农产品评价观点抽取和情感识别系统设计实现

陈　杰,周梓豪*,吴军辉*
(同济大学 电子与信息工程学院,上海 201804)

摘　要:电商平台上的评价数据蕴藏着消费者的情感观点,识别评价情感表达的关键是挖掘其在产品属性方面级别的观点,并判别情感倾向。先前的有监督学习模型需要相关领域的大量人工标注数据进行训练,耗费较多的人力成本,因此,构建了无监督学习框架的农产品评价观点抽取和情感识别系统。通过爬虫获取多源电商平台的评价数据,首先通过 LDA 模型确定领域主题属性,结合 SO-PMI 算法构建领域情感词典,然后通过 LTP 库的依存句法分析和词嵌入相似度制定方面观点的抽取规则,并提出情感强度值计算方法识别评价的方面情感倾向。实验证明,该框架的查准率为 85.08%,召回率为 78.50%,F1 值为 81.66%,性能优于传统模型。根据观点抽取和情感识别结果构建可视化平台,从多个角度挖掘消费者对农产品的偏好。该系统已实际部署在农资农产品在线服务交易平台的项目中,致力于服务消费者、经销商、电商平台和监管部门四个主体,取得了良好的应用效果。

关键词:观点挖掘;自然语言处理;无监督学习;领域词典;依存句法规则;农产品评价
中图分类号:TP391.1　　**文献标识码**:A　　**文章编号**:1673-629X(2023)08-0116-08
doi:10.3969/j.issn.1673-629X.2023.08.017

Design and Implementation of Agricultural Product Review Opinion Mining and Sentiment Recognition System

CHEN Jie,ZHOU Zi-hao*,WU Jun-hui*
(School of Electronic & Information Engineering,Tongji University,Shanghai 201804,China)

Abstract:The review data on the e-commerce platform contains sentiment opinions of consumers,and the key to identifying sentiment expression of reviews is to extract their opinions at the product attribute level and identify sentiment tendencies. Previous supervised learning models require a large amount of manually labeled data in related fields for training,which consumes a lot of labor costs. Therefore,an unsupervised learning framework is constructed for agricultural product review for opinion mining and sentiment recognition. The review data of the multi-source e-commerce platform is obtained through the crawler. Firstly,the domain theme attribute is determined by the LDA model,and the domain sentiment dictionary is constructed in combination with the SO-PMI algorithm,then the extraction rules of aspect-level opinions are formulated through the dependency syntactic analysis of the LTP library and the similarity of word embedding,and the sentiment intensity value calculation method is proposed to identify the aspect-level sentiment tendency of the review. Experiments show that the accuracy of the proposed framework is 85.08%,the recall rate is 78.50%,and the F1 value is 81.66%,which is better than that of traditional models. According to the opinion mining and sentiment recognition results,a visualization platform is built to explore consumers' preferences for agricultural products from multiple angles. The system has been actually deployed in the project of the online service trading platform for agricultural materials and agricultural products,and is committed to serving consumers,distributors,e-commerce platforms and regulatory departments,which achieved good application results.
Key words:opinion mining;natural language processing;unsupervised learning;domain dictionary;dependency parsing rule;agricultural product review

0　引　言

随着电子商务的普及,更多的消费者通过在线平台购买农产品,例如京东生鲜、天猫超市、每日优鲜等,并持续地产生海量的评价信息。观点挖掘作为 NLP 领域的分支,是典型的文本数据挖掘技术,能够发挥大数据的“4V”特征优势[1]。将该方法运用于农产品评价分析中,有助于消费者了解产品画像,经销商和电商平台掌握消费者的需求,监管部门及时发现产品的

收稿日期:2022-10-21　　**修回日期**:2023-02-23
基金项目:国家重点研发计划(2020YFD1100603)
作者简介:陈　杰(1968-),男,副教授,研究方向为大数据应用、自然语言处理;通讯作者:周梓豪(1998-),男,硕士研究生,研究方向为机器学习、自然语言处理;通讯作者:吴军辉(1974-),男,副教授,研究方向为自然语言处理、数据挖掘。

问题。

观点挖掘被定义为判断作者对特定实体发表意见的情感倾向的任务[2]。按照面向对象的不同可以划分为粗粒度和细粒度目标[3]。先前的研究主要集中在对评价整体进行粗粒度的分析[4]。细粒度任务是一个新兴的方向,目前已被应用在旅游景点、在线论坛等领域[5-6],其主要分为方面观点目标提取和目标情感识别两个子任务[7]。在模型的选用上可以分为基于知识的方法和学习的方法[8]。在基于学习的方法中,先前研究通常使用支持向量机、隐马尔可夫模型等统计学习方法[9]。随后,基于目标依赖和关联的长短期记忆神经网络和注意力机制运用于该任务中[10]。但基于学习的方法需要大量人工标注的领域数据集进行训练,且对于结构复杂句子的拟合效果不佳。在基于知识的方法中,Qin 等[5]根据外部知识库提炼出游客关心的旅游景点属性,构建三级评价体系的决策系统进行观点挖掘。万岩等[11]在微博评论分析中构建了融合情感词典和语义规则的模型。上述方法能够较好地提取并识别显式观点,但忽略了评价语法表达的多样性而导致隐式观点的遗漏,在短评价观点挖掘任务中最为常见[12]。目前已有通过有监督、半监督和无监督学习的方法识别隐式观点的研究。Li 等[13]从领域内语料库中检索出带情感标注的数据集进行有监督的对比预训练,通过将隐含情感特征表示和具有相同标签的情感表达对齐,来捕捉评价中的隐式观点特征。Xu 等[12]构建融合 SVM 和主题模型的半监督方法提取隐含观点。相比之下,无监督方法具有无需人工标注训练集的优势,在该方法中往往使用包括层次结构、本体识别、主题建模、共现、依存关系分析、关联规则挖掘和聚类等各种 NLP 模型[14]。Hu 等[15]根据近邻关系将评价中的属性词与情感词对应,但忽略了语法元素的依赖而造成抽取观点的效果较差。在此基础上,Sun 等[16]提出属性词和情感词之间的上下文语境关系对识别隐式观点具有一定帮助。也有研究将词对依存规则与词法分析结合,以更精确地抽取方面观点[17]。

目前的观点挖掘技术已取得较多应用成果,但少有在农产品领域开展应用的文献和开源数据集。为解决上述问题,该文结合爬虫技术获取多源电商农产品评价数据,构建了一种基于无监督学习框架的评价观点抽取和情感识别系统。在算法层面,对传统单一模型进行改进,组合了 LDA 主题模型、SO-PMI 算法、依存句法规则和词嵌入相似度计算的 NLP 技术,以更好地识别隐式的方面观点并进行情感识别。在应用层面,根据分析结果构建可视化平台,从品类、品种、品牌和店铺的不同角度挖掘消费者偏好,通过词频统计比较产品的优势和缺陷,以期服务于消费者、经销商、电商平台和监管部门四个主体。

1　观点抽取和情感识别方法实现

该文构建无监督学习框架的观点抽取和情感识别方法,如图 1 所示,包括了数据采集和预处理、主题识别和词典构建、方面观点抽取和情感识别算法以及可视化平台搭建应用四个模块。以下将详细介绍上述流程运用的技术。

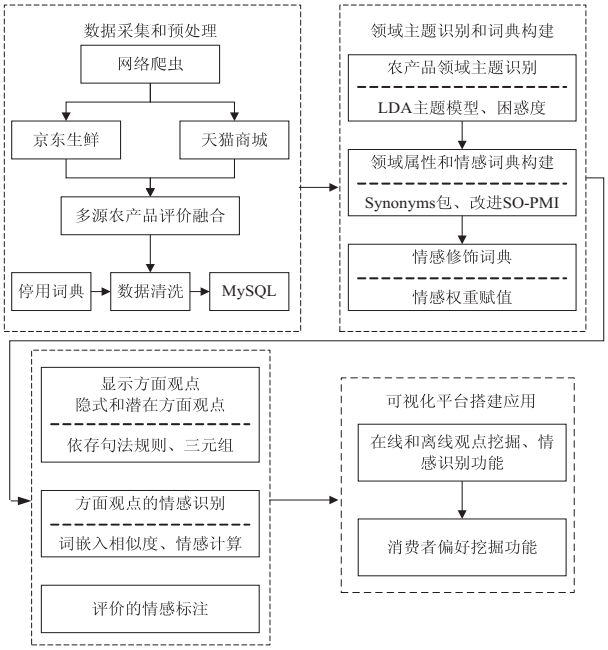

图 1　基于无监督学习的分析技术框架

1.1　数据采集与预处理

该文采用 Python 语言的 Selenium 框架爬虫技术,分别从京东生鲜和天猫商城网站获取多源农产品评价数据。根据产品销量排序分别获得包括花生、玉米、苹果、梨、猕猴桃、橘子、芒果、竹笋、茶叶、茶油的 10 个农产品品类商品的评价数据。在进行数据融合的同时进行数据清洗,首先去除重复评价和空评价,然后运用正则化表达式去除评价中的数字和表情符号,导入停用词典去除评价中的无效信息,并去除字符数量少于 3 的过短评价和大于 200 的过长评价,最终得到 1 147 861 条评价作为该文的数据集,并附带有评价对应的商品链接、店铺、品牌和品种信息,采集结果写入 MySQL 中。

1.2　领域主题识别和词典构建方法

1.2.1　基于 LDA 模型的领域主题识别

为确定消费者所关注的电商农产品主题属性,该文根据获取的评价数据使用 LDA 模型进行主题挖掘,通过调用 Gensim 库下的 LdaModel 方法完成上述任务。在构建 LDA 模型的过程中,采用基于困惑度的方式确定最优主题数 K 的取值,困惑度取值越小说明生成模型效果越好[18]。其计算方法如公式(1)所示:

$$\text{perplexity}(D) = \exp\left(-\frac{\sum \log p(z \mid d) * p(w \mid z)}{\sum_{d=1}^{M} N_d}\right) \tag{1}$$

式中，$p(z \mid d)$ 表示文档 d 中每个主题出现的概率，$p(w \mid z)$ 表示每一个词汇 z 在某个主题下出现的概率，$\sum_{d=1}^{M} N_d$ 表示语料中所有词汇的总长度。设定基于困惑度计算的参数 K 寻优区间为 [2,20]，最终得到当 K = 8 时模型困惑度达到最小，即划分出 8 个领域主题，并分别命名为口感、品质、色泽、价格、分量、包装、物流服务、售后服务。其中口感、品质和色泽主题代表了农产品自身画像的特性，价格和分量主题代表了农产品的性价比特性，包装、物流服务和售后服务主题代表了农产品的配套服务特性。该文根据划分出的主题构建领域词典，以更好地进行方面意见抽取。

1.2.2 领域属性词典和情感词典构建

筛选出 LDA 模型输出结果中每个主题的名词、动词关键词作为核心属性词，在此基础上，引入中文工具包 Synonyms 对每个核心属性词进行近义词扩充，对部分没有主题意义的噪声词进行过滤，最终形成包含 275 个词汇的领域属性词典，示例如表 1 所示。

表 1 领域属性词典构建结果

意义	主题属性	属性词举例
农产品画像	口感	口感、味道、口味、气味、甜度…
	品质	品质、质量、品控、食材、用料…
	色泽	颜色、色泽、成色、品相、外观…
性价比	价格	价格、价钱、价位、单价、定价…
	分量	分量、份量、量、个头、大小…
配套服务	包装	包装、外包装、罐子、袋子、箱子…
	物流	物流、发货、送货、配送、小哥…
	售后	售后、服务、客服、退款、退换…

另一方面，需要构建的情感词典包括基础情感词典、领域情感词典和情感修饰词典。首先选用 HowNet 发布的“情感分析用词语集”中的中文正负面评价词典作为基础情感词典 ，包含 3 730 个正面情感词和 3 116个负面情感词。为了识别评价中未与属性词搭配的情感词，即隐式方面观点，构建有效的领域情感词典是必要的。根据领域属性词典，运用分句和字符串匹配将农产品评价按照 8 个主题属性切分形成子句集，使用 TF-IDF 算法识别各属性下的词频排序前 100 的形容词和名词，公式如式(2)所示：

$$\text{TF-IDF}_{w,D_i} = \text{TF}_{w,D_i} \times \text{IDF}_w = \frac{\text{count}(w)}{|D_i|} \times \log \frac{N}{\sum_{i=1}^{N} I(w, D_i)} \tag{2}$$

式中，TF_{w,D_i} 表示词汇 w 在文档 D_i 中出现的频率，IDF_w 表示其逆文档频率，$\text{count}(w)$ 统计词汇 w 出现的频次，$|D_i|$ 为 D_i 中所有词的频数，N 为文档总数，$I(w, D_i)$ 表示 D_i 中是否包含词汇的 0/1 变量。剔除在多种属性下共同出现的高频词汇，得到相应属性的候选情感词集。然后根据基础情感词典判断词汇的情感倾向，筛选出对应属性的种子情感词。针对部分未被基础情感词典收录的词汇，结合种子情感词运用 SO-PMI 算法确定情感极性[19]。基础 SO-PMI 公式如式(3)(4)所示：

$$\text{PMI}(\text{word1}, \text{word2}) = \log\left(\frac{P(\text{word1\&word2})}{P(\text{word1})P(\text{word2})}\right) \tag{3}$$

$$\text{SO\_PMI}(\text{word1}) = \sum_{p \in \text{pwords}} \text{PMI}(\text{word1}, p) - \sum_{n \in \text{nwords}} \text{PMI}(\text{word1}, n) \tag{4}$$

式(3)中，$P(\text{word1})$ 和 $P(\text{word2})$ 分别表示候选情感词汇和种子情感词出现的概率，$P(\text{word1\&word2})$ 表示它们共现的概率，式(4)中 pwords 和 nwords 分别表示正面和负面的种子情感词集。先前普遍认为 SO-PMI 算法得到的 SO_PMI 值大于 0 标注词汇为正面情感，小于 0 标注为负面情感，但实际情况中存在大量情感值接近于 0 的中性情感词，因此对 SO_PMI 值进行 Min-Max 标准化线性转换，公式如式(5)所示：

$$\text{senti}_i = \frac{\text{SO\_PMI}_i - \min\limits_{1 \leqslant j \leqslant n}\{\text{SO\_PMI}_j\}}{\max\limits_{1 \leqslant j \leqslant n}\{\text{SO\_PMI}_j\} - \min\limits_{1 \leqslant j \leqslant n}\{\text{SO\_PMI}_j\}} \tag{5}$$

式中，SO_PMI_i 表示第 i 个词汇的初始情感值，$\min\limits_{1 \leqslant j \leqslant n}\{\text{SO\_PMI}_j\}$ 和 $\max\limits_{1 \leqslant j \leqslant n}\{\text{SO\_PMI}_j\}$ 分别表示所有词汇中的最小和最大情感值，senti_i 表示经过标准化处理得到的情感值。然后需要确定阈值 T，剔除情感值在 $[-T, T]$ 区间内的中性词汇，将大于 T 的词汇归类为正面，小于 T 的词汇归类为负面，以此构建领域情感词典。经过反复实验，设置 SO-PMI 算法的共现窗口为 5，阈值 T = 0.2，最终形成包含 194 个正面词汇和 190 个负面词汇的领域情感词典，示例如表 2 所示。

表 2 领域情感词典构建结果

领域属性	正面词汇举例	负面词汇举例
口感	好吃、美味、新鲜、甜…	淡、苦、酸、硬…
品质	正品、优质、营养、有机、维生素…	次品、发霉、假冒、虫子、变质、转基因…
色泽	鲜艳、好看、光泽、嫩绿…	丑、皱、发黄、发黑…
价格	划算、实惠、便宜、值得…	提价、涨价、不值、贵…
分量	多、足、饱满、厚实…	少、小、细、干瘪…

续表 2

领域属性	正面词汇举例	负面词汇举例
包装	完整、精美、真空、密封…	漏气、破损、简陋、破…
物流	快、快捷、迅速、当日达…	慢、拖拉、丢件、损毁…
售后	耐心、真诚、周到、诚信…	消极、欺骗、厌烦、敷衍…

情感修饰词包含程度副词和否定词,其作用是影响评价情感语义强度[20]。例如程度副词“非常”能够加强情感词的强度,“略微”对情感词强度起减弱作用,而否定词“没有”逆转了情感词的语义。该文以 HowNet 词汇为基础,选取了 220 个程度副词和 58 个否定词构成情感修饰词典,根据各个情感修饰词对情感表达的加强、减弱或逆转的作用,将词典划分为 7 个等级并赋予从-1 到 2 不等的情感权重。例如“非常”“极其”“极度”等 69 个词的情感权重设定为 2,“很”“太”“特别”等 42 个词的情感权重设定为 1.5,“稍微”“相当”“有点”等 29 个词的情感权重设定为 0.6,“不”“没有”“并非”等 58 个词的情感权重设定为-1。

1.3　方面观点抽取和情感识别

1.3.1　基于依存句法规则的方面观点抽取

方面观点是评价中针对领域属性的情感特征表示,一般由属性词、情感词和情感修饰词组成。先前研究总结方面观点表达通常存在于主谓关系、动宾关系、定中关系、并列关系、动补结构、状中结构的依存句法关系中[21],该文额外地补充了省略属性词的隐式观点表达和同义转述的潜在观点表达两种情况,示例如表 3 所示。

表 3　方面观点表达形式举例

标签	关系类型	评价举例	方面观点
SBV	主谓关系	价格很便宜	<价格,很,便宜>
VOB	动宾关系	没有包装	<包装,没有>
ATT	定中关系	相对实惠的价格	<价格,相对,实惠>
COO	并列关系	口感和品质都棒	<口感,棒><品质,棒>
CMP	动补结构	商家服务态度好	<服务态度,好>
ADV	状中结构	口感很不错	<口感,很,不错>
IMP	隐式表达	买贵了	<null,贵>==<价格,贵>
POT	潜在表达	煮着汤色好看	<汤色,好看>==<色泽,好看>

该文使用哈工大语言技术平台(LTP)提供的功能,首先将评价按照标点符号划分为子句后进行分词和词性标注,通过识别句子的核心词成分和句法结构来分析词汇之间的语法依赖关系。然后融合领域属性词典和情感词典制定了 7 条依存句法规则,分别对显式方面观点、隐式方面观点、潜在方面观点和情感修饰词进行抽取,尽可能地挖掘评价中的情感信息。

规则 1:当评价同时存在属性词和情感词时,若满足 SBV、ATT 或 CMP,提取<属性词,情感词>作为显式方面观点。若有多个满足条件则按照最近邻关系搭配。

规则 2:当评价中仅存在领域情感词典中的词汇,而没有关联的属性词时,将所有的领域情感词分别抽取出来,形成<null,情感词>作为隐式方面观点。

规则 3:识别未被属性词典概括的未登录词。当核心词词性在[‘n’,‘v’,‘nz’]中,提取作为属性词。然后抽取满足 VOB 或 CMP 的形容词作为情感词,形成<属性词,情感词>表达潜在方面观点,若匹配失败则舍弃属性词。

规则 4:识别未被情感词典概括的未登录词。当核心词词性在[‘a’,‘u’,‘d’]中,提取作为情感词。然后抽取满足 SBV 或 ATT 且词性在[‘v’,‘n’,‘j’,‘nz’]的词汇作为属性词。若匹配失败也将情感词保留,形成<属性词,情感词>或<null,情感词>表达潜在方面观点。

规则 5:识别核心词周围的潜在情感表达。首先提取和核心词满足 VOB 且词性在[‘v’,‘n’,‘j’,‘nz’]的词汇作为属性词,匹配和该词关联的形容词作为情感词,若匹配失败则舍弃。然后提取和核心词满足 VOB 或 CMP 的形容词作为情感词,匹配和该词关联的词性在[‘v’,‘n’,‘j’,‘nz’]的词汇作为属性词,若匹配失败也将其保留,形成<属性词,情感词>或<null,情感词>表达潜在方面观点。

规则 6:在规则 1 到规则 5 的基础上,寻找与方面观点满足 COO 的元素。抽取出的元素与属性词共享情感词和情感修饰词,或与情感词共享属性词,形成新的方面观点。

规则 7:在规则 1 到规则 6 的基础上,提取和情感词满足 CMP 或 ADV 的副词作为情感修饰词,最终的方面观点列表以弹性三元组格式存储。在该格式中,属性词可能不存在而情感词必须存在,情感修饰词可能不存在或存在多个。其表达方式有<属性词,情感修饰词,情感词><属性词,null,情感词><null,情感修饰词,情感词>和<null,null,情感词>四种形式。

1.3.2　基于词嵌入和情感值计算的情感识别

(1)词嵌入模型。

抽取评价的方面观点列表后,需要识别每个方面观点表达的主题属性和情感倾向。对于显式和隐式的方面观点,可以通过词典匹配得出识别结果。对于潜在方面观点,需要判断属性词、情感词与领域词汇是否为同义词,区分其是有效的情感表达还是噪声。首先需要把文本数据转化为词向量表示,word2vec 是一种经典的词嵌入方法,与 one-hot 编码相比解决了维度灾难问题[22],该文使用 Gensim 库建立 word2vec 模型。将农产品评价数据集与开源的 wiki 百科语料库

合并对模型进行预训练，设置参数 size = 250，min_count = 5。然后导入模型将词汇转化成词向量的形式，计算潜在方面观点中属性词、情感词与领域词典词汇的余弦相似度，计算方法如公式(6)所示。

$$\cos(w_1,w_2)=\frac{\sum_{i=1}^{size} w_1\times w_2}{\sqrt{\sum_{i=1}^{size} w_1^2}\times\sqrt{\sum_{i=1}^{size} w_2^2}} \tag{6}$$

式中，w_1 和 w_2 表示两个词汇的词向量，size 表示维数，$\cos(w_1,w_2)$ 表示两词的余弦相似度。经过反复实验，设定相似度阈值为 0.85，匹配相似度最大且大于 0.85 的领域属性词汇对应的属性作为潜在方面观点属性，若匹配失败，进一步选取相似度最大且大于 0.85 的领域情感词汇对应的属性作为其属性，若匹配失败则将该潜在方面观点判定为噪声数据。

(2)情感强度计算。

实际情况中，方面观点可能存在多个情感修饰词同时搭配一个情感词的情况，且“否定词+程度副词+情感词(ne+dg+se)”与“程度副词+否定词+情感词(dg+ne+se)”的情况表达不同的情感态度。例如评价句“口感不是很好”与“口感很不好”相比，后者的负面情感强度更大。针对单个方面观点中 5 种不同搭配情况，情感强度的计算方法如公式(7)所示。

$$q_i=\begin{cases} V_{ne+dg+se}=sen(se)\times(-1)^{|ne|}\times\left(1+0.2\times\left(-\frac{1}{\sum_{j=1}^{n}W_{dg}^{j}}\right)\right)\\ V_{dg+ne+se}=sen(se)\times(-1)^{|ne|}\times\left(1+0.2\times\sum_{j=1}^{n}W_{dg}^{j}\right)\\ V_{dg+se}=sen(se)\times\left(1+0.2\times\sum_{j=1}^{n}W_{dg}^{j}\right)\\ V_{ne+se}=sen(se)\times(-1)^{|ne|}\\ V_{se}=sen(se)\end{cases} \tag{7}$$

式中，q_i 表示评价句中的第 i 个方面观点情感强度，$sen(se)$ 表示情感词 se 的情感极性，W_{dg}^{j} 表示第 j 个程度副词的权重，n 表示程度副词的个数，$|ne|$ 表示否定词的个数。另一方面，评价中多个方面观点可能同时表达一个领域属性，因此有必要对所有方面观点情感强度进行聚合，计算方法如公式(8)所示。

$$V_b=\frac{1}{N_b}\sum_{c=1}^{N_b} q_{i\_c} \tag{8}$$

式中，N_b 表示评价中提及属性 b 的方面观点个数，q_{i_c} 表示第 i 个方面观点的情感强度，V_b 表示评价在属性 b 的方面情感强度，若 V_b 大于 0 表示评价在属性 b 的正面表达，小于 0 表示负面表达，等于 0 表示无情感表达。根据计算结果对每条评价进行方面情感倾向标注。

1.4 消费者偏好挖掘

1.4.1 关注度分析

统计评价的观点抽取和情感识别结果，可以反映消费者的偏好。一方面，现有研究指出消费者对不同主题属性有不同的重视程度[23]，该文引入关注度的概念衡量这方面的差异，计算方法如公式(9)所示。

$$A(u_j^i)=\frac{c_j^i}{c_j} \tag{9}$$

式中，$A(u_j^i)$ 表示第 j 类产品第 i 个主题属性的消费者关注度，c_j^i 表示第 j 类产品第 i 个主题属性被提及的次数，c_j 表示第 j 类产品的评价个数。

1.4.2 满意度分析

另一方面，现有研究集中在从评价中提取消费者偏好信息衡量对产品的满意度[24]。该文计算在每个领域主题属性下的正面情感评价数量与评价总数的比率，量化消费者对农产品属性满意度差异，计算方法如公式(10)所示。

$$p(u_j^i)=\frac{\sum C_j^{i>0}}{\sum C_j^{i>0}+\sum C_j^{i<0}} \tag{10}$$

式中，$p(u_j^i)$ 表示消费者对第 j 类产品第 i 个主题属性的满意度，$C_j^{i>0}$ 表示第 j 类产品第 i 个主题属性的正面情感评价，$C_j^{i<0}$ 表示第 j 类产品第 i 个主题属性的负面情感评价。

1.4.3 产品优势与缺陷分析

将方面观点三元组的抽取结果按照主题属性和情感值是否大于 0 进行划分，分别对属性词、情感词与整体方面观点表达进行词频统计，选取正面方面观点的高频表达作为农产品的主题属性优势，选取负面方面观点的高频表达作为农产品的主题属性缺陷。

2 模型性能分析与实际应用

2.1 实验设计

2.1.1 验证集来源

由于目前还没有开源的农产品评价标注数据，该文从农产品评价数据集中随机抽取 1 000 条数据，提取其中包含的方面观点后，人工标注每条评价的情感标签作为验证集，1 表示正面情感，-1 表示负面情感，-2 表示无情感，验证集的示例如表 4 所示。

2.1.2 实验方案

一方面，将文中模型与传统的无监督识别方面级情感的方法进行对比，对比模型选用 Zhang 等[25]提出的基于 PMI 统计方法模型(Baseline_pmi)，Sun 等[16]

提出的基于共现矩阵相似度模型(Baseline_co_sim),以及张林[26]提出的基于 Hanlp 工具的依存句法规则和情感值计算的模型(Hanlp_model)。另一方面,通过消融试验验证文中模型中各个方法所起的作用。在依存句法规则的基础上使用领域情感词典和词嵌入是关键的,前者侧重于识别隐式观点,后者侧重于识别潜在观点。因此设定其它三种模型分别为:仅运用依存句法规则的模型(Ltp_dp)、加入词嵌入相似度判别的模型(Ltp_dp_vec)、加入 SO-PMI 构建领域情感词典的模型(Ltp_dp_sopmi),并与文中模型进行对比。

表 4　验证集的示例

评价文本	包含的方面观点	情感标签
口感挺好的,整体上很满意,茶叶的汤色和味道都比较满意,性价比还是非常不错,第一泡的时候汤色碧绿	<口感,挺,好><汤色,比较,满意><味道,比较,满意><性价比,非常,不错><汤色,碧绿>	口感:1,品质:-2,色泽:1,价格:1,分量:-2,包装:1,物流:-2,售后:-2

2.1.3　验证指标选取

选取机器学习验证指标(查准率、召回率和 F1 值)对模型进行性能度量[27]。查准率指正确识别情感标签占所有识别出情感标签数量的比例,召回率指正确识别情感标签占所有情感标签的比例。F1 值是准确率与召回率的平均值,并根据每类情感标签分类结果计算宏平均值反映模型整体性能。上述验证指标的计算方法如公式(11)~(13)所示。

$$P = \frac{\text{识别正确的情感标签数量}}{\text{识别出的情感标签数量}} \times 100\% \quad (11)$$

$$R = \frac{\text{识别正确的情感标签数量}}{\text{情感标签总数量}} \times 100\% \quad (12)$$

$$\text{Macro\_F1} = \frac{1}{n}\sum_{i=1}^{n}\frac{2 \times P \times R}{P + R} \times 100\% \quad (13)$$

式中,P 表示查准率,R 表示召回率,Macro_F1 表示所有情感标签类别的宏平均 F1 值。

2.2　模型性能比较

模型性能比较结果如表 5 所示。传统 Baseline_pmi 和 Baseline_co_sim 模型查准率和召回率较低,原因是基于统计或共现的方法忽略了词汇句法关系造成大量噪声,且无法较好地关注到低频词。单独使用依存句法的模型 Ltp_dp 召回率较低,这是因为忽略了评价中的隐式和潜在情感信息。加入词嵌入相似度的模型 Ltp_dp_vec 与加入 SO-PMI 构建领域情感词典的模型 Ltp_dp_sopmi,召回率和 F1 值均有所提高,前者召回了部分潜在情感信息,后者抽取了更多隐式情感信息。文中模型的查准率、召回率和 F1 值分别为 85.08%,78.50% 和 81.66%,尤其是召回率和 F1 值提升显著,达到了最优的模型性能。相比之下,基于 Hanlp 工具的模型在分词和依存句法标注上的精确性不如 LTP,且忽略 SO-PMI 和词嵌入相似度最优阈值的设定也将降低模型性能。

表 5　不同模型试验结果对比　　%

试验语料	模型	查准率	召回率	F1 值
农产品评价	Baseline_pmi	75.91	59.20	66.57
	Baseline_co_sim	78.22	67.27	72.33
	Ltp_dp	83.27	51.67	63.77
	Ltp_dp_vec	82.85	55.23	66.28
	Ltp_dp_sopmi	85.94	72.44	78.61
	Hanlp_model	83.02	74.87	78.81
	文中模型	85.08	78.50	81.66

2.3　系统的搭建与应用

基于该文设计的算法,使用 Django 后端 + Booststrap 前端框架作为主体框架进行观点抽取和情感识别系统的搭建,使用 Layui 和 Echarts 组件进行分析结果可视化展现。在系统的前端部分,分别设计了文本输入模块与根据农产品品类、品种、品牌、店铺名称进行条件筛选模块,用户提交表单后实时返回查询的结果。在系统的后端部分,设计了数据接收、数据处理和数据提交功能,前后端之间通过 ajax 进行 JSON 数据的传输。根据系统功能的不同,可分为观点抽取和情感识别功能、消费者偏好挖掘功能两类。

2.3.1　观点抽取和情感识别功能

该功能可以分为在线分析与离线分析,在线分析功能的界面展示如图 2 所示。系统用户输入需要判别的农产品评价数据、选择品类并点击提交后,前端向服务器发送请求,Django 后端实时调用该文的算法接口,返回面向农产品主题属性的有效方面观点三元组抽取结果和对应的情感强度得分结果。

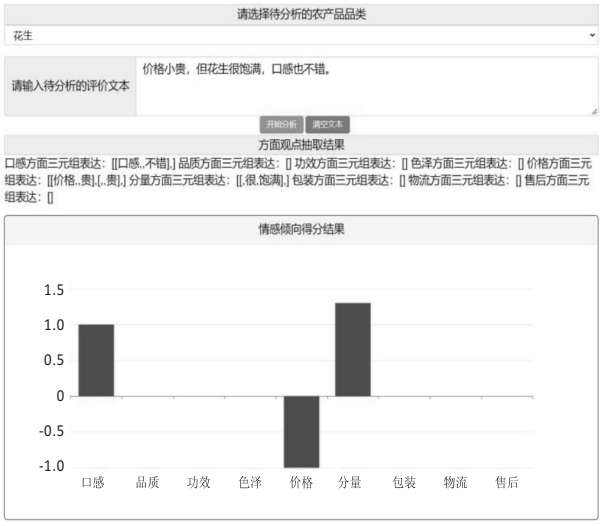

图 2　在线分析功能界面展示

另一方面,系统启动时后端通过调用评价爬虫模

块、观点抽取和情感识别模块对 MySQL 中的评价数据进行持续的增量更新。离线分析功能的界面展示如图 3 所示，系统用户可在前端通过下拉框筛选需要查询的农产品品类、来源、店铺、品种、商品链接所对应的评价数据，每条评价都展现所抽取的方面观点三元组和生成的方面情感倾向信息。

ID	用户昵称	用户评价	粗粒度情感标签	细粒度情感标签	方面观点三元组	所属URL	所属店铺	评价时间	情感判别时间
1121484	None	优雅清高的自然花香气；浓郁、甘醇、爽口、回甘的滋味；橙黄清澈明亮的汤色；青蒂绿腹红镶边的叶底和极耐冲泡的底力，构成凤凰单丛茶特有的色、香、味特点。	positive	口感:positive,品质:null,功效:null,色泽:positive,价格:null,分量:null,包装null,物流:null,服务:null	[滋味,,回甘],[,,浓郁],[,,爽口],[,,香],[,,清澈],[汤色,,明亮],	https://item.jd.com/100031668416.html	传奇会京东自营旗舰店	2021年9月7日 20:53	2022年10月29日 22:12
1121579	None	机器采摘的断叶，此茶贵在不是正宗福建产	negative	口感:null,品质:negative,功效:null,色泽:null,价格:negative,分量:null,包装null,物流:null,服务:null	[,不是,正宗],[,,贵],	https://item.jd.com/100031668416.html	传奇会京东自营旗舰店	2020年5月1日 19:42	2022年10月29日 22:11
1121486	None	包装：包装结实，内赠礼品袋，包装无破损叶底：茶底肥厚香味：第一次喝这种香型茶感觉还可以汤色：汤色很好耐泡度：比较耐泡	positive	口感:positive,品质:positive,功效:null,色泽:positive,价格:null,分量:positive,包装:positive,物流:null,服务:null	[,,香味],[,比较,耐泡],[汤色,很,好][,,肥厚],[包装,,结实],[包装,无,破损],	https://item.jd.com/100031668416.html	传奇会京东自营旗舰店	2020年5月17日 08:13	2022年10月29日 22:10

图 3 离线分析结果查询界面展示

2.3.2 消费者偏好挖掘功能

对农产品评价语料进行观点抽取和情感识别后，可根据公式(9)和(10)分别计算不同农产品品类、品种、品牌和店铺下消费者对领域主题属性的关注度和满意度差异。图 4 以农产品品类为例分析了上述区别。在关注度方面，与其它属性相比，消费者更侧重于关注口感和价格属性，平均值分别达到 0.29 和 0.26。但不同品类之间存在关注度的差异。例如，消费者对于竹笋和花生的口感属性关注度分别为 0.40 和 0.26，显著高于其它农产品；对于猕猴桃的价格属性关注度达到了 0.30，与之相比消费者对茶叶的价格属性关注度较低，仅为 0.19。在满意度方面，在线农产品的包装和物流属性被多数消费者认可，满意度均值分别达到了 0.84 和 0.86。在具体的口感属性方面，花生品类得到了最多消费者的喜爱，满意度达到了 0.90，但猕猴桃的口感满意度仅为 0.61；在色泽属性方面，茶叶品类的满意度最高达到了 0.89，但芒果和猕猴桃的满意度较低，分别仅为 0.54 和 0.57。

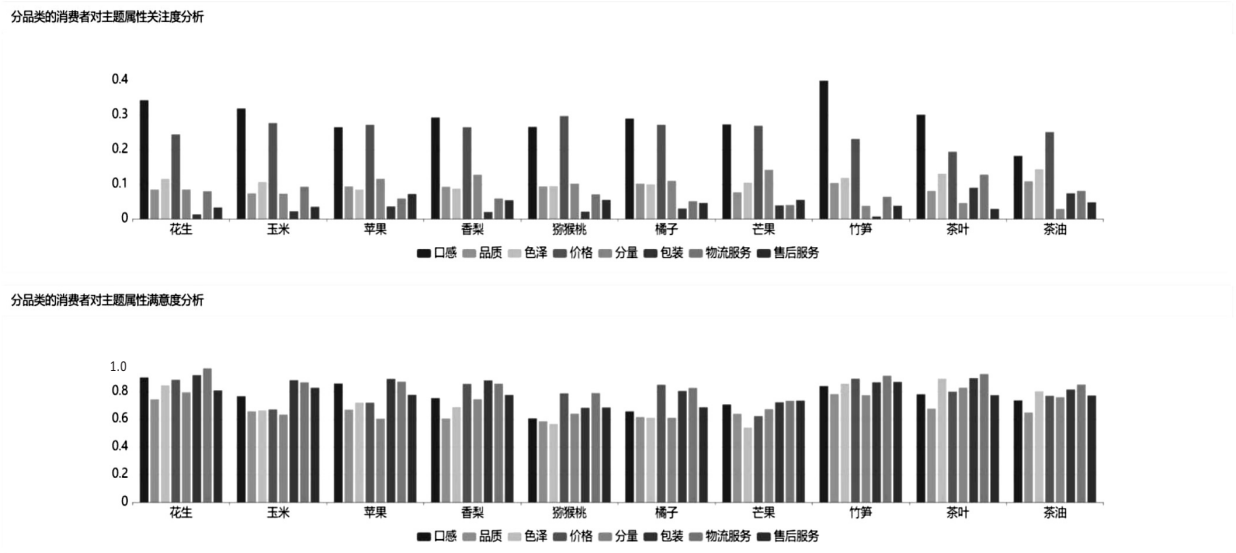

图 4 消费者关注度和满意度分析界面展示

另一方面，对海量评价数据抽取的有效方面观点三元组进行词频统计和排序，为平台用户提供产品的属性级优势与缺陷分析，以售卖猕猴桃品类的店铺“绿美生鲜专营店”的口感属性缺陷的高频情感词为例进行分析，图 5 展示了分析结果，可以看出该店铺在口感方面的缺陷主要体现在“硬”“酸”“不甜”“难吃”等。

该系统面向的用户主要分为消费者、经销商、平台和监管部分四个主体。消费者可以通过历史消费者的评价观点挖掘结果进行购买选择，农产品经销商和平台可以根据消费者偏好挖掘结果更有针对性地对产品进行具体属性的改进，监管部门通过查询具体店铺的农产品属性级缺陷，及时发现问题并进行管控。该平台上线后取得了良好的应用效果。

按品类搜索 猕猴桃　按商品所属店铺搜索 绿美鲜生鲜专营店　按商品优缺点搜索 商品缺陷　按商品主题搜索 口感　按词性搜索 情感词　多条件搜索

农产品优势与缺陷统计列表

词汇/词组ID	高频词汇/词组表达	所属词性类型	词汇/词组频率统计	情感极性	所属主题	所属店铺	品类	信息更新时间
35677	硬	情感词	64	商品缺陷	口感	绿美鲜生鲜专营店	猕猴桃	2022年10月18日 12:47
35678	酸	情感词	51	商品缺陷	口感	绿美鲜生鲜专营店	猕猴桃	2022年10月18日 12:47
35679	很硬	情感词	12	商品缺陷	口感	绿美鲜生鲜专营店	猕猴桃	2022年10月18日 12:47
35680	不甜	情感词	12	商品缺陷	口感	绿美鲜生鲜专营店	猕猴桃	2022年10月18日 12:47
35681	难吃	情感词	6	商品缺陷	口感	绿美鲜生鲜专营店	猕猴桃	2022年10月18日 12:47
35682	太硬	情感词	6	商品缺陷	口感	绿美鲜生鲜专营店	猕猴桃	2022年10月18日 12:47
35683	一般	情感词	6	商品缺陷	口感	绿美鲜生鲜专营店	猕猴桃	2022年10月18日 12:47
35684	还硬	情感词	6	商品缺陷	口感	绿美鲜生鲜专营店	猕猴桃	2022年10月18日 12:47

图 5　具体店铺的口感属性缺陷分析界面展示

3　结束语

针对如何在缺少标注数据集的情况下对农产品评价进行观点抽取和情感识别的问题,提出了一种组合 NLP 方法的无监督学习框架。采集多源评价数据后,首先采用 LDA 模型对领域主题进行识别,运用改进的 SO-PMI 算法构建领域词典,然后结合依存句法规则和词嵌入相似度对评价中的隐式和潜在方面观点进行更好地召回,并定义了情感强度的计算方法标注方面级情感倾向。相较于传统模型,该文的分析识别框架在性能上得到明显的提升,在农产品评价语料中查准率、召回率和 F1 值分别达到 85.08%、78.50%、81.66%。基于该算法框架搭建的可视化系统具有在线、离线识别功能和基于不同角度的消费者偏好分析功能,分析结果致力于为不同的主体提供信息价值。

参考文献:

[1] AKTER S, WAMBA S F. Big data analytics in E-commerce: a systematic review and agenda for future research [J]. Electronic Markets, 2016, 26(2): 173-194.

[2] MULLEN T, COLLIER N. Sentiment analysis using support vector machines with diverse information sources[C]//Proceedings of the 2004 conference on empirical methods in natural language processing. Barcelona: ACM, 2004: 412-418.

[3] FINK C R, CHOU D S, KOPECKY J J, et al. Coarse- and fine-grained sentiment analysis of social media text[J]. Johns Hopkins Apl Technical Digest, 2011, 30(1): 22-30.

[4] LUO F, LI C, CAO Z. Affective-feature-based sentiment analysis using SVM classifier[C]//2016 IEEE 20th international conference on computer supported cooperative work in design (CSCWD). Nanchang: IEEE, 2016: 276-281.

[5] QIN Y, WANG X, XU Z. Ranking tourist attractions through online reviews: a novel method with intuitionistic and hesitant fuzzy information based on sentiment analysis[J]. International Journal of Fuzzy Systems, 2022, 24(2): 755-777.

[6] 孙玲玲,胡彦蓉,刘洪久. 基于产品特征细粒度情感分析的在线品牌社群用户评论挖掘[J]. 数学的实践与认识, 2021, 51(24): 83-95.

[7] SCHOUTEN K, FRASINCAR F. Survey on aspect-level sentiment analysis[J]. IEEE Transactions on Knowledge and Data Engineering, 2015, 28(3): 813-830.

[8] JIN W, HO H H, SRIHARI R K. OpinionMiner: a novel machine learning system for web opinion mining and extraction [C]//Proceedings of the 15th ACM SIGKDD international conference on knowledge discovery and data mining. Paris: ACM, 2009: 1195-1204.

[9] CAMBRIA E, SCHULLER B, LIU B, et al. Knowledge-based approaches to concept-level sentiment analysis[J]. IEEE Intelligent Systems, 2013, 28(2): 12-14.

[10] 贾　川,方　睿,浦　东,等. 基于循环实体网络的细粒度情感分析[J]. 中文信息学报, 2019, 33(9): 123-128.

[11] 万　岩,杜振中. 融合情感词典和语义规则的微博评论细粒度情感分析[J]. 情报探索, 2020(11): 34-41.

[12] XU H, ZHANG F, WANG W. Implicit feature identification in Chinese reviews using explicit topic mining model[J]. Knowledge-Based Systems, 2015, 76: 166-175.

[13] LI Z, ZOU Y, ZHANG C, et al. Learning implicit sentiment in aspect-based sentiment analysis with supervised contrastive pre-training[J]. arXiv: 2111.02194, 2021.

[14] TUBISHAT M, IDRIS N, ABUSHARIAH M. Implicit aspect extraction in sentiment analysis: review, taxonomy, oppportunities, and open challenges[J]. Information Processing & Management, 2018, 54(4): 545-563.

[15] HU M, LIU B. Mining and summarizing customer reviews [C]//Proceedings of the tenth ACM SIGKDD international conference on knowledge discovery and data mining. Seattle: ACM, 2004: 168-177.

[16] SUN L, LI S, LI J Y, et al. A novel context-based implicit feature extracting method[C]//2014 international conference on data science and advanced analytics (DSAA). Shanghai: IEEE, 2014: 420-424.

[17] 华静伟. 基于依存关系改进的方面级观点挖掘算法及可视分析研究[D]. 燕山:燕山大学, 2021.

(下转第 130 页)