doi:10.6041/j.issn.1000-1298.2021.S0.063

基于 DSLML 的鸡蛋消费在线评论情感分析

包乾辉　李佳利　石淑珍　戴　引　刘　雪
(中国农业大学信息与电气工程学院, 北京 100083)

摘要: 随着信息技术、包装和物流技术的快速发展,包括农产品在内的电商产品范围和规模越来越大,同时,网上购物在线评论数据也呈指数级增加。在线评论成为关注的热点。以京东电商平台为例,挖掘鸡蛋消费在线评论文本,深入分析消费者鸡蛋消费情感倾向,提出了一种结合机器学习的领域情感词典(Domain sentimental lexicon with machine learning, DSLML)分类方法,该方法通过情感倾向逐点互信息(Semantic orientation pointwise mutual information,SO-PMI)方法构建领域情感词典,并选择机器学习模型作为情感分类器,实现对鸡蛋在线评论的情感倾向分类;然后构建 LDA 主题模型挖掘出鸡蛋评论中的正、负向主题。实验结果表明,与单独的机器学习模型和领域情感词典(Domain sentimental lexicon,DSL)相比,DSLML 分类模型在文本情感倾向分类中的各指标均有所提升;主题挖掘结果表明,消费者最为关心的是鸡蛋品质和包装。本研究结论可以为鸡蛋电商经营者有针对性提升经营策略、提高服务质量提供数据支持和理论支撑。

关键词: 情感分析; 鸡蛋; 在线评论; 情感倾向分类; 领域情感词典; LDA 主题模型
中图分类号: O433.4　　**文献标识码**: A　　**文章编号**: 1000-1298(2021)S0-0496-08

Sentimental Analysis of Online Reviews of Egg Consumption Based on DSLML

BAO Qianhui　LI Jiali　SHI Shuzhen　DAI Yin　LIU Xue
(*College of Information and Electric Engineering*, *China Agricultural University*, *Beijing* 100083, *China*)

Abstract: With the rapid development of information technology, packaging and logistics technology, the range and scale of E-commerce products, including agricultural products, are getting larger and larger. At the same time, the online shopping review data has grown exponentially. The online reviews has become a hotspot. Taking JD's E-commerce platform as an example, the online reviews were mined out and the sentimental tendency of consumers about eggs consumption was analyzed deeply. The main research contents included proposing a domain sentimental lexicon with machine learning (DSLML) classification method. The semantic orientation pointwise mutual information (SO-PMI) method was used to construct the domain sentimental lexicon, and then a machine learning model was selected as the classifier to achieve the classification of sentimental orientation of online egg reviews. Then the LDA topic model was constructed to mine out the positive and negative topics in egg reviews. The experimental results showed that the DSLML classification model was improved in each indicator of text sentimental tendency classification, compared with machine learning models and domain sentimental lexicon (DSL) alone. From the results of the theme mining, the quality of eggs and the packaging of goods were the two aspects that consumers mostly concerned about. The conclusion of this research can provide data support and theoretical support for egg E-commerce operators to improve business strategies and service quality.

Key words: sentiment analysis; egg; online reviews; sentimental classification of tendency; domain sentimental lexicon; LDA theme model

收稿日期: 2021-07-16　修回日期: 2021-09-11
基金项目: 现代农业产业技术体系北京市家禽创新团队建设项目(2021)
作者简介: 包乾辉(1997—),男,硕士生,主要从事文本情感分析研究,E-mail: S20203081479@cau.edu.cn
通信作者: 刘雪(1968—),女,副教授,博士,主要从事农业信息管理、物流与供应链管理研究,E-mail: liusnow@cau.edu.cn

0　引言

随着信息技术、包装和物流技术的快速发展，包括鸡蛋在内的鲜活农产品电商范围和规模越来越大。在为消费者提供便捷购物通道的同时，电商平台也提供了信息共享渠道，网上购物在线评论数据也呈指数级增加。作为网络购物的衍生物，在线评论是消费者对产品和服务的直接反馈[1]。据调查[2]，90% 以上的用户在消费前会参考相关的在线评价信息，之后又通过给出自己的评价为后面的用户提供参考。要在激烈的竞争中实现营业额的增长，企业不仅需要了解行业、竞争对手及产品，更需要了解用户的需求和关注点[3]。因此，从评论中凝练出用户对产品的情感倾向和需求已成为热点。而文本情感分析技术是最常用的手段[4]。

文本情感分析（Sentiment analysis，SA）又称观点挖掘（Opinion mining，OM）[5]，是指利用额外文本资源（如词表、基于情感的词典、复杂的词典和词汇本体），采用自然语言处理（Natural language processing，NLP）技术从特定领域的非结构化评论文本中获取评论者的主观性、观点、情绪以及极性提取、分析、处理、归纳的过程[6]。目前，文本情感分析技术已经广泛应用到在线评论研究中[7-16]。研究方法主要是基于情感词典或机器学习的文本情感倾向分类，并通过构建主题模型挖掘出用户对产品的细粒度情感。其中，基于情感词典的分类方法耗时耗力，且过度依赖于情感词库的覆盖程度；基于机器学习的情感分类模型则需要对数据进行人工标注，且在训练阶段需要高质量的特征构造和选取。

因此，本文提出一种结合机器学习的领域情感词典（Domain sentimental lexicon with machine learning，DSLML）分类方法，通过情感倾向逐点互信息（Semantic orientation pointwise mutual information，SO－PMI）方法进行情感词扩展来构建领域情感词典，增强情感词的覆盖程度；再选择机器学习模型作为分类器，进一步理解文本情感语义，从而达到效果更好的情感倾向分类；最后构建隐狄利克雷分布（Latent Dirichlet allocation，LDA）模型，挖掘出正、负评论文本的主题信息，实现对鸡蛋消费者喜恶点与关注点的有效提取。

1　研究方法与数据

1.1　逻辑框架

本文的逻辑框架如图 1 所示。首先，进行数据获取和预处理，采用 Python 爬虫技术获取京东电商平台的鸡蛋评论数据，并进行文本预处理；其次，构建基于 SO－PMI 的领域情感词典；再次，选择最佳的机器学习模型作为分类器；最后构建 LDA 主题挖掘模型，实现对鸡蛋评论数据的主题挖掘。

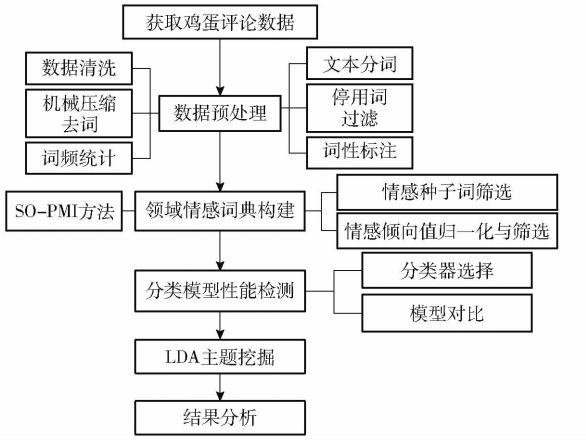

图 1　逻辑框架

Fig. 1　Logical framework

1.2　研究方法

1.2.1　情感倾向分类

情感倾向分类是文本情感分析的一项基本任务，用于判断文本的情感极性，以期了解消费者对某个事件或某件产品的态度[17]。研究者通常将文本的主观情感归为两类（正向和负向）或 3 类（正向、负向和中性）[18]。目前，针对文本情感倾向分类的方法主要有基于情感词典的方法和基于机器学习的方法[19]。

基于情感词典的方法是一种属于无监督学习的方法，主要思想是通过情感词典对文本中的情感词进行匹配，查找文本中出现的积极或消极情感词汇，并结合特定语法规则计算出各文本的情感倾向值，从而达到对文本情感进行判断的效果[20]。文献[21]构建了中文文本情感词语本体库，为基于情感词典的中文文本情感倾向分类奠定了基础；文献[22]通过使用句法分析和情感词典，提出了基于规则的文本情感倾向分类方法；文献[23]通过抽取情感短语，然后对抽取出的短语计算平均语义得分来判断句子的总体情感倾向。然而基于情感词典方法的弊端随着深入研究逐渐暴露：构建情感词典时所采用的情感词典质量会直接影响分类结果；词汇的具体语义往往随着应用领域的变化而变化，这导致情感词典的通用性大大降低。随着机器学习技术的快速发展，通过文本特征提取和训练模型对文本进行情感倾向分类，弥补了基于情感词典方法的许多不足，逐渐成为文本情感倾向分类任务的主流方案[24]。

基于机器学习的方法主要使用了经典的机器学

习算法,如:支持向量机、随机森林、朴素贝叶斯等来完成对文本的情感倾向分类[25]。研究者通常使用大量标注好的数据来训练机器学习模型,使模型学习各特征之间的关系,最终将该模型应用到实际的情感倾向分类场景中[26]。文献[27]分别使用朴素贝叶斯模型和最大熵模型在 IMDB 数据集上进行情感倾向分类实验,为此后基于机器学习的文本情感倾向分类研究奠定了基础;文献[28]采用集成模型对 8 个主题的微博舆情数据进行情感倾向分类;文献[29]通过对特征词进行加权,并用 SVM 算法进行分类。基于机器学习的方法通过特征工程和分类模型可进一步理解文本情感语义,从而提升分类性能[30]。

通常情况下,基于机器学习的分类方法准确度要高于情感词典,但分类稳定性不如情感词典,且需事先对海量数据进行人工标注。而情感词典不仅不需要数据标注,使用成本低,且容易进行词典扩展,但对于文本语义特征的学习能力较弱。因此,本文结合情感词典以及机器学习分类模型各自的优点,提出了一种结合机器学习的领域情感词典(DSLML)分类方法,并用该方法对鸡蛋评论文本进行情感倾向分类。

1.2.2 LDA 主题挖掘

主题模型是一种用于在一系列文档中发现隐含主题的语言概率模型[31],属于无监督的文本聚类算法[32]。LDA 主题模型是目前被普遍使用的主题模型,其本质是一个“文档(d)-主题(z)-词(t)”的三层贝叶斯生成模型,文本集中每条文本所提及的主题可以按照一定的概率分布形式给出[33]。在识别大规模文档集或语料库中潜藏的主题信息具有一定的优势[34]。

LDA 模型的基本思想是将文档描述为主题概率分布并进一步将主题描述为词项概率分布。模型的原理示意图如图 2 所示。

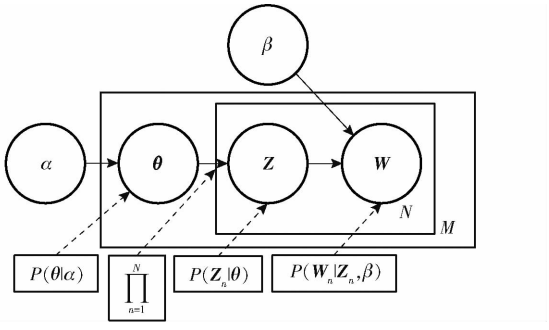

图 2 LDA 模型原理示意图

Fig. 2 LDA model schematic

假设任何文本都可以表示成一系列主题的混合分布,记为 $P(\boldsymbol{Z})$,计算公式为

$$P(D|\alpha,\beta)=\sum_{k\in K}P(\boldsymbol{W}_n|K_k)P(K_k|M_m) \quad (1)$$

其中,$P(\boldsymbol{W}_n|K_k)$表示在主题 k 条件下出现单词 $\boldsymbol{W}_n$ 的概率,$P(K_k|M_m)$则表示在文档 m 条件下出现主题 k 的概率。对于一个含有 M 个文本的语料库$D=\{d_1,d_2,\cdots,d_m\}$,d_m 表示语料库中的第 m 个文本,D 中的一条文本的生成过程如下[35]:

(1)选择主题向量 $\boldsymbol{\theta}$ 服从参数为 α 的狄利克雷分布,$\theta_i\sim\mathrm{Dir}(\alpha)$,其中 $i\in\{1,2,\cdots,m\}$,表示语料库 D 中评论个数。

(2)选择 ϕ 服从参数为 β 的狄利克雷分布,$\phi_k\sim\mathrm{Dir}(\beta)$,其中 $k\in\{1,2,\cdots,m\}$。

(3)对于文本中的每个词语 $\boldsymbol{W}_n$:生成主题 Z_{ij} 服从参数为 ϕ_i 的狄利克雷分布,$Z_{i,j}\sim\mathrm{Multi}(\phi_i)$,其中 ϕ_i 是多项分布的参数,生成词语 W_{ij} 服从参数为 $\phi_{z_{i,j}}$ 的狄利克雷分布,$Z_{i,j}\sim\mathrm{Multi}(\phi_{z_{i,j}})$。

LDA 模型目前已被许多研究者应用于各领域,但将该方法应用于农产品领域还较为少见,因此,本文在对数据集进行情感倾向分类后,构建 LDA 主题模型,挖掘出消费者对电商平台鸡蛋商品的喜恶点以及关注点。

1.3 数据

1.3.1 获取方法

用户评论语料可通过现有的爬虫工具或编写的爬虫程序从相关网站爬取[36]。现有的爬虫工具可操作性和适用性较差,因而爬虫技术的使用在网络数据获取任务中较为常见。本文通过设计一个垂直式的 Python 爬虫程序来获取京东电商平台的鸡蛋商品评论数据,并将其储存到数据库中。

1.3.2 预处理方法

在获取到的数据中存在着大量价值低甚至毫无意义的评论文本,直接使用文本数据必然会对分析造成很大的影响。同时,由于文本属于非结构化数据,计算机无法直接进行处理,需将其转化为计算机可直接处理的结构化数据,以便快速对文本进行相关方向的信息挖掘[37],故需要对文本数据进行分词处理,以方便之后的词向量转换。因此,在使用这些文本评论之前需进行文本预处理,包括:数据清洗、机械压缩去词、文本分词、停用词过滤、词性标注以及词频统计。

2 领域情感词典构建

构建领域情感词典的步骤为:从通用情感词典中筛选出正、负向情感词来确定所构词典的情感种子词;采用 SO - PMI 方法生成所有词语的情感值,并从中筛选一部分作为情感词;对各单词对

应情感值进行归一化处理，并筛选出部分情感词为中性词。

2.1　情感种子词筛选

本文从苏州大学电商领域情感词典、大连理工情感词典以及 HowNet 情感词典中选取积极和消极情感词，分别构建成正、负情感词集。将评论语料库中的词按词频从高到低进行挑选，从正、负情感词集中选出相同数量的情感词作为正、负向情感种子词，将数量定在 30～100 之间，步长为 10。通过不断测试发现，当数量为 60 时，情感词典在后面实验结果中的各指标分数最高，分类效果最佳。因此，正、负向情感种子词的数量分别为 60 个。

2.2　情感值生成与情感词扩展

在筛选出正、负向情感种子词后，此时情感词典中的情感词仅为基础情感词，仍缺乏大量涉及鸡蛋领域的情感词，因此需要进行情感词的扩展。本文采用 SO－PMI 方法对评论语料库中各词语的情感值进行计算，并从中分别筛选出最终的正、负情感词。

SO－PMI 方法为：通过计算各个单词的逐点互信息（Pointwise mutual information，PMI）指标来评估词语的情感倾向（Semantic orientation，SO）[38]。V_{PMI} 指标用来衡量两个词语间的独立性，计算公式为

$$V_{PMI}(\boldsymbol{W}_1,\boldsymbol{W}_2)=\lg\frac{P(\boldsymbol{W}_1,\boldsymbol{W}_2)}{P(\boldsymbol{W}_1)P(\boldsymbol{W}_2)} \tag{2}$$

式中　$P(\boldsymbol{W}_1)$、$P(\boldsymbol{W}_2)$——词在文本中单独出现的概率

$P(\boldsymbol{W}_1,\boldsymbol{W}_2)$——两个词在文本中同时出现的概率

$V_{PMI}(\boldsymbol{W}_1,\boldsymbol{W}_2)$——用来衡量两个词的共现程度，值越大表明两个词的关系越紧密，情感倾向越一致

用 $P(\boldsymbol{W}_i)$ 和 $N(\boldsymbol{W}_i)$ 分别表示第 i 个积极或消极情感词，则评论中某词 W 的情感倾向值计算公式为

$$V_{SO\text{-}PMI(W)}=\sum_{i=1}\left[V_{PMI}(\boldsymbol{W},P(\boldsymbol{W}_i))-V_{PMI}(\boldsymbol{W},N(\boldsymbol{W}_i))\right] \tag{3}$$

$V_{SO\text{-}PMI(W)}$ 会出现 3 种情况，即大于、小于和等于 0，分别对应 $\boldsymbol{W}$ 具有积极情感倾向、$\boldsymbol{W}$ 具有消极情感倾向和 $\boldsymbol{W}$ 不具任何情感倾向。

在使用 SO－PMI 方法生成了语料库中各分词的情感值后，分别将 SO－PMI 值（情感倾向值）大于 0 和小于 0 的情感词划分为正向和负向情感词。同时，也存在以下问题：同类情感倾向值中最大值和最小值之间相差较大；存在大量情感倾向值接近 0 的情感词，这些词大多为中性词，但 SO－PMI 方法也将其归为情感词。

因此，需要对各个分词的情感倾向值进行归一化处理，从而消除奇异样本数据导致的不良影响；同时设定一个区间，将情感倾各值位于该区间的情感词划为中性词，以便后续工作的进行。

2.3　情感倾向值归一化与筛选

本次实验使用线性归一化（Min－Max 标准化）对各个词的情感倾向值进行线性变换。使最终得到的数值可以映射到［0，1］区间内。

进行归一化处理后，需要确定一个阈值 t，令情感倾向值位于［$-t$，t］内的情感词为中性词。本文将 t 设为 0～0.2，步长为 0.01，通过实验发现，当 t 为 0.07 时各项指标的结果最好，故将 t 值设为 0.07，则情感值位于［－0.07，0.07］内的情感词为中性词，不具备情感倾向性，最终得到情感词 8 704 个，其中正向情感词为 5 313 个，负向情感词为 3 191 个。将这些词存入情感词典中，最终完成情感词典的构建，部分情感词及其对应情感倾向值如表 1 所示。

表 1　鸡蛋领域情感词典（部分）

Tab.1　Part of eggs domain sentimental lexicon

情感词	情感倾向值	情感词	情感倾向值
一流	1	臭	－1
服务周到	0.86	退货	－0.97
很快	0.85	发臭	－0.966
正品	0.847	最差	－0.937
大爱	0.801	垃圾	－0.901

3　实验与结果分析

3.1　实验数据

3.1.1　数据获取

本次实验通过设计一个垂直式的 Python 爬虫程序来获取京东电商平台鸡蛋商品的评论数据，具体步骤如下：

（1）建立待抓取的网页 URL 列表，使用 requests 模块中模拟用户登录状态，从而对网页进行访问。

（2）通过 post 请求解析商品预览页面，设置循环机制访问各个预览页面。

（3）利用正则表达式从爬取的页面中提取所需要的信息并存储在本地中。

为避免京东电商平台的反爬虫机制，本次爬虫程序设计采取以下两种措施来应对：

（1）设置爬虫时间间隔。正常用户访问网站的频率不会过快，因此使用 time 模块设置每次爬取的时间间隔，这样虽然爬取速度下降，但保证了程序顺

利运行。

(2)设置多个 IP 地址。一个 IP 地址频繁访问可能会引起网站的反爬虫机制,因此本文设置多个 IP 地址规避网站的检测。

本次实验从京东电商平台爬取了 62 342 条鸡蛋评论数据,包含“品牌”和“评论文本”两个特征,并将获取的数据保存至数据库中。

3.1.2 数据预处理

需对评论文本进行预处理操作,具体步骤如下:

(1)本文使用正则表达式来识别并剔除数据中的异常内容,包括符号、数字、字母、空格等。

(2)去除评论文本中无实际意义的重复词汇,从而降低词频统计时的噪声。

(3)将文本按照一定规则分割成若干个词,将文本转化为结构化数据,便于之后的词频统计和建模。

(4)加载哈工大停用词表,剔除无意义词汇。

(5)对每个词的词性进行标注。

(6)统计各个词的出现频次。

通过对评论文本的预处理,最终得到 45 403 条数据。

3.2 评价指标

为验证本文所提模型的情感倾向分类性能,选取准确率、召回率和 F1 值作为评价标准,这些指标的值越大,表明模型的分类效果越好。

3.3 分类器选取

构建完情感词典后,需要选取合适的机器学习模型作为最终的分类器。在此之前,从预处理完的数据集中随机抽取 3 000 条评论数据,按照 4∶1的比例来划分为训练集和测试集,其中,情感倾向包括正向、负向和中性 3 类,由人工进行标注。最后使用不同的机器学习模型分别与构建好的领域情感词典相结合,对测试集进行情感倾向分类,实验 5 次,每次随机将训练集和测试集按照 4∶1的比例进行分配。最终选取 5 次实验各指标的平均值作为最终结果。

由表 2 中各项评价指标可知,极限梯度提升模型(eXtreme gradient boosting,XGBoost)与领域情感词典结合使用时的分类效果最佳,故将其选为分类器,则 XGBoost 与领域情感词典相结合构成了本文所提出的模型 DSLML。

3.4 结果对比

选取好分类器后,从数据集中随机抽取出 10 000 条数据进行标注,分为 5 组,每组 2 000 条评论,从中选取 1 500 条进行训练,其余用来测试不同分类模型的性能。对DSLML分类模型、领域情感词典(Domain sentimental lexicon,DSL)以及 5 个在情感倾向分类任务中表现效果较优的传统机器学习模型在这 5 组数据集上进行性能对比,各个指标的最终结果取该模型在 5 组数据上的平均值。对比结果如表 3 所示。

表 2 各机器学习模型与领域情感词典结合使用的结果对比

Tab. 2 Comparison of results of each machine learning model combined with domain sentiment lexicon

模型	准确率	召回率	F1 值
MultinomialNB + 领域情感词典	0.905	0.640	0.631
SVC(linear) + 领域情感词典	0.901	0.717	0.715
Random Forest + 领域情感词典	0.885	0.652	0.657
AdaBoost + 领域情感词典	0.890	0.678	0.682
XGBoost + 领域情感词典	0.909	0.715	0.724

表 3 各模型实验结果对比

Tab. 3 Comparison of experimental results of each model

模型	准确率	召回率	F1 值
MultinomialNB	0.878	0.626	0.621
SVC(linear)	0.880	0.671	0.652
Random Forest	0.867	0.638	0.636
AdaBoost	0.855	0.630	0.632
XGBoost	0.867	0.642	0.665
DSL	0.845	0.629	0.645
DSLML	0.912	0.717	0.715

从对比结果来看,不论是单独的领域情感词典还是传统的机器学习模型,在各个指标的表现都不如本文所提出的分类模型效果好。通过将 DSL 与 DSLML 的结果对比可知,DSLML 在准确率、召回率、F1 值上比 DSL 分别高 6.7%、8.8% 和 7%,说明运用特征工程可以进一步地理解文本情感语义,从而提升文本情感分类的性能。从 DSLML 与单一机器学习模型的对比来看,DSLML 虽然在各项指标均表现最优,但二者差距并不大,说明机器学习模型在情感倾向分类方面的效果确实很理想,但由于稳定性较差、特征词难题取等问题,表现不如与情感词典相结合的分类模型。

3.5 情感分类的主题挖掘

在使用 DSLML 分类模型对所有评论进行情感倾向分类后,为探究消费者对于鸡蛋商品的喜恶点,分别对正、负向评论文本使用 LDA 模型进行主题挖掘。在构建 LDA 模型的过程中需要使用“文档-主题”分布的超参数 α 和“主题-词语”分布的超参数 β,以及迭代次数 G 和最优主题数 K。通过不断调参发现,在 $\alpha = 50/K$、$\beta = 0.01$、$G = 50$ 时,LDA 模型在评论数据集上有较好的主题识别效果。以正向情感评论集的主题挖掘为例(图 3),当代表主题的圆圈

相隔较远时，说明主题的凝聚效果良好。各主题的关键词会在右侧以柱状列表的形式显示出来，红色柱条的长短表示在某一主题中对应关键词的词频大小，总长表示该词在整个文档中出现的频率。

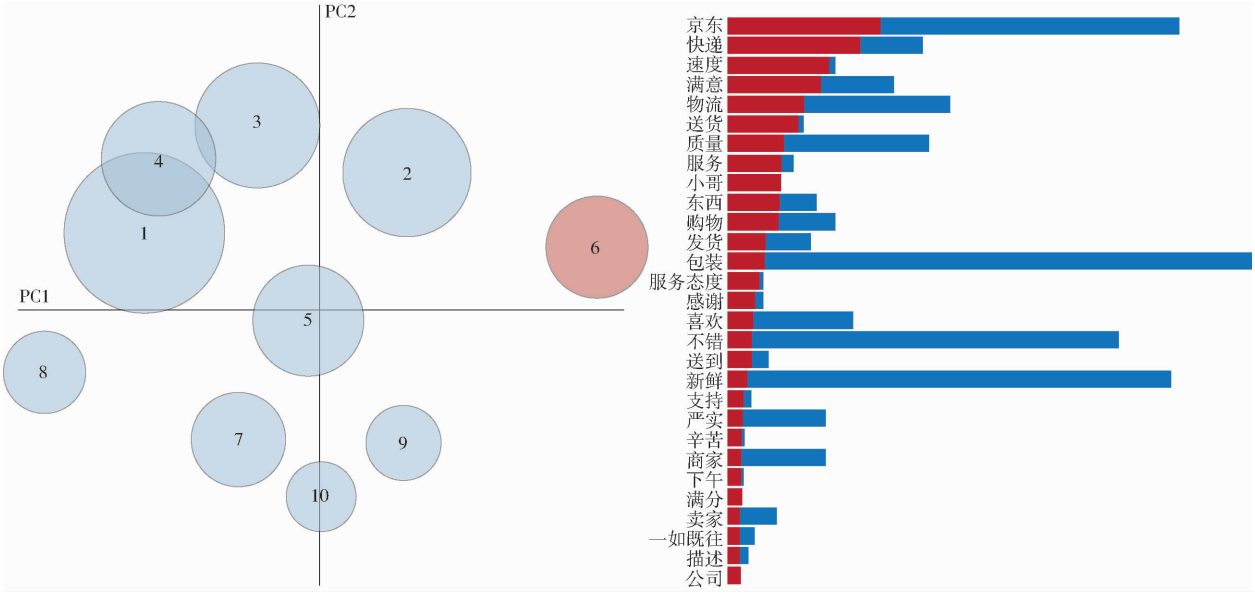

图 3　LDA 模型示意图

Fig. 3　Schematic of LDA model

根据主题模型的主题分布情况，最终将 LDA 模型的主题数量 K 设置为 10。将每个主题中的词语提取前 20 ~ 30 个，最终将主题归纳为 7 个，分别是：商品包装、食用、新鲜度、物流服务、鸡蛋外观、商家服务和性价比。过滤没有准确含义的词汇，选出各个主题具有代表性的前 8 个关键词。各主题对应的核心关键词以及涉及该主题的评论占同类情感倾向的评论百分比如表 4 所示。

表 4　鸡蛋正、负向评论主题挖掘结果

Tab. 4　Topic mining results of positive and negative comments on eggs　%

主题	情感倾向	核心关键词	同类情感倾向文本中百分比
新鲜度	正向	新鲜、蛋黄、浓稠、清晰、透亮、优质、浓郁、蛋白	41. 54
	负向	腥、长毛、发霉、异味、坏、臭、烂、难闻	45. 71
商品包装	正向	好看、细腻、紧实、轻、纸箱、牢固、填充物、环保	42. 46
	负向	米糠、受损、简陋、清理、陈旧、满屋子、脏、湿	28. 04
食用	正向	香、卫生、好吃、水煮蛋、嫩、营养、早餐、丰富	42. 29
	负向	难吃、人造、变质、咸、散、剥不开、拉肚子、浑浊	27. 10
鸡蛋外观	正向	饱满、大、洁净、完好、匀称、鹅蛋、光滑、圆润	32. 60
	负向	破蛋、裂纹、黑斑、小、袖珍、浅、鸽子蛋、爆裂	41. 87
物流服务	正向	迅速、轻放、到位、送货上门、方便快捷、快、包邮、时效	34. 38
	负向	拒收、野蛮、高温、耽误、慢、砸、远、压坏	16. 03
商家服务	正向	礼品、负责、良心、仔细、保证、信誉、热情、专业	31. 33
	负向	退款、恶劣、赔付、欺诈、勉强、一拖再拖、搭理、投诉	21. 91
性价比	正向	折扣、一块钱、划算、低廉、物超所值、实惠、经济、活动	25. 52
	负向	坑、差太多、堪忧、教训、不值、慎重、差、贪便宜	33. 78

通过表 4 可以看到，消费者对于京东电商平台鸡蛋商品的正面评价主要集中在商品的包装、食用和鸡蛋新鲜度等方面，与这 3 个主题相关的评论在所有正向评论中的百分比超过 40%。在商品的包装方面，消费者对于具备美观、牢固、环保、轻便等包装特点的商品更为青睐；在食用方面，消费者主要关注鸡蛋的食用口感和营养情况，对香嫩、富含营养的鸡蛋较为钟爱；在正向评论中，涉及性价比这类主题的评论数量占比最少，说明与商品性价比相比，商品的优良品质以及愉快的购物体验更能让消费者产生好感。

对于鸡蛋商品的负面评价主要体现在鸡蛋的新鲜度和外观上。腥臭、发霉的鸡蛋最容易引起消费者的不满情绪；其次是外观有黑斑、出现破裂、个头

较小的鸡蛋;消费者很少会因为商家服务与物流服务等原因而给出负面评论。

综合正、负评论中各主题的占比情况来看,消费者对于京东电商平台鸡蛋商品的关注点主要集中于鸡蛋的品质以及商品的包装上,在鸡蛋品质方面,消费者最关心的是新鲜度,可以看到,在正、负面评论的众多主题中,新鲜度一直是评论数量占比非常大的一个主题;其次是食用口感和鸡蛋的外观,而商品性价比以及服务体验等方面相对上述几个主题而言并没有过多引起消费者注意。

从不同主题在正、负向情感评论中的占比情况来看,京东电商平台鸡蛋商品的优势主要集中在商品的包装、物流服务以及鸡蛋的食用上,这 3 个主题在正向评论中占比较大,在负向评论中占比较小;与之相反的是新鲜度、外观及性价比,这 3 个主题在正向评论中的占比要小于在负向评论中的占比,因此,京东电商平台的鸡蛋经营者应对鸡蛋的新鲜度以及外观做好质量把控,同时可以适当开展一些促销活动。

4 结束语

本文提出了一种将机器学习模型和领域情感词典相结合的情感倾向分类方法,对京东电商平台的鸡蛋评论文本进行情感倾向分类和主题挖掘。结果表明,与传统的机器学习模型和领域情感词典相比,本文所提出的 DSLML 模型的分类性能有了较大提升;从主题挖掘结果来看,鸡蛋商品的正、负面评价主要集中在鸡蛋新鲜度和包装两个方面,鸡蛋经营者应对鸡蛋的新鲜度以及外观做好质量把控,同时可以适当开展一些促销活动。

参 考 文 献

[1] 尹丽春,王悦. 基于在线评论的图书消费者满意度影响因素与作用机理[J]. 图书情报工作, 2019, 63(22): 106 - 117.
YIN Lichun, WANG Yue. Influencing factors and mechanism of book consumer satisfaction based on online comments[J]. Library and Information Service, 2019, 63(22): 106 - 117. (in Chinese)

[2] 韩发丽. 在线评论对消费者购买决策的影响[J]. 现代企业, 2018(5): 51 - 52.

[3] 朱晓红,陈寒松,张腾. 知识经济背景下平台型企业构建过程中的迭代创新模式:基于动态能力视角的双案例研究[J]. 管理世界, 2019(3): 142 - 156.

[4] 李然,林政,林海伦,等. 文本情绪分析综述[J]. 计算机研究与发展, 2018, 55(1): 30 - 52.
LI Ran, LIN Zheng, LIN Hailun, et al. Text emotion analysis: a survey[J]. Journal of Computer Research and Development, 2018, 55(1): 30 - 52. (in Chinese)

[5] HEMMATIAN F, SOHRABI M K. A survey on classification techniques for opinion mining and sentiment analysis[J]. Artificial Intelligence Review, 2019, 52(3): 1495 - 1545.

[6] LIU Z, WANG Y, CHEN T. Audio feature extraction and analysis for scene segmentation and classification[J]. The Journal of VLSI Signal Processing, 1998, 20(1 - 2):61 - 79.

[7] 左韶泽. 对网购评论数据的文本挖掘——以某款式冰箱的评论为例[D]. 秦皇岛:燕山大学, 2019.
ZUO Shaoze. Text mining for online purchase of comment data: take the review of a refrigerator as an example[D]. Qinhuangdao: Yanshan University, 2019. (in Chinese)

[8] 高思梦. 基于机器学习的老年手机用户中文评论情感分析——以京东数据为例[D]. 兰州:兰州财经大学, 2019.
GAO Simeng. Sentiment analysis for elderly mobile phone users based on machine learning[D]. Lanzhou: Lanzhou University of Finance and Economics, 2019. (in Chinese)

[9] 黄文波. 电商平台用户评论的分析及应用[D]. 广州:华南理工大学, 2020.
HUANG Wenbo. The analysis and application of user comments base on E-commemerce[D]. Guangzhou: South China University of Technology, 2020. (in Chinese)

[10] DASGUPTA S, SENGUPTA K. Analyzing consumer reviews with text mining approach[J]. Paradigm, 2016, 20(1): 56 - 68.

[11] 冯志骥. 基于文本挖掘的新能源汽车市场研究[D]. 石家庄:河北经贸大学, 2020.
FENG Zhiji. Research on new energy vehicle market based on text mining[D]. Shijiazhuang: Hebei University of Economics and Business, 2020. (in Chinese)

[12] 曹紫琰,冯敏萱,毛雪芬,等. 细颗粒度汽车评论语料库的构建和分析[J]. 中文信息学报, 2020, 34(9): 28 - 35.
CAO Ziyan, FENG Minxuan, MAO Xuefen, et al. Construction and analysis of fine-grained car review corpus[J]. Journal of Chinese Information Processing, 2020, 34(9): 28 - 35. (in Chinese)

[13] 潘艳茜,姚天昉. 微博汽车领域中用户观点句识别方法的研究[J]. 中文信息学报, 2014, 28(5): 148 - 154.
PAN Yanxi, YAO Tianfang. Recognition of microblog customer opinion sentences in automobiles domain[J]. Journal of Chinese Information Processing, 2014, 28(5): 148 - 154. (in Chinese)

[14] 李宏媛,陶然. 服装电商评论情感分析研究[J]. 智能计算机与应用, 2017, 7(1): 27 - 30, 34.
LI Hongyuan, TAO Ran. Sentiment analysis on electronic commerce reviews in the field of clothing[J]. Intelligent Computer and Applications, 2017, 7(1): 27 - 30, 34. (in Chinese)

[15] 赵志杰,刘岩,张艳荣,等. 基于 Lasso - LDA 的酒店用户偏好模型[J]. 计算机应用与软件, 2021, 38(2): 19 - 26.
ZHAO Zhijie, LIU Yan, ZHANG Yanrong, et al. Hotel user preference model based on Lasso - LDA[J]. Computer Applications and Software, 2021, 38(2): 19 - 26. (in Chinese)

[16] 杨鑫,杨云帆,焦维,等. 基于领域词典的民宿评论情感分析[J]. 科学技术与工程, 2020, 20(7): 2794 - 2800.

YANG Xin, YANG Yunfan, JIAO Wei, et al. Sentiment analysis of homestay comments based on domain dictionary[J]. Science Technology and Engineering, 2020, 20(7): 2794 – 2800. (in Chinese)
[17] 李勇敢,周学广,孙艳,等. 中文微博情感分析研究与实现[J]. 软件学报, 2017, 28(12): 3183 – 3205.
LI Yonggan, ZHOU Xueguang, SUN Yan, et al. Research and implementation of Chinese microblog sentiment classification [J]. Journal of Software,2017, 28(12): 3183 – 3205. (in Chinese)
[18] 赵妍妍,秦兵,刘挺. 文本情感分析[J]. 软件学报, 2010, 21(8): 1834 – 1848.
ZHAO Yanyan, QIN Bing, LIU Ting. Sentiment analysis[J]. Journal of Software,2010, 21(8): 1834 – 1848. (in Chinese)
[19] PANDARACHALIL R, SENDHILKUMAR S, MAHALAKSHMI G S. Twitter sentiment analysis for large-scale data: an unsupervised approach[J]. Cognitive Computation, 2015, 7(2): 254 – 262.
[20] 洪巍,李敏. 文本情感分析方法研究综述[J]. 计算机工程与科学, 2019, 41(4): 750 – 757.
HONG Wei, LI Min. A review: text sentiment analysis methods[J]. Computer Engineering & Science,2019, 41(4): 750 – 757. (in Chinese)
[21] 徐琳宏,林鸿飞,赵晶. 情感语料库的构建和分析[J]. 中文信息学报, 2008, 22(1): 116 – 122.
XU Linhong, LIN Hongfei, ZHAO Jing. Construction and analysis of emotional corpus[J]. Journal of Chinese Information Processing,2008, 22(1): 116 – 122. (in Chinese)
[22] QIU G, HE X, ZHANG F, et al. DASA: dissatisfaction-oriented advertising based on sentiment analysis[J]. Expert Systems with Application, 2010, 37(9):6182 – 6191.
[23] PETER D T. Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews[C]//40th Annual Meeting of the Association for Computational Linguistics:(CD:CD – CNF – 0517), 2002:417 – 424.
[24] 刘磊. 基于生成式对抗网络与异质集成学习的文本情感分类研究[D]. 南京:南京邮电大学, 2020.
LIU Lei. Research on text sentiment classfication based on generative adversarial network and heterogenous ensemble learning [D]. Nanjing:Nanjing University of Posts and Telecommunications,2020. (in Chinese)
[25] 陈平平,耿笑冉,邹敏,等. 基于机器学习的文本情感倾向性分析[J]. 计算机与现代化, 2020(3): 77 – 81, 92.
CHEN Pingping, GENG Xiaoran, ZOU Min, et al. Analysis of text sentiment orientation based on machine learning[J]. Computer and Modernization,2020(3): 77 – 81, 92. (in Chinese)
[26] KHAN K, LEE L H, BAHARUDIN B. A review of machine learning algorithms for text-documents classification[J]. Journal of Advances in Information Technology, 2010, 1(1):4 – 20.
[27] PANG B, LEE L, VAITHYANATHAN S. Thumbs up? Sentiment classification using machine learning techniques[C]// Proceedings of the ACL – 02 Conference on Empirical Methods in Natural Language Processing-volume 10. Stroudsburg: Association for Computational Linguistics,2002:79 – 86.
[28] 杨莉,王敏,程宇. 基于 LDA 和 XGBoost 模型的环境公共服务微博情感分析[J]. 南京邮电大学学报(社会科学版), 2019, 21(6): 23 – 39.
YANG Li, WANG Min, CHENG Yu. Microblog sentiment analysis of Jiangsu environmental public service based on LDA and XGBoost models[J]. Journal of Nanjing University of Posts and Telecommunications(Social Science),2019, 21(6): 23 – 39. (in Chinese)
[29] 魏广顺,吴开超. 基于词向量模型的情感分析[J]. 计算机系统应用, 2017, 26(3): 182 – 186.
WEI Guangshun, WU Kaichao. Sentiment analysis based on word vector model[J]. Computer Systems & Applications,2017, 26(3): 182 – 186. (in Chinese)
[30] 陈龙,管子玉,何金红,等. 情感分类研究进展[J]. 计算机研究与发展, 2017, 54(6): 1150 – 1170.
CHEN Long, GUAN Ziyu, HE Jinhong, et al. A survey on sentiment classification[J]. Journal of Computer Research and Development,2017, 54(6): 1150 – 1170. (in Chinese)
[31] 邹晓辉,孙静. LDA 主题模型[J]. 智能计算机与应用, 2014, 4(5): 105 – 106.
ZOU Xiaohui, SUN Jing. Latent Dirichlet allocation topic model[J]. Intelligent Computer and Applications,2014, 4(5): 105 – 106. (in Chinese)
[32] MOMTAZI S. Unsupervised latent Dirichlet allocation for supervised question classification[J]. Information Processing & Management, 2018, 54(3): 380 – 393.
[33] 石晶,胡明,石鑫,等. 基于 LDA 模型的文本分割[J]. 计算机学报, 2008, 31(10): 1865 – 1873.
SHI Jing, HU Ming, SHI Xin, et al. Text segmentation based on model LDA[J]. Chinese Journal of Computers,2008, 31 (10): 1865 – 1873. (in Chinese)
[34] CHIRU C, REBEDEA T, CIOTEC S. Comparison between LSA – LDA – Lexical chains[C]//Barcelona, 2014.
[35] BLEI D M, NG A Y. JORDAN M I. Latent Dirichlet allocation[J]. Journal of Machine Learning Research, 2003, 3: 993 – 1022.
[36] 罗安然,林杉杉. 基于 Python 的网页数据爬虫设计与数据整理[J]. 电子测试, 2020(19): 94 – 95, 31.
LUO Anran, LIN Shanshan. Design of legal document crawler based on Python[J]. Electronic Test,2020(19): 94 – 95, 31. (in Chinese)
[37] 孔雪娜,孙红. 中文微博文本采集与预处理综述[J]. 软件导刊, 2017, 16(2): 186 – 189.
[38] 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(增刊 2): 78 – 82, 91.
JING Li, LI Manman, HE Tingting. Sentiment classification of network reviews combining extended dictionary and self-supervised learning[J]. Computer Science,2020, 47(Supp. 2): 78 – 82, 91. (in Chinese)