Operations Research and Fuzziology 运筹与模糊学, 2023, 13(2), 528-542
Published Online April 2023 in Hans. https://www.hanspub.org/journal/orf
https://doi.org/10.12677/orf.2023.132052

基于在线评论的决策支持框架

耿瑞娟[1*]，纪　颖[2]，张　洋[1]

[1]上海理工大学理学院，上海
[2]上海大学管理学院，上海

收稿日期：2023年2月24日；录用日期：2023年4月3日；发布日期：2023年4月10日

摘　要

在数字信息化时代，消费者在网上购物时越来越依赖在线评论，大数据爆炸式增长也导致消费者要花费大量的时间来阅读在线评论，筛选信息并做出决策。所以本研究旨在提出一个新的基于在线评论的决策支持框架，用于帮助消费者依据在线评论对可替代产品进行评估和选择。决策支持框架主要包括三个部分，1) 数据处理，用python抓取在线消费者评论进行数据清洗和预处理，提取出关键特征作为评价标准；2) 情感分析，利用朴素贝叶斯对在线评论进行情感分析，用积极意见的优势比作为模型的输出数据；3) 基准分析，利用RDEA模型来计算可替代产品的效率得分，根据效率得分进行排名。最后，对京东平台上爬取的15款笔记本电脑的在线评论进行实证分析，来验证所提出的决策支持框架有用性和适用性，并进行了对比分析，结果证明提出的方法更符合客观实际情况，并且步骤更简单，易于操作。

关键词

在线评论，情感分析，数据包络分析，最优化决策

Decision Support Framework Based on Online Review

Ruijuan Geng[1*], Ying Ji[2], Yang Zhang[1]

[1]College of Science, University of Shanghai for Science and Technology, Shanghai
[2]School of Management, Shanghai University, Shanghai

Received: Feb. 24th, 2023; accepted: Apr. 3rd, 2023; published: Apr. 10th, 2023

Abstract

In the era of digital information, consumers increasingly rely on online reviews when shopping

*通讯作者。

文章引用: 耿瑞娟, 纪颖, 张洋. 基于在线评论的决策支持框架[J]. 运筹与模糊学, 2023, 13(2): 528-542.
DOI: 10.12677/orf.2023.132052

online. The explosive growth of big data also leads to consumers spending a lot of time reading online reviews, screening information, and making decisions. Therefore, this study aims to propose a new decision support framework based on online reviews to help consumers evaluate and select alternative products based on online reviews. The decision support framework mainly includes three parts: 1) Data processing, which uses python to capture online consumer reviews for data cleaning and preprocessing, and extracts key features as evaluation criteria; 2) Emotional analysis, which uses naive Bayes to conduct emotional analysis on online reviews, and uses the advantage ratio of positive opinions as the output data of the model; 3) Benchmark analysis, which uses RDEA model to calculate the efficiency score of alternative products, and rank according to the efficiency score. Finally, an empirical analysis is conducted from the online comments of 15 laptops crawled on the JD platform to verify the usefulness and applicability of the proposed decision support framework, and a comparative analysis is conducted. The results show that the proposed method is more in line with the objective reality, and the steps are simpler and easier to operate.

Keywords

Online Review, Sentiment Analysis, Data Envelopment Analysis, Optimization Decision

1. 引言

随着信息技术的快速发展和 Web 2.0 技术的深入应用，加快了电子商务发展的速度。为了应对 COVID-19 的传播而实施的封锁隔离等措施改变了消费者的消费方式，人们的购物场所逐渐从线下实体店转向线上消费平台。但由于网络的虚拟性，导致产品的质量难以得到保证。消费者可以通过商家给出的产品信息和其他消费者的在线评论来了解和挑选产品，相对于商家，人们会更愿意相信其他消费者[1]。尽管现在有很多购物网站也提供了产品的量化标准，由于其主观性和可变性，还是建议消费者通过在线评论来做出购买决策。在线评论不仅可以帮助消费者了解产品的具体信息，降低网上购物的不确定性；同时也能帮助企业获得消费者的真实反馈，从而了解消费者的偏好和需求[2]。

根据中国互联网络信息中心(CNNIC)在京发布第 50 次《中国互联网络发展状况统计报告》显示，仅 2022 年上半年，全国网上零售额 6.3 万亿元，同比增长 3.1%。网上购物量的增大，会导致每天都会有新增的大量的评论，这会导致消费者和商家从在线评论中提取信息是非常困难的。所以提出一种基于在线评论的决策支持框架，就显得极具意义和价值。

目前对于通过在线评论的研究主要分为两个部分，第一个部分就是如何挖掘出在线评论的信息，第二个部分就是如何利用在线评论来帮助决策。从在线评论文本中挖掘信息的过程也就是情感分析，最早是由 Sanjiv 提出，从股票留言板中提取投资者情绪，用于评估管理公告、新闻稿等对投资者意见的影响[3]。早期比较传统的情感分析方法主要分为两类：基于字典的情感分析技术[4] [5] [6]，基于语料库的情感分析技术[7] [8] [9]。但是其本质上依赖于情感词典和判断规则的质量，结果的准确度受限于情感词典的覆盖率和准确率。随着在线评论数据量的增加和网络用词的出现，扩充语料库需要花费大量的时间和资源，情感分类时的灵活度不高，导致其无法跟上日益复杂的文本情感分类问题。这也就诞生了基于机器学习的情感分析方法，经典的分类模型包括支持向量机(SVM) [10] [11]，朴素贝叶斯(NB) [12]，最大熵模型[13]等。目前已有研究证明在情感分析任务上，基于机器学习的情感分析方法相对于基于词典的情感

分析方法，可以获得更高的准确率[14]，所以本文采取机器学习中的加权朴素贝叶斯对在线评论进行情感分析。第二部分是如何利用在线评论来帮助决策的研究，主要是通过分析在线评论，来帮助商家分析产品需要改进的地方，以及帮助消费者做出购买决策。常见的研究方法有 TOPSIS [15] [16]，VIKOR [17]，TODIM [18]和其他对产品进行排名的方法。但是这些方法的评价指标通常都是给定的，不符合客观实际情况，并且操作步骤比较复杂，消费者难以实现。DEA 是一种衡量具有多输入和多输出的决策单元效率的非参数方法，我们也可以根据效率对决策单元进行排名。

针对上述情况，本文提出一种基于在线评论的决策支持框架，该框架利用机器学习的方法从在线评论文本中挖掘有用信息，再通过基准分析来帮助消费者和商家做出正确的决策，主要贡献如下：

1) 利用机器学习中的加权朴素贝叶斯对在线评论进行情感分析，RDEA 模型考虑数据的不确定性对可替代产品进行基准分析；

2) 从在线评论中提取关键属性作为评价指标，从消费者的角度出发，更加客观和符合实际情况；

3) 利用 python 从京东(JD.COM)爬取 15 款笔记本电脑的 101,405 条在线评论进行数值实验，验证提出的决策框架的有效性和适用性。

本文的其余部分如下所示。第二节是预备知识；第三节是提出的决策支持框架；第四节是实证分析，以从京东上爬取的 15 款笔记本电脑的在线评论为例，验证提出的决策框架的有用性和适用性；第五节就是比较分析；第六节是结论，强调本文的主要贡献和研究的局限性以及未来的工作。

2. 预备知识

2.1. 朴素贝叶斯(NB)相关知识

朴素贝叶斯(NB)是一种机器学习的情感分析方法，属于监督学习。加权朴素贝叶斯是朴素贝斯的一个扩展，其中的属性具有不同的权重[19]。情感分析的过程为：首先对向量进行转化，再对分类器进行训练，为了防止模型过度拟合，按照 80%:20%的比例将数据集划分为训练集和测试集，最后进行预测分类。本文利用的是 pysenti 库(利用的是加权朴素贝叶斯)，结合句子结构给各情感词语的情感极性赋予权重，然后加权求和得到文本的情感极性得分。

朴素贝叶斯基于贝叶斯定理的概率分类技术，假设属性之间相互独立，互不干扰，利用带类别标签的训练集文本计算得到数据的先验概率，然后基于贝叶斯定理求出测试集文本属于某一类别的概率，公式如下所示：

$$P\left(C_i \mid X\right)=\frac{P\left(X \mid C_i\right)P\left(C_i\right)}{P\left(X\right)} \tag{1}$$

其中，$C_i\left(i=1,2,\cdots,m\right)$ 说明数据被分为 m 个类别，本文设为 3 个类别，分为积极、中性和消极；X 表示属性集合，有 $X=\{x_1,x_2,\cdots,x_n\}$，即说明总共有个 n 属性，对应到文本数据中为特征词的数量；若 $P\left(C_i \mid X\right)=\max\left\{P\left(C_j \mid X\right)\right\}\left(j=1,2,\cdots,m,i\neq j\right)$ 时，未知样本的类别就判断为 C_i 类别。$P\left(C_i\right)$ 表示训练集中各类别数据出现的概率，可通过计算 C_i 类别数据数量 N_i 占总样本数量 N 的比例来获得，$P\left(C_i\right)=N_i/N$。$P\left(X \mid C_i\right)$ 可通过训练文本中某类别下各属性出现的先验概率来计算得到，因朴素贝叶斯中假定不同属性之间相互独立，因此 $P\left(X \mid C_i\right)$ 可表示为：$P\left(X \mid C_i\right)=P\left(x_1x_2\cdots x_n \mid C_i\right)=\prod_{k=1}^{n}P\left(x_k \mid C_i\right)$。分母 $P\left(X\right)$ 对于 x_k 均为固定值，$P\left(X\right)=P\left(x_1x_2\cdots x_n\right)=\prod_{k=1}^{n}P\left(x_k\right)$。

由于上式具有相同的分母 $P\left(X\right)=P\left(x_1x_2\cdots x_n\right)=\prod_{k=1}^{n}P\left(x_k\right)$，可以将 $P\left(X\right)$ 可看成是标准化因子，所以将上式化简可以得到简化的朴素贝叶斯分类器。

令 $V_{nb}\left(C\right)$ 为基于朴素贝叶斯归类后的最大类别，定义如下：

$$V_{nb}(C)=\arg\max_{x_k} P(C_i)\prod_{k=1}^{n}P(x_k\mid C_i) \tag{2}$$

由于条件独立性假设在现实中很少成立，因此需要扩展朴素贝叶斯来放松条件独立性假设，其中一种方法就是属性的加权方式不同，由此产生的模型称为加权朴素贝叶斯。加权朴素贝叶斯(WNB)的定义如下：

$$V_{wnb}(C)=\arg\max_{x_k} P(C_i)\prod_{k=1}^{n}P(x_k\mid C_i)^{w_i} \tag{3}$$

其中，$V_{wnb}(C)$为加权朴素贝叶斯归类后的最大类别，并且w_i为属性X_i的权重。

2.2. 区间 DEA 相关知识

传统 DEA 模型中的输入输出数据使用的是标称数据，考虑的是数据是确定性的情况。但在实际生活中，决策单元的输入输出通常是不确定性的。因此 Wang 等人[20]提出了区间 DEA 模型，与传统的 DEA 模型不同，区间 DEA 模型的输出和输入处于一定的有界区间内，假设输出的取值范围为$\left[y_{ij}^{L},y_{ij}^{U}\right]$，输入的取值范围为$\left[x_{ij}^{L},x_{ij}^{U}\right]$，并且$x_{ij}^{L}>0,y_{rj}^{L}>0$。

首先，考虑对目标决策单元最有利的情况，也就是使得决策单元的输出最大化，输入最小化，此时传统 DEA 模型可以转化为下面的模型(4)，我们可以根据模型(4)得到区间 DEA 模型的效率值的上限θ^{U}，θ^{U}是在最有利的条件下最差的相对效率。

$$\begin{aligned}
&\max \theta^{U}=\sum_{r=1}^{s}u_r y_{ro}^{U}\\
&\text{s.t.}\quad \sum_{i=1}^{m}v_i x_{io}^{L}\le 1\\
&\qquad \sum_{r=1}^{s}u_r y_{rj}^{L}-\sum_{i=1}^{m}v_i x_{ij}^{U}\le 0,\forall j\\
&\qquad u_r,v_i\ge 0.
\end{aligned} \tag{4}$$

同理，当决策单元的输出最小化，输入最大化时，传统 DEA 模型可以转化为以下的模型(5)，此时是对目标决策单元最不利的情况，我们可以根据模型(5)来得出区间 DEA 模型效率值的下限θ^{L}，θ^{L}是在最不利的条件下最差的相对效率。

$$\begin{aligned}
&\max \theta^{L}=\sum_{r=1}^{s}u_r y_{ro}^{L}\\
&\text{s.t.}\quad \sum_{i=1}^{m}v_i x_{io}^{U}\le 1\\
&\qquad \sum_{r=1}^{s}u_r y_{rj}^{U}-\sum_{i=1}^{m}v_i x_{ij}^{L}\le 0,\forall j\\
&\qquad u_r,v_i\ge 0.
\end{aligned} \tag{5}$$

也就是说，区间 DEA 模型悲观的效率区间是$\theta\in\left[\theta^{L},\theta^{U}\right]$。

2.3. RDEA 相关知识

传统 DEA 模型中的输入输出数据使用的是标称数据，考虑的是数据是确定性的情况。但在实际生活中，决策单元的输入输出通常是不确定性的。鲁棒优化是一种求解不确定问题中常用的方法，也就是在最坏的情况下寻找最优解。RDEA 模型就是考虑不确定的输入输出，使得在最坏的情况下仍能保持鲁棒性。具有不确定性的输入输出变量通常表示为：

$$U=\left\{\tilde{x}_{ij}=x_{ij}+\xi_{ij}^{x}\hat{x}_{ij},\tilde{y}_{rj}=y_{rj}+\xi_{rj}^{y}\hat{y}_{rj},\hat{x}_{ij},\hat{y}_{rj}\in Z\right\}$$

其中 $\hat{x}_{ij}=\delta^{x}x_{ij}$，$\hat{y}_{rj}=\delta^{y}y_{rj}$，$\xi_{ij}^{x},\xi_{ij}^{y}$ 是给定的偏离标称值，$\hat{x}_{ij},\hat{y}_{rj}$ 与标称值 x_{ij},y_{rj} 的扰动百分比，δ^{x},δ^{y} 是输入和输出的不确定扰动因子。

所以具有不确定输入输出 $\tilde{x}_{ij},\tilde{y}_{rj}\in U$ 的决策单元的 RDEA 模型如下所示[21]：

$$\begin{aligned}&\max\theta=\sum_{r=1}^{s}u_{r}\tilde{y}_{ro}\\&\text{s.t.}\quad\sum_{i=1}^{m}v_{i}\tilde{x}_{io}\le 1,\forall\tilde{x}_{io}\in U\\&\qquad\sum_{r=1}^{s}u_{r}\tilde{y}_{rj}-\sum_{i=1}^{m}v_{i}\tilde{x}_{ij}\le 0,\forall\tilde{x}_{ij},\tilde{y}_{rj}\in U\\&\qquad u_{r},v_{i}\ge 0.\quad\forall_{r},\forall_{i}\end{aligned}\tag{6}$$

3. 研究框架

本研究引入了一个全面的基于在线评论决策支持框架，通过在线消费者评论来对替代产品进行排名，以供决策者更好的做出购买决策。本文提出的决策支持框架主要分为三个部分，即数据预处理、情感分析和基准分析，如图 1 所示。每个模块的过程和基本作用会在下面进行详细的描述。

1) 数据处理：使用 python 从京东平台抓取产品的在线消费者评论，对抓取到的数据进行数据清洗和预处理，然后提取出关键属性，作为判断替代产品排名的评价指标。

2) 情感分析：将清洗后的在线评论句子聚类到提取出的关键属性群组中，然后基于朴素贝叶斯对在线评论进行情感分析。

3) 基准分析：利用提出的 RDEA 模型来计算可替代方案的效率得分，然后根据效率得分对可替代产品进行排名。

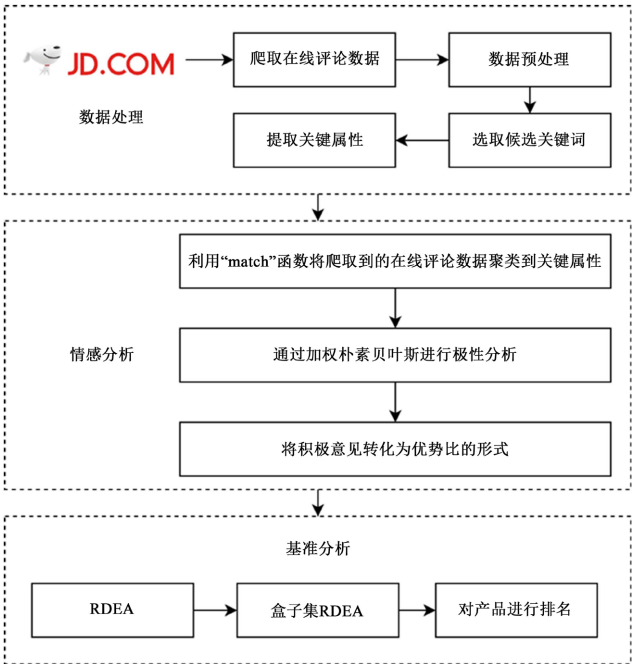

Figure 1. Flow chart of decision-making framework
图 1. 决策框架流程图

3.1. 数据处理

尽管现在很多购物网站上都提供了对商品量化标准，但是由于其主观性和可变性导致的意见两极化，还是建议消费者通过在线评论来了解商品的具体情况。数据处理的过程如下：

1) 使用 python 从京东 JD.COM (https://www.jd.com/)爬取了戴尔、宏基、华为、惠普和联想等品牌的 15 款笔记本电脑的在线评论数据，可替代商品的集合表示为 $A=\{A_1,A_2,\cdots,A_n\}$。本文主要是爬取前 100 页的产品在线评论(京东限制最多查看前 100 页，不足则全部获取)，得到的在线评论数据以 xls 格式进行存储，方便计算机程序导入数据。

2) 对在线评论数据进行预处理。先进行数据清洗，进行重新审查和校验，保证数据的一致性；然后使用 jieba 中文分词程序对文本的句子进行中文分词；去停用词(通常包括连词，介词、代词、标点符号、逻辑字符和特殊字符)，参考哈工大停用词库；结合上下文语境对在线评论中的词语进行性质的确定以及标注，将词语分类为形容词、名词、动词等。

3) 提取关键属性。通过计算每条评论对应的 Term-Frequency-Inverse Document Frequency (TF-IDF)值，选取在线评论中的候选关键词。使用 word2vec，训练词向量权重，之后通过分词，去除不符合条件的词，获取频率最高的 Top 200 词。然后利用 K-means 聚类算法对获取词的词向量进行聚类，提取出关键属性，创建归一化标签。

3.2. 情感分析

情感分析是对人们的意见、情感、情绪和态度的研究，通过对文本上下文的挖掘，来识别和提取文本数据中的主观信息[22]。朴素贝叶斯是用来进行情感分析常用的一种机器学习的方法，而加权朴素贝叶斯是朴素贝斯的一个扩展，其中的属性具有不同的权重[19]。

本节旨在计算产品各关键属性的情感得分，所以首先是利用 R 语言中的模式匹配“match”函数，利用算法 1 根据关键属性对爬取的在线评论数据进行聚类。然后利用朴素贝叶斯对在线评论进行情感分析。首先对向量进行转化，再对分类器进行训练，为了防止模型过度拟合，按照 80%:20%的比例将数据集划分为训练集和测试集，最后进行预测分类。本文利用的是 pysenti 库(加权朴素贝叶斯)，结合句子结构给各情感词语的情感极性赋予权重，然后加权求和得到文本的情感极性得分。我们将情感极性得分位于 $[0,0.45],[0.45,0.55],[0.55,1]$，分别分类为消极中性和积极意见。

从原始的情感极性结果中(图 3)可以看出，更多的积极意见反映了更高的客户满意度，所以本文采用优势比[23]的定义来表示积极意见相对于其他类型意见的程度，用 Ωpos 来表示。

$$\Omega pos=\frac{Ppos}{1-Ppos} \tag{7}$$

其中，$Ppos$ 代表积极意见在在线消费评论中的概率，$1-Ppos$ 代表极性为中性和消极的概率之和。Ωpos 的值越大，代表积极意见的占比也就越高。$\Omega pos=1$ 时，代表积极意见的概率等于中性意见和消极意见的概率之和。根据公式(7)变形，可以得到 $Ppos=\Omega pos/(1+\Omega pos)$。

算法 1. 根据产品对在线评论进行聚类

索引，集合和参数：
i：评论中的句子索引 $(i=1,2,\cdots,I)$
j：产品索引 $(j=1,2,\cdots,J)$

Continued

k：关键特征索引 $(k=1,2,\cdots,K)$

x_i：提取出的 i^{th} 词向量

C_k：k^{th} 关键特征

S_{ij}：i^{th} 评论中关于 j^{th} 个产品的句子

F_{kj}：j^{th} 个产品的 k^{th} 关键特征

方法：

While $j \le J$ do{

While $k \le K$ do{

应用“match”函数来检测是否 C_k 存在于 S_{ij}

分类 S_{ij} 到 F_{kj}，如果 C_k 匹配到其中的某个词典 S_{ij}.

}

}

3.3. 基准分析

DEA 是一种典型的非参数线性规划性能评估模型，鲁棒数据包络分析(RDEA)是一种基于 DEA 的保守的方法，用于对决策单元的输入和输出数据中的不确定性建模。输入输出数据的不确定性有多种情况，结合在线评论的实际情况，本文主要考虑输出数据的不确定性，由两部分组成，一部分是确定的值，另一部分是不确定的值。输出的不确定性描述如下：

$$U=\left\{\tilde{y}_{rj}=y_{rj}+\sum_{l=1}^{L}\delta_l y_{rj}^{R},\delta_l\in Z\right\}$$

本文主要考虑输出数据的是盒子集不确定集合的情况，然后构造了输出数据是盒子集的 RDEA 模型。

定理 1：基于盒子不确定集合的 RDEA 可以构造为：

$$\begin{aligned}&\max\theta\\&\text{s.t. }\theta-\sum_{r=1}^{s}u_r y_{ro}-\Phi\sum_{l=1}^{L}\sum_{r=1}^{s}u_r y_{ro}^{R}\le 0\\&\sum_{i=1}^{m}v_i x_{io}\le 1\\&\sum_{r=1}^{S}u_r y_{rj}-\sum_{i=1}^{m}v_i x_{ij}+\Phi\sum_{l=1}^{L}\sum_{r=1}^{s}u_r y_{rj}^{R}\le 0\\&u_r,v_i\ge 0.\quad\forall_r,\forall_i\end{aligned}\tag{8}$$

其中的盒子集不确定集合定义为：$Z^{box}=\left\{\delta_l\in\mathbb{R}^L:\left\|\delta_l\right\|\le\Phi\right\}$，其中 Φ 是不确定输出的鲁棒参数，用于衡量盒子不确定集的不确定度。$y_{rj}^{R}=\xi_{rj}^{y}y_{rj}$，$\xi_{rj}^{y}$ 是给定的偏离标称值，$\hat{x}_{ij},\hat{y}_{rj}$ 与标称值 x_{ij},y_{rj} 的扰动百分比。δ_l 是输出的不确定扰动因子，L 是不确定因子的个数。x_{ij},y_{rj} 分别表示第 j 个决策单元的输入和输出。具体的证明如下：

证明：该模型仅考虑输出数据是不确定性的，也就是说 $\tilde{y}_{rj}=y_{rj}+\sum_{l=1}^{L}\delta_l y_{rj}^{R}$，所以模型(4)可转化为：

$$\begin{aligned}&\max \theta\\ &\text{s.t.}\quad \theta-\sum_{r=1}^{s}u_r y_{ro}-\sum_{l=1}^{L}\sum_{r=1}^{s}u_r\delta_l y_{ro}^{R}\le 0\\ &\sum_{i=1}^{m}v_i x_{io}\le 1\\ &\sum_{r=1}^{s}u_r y_{rj}-\sum_{i=1}^{m}v_i x_{ij}+\sum_{l=1}^{L}\sum_{r=1}^{s}u_r\delta_l y_{rj}^{R}\le 0\\ &u_r,v_i\ge 0.\quad \forall_r,\forall_i\end{aligned}\tag{9}$$

并且 δ_l 属于盒子不确定集合，满足 $Z^{box}=\left\{\delta_l\in\mathbb{R}^L:\left\|\delta_l\right\|\le\Phi\right\}$，模型(6)中的约束 $\theta-\sum_{r=1}^{s}u_r y_{ro}-\sum_{l=1}^{L}\sum_{r=1}^{s}u_r\delta_l y_{ro}^{R}\le 0,\forall\left\{\delta_l:\left\|\delta_l\right\|\le\Phi\right\}$，等价于下面的问题：

$$\begin{aligned}&\max\sum_{l=1}^{L}\sum_{r=1}^{s}u_r\delta_l y_{ro}^{R}\le\theta-\sum_{r=1}^{s}u_r y_{ro},\forall\left\{\delta_l:\left\|\delta_l\right\|\le\Phi\right\}\\ &\max_{\left\|\delta_l\right\|\le\Phi}\sum_{l=1}^{L}\sum_{r=1}^{s}u_r\delta_l y_{ro}^{R}=\Phi\sum_{l=1}^{L}\sum_{r=1}^{s}u_r y_{ro}^{R}\end{aligned}\tag{10}$$

所以第一个约束可以转化为 $\theta-\sum_{r=1}^{s}u_r y_{ro}-\Phi\sum_{l=1}^{L}\sum_{r=1}^{s}u_r y_{ro}^{R}\le 0$，同理第三个约束可以转化为 $\sum_{r=1}^{S}u_r y_{rj}-\sum_{i=1}^{m}v_i x_{ij}+\Phi\sum_{l=1}^{L}\sum_{r=1}^{s}u_r y_{rj}^{R}\le 0$。□

4. 实证分析

为了验证所提出的决策框架的有效性和适用性，将其应用于基于在线评论数据的实际案例中。利用 python 从京东 JD.COM (https://www.jd.com/)抓取了 15 款笔记本电脑的在线消费者评论数据，包括戴尔灵越 5000，戴尔游匣 5515，戴尔游匣 G15，宏基暗影骑士、宏基非凡 S3、宏基掠夺者、华为 mate book 14S 2021、华为 mate book D15、华为 mate book X pro 2021、惠普暗影精灵、惠普星 15、惠普战 99、联想小新 Air 14 2021、联想拯救者 Y9000K2021、联想拯救者 Y9000P，可替代商品的集合表示为 $A=\left\{A_1,A_2,\cdots,A_{15}\right\}$，表 3 显示了爬取到的在线消费者评论在数据清洗之后的 101,405 条数据集。

4.1. 数据处理

首先对数据清洗之后的在线消费者评论数据进行预处理。预处理过程需要先进行中文分词，去停用词，以及对词性进行标注。然后用 TF-IDF 算法提取在线消费者评论中的候选关键词，留取频率最高的 Top 200 词。去除不符合条件的词之后，用 K-means 聚类算法，基于点与点之间的距离的相似度计算最佳类别归属，创建归一化标签，K-means 聚类图如下图 2 所示。

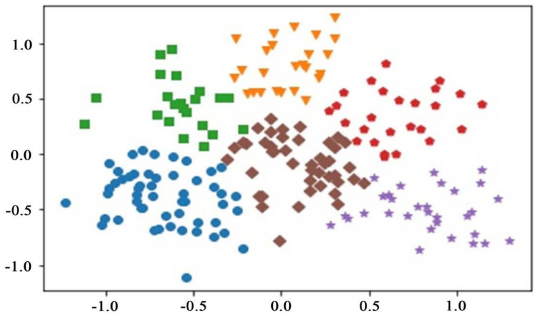

Figure 2. Clustering effect when $K=6$
图 2. $K=6$ 时，聚类效果图

从图 2 中可以看出当 $K = 6$ 时，聚类的效果比较好，分类也比较明显。聚类过程将具有相似含义的单词分类到同一组中，总共聚类为六个关键属性：售后服务、质量、物流、价格、外观、赠品，分别用 $C=\{C_1,C_2,\cdots,C_6\}$ 表示。

4.2. 情感分析

本节旨在对抓取到的可替代产品的在线消费者评论进行情感分析，情感极性分为积极、中性和消极。情感分析的过程就是在将线消费者评论的文本信息转化为可以进行分析和处理的数据。

首先利用 R 语言中的“match”函数，根据关键属性对数据清洗后的可替代产品的在线消费者评论进行聚类。然后对预处理后的词向量进行转化，对分类器进行训练。最后利用加权朴素贝叶斯，结合句子结构给各情感词语的情感极性赋予权重，然后加权求和得到线消费者评论的情感极性得分。情感分析的极性结果如表 1 所示，其中 Pos，Neu，Neg 分别代表情感极性为积极，中性和消极的在线消费者评论的数量。

本文对表 1 中的可替代产品的在线消费者评论数据进行归一化处理，为了说明在线消费者评论中，积极意见更能代表顾客的满意度。归一化的处理结果如表 2。

Table 1. Results of emotional polarity
表 1. 情感极性的结果

属性	极性	可替代产品														
		A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8	A_9	A_{10}	A_{11}	A_{12}	A_{13}	A_{14}	A_{15}
C_1	Pos	3218	730	25	1514	81	54	2246	2663	352	899	1880	424	1907	2173	1458
	Neu	66	51	3	67	3	4	62	42	0	35	67	7	41	92	60
	Neg	413	173	132	353	57	7	365	264	775	1300	859	308	1081	252	362
C_2	Pos	22	3634	909	16,250	9599	2712	19,168	18,671	9413	16,631	12,476	13,755	12,414	10,304	9587
	Neu	202	8	10	126	45	32	30	84	244	188	149	35	61	40	172
	Neg	546	141	229	1064	526	467	875	600	1760	2578	1678	1190	1565	464	1009
C_3	Pos	8	736	118	5301	345	746	1234	1783	1098	4289	1025	1848	1064	3115	2975
	Neu	82	10	0	52	3	12	26	16	0	44	24	35	17	56	67
	Neg	430	51	18	199	51	109	111	126	319	518	165	210	262	191	748
C_4	Pos	0	126	46	1435	1104	172	378	505	865	1281	1191	1029	1379	533	519
	Neu	22	0	0	10	6	6	0	4	18	45	10	0	8	8	0
	Neg	5284	4	7	63	45	22	35	10	339	196	147	105	188	46	89
C_5	Pos	10	1732	302	6689	5827	1080	14,690	14,490	6540	6814	8840	8309	6914	3705	3986
	Neu	119	4	1	25	24	10	19	22	63	49	35	63	24	12	24
	Neg	530	11	12	115	74	88	143	160	386	297	257	161	156	34	118
C_6	Pos	3	120	10	214	147	22	819	1450	197	152	757	98	841	72	278
	Neu	127	16	0	4	3	0	0	27	18	2	31	0	26	4	2
	Neg	22	41	6	60	24	7	108	180	169	88	254	126	394	28	83

Table 2. Normalized results of emotional polarity
表 2. 情感极性归一化结果

极性	可替代产品														
	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8	A_9	A_{10}	A_{11}	A_{12}	A_{13}	A_{14}	A_{15}
Pos	0.29	0.93	0.77	0.94	0.95	0.86	0.95	0.96	0.82	0.85	0.88	0.92	0.87	0.94	0.87
Neu	0.06	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.02	0.01	0.01	0.001	0.01	0.01	0.02
Neg	0.65	0.06	0.22	0.06	0.04	0.13	0.04	0.03	0.17	0.14	0.11	0.08	0.13	0.05	0.11

将表 2 中在线消费者评论的归一化结果画出如下条形图，能更加清楚地看出在线消费者评论中，积极意见占比最多，总的积极意见占比为85.407%，中性意见的占比为1.202%，消极意见的占比为13.391%。

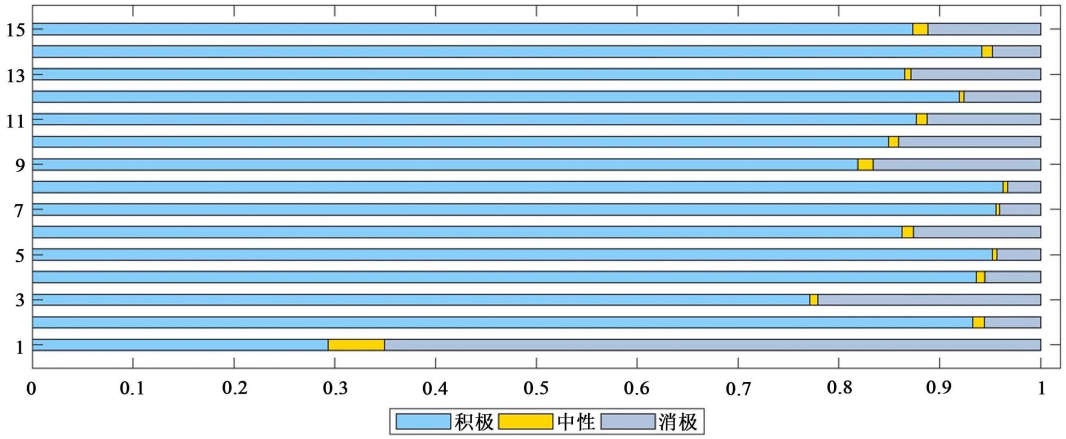

Figure 3. Polarity results of online reviews of alternative products before clustering
图 3. 聚类之前可替代产品在线评论的极性结果

从图 3 中可以看出积极意见的占比最大，所以根据公式(6)将上述表中积极意见的数据转化成积极意见的优势比，结果如下表 3 所示：

Table 3. Advantage ratio of key attributes
表 3. 关键属性的优势比

产品	关键属性					
	C_1	C_2	C_3	C_4	C_5	C_6
A_1	6.72	0.03	0.02	0.00	0.02	0.02
A_2	3.26	24.39	12.07	31.50	115.47	2.11
A_3	0.19	3.80	6.56	6.57	23.23	1.67
A_4	3.60	13.66	21.12	19.66	47.78	3.34
A_5	1.35	16.81	6.39	21.65	59.46	5.44
A_6	4.91	5.43	6.17	6.14	11.02	3.14
A_7	5.26	21.18	9.01	10.80	90.68	7.58
A_8	8.70	27.30	12.56	36.07	79.62	7.00
A_9	0.45	4.70	3.44	2.42	14.57	1.05
A_{10}	0.67	6.01	7.63	5.32	19.69	1.69
A_{11}	2.03	6.83	5.42	7.59	30.27	2.66
A_{12}	1.35	11.23	7.54	9.80	37.09	0.78
A_{13}	1.70	7.63	3.81	7.04	38.41	2.00
A_{14}	6.32	20.44	12.61	9.87	80.54	2.25
A_{15}	3.45	8.12	3.65	5.83	28.07	3.27

4.3. RDEA 结果

本文仅考虑了输出数据的不确定性，将表 3 中关键属性积极意见的优势比作为 RRDEA 模型一类输出。同时 DEA 模型要求决策单元必须有输入变量，所以本文将所有决策单元均赋予相同的虚拟输入变量 $x_{ij}=1$。因为本文考虑所有输出的不确定性，所以 $L=6$。从以前的参数设置中，我们可以知道扰动范围从 0 到 0.1，所以本文设置扰动变量 $\xi_{rj}^{y}=0.02$。

当不确定集合为盒子集时，15 个可替代商品的 RDEA 效率如表 4 所示。$\Phi=0$ 时，此时的模型(6)等价于传统的 DEA，也就是输入输出数据没有受到扰动的标称值问题，可替代产品的效率值和排名结果在表 5 中的第 2 列和第 3 列。当不确定参数 $\Phi=1$ 时，可替代产品的效率值和排名结果在表 5 中的第 4 列和第 5 列。

Table 4. Efficiency and ranking results
表 4. 效率和排名结果

产品	$\Phi=0$ 效率	排名	$\Phi=1$ 效率	排名	区间 DEA 效率	排名
A_1	0.7722	4	0.7719	3	[0.3708, 0.4017]	3
A_2	0.9999	2	1	1	[0.4803, 0.5204]	1
A_3	0.3707	12	0.3707	11	[0.1780, 0.1929]	11
A_4	1	1	1	1	[0.4803, 0.5204]	1
A_5	0.7589	5	0.7585	4	[0.3644, 0.3947]	4
A_6	0.5642	6	0.5640	5	[0.2709, 0.2935]	5
A_7	1	1	1	1	[0.4803, 0.5204]	1
A_8	1	1	1	1	[0.4803, 0.5204]	1
A_9	0.2155	13	0.2155	12	[0.1035, 0.1121]	12
A_{10}	0.4003	10	0.4002	9	[0.1922, 0.2082]	9
A_{11}	0.4021	9	0.4020	8	[0.1931, 0.2092]	8
A_{12}	0.4968	7	0.4968	6	[0.2387, 0.2585]	6
A_{13}	0.3804	11	0.3804	10	[0.1827, 0.1979]	10
A_{14}	0.9422	3	0.9422	2	[0.4526, 0.4903]	2
A_{15}	0.4550	8	0.4548	7	[0.2184, 0.2367]	7

5. 比较分析

为了证明提出方法的有效性，在本节进行两个方面的比较，首先是是否考虑数据的不确定性，其次是在考虑数据的不确定时的区间 DEA 之间的比较。

正如前面所说，传统的 DEA 没有考虑数据扰动，现实生活的不确定性使得结果的准确性很难保证。本文提出的一种根据在线评论对可替代产品进行排序的决策支持框架，不仅考虑了数据的不确定性，而且避免了人为给定关键属性权重的主观性。本文提出的该框架可以帮助消费者在在线评论信息过载的情况下，在多种商品之间轻松的做出购买决策。

为了考虑数据的不确定性，有研究提出了区间 DEA 模型，为了保证对比结果的公平性，本文在求解区间 DEA 模型的结果时，输出数据是在表 3 中积极意见的优势比的基础上，

$y_{ij} \in \left[y_{ij} - 0.02y_{ij}, y_{ij} + 0.02y_{ij} \right]$。区间 DEA 的输出数据如表 5 所示。将表 5 中的数据代入模型 4 和模型 5 中，计算得出的区间 DEA 效率和排名结果在表 4 中的第 6 列和第 7 列。

Table 5. Interval DEA output data
表 5. 区间 DEA 输出数据

产品	关键属性					
	C_1	C_2	C_3	C_4	C_5	C_6
A_1	[6.5837, 6.8525]	[0.02882, 0.0300]	[0.1531, 0.1594]	[0.0000, 0.0000]	[0.1960, 0.02040]	[0.01973, 0.02040]
A_2	[3.1937, 3.3241]	[23.9014, 24.8770]	[11.8242, 12.3068]	[30.8700, 32.1300]	[113.1573, 117.7760]	[2.0631, 2.1473]
A_3	[0.1814, 0.1888]	[3.7272, 3.8794]	[6.4244, 6.6866]	[6.4400, 6.7028]	[22.7661, 23.6953]	[1.6333, 1.7000]
A_4	[3.5326, 3.6768]	[13.3823, 13.9285]	[20.6971, 21.5419]	[19.26438, 20.0506]	[46.8229, 48.7341]	[3.2768, 3.4106]
A_5	[1.3230, 1.3770]	[16.4746, 17.1470]	[6.2611, 6.5166]	[21.2141, 22.0800]	[58.2700, 60.6483]	[5.3355, 5.5533]
A_6	[4.8109, 5.0072]	[5.3261, 5.5435]	[6.0419, 6.2885]	[6.0200, 6.2657]	[10.8000, 11.2408]	[3.0800, 3.2057]
A_7	[5.1547, 5.3651]	[20.7565, 21.6037]	[8.8271, 9.1874]	[10.5840, 11.0160]	[88.8654, 92.4925]	[7.4316, 7.7350]
A_8	[8.5285, 8.8766]	[26.7508, 27.8427]	[12.3052, 12.8074]	[35.3500, 36.7928]	[78.0230, 81.2076]	[6.8647, 7.1449]
A_9	[0.4451, 0.4632]	[4.6031, 4.79104]	[3.3731, 3.5108]	[2.3745, 2.4714]	[14.2743, 14.8570]	[1.0324, 1.0745]
A_{10}	[0.6599, 0.6868]	[5.8924, 6.1329]	[7.4790, 7.7843]	[5.2090, 5.4216]	[19.2997, 20.0875]	[1.6551, 1.7226]
A_{11}	[1.9896, 2.0708]	[6.6921, 6.9652]	[5.3148, 5.5317]	[7.4342, 7.7377]	[29.6684, 30.8794]	[2.6030, 2.7092]
A_{12}	[1.3191, 1.3729]	[11.0040, 11.4531]	[7.3920, 7.6937]	[9.6040, 9.9960]	[36.3518, 37.8356]	[0.7622, 0.7933]
A_{13}	[1.6656, 1.7336]	[7.4819, 7.7873]	[3.7373, 3.8898]	[6.8950, 7.1764]	[37.6428, 39.1793]	[1.9623, 2.0424]
A_{14}	[6.1905, 6.4431]	[20.0355, 20.8533]	[12.3591, 12.8635]	[9.6729, 10.0677]	[78.9326, 82.1543]	[2.2050, 2.2950]
A_{15}	[3.3858, 3.5240]	[7.9553, 8.2800]	[3.5773, 3.7233]	[5.7148, 5.9480]	[27.5090, 28.6318]	[3.2051, 3.3360]

为了能更清晰地展示不同方法的排名结果，本文画了折线图如下图 4 所示。

从图 4 排名结果折线图中可以看出，本文提出的方法与区间 DEA 的排名结果重合，与传统 DEA 的方法得到的排名结果的趋势相同，这能说明本文提出的决策支持框架的基准分析部分的方法的适用性。

本文提出决策支持框架的基准分析部分的方法的优越性体现在，相对于传统 DEA 方法来说，本文提出的方法考虑了数据的不确定性，更符合客观现实情况；相对于区间 DEA 方法来说，本文提出的方法更加简便，操作更易上手。而且区间 DEA 方法需要知道数据的具体分布，但是这在现实中的实现是比较困难的。而 RDEA 方法不需要提前知道数据的分布情况，并且这些方法之间的比较是基于本文提出的决策支持框架的数据处理和情感分析之上。

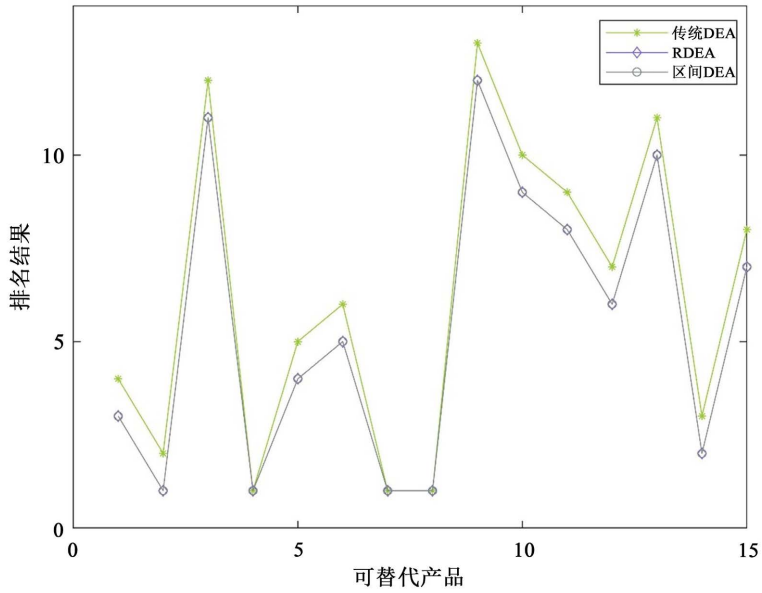

Figure 4. Ranking result line chart
图 4. 排名结果折线图

6. 结论

随着大数据爆炸式的增长，在线消费者评论数量也越来越多，如何使用这些在线评论来帮助决策变得越来越复杂。消费者和商家需要耗费大量的时间来阅读在线评论，识别在线评论中的有效信息之后，再从众多可替代产品中做出决策。所以在本研究中，我们提出了一种新的决策支持框架，该方法充分利用情感分析和 RDEA 通过在线评论对替代产品进行排序，帮助消费者和商家做出决策。

本文提出的决策支持框架总共包括三个部分，分别是数据处理、情感分析和基准分析。首先我们使用 python 从京东平台上抓取产品的在线消费者评论，对抓取到的数据进行预处理并且提取出关键特征作为评价指标。然后基于朴素贝叶斯对在线评论进行情感分析，由于积极意见的在线消费者评论更能代表消费者的满意度，所以本文将积极意见的优势比作为模型的输出。最后利用提出的 RDEA 模型求出可替代产品的 RDEA 效率，然后根据效率得分对可替代产品进行排名。比较研究的分析结果也说明我们提出的方法考虑的问题更加全面客观，也更符合客观实际情况。本文提出的基于在线评论的决策支持框架的主要贡献如下：

1) 使用机器学习中的加权朴素贝叶斯对在线评论进行情感分析，可以获得更高的准确率；通过 RDEA 模型考虑数据的不确定性对产品进行排名，鲁棒优化是处理不确定性的一种常见的方法，我们通过盒子不确定集来考虑数据的不确定性，最后通过求出的 RDEA 效率来对可替代产品进行排名；

2) 从在线评论中提取关键属性作为评价指标。相对于以往给定的评价指标，本文从消费者的角度出发，利用 TF-IDF 算法提取在线消费者评论中的消费者关心的商品的关键词，再通过 K-means 聚类出关键属性作为评价指标。考虑消费者的偏好，更加客观符合实际情况；

3) 从京东(JD.COM)上爬取的 15 款笔记本电脑的 101,405 条在线评论进行数值实验，来验证提出模型的有效性和适用性。本文提出的决策框架的基准分析部分，与传统的 DEA 模型相比，本文提出的方法考虑了数据的不确定；与区间 DEA 模型相比，本文提出的方法步骤更加简便，易于操作。在信息爆炸的时代，帮助消费者从商品的在线消费者评论中做出购买决策。

目前，在线评论已用于不同场景现实生活中的。本文中提出的方法不仅适用于替代产品排名基于电

子商务中的在线评论来帮助消费者做出购买决策，还对商家提供低成本和时效性的信息来帮助做出管理决策，还可以应用于具有类似流程的，比如在旅游业，在医疗行业、电影和电视等其他行业。本研究也有一些局限性，只识别了在线评论的文本信息，未来可以改进该方法以识别更多形式的在线评论，如表情符号和视频等。

参考文献

[1] Park, D.H., Lee, J. and Han, I. (2007) The Effect of On-Line Consumer Reviews on Consumer Purchasing Intention: The Moderating Role of Involvement. *International Journal of Electronic Commerce*, **11**, 125-148. https://doi.org/10.2753/JEC1086-4415110405

[2] Zhan, Y., Tan, K.H., Li, Y. and Tse, Y.K. (2018) Unlocking the Power of Big Data in New Product Development. *Annals of Operations Research*, **270**, 577-595. https://doi.org/10.1007/s10479-016-2379-x

[3] Das, S.R. and Chen, M.Y. (2007) Yahoo! for Amazon: Sentiment Extraction from Small Talk on the Web. *Management Science*, **53**, 1375-1388. https://doi.org/10.1287/mnsc.1070.0704

[4] Moreo, A., Romero, M., Castro, J.L. and Zurita, J.M. (2012) Lexicon-Based Comments-Oriented News Sentiment Analyzer System. *Expert Systems with Applications*, **39**, 9166-9180. https://doi.org/10.1016/j.eswa.2012.02.057

[5] Jiao, J. and Zhou, Y. (2011) Sentiment Polarity Analysis Based Multi-Dictionary. *Physics Procedia*, **22**, 590-596. https://doi.org/10.1016/j.phpro.2011.11.091

[6] Jurek, A., Mulvenna, M.D. and Bi, Y. (2015) Improved Lexicon-Based Sentiment Analysis for Social Media Analytics. *Security Informatics*, **4**, Article No. 9. https://doi.org/10.1186/s13388-015-0024-x

[7] Medhat, W., Hassan, A. and Korashy, H. (2014) Sentiment Analysis Algorithms and Applications: A Survey. *Ain Shams Engineering Journal*, **5**, 1093-1113. https://doi.org/10.1016/j.asej.2014.04.011

[8] Zhang, W., Xu, H. and Wan, W. (2012) Weakness Finder: Find Product Weakness from Chinese Reviews by Using Aspects Based Sentiment Analysis. *Expert Systems with Applications*, **39**, 10283-10291. https://doi.org/10.1016/j.eswa.2012.02.166

[9] Xu, K., Liao, S.S., Li, J. and Song, Y. (2011) Mining Comparative Opinions from Customer Reviews for Competitive Intelligence. *Decision Support Systems*, **50**, 743-754. https://doi.org/10.1016/j.dss.2010.08.021

[10] Zhang, D., Xu, H., Su, Z. and Xu, Y. (2015) Chinese Comments Sentiment Classification Based On Word2vec and SVM^{perf}. *Expert Systems with Applications*, **42**, 1857-1863. https://doi.org/10.1016/j.eswa.2014.09.011

[11] Tian, F., Wu, F., Chao, K.-M., Zheng, Q., Shah, N., Lan, T. and Yue, J. (2016) A Topic Sentence-Based Instance Transfer Method for Imbalanced Sentiment Classification of Chinese Product Reviews. *Electronic Commerce Research and Applications*, **16**, 66-76. https://doi.org/10.1016/j.elerap.2015.10.003

[12] Kang, H., Yoo, S.J. and Han, D. (2012) Senti-Lexicon and Improved Naïve Bayes Algorithms for Sentiment Analysis of Restaurant Reviews. *Expert Systems with Applications*, **39**, 6000-6010. https://doi.org/10.1016/j.eswa.2011.11.107

[13] Duric, A. and Song, F. (2012) Feature Selection for Sentiment Analysis Based on Content and Syntax Models. *Decision Support Systems*, **53**, 704-711. https://doi.org/10.1016/j.dss.2012.05.023

[14] Li, F. (2010) The Information Content of Forward-Looking Statements in Corporate Filings—A Naïve Bayesian Machine Learning Approach. *Journal of Accounting Research*, **48**, 1049-1102. https://doi.org/10.1111/j.1475-679X.2010.00382.x

[15] Liu, Y., Bi, J.-W. and Fan, Z.-P. (2017) A Method for Ranking Products Through Online Reviews Based on Sentiment Classification and Interval-Valued Intuitionistic Fuzzy TOPSIS. *International Journal of Information Technology & Decision Making*, **16**, 1497-1522. https://doi.org/10.1142/S021962201750033X

[16] Yang, L. and Li, Y. (2022) A New Method for Ranking the Usefulness of Negative Online Reviews Based on Combined Weighting Method and Improved TOPSIS. *Journal of Intelligent & Fuzzy Systems*, **42**, 3719-3736. https://doi.org/10.3233/JIFS-211928

[17] Liang, X., Liu, P. and Wang, Z. (2019) Hotel Selection Utilizing Online Reviews: A Novel Decision Support Model Based on Sentiment Analysis and Dl-Vikor Method. *Technological and Economic Development of Economy*, **25**, 1139-1161. https://doi.org/10.3846/tede.2019.10766

[18] Zhang, D., Li, Y. and Wu, C. (2020) An Extended TODIM Method to Rank Products with Online Reviews under Intuitionistic Fuzzy Environment. *Journal of the Operational Research Society*, **71**, 322-334. https://doi.org/10.1080/01605682.2018.1545519

[19] Zhang, H. and Sheng, S. (2004) Learning Weighted Naive Bayes with Accurate Ranking. *Proceedings of the Fourth IEEE International Conference on Data Mining*, Brighton, 1-4 November 2004, 567-570.

[20] Wang, Y.-M., Greatbanks, R. and Yang, J.-B. (2005) Interval Efficiency Assessment Using Data Envelopment Analysis. *Fuzzy Sets and Systems*, **153**, 347-370. https://doi.org/10.1016/j.fss.2004.12.011

[21] Mensah, E.K. (2020) Robust Data Envelopment Analysis via Ellipsoidal Uncertainty Sets with Application to the Italian Banking Industry. *Decisions in Economics and Finance*, **43**, 491-518. https://doi.org/10.1007/s10203-020-00299-3

[22] Zhao, J., Liu, K. and Xu, L. (2016) Sentiment Analysis: Mining Opinions, Sentiments, and Emotions. *Computational Linguistics*, **42**, 595-598. https://doi.org/10.1162/COLI_r_00259

[23] Park, J. and Lee, B.K. (2021) An Opinion-Driven Decision-Support Framework for Benchmarking Hotel Service. *Omega*, **103**, Article ID: 102415. https://doi.org/10.1016/j.omega.2021.102415