OPEN:一个基于评论的商品特征抽取及情感分析框架

卿 勇[1,2] 刘梦娟[2] 薛 浩[2] 刘冰冰[2] 秦志光[2]

[1](达州职业技术学院 四川 达州 635001)

[2](电子科技大学信息与软件工程学院 四川 成都 610054)

摘 要 针对电商平台提出一个基于评论的商品特征抽取及情感分析框架,并将该框架在京东生鲜类商品的评论中进行应用。实验结果表明该框架确实能够成功抽取出商品的典型特征及该特征对应的情感极性,且在小样本数据集上测试了特征词和观点词抽取算法以及情感极性计算方法的性能,其中显式 < 特征词, 观点词 > 词对抽取的准确率达到了 53.6%,召回率达到了 81.5%,极性判断的准确率达到了 98.3%。主要贡献包括:提出一种依据观点词与特征词关联度的隐含特征词映射方法;基于 word2vec 词向量模型计算特征词相似度,并利用改进的半监督层次聚类算法对特征词进行典型特征聚类,建立特征词关联表。

关键词 特征提取 依存关系分析 word2vec 特征聚类 情感分析

中图分类号 TP391 **文献标识码** A DOI:10.3969/j.issn.1000-386x.2018.01.011

OPEN: A FRAMEWORK FOR PRODUCT FEATURE EXTRACTION AND SENTIMENT ANALYSIS BASED ON PRODUCT COMMENTS

Qing Yong[1,2] Liu Mengjuan[2] Xue Hao[2] Liu Bingbing[2] Qin Zhiguang[2]

[1](*Dazhou Vocational and Technical College, Dazhou 635001, Sichuan, China*)

[2](*School of Information and Software Engineering, University of Electronic Science and Technology of China, Chengdu 610054, Sichuan, China*)

Abstract This paper dedicates our work to propose a framework of product feature extraction and sentiment analysis based on comments of e-commerce platform. By applying this framework to the comments of fresh goods in JD.COM, the results of the experiments show that our framework can successfully extract the typical features and their corresponding sentiment polarities of every fresh goods. Also, we evaluated the performance of the feature and opinion extraction algorithms on a small data set; the accuracy and recall of extraction the explicit < feature, opinion > pairs reach 53.6% and 81.5% respectively. Furthermore, the accuracy of sentiment analysis reaches 98.3%. Overall, we make the following contributions: we proposed an implicit product feature mapping method, which is based on the correlations of opinions and product features; we proposed an improved semi-supervised hierarchical clustering algorithm to cluster product features and then establish an associative table of similar product features, where we use a new toolkit word2vec, to compute the similarity between any two feature words.

Keywords Feature extraction Dependence analysis word2vec Feature clustering Sentiment analysis

0 引 言

随着在线购物逐渐成为人们日常购物的重要途径,其质量安全问题日益受到政府监管部门、平台管理方、以及消费者的重视。然而由于电商平台本身低门槛、虚拟化等特点,导致电商平台上商品的质量安全难以得到有效监管。目前,大多数的电商平台都只提供商家的信用信息,例如商品质量、服务态度、物流速度、商品描述等的综合评分。针对单个商品,只有商家提

收稿日期:2017-04-10。国家自然科学基金重点项目和青年项目(61133016,61202445);达州市科技惠民计划项目(2015 1004)。**卿勇**,讲师,主研领域:物联网,数据挖掘。**刘梦娟**,副教授。**薛浩**,硕士生。**刘冰冰**,本科生。**秦志光**,教授。

供的商品信息,以及买家的评分和评论,导致用户在选购商品时只能根据好评率、以及对评论的浏览来了解商品的质量和特点。然而随着评论数的快速增长,评论内容越来越庞杂,甚至不同评论出现矛盾的观点,导致用户难以从评论中获得有效信息。为此,研究者提出利用情感分析技术进行评论总结[1],帮助用户从大规模评论集中挖掘商品的有效信息。

虽然已有大量针对商品评论意见抽取的研究成果,但仍然在以下方面存在改进空间:(1) 由于商品评论存在口语化、随意化等特点,因此一些隐含特征难以通过词频、依存关系等方法提取。例如评论"好吃,新鲜"中只有观点词,没有特征词,但这些观点词修饰的特征也是比较明确的。(2) 从评论中提取的特征词和观点词是多样化的,而这些特征词通常都可隐含地归纳为几类典型特征。例如"物流、速度、快递"都隐含对应了物流特征,"口感、味道"都隐含对应了品质特征。因此如果能将提取的特征词聚类为几类典型特征,提供这几类典型特征的情感分析,将使评论体现的商品特征及情感表述更为简洁。

本文针对上述问题,提出一个基于评论的商品特征抽取及情感分析框架(OPEN)。OPEN 首先利用依存关系和词性搭配规则提取每条评论中包含的 < 特征词, 观点词 > 词对;然后利用特征词的深度表示模型计算特征词的相似度,并基于一个改进的半监督层次聚类算法对特征词进行聚类,得到这些特征词属于的典型特征类别;最后计算商品每个特征词对应的情感极性,以及典型特征类别的情感极性。本文利用京东生鲜的评论数据进行验证,实验结果表明该框架确实能够成功抽取出每个生鲜商品的典型特征及该特征对应的情感极性。图 1 是利用 OPEN 框架对本文实验中的猪肉商品进行评论分析后的结果示意。表 1 是利用半监督聚类算法对该猪肉商品的高频特征词进行聚类的结果。

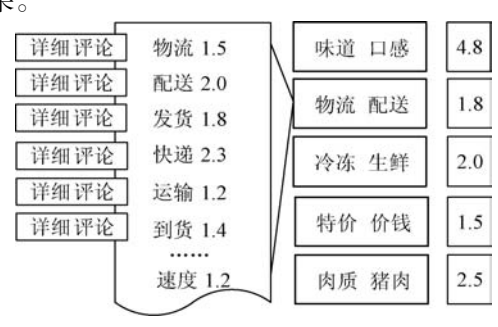

图 1 利用 OPEN 框架评论分析的结果示

表 1 OPEN 对高频特征词的聚类结果

特征类别	特征词
C1	味道 口感 炒菜 肉香 炖肉 香味 口味 肉味
C2	配送 发货 物流 快递 运输 到货 速度 师傅
C3	冷冻 生鲜 冰冻 保鲜 冷链
C4	特价 价钱 价格 经济 性价比 品质
C5	肉质 猪肉 肉馅 肉块

1 相关工作

情感分析又称为观点挖掘,是对带有情感色彩的主观性文本进行分析、处理、归纳和推理的过程。情感分析按照粒度可分为文本级、句子级、短语表达式级和单词级。本文针对商品特征的情感分析属于短语表达式级。最经典的面向产品特征的情感分析方法是 Hu 等在文献[1-2]中提出的,是目前产品特征情感分析的基本框架:首先通过关联规则算法提取频繁特征词;其次利用最邻近原则对频繁特征词邻近的观点词进行提取(通常为形容词),利用观点词在 WordNet[18] 词库里面搜索同义词和反义词,将其加入到观点词库;然后从观点词出发重新发现与观点词最邻近的不频繁特征词;最后根据所有抽取的特征词,利用其对应的观点词,统计每个特征词对应的正负极性值。Hu 的方法简单有效,但是只能针对评论中明确出现的特征词(名词或名词短语)进行提取,不能处理描述中包含的隐含特征词。

文献[3]在 Hu 的基础上,提出使用 Web PMI 指标来帮助提取与产品相关的特征词,例如产品的组成部件、特征等,通过语法依存关系来寻找特征词对应的观点词。文献[4]继续利用依存关系来提取特征词和观点词,作者指出该方法的关键是需要一个准确的依存关系分析工具。文献[5]的作者在比较了三款主流支持中文的依存关系分析工具后,提出可结合词性搭配规则来提取特征词和 < 特征词, 观点词 > 词对。文献[6]根据具体的餐饮点评场景提出了自己的词性搭配规则,其特色是首先将菜名进行了抽取,并通过建立领域知识库来帮助提升特征词和观点词的准确率。此外文献[7]还通过在文献[4]的基础上通过特征词-实词共现的频率矩阵来计算隐含特征,即分句中如果只有观点词不包含特征词,希望结合该观点词的上下文判断出该观点词对应的隐含特征词。

随着主题模型在自然语言处理 NLP(Natural Language Processing)领域的广泛应用,文献[8]提出一种利用主题模型来提取特征词的方法,即将每条评论以词向量的模式输入 LDA(Latent Dirichlet Allocation)模

型,可以得到每条评论在若干主题上的概率分布,以及每个主题的词向量的概率分布,通过选择与主题最相关的词作为特征词。文献[9]进一步利用概率话题模型和深度学习模型来提取评论特征。此外,文献[10]引入深度学习模型来完成文本分类任务,利用深度信念网络自动提取文本特征;文献[11]设计了一个具有三种不同大小卷积核的神经网络结构,来完成局部抽象特征的自动提取。word2vec[16]是 2013 年谷歌开源的一款能够将词表征为实数值向量的 NLP 工具。其利用深度学习的思想,通过训练可以把对文本内容的处理简化为 K 维向量空间中的向量运算,而向量空间上的相似度可以用来表示文本语义上的相似度。本文将尝试利用该工具计算特征词之间的相似度,以进行特征词聚类。

综上所述,在提取特征词和观点词之后,需要对观点词的极性进行判断。通常来说,观点词的极性确定需要查阅情感词典,而情感词典的构建方法大致可分为人工收集法、基于词典的方法[12]和基于语料的方法[13]。本文不对情感词典的构建进行研究,直接利用已有的情感词典 HowNet[17],结合所提出的特征词极性计算方法完成特征词以及典型特征的极性计算。

2　OPEN 框架

图 2 展示了 OPEN 框架的主要步骤:(1) 对收集到的原始评论进行清洗,同时利用自然语言分析工具对清洗后的每条评论进行分词、词性标注以及依存关系分析,利用 word2vec 工具包基于京东生鲜商品的评论数据库学习词的隐含向量表示;(2) 提取特征词,利用词频提取高频特征词,利用依存关系和词性搭配规则提取低频特征词,利用观点词与特征词的关联度抽取隐含特征词;(3) 提取观点词,包括利用通用情感词典提取,以及从高频特征词出发寻找新的观点词,同时对每个分句提取所包含的 < 特征词, 观点词 > 词对;(4) 根据提取的每条评论包含的 < 特征词, 观点词 > 词对,通过在情感词典查找观点词的极性,从而确定每条评论特征词的极性,同时根据修饰观点词的程度副词及否定副词对特征词的极性进行修正;(5) 利用改进的半监督层次聚类算法对特征词进行聚类,得到典型特征及每类典型特征所包含的特征词,从而计算每类典型特征的情感极性。上述步骤并不是完全的顺序执行,其中步骤(2)和步骤(3)存在交互迭代,步骤(5)特征词聚类通常是在特征词提取完成后就执行。

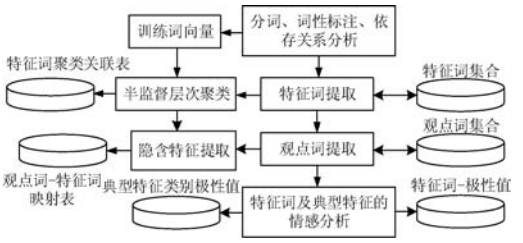

图 2　OPEN 框架的主要步骤

为了便于描述,本节首先对其中的部分术语给出定义。假设一款商品的所有评论样本构成集合记为 $D = \{d_1, d_2, \cdots, d_N\}$,$d_i$ 表示第 i 条评论,N 是评论总数,通过 OPEN 框架可以达到如下目标:

1) 提取特征词集合,记为 $F = \{f_1, f_2, \cdots, f_s\}$,$f_i$ 表示第 i 个特征词,s 是特征词总数,其中包含高频特征子集 $HF \subseteq F$,低频特征子集 $LF \subseteq F$。

2) 基于相似度对特征词进行聚类,典型特征类别记为 $C_1, C_2, \cdots, C_M$,M 为类别个数,其中每个典型特征类别包含的特征词为 $C_i = \{f_1, f_2, \cdots, f_n\}$。

3) 提取观点词集合,记为 $O = \{o_1, o_2, \cdots, o_l\}$,$o_i$ 表示第 i 个观点词,l 是观点词总数。

4) 针对每条评论 d_i,提取包含的{特征词, 观点词} 词对,记为 d_i:{ $< f_1, o_1 >$, $< f_2, o_2 >, \cdots$},所有评论样本的词对集合记为 FO。

5) 计算评论 d_i 中每个特征词的极性,记为 $p(f, d_i)$;对于样本集合 D,计算每个特征词 $f(\in F)$ 的情感极性,记为 $p(f, D)$;计算每个典型特征类别的极性,记为 $p(F, D)$。

2.1　分词、词性标注及依存关系分析

由于商品的用户评论存在口语化、随意化的特点,因此首先需要对原始评论集合进行清洗,包括用模糊匹配算法纠正其中的错别字,对于其中的标点符号和空格进行规范化处理等;然后 OPEN 框架使用支持中文的 NLP 工具[14]对清洗后的评论样本集进行分词、词性标注、依存关系分析,以及去除停用词处理,得到新的评论样本集合 $D = \{d_1, d_2, \cdots, d_N\}$,其中每条评论由词向量组成。需要注意的是准确的词性标注及依存关系分析对于后续特征词和观点词的提取非常关键,因此建议选择词性标注及依存关系分析尽可能准确的 NLP 工具。

2.2　基于 word2vec 训练词的表示向量

为了训练评论集中每个词的表示向量,在 OPEN 中通过利用 word2vec 工具包来实现,为了使得词向量更为准确地表示该词在电商评论中的语义,本文采用了京东生鲜类商品的 395 760 条评论作为训练语料。训练后可以得到每个词的表示向量,例如设置隐含空

间维度为 10 时,“价格”、“特价”、“味道”三个词的隐含表示向量分别为:

[0.256 -0.495 -0.993 0.794 0.492 -1.260 0.214 0.231 -0.030 0.750]

[0.284 -0.235 -1.014 0.423 -0.215 -1.127 0.268 0.011 0.423 0.041]

[0.533 -0.153 -0.576 1.012 0.041 -0.140 0.121 0.367 -0.054 0.409]

因此,采用余弦距离很容易得到基于词向量计算三个词的两两相似度:{价格, 特价}相似度为 0.815,{价格, 味道}相似度为 0.740,{特价, 味道}相似度为 0.571。

2.3 依存关系及词性搭配规则

在提取特征词、观点词以及修饰观点词的程度副词和否定副词时都需要用到依存关系和词性搭配规则。本节介绍 OPEN 中用到的主要依存关系和词性搭配规则。表 2 展示的是提取观点词和低频特征词所采用的主要依存关系及词性搭配规则。表 3 和表 4 分别列出了提取程度副词和否定词时的依存关系和词性搭配规则。

表 2 用于特征词和观点词提取的依存关系及词性搭配规则

依存关系	词性搭配规则	示例
主谓关系(SBV)	名词 + 形容词	价格(F)划算(O)
主谓关系(SBV)	名词 + (副词) + 动词	味道(F)可以(O)
主谓关系(SBV)	名词 + 副词 + 形容词	包装(F)很好(O)
动补关系(CMP)	动词 + 形容词	送货(F)及时(O)
动宾关系(VOB)	动词 + 名词	喜欢(O)味道(F)
动宾关系(VOB)	动词 + 动词	看(F)着可以(O)
状中关系(ADV)	形容词 + 动词	及时(O)送货(F)

表 3 用于提取程度副词的依存关系及词性搭配规则

依存关系	词性搭配规则	示例
状中关系(ADV)	副词 + 形容词	非常(ad.)好吃(O)
状中关系(ADV)	副词 + 动词	比较(ad.)满意(O)
定中关系(ATT)	副词 + 的 + 形容词	非常(ad.)的好吃(O)
动补关系(CMP)	动词 + 量词	少(O)很多(ad.)

表 4 用于提取否定副词的词性搭配规则

词性搭配规则	示例
否定词 + (副词) + 形容词	不(ad.)新鲜(O)
否定词 + (副词) + 动词	不(ad.)满意(O)
(副词) + 否定词 + 形容词	很(ad.)不(ad.)及时(O)
(副词) + 否定词 + 动词	很(ad.)不(ad.)满意(O)

2.4 特征词和观点词提取

在 OPEN 框架中,特征词抽取包括三个步骤:首先是基于词频选择高频名词(动名词)作为高频特征词;然后是基于依存关系和词性搭配规则,提取观点词和低频特征词;最后是根据已有的高频和低频特征词与观点词的修饰频率分析观点词对应的隐含特征词。将词频大于设定阈值的名词(动名词)作为高频特征词提取,得到高频特征词集合 HF。再从高频特征词出发,依据表 2 中的依存关系,提取修饰高频特征词的观点词 o,将其加入到观点词集合 O 中。

接着,利用 HowNet[17] 情感词典 S_dict,提取具有情感色彩的备选观点词,并从备选观点词出发,依据表 2 中的依存关系和词性搭配规则提取低频特征词和观点词,分别加入到 LF 和 O 中。具体描述如下:基于情感词典判断评论中的词语是否具有感情色彩,将具有感情色彩的词语加入到备选观点词集合 O^*;判断分句中是否包含备选观点词,如果包含备选观点词,继续判断该备选观点词在分句中存在的依存关系;根据依存关系和词性搭配规则,提取对应的特征词,将备选观点词加到观点词集合 O 中。需要说明的是,HowNet 情感词典中的词语只有情感极性,没有标注词性,因此在实际过滤情感词的时候会有导致部分误差。特征词和观点词提取过程如算法 1 所示。

算法 1 特征词和观点词提取算法

输入:$D = \{d_1, d_2, \cdots, d_N\}$, S_dict

输出:HF, LF, O^*, O, FO

```
for d_i in D = {d_1, d_2, ..., d_N}
    for w_j in d_i = {w_1, w_2, ..., w_m}
        if w_j 的词性为名词或动名词 then
            w_j 的计数器加 1;
    选择词频大于阈值的高频特征词加到 HF;
    根据表 2 提取修饰高频特征词的观点词加到 O;
    将提取的 < f, o > 词对加到 FO;
    for w_j in d_i = {w_1, w_2, ..., w_m}
        if w_j ∈ S_dict then
            if w_j 的词性不为副词 then
                将 w_j 加入到备选观点词集合 O*;
for d_i in D = {d_1, d_2, ..., d_N}
    以","将 d_i 分割为分句{cd_1, cd_2, ..., cd_z};
    for cd_j in {cd_1, cd_2, ..., cd_z}
        if cd_j 中包含 o* 且 o* ∈ O* then
            根据表 2 寻找低频特征词 f;
            if 找到对应的特征词 f then
                if 特征词不属于 HF then
                    将该特征词作为低频特征词加到 LF;
                将 o* 加到 O;
                将 < f, o* > 词对加到 FO;
```

由于用户在评论中的随意性,因此很多的备选观点词在进行依存关系匹配时,对应的特征词是缺失的,

例如“很好,新鲜干净”,“贵死了”等。这里观点词“好”、“新鲜”、“干净”、“贵”都无法提取对应的特征词。对于这种情况 OPEN 可以帮助那些能够明显反映特征信息的观点词提取隐含特征,即在当前商品的评论集中,如果该观点词总是修饰一个特征词,或者总是修饰一个典型特征类别的特征词,那么可推断该特征词或典型特征类别是观点词的隐含特征。基于这一思想,本论文提出一个依据观点词与特征词关联度的隐含特征映射方法。基本思想如下:首先针对观点词 o_j,分析其在 < 特征词, 观点词 > 修饰频率矩阵 $\boldsymbol{M}_{s\times l}$ 中与对应的特征词的修饰次数,这里 $\boldsymbol{M}_{s\times l}$ 矩阵的元素 m_{ij} 表示特征词 f_i 与观点词 o_j 在所有评论中存在修饰关系的次数;假设与 o_j 存在修饰关系的特征词个数为 τ,将这些特征词根据修饰次数降序排列,$f_1, f_2, \cdots f_M, \cdots, f_\tau$,如果 M 为使式(1) 成立的最小特征词个数,且所有 M 个特征词均属于同一个典型特征类别,则推断 o_j 修饰的隐含特征词为 f_1,否则不能够给 o_j 推断隐含特征。在式(1) 中,*IFthresh* 为隐含特征阈值, 范围为[0,1], *IFthresh* 值越大,可推断隐含特征词的要求越严格,通常取阈值为 0.5。

$$\text{针对观点词 } o_j: \frac{(\sum_{i=1}^{M} m_{ij})}{(\sum_{i=1}^{s} m_{ij})} \geqslant IFthresh \qquad (1)$$

举一个简单的例子。假设观点词“便宜”在该商品的所有评论中,只修饰了特征词“价格”,因此可推断“便宜”修饰的隐含特征词为“价格”;假设观点词“不错”在评论集中修饰过若干特征词,在满足式(1) 的情况下,修饰次数最多的特征词依次为:味道、质量、肉质、服务、包装,而这些特征词不属于同一个典型特征类别,因此在商品中,如果只出现观点词“不错”不能为其推断隐含特征词。

2.5　典型特征聚类

由于词汇本身的多样性导致基于依存关系和词性搭配规则方法提取出的特征词的个数是比较多的,使得用户浏览非常繁琐。在本文的实验中一个 2 596 条评论的数据集,就能提取出 297 个特征词,而其中大多数特征词都可从语义上聚类为几个典型的特征。然而遗憾的是,目前的无监督聚类算法或者基于主题模型的聚类算法的效果都不太理想,究其原因主要是难于准确地衡量两个特征词的距离(相似度)。为此,本文引入 word2vec 来训练每个特征词的隐含表示向量,从而计算特征词之间的相似度;并提出一个改进的半监督层次聚类算法对特征词进行聚类。聚类过程如算法 2 所示,其中 $\{f_1, f_2, \cdots, f_s\}$ 表示 s 个特征词的 k 维隐含表达向量,$f_i = [r_{i1}, r_{i2}, \cdots, r_{ik}]$ $(r_{ij} \in R)$。首先利用余弦距离计算任意两个特征词之间的相似度,如式(2) 所示;然后以每个特征词作为一个初始类别开始聚类,每次只将相似度最大的两个类别进行合并,合并时需要满足输入的聚类约束条件;重复执行合并过程直到满足聚类终止条件为止。

$$sim(f_x, f_y) = \frac{\sum_{i=1}^{k}(r_{xi} \cdot r_{yi})}{\sqrt{\sum_{i=1}^{k} r_{xi}^2} \cdot \sqrt{\sum_{i=1}^{k} r_{yi}^2}} \qquad (2)$$

为了提升聚类效果,论文引入了约束条件限制的半监督聚类算法,通过先验知识设计少量特征词对之间的 must-link 约束和 cannot-link 约束来辅助聚类。其中存在 must-link 约束的两个特征词必须在同一个类别中,而存在 cannot-link 约束的两个特征词不能聚类在同一个类别中。由于本论文的实验采用京东肉类商品的评论数据集,因此初始约束条件设计如下:

must-link 约束:{味道, 口感},{配送, 物流}

cannot-link 约束:{味道, 配送},{价格, 味道}

由于上述约束存在异类传递特性,即:

$(f_i, f_j) \in must\text{-}link \& (f_i, f_k) \in cannot\text{-}link \Rightarrow (f_j, f_k) \in cannot\text{-}link$

因此属于 cannot-link 约束的词对还包括:{口感, 配送},{口感, 物流},{味道, 物流},{价格, 口感}。在实际类别合并时,如果 $(f_i, f_j) \in cannot\text{-}link, f_i \in C_1, f_j \in C_2$,则 C_1 和 C_2 不能合并。

本文采用两个类别中任意两个特征词的平均相似度作为两个类别的相似度,如式(3) 所示。聚类终止条件可以使用任意两个类之间的平均相似度低于设定的阈值。

$$sim_{\text{avg}}(C_i, C_j) = \frac{\sum_{f_x \in C_i} \sum_{f_z \in C_j} \text{sim}(f_x, f_z)}{|C_i| \cdot |C_j|} \qquad (3)$$

算法 2　基于词向量的半监督层次聚类算法

输入:$F = \{f_1, f_2, \cdots, f_s\}$,约束条件, 终止条件

输出:$C_1, C_2, \cdots, C_M$

1. 将每个特征词初始化为一个类别,$C_i = \{f_i\}$;
2. 根据 must-link 约束条件,将特征词进行聚类;
3. 利用式(3) 计算任意两个类别的相似度 $sim(C_i, C_j)$;
4. 在满足 cannot-link 约束条件的情况下,将相似度最大的两个类别进行合并;
5. 重复执行步骤 3 和步骤 4 直到终止条件满足。

2.6　特征词的情感分析

在 < 特征词, 观点词 > 提取完成后,OPEN 设计了一个简单的方法来计算每条评论中每个特征词对应的情感极性。首先在情感词典中查找该观点词的极性(正/负),然后根据具体的词性搭配规则计算修饰的

特征词的极性值,本文主要采用以下三种搭配规则:

<特征词,[程度词1]…[程度词 n]观点词>:

$$p(f,d_i) = \prod_{j=1}^{n} \deg(adv_j) \times p(o)$$

<特征词,[否定词]观点词>:

$$p(f,d_i) = (-1) \times p(o)$$

<特征词,[程度词][否定词]观点词>:

$$p(f,d_i) = deg(adv) \times (-1) \times p(o)$$

这里 $deg(adv)$ 表示程度副词对应程度的权重,在本论文中使用的程度词典将程度分为五个等级,每个等级有自己的权重值,如果修饰观点词的程度副词有多个,则将多个程度词的程度权重相乘。在 OPEN 中,可以针对每个特征词 f ,计算其在 $\boldsymbol{D}$ 中的极性值,如式(4)所示,这里 $|D(f)|$ 表示样本集中包含特征 f 的评论数;也可以计算每个典型特征类别的极性值,如式(5)所示,这里 $\boldsymbol{F} = \{f_1, f_2, \cdots, f_n\}$ 。

$$p(f,\boldsymbol{D}) = \frac{1}{|D(f)|} \sum_{i=1}^{N} p(f,d_i) \tag{4}$$

$$p(\boldsymbol{F},\boldsymbol{D}) = \frac{1}{\sum_{i=1}^{n} |f_i|} \left(\sum_{i=1}^{n} \sum_{j=1}^{N} p(f_i,d_j) \right) \tag{5}$$

3 实验及结果分析

本文将 OPEN 在京东的生鲜类商品的评论中进行应用,得到了较好的效果。为了对 OPEN 的性能进行量化评价,本文在一个小样本猪肉商品的评论数据集上设计了 3 组实验:实验 1 验证 OPEN 提取特征词和观点词的性能;实验 2 验证准确提取 < 特征词, 观点词 > 词对的性能;实验 3 验证对每条评论中包含的特征词的情感极性的分析性能。评价指标分别采用准确率和召回率。对比方案包括 Hu 等[1]提出的经典方案,Luo 等[6]提出的依赖规则和知识库的特征词提取方案,以及 Zhang 等[7]提出的提取隐含特征词的方案。

3.1 特征词和观点词抽取算法的性能

实验 1 用于验证本论文提出的特征词和观点词抽取算法的性能,实验结果如表 5 所示。结果显示 Hu 的方案提取的特征词和观点词数量明显少于其他三种方案。这是因为 Hu 的方案中观点词只考虑了形容词,特征词只考虑了名词,这是一个比较严格的规则,因此准确率较高,召回率较低,说明提取特征词和观点词都有一定的遗漏。在中文评论中观点词还可以是动词或者名词,因此在 Luo 的方案中放宽了观点词的词性,不仅包括形容词,还可以包括有情感倾向的动词和名词。Luo 的方案提取的特征词和观点词数量明显高于 Hu 的方案,但是增加的特征词和观点词不一定都是正确的,因此准确率有一定的下降,但是召回率大幅上升。OPEN 提取的观点词数量高于其他方案,这是因为 OPEN 中把所有包含的具有情感色彩的词都作为了观点词,但是正确的观点词数量却有少量下降。

表 5 特征词和观点词提取的性能对比

方案	Hu	Luo	Zhang	OPEN
特征词数量	179	297	297	297
正确特征词数量	103	140	140	144
观点词数量	124	332	332	360
正确观点词数量	100	207	207	196
准确率(特征词)	0.575	0.471	0.471	0.485
召回率(特征词)	0.696	0.946	0.946	0.973
准确率(观点词)	0.806	0.623	0.623	0.544
召回率(观点词)	0.454	0.941	0.941	0.891

3.2 词对抽取的性能

实验 2 用于验证 OPEN 提取 < 特征词, 观点词 > 词对的能力。上述四种方案对于每条评论都可以提取出每个分句中包含的 < 特征词, 观点词 > 词对,结果如表 6 和表 7 所示。

表 6 < 特征词, 观点词 > 词对提取算法的性能对比(不含隐含特征词)

方案	Hu	Luo	Zhang	OPEN
词对数	2 624	4 864	4 864	6 491
正确词对数	1 724	2 350	2 350	3 477
准确率	0.657	0.483	0.483	0.536
召回率	0.404	0.551	0.551	0.815

表 7 对隐含特征词的词对提取算法的性能对比

方案	手工	Zhang	OPEN
明确含义的隐含特征词对数	804	402	469
正确映射的词对数	—	290	436
准确率	—	0.721	0.930
召回率	—	0.360	0.542

表 6 中结果显示 OPEN 正确提取的词对数明显高于其他三种方案,召回率有较大幅度的提升,达到 81.5%。OPEN 提取大量无关词对的原因主要是有些评论是一些无关产品特征的描述,但是符合本文的依存关系和词性搭配原因,因此仍然被提出,一个简单的改进方法是将符合规则但低频的词对删去。

表 7 展示了 Zhang 的方案和 OPEN 对隐含特征词提取的性能对比。在小样本集的评论中有 804 个观点词是没有(显式)特征词,但是具有明确的隐含含义。

在人工标注的 804 个隐含特征词对中,Zhang 的方案正确提取出了其中的 402 个观点词,进一步隐含特征词映射正确的个数为 290 个,准确率为 72.1%。OPEN 正确映射的隐含特征词的准确率为 93.0%,明显优于 Zhang 的方案。这说明 OPEN 的隐含特征映射方法确实更有效。

3.3　词对极性判断的能力

实验 3 用于验证 OPEN 对词对极性的判断能力,准确地说是对观点词的情感极性判断,以及程度副词和否定副词的提取能力的评价。实验结果表明本文提出的程度词及否定词提取方法能够获得较高的准确率,达到 98.3%。同时本文提出的极性值计算方法不仅能够反映出观点的正负极性,而且能够较为准确地反映出观点的极性强度。这个方法的问题是,性能取决于情感词典和程度词典的完整性和准确性,实验中暂时把没有找到情感极性的观点词都算作正向情感,更完善的方法是针对评论商品的特点建立有针对性的情感词典和程度词典,将在后续工作中完成。

4　结　语

本文针对电商平台的评论挖掘展开研究,提出了一个基于评论的商品特征抽取及情感分析框架 OPEN,并将该框架在京东的生鲜类商品的评论中进行应用。实验结果表明该框架确实能够成功抽取出每个生鲜商品的典型特征及该特征对应的情感极性,且在小样本数据集上测试了特征抽取算法的性能,特征词的准确率和召回率分别达到48.5%和97.3%,观点词的准确率和召回率分别达到 54.4%和 89.1%,<特征词, 观点词>词对的准确率和召回率分别达到54.4%和81.5%,词对极性判断的准确率为 98.3%。另一方面本文的情感极性值计算使用的是 HowNet 的通用情感词典,在不同的商品场景下,某些词可能表现出完全不同的情感,例如“师傅辛苦”,观点词“辛苦”在情感词典中是负向极性,但是在电商评论中,这明显是正向极性的词。因此后续工作的一个重点是在通用情感词典的基础上,建立面向应用场景的情感词典和程度词典。

参　考　文　献

[1] Hu Mingqing, Liu Bing. Mining opinion features in customer reviews[C]//Proceedings of the 19th national conference on Artificial intelligence. AAAI Press, 2004: 755 - 760.

[2] Hu Minqing, Liu Bing. Mining and summarizing customer reviews[C]//Proceedings of the 10th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2004: 168 - 177.

[3] Popescu A M, Nguyen B, Etzioni O. OPINE: extracting product features and opinions from reviews[C]//Proceedings of HLT/EMNLP on Interactive Demonstrations. ACM, 2005: 32 - 33.

[4] Qiu Guang, Liu Bing, Bu Jiajun, et al. Opinion Word Expansion and Target Extraction through Double Propagation[J]. Computational Linguistics, 2011, 37(1): 9 - 27.

[5] Zhai Zhongwu, Liu Bing, Zhang Lei, et al. Identifying evaluative sentences in online discussions[C]//Proceedings of the 26th national conference on Artificial intelligence. AAAI Press, 2011.

[6] 罗熹. 基于评论信息的内容感知方法研究[D]. 成都:电子科技大学, 2015.

[7] Zhang Yu, Zhu Weixiang. Extracting implicit features in online customer reviews for opinion mining[C]//Proceedings of International Conference on World Wide Web Companion. 2013: 424 - 32.

[8] Liu K, Xu L, Zhao J. Co-Extracting Opinion Targets and Opinion Words from Online Reviews Based on the Word Alignment Model[J]. Knowledge & Data Engineering IEEE Transactions on, 2015, 27(3): 636 - 650.

[9] Lao X, Ma B, Zhang N, et al. Public Opinion Analysis Based on Probabilistic Topic Modeling and Deep Learning (in Chinese)[C]//CNAIS National Congress, 2015.

[10] 张庆庆,刘西林. 基于深度信念网络的文本情感分类研究[J]. 西北工业大学学报(社会科学版), 2016, 36(1): 62 - 66.

[11] 蔡慧苹,王丽丹,段书凯. 基于 word embedding 和 CNN 的情感分类模型[J]. 计算机应用研究, 2016, 33(10): 2902 - 2905.

[12] Khalifa K, Omar N. A hybrid method using lexicon-based approach and Naive Bayes classifier for Arabic opinion question answering[J]. Journal of Computer Science, 2014, 10(10): 1961 - 1968.

[13] 陈铁明,缪茹一,王小号. 融合显性和隐性特征的中文微博情感分析[J]. 中文信息学报, 2016, 30(4): 184 - 192.

[14] 哈尔滨工业大学语言云平台[OL]. http://www.ltp-cloud.com/.

[15] Ebbinghaus H. Memory: A Contribution to Experimental Psychology[J]. Annals of Neurosciences, 2013, 20(4): 155 - 156.

[16] Mikolov T, Chen K, Corrado G, et al. Efficient Estimation of Word Representations in Vector Space[C]//Proceedings of Workshop at ICLR, 2013.

[17] HowNet[OL]. http://www.keenage.com/html/c_index.html.

[18] WordNet [OL]. http://wordnet.princeton.edu/wordnet/download/.