一种情感判别分析体系在汽车品牌舆情管理中的应用

宋云生
深圳联友科技有限公司，广东 深圳 518031

摘要
品牌舆情管理涉及文本、语音等自然语言产物的处理，如挖掘文本内涵的情感、观点等并对其量化，才能进一步分析品牌所处的舆论环境。对自然语言中情感的量化即情感判别分析，针对传统的基于词典的情感分析和基于监督模型的情感分析存在的不足，提出了一种新的情感分析系统，并结合朴素贝叶斯分类算法，提高了情感分析的准确率，并增强了量化分析情感强度的能力。经测试，提出的文本情感分析引擎的情感判别准确率高于常见的分析方法，且不具有非常明显的行业特异性。

关键词
情感分析；监督模型；朴素贝叶斯；自然语言处理

中图分类号：TP391 文献标识码：A doi: 10.11959/j.issn.2096-0271.2017061

Application of an emotion discriminant analysis system in the management of automobile brand

SONG Yunsheng
Shenzhen Lan-You Technology Co., Ltd., Shenzhen 518031, China

Abstract
Brand public opinion management involves text, voice and other natural language processing, such as mining the emotions and views of the text and quantifies it. The quantification of emotion in natural language is the emotion discriminant analysis. Considering the disadvantage in the traditional sentiment analysis that based on emotional dictionary and supervision model based sentiment analysis system, a new sentiment analysis system was proposed, and combined with the Naive Bayesian classification algorithm, the accuracy of sentiment analysis was improved, and the ability of quantitative analysis of emotional strength was enhanced. The sentiment discrimination accuracy of the proposed text sentiment analysis engine is higher than that of the common analysis method, and there is no decent of accuracy in out-of-sample texts from different industries.

Key words
sentiment analysis, supervised model, naive Bayes, natural language processing

1 引言

随着互联网以及各类新兴网络社交媒体的快速发展与普及，由用户发表的文字信息也在暴增，如论坛帖子、微博、博客、产品评论等。如何有效地对这些海量文本信息进行挖掘，识别其中的情感倾向，并加以合理有效地利用，是非常值得探讨的问题。情感分析又称倾向性分析，是人们对事物以及事物的属性持有的意见、情绪和情感的计算研究[1]。事物可以是产品、服务、组织、个人、事件、问题或者话题。情感分析也可以被定义为通过自然语言处理（natural language processing，NLP）技术从文本、演讲、微博等数据源中自动挖掘态度、观点、意见和情绪的过程[2]。文本情感分析就是分析一段文字的情感倾向，作为舆情监控的基础工作，用途广泛。社交网络越来越火，“意见领袖”越来越多，允许用户对商品和服务评价打分的站点更是如雨后春笋，用户的评价和建议可以全网传播。这些文本类型的数据毫无疑问是精准营销的动力来源。企业可以根据情感分析建立自己的数字形象，识别新的市场机会，做好市场细分，进而推动产品成功上市，但抓住这些评论的价值部分也是企业的巨大挑战。政府同企业一样，需要通过情感分析监控、缓解、引领舆情，消弭社会矛盾，上述正是情感分析的应用背景。

但与如此重要的背景背道而驰的是中文情感分析系统的弱势，常见的情感分析分为基于词典的情感分析和基于监督模型的情感分析。基于词典的情感分析，顾名思义，非常依赖于情感词典的构建，Ku L W等人[3]和Kaji N等人[4]对情感词典的构建开展了深入的研究。通常先将情感词分为正向（褒义）和负向（贬义），然后统计一条待分析的中文文本分词的正向词个数和负向词个数，如果正向词个数大于负向词个数，则这条文本属于情感正向，否则属于情感负向。有些研究者对情感词典进行了人工加权，比如“爱”和“喜欢”的权重不一样，人工给予“爱”更高的权重。但是无论怎么改变，这种分析方式都存在以下缺陷：首先，准确率非常低，一般为50%左右，几乎不能支撑舆情监控要求；其次，人工定义情感词的正负倾向或权重，工作量巨大，而且非常武断；最后，这种方式对于否定句和程度副词加强的语句几乎无效，从而丧失了分析情感细腻性（程度）的能力。另一种是基于监督模型的情感分析，即通过人工标注一个训练集（训练集的每一条文本都要人工将其分为情感正向或情感负向），然后使用训练集训练模型，模型训练完成后，预测待分析文本。这种方法虽然基于大量的训练集暂时提高了准确率（一般75%左右），但是标注训练集等如此浩繁的工作让使用者望而却步，另外，人工标注训练集的粒度导致了这种方式同样不具有分析情感细腻性的能力，或者能力较弱。

本文构建了一种新的情感分析系统，解决了这些比较具体的问题，化繁就简，提高了情感分析的准确率，并具有细腻分析情感的能力，为各大行业的舆情分析提供了一种新的实践路径。

2 舆情管理在各个行业的需求

早在互联网普及之前，人们就让朋友推荐一个汽车修理工或者在地方选举投票给谁，又或者向消费者咨询买什么样的洗碗机。“别人怎么想在我们做决策的过程中是一个非常重要的信息”[5]。随着Web2.0平台的爆发式增长，博客、论坛、

点对点网络等其他各种类型的社交媒体的出现，个人用户在网上表现出对产品和服务的兴趣（积极或消极）会产生一些潜在的影响，通过互联网的传播放大，能够产生前所未有的影响力，商品供应商也越来越关注网络用户的评论。目前舆情分析已渗透到生活的方方面面，几乎在各大行业中都有应用，包括政府、高校、企业、媒体、医疗、电力等领域。

政府对舆情的分析主要关注民生民意、行业动态以及危机公关，如通过分析网络上的评论可以非常准确地了解大众对政策的理解和情感倾向。德国慕尼黑大学的研究表明，推特（Twitter）上的信息能够非常准确地反映选民的政治倾向，通过分析2009年德国大选期间选民涉及政党和政客的10万条推特，结论是推特的信息能够预示大选的结果，其准确性不亚于传统的民意调研[6]。

2009年7月，一则《应届毕业生怒问：谁替我签的就业协议？注水的就业率！》[7]刷爆网络，“被就业”获得了社会各界的广泛关注。通过网络舆情分析，能够获取广大师生对高校就业的观点和建议，可以监测社会民众的情感走向，及时采取相应的政策引导舆论向有利于构建和谐健康的社会主义社会发展。

舆情分析在企业中的运用主要包括两个方面：品牌危机管理和营销管理。品牌危机是指突然发生的并能对企业声誉和生产经营活动构成重大威胁或造成破坏和损失的紧急事件[8]。通过对社会媒体的监测和分析，对危机发生前的环境进行监测和预警，第一时间掌握舆论导向，制定相应的对策和方法化解危机。目前企业越来越热衷于使用用户针对产品留下的评论内容等数据，帮助改进市场营销、品牌定位、产品开发和制定相应的优惠政策等活动。例如，北京小米科技有限责任公司想知道客户对他们的新机型的评价，在社交媒体和数据挖掘兴起之前，他们只能通过市场调研的方式解决。而数据分析则可以抓取消费者在各大消费网站（如亚马逊、京东、天猫、社交媒体）留下的评论数据，通过分析这些文本内容，从而获知消费者对某个新机型的情感倾向。通过机器学习量化文本中消费者对品牌或产品属性喜好的数据挖掘方式，即情感分析。情感分析作为一种数据挖掘的方式，可以用于采集竞争对手的竞争优势，例如企业可以轻易地跟踪社交媒体的情感倾向和社交媒体对竞争对手的情感倾向，了解消费者对竞争品牌的印象及对其产品的情感倾向。另外，情感分析的指数和结果还可以作为变量应用到其他数据挖掘项目，例如预测用户流失的概率时就可以添加情感指数作为变量。

目前，情感分析仍然面临着很多挑战，其中主要包括：人们表达态度的方式非常复杂，很难识别真正的情感倾向；另外，仅仅使用词汇并不能非常准确地识别一条文本表达的情感倾向；一些修辞手法（如反讽、欲扬先抑等）也会给情感分析造成困难。

3 情感分析的种类和方法介绍

一般情感分析分为两个层次：主/客观分析（subjectivity/objectivity identification）和情感/主体分析（feature/aspect-based sentiment analysis）。前者主要分析一个文本或片段是主观表述还是客观表述，需要注意的是在做这类分析时同样面临挑战，因为具体的语境可能会改变句子的意思，原本的主观评价可能就变成了客观描述，如“我买的锤子手机外表像广告一样光鲜，但速度就像他的名字一样，就是个锤子”。而情感/

主体分析将文本中表现的情感和具体的主体联系起来，即确定情感的归属，显然后者对文本情感的分析更加细腻。

目前情感分析方法主要分为两大类：基于词典的情感分析方法、基于机器学习的情感分析方法[9]。

基于词典的情感分析法起源于基于语法规则的文本分析，方法比较简单纯朴，首先需要具有语法敏感性的专业人士构建情感分析的词典：正向情感词典和负向情感词典，即将某语言中用于表达情感的词汇分为两个类别，然后比对文本中正负情感词的个数、频度，评估文本的情感倾向，这种方法非常容易理解。Taboada M等人[10]通过创建带有语义倾向标注的词典（极性和强度），并应用于极性分类任务，即可判断一个文本是正向还是负向。张成功等人[11]通过构造极性词典，包括基础词典、领域词典、网络词词典以及修饰词词典，深入探究了修饰词对极性词的影响，提出一种基于极性词典的情感分析方法，并验证了该方法的有效性。然而情感词也分轻重缓急，比如喜欢和爱虽然都是正向，但其程度不一样，因此根据语言专家的分析，给予情感词不同的情感级别或权重，即对上述分析方法的改进，毫无疑问这种方法包含一定的语法分析的成分，谷歌翻译的早期版本就是基于语法的方式，其效果可见一斑。

基于机器学习的情感分析过程首先制作一个规模庞大的训练集，人工标注文本的正负向，然后通过机器学习或算法等方式训练模型，得出模型后，再用来识别新文本的情感倾向，比较像垃圾邮件的分类方法。首先精挑一些垃圾邮件和正常邮件让模型学习，然后再将模型用于垃圾邮件的分选。基于机器学习的情感分析方法本质上是一个监督分类的方法，当然现在也有非监督分类的尝试。机器学习技术，如朴素贝叶斯（naive Bayes）、最大熵（maximum entropy）、支持向量机（support vector machine，SVM）等已经成功运用在情感分析中。Firmino A A等人[12]进行了一个案列研究，对比SVM和朴素贝叶斯分类器的性能，结果表明SVM性能更优。孙建旺等人[13]提出基于特征极性值的位置权重计算方法，将SVM作为机器学习模型，能够有效地对微博文本进行情感分类。关雅夫等人[14]提出了基于主动学习的微博情感分析，并结合SVM进行二级分类，结果表明该方法在提高准确率、降低过拟合及错误级联等方面有着良好的表现。樊娜等人[15]通过对文本结构和情感表达的特点进行分析，提出一种基于K-近邻的文本情感分析方法，实验表明该方法优于传统的机器学习。

4 情感分析的新分析体系介绍

本文提出了一种新的中文文本情感分析系统，主要创造了情感分析训练集的标注方式、加权情感词典的生成方式、汉语言语法规则的调整方式及基于朴素贝叶斯理论的情感得分计算方法。文本情感倾向值计算主要包括以下几个步骤文本预处理；文本特征提取，主要包括：提取文本情感主体、情感词、情感词前后的程度词和否定词；加权情感词典构建，情感词匹配；汉语规则构建，调整情感文本矩阵；模型训练；文本情感值计算。

4.1 系统分析流程

文本a进入系统后，首先对文本a进行分词，然后将文本分出来的词汇与加权情感词典中的词汇进行匹配，这样不仅筛选出了文本a中包含的情感词，而且给情感词

汇添加了正负向情感权重，即获得了文本a带有权重的情感词矩阵。为了分析文本a情感的强度，本文根据汉语语法构建了两个比较基本但很重要的规则：程度副词加权规则和否定词逆转规则，用于调整文本的情感词矩阵，将经过汉语言规则调整后的情感词矩阵输入算法模型，计算情感得分。情感得分的区间为[0,1]，越靠近0，负向情感越强；越靠近1，正向情感越强。整个流程如**图1**所示。

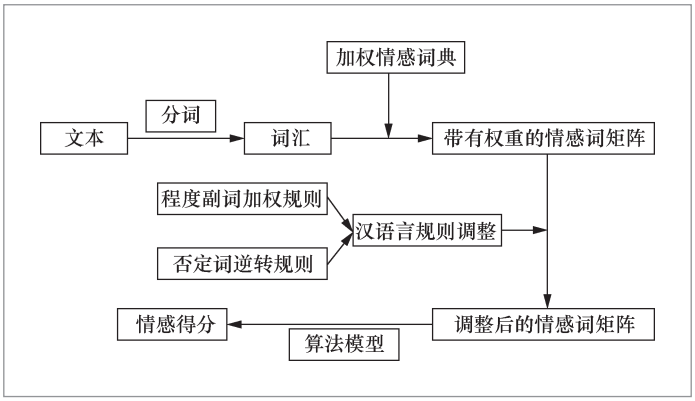

图1 情感分析系统流程

4.2 加权情感词典构建

随着汉语的演化，情感词还在不断增加。在文本分析的过程中，笔者积累了大量的情感词，并构建了情感词典，大约包括中文情感词20 000个左右。原始的情感词典见**表1**（其中1代表正向词汇，-1代表负向词汇），其仅仅是武断地将情感词汇分为正向和负向，这样的词典除了带有主观性以外，而且无法满足分析情感程度的目的，所以需要一种更加快速、客观的加权方式。

表1 原始的情感词典样例

情感词	类型
哀艳	1
安顿	1
爱不忍释	1
赞助	1
责任	1
阿Q	1
暗弱	-1
昂贵	-1
浅陋	-1

在构建加权情感词典之前，首先要有一个标注的情感分析文本集，这本来是一个需要人工标注的过程，工作量巨大，而且具有行业局限性。在绝大多数情况下，人们用于表达情感的词汇是相似的，仅仅有个别词汇具有行业特征。而且现在有大量的网站留下了用户的评论数据，有些网站，如汽车行业的汽车之家要求用户发表口碑评论时分为两个部分：最满意的部分和最不满意的部分，笔者抓取了大量的评论，并将“最满意的部分”标注为正向文本，把“最不满意的部分”标注为负向文本（类似的方法还可以使用用户评分进行文本标注），加上其他研究者已经公布的标注文本，共获得了大约30万条正负向文本标注训练集，通过这种批量方法可以节省大量的标注时间，而且扩大了文本的行业来源，还可以随着数据量的增加持续更新情感分析文本集，进而更新加权情感词典。

有了标注训练集，就需要基于标注训练集对情感词典加权。

情感加权规则：一个情感词在正向文本集出现的文档频率（document frequency，DF）作为它的正向权重，在负向文本集出现的文档频率作为它的负向权重，所谓DF，即包含某词的文档数/语料库的文档总数。

对于一些一般人无法判断的中性词，也能非常快速、合理地获得正负向情感权重，因此依据以上这种数据驱动的规则获得情感词的权重，不仅工作量锐减，而且更加客观（见**表2**）。通过以上方法获得了加权情感词典。

表 2 加权情感词典样例

情感词	在正向文本中的DF（pDF）	在负向文本中的DF（nDF）
哀艳	0.000 255 102	8.430 28E-5
安顿	0.000 340 136	0.000 590 12
爱不忍释	0.000 255 102	8.430 28E-5
赞助	8.503 4E-5	8.430 28E-5
责任	0.001 530 612	0.003 877 93
阿Q	0.000 255 102	8.430 28E-5
暗弱	8.503 4E-5	0.000 337 211
昂贵	0.000 680 272	0.000 421 514
浅陋	8.503 4E-5	0.000 337 211

4.3 构建汉语规则

本文系统构建了两种汉语语法规则，其一用于处理程度副词出现的情况，比如“我非常喜欢夏天”；其二用于处理否定词出现的情况，比如“我不喜欢夏天”。一般的基于词典和监督模型的情感分析系统基本上无法有效处理上述两种现象，而且上述现象是在汉语中非常常见的情感表达方式，所以针对上述两种情况本文提出了两套规则，按照先后顺序调整情感词矩阵即可。

4.3.1 程度副词加权规则

程度副词加权规则要求首先准备一张程度词加权词典，汉语中的程度副词比较少，通过人工整理并给予相应的权重可得部分词典，见**表3**。

程度副词加权规则：如果情感词前后不远处（可以根据标点符号和需求自定义）出现了任意一个程度副词，那么在该情感词的正负权重中，较大者加倍。例如“我非常不喜欢喝茶”，“喜欢”这个词的正向情感权重为0.05，负向情感权重为0.02，它的前方出现了“非常”程度副词，所以“喜欢”在本文本里的正负向权重就变成了0.1和0.02。

表 3 程度副词加权词典样例

程度副词	权重
略微	0.8
稍微	0.8
非常	2
极端	2
至极	2

4.3.2 否定词逆转规则

一个文本的情感词矩阵经过程度副词加权规则调整后，需要根据否定词规则进一步调整，本文构建了否定词逆转规则。所谓否定词逆转规则，即如果情感词前面不远处（可以根据标点符号和需求自定义）出现了否定词，且否定词的个数为奇数，那么该情感词的正负权重进行一次对调。例如“我非常不喜欢喝茶”，“喜欢”这个词的正负向情感权重经过程度副词加权后变成了0.1和0.02，但它的前方出现了“不”字且为否定词，并只出现了1次，所以“喜欢”的正负向权重就变成了0.02和0.1。那么经过调整后，“我非常不喜欢喝茶”的文本情感词矩阵就变成了**表4**。

4.4 构建模型

根据上文的基础词库和规则，可以获得任何一条文本的情感词矩阵，稍作矩阵变换，就可以作为构建各种监督型机器学习算法的输入数据，得出文本情感值，加上强大的训练集标注方法，各种监督模型（随机森林、SVM、逻辑回归等）均可以使用上述矩阵进行模型训练和测试，准确率相较普通系统大幅提高。经过程序测试，本文选择了朴素贝叶斯分类器算法，并集合汽车行业特有的标注数据，应用于汽车行业品牌情感分析。

表 4 文本情感词矩阵样例

文本	情感词	在正向文本中的DF（pDF）	在负向文本中的DF（nDF）
我非常不喜欢喝茶	喜欢	0.02	0.1

5 情感分析在汽车行业品牌舆情管理的应用

本文仅挑选朴素贝叶斯算法作为分类算法演示分类体系，具体实现流程如**图2**所示。所谓朴素贝叶斯分类器在本文中可以通俗地进行如下解释：一条文本中的所有情感词在正向文本中出现的概率连乘积如果大于这条文本中所有情感词在负向文本中出现的概率连乘积，则这条文本属于正向，否则属于负向，计算式如下：

$$P(a_1|y_i)P(a_2|y_i)\cdots P(a_m|y_i)P(y_i) = P(y_i)\prod_{j=1}^{m}P(a_j|y_i) \quad (1)$$

其中，a表示文本的情感词正负向权重，y代表文本属于正负向分类的概率。

5.1 实验数据采集

本文通过网络爬虫技术抓取发表在汽车之家、凤凰汽车网、太平洋汽车网和新浪汽车网的数据约2 100万条，取数周期为2016年1月1日—4月21日。在进行情感分析之前，首先对获取的数据集进行数据清洗[16,17]，主要包括删除特殊符号、分词、去除停用词，然后对文本属性特征进行提取，去除不能反映文本主题的词语。选择35万条口碑数据作为标注数据集，口碑数据具有两个非常明显的模式片段：最满意的一点和最不满意的一点。本文将最满意的一点作为情感正向标注，将最不满意的一点作为负向标注，从而丰富了数据标注，增加了行业特异性。

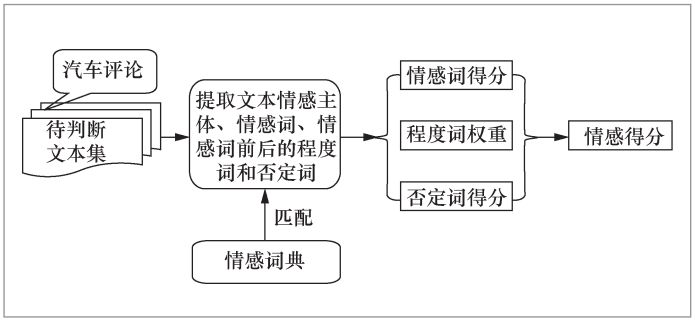

图 2 汽车行业的情感分析流程

5.2 加权情感词典构建

本文将知网、清华大学①、台湾大学②发布的基础情感词典作为基础词典，补充汽车行业情感词，通过训练汽车行业评论文本，整理出情感词词典。部分情感词典见**表5**。

① http://nlp.csai.tsinghua.edu.cn/~lj/sentiment.dict.v1.0.zip

② http://www.datatang.com/data/11837

5.3 实验结果

以“我非常不喜欢涡轮增压，保养贵”这句评论为例。第一步，通过数据清洗后，使用分词结果与加权情感词典进行匹配，获得带有权重的情感词矩阵，流程如**图3**所示。

第二步，在带有权重的情感词矩阵中，

表 5 情感词典

情感词	正向情感得分	负向情感得分
喜欢	0.051 161 3	0.009 133 5
保养贵	0.000 020 8	0.001 808 3
乱七八糟	0.000 541 6	0.001 164 7
发疯	0.000 041 7	0.000 122 6
创新	0.000 312 5	0.000 245 2
渗油	0.000 020 8	0.001 409 9
漏油	0.000 020 8	0.002 114 8
…	…	…

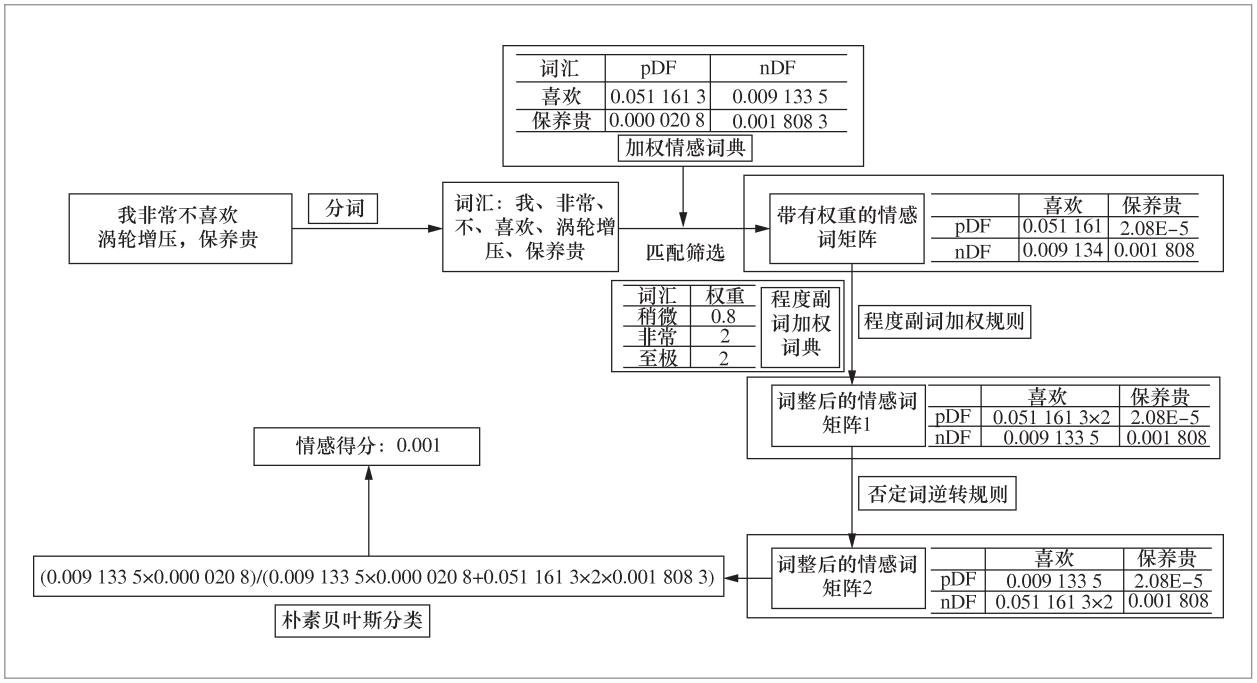

图 3 基于朴素贝叶斯算法情感分析流程示例

根据情感词的位置，在原文本中向前或向后查找程度副词。如果找到程度副词，则根据规则调整情感词矩阵，如本例在“喜欢”的前面找到了程度副词“非常”，且“喜欢”的正（pDF）负（nDF）情感倾向中较大者为pDF，所以将其乘以程度副词“非常”的权重2，其nDF不作改变；情感词“保养贵”前后均未找到程度副词，所以其权重不作调整，这样就获得了调整后的情感词矩阵1。

第三步，调整后的情感词矩阵1中，根据情感词的位置，在原文本中向前查找否定词，如果找到否定词，则根据规则调整情感词矩阵，如本例在“喜欢”的前面找到了否定词“不”，“喜欢”的正负情感倾向进行逆转，即将喜欢的pDF替换为nDF，nDF替换为pDF，双方互换在情感词矩阵中的位置；情感词“保养贵”前面没找到否定词，所以其权重不作调整，这样就获得了调整后的情感词矩阵2。

最后根据调整后的情感词矩阵2，构建朴素贝叶斯分类器计算情感得分，求出所有情感词pDF的乘积，然后计算其与所有情感词pDF的乘积加上所有情感词nDF的乘积之和的商值作为情感得分，可以得到文本的情感得分为0.01，较严重的负向倾向。

经测试集测试，朴素贝叶斯文本情感分析引擎的情感判别准确率较高，达到86.7%，并能准确应对否定句、双重否定及程度副词等在网络语言中较为普遍的句法、语法类型。

6 结束语

本文提出了一种获得情感特征词权重的量化方法，并设计了两个比较常见的汉语规则，用于调整情感权重，结合常见的监督型机器学习算法取得了86.7%的分类准确率。对比其他企业落地实施的情感分析引擎，本系统取得了不错的成绩。系统弱化了行业特异性，更加易于移植到其他行

业，整个分析体系弱化了人工干预和标注的工作，更加符合企业应用减少人工的需求，就其分析逻辑而言，很容易与自然语言理解领域其他研究模块结合，比如与句法解析结合，解决情感归属问题。

情感分析正在向语义级别发展，但其在企业应用中的需求至少满足两个方面：其一，情感越来越细腻；其二，情感归属问题。情感越来越细腻，包括实际情感的细化，但随着品牌舆情管理的细化，可能需要更加细粒度的情感分类，比如喜欢、高兴、伤心、厌恶、憎恨等。每一种情感背后蕴含的看法和观点存在很大的不同，其中参考文献[18,19]对情感分析进行了更加细腻的探索研究。不同文本的情感程度是不一样的，而且其带来的社会影响也不同，因此除了区分情感的细分分类以外，企业需要更加细腻的情感程度衡量方式，即传统的二分类问题或多分类问题，转化为分类和连续的程度衡量问题。

除了细腻的情感分析以外，情感归属也是一个亟待解决的问题。情感归属正逐步深入自然语言理解的句法分析领域，它不仅要求句法分析做得优秀，而且情感分析做得也同样优秀，才能做到准确的情感归属。

参考文献:

[1] ZHAO Y Y, QIN B, LIU T. Sentiment analysis[J]. Journal of Software, 2010, 21(8): 1834-1848.

[2] KHARDE V A, SONAWANE S. Sentiment analysis of twitter data: a survey of techniques[J]. Computer Science, 2016: arXiv:1601.06971.

[3] KU L W, LO Y S, CHEN H H. Using polarity scores of words for sentence-level opinion extraction[C]//The 6th NTCIR-6 Workshop Meeting, May 15-18, 2007, Toyko, Japan. [S.l.:s.n.], 2007.

[4] KAJI N, KITSUREGAWA M. Building lexicon for sentiment analysis from massive collection of HTML documents[C]//The 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, June 28-30, 2007, Prague, Czech Republic. [S.l.:s.n.], 2007.

[5] PANG B, LEE L. Opinion mining and sentiment analysis[J]. Foundations & Trends in Information Retrieval, 2008, 2(1): 459-526.

[6] JUNGHERR A. Twitter use in election campaigns: a systematic literature review[J]. Journal of Information Technology & Politics, 2016, 13(1): 72-91.

[7] 宋玮，訾红.“被就业”事件网络舆情分析[J]. 河北广播电视大学学报，2010，15(3): 106-108.
SONG W, ZI H. Analysis of online public opinion on the “be job hunted” scandal[J]. Journal of Hebei Radio & TV University, 2010, 15(3): 106-108.

[8] 赵晋. 浅析网络舆情分析在企业品牌危机管理中的应用[J]. 新闻世界，2008(12): 97-98.
ZHAO J. Analysis of the application of network public opinion in enterprise brand crisis management[J]. News World, 2008(12): 97-98.

[9] AKKAYA, C, CONRAD A, WIEBE J, et al. Amazon mechanical Turk for subjectivity word sense disambiguation[C]//The NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk, June 6, 2010, Los Angeles, California. Stroudsburg: Association for Computational Lin guistics, 2010.

[10] TABOADA M, BROOKE J, TOFILOSKI M, et al. Lexicon-based methods for sentiment analysis[J]. Computational Linguistics, 2011, 37(2): 267-307.

[11] 张成功，刘培玉，朱振方，等. 一种基于极性词典的情感分析方法[J]. 山东大学学报(理学版)，2012，47(3): 50-53.

ZHANG C G, LIU P Y, ZHU Z F, et al. A sentiment analysis method based on a polarity lexicon[J]. Journal of Shandong University, 2012, 47(3): 50-53.

[12] FIRMINO A A, PAIVA A C D. A comparison of SVM versus naive-Bayes techniques for sentiment analysis in tweets: a case study with the 2013 FIFA confederations cup[C]//The 20th Brazilian Symposium on Multimedia and the Web, November 18-21, 2014, João Pessoa, Brazil. New York: ACM Press, 2014: 123-130.

[13] 孙建旺, 吕学强, 张雷瀚. 基于词典与机器学习的中文微博情感分析研究[J]. 计算机应用与软件, 2014, 31(7): 177-181.
SUN J W, LV X Q, ZHANG L H. On sentiment analysis of Chinese microblogging based on lexicon and machine learning[J]. Computer Applications and Software, 2014, 31(7): 177-181.

[14] 关雅夫. 基于主动学习的微博情感分析方法研究[D]. 长春: 吉林大学, 2017.
GUAN Y F. Research on microblog sentiment analysis based on active learning[D]. Changchun: Jilin University, 2017.

[15] 樊娜, 安毅生, 李慧贤. 基于K-近邻算法的文本情感分析方法研究[J]. 计算机工程与设计, 2012, 33(3): 1160-1164.
FAN N, AN Y S, LI H X. Research on analyzing sentiment of texts based on k-nearest neighbor algorithm[J]. Computer Engineering and Design, 2012, 33(3): 1160-1164.

[16] 李英. 基于词性选择的文本预处理方法研究[J]. 情报科学, 2009, 27(5): 717-719.
LI Y. Research on the text pretreatment based on part of speech selection[J]. Information Science,2009, 27(5): 717-719.

[17] 张宁. 基于语义的中文文本预处理研究[D]. 西安: 西安电子科技大学, 2011.
ZHANG N. Research of chinese text preprocessing based on semantic[D]. Xi'an: Xidian University, 2011.

[18] TOKUHISA R, INUI K, MATSUMOTO Y. Emotion classification using massive examples extracted from the Web[C]//The 22nd International Conference on Computational Linguistics, August 18-22, 2008, Manchester, United Kingdom. Stroudsburg: Association for Computational Linguistics, 2008: 881-888.

[19] YANG Y H, LIU C C, CHEN H H. Music emotion classification: a fuzzy approach[C]//The 14th ACM international conference on Multimedia, October 23-27, 2006, Santa Barbara, USA. New York: ACM Press, 2006: 81-84.

作者简介

宋云生(1985-), 男, 深圳联友科技有限公司数据挖掘工程师, 主要研究方向为自然语言理解及深度学习。

收稿日期: 2017-09-07