.98% | 74.78% |
| 焦虑 | 61.11% | 67.34% | 64.07% |
| 愤怒 | 80.80% | 79.42% | 80.10% |

从表 6 中可以看出,基于 PAD 模型的微博情感,与主观判断结果比较,更符合评测者的主观评测判断结果,10 类情感的判断总体上高于目前多类情感分析系统。因为采用聚类方法处理各个情感分类中的情感词,使得各类情感的精度较为均衡;表中也可以看出负向情感精度总体上高于正向情感,也符合主观判断规律。表中的愤怒情感的判断精度最高,部分因为愤怒的 PAD 模型中词汇较为明显,与其他情感词汇的距离较远,据此也可以引入深层语义特征来进一步判断 PAD 模型中词汇间的距离。基于语义进行聚类分析可以进一步提高情感判断的精度,因目前语义分析的精度普遍不高,达不到实用效果,因此本文未引入语义判断情感。

根据表 7 分析可知,因为只考虑文本属性分析是否正确,各个情感分类的综合 *F* 值有所提高。因为文本情感属性与情感词本身存在较强的相关性,但同时受到词法和句法的影响,因此文本情感判断的 *F* 值提高有限。通过引入词法和句法信息,可以进一步提高情感分类的综合指标,但目前语法和句法信息本身精确

率不高,引入后反而会影响判断精度。

表 7 实验 2 文本情感属性判断

Table 7 Text emotional judgment result of Experiment 2

| | P | R | F |
|---|---|---|---|
| 放松 | 79.56% | 71.54% | 75.34% |
| 温顺 | 77.45% | 72.11% | 74.68% |
| 惊奇 | 70.12% | 72.67% | 71.37% |
| 喜悦 | 82.21% | 73.43% | 77.57% |
| 轻蔑 | 70.11% | 76.67% | 73.24% |
| 厌恶 | 81.34% | 77.76% | 79.51% |
| 恐惧 | 70.53% | 88.75% | 78.60% |
| 悲伤 | 78.57% | 75.45% | 76.99% |
| 焦虑 | 62.67% | 68.65% | 65.52% |
| 愤怒 | 81.34% | 80.54% | 80.37% |

## 3 结论及展望

随着微博在中国社会交互领域的流行,微博文本的分析研究也越发成为时下的研究热点。情感计算技术的出现使得情感状态具有了可计算性。为了提高微博文本的情感分析精确程度,获得更加准确的文本情感。本文通过对情感词进行连续情感空间建模,利用情感词典等文本情感特征提取方法,提出了基于 PAD 情感模型的微博情感分析方法。通过建立 PAD 情感模型对不同文本语料的情感状态进行了分析和描述。具体来说,本文通过提取微博文本中的情感词和情感特征,并对其进行分析判断,获得 PAD 空间中各个情感词位置,再通过 PAD 模型对微博文本中所包含情感进行分析,以使得计算机能够获得更加准确的微博情感,达到提高微博情感分析准确率和精确度的目标。实验结果表明,本文所用方法能够更准确地获得微博文本的情感状态,同以往的情感极性判断方法相比较,本方法能够获得更加准确的情感归属,量化微博文本所表达的情感状态,具有了一定的可计算能力。

目前,本方法尚有很多不足之处。在未来的研究中,将从如下几个方面进行进一步的研究和探索。首先,在语音的 PAD 情感模型中,PAD 空间中的情感状态并不是呈均匀的标准正态分布的。在文本情感中也有类似的问题。如何获取文本情感的分布关系,将是未来研究的方向;其次,在微博文本中,很多情感状态与其上一条或下一条微博的内容、情感有关,有的甚至与发送微博当时的环境存在较大的联系,如何确定这种环境的影响,在多通道下获取人们所表达的情感状态,将有助于提高本文情感分析的正确率;同时,也看到在进行文本情感分析的研究中,语言学、心理学等学科的研究成果都会对本领域的研究产生重大影响。作为交叉学科领域,下一步将对语言学相关论述进行研究分析,寻找一种更好的方法更加准确地提取文本内所蕴含的情感特征。

**参考文献:**

[1] Pang B, Lee L. Opinion Mining and Sentiment Analysis[J]. *Foundations and Trends in Information Retrieval*, 2008, **2**(1-2): 1-135.

[2] Sun Xiao, Ye Jiaqi, Ren Fuji. Real Time Early-stage Influenza Detection with Emotion Factors from Sina Microblog[C]// Proceedings of the 5th Workshop on South and Southeast Asian NLP, 25th International Conference on Computational Linguistics(Coling 2014), Dublin, Ireland, August, 2014, 23-29: 80-84.

[3] Hu M, Liu B. Opinion Extraction and Summarization on the Web[C]// AAAI, 2006, **7**: 1621-1624.

[4] 王志良. 人工心理[M]. 北京: 机械工业出版社, 2007.

[5] Mehrabian A. Pleasure-arousal-dominance: A general Framework for Describing and Measuring Individual Differences in Temperament[J]. *Current Psychology*, 1996, **14**(4): 261-292.

[6] 毛峡, 江琳. 一种基于 PAD 的汉语词汇情感建模方法, 国家发明专利. CN102184232A[P], 2011-09-14.

[7] 周胜臣, 瞿文婷, 石英子, 等. 中文微博情感分析研究综述[J]. 计算机应用与软件, 2013, **30**(3): 161-164.

[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.

[9] 王素格,李德玉,魏英杰,等. 基于同义词的词汇情感倾向判别方法[J]. 中文信息学报,2009,**23**(5):68-74.

[10] 孙晓,李承程,叶嘉麒,等. 基于重复字串的微博新词非监督自动抽取[J]. 合肥工业大学学报,2014,**37**(6):674-678.

[11] 王岩. 基于共现链的微博情感分析技术的研究与实现[D]. 长沙:国防科技大学,2011.

[12] Zhao J,Liu K,Wang G. Adding Redundant Features for CRFs-based Sentence Sentiment Classification[C]// Proceedings of the Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2008:117-126.

[13] Li F,Liu N,Jin H,*et al*. Incorporating Reviewer and Product Information for Review Rating Prediction[C]// IJCAI, 2011:1820-1825.

[14] Dasgupta S,Ng V. Mine the Easy,Classify the Hard:a Semi-supervised Approach to Automatic Sentiment Classification [C]// Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP:Volume 2-Volume 2. Association for Computational Linguistics, 2009:701-709.

[15] 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.

[16] Sun Xiao. Semantic Polarity Detection of Chinese Multiword Expression in Microblogging based on Discriminative Latent Model[J]. Journal of Intelligent & Fuzzy Systems,2014,**27**:753-759.

[17] Nasukawa T,Yi J. Sentiment analysis:Capturing Favorability Using Natural Language Processing[C]// Proceedings of the 2nd international conference on Knowledge capture. ACM,2003:70-77.

[18] Ding X,Liu B. The utility of linguistic rules in opinion mining[C]// Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM,2007:811-812.

[19] Soleymani M,Chanel G,Kierkels J J M,*et al*. Affective Characterization of Movie Scenes Based on Multimedia Content analysis and user's Physiological Emotional Responses[C]// Multimedia,2008. ISM 2008. Tenth IEEE International Symposium on. Ieee,2008:228-235.

[20] Mikolov T,Chen K,Corrado G,*et al*. Efficient Estimation of Word Representations in Vector Space[J]. arXiv preprint arXiv:1301.3781,2013.

[21] Mikolov T,Sutskever I,Chen K,*et al*. Distributed Representations of Words and Phrases and Their Compositionality[C] // Advances in Neural Information Processing Systems,2013:3111-3119.

[22] Tomas Mikolov,Wen-tau Yih,Geoffrey Zweig. Linguistic Regularities in Continuous Space Word Representations[C]. ProceediZgs of NAACL HLT,2013.