基于 TF-IDF 算法的农产品消费者购买情感分析
——来自京东电商平台在线点评数据

邓颖仪[1],邱秀芳[1]*,黄华乾[1],庞 青[2] (1. 广东工贸职业技术学院,广东广州 510510; 2. 京东教育研究院,北京 100000)

摘要 “互联网+农产品”模式大大拓宽了农产品的销售渠道,为乡村振兴注入活力。于京东电商平台甄选出点评数达 200 条以上的农产品共 2 090 种 112 779 条评论,借助 Hanlp 工具对评论文本进行分词,并利用 TF-IDF 算法对特征词进行关注度分析。研究表明,消费者网购农产品过程中,较关心农产品的新鲜度、品质、快递服务及包装;消费者网购水果类、蔬菜类等农产品的比例最高,其中对水果类“甜”“熟”“酸”等特征词的关注度最高,对蔬菜类“糯”“好吃”“香甜”等特征词关注度最高。以上结论既丰富了农产品网络营销的相关研究,也可用于指导农产品的网络营销实践。

关键词 农产品;网络营销;TF-IDF 算法;京东
中图分类号 S-058 **文献标识码** A
文章编号 0517-6611(2022)11-0203-04
doi:10.3969/j.issn.0517-6611.2022.11.051
开放科学(资源服务)标识码(OSID):

Analysis on the Consumption Emotions of Agricultural Products Based on TF-IDF Algorithm—From Online Review Data of JD E-Commerce Platform

DENG Ying-yi, QIU Xiu-fang, HUANG Hua-qian et al (Guangdong Polytechnic of Industry and Commerce, Guangzhou, Guangdong 510510)

Abstract The Internet plus agricultural products marketing model has greatly broadened the distribution channels of agricultural products and injected vitality into the development and revitalization of rural economy. Based on the JD e-commerce platform, we selected a total number of 112 779 comments from 2 090 kind of agricultural products with online comments number over 200. The Hanlp tool was used to carry out the word segmentations of these comments. Finally TF-IDF algorithm was used to analyze the attention-degree of the feature words. Research results showed that consumers paid more attention to the freshness, quality, express service and packaging of agricultural products while shopping online; the best-selling agricultural products online were fruits, vegetables, among which consumers paid the highest attention to the fruit sweetness, ripeness and sourness, as well the vegetable waxiness, deliciousness and sweetness. The above conclusion not only enriched the research of agricultural products network marketing, but also could be used to guide the network marketing practice of agricultural products.

Key words Agricultural products; Network marketing; TF-IDF algorithm; JD e-commerce platform

根据《中华人民共和国农产品质量安全法》,农产品是指来源于农业的初级产品,即在农业活动中获得的植物、动物、微生物及其产品。农产品电子商务就是消费者和销售商利用电子数据传输技术,在线上完成农产品交易的商务活动。在非接触的农产品网购环境下,消费者的购后在线评论会影响农产品电商厂家的销量和发展[1]。由于信息不对称,消费者在电商平台上选购不同类型的农产品将面临决策风险[2]。消费者在线选购农产品时,不能直观感受其性价比和鲜活度,只能参考在线评论做出购买决策[3]。因此,在线评论属于网络口碑的范畴,是指消费者通过电商平台选购商品,在商品送达后根据其性价比做出相应的评价,并在商家评论区与其他消费群体进行互动和交流[4]。

在线评论的各项内容能够加深消费者对商品的了解,减少消费者心中对商品出现的不确定性,帮助消费者做出相应的选购决策[5]。在线评论作为网购环境的一种有效的信任机制,已成为学界和业界关注的热点话题[6]。相关研究表明,在线评论不仅影响消费群体的购买行为,而且更会影响电商平台商品总体的销售量[7]。消费者所购买的产品种类不同,则给出的在线评价内容也不尽相同,评价结果产生的影响力同样也会有所差异[8],因此产品类型能够对在线评论呈现出的有用性程度进行有效的调节[9]。从农产品角度来看,消费者在购买这类产品时,会非常注重产品质量和产品安全性[10],而在线评论作为消费者了解和熟知农产品属性的重要渠道,对于电商平台销售商家的发展有一定影响。在线评论次数作为核心评价指标,对产品销量有显著的影响[11]。目前,电商平台的在线评论机制主要面向已经购买商品的消费者。因此,某种商品的评论次数愈多则销量愈好,间接表明该商品深受广大消费者的支持和认可。在这种情况下,购买决策面临的不确定风险较低,且交易环节的成本支出也较少[12]。因此,挖掘电商平台关于农产品的在线评论数据,并对消费者情感进行分析,具有重要的理论意义和实践意义。鉴于此,笔者于京东电商平台甄选出点评数达 200 条以上的农产品共 2 090 种 112 779 条评论,借助 Hanlp 工具对评论文本进行分词,并利用 TF-IDF 算法对特征词进行关注度分析。

1 TF-IDF 算法

TF-IDF 算法的主要原理是如果一个单词在该文章出现的频率(TF)高,并且在其他文章中出现频率很低,则认为该单词具有很好的区分能力,适合用来进行分类。

1.1 词频(Term Frequency) 词频表示单词在该类农产品评论中出现的频率。

基金项目 广东省 2020 年度普通高校重点科研平台和项目“乡村振兴战略下校政企协同开展农村电商人才培养的长效机制研究”(2020ZDZX1072)。
作者简介 邓颖仪(1988—),女,广东广州人,讲师,博士,从事消费者行为研究。*通信作者,教授,硕士,从事高职教育、创新创业教育、教师发展研究。
收稿日期 2021-12-25;**修回日期** 2022-01-06

$$TF_{ij}=\frac{n_{ij}}{\sum_k n_{kj}} \quad (1)$$

式中,TF_{ij} 表示分词 t_i 在文档编号 d_j 中出现的频率。分子代表分词 t_i 在文档 d_j 中出现的次数,分母表示文档 d_j 中所有词出现次数的总和。

1.2 逆向文档频率(Inverse Document Frequency) 表示某一个特定单词 IDF 可以由总文章数除以包含该单词的文章数,再将得到的商取对数。如果包含该单词的文章越少,则 IDF 越大,表明该单词具有很好的文章区分能力。

$$IDF_i=\log\frac{|D|}{|\{j:t_i\in d_j\}|} \quad (2)$$

式中,$|D|$代表文档总数,$|\{j:t_i\in d_j\}|$代表包含了分词 t_i 的文件数。由此,可以计算某一个词语的 TF-IDF 值:

TF-IDF=词频(TF)×逆文档频率(IDF) (3)

TF-IDF 算法可用来提取点评文本内容的关键词、摘要、核心关注点,该算法简单快速、普适性强,可以推广到各种文本类分析场景。TF-IDF 值越大,说明该词在该文档中出现的频率越高,可以作为该文档的关键。

2 数据采集与清洗

2.1 数据采集 通过比较淘宝、天猫、拼多多等电商平台,发现京东电商平台(http://www.jd.com)在售农产品具有种类丰富、交易量大、在线评论多等特点,因而以此作为数据采集来源。截至 2021 年 9 月 21 日,京东电商平台以农产品为关键词共检索到 6 000 余条农产品广告展示,采用 Scrapy 爬虫框架爬取数据,并甄选出点评数达 200 以上的农产品 2 090 种。为保证数据样本质量和网络点评分析的准确性,过滤重复、无实质性内容的评论。

2.2 数据清洗 由于在线评论文本数据中存在着大量的无关数据、重复数据、无效数据等,这些数据没有实际意义,还可能对结果产生影响。基于 Spark 计算框架结合 Scala 编程语言,实现对采集后的原始评论数据进行清洗、规整、补录和统计。第一,针对搜集到的评论文本进行清洗,删除针对研究没有意义的无效、重复和缺失评论数据;第二,将搜集到的部分半格式化的信息进行格式化,如将包含中文的评论数字段规整为整型字段等;第三,补录部分空缺的关键字段,针对空缺的商品种类、名称等关键字段进行补录。数据清洗流程为之后的数据分析、数据建模提供优质的基础数据。最后,针对评论文本的分词结果进行清洗,删除低频、无意义的词语,总共搜集的评论数 114 724 条,其中好评数据为 80 849 条,差评为 33 875 条。经清洗过后,有效好评数据为 79 476 条,差评为 33 303 条。清洗部分无效数据及其对应分类如表 1 所示。

表 1 无效评论文本数据示例

Table 1 Text data of invalid comments

序号 Code	类别 Type	评论文本 Comment text
1	数字类	5560900
2	符号类	&+&￥#@
3	未填写类	此用户未填写评价内容
4	文本简短	好、不错、可以

3 数据分析结果

3.1 数据分词 运用 Hanlp 分词工具对在线评论文本进行分词,该框架分词性能较好,且支持用户自定义词语,比如“不好吃”“不便宜”等单词,用普通分词会切分为“不”“好吃”“不”“便宜”,导致切分后的语义发生严重误导,利用 Hanlp 分词工具可以将“不好吃”“不便宜”设置为自定义单词,使切词后不失去原来的语义。部分处理样例结果如表 2 所示。

表 2 分词结果

Table 2 Word segmentation results

序号 Code	类别 Type	在线评论文本数据 Text data of online comment	分词结果 Word segmentation
1	水果类	包装严实,内部层层包裹,运输安全,无坏果,口感适中	包装/严实/内部/层层/包裹/运输/安全/无/坏/果/口感/适中
2	粮油类	物流很快,真空包装很好,蒸饭时能闻到米香味	物流/很快/真空/包装/很好/蒸饭/时/能/闻到/米/香味
3	蔬菜类	板栗南瓜,京东购物真的很划算	板栗/南瓜/京东/购物/真的/很/划算
4	肉禽类	配有冰袋,肉肥瘦均匀,口感很好	配有/冰袋/肉/肥瘦/均匀/口感/很好
5	药材类	包装很好,味道很棒,药材很新鲜	包装/很好/味道/很棒/药材/很/新鲜
6	干货类	香菇挺不错,看着就好,这个价格划算	香菇/挺不错/看着/就/好/这个/价格/划算
7	水产类	顺丰速递真的快,带鱼共 10 条,家乡的味道	顺丰/速递/真的/快/带鱼/共/十/条/家乡/的/味道
8	茶	很香,很干,无杂质	很香/很干/无杂质

为了研究的普遍性,将分词结果中无意义的语气词、副词等,如“的”“么”“一方面”“快”“可以”等加入停用词表进行过滤。

3.2 消费者情感分析 运行 Spark 计算框架,得到有效正面评价特征词 1 323 个,有效负面评价特征词 693 个,表 3 为排名前 5 的分词结果。

分别提取正面评价和负面评价词频前 40 的特征词,制作词云分布图,结果如图 1 所示。

图 1a 正面评价词云中“很好吃”“包装”“顺丰”“味道”等特征词的词频较高,说明消费者对网购农产品的口感、顺丰快递服务及包装等给予较高的好评。图 1b 负面评价词云中“坏”“烂”“差”“发货慢”等特征词的词频较高,说明消费者对网购农产品的品质、快递服务等给予较低的评价。总体而言,消费者网购农产品过程中,较关心农产品的新鲜度、品质、快递服务及包装。

3.3 农产品分类情况 通过对在线评论数据各个维度进行

表 3　正面与负面评价排名前 5 的特征词

Table 3　Characteristic words of top 5 positive and negative evaluation

评价 Evaluation	特征词 Feature word	词频 Word frequency	排名 Rank
正面 Positive	很好吃	26 553	1
	很不错	20 565	2
	味道	12 521	3
	包装	12 141	4
	顺丰	11 948	5
负面 Negative	坏	4 733	1
	差	4 181	2
	包装破损	3 163	3
	烂	3 159	4
	服务态度差	3 018	5

统计分析,将农产品划分为 8 大类,包括水果类、蔬菜类、肉禽类、药材类、粮油类、水产类、干货类、茶叶类(表 4)。

平均评论数越多则可以认为购买人数越多,表 4 中可见数据情况基本符合现实的消费者在电商平台农产品的购买行为。消费者购买水果类、蔬菜类占多数,水果类相比蔬菜类具有地域性强的特点;其中茶叶类购买人数最少,茶叶是属于感官体验型产品,对于品茶的人来说,线下实体店适合品鉴是必不可少的环节,所以与其他农产品相比,消费者在电商平台上购买茶叶较少。

3.4　基于 TF-IDF 算法的关注度分析　该研究基于搜集的 8 大类农产品,对归属于相同产品种类的商品好评内容进行聚类,形成 8 个不同的大文档,并创新使用 TF-IDF 算法对这 8 个文档进行分析,挖掘不同种类的农产品影响消费者积极消费行为的个性化核心主题,这些主题代表着消费者针对该类农产品有异于非同类产品的观点和关注重点,以便根据不同的商家推荐更符合该产品的营销维度,为农产品的个性化营销模式提供决策方向。

a. 正面评价词云　　b. 负面评价词云

注:词频越高字号越大

Note:The bigger word size indicated higher frequency

图 1　词云分布

Fig. 1　The word cloud distribution

表 4　不同类型农产品的评论数比较

Table 4　Comparison of the comment number of different types of agricultural products

序号 Code	农产品类型 Agricultural products type	评论数 Comment number
1	水果类	13 830
2	蔬菜类	8 677
3	肉禽类	5 405
4	药材类	2 465
5	粮油类	1 842
6	水产类	1 306
7	干货类	1 213
8	茶叶类	1 084

基于 TF-IDF 算法的分析结果如表 5 所示,按照概率大小排序,分别把 8 大类农产品排名前 10 个特征词抽取出来(表 5)。关键词权重越大,说明其针对该类产品的关注度越高,商家在做广告以及推销产品时候可以用此类情感词,达到吸引消费者购买的效果。

3.4.1　水果类。该类农产品的特征词是熟、水分、维生素、物超所值、完好无损等,说明了消费者对水果最关心的是是否含有丰富的维生素和水分,因为丰富的维生素和水分可以给身体提供营养成分。销售者可以多推销苹果、橙子、猕猴桃等富含维生素的水果,并且强调层层包裹完好无损,如有损坏可免费退还。

3.4.2　蔬菜类。该类农产品的特征词是糯、好吃、品质、绿色健康等,说明了随着生活水平的提高,消费者对蔬菜的需求不仅仅局限于够吃,还追求口感品质、绿色健康等方面。销售者可以主打绿色主题,捕捉消费者心态,多推销有机蔬菜,另外在包装处可附有产地、采摘时间、营养价值等详情,增加可信度。

3.4.3　肉禽类。该类农产品的特征词是肉质、生鲜、顺丰等,其中肉质这个特征词说明消费者对家禽类养殖的焦点在于养殖过程中是否用激素催生,是否采用工业化养殖饲料。销售者在推销家禽类时可以重点强调用谷、饭、菜叶或小虫子喂养,养殖地空气质量高,因此肌肉质地结实,加上配送只限顺丰速递,保障肉质鲜美。

3.4.4　药材类。该类农产品的特征词是泡水、味浓、货真价实、方便等,说明消费者偏向喜欢药材能够直接泡水代替煎熬,方便服用。最近几年药材类为了迎合消费者需要,特别

是中药类研制了颗粒,可直接泡水服用,代替瓦煲煎熬,销售者在推销药材的时候可以在这方面多加强调。

表 5 不同种类农产品特征词 TF-IDF 比较

Table 5 Comparison of TF-IDF of characteristic words of different types of agricultural products

种类 Type	特征词 Characteristic words	TF-IDF	种类 Type	特征词 Characteristic words	TF-IDF
水果类	甜	17.627	粮油类	精美	6.496
Fruit	熟	14.162	Grain and	产品包装	6.070
	酸	12.993	oil	拌饭	5.197
	水分	8.486		好吃	4.861
	维生素	6.496		日期	4.046
	新鲜	3.908		赠品	3.154
	饱满	3.637		家乡	3.031
	价格便宜	3.154		老人	2.599
	物超所值	3.035		价廉物美	2.023
	完好无损	2.425		新疆	1.577
蔬菜类	糯	15.819	水产类	鲜美	5.519
Vegetable	好吃	5.337	Aquatic	个头	3.185
	香甜	5.058	product	很给力	2.425
	品质	3.154		冰块	2.365
	绿色健康	3.035		存活率	2.229
	熟	3.035		鲜嫩	2.023
	均匀	2.365		分量	1.605
	外包装	2.023		肉质	1.577
	品种	1.808		运输	1.356
	生产日期	1.577		广东	1.299
肉禽类	肉质	7.096	干货类	厚	7.081
Meat and	生鲜	6.070	Dried	饱满	5.455
poultry	结实	3.898	goods	信得过	3.898
	缺斤少两	3.898		炖肉	3.154
	顺丰	3.637		产品质量	3.035
	美味	3.154		试吃	2.599
	日期	3.035		煲汤	2.365
	弹性	2.599		新疆	2.365
	冰袋	2.365		密封	1.577
	嫩	2.229		解馋	1.299
药材类	泡水	4.046	茶叶类	天然	2.599
Medicinal	精神	2.599	Tea	泡水	2.023
material	味浓	2.599		减肥	1.577
	货真价实	2.365		无杂质	1.299
	质量	2.192		色彩	1.299
	好喝	1.808		清香	1.212
	正品	1.577		降血压	1.012
	干燥	1.299		甘甜	1.012
	方便	1.274		降尿酸	1.818
	味十足	1.012		美丽	1.012

3.4.5 粮油类。该类农产品的特征词是精美、产品包装、拌饭、老人、家乡等,一方面说明消费者偏好精美的包装,由此可推测粮油类除了家庭日常所需之外,还是节日送礼的首选;另一方面深受家里老人喜欢,具有家乡的味道。销售者可以在包装方面发挥想象,设计别与其他商家的包装,吸引消费者的眼球。

3.4.6 水产类。该类农产品的特征词是鲜美、冰块、存活率、广东等,说明消费者偏好广东的海鲜,在配送过程中喜欢有冰块保鲜,这样可以提高海鲜的存活率。销售者可以多推销产地是广东的海鲜,配送中加入冰块,确保存活率。

3.4.7 干货类。该类农产品的特征词是饱满、炖肉、煲汤、解馋、新疆等,说明消费者偏好新疆产的干货,用途大多用于煲汤或者闲余时间解馋。销售者可以多入手产地是新疆的干货类,满足消费者的需求。

3.4.8 茶叶类。该类农产品的特征词是泡水、减肥、无杂质、降血压等,泡水喝能够起到到扩张血管,清热解暑以及降血压的作用,由此对于女性消费者而言,通过喝茶能够达到减肥的效果。销售者可以在推销茶叶的时候,针对具有减肥、降血压等功效的茶类品种多做广告,加大口碑宣传。

4 结论与展望

该研究采用 Scrapy 爬虫框架爬取京东电商平台农产品网络消费者在线评论数据,利用 Spark 计算框架分析消费者正负面的情感评价,并进一步采用 TF-IDF 算法对不同类型农产品的消费者关注度进行分析。结果表明,①消费者网购农产品过程中,较关心农产品的新鲜度、品质、快递服务及包装;②消费者网购水果类、蔬菜类等农产品的比例最高,其中对水果类甜、熟、酸等特征词的关注度最高,对蔬菜类糯、好吃、香甜等特征词关注度最高。

理论上,该研究丰富了农产品网络营销的相关研究,创新性地引入 TF-IDF 算法对不同类型农产品的消费者关注度进行分析,拓展了 TF-IDF 的应用领域。实践上,该研究结果和结论可为农产品网络营销实践提供指导,即农产品电商应注重农产品的保鲜、储存和包装,并运送及时、服务优质的快递服务商。

乡村振兴战略背景下,农产品电商纷纷入驻京东、淘宝、天猫、拼多多、抖音等各大电商平台,促使我国农产品网络营销蓬勃发展。该研究仅以京东电商平台为案例,消费者在线评论数据覆盖面存在一定的局限性,未来研究可进一步拓展数据获取范围,如增加淘宝、天猫、抖音等电商平台的数据,使研究结果和结论更具代表性。

参考文献

[1] 王秀清,孙云峰. 我国食品市场上的质量信号问题[J]. 中国农村经济,2002(5):27-32.

[2] 李爱国,邓召惠,毛冰洁. 在线负面评论对体验型产品销量的影响:基于商家回复视角[J]. 商业研究,2016(7):138-144.

[3] DE PELSMACKER P,VAN TILBURG S,HOLTHOF C. Digital marketing strategies,online reviews and hotel performance[J]. International journal of hospitality management,2018,72:47-55.

[4] 游浚,张晓瑜,杨丰瑞. 在线评论有用性的影响因素研究:基于商品类型的调节效应[J]. 软科学,2019,33(5):140-144.

[5] 张耕,刘震宇. 在线消费者感知不确定性及其影响因素的作用[J]. 南开管理评论,2010,13(5):99-106.

[6] 殷国鹏. 消费者认为怎样的在线评论更有用? ——社会性因素的影响效应[J]. 管理世界,2012(12):115-124.

[7] WEATHERS D,SHARMA S,WOOD S L. Effects of online communication practices on consumer perceptions of performance uncertainty for search and experience goods[J]. Journal of retailing,2007,83(4):393-401.

[8] 宋鹏,郭勤勤. 异质产品的在线评论特征对产品销量的影响[J]. 山西大学学报(哲学社会科学版),2019,42(4):105-112.

[9] 张艳辉,李宗伟. 在线评论有用性的影响因素研究:基于产品类型的调节效应[J]. 管理评论,2016,28(10):123-132.

[10] 周小梅,范鸿飞. 区域声誉可激励农产品质量安全水平提升吗? ——基于浙江省丽水区域品牌案例的研究[J]. 农业经济问题,2017,38(4):85-92,112.

[11] FINK L,ROSENFELD L,RAVID G. Longer online reviews are not necessarily better[J]. International journal of information management,2018,39:30-37.

[12] LIU Y M,DU R. The effects of image-based online reviews on customers' perception across product type and gender[J]. Journal of global information management,2019,27(3):139-158.