计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.210200089

考虑语境的微博短文本挖掘:情感分析的方法

史 伟[1] 付 月[2]
1 湖州师范学院经济管理学院 浙江 湖州 313000
2 湖州师范学院求真学院 浙江 湖州 313000

摘 要 传统基于词典的情感分析方法中情感词语的极性和强度是固定和静态的,没有考虑情感词语随不同语义环境极性和强度的变化。为此,提出一种考虑语境的基于情感本体和情感圈的微博短文本情感分析方法。采用情感圈方法考虑不同语境中词语的共现模式,以捕获它们的语义并更新情感词语的极性和强度。结合已构建的情感本体和语义量化规则,建立考虑语义环境的微博短文本挖掘方法。实验结果表明,该方法从实体级和微博级两个层面,在精度、召回率、F值和准确率几个指标上都明显优于基线方法。

关键词:微博短文本;情感本体;情感圈;情感分析
中图法分类号 TP391.1

Microblog Short Text Mining Considering Context:A Method of Sentiment Analysis

SHI Wei[1] and FU Yue[2]
1 School of Economics and Management,Huzhou University,Huzhou,Zhejiang 313000,China
2 Qiuzhen College,Huzhou University,Huzhou,Zhejiang 313000,China

Abstract In the traditional dictionary based sentiment analysis,the polarity and intensity of sentiment words are fixed and static,without considering the change of polarity and intensity of sentiment words with different semantic environments. This paper proposes a sentiment analysis method of microblog short text based on sentiment ontology and sentiment circle considering context semantics. In order to capture their semantics and update the polarity and intensity of emotional words,we use the sentiment circle method to consider the co-occurrence patterns of words in different contexts. Combined with the constructed emotion ontology and semantic quantitative rules,a method of microblog short text mining considering semantic environment is established. The experimental results show that the proposed method is superior to the baseline method in terms of accuracy,recall,*F* value and accuracy from both entity level and microblog level.

Keywords Microblog short text,Sentiment ontology,Sentiment circle ,Sentiment analysis

1 引言

微博情感分析引起了人们的广泛关注,因为微博为人们表达对各种话题的看法和态度提供了一种平民平台。微博中的信息主要以短文本的形式存在,微博短文本情感分析的方法主要集中于个体微博情感的识别(即单条微博级情感检测)。一般而言,目前的微博级情感检测工作主要采用两种方法:基于机器学习的方法和基于词典的方法。

机器学习方法需要为情感分类器学习训练数据。在微博中,训练数据有些通过情感符号假设微博的极性(正面、负面和中立),有些则从情感检测网站返回的结果中获得共识。而且监督方法是领域依赖的,需要对新的数据进行重新训练。鉴于微博中不断涌现的不同的主题,领域依赖限制了这种方法的应用。另一方面,基于词典的方法则不需要训练数据,相反,它们使用所有情感词汇加权来确定给定文本的整体情感倾向。这些方法在常规文本中显示出了它们的有效性[1]。然而,传统的词汇往往不适合微博文本的分析,因为微博文本中包含大量的畸形词和口语表达(例如"ky""ssfd""猴腮雷")。此外,许多基于词典的方法还利用句子的词汇结构来确定其情感,这在微博中是有问题的,因为微博一般都是短文本,非语法的句子非常常见。为了解决这些问题,Shi 等于 2015 年构建了一种基于情感本体和语义的社交化短文本情感分析方法,称作 EOSentiMiner[2]。虽然构建的 EOSentiMiner 和情感本体在相应的数据集中取得了良好的情感分析效果,但是和其他基于词典的方法类似,其中的情感本体同样面临两个主要的问题。首先,EOSentiMiner 的准确性召回率受限于情感本体中的固定词集,如果词语未在情感本体中则在情感分析中就很难被考虑,这在处理微博文本时会成为一个问题,因为微博中新的表达和隐语不断涌现。其次,更为重要的是,EOSentiMiner 提供的是固定的、上下文语境无关的情感词的极性和强度,但是在实际的很多微博文本中不同的词语在不同的上下文语境中却表现出了不同的情感极性和强度。如下

基金项目:国家社会科学基金一般项目"重大突发事件中网民情感状态演变规律及引导研究"的阶段性成果(20BXW013)
This work was supported by the General Program of National Social Science Foundation of China "Research on the Evolution Law and Guidance of Netizens' Sentiment State in Major Emergencies"(20BXW013).
通信作者:史伟(shiwei@zjhu.edu.cn)

面两个文本“这款手机价格真便宜啊！”和“这款手机的设计真便宜！”，其中情感词“便宜”在两个文本中体现出了不同的情感极性和强度。为此，获取情感词极性和强度在上下文语境中的变化，并以此构建更为准确和高效的微博情感分析系统成为了本文的主要动机。

本文第 2 节介绍了关于微博级和实体级情感分析的相关工作；第 3 节提出了词语的情感圈表示；第 4 节描述了如何应用情感圈进行情感分析；第 5 节和第 6 节主要介绍了实验的设置和结果分析；最后总结全文并展望未来。

2　相关工作

大多数现有的微博情感分析方法的主要工作是将个人微博进行情感分类（正面/负面，或细粒度分类）。这些方法可以被归为两类：基于机器学习的方法（需要训练数据）和基于词典或语义的方法（预先设定情感极性和强度的词语集和情感语义规则）。

2.1　基于机器学习的方法

机器学习方法是基于训练分类的，如支持向量机（SVM）、贝叶斯（NB）、卷积神经网络（CNN）等，考虑各种组合特征，如 n-grams、词性（POS）、情感主题特征、语义模式、微博语法特征等。Yi 等应用机器学习算法对基于顾客体验的产品信息和商店信息进行学习和分类，分析结果发现，机器学习算法的性能优于其他方法[3]。Fusilier 等基于 PU 学习算法提出了新的学习模型，并在此基础下训练朴素贝叶斯分类器，获得了较好的结果[4]。秦锋等将微博情感分析问题看作标签序列学习任务，使用隐马尔可夫支持向量机（SVM）把微博上下文语境融入微博情感分析问题中，实验结果表明，该方法相比基于朴素贝叶斯或支持向量机的微博情感分析模型可以更好地分析微博情感极性[5]。卢欣等提出了一种融合语言特征的卷积神经网络（CNN）的反讽识别方法，该方法将反讽特征和句子分别采用 Word Embdding 作为输入，在卷积、池化后，将其全连接融合，构建了新的卷积神经网络模型，实验结果表明该方法在反讽识别的性能上优于传统的基于机器学习的方法[6]。

2.2　基于词典语义的方法

基于词典的方法试图通过在情感词典中预先设定好情感极性和强度的词语，来计算给定文本的整体情感。不同于机器学习方法需要使用训练数据，基于词典方法依赖于预先构建的具有相关情感极性的词汇词典，如英文词典 SentiWordNet、LIWC 词典和 MPQA 主观性词典，中文词典如知网（HowNet）、大连理工大学中文情感词典库和台湾大学的情感词典等。相关学者基于基础词典做了扩展和应用工作，取得了一些研究进展，吴杰胜等对传统情感词典进行了改进与扩展，包括构造了程度副词、否定词词典、微博领域词典等相关词典，结合多部情感词典和规则集实现了对微博的情感分析，取得了一定效果[7]。李继东等提出了基于扩展词典与语义规则的中文微博情感分析方法，构建基础情感词典，结合微博表情词典、否定词典、程度词典等扩展词典，结合语义规则对中文微博进行了情感分析，实验结果证明了该方法的有效性[8]。Ahmed 等提出了一种建立领域相关情感词典的新方法 sentidomain，该方法是弱监督神经模型，主要目的是从目标域的句子全局表示中学习一组嵌入的情感聚类，实验结果表明该方法有效提高了极性检测的性能[9]。Ismail 等将词典方法和机器学习方法进行结合，定义了一种新的、全自动的与领域无关的方法，从 Twitter 文本语料库中构建特征向量，用于基于模糊词表和情感替换的情感分析中，实验结果显示该方法领先于各种基线方法[10]。景丽等也综合了情感词典和机器学习两种方法的特点，构建了一个网络评论情感分类模型，该模型的分类正确率相比其他方法有所提升[11]。

2.3　研究评述

总结相关研究发现，基于微博的情感分析还存在以下一些局限：1）缺乏对情感词语在上下文语义中情感极性和强度动态变化的研究；2）考虑上下文语义的细粒度微博情感分析的研究工作还很少，情感分析的有效性有待提高；3）缺乏从实体级和微博级同时展开情感分析比较研究的工作，相应的中文语料库的构建有待加强。

本文将采用情感圈（SentiCircles）方法[12]，捕捉词汇上下文语义（如文本中词语的语义共现模式）以建立词语的动态表示，适时调整已构建的情感本体中情感词的情感极性和强度。Wittgenstein 于 1953 年提出的上下文语义（aka 统计语义）已经被广泛应用于计算机科学的不同领域，如自然语言处理和信息检索方面[13]。上下文语义背后的主要理论是在一个给定文本中同时出现的词语倾向于具有一定的关系或语义影响，本文就用情感圈进行这种语义关系的捕捉工作。在两个不同的情感分析任务中对提出的情感圈的方法进行评价：1）实体级情感检测，针对特定的实体或主题检测情感（如手机、电脑、电影等）；2）微博级情感检测，针对单条微博文本进行总体情感极性和强度的判断。利用情感圈表示上述几种三角恒等式来执行这两类情感分析任务。

本文的主要贡献如下：

（1）介绍和发展了一种新的基于情感本体和情感圈的情感分析方法，运用情感圈从词语所处语境的共现模式中捕获词语的潜在语义，并相应地更新它们在情感本体中的情感取向。

（2）结合语义量化规则，提出了中文情感圈的语义表示和情感值计算方法。

（3）进行了一系列实验，证明了本文方法在实体级和微博级情感检测中的有效性。

3　词语的情感圈表示

本节主要介绍情感圈方法和应用它进行词语的上下文语义和情感捕获的过程。情感圈方法主要词语的上下文语义中获取它的情感倾向。这个背后主要的观念认为词汇的情感不同于传统基于词典的方法中的固定和静止，词汇的情感依赖于词汇的上下文，如依赖于它的上下文语义。我们将上下文语境定义为一个文本库或一组微博集。为了获取词语的上下文语义，遵循如下的分布假说：在相似上下文语境中出现的词语往往具有相似的意义[14]。因此，在本文方法中词语 m 的上下文语义是通过它与其他词语的共现模式计算出来的。

图 1 给出了本文方法的系统工作流程，可以概括为如下几个步骤：

（1）词语索引。该步骤主要从微博文本集中创建词语索引。多个文本处理的程序被应用到这个过程中，如对在空白边界上的个别词进行分离；从词语中去除所有非文字的数字字符；去除 1208 个标准停用词，包括常见的一些动词；为了避

免垃圾信息和其他一些不相关的微博信息，从微博中过滤掉额外的链接，如含有“http:”或者“www.”的表达和用户的名字(用符号@标志的)；移除“回复”“转发微博”等词和转发的内容(只是转发没有增加任何评论的帖子)；基本词性标注(POS)和否定处理(见 3.2 节)。

(2)词语上下文语境向量的生成。该步骤主要将词语 m 表示成由其微博语境中所有的词语所组成的一个向量(即与词语 m 同时出现在相同的语境中)。具体见后文的定义。

(3)上下文语境特征的生成。计算每一个词语与语境中其他词语的相关度。同时使用外部情感本体为这些语境中的词语分配初始情感值。

(4)情感圈的生成。该步骤将 m 的词语语境向量转换为 2D 几何圆，它由表示 m 的上下文语境词语的点组成。每个语境词语在圆中的位置是基于它的角度(由其先前的情感决定)和它的半径(由它和词语 m 的相关度决定)(详见 3.1 节)来确定的。

(5)情感识别。本文采用不同的方法，主要是利用情感圈上的几个三角恒等式在实体级或微博级进行情感识别(详见第 4 节)。

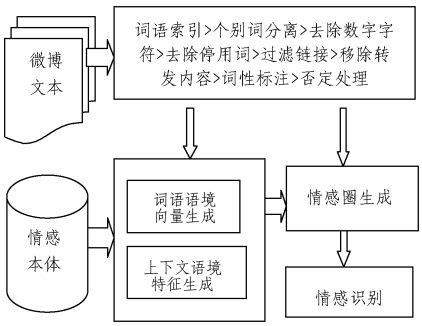


图 1 基于情感圈的情感分析方法系统工作流程图

Fig. 1 Workflow of sentiment analysis method system based on sentiment circle

定义 1(词语上下文语境向量) 给定一组微博文本集 T，一个词语 m 的语境向量 $\vec{c}=(c_1,c_2,\cdots,c_n)$，$c$ 是在 T 的任何一条微博文本中与词语 m 共同出现的语境词语。M 的语境语义是由它与每个语境词语 $c_i\in\vec{c}$ 的语义关系决定的。通过计算 c_i 的如下两个主要特征来确定 m 和语境词语 c_i 间的语义关系。

(1)先验情感值：基于已经构建的情感本体，确定每个语境词 c_i 的初始情感值。

(2)词语的相关度(CDOT)：这个特征表示词语 m 和它的语境词语 $c_i\in\vec{c}$ 的相关程度(即 c_i 相对于 m 的重要程度)。参考 TF-IDF 加权方法，该特征值的计算方法如下：

$$CDOT(m,c_i)=f(c_i,m)\times\log\frac{N}{N_{c_i}} \tag{1}$$

其中，$f(c_i,m)$ 为 c_i 和 m 共同出现在微博文本中的次数，N 是微博文本中所有词语的总数，N_{c_i} 是微博文本中所有 c_i 的总数。

3.1 语义的情感圈表示

现在每个词语 m 都有一个上下文语境词 $\vec{c}$ 的向量以及 m 和 $c_i\in\vec{c}$ 之间的两种语义相互特征。根据这些信息，将词语 m 的上下文语义表示为一个几何圆——情感圈，其中词语 m 位于圆的中心，围绕它的每个点表示语境词 c_i。c_i 的位置由它的先验情感和词语相关度(CDOT)共同决定。使用这种圆形表示词语上下文语义，主要基于它能提供三角属性评估词语的情感极性和强度。它还能够分别计算上下文词语对目标词语的情感极性和强度的影响，这是传统的向量表示方法难以做到的。情感圈在极坐标系统中可以用如下计算式表示：

$$r^2-2rr_0\cos(\theta-\phi)+r_0^2=a^2 \tag{2}$$

其中，a 是圆的半径，(r_0,ϕ) 是圆中心的极坐标，(r,θ) 是一个语境词语在圆上的极坐标。为简单起见，假设情感圈的中心在原点(即 $r_0=0$)。

因此，若要为词语 m 构建一个情感圈则只需计算语境词语 c_i 的半径 r_i 和角度 θ_i。这里用 c_i 的先验情感值 PS(Prior Sentiment)和词语相关度(CDOT)来表示。

$$\begin{aligned}r_i&=CDOT(m,c_i)\\\theta_i&=PS(c_i)*\pi\end{aligned} \tag{3}$$

在一个情感圈中将所有词语的半径都标准化为 0 到 1 之间，因此任何一个情感圈的半径 a 都是 1，同样所有角度值都是弧度。情感圈在极坐标系中可被分为 4 个情感象限，如图 2 所示。其中位于两个上象限中的词语具有正面情感($\sin\theta>0$)，左上象限表示更强的情感，因为它比右上象限中的词语具有更大的角度值。同样位于两个下象限中的词语具有负面情感值($\sin\theta<0$)。尽管对于任何一个词语 m 的情感圈的半径都等于 1，但是表示 m 语境词语的各个点在圆中的半径就各不一样了($0\leqslant r_i\leqslant 1$)，半径的大小反映了语境词语对于 m 的重要性，这里定义半径越大，语境词语对于 m 就越重要。

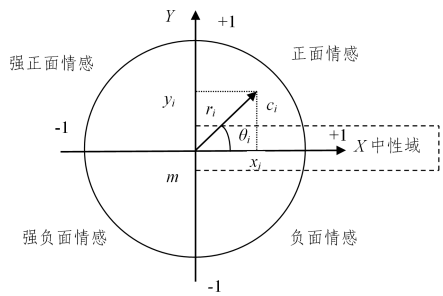


图 2 词语 m 的情感圈

Fig. 2 Sentiment circle of word m

可以通过使用三角函数正弦和余弦将极坐标系转换到平面直角坐标系(笛卡尔坐标系)，如式(4)所示，将坐标系转换为笛卡尔坐标系后就可以使用圆的三角形属性对词语的上下文语义进行编码，以此作为情感极性和情感强度值。笛卡尔坐标系中的 Y 轴表示词语的情感极性，即如果 y 为正值则表示为正面情感，反之亦然。X 轴表示词语的强度，x 值越小，则情感越强。此外，还定义了一个叫作“中性域”的小区域，如图 2 所示，这个区域位于“正”和“负”象限中非常接近 X 轴的位置，位于该区域的词语情感非常弱(即 $|\theta|\approx 0$)，这个“中性域”在对给定情感圈进行整体情感衡量时具有关键的作用，后文将进行具体介绍。在极端情况下，当 $r_i=1$ 和 $\theta_i=\pi$ 同时发生时，语境词语 c_i 是位于“非常正”还是“非常负”象限主要取决于其先前的情感极性。

$$x_i=r_i\cos\theta_i,\ y_i=r_i\sin\theta_i \tag{4}$$

图 3 给出了实体“华为 P20”的情感圈。情感圈中词语(即点)所在的位置表示了它们对于实体“华为 P20”的情感值和重要性(相关度)。位于情感圈上半部分的点(菱形)表示带

有正面情感的词语，而下半部分的点（圆形）则表示带有负面情感的词语。例如情感圈中的“爱”具有正面情感和较高的重要性，因为表示“爱”的点位于强正面情感象限，离原点“P20”的距离较远。词语“生动”也具有正面情感，但它的情感强度和重要性都不如词语“爱”，因为表示“生动”的点位于正面情感象限，而且离原点“P20”的距离较近。

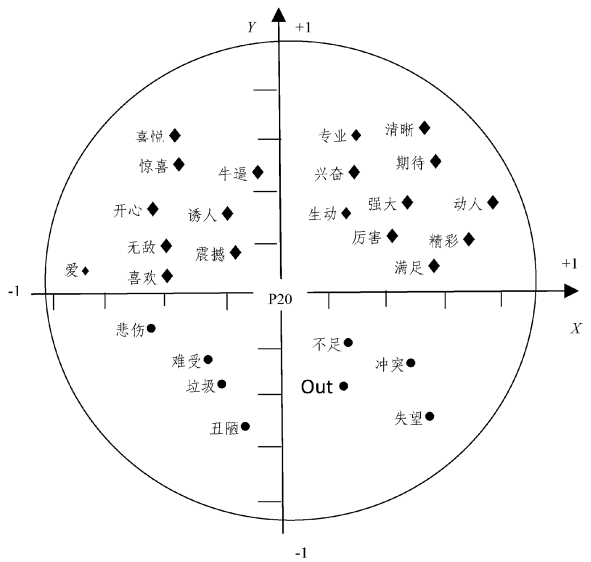


图 3　“华为 P20”的情感圈

Fig. 3　Sentiment circle of “Huawei P20”

3.2　情感圈中的否定和程度表达处理

当构建情感圈时，如果词语 t 周边有否定表达出现，则 t 的相关情感值 SO_t 在情感圈构建中取反，参考 Shi 等[2]的处理方式，从 HowNet 中人工抽取出 22 个否定词，当句子中出现了满足规则的否定词时，则利用式(5)对词组的情感类进行调整。本文为情感词上下文设置了一个大小为 5 的检测窗口。若在检测窗口内出现否定词，则对词语情感极性取反。

$$SO_t = (-1)^n \times SO_t' \tag{5}$$

其中，SO_t'为词语 t 在情感本体中的原始情感值；n 为满足否定规则时对于词语 t 而言否定词的出现次数。例如，微博“我对这款华为 P20 不满意！”，其中情感词“满意”（情感本体中的情感值为 0.44，情感类为高兴）前有否定词，则它的情感值＝(－1) * 0.44＝－0.44，属于情感类“高兴”的强度是－0.44。

当构建情感圈时，如果词语 t 周边有程度表达出现，则对 t 的相关情感值 SO_t 在情感圈构建中进行相应的调整，为了准确地衡量微博的情感强度，在情感词的上下文设置一个检测窗口，本文采用的窗口大小为 5。如果在检测窗口内有程度词出现，则按程度词的等级差别相应增加情感词的情感强度，从高到低依次增加 1.5 到 0.8 倍。从 HowNet 中抽取 60 个程度词并将其分成 7 类，具体设置如表 1 所列。

表 1　程度词赋值表

Table 1　Assignment table of degree words

程度词	1.5	最、最为、极、极为、极其、极度、极端
	1.4	太、至、至为、顶、过、过于、过分、分外、万分、何等
	1.3	很、挺、怪、老、非常、特别、相当、十分、甚、甚为、异常、深为、蛮、满、够、多、多么、殊、何其、尤其、无比、尤为、超
	1.2	不甚、不胜、好、好不、颇、颇为、大、大为
	1.1	稍稍、稍微、稍许、略微、略为、多少
	0.9	较、比较、较为、还
	0.8	有点、有些

利用式(6)计算程度词结合情感词得到的情感值为：

$$SO_t = value_{deg} \times SO_t' \tag{6}$$

其中，SO_t'表示词语 t 的原始情感值；$value_{deg}$ 表示程度词 deg 的强度值。例如，“非常满意”的情感值＝1.3 * 0.44＝0.57 属于情感类“高兴”的强度是 0.57。

3.3　情感圈的语境情感值

如前文所述，使用情感本体对语境词语进行了情感值初始赋值，构建的情感圈能够根据上下文语境对这些词语的情感值进行修正。这里基于情感圈用语境情感计算词语新的情感值。词语 m 的情感圈由它的所有上下文语境词语的笛卡尔坐标(x,y)组成，其中 y 值表示情感极性，x 值表示情感强度。可以通过计算圈中所有点的几何中值来估计给定情感圈的整体情感。对于给定的情感圈中的 n 个点$(p_1,p_2,\cdots,p_n)$，它的二维几何中值 g 定义为：

$$g = \arg\min_{g\in R^2} \sum_{i=1}^{n} \| p_i - g \|_2 \tag{7}$$

其中，几何中值为点 $g=(x_k,y_k)$，该点到所有其他点 p_i 的欧氏距离是最小的。这里称几何中值 g 为情感中值，因为它可以表示给定词语 m 的情感圈的情感极性（y 坐标）和情感强度（x 坐标）。

4　基于情感圈的情感分析

本节将介绍基于情感圈进行两种不同的情感分析任务：实体级和单条微博级情感检测。

4.1　实体级情感检测

给定一个实体 $e_i \in \varepsilon$ 和它相应的情感圈，这个实体的情感可由情感圈的情感中值 g 表示（即组成情感圈的所有点的几何中值）。根据图 2 的描述，如果情感中值 g 位于“中性域”，则实体具有中性情感；如果 g 位于正面情感象限，则实体具有正面情感；如果 g 位于负面情感象限，则实体具有负面情感。给定一个实体 e 的情感中值 g_e，则实体情感函数 γ 为：

$$\gamma(g_e)=\begin{cases} positive, & \text{if } y_g > +\sigma \\ neutral, & \text{if } |y_g| \leqslant \sigma \,\&\, x_g \geqslant 0 \\ negative, & \text{if } y_g < -\sigma \end{cases} \tag{8}$$

其中，σ 是定义“中性域”Y 轴边界的阈值，后文会介绍如何计算这个阈值。

4.2　单条微博级情感检测

给定单条微博 $t_i \in T$，有几种方法可以运用微博中词语的情感圈进行整体情感的确定。例如，“华为 P20 的拍照功能很好”包括 3 个实词：“华为 P20”“拍照功能”“很好”，每个词语都有一个相关的情感圈的表示。这 3 个情感圈可以进行不同的组合以提取与这条微博相关的情感。这里介绍 3 种不同的方法来探讨用情感圈进行单条微博级的情感检测。

4.2.1　中值法

该方法主要是将每条微博 $t_i \in T$ 表示为情感中值$\vec{g}=(g_1,g_2,\cdots,g_n)$的一组向量，其中 n 为组成该条微博的实词的数量，g_j 为实词 m_j 的情感圈的情感中值。运用式(7)计算情感中值 g_j，然后对向量 $\vec{g}$ 中的所有情感中值求取平均值，最后运用式(8)确定微博 t_i 的整体情感。

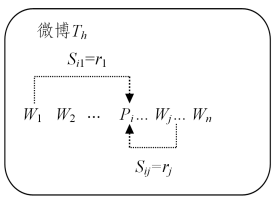
4.2.2　关键词法

该方法将关注点放在微博中的一些关键词上，假设微博中的情感表达总是针对一个或多个特定目标，则称这些特点

目标为关键词。在前面提到的微博例子中，有两个关键词“华为 P20”和“拍照功能”，情感词“很好”用来描述它们。因此，该方法可通过以下两种方式进行操作。1)提取微博中所有的关键词，关键词提取过程为：对微博文本进行分词和词性标注，去除停用词，采用文献[15]中的方法，运用规范化的 TFIDF 加权方法提取出有代表性的名词和代词作为关键词。2)为每个关键词累计它接收到的来自其他词语的情感影响。微博的整体情感对应于关键词接收到的那个最高情感影响。针对每个候选关键词，构建相应的情感圈，以此计算微博中其他词语对于关键词的情感影响。最大情感影响 $\hat{s}$ 的计算式如下：

$$\hat{s}=\arg\max_{s\in S}H_s(\vec{p})=\arg\max_{s\in S}\sum_{i}^{N_{\vec{p}}}\sum_{j}^{N_{\vec{w}}}H_s(p_i,w_j) \tag{9}$$

其中，$s\in S=\{$正面，中性，负面$\}$ 表示情感极性，$\vec{p}$ 是单条微博中所有关键词的向量，$N_{\vec{p}}$ 和 $N_{\vec{w}}$ 分别表示单条微博中关键词集和剩余词集，$H_s(p_i,w_j)$ 表示情感影响函数，即情感圈中词语 w_j 对关键词 p_i 的情感影响，情感影响程度(情感强度)为词语 w_j 到 p_i 的欧氏距离(即词语 w_j 的半径)，如图 4 所示。如果词语 w_j 落在“强正面情感”或“强负面情感”象限，则情感影响值加倍。



注：s_{ij} 为词语 w_j 对于 p_i 的情感强度，r_j 为在 p_i 的情感圈中词语 w_j 到 p_i 的半径

图 4 关键词法

Fig. 4 Key words method

4.2.3 混合法

该方法主要是将前面介绍的两种方法结合起来使用。如 4.2.2 节的介绍，关键词法主要依赖微博的句法结构和词语间的情感关系，但当有些微博文本过短、缺乏关键词或者微博中包含大量病态词语时，这种方法就无法使用。若遇到这种情况我们就转而应用中值法，将两者结合起来使用的方法就叫作混合法。

5 实验和数据分析

如第 4 节所述，通过情感圈表示获取上下文语义主要基于语料库中的词语共现和情感词典中的初始情感权重集。本文提出一种使用两个不同语料库(微博集合)和一个通用情感词典的评价设置，以使我们能够评价不同语料库和词典对情感圈方法表现的影响。

5.1 数据集

本节将介绍用于评估的两个数据集：“电影评论”[16]和“手机微点评”[15]。使用“电影评论”(由史伟等于 2015 年提出)数据集来评估本文方法在单条微博水平的性能，因为它们只为单条微博而不是实体提供人工标注(即每条微博都被赋予了 8 类情感类和 2 类评价类，将期待、高兴、喜爱、惊讶和 G(好)类评价归为正面情感，将焦虑、悲伤、生气、讨厌和 B(坏)类评价归为负面情感)。

使用“手机微点评”(由史伟等于 2014 年提出)数据集评估实体层面情感。该数据集包含微博和实体情感评价，因此我们在本文中使用它来评估情感圈在实体层面和微博层面的表现。

表 2 列出了两个数据集中的正面和负面微博的数量。

表 2 微博评估的数据集

Table 2 Data set of microblog evaluation

数据集	微博数	正面情感	负面情感
“电影评论”	92701	49652	43049
“手机微点评”	9878	7985	1893

5.2 情感本体

如第 3 节所描述的，情感圈中词语的初始情感值是由某个情感词典赋值的(先验情感值)。这里使用已经构建的模糊情感本体库来评估本文方法。在前期研究中已详细论述了情感本体的构建过程[17]，创建了可用于在线评论情感分析的情感词本体库，并基于此本体库进行了系列情感分析研究[2,15-16]，并取得了非常好的效果。主要创新之处是将情感本体划分为评价词本体和情感词本体，利用模糊理论和知网模型，构建情感本体的基本模型。根据评价词和情感词的各自特点，运用模糊化处理和语义相似度的相关理论，分别对评价词本体和情感词本体的情感类型和隶属度进行了相应处理。情感本体的具体形式如下：

$FEO=((18;$开心$;$happy$;$adj$;$张三$;$知网 2007 版情感分析用词语集$),($快乐$;$愉快$),($高兴$;1.00))$

最终的情感本体收录了 9952 个词条，各类情感(2 种评价类和 8 种情感类)统计如表 3 所列。

表 3 各情感类的词汇数量

Table 3 Number of sentiment words

情感类	G(好)类评价词	B(坏)类评价词	期待	高兴	喜爱	惊讶	焦虑	悲伤	生气	讨厌
词汇数	3715	3147	170	395	339	65	271	220	201	429

各情感类词汇分别赋予了相应的情感类和情感隶属度值(情感值)，情感隶属度取值范围为$[0,1]$，可用于分析微博的情感极性和强度。情感有正面和负面之分，即情感极性。上述 8 类情感中期待、愉快、喜爱属于正面情感，而悲伤、生气和讨厌则属于负面情感，惊讶和焦虑在不同的语境下既可能表现为正面也可能为负面。

5.3 基线方法

为了比较所提出的情感圈在微博和实体情感分析中的表现，这里的基线方法考虑两个层次：基于情感词典的方法和基于情感计算的方法。

基于情感词典的方法，采用知网情感分析用到的词语集[18]，从给定文本中提取情感。如果一条微博文本包含的正面情感词多于负面情感词，则该条微博标记为正面，反之亦然。对于实体级情感检测，实体情感的标记是基于与实体共同出现在相关微博中的正面和负面情感词的数量决定的。

基于情感计算的方法[2,14]是基于情感本体和语义的比较

先进的情感检测方法。采用此方法对单条微博进行情感计算，如果通过计算得知正面情感强度大于负面情感强度，则该条微博被认为是正面的，反之亦然。对于实体级情感检测，一个实体的情感值是基于在一定窗口内与实体共同出现的正面情感词和负面情感词。这里微博情感强度的计算需要人工构建相应的语义规则，如程度词、否定词、标点符号、修辞句、表情符号等情感语义的量化处理。

6　实验结果

在实体级和微博级两个层面的情感检测任务中与基线方法进行比较，以验证所构建方法的性能。对于实体级的情感检测，我们在“手机微点评”数据集上进行实验，而对于微博级的情感检测，我们使用“电影评论”和“手机微点评”两个数据集。

6.1　对词语先验情感的影响

构建基于情感圈的情感分析法主要是因为词语的情感会随上下文语境的变化而变化。为了获得这些词语的语境语义，我们运用情感圈表示方法去调整词语的情感极性和强度。图 5 给出了在两个数据库中由情感圈改变词语初始情感极性和强度的平均比率，其中在我们的语料库中平均 68% 的词语被情感词本体库所覆盖，并被赋予了先验情感极性和强度，32% 的词语未在情感词本体库中找到，运用情感圈表示方法使得 45% 的词语重新调整了他们的情感极性（如从正面转变到负面或转变到中性），51% 的词语在未改变情感极性的情况下情感强度发生了变化，因此有 16% 的词语的原始的情感极性和强度未变化。另外，本文模型对 23% 未被情感词本体库覆盖的隐藏的词语赋予了情感极性和强度。

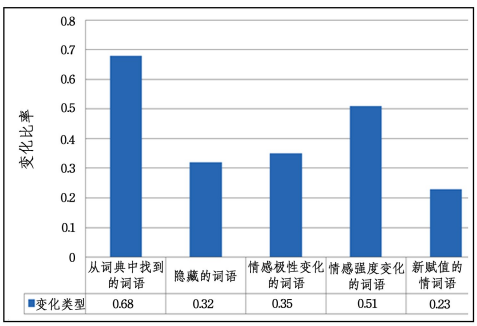


图 5　情感圈改变词语初始情感极性和强度的平均比率

Fig. 5　Sentiment circle changes the average ratio of initial sentiment polarity and intensity of words

6.2　实体级情感检测

对于实体级的情感检测，使用所提出的中值方法（见第 4.1 节），结合情感本体和语义规则来识别给定实体的情感圈的整体情感。我们用准确率、精度、召回率和 F 值来衡量两个识别任务的结果：主观性检测，它识别给定实体是主观的（正面的还是负面的）还是客观的（中立的）。第二个任务是情感极性检测，它识别实体是否有正面或负面情感。两种识别任务都应用于 10 个不同的实体（产品特征）[15]。

从表 4 可以发现，对于主观性识别，我们提出的基于情感圈的情感分析方法在 4 个指标上都大幅度领先基线方法。表 5 列出了实体级情感极性识别（正面或负面）的结果，基于情感圈的情感分析方法虽然没有全面大幅度领先基线方法，甚至在召回率上还落后基线方法 0.01，但是在其他 3 个指标上还是都略有提高。

表 4　实体级情感分析的结果比较（主观性检测）

Table 4　Comparison of entity level sentiment analysis results (subjectivity test)

方法	精度	召回率	F 值	准确率
基于情感词典方法	0.62	0.46	0.58	0.72
基于情感计算方法	0.63	0.56	0.62	0.75
基于 SentiCircle 的情感分析法（中值法）	0.80	0.83	0.81	0.82

表 5　实体级情感分析的结果比较（情感极性检测）

Table 5　Comparison of sentiment analysis results at entity level (sentiment polarity detection)

方法	精度	召回率	F 值	准确率
基于情感词典方法	0.68	0.72	0.79	0.69
基于情感计算方法	0.82	0.87	0.84	0.80
基于 SentiCircle 的情感分析法（中值法）	0.87	0.86	0.87	0.88

6.3　微博级情感检测

对于微博级情感检测，运用基于情感圈情感分析法中的中值法、关键词法和混合法，在“电影评论”和“手机微点评”两个数据库中进行了实验。同时将这些实验结果同两种基线方法（词典方法和情感计算方法）进行了比较。

图 6 给出了在“电影评论”语料库中微博级不同情感检测方法的准确率，基于情感圈情感分析法在准确率方面的表现普遍比基线方法更出色。同时还观察到，3 种情感圈方法中混合法比关键词法和中值法的准确率高，达到了 0.87。基线方法中情感计算方法的准确率与 3 种情感圈方法比较接近，都能达到 0.8 以上，词典方法由于考虑因素过少，因此准确率未能达到 0.6。图 7 给出了在“手机微点评”语料库中的几种情感检测方法的准确率，基本情况与图 6 中的表现相当，3 种情感圈方法的平均准确率达到了 0.84，比情感计算方法的表现略优。

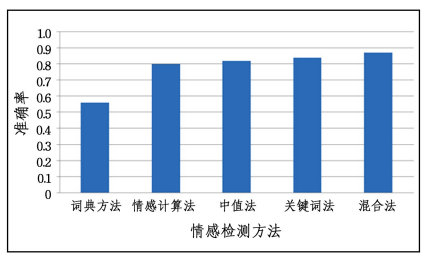


图 6　微博级不同情感检测方法的准确率比较（电影评论数据集）

Fig. 6　Accuracy comparison of different sentiment detection methods at micro blog level (movie review data set)

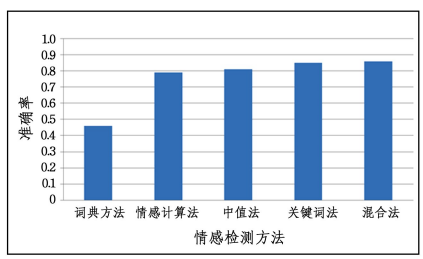


图 7　微博级不同情感检测方法的准确率比较（手机微点评数据集）

Fig. 7　Accuracy comparison of different sentiment detection methods at micro blog level

结束语　本文介绍了一种新颖的词语情感语义的表示方法，可为词语确定上下文语境中的情感倾向。结合已构建的

情感本体和情感语义量化规则，建立了基于上下文语境的微博短文本情感分析方法，通过实验描述了情感圈方法在情感检测（实体级和微博级）中的应用，并与基线方法进行了比较。本文方法无论在实体级和微博级多个衡量指标上都优于基线方法。本文构建的方法比基于情感计算的方法的表现更为优异，基于情感计算的方法主要是基于情感本体和语义规则对微博文本进行情感分析，词语的情感极性和强度为情感本体中预先设定的值，未体现出在不同上下文和不同语料库中的变化，而本文的基于情感圈的情感分析方法根据词语在微博上下文语境中不同的情感极性和强度动态地进行了更新和调整，使得情感检测的准确率更高。

本文选择的基线方法是两种基于语义的情感分析法，将来可将构建方法与一些机器学习方法进行比较，如支持向量机方法（SVM）等，同时也可选择新的语料库作为实验场景，以提高方法验证的全面性。未来也可考虑不同情感词典和微博中情感正负分布对构建方法有效性的影响。

参 考 文 献

[1] LIU B. Sentiment analysis and subjectivity [J]. Handbook of Natural Language Processing, 2010, 2: 568.

[2] SHI W, WANG H W, HE S Y. EOSentiMiner: an opinion-aware system based on emotion ontology for sentiment analysis of Chinese online reviews[J]. Journal of Experimental & Theoretical Artificial Intelligence, 2015, 27(4): 423-448.

[3] YI S S, LIU X F. Machine learning based customer sentiment analysis for recommending shoppers, shops based on customers' review[J]. Complex & Intelligent Systems, 2020, 6: 621-634.

[4] FUSILIER D H, MONTES-Y-GOMEZ M, ROSSO P, et al. Detecting positive and negative deceptive opinions using PU-learning[J]. Information processing & management, 2015, 51(4): 433-443.

[5] 秦锋，王恒，郑啸. 基于上下文语境的微博情感分析[J]. 计算机工程，2017，43(3)：241-246，252.

[6] 卢欣，李旸，王素格. 融合语言特征的卷积神经网络的反讽识别方法[J]. 中文信息学报，2019，33(5)：31-38.

[7] 吴杰胜，陆奎. 基于多部情感词典和规则集的中文微博情感分析研究[J]. 计算机应用与软件，2019，36(9)：93-99.

[8] 李继东，王移芝. 基于扩展词典与语义规则的中文微博情感分析[J]. 计算机与现代化，2018，270(2)：90-95.

[9] AHMED M, CHEN Q, LI Z H. Constructing domain-dependent sentiment dictionary for sentiment analysis[J]. Neural Computing and Applications, 2020, 32: 14719-14732.

[10] ISMAIL H M, BELKHOUCHE B, ZAKI N. Semantic Twitter sentiment analysis based on a fuzzy thesaurus[J]. Soft Computing, 2018, 22: 6011-6024.

[11] 景丽，李曼曼，何婷婷. 结合扩充词典与自监督学习的网络评论情感分类[J]. 计算机科学，2020，47(S2)：78-82.

[12] SAIF H, FERNANDEZ M, HE Y, et al. Senticircles for contextual and conceptual semantic sentiment analysis of Twitter [C]// Proc. 11th Extended Semantic Web Conf. (ESWC). Crete, Greece, 2014.

[13] WITTGENSTEIN L. Philosophical investigations[D]. London, UK: Blackwell, 1953.

[14] TURNEY P D, PANTEL P. From frequency to meaning: Vector space models of semantics[J]. Journal of Artificial Intelligence Research, 2010, 37(1): 141-188.

[15] 史伟，王洪伟，何绍义. 基于微博的产品评论挖掘：情感分析的方法[J]. 情报学报，2014，33(12)：1311-1321.

[16] 史伟，王洪伟，何绍义. 基于微博的情感分析的电影票房预测研究[J]. 华中师范大学学报（自然科学版），2015，49(1)：66-72.

[17] 史伟，王洪伟，何绍义. 基于知网的模糊情感本体的构建研究[J]. 情报学报，2012，31(6)：595-602.

[18] HowNet[R/OL]. HowNet sHomePage. http ://www. keenage. com.

SHI Wei, born in 1981, Ph.D, professor. His main research interests include business intelligence and affective computing.