



Importance-performance analysis to develop product/service improvement strategies through online reviews with reliability

Xingli Wu¹ · Huchang Liao² · Chonghui Zhang²

Received: 1 September 2022 / Accepted: 31 August 2023

© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2023

Abstract

Online reviews are important data for developing product/service improvement strategies. Relevant studies treated different online reviews as equally important, and the validity of the results was vulnerable to unreliable online reviews. To solve this challenge, this study proposes an importance-performance analysis model that considers the reliability of online reviews. First, the reliability degree of online reviews is defined based on the quality and timeliness of online reviews and the credibility of reviewers. To estimate the importance of product/service attributes from online reviews, a preference learning model is designed based on the reliability degrees of online reviews, where the online reviews with higher reliability have a greater impact on the learning results. In addition, the attribute performance is determined by aggregating the satisfaction of online reviews for the attribute. Finally, we verify the practicability of the proposed importance-performance analysis model by a case study on four five-star hotels.

Keywords Importance-performance analysis · Online reviews · Consumer satisfaction · Preference learning · Reliability

1 Introduction

Consumer satisfaction analysis plays a key role in product/service defect identification and upgradation. An important topic in consumer satisfaction analysis is to identify the factors that influence consumer satisfaction (Piris & Gay, 2021). Multi-attribute analysis methods have been widely used to measure consumers' comprehensive satisfaction with products through attribute-level satisfaction (Bi et al., 2020). Importance-performance analysis (IPA) (Martilla & James, 1977) is a multi-attribute analysis method that aims to identify the most crucial attributes in terms of strengths and weaknesses by comparing the importance and performance of different attributes. It prioritizes attributes and suggests an effective allocation of limited

✉ Huchang Liao
liaohuchang@163.com

¹ Business School, Sichuan University, Chengdu 610064, China

² College of Statistics and Mathematics, Zhejiang Gongshang University, Hangzhou 310018, China

resources to improve and maintain consumer satisfaction. Owing to its simplicity and ease of application, the IPA has been applied in many fields such as tourism (Chen et al., 2022), food (Ji & Han, 2022), and education (Phadermrod et al., 2019).

In the framework of IPA, an important process is to determine the importance and performance of different attributes. Attribute importance can be considered as a measure of the perceived value or significance of an attribute to consumers. It is also associated with consumer expectations. The more important an attribute is, the greater the impact of the performance of the attribute on consumer satisfaction or dissatisfaction is. Attribute performance is inherent in a product/service. It is usually not quantifiable and objectively determined, but can be subjectively evaluated by consumers. Existing studies on IPA mainly used questionnaire-based surveys to obtain information about attribute importance and performance (Dueñas et al., 2021; Sun et al., 2020). For example, consumers who purchased a target product/service are invited to rate the importance of a set of predetermined attributes and their satisfaction with the performance of the attributes based on Likert scales (Mejia et al., 2022). Grigorioudis and Siskos (2002) developed the MUSA (MULTicriteria Satisfaction Analysis) method to estimate attribute importance using an aggregation-disaggregation process based on consumers' overall satisfaction and attribute-level satisfaction gathered through questionnaires. Nevertheless, questionnaire-based approaches do not offer advantages in terms of volume, time, and cost of data acquisition (Bi et al., 2019). In addition, due to respondents' limited cognition of attributes predefined by managers, the evaluation of these attributes can also be biased. Therefore, the results obtained by surveys may not reflect the real preferences of consumers.

Online reviews carry lots of information about consumer satisfaction with product attributes (Yang et al., 2021). They are positive or negative evaluations posted by consumers on e-commerce platforms in the form of ratings (e.g., 1-star to 5-star ratings) or unstructured text. Their characteristics of easy to collect, large amount of data and wide coverage can compensate for the disadvantages of survey data. Therefore, in recent years, many researchers have used online reviews to conduct the IPA. For example, Bi et al. (2019) conducted the IPA on two five-star hotels and verified that using online reviews provides valid results at a lower cost and in less time than using questionnaires. Hu et al. (2020) proposed an asymmetric impact-sentiment-performance analysis method in which key product attributes were identified by the terms frequently mentioned in online reviews. They pointed out that the importance/performance ratings of predefined product attributes estimated from questionnaires hardly reflect competitive situations from the consumer perspective. In the same spirit, Albayrak et al. (2021) evaluated the competitiveness of a theme park with its competitor in Hong Kong by analyzing online reviews based on the IPA. Zhang et al. (2021) made the prioritization of improvements in Four Seasons Hotel by an online reviews-driven method. It is observed that these studies ignored the impact of the reliability of online reviews on judging attribute importance and performance. The results may be misled by online reviews with poor reliability. In fact, not all online reviews of a product/service are completely reliable for measuring its current attribute performance and importance. For example, sellers often resort to hiring fake reviewers, swiping orders, sending red envelopes, and discounts to obtain extreme praise, resulting in distortion of information. The reliability of online reviews varies depending on information quality, source credibility and timeliness (Ngo-Ye & Sinha, 2014). Information quality refers to the content quality of online reviews from the perspective of review textual features, including review length, readability, expressiveness and so on (Korfiatis et al., 2012). Source credibility is related to reviewer characteristics, including personal information disclosure, professionalism, reputation, and so on. Timeliness also affects the reliability of online reviews as a measure of attribute performance and real-time consumer

preferences (Yeap et al., 2014). How to conduct the IPA based on reliable online reviews is a research gap that we want to solve in this study.

We present a reliability-based IPA model through online reviews to determine the prioritization of improvements for products/services. The proposed model aims to answer the following research questions: (1) How to determine the reliability of each online review? (2) How to infer the importance and performance of each attribute from online reviews while considering the reliability of online reviews? (3) How to prioritize attributes based on their importance and performance?

The main contributions of the paper include: (1) The reliability of online reviews is defined based on information quality, source credibility, timeliness, and usage period before reviewing. (2) A mathematical programming model is established to learn consumers' preference models from online reviews by minimizing estimation errors of consumers' satisfaction with products. The reliability degree of online reviews is used as a discounting factor for the estimation errors, so that less reliable online reviews have less impact on estimation results. (3) A reliability-based information fusion operator is proposed to aggregate attribute performance values determined by multiple reviewers into a comprehensive performance value. An IPA model that considers the reliability of online reviews is proposed and further validated by a case study on four five-star hotels.

The remainder of this paper is organized as follows. Section 2 presents the methodology for conducting reliability-based IPA through online reviews. Section 3 applies the proposed model to hotel improvement, and the results are discussed in Sect. 4. Finally, Sect. 5 draws conclusions.

2 Methodology

We consider m same-type products/services, a_i ($i = 1, 2, \dots, m$), sold on an e-commerce platform. Each of them has n attributes c_j ($j = 1, 2, \dots, n$) that influence consumer satisfaction, and L_i online reviews posted on the e-commerce platform. The proposed IPA framework to develop an improvement strategy for each product/service contains four stages: (1) determining the reliability of each online review, (2) learning attribute importance from online reviews, (3) estimating attribute performance based on online reviews, and (4) identifying to-be-improved attributes through IPA plot. These stages are described in the following four subsections.

2.1 Determining the reliability of online reviews

Reliability is an objective property of an information source. It refers to the ability to provide a correct measurement or assessment for an issue under consideration (Wu & Liao, 2022). The reliability of an information source is independent of the reliability of other information sources. The reliability or confidence of reviews has been a focus in the field of machine learning, especially natural language processing. For example, researchers (Bilal & Almazroi, 2023; Xu et al., 2020) have introduced the BERT (Bidirectional Encoder Representations from Transformers) base model to classify helpful and unhelpful reviews. These studies can predict the helpfulness of recent reviews that did not have enough time to receive helpful votes. While "helpfulness" may reflect the quality of readers' perceptions of online reviews, it is not sufficient to reflect the reliability of online reviews in measuring product/service performance. This section considers four indicators to measure the reliability of both old and

new online reviews, including the quality of online reviews, the credibility of reviewers, the timeliness of online reviews, and the usage period before reviewing:

(1) The reliability defined based on the quality of online reviews.

To identify useful online reviews, most e-commerce platforms, such as Amazon.com and TripAdvisor.com, place a question at the end of each review to ask whether the review is “helpful”. Korfiatis et al. (2012) pointed out that voting on the helpfulness of online reviews is important since it reflects the focus of consumers’ purchase decisions. Helpful votes of online reviews represent the degree of consumer approval of online reviews, which is the perceived utility of information contained in the reviews (Huang et al., 2015). Therefore, helpful votes of online reviews have become a quality signal for online reviews (Li et al., 2021).

We consider the following three points for defining the reliability degree of an online review based on its quality.

1. Most online reviews get few helpful votes. Online reviews without votes do not mean they are completely unreliable. Therefore, when defining the reliability degree of an online review based on its helpful votes, a parameter greater than 0 needs to be set to reflect the reliability degrees of online reviews without votes. The more helpful votes an online review receives, the higher the reliability degree of the online review is.
2. A small number of online reviews can obtain a large number of helpful votes. To avoid the impact of too large values on the definition of reliability, a threshold can be set such that when an online review receives votes greater than or equal to this threshold, its reliability is 1. This threshold can be defined according to the distribution of the number of helpful votes of the considered online reviews such that most online reviews does not exceed this threshold.
3. According to the Weber-Fechner law (Lootsma, 1993), the closer a value to the reference point is, the greater the change in “psychological intensity” caused by the same amount of “stimulus quantity” is. The mathematical expression of the Weber-Fechner law is $PI = K \lg S$, where PI is the psychological intensity, S is the stimulus quantity, and K is a consistent. It demonstrates how the logarithmic relationship of the stimulus quantity to the psychological intensity. That is, the psychological intensity should grow in arithmetic steps as the geometric steps of the stimulus quantity increase. The number of helpful votes can be regarded as a “stimulus” to measure the reliability of online reviews, and a review with zero helpful vote is a reference point. The difference in reliability among reviews with helpful votes close to zero is greater than the difference in reliability among reviews far from zero. For example, online reviews with five helpful votes are significantly more reliable than those with zero helpful vote, while online reviews with 100 votes are not significantly more reliable than those with 105 votes. In addition, when the number of helpful votes is large enough, an increase in the number of helpful votes does not increase people’s perceptions of reliability degrees. Therefore, in this paper, we use the logarithm function to convert the number of helpful votes to the quality of online reviews.

Combining the above three points, we expand the mathematical expression of the Weber-Fechner law to define the reliability of online reviews as follows. Let h_{li} be the number of helpful votes received by the l -th online review of product/service a_i . Let H be a threshold of helpful votes such that most online reviews receive fewer helpful votes than

it. The reliability degree of the l -th online review can be defined based on the number of helpful votes (i.e., review quality) as

$$\sigma_{li}^H = \begin{cases} \sigma_H^* + (1 - \sigma_H^*) \times \frac{\log_{\lambda_1}(h_{li}+1)}{\log_{\lambda_1}(H+1)}, & \text{if } h_{li} < H \\ 1, & \text{if } h_{li} \geq H \end{cases} \quad (1)$$

where $\sigma_H^* \in [0, 1]$, a user-defined parameter, indicates the reliability degree of the online reviews without helpful votes. If $h_{li} = 0$, then $\sigma_{li}^H = \sigma_H^*$; if $h_{li} \geq H$, then $\sigma_{li}^H = 1$; otherwise, $\sigma_{li}^H \in (\sigma_H^*, 1)$. The larger the value of h_{li} is, the larger the reliability degree σ_{li}^H is. The base of the logarithmic function $\lambda_1 > 1$ is a user-defined parameter. A smaller value of λ_1 indicates a greater difference in reliability among the reviews with different overall votes that close to zero.

- (2) The reliability defined based on the credibility of reviewers.

To the same spirit, the total number of helpful votes a reviewer receives can reflect his/her credibility. The more thoughtful, truthful, or constructive a reviewer's comments are, the more helpful they are to other readers in making purchase decisions, and the more helpful votes the reviewer can get. This perspective leads us to the conclusion that the more overall votes a reviewer receives, the more reliable (s)he is. Therefore, we use the number of helpful votes to measure the credibility of reviewers.

Let r_{li} be the number of helpful votes received by the reviewer who posted the l -th online review of a_i . Let R be a threshold of helpful votes such that most reviewers receive fewer helpful votes than it. According to the Weber-Fechner law, the reliability degree of the l -th online review can be defined based on the number of helpful votes received by the reviewer (i.e., reviewer credibility) as

$$\sigma_{li}^R = \begin{cases} \sigma_R^* + (1 - \sigma_R^*) \times \frac{\log_{\lambda_2}(r_{li}+1)}{\log_{\lambda_2}(R+1)}, & \text{if } r_{li} < R \\ 1, & \text{if } r_{li} \geq R \end{cases} \quad (2)$$

where $\sigma_R^* \in [0, 1]$, a user-defined parameter, indicates the reliability degree of online reviews posted by reviewers who do not receive any helpful votes. If $r_{li} = 0$, then $\sigma_{li}^R = \sigma_R^*$; if $r_{li} \geq R$, then $\sigma_{li}^R = 1$; otherwise, $\sigma_{li}^R \in (\sigma_R^*, 1)$. The larger the value of r_{li} is, the larger the reliability degree σ_{li}^R is. λ_2 has the same meaning as λ_1 .

- (3) The reliability defined based on the timeliness of online reviews.

The timeliness of online reviews, i.e., the gap between the time of posting online reviews and the current time, is also a factor that affects the reliability of online reviews. Huang et al. (2018) pointed out that specific product reviews are more helpful when temporal cues are displayed at close range. Due to technological updates and improvements, there may be differences between currently selling products and previously sold products/services. As a result, sentiments in new online reviews usually are more consistent with the quality of currently selling products than sentiments in earlier online reviews (Liu et al., 2017; Zhang et al., 2020).

Let t_{li}^T be the time of the l -th online review posted for a_i , and T_i be the time of the first online review posted in the considered time period. Let T_C be the current time, which is regarded as a reference point. According to the Weber-Fechner law, when an online review is posted long enough from the current time, an increase in time does not cause an equal decrease in the perception of its reliability. For this reason, we use an exponential function to define the reliability degree the l -th online review based on the

posting time (i.e., review timeliness), that is

$$\sigma_{li}^T = \sigma_T^* + (1 - \sigma_T^*) \times \frac{\lambda_3^{(t_{li}^T - T_i)/(T_C - T_i)} - 1}{\lambda_3 - 1} \quad (3)$$

where $\sigma_T^* \in [0, 1]$, a user-defined parameter, indicates the reliability degree of the first online review posted in the considered time period. If $t_{li}^T = T_i$, then $\sigma_{li}^T = \sigma_T^*$; if $t_{li}^T = T_C$, then $\sigma_{li}^T = 1$; otherwise, $\sigma_{li}^T \in (\sigma_T^*, 1)$. The closer the value of t_{li}^T to the current time T_C is, the larger the reliability degree σ_{li}^T is. The unit of time can be set to day, month, year, etc. The base of the exponential function $\lambda_3 > 1$ is a user-defined parameter.

- (4) The reliability defined based on the usage period before reviewing.

It is true that new reviews are more responsive to current performance than old ones for short-term services and non-durable products (e.g., food, beverages, cosmetics, daily essentials, etc.), as they can be evaluated immediately after provision or use. However, when it comes to long-term services and durable products (e.g., running shoes, car tires, furniture, appliances, etc.), we can have a mixed picture. Very recent reviews might not be helpful either because the user had no time to test the service/product (e.g., “I bought these running shoes yesterday and they look great. Can’t wait to run a few miles on them”). To resolve this issue, we further measure the gap between the purchase time of a product/service and the publishing time of its review. It can represent the usage period of the user before reviewing. The longer the usage period, the more thorough and accurate the user’s comments will be. In this regard, the reliability of an online review is also related to the usage period before reviewing.

Let t_{li}^P be the purchase time of a_i corresponding to its l -th online review. The usage period before reviewing can be defined as $\mu_{li} = t_{li}^T - t_{li}^P$. Let P be a threshold of usage periods such that the gap between the posting time of most online reviews and the purchase time of the corresponding product/service is less than it. Overall, when only considering the usage period of long-term services and durable products, online reviews posted after a longer period of use are more reliable than those published after a shorter period of use. Similar to the definition of the first two reliability indicator σ_{li}^H and σ_{li}^R , we define the reliability degree of the l -th online review based on the usage period through a logarithm function, that is

$$\sigma_{li}^P = \begin{cases} \sigma_P^* + (1 - \sigma_P^*) \times \frac{\log_{\lambda_4}(\mu_{li} + 1)}{\log_{\lambda_4}(P + 1)}, & \text{if } \mu_{li} < P \\ 1, & \text{if } \mu_{li} \geq P \end{cases} \quad (4)$$

where $\sigma_P^* \in [0, 1]$, a user-defined parameter, indicates the reliability degree of an online review posted immediately after the purchase of a product/service. If $\mu_{li} = 0$, then $\sigma_{li}^P = \sigma_P^*$; if $\mu_{li} \geq P$, then $\sigma_{li}^P = 1$; otherwise, $\sigma_{li}^P \in (\sigma_P^*, 1)$. The larger the value of μ_{li} is, the larger the reliability degree σ_{li}^P is. $\lambda_4 > 1$ is a user-defined parameter.

- (5) The comprehensive reliability.

To define the comprehensive reliability degree of the l -th online review, we need to make a trade-off regarding the four indicators: the quality of online reviews, the credibility of reviewers, the timeliness of online reviews and the usage period before reviewing. In this regard, the weighted averaging operator is used:

$$\sigma_{li} = w_H \sigma_{li}^H + w_R \sigma_{li}^R + w_T \sigma_{li}^T + w_P \sigma_{li}^P \quad (5)$$

where $\sigma_{li} \in [0, 1]$. A larger value of σ_{li} indicates a more reliable online review. $w_H, w_R, w_T, w_P \in [0, 1]$ are the weights of the four reliability indicators, respectively, such that $w_H + w_R + w_T + w_P = 1$. These weights can be determined by decision makers on a case-by-case basis. Especially, for short-term services or non-durable products, we should set $w_P = 0$; otherwise, we set $w_P > 0$.

2.2 Learning attribute importance from online reviews

Most e-commerce platforms, such as Amazon.com, TripAdvisor.com, Taobao.com and JD.com, encourage consumers to post online reviews using star ratings and unstructured text. Star ratings and text reviews contain different values of information. Star ratings visually reflect consumers' overall satisfaction with a product. Text reviews contain a richer amount of information, reflecting consumers' satisfaction with different product attributes. Researches in the field of natural language processing (Hirschberg & Manning, 2015) provided support for the analysis of text reviews. Sentiment analysis aims to extract the tendency and intensity of consumer preferences for product attributes from text reviews, enabling a structured representation of sentiment information.

Stanford CoreNLP (Manning et al., 2014) is one of the most widely used natural language processing toolkits. This open-source toolkit is based on the Stanford CoreNLP lexicon and Recursive Neural Tensor Network for sentiment analysis (Socher et al., 2013). The Recursive Neural Tensor Network is considered as one of the most advanced sentiment analysis models that can analyze the grammatical structure of text and quantify the sentiment information effectively (Baly et al., 2017). The steps of sentiment analysis by Stanford CoreNLP include (Zhao et al., 2021): (1) sentences related to each attribute are extracted from online reviews based on the keywords of each attribute; (2) the grammatical structure, sentiment words and their categories in the sentences are identified for each attribute-related set of sentences; (3) the sentiment intensity distribution of each group of sentences is calculated separately as the sentiment analysis results of the attributes. The sentiment intensities can be divided into five types, including: very negative (VN), negative (N), neutral (Ne), positive (P), and very positive (VP), which constitute a linguistic term set $\{s_1 = \text{VN}, s_2 = \text{N}, s_3 = \text{Ne}, s_4 = \text{P}, s_5 = \text{VP}\}$. The sentiment analysis result of a sentence derives a distribution of five sentiment intensities, where the probability of a linguistic term indicates the closeness of the sentiment word to that linguistic term. Sentiment analysis for an attribute is the analysis of all sentences related to that attribute.

The sentiment intensity of the l -th online review of product a_i regarding attribute c_j can be expressed as

$$S^{lij} = \left\{ s_1 \left(p_1^{lij} \right), s_2 \left(p_2^{lij} \right), \dots, s_5 \left(p_5^{lij} \right) \right\} \quad (6)$$

where $\{p_1^{lij}, p_2^{lij}, \dots, p_5^{lij}\}$ represents the probability distribution of the five linguistic terms. If the l -th online review does not contain any sentiment information about c_j , then $S^{lij} = \emptyset$; otherwise, $p_\alpha^{lij} \geq 0$ for $\alpha = 1, 2, \dots, 5$, and $\sum_{\alpha=1}^5 p_\alpha^{lij} = 1$.

According to the expected utility theory, the utility of S^{lij} can be calculated by (Wu & Liao, 2021)

$$U(S^{lij}) = \sum_{\alpha=1}^5 u(s_\alpha) p_\alpha^{lij} \quad (7)$$

where $u(s_\alpha)$ is the numerical semantics of s_α , satisfying if $\alpha < \beta$, then $0 \leq u(s_\alpha) < u(s_\beta) \leq 1$. Without loss of generality, we can set $u(s_\alpha) = (\alpha - 1)/4$. In this way, the utility of “neutral” is 0.5. $U(S^{lij})$ is a signal of the performance of a_i . The higher the value of $U(S^{lij})$ is, the better the performance of a_i under c_j is, *i.e.*, the more the reviewer is satisfied with this product attribute.

Consumers usually have different preferences for different attributes (Oliveira & Dias, 2020). For example, when choosing a restaurant, some consumers value the taste of the “food”, and some place more importance on the quality of “service”. A consumer’s overall star rating of a product is the result of the trade-off over the performance of the attributes of that product (Wu & Liao, 2021). This trade-off process can be portrayed by a multi-attribute aggregation operator. The consumer’s preferences to different attributes are represented by attribute weights $w_j \in [0, 1]$, $j = 1, 2, \dots, n$. The greater the weight of an attribute, the greater the influence of the performance of that attribute on the overall performance of the product. In addition, the more positive the sentiments of online reviews, the better the attribute performance. Therefore, based on the l -th online review, the performance value of product a_i can be determined by aggregating the sentiment intensities of different attributes through an additive value function according to the multi-attribute utility theory (Keeney & Raiffa, 1993), *i.e.*,

$$U_l(a_i) = \frac{\sum_{j=1}^n \pi_{lj} w_j U(S^{lij})}{\sum_{j=1}^n \pi_{lj} w_j} \quad (8)$$

where π_{lj} is an indicator parameter such that

$$\pi_{lj} = \begin{cases} 1, & \text{if the } l\text{-th online review contains the sentiment information of } c_j \\ 0, & \text{otherwise} \end{cases} \quad (9)$$

The operator $\pi_{lj} w_j / \sum_{j=1}^n \pi_{lj} w_j$ indicates the weight of c_j corresponding to the l -th online review. If the l -th online review does not contain the sentiment information of c_j , then $\pi_{lj} w_j / \sum_{j=1}^n \pi_{lj} w_j = 0$; if the l -th online review only contains the sentiment information of c_j , then $\pi_{lj} w_j / \sum_{j=1}^n \pi_{lj} w_j = 1$; otherwise, $\pi_{lj} w_j / \sum_{j=1}^n \pi_{lj} w_j \in (0, 1)$.

Intuitively, the performance value of a product with a t -star rating should be larger than that of the product with a $t - 1$ -star rating ($\forall t \in \{2, \dots, 5\}$). To relate star ratings to product performance as determined by text reviews, we define a series of classification thresholds, θ_t , $t = 2, \dots, 5$, satisfying $0 \leq \theta_{t-1} < \theta_t \leq 1$, $\theta_1 = 0$ and $\theta_5 = 1$. We can set that $\theta_{t-1} \leq U_l(a_i) < \theta_t$ if a_i is a t -star rating as determined by its l -th online review. Ideally, all online reviews of a product should meet this constraint. Therefore, the goal is to find an aggregation operator (as defined by Eq. (8) with unknown parameters) such that as many online reviews as possible satisfy this constraint. To do so, a pair of estimation errors are introduced to this constraint, *i.e.*,

$$\theta_{t-1} \leq U_l(a_i) - \eta_{il}^+ + \eta_{il}^- < \theta_t \quad (10)$$

where η_{il}^+ is an overestimation error such that if $U_l(a_i) \geq \theta_t$, then $\eta_{il}^+ > 0$; η_{il}^- is an underestimation error such that if $U_l(a_i) < \theta_{t-1}$, then $\eta_{il}^- > 0$. If $\theta_{t-1} \leq U_l(a_i) < \theta_t$, then $\eta_{il}^+ = \eta_{il}^- = 0$.

Model 1 is established to estimate the attribute weights and classification thresholds for product/service a_i ($i \in \{1, 2, \dots, m\}$). The objective function of Model 1 is to minimize the sum of estimation errors. We use the reliability degree of online reviews as a coefficient of

estimation errors. The greater the reliability degree of an online review, the more it needs to satisfy the constraint. In other words, the estimated preference model should reproduce online reviews with high reliability as much as possible.

$$\begin{aligned}
 \textbf{Model 1 } F &= \sum_{l=1}^{L_i} \sigma_{li} (\eta_{li}^+ + \eta_{li}^-) \\
 \text{s.t. } &\begin{cases} \theta_{t-1} \leq U_l(a_i) - \eta_{li}^+ + \eta_{li}^- < \theta_t, \quad t \in \{2, 3, 4, 5\}, \forall l = 1, 2, \dots, L_i \\ w_j \in [0, 1], \forall j = 1, 2, \dots, n \\ 0 \leq \theta_t < \theta_{t+1} \leq 1, \forall t = 1, 2, 3, 4, \theta_1 = 0, \theta_5 = 1 \\ \eta_{li}^+ \geq 0, \eta_{li}^- \geq 0, \forall l = 1, 2, \dots, L_i \end{cases}
 \end{aligned}$$

where σ_{li} is the reliability degree of the l -th online review of a_i . $U_l(a_i)$ is the overall performance value of a_i determined by the l -th text review, which is defined based on attribute weights $w_j, j = 1, 2, \dots, n$. θ_{t-1} and θ_t are two adjacent classification thresholds for defining the performance value of a product/service when its overall star rating is t -star ($t \in \{2, 3, 4, 5\}$). η_{li}^+ and η_{li}^- are overestimation and underestimation errors, respectively. The first constraint of Model 1 defines the performance value of the product based on its text reviews and star ratings. The next three constraints define the value range of attribute weights $w_j, j = 1, 2, \dots, n$, classification thresholds $\theta_t, t = 1, 2, \dots, 5$, and overestimation and underestimation errors $\eta_{li}^+, \eta_{li}^-, l = 1, 2, \dots, L_i$, respectively. They are the decision variables of Model 1. Model 1 is a linear programming model with more decision variables than constraints, hence it always has a solution.

To obtain robust results, we construct multiple datasets to input into Model 1 for preference learning. Each dataset is randomly generated as a subset of the entire dataset consisting of all online reviews of a product. To ensure the validity of the learning results, the number of online reviews in the subset should be large enough (e.g., accounting for 50–100% of the entire set). Let D be the number of generated subsets for a product, and w_j^d be the weight of attribute c_j obtained based on the d -th subsets, $d \in \{1, 2, \dots, D\}$. To combine the results obtained by different subsets, attribute weights are normalized so that the largest weight in each weight vector is 1. The weight of c_j regarding the d -th subsets can be normalized by

$$w_j^{d*} = \frac{w_j^d}{\max_j w_j^d} \tag{11}$$

Through an average operator, the weight of c_j can be estimated by aggregating the results obtained based on all generated subsets, *i.e.*,

$$w_j^* = \frac{1}{D} \sum_{d=1}^D w_j^{d*} \tag{12}$$

The weight of each attribute is between 0 and 1. We can use 0.5 as a benchmark to judge the importance of attributes. If $w_j^* \geq 0.5$, then c_j is an important attribute; otherwise, it is unimportant.

2.3 Estimating attribute performance based on online reviews

This section aims to estimate attribute performance based on online reviews. This is an information fusion problem, that is, integrating the sentiments of different online reviews on the same product attribute to measure its performance. A weighted averaging operator

is used for information fusion, where the weight of each online review is determined by its reliability degree. The more reliable an online review is, the greater its impact on attribute performance measurements.

The overall performance value of product/service a_i regarding attribute c_j can be determined by

$$U(a_i(c_j)) = \frac{\sum_{l=1}^{L_i} \sigma_{li} \times \pi_{li} \times U(S^{lij})}{\sum_{l=1}^{L_i} \sigma_{li} \times \pi_{li}} \quad (13)$$

where $U(S^{lij})$ is the performance value of a_i under attribute c_j determined by the l -th online review. If the l -th online review does not contain any sentiment information of c_j , then this review is invalid for measuring the performance of c_j , *i.e.*, $U(S^{lij}) = \text{null}$, $l \in \{1, 2, \dots, L_i\}$. In this case, c_j does not affect the measurement of product/service performance. σ_{li} is the reliability degree of the l -th online review of a_i . The operator $\sigma_{li} \times \pi_{li} / \sum_{l=1}^{L_i} \sigma_{li} \times \pi_{li}$ is used to determine the weight of the l -th online review, such that the sum of weights is 1. If the l -th online review does not contain the sentiment information of c_j , then its weight is zero.

Given that the utility of “neutral” is 0.5, we can take 0.5 as a benchmark to judge the performance of products. If $U(a_i(c_j)) \geq 0.5$, the performance of product a_i under attribute c_j is positive; otherwise, its performance is negative.

2.4 Framework of the IPA based on online reviews

As a tool for developing marketing strategies, the IPA divides attributes into four quadrants based on consumer perceptions of performance and importance to set priorities for allocating limited resources. These four quadrants are defined as “keep up the good work”, “possible overkill”, “low priority” and “concentrate here” (Martilla & James, 1997). The interpretation for each of the four quadrants is below.

Q1: Keep up with the good work. The attributes are important (*i.e.*, $w_j^* \geq 0.5$) and their performance is positive (*i.e.*, $U(a_i(c_j)) \geq 0.5$). They are the main attributes leading to consumer satisfaction. Managers should maintain their investment in such attributes so that their performance meets consumer expectation.

Q2: Possible overkill. The attributes are not important (*i.e.*, $w_j^* < 0.5$) but perform well (*i.e.*, $U(a_i(c_j)) \geq 0.5$). Such attributes are not key attributes to increase consumer satisfaction. At the same time, reducing the performance of such attributes would not significantly reduce consumer satisfaction. Therefore, investments in such property improvements can be appropriately reduced.

Q3: Low priority. The attributes are not important (*i.e.*, $w_j^* < 0.5$) and do not perform well (*i.e.*, $U(a_i(c_j)) < 0.5$). Such attributes do not have a significant impact on consumer satisfaction, *i.e.*, they are not key factors contributing to consumer dissatisfaction. Improvements to such attributes do not significantly increase consumer satisfaction. Therefore, it is not necessary to increase investment in improving such attributes.

Q4: Concentrate here. The attributes are important (*i.e.*, $w_j^* \geq 0.5$) but do not perform well (*i.e.*, $U(a_i(c_j)) < 0.5$). They are the main attributes that lead to consumer dissatisfaction. Managers should increase their investment in improving such attributes so that they perform to consumer expectation.

3 A case study

This section conducts a case study of hotel improvement to validate the proposed model.

3.1 Data collection

The analysis was done on four five-star hotels located near the Eiffel Tower in Paris, France. The data was collected from TripAdvisor.com, one of the world's leading travel websites. In early June 2022, 7,038 online reviews expressed in English were collected using a Web crawler (<https://www.houyicaiji.com/>). As shown in Table 1, each record contains the overall star rating, text review, time of posting, number of helpful votes given to the review, and number of helpful votes received by the reviewer.

A topic modeling technique, Latent Dirichlet allocation (LDA) (Jelodar et al., 2019), was used to mine the topics of the text reviews as well as the keywords present within each topic. The LDA was built in Python using the Gensim module. However, the topic extraction outputs were not totally compatible with hotel attributes that influence consumers purchase decisions and those for which hotels may implement improvements. By merging the literature (Zhang et al., 2021; Zhao et al., 2020), we manually adjusted the resulting topics and keywords until we had a list of nine topics, including staff attitude (c_1), food (c_2), price (c_3), cleanliness (c_4), location (c_5), decoration (c_6), room (c_7), bathroom (c_8) and recreation (c_9). These topics are hotel attributes considered in this case study. Using Stanford CoreNLP, we obtain the sentiment intensities of text reviews concerning different attributes. The results are listed in Table 2. A blank space in the table indicates that the corresponding attribute is not evaluated in this text review.

We then filter online reviews based on the sentiment analysis results. Firstly, records that do not contain sentiment intensity for any of the attribute are deleted. Secondly, we remove records where the star ratings are extremely inconsistent with text reviews, e.g., an online review has a 5-star rating (or 1/2-star rating) but shows negative (or positive) sentiment for all mentioned attributes. Specifically, according to Eq. (7), the utility of the “neutral” sentiment is approximately 0.5. Intuitively, when the overall rating is 1 or 2 stars, the utility of the corresponding text review should be less than 0.5; when the overall rating is 4 or 5 stars, the utility of the corresponding text review should be greater than 0.5. Considering that there may

Table 1 Basic information about online reviews

| Review | Star rating | Text review | Time | Votes of the review | Votes of the reviewer |
|------------|-------------|---|------|---------------------|-----------------------|
| r_1 | 3 stars | Not the best, but definitely not the worst either... | 2022 | 1 | 101 |
| r_2 | 4 stars | Nice but not luxurious... | 2022 | 0 | 247 |
| ... | ... | ... | ... | ... | ... |
| r_{7038} | 4 stars | Lovely hotel and location. Missing a bit with the service and the prices... | 2017 | 2 | 69 |

Table 2 The sentiment intensity of text reviews concerning different attributes

| Review | c_1 | c_2 | c_3 | c_4 | c_5 | c_6 | c_7 | c_8 | c_9 |
|------------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| r_1 | 0.23 | 0.23 | 0.00 | 0.00 | 0.84 | 0.00 | 0.31 | 0.00 | 0.00 |
| r_2 | 0.68 | 0.44 | 0.00 | 0.00 | 0.36 | 0.00 | 0.67 | 0.53 | 0.00 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| r_{7038} | 0.48 | 0.43 | 0.29 | 0.00 | 0.79 | 0.83 | 0.00 | 0.00 | 0.83 |

Table 3 The distribution of star ratings received by each hotel

| Hotel | Time | Number of reviews | 1-star | 2-star | 3-star | 4-star | 5-star |
|-------|-----------|-------------------|--------|--------|--------|--------|--------|
| H1 | 2017–2022 | 833 | 0.09 | 0.09 | 0.14 | 0.24 | 0.43 |
| | 2012–2016 | 319 | 0.03 | 0.03 | 0.11 | 0.31 | 0.53 |
| H2 | 2017–2022 | 378 | 0.04 | 0.04 | 0.08 | 0.17 | 0.67 |
| | 2012–2016 | 533 | 0.02 | 0.04 | 0.08 | 0.27 | 0.60 |
| H3 | 2017–2022 | 907 | 0.11 | 0.07 | 0.13 | 0.28 | 0.41 |
| | 2012–2016 | 730 | 0.05 | 0.05 | 0.09 | 0.34 | 0.47 |
| H4 | 2017–2022 | 749 | 0.06 | 0.08 | 0.16 | 0.32 | 0.39 |
| | 2012–2016 | 1416 | 0.05 | 0.06 | 0.14 | 0.36 | 0.40 |

Table 4 The frequency of each attribute mentioned in online reviews

| Hotel | Time | c_1 | c_2 | c_3 | c_4 | c_5 | c_6 | c_7 | c_8 | c_9 |
|-------|-----------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| H1 | 2017–2022 | 0.78 | 0.52 | 0.26 | 0.26 | 0.50 | 0.09 | 0.60 | 0.16 | 0.20 |
| | 2012–2016 | 0.79 | 0.54 | 0.28 | 0.25 | 0.59 | 0.13 | 0.54 | 0.17 | 0.25 |
| H2 | 2017–2022 | 0.83 | 0.33 | 0.15 | 0.25 | 0.59 | 0.21 | 0.56 | 0.19 | 0.10 |
| | 2012–2016 | 0.88 | 0.39 | 0.20 | 0.26 | 0.64 | 0.25 | 0.61 | 0.21 | 0.14 |
| H3 | 2017–2022 | 0.76 | 0.47 | 0.26 | 0.29 | 0.55 | 0.16 | 0.71 | 0.19 | 0.14 |
| | 2012–2016 | 0.75 | 0.49 | 0.28 | 0.33 | 0.59 | 0.23 | 0.73 | 0.19 | 0.15 |
| H4 | 2017–2022 | 0.76 | 0.41 | 0.24 | 0.17 | 0.75 | 0.10 | 0.68 | 0.19 | 0.17 |
| | 2012–2016 | 0.75 | 0.39 | 0.27 | 0.16 | 0.76 | 0.11 | 0.71 | 0.19 | 0.16 |

be errors in sentiment analysis and inconsistencies in the performance of different attributes, we set the filtering rules as follows: the record should be removed if the utility of its text review is greater than 0.6 and the overall rating is 1 or 2 stars, or the utility is beyond the interval $[0.3, 0.7]$ and the overall rating is 3 stars, or the utility is smaller than 0.4 and the overall rating is 4 or 5 stars. Through filtering, 5865 records were obtained for the IPA. The statistical information of online reviews is shown in Table 3. Table 4 lists the frequency of each attribute mentioned in online reviews. Attributes c_1 , c_2 , c_5 and c_7 are mentioned more frequently, and the other five attributes (c_3 , c_4 , c_6 , c_8 and c_9) are mentioned less frequently.

3.2 Solutions

Considering the timeliness, we select online reviews from 2017 to 2022 for analysis. The detailed analysis steps are shown as follows:

- (1) Determining the reliability degrees of online reviews.
 - 1) Determining the reliability degrees of online reviews based on the number of helpful votes for online reviews. The third quartile of helpful votes for the online reviews of the four hotels are 2, 1, 2 and 2, respectively. They are set as the thresholds of helpful votes concerning online reviews. Let $\lambda_1 = e$, *i.e.*, the base of natural logarithms. Using Eq. (1), we obtain the first kind of reliability degrees of online reviews.
 - 2) Determining the reliability degree based on the number of helpful votes received by reviewers. The third quartile of helpful votes received by the reviewers of the four hotels are 51, 47, 55 and 85, respectively. They are set as the thresholds of helpful votes concerning reviewers. Let $\lambda_2 = e$. Using Eq. (2), we obtain the second kind of reliability degrees of online reviews.
 - 3) Determining the reliability degree based on the timeliness. Let 2022 be the current time and 2017 be the initial time. Let $\lambda_3 = e$. Using Eq. (3), the third kind of reliability degrees of online reviews published for each year from 2017 to 2022 is obtained as 0.5, 0.56, 0.64, 0.74, 0.86 and 1, respectively.
 - 4) Since hotels are short-term services, we do not consider here the effect of the usage period before reviewing on the reliability of online reviews. It is assumed that the above three reliability indicators have the same importance. The comprehensive reliability of each online review is obtained by aggregating these three indicators using Eq. (5). The results are shown in Table 5.

- (2) Learning attribute importance from online reviews.

The online reviews of each hotel are taken as an entire dataset. Since consumers have diverse preferences for hotel attributes, the reviews they leave vary significantly. Learning a comprehensive consumer preference model from different datasets of online reviews would produce varying outcomes. To avoid one-sided results caused by using a single dataset, we create different subsets of data from the entire dataset for preference learning. In each subset, the online reviews are randomly selected from the entire dataset according to a certain proportion. To ensure that the amount of data utilized for preference learning is sufficient, we set the ratios of online reviews in the subset to the

Table 5 The reliability degrees of online reviews

| Review | Reliability determined based on the quality of online reviews | Reliability determined based on the credibility of reviewers | Reliability determined based on the timeliness of online reviews | Comprehensive reliability degree |
|------------|---|--|--|----------------------------------|
| r_1 | 0.815 | 1 | 1 | 0.938 |
| r_2 | 0.5 | 1 | 1 | 0.833 |
| ... | ... | ... | ... | ... |
| r_{5865} | 1 | 0.979 | 0.5 | 0.826 |

entire dataset to be greater than 0.5. We take into account three different typical ratios: 0.5, 0.7, and 0.9. Approximately 33 subsets were created using each ratio and a total of 100 subsets are generated.

Each subset is input into Model 1 to learn attribute weights. A total of 100 sets of attribute weights are obtained for each hotel. The estimated attribute weights are normalized using Eq. (11). The normalized weights are then aggregated by Eq. (12) to obtain the comprehensive weight of each attribute. The results are listed in Table 6. It indicates that an attribute has different importance for different hotels. For example, c_5 is important for H1 but not important for H2. In addition, from Tables 4 and 6, we can find that the importance of an attribute is not directly related to how often the attribute is mentioned in online reviews. For example, attribute c_8 is not mentioned as frequently in online reviews, yet it is a significant factor in how customers rate hotels. This is due to the fact that attributes that are frequently mentioned by consumers in reviews, such as hotel service, food, rooms, etc. usually are typical attributes of the product/service and are well described. Although a consumer's opinions on a particular attribute of a product/service affect his/her overall evaluation, frequent attributes may not always have a greater influence on that consumer's decisions than infrequent attributes (Zha et al., 2013).

(3) Estimating attribute performance based on online reviews.

The performance values of each hotel regarding different attributes are determined by aggregating the sentiment intensities of online reviews using Eq. (13). The results are shown in Table 7.

(4) Conducting the IPA.

The IPA plots of the four hotels are constructed based on their attribute importance and performance, as shown in Fig. 1. For H1, H3 and H4, c_3 , c_7 and c_8 belong to Q4 (Concentrate here). They are important attributes that influence consumer satisfaction but their performance is negative. For H2, c_3 and c_8 belong to Q4; c_7 belongs to Q1,

Table 6 The comprehensive weight of each attribute estimated from online reviews

| Hotel | c_1 | c_2 | c_3 | c_4 | c_5 | c_6 | c_7 | c_8 | c_9 |
|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| H1 | 0.72 | 0.76 | 1.00 | 0.84 | 0.70 | 0.21 | 0.94 | 0.92 | 0.25 |
| H2 | 0.73 | 0.38 | 0.54 | 0.25 | 0.38 | 0.31 | 0.71 | 1.00 | 0.39 |
| H3 | 0.98 | 0.64 | 1.00 | 0.41 | 0.52 | 0.49 | 0.79 | 0.80 | 0.41 |
| H4 | 0.56 | 0.33 | 0.84 | 0.64 | 0.56 | 0.31 | 0.97 | 1.00 | 0.19 |

Table 7 The performance values of each hotel regarding different attributes estimated from online reviews

| Hotel | c_1 | c_2 | c_3 | c_4 | c_5 | c_6 | c_7 | c_8 | c_9 |
|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| H1 | 0.55 | 0.54 | 0.40 | 0.54 | 0.59 | 0.65 | 0.45 | 0.43 | 0.57 |
| H2 | 0.63 | 0.61 | 0.45 | 0.62 | 0.63 | 0.67 | 0.50 | 0.48 | 0.56 |
| H3 | 0.53 | 0.57 | 0.42 | 0.62 | 0.59 | 0.59 | 0.47 | 0.48 | 0.52 |
| H4 | 0.53 | 0.55 | 0.40 | 0.50 | 0.59 | 0.55 | 0.43 | 0.45 | 0.53 |

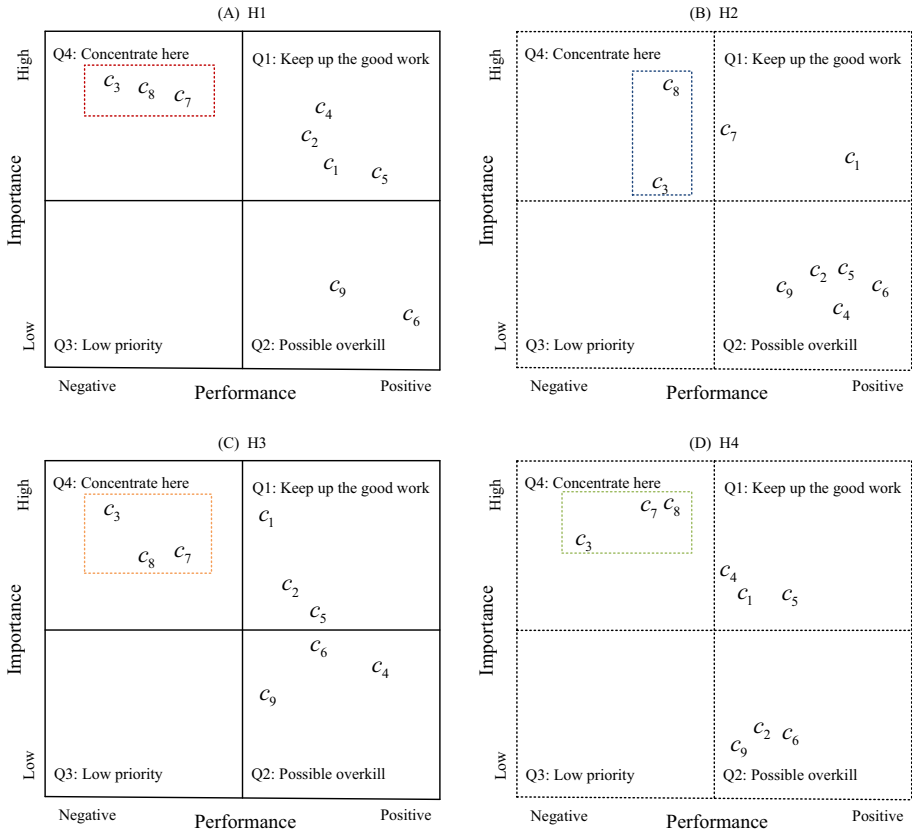


Fig. 1 The IPA plots of the four hotels

but its performance is still not very well. Therefore, the four hotels should increase the investment in attributes c_3 , c_7 and c_8 to improve their performance to enhance consumer satisfaction.

4 Discussions

4.1 The impact of the reliability of online reviews on results

To illustrate the impact of reliability of online reviews on results, attribute importance and attribute performance are determined below without considering the reliability degrees of online reviews. In this case, the objective function of Model 1 is $F = \sum_{l=1}^{L_i} (\eta_{il}^+ + \eta_{il}^-)$, and the constraints do not change. The estimation results are shown in Table 8.

From Table 8, we can find that the results obtained with and without reliability are slightly different. For example, the ranking of attributes in terms of their importance regarding H2 is $c_8 > c_1 > c_7 > c_3 > c_9 > c_2 > c_5 > c_6 > c_4$ when considering reliability, while it is $c_8 > c_1 > c_7 > c_3 > c_2 > c_5 > c_9 > c_6 > c_4$ without considering reliability. In addition, we

Table 8 The comprehensive weight of each attribute estimated from online reviews without considering reliability

| Hotel | c_1 | c_2 | c_3 | c_4 | c_5 | c_6 | c_7 | c_8 | c_9 |
|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| H1 | 0.72 | 0.74 | 1.00 | 0.91 | 0.66 | 0.25 | 0.92 | 0.94 | 0.25 |
| H2 | 0.75 | 0.40 | 0.56 | 0.24 | 0.36 | 0.28 | 0.70 | 1.00 | 0.35 |
| H3 | 0.97 | 0.63 | 1.00 | 0.33 | 0.56 | 0.49 | 0.75 | 0.85 | 0.44 |
| H4 | 0.55 | 0.35 | 0.85 | 0.58 | 0.56 | 0.36 | 0.99 | 1.00 | 0.21 |

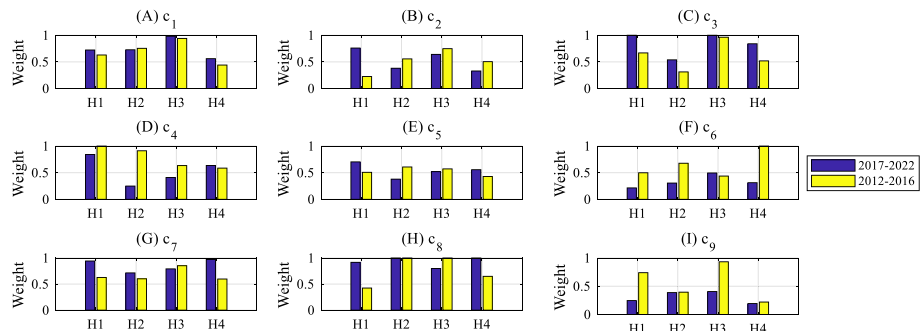
Table 9 The performance values of hotels estimated from online reviews without considering reliability

| Hotel | c_1 | c_2 | c_3 | c_4 | c_5 | c_6 | c_7 | c_8 | c_9 |
|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| H1 | 0.55 | 0.54 | 0.40 | 0.54 | 0.59 | 0.65 | 0.45 | 0.42 | 0.57 |
| H2 | 0.63 | 0.61 | 0.45 | 0.63 | 0.63 | 0.67 | 0.51 | 0.47 | 0.56 |
| H3 | 0.54 | 0.58 | 0.42 | 0.62 | 0.58 | 0.59 | 0.48 | 0.48 | 0.53 |
| H4 | 0.53 | 0.55 | 0.41 | 0.51 | 0.59 | 0.55 | 0.43 | 0.45 | 0.53 |

use an averaging aggregation operator to calculate the performance values of the four hotels regarding different attributes. As shown in Table 9, the results are slightly different from those obtained when the reliability degrees of online reviews are considered. Overall, the reliability of online reviews has an impact on the estimation results of attribute importance and attribute performance.

4.2 Changes in consumer preferences and hotel performance over time

To understand how the consumer preferences and attribute performance of hotels change over time, online reviews posted between 2012 and 2016 are used to estimate attribute importance and performance. The comparison of attribute importance and attribute performance under two time periods, i.e., 2017–2022 (present) and 2012–2016 (previous), is shown in Figs. 2 and 3, respectively.

**Fig. 2** The attribute importance of the four hotels under two time periods

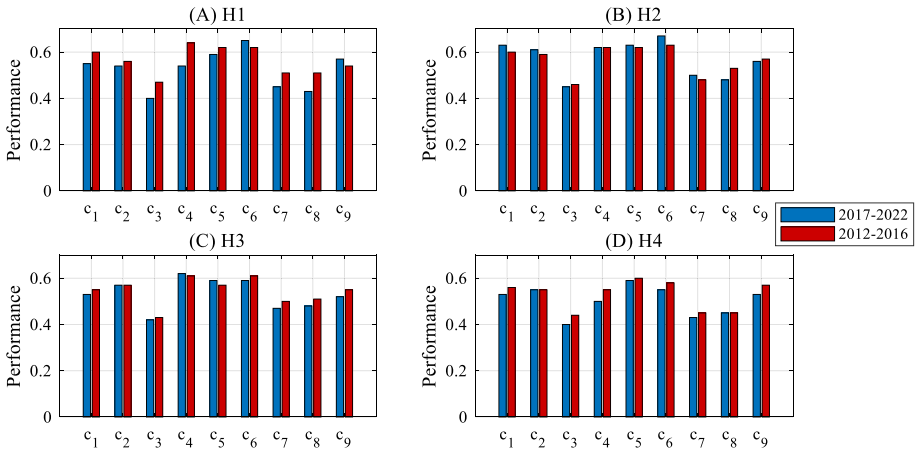


Fig. 3 The attribute performance of the four hotels under two time periods

From Fig. 2, we can find that, except for attributes c_1 and c_5 , the importance of the other attributes changed over time. For example, c_4 used to be an important attribute for H2, but now it is unimportant. Overall, for the four hotels in the present and before, c_1 , c_3 , c_7 and c_8 are relatively important attributes; while c_2 , c_5 , c_6 and c_9 are relatively unimportant attributes.

As can be seen in Fig. 3, except for H2, the current performance of each of the other three hotels is slightly worse than before. We can draw similar conclusions according to their star ratings (See Table 3). For the four hotels, attributes c_3 , c_7 and c_8 are always important, but they perform poorly. This further suggests that hotel managers should focus on improving the performance of these three attributes. It also implies that the attribute performance and importance of hotels can be obtained in a timely manner through online reviews.

5 Conclusion

This paper proposed an IPA framework that considers the reliability of online reviews, aiming to develop product improvement strategies by exploring the relationship between attribute importance, attribute performance and consumer satisfaction. Although online reviews have the advantages of easy collection, large data volume and low cost compared with survey data, the quality of online reviews varies, especially when there are fake online reviews. To address this issue, this paper defined the reliability degrees of online reviews based on the quality of online reviews, the credibility of reviewers, the timeliness of online reviews and the usage period before reviewing. Both measures of attribute importance and attribute performance took reliability into account, allowing online reviews with higher reliability to have a greater impact on decision-making results. We conducted a case study for four five-star hotels using online reviews collected from TripAdvisor.com. We compared the results obtained for two scenarios where reliability was considered and where it was not considered. In addition, the current and previous attribute importance and attribute performance of the hotels were compared.

According to the three-factor theory (Kano, 1984), the importance of an attribute varies with its performance. For example, negative (positive) performance of a basic (excitement)

factor has a greater impact on consumer satisfaction than its positive (negative) performance. In this sense, it would be an interesting research issue for future study to consider the functional relationship between attribute performance and attribute importance when constructing preference models. In addition, to improve the accuracy of the IPA, we can integrate the information obtained from managers, experts and consumers about attribute performance and importance to develop product improvement strategies.

Acknowledgements The work was supported by the National Natural Science Foundation of China (71971145, 72171158, 72301186, 72371173), the Sichuan Planning Project of Social Science (SC22A021), the Sichuan University (2022CX24) and the Fundamental Research Funds for the Central Universities (YJ202250).

Data availability The data that support the finding of this study are available. The original data of the paper are included in this paper.

Declarations

Conflict of interest We wish to confirm that there are no known conflicts of interest associated with this publication.

Ethical approval This article does not contain any studies with human participants or animals performed by any of the authors.

References

- Albayrak, T., Cengizci, A. D., Caber, M., & Fong, L. H. N. (2021). Big data use in determining competitive position: The case of theme parks in Hong Kong. *Journal of Destination Marketing & Management*. <https://doi.org/10.1016/j.jdmm.2021.100668>
- Baly, R., Hajji, H., Habash, N., Shaban, K. B., & El-Hajji, W. (2017). A sentiment treebank and morphologically enriched recursive deep models for effective sentiment analysis in Arabic. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 16(4), 1–21.
- Bi, J. W., Liu, Y., Fan, Z. P., & Zhang, J. (2019). Wisdom of crowds: Conducting importance-performance analysis (IPA) through online reviews. *Tourism Management*, 70, 460–478.
- Bi, J. W., Liu, Y., Fan, Z. P., & Zhang, J. (2020). Exploring asymmetric effects of attribute performance on customer satisfaction in the hotel industry. *Tourism Management*. <https://doi.org/10.1016/j.tourman.2019.104006>
- Bilal, M., & Almazroi, A. A. (2023). Effectiveness of fine-tuned BERT model in classification of helpful and unhelpful online customer reviews. *Electronic Commerce Research*. <https://doi.org/10.1007/s10660-022-09560-w>
- Chen, J., Becken, S., & Stantic, B. (2022). Assessing destination satisfaction by social media: An innovative approach using importance-performance analysis. *Annals of Tourism Research*. <https://doi.org/10.1016/j.annals.2022.103371>
- Dueñas, X., Rodríguez, M., & Pérez, L. M. (2021). Asymmetric importance-performance analysis: Measuring classification changes of destination attributes into basic, performance and excitement factors according to the segmentation criterion. *Tourism and Hospitality Research*, 21(4), 418–425.
- Grigoroudis, E., & Siskos, Y. (2002). Preference disaggregation for measuring and analysing customer satisfaction: The MUSA method. *European Journal of Operational Research*, 143(1), 148–170.
- Hirschberg, J., & Manning, C. D. (2015). Advances in natural language processing. *Science*, 349, 261–266.
- Hu, F., Li, H. X., Liu, Y., & Teichert, T. (2020). Optimizing service offerings using asymmetric impact-sentiment-performance analysis. *International Journal of Hospitality Management*. <https://doi.org/10.1016/j.ijhm.2020.102557>
- Huang, A. H., Chen, K. C., Yen, D. C., & Tran, T. P. (2015). A study of factors that contribute to online review helpfulness. *Computers in Human Behavior*, 48, 17–27.
- Huang, L., Tan, C. H., Ke, W., & Wei, K. (2018). Helpfulness of online review content: The moderating effects of temporal and social cues. *Journal of the Association for Information Systems*, 19(6), 503–522.
- Jelodar, H., Wang, Y. L., Yuan, C., Feng, X., Jiang, X. H., Li, Y. C., & Zhao, L. (2019). Latent Dirichlet allocation (LDA) and topic modeling: Models, applications, a survey. *Multimedia Tools and Applications*, 78, 15169–15211.

- Ji, Y., & Han, J. (2022). Sustainable home meal replacement (HMR) consumption in Korea: exploring service strategies using a modified importance–performance analysis. *Foods*, 11(6), 889. <https://doi.org/10.3390/foods11060889>
- Kano, N. (1984). Attractive quality and must be quality. *Hinshitsu (Quality)*, 14(2), 147e156.
- Keeney, R., & Raiffa, H. (1993). *Decisions with multiple objectives: Preferences and value trade-offs*. Cambridge University Press.
- Korfiatis, N., García-Bariocanal, E., & Sánchez-Alonso, S. (2012). Evaluating content quality and helpfulness of online product reviews: The interplay of review helpfulness vs review content. *Electronic Commerce Research and Applications*, 11(3), 205–217.
- Li, J., Xu, X., & Ngai, E. W. T. (2021). Does certainty tone matter? Effects of review certainty, reviewer characteristics, and organizational niche width on review usefulness. *Information & Management*, 58(8), 103549. <https://doi.org/10.1016/j.im.2021.103549>
- Liu, Y., Bi, J. W., & Fan, Z. P. (2017). Ranking products through online reviews: A method based on sentiment analysis technique and intuitionistic fuzzy set theory. *Information Fusion*, 36, 149–161.
- Lootsma, F. A. (1993). Scale sensitivity in the multiplicative AHP and SMART. *Journal of Multi-Criteria Decision Analysis*, 2(2), 87–110.
- Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., & McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. In: Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, Baltimore, Maryland, America, 55–60.
- Martilla, J. A., & James, J. C. (1977). Importance-performance analysis. *Journal of Marketing*, 41(1), 77–79.
- Mejia, C., Båk, M., Zientara, P., & Orłowski, M. (2022). Importance-performance analysis of socially sustainable practices in US restaurants: A consumer perspective in the quasi-post-pandemic context. *International Journal of Hospitality Management*. <https://doi.org/10.1016/j.ijhm.2022.103209>
- Ngo-Ye, T. L., & Sinha, A. P. (2014). The influence of reviewer engagement characteristics on online review helpfulness: A text regression model. *Decision Support Systems*, 61, 47–58.
- Oliveira, G. D., & Dias, L. C. (2020). The potential learning effect of a MCDA approach on consumer preferences for alternative fuel vehicles. *Annals of Operations Research*, 293, 767–787.
- Phadermod, B., Crowder, R. M., & Wills, G. B. (2019). Importance-performance analysis based SWOT analysis. *International Journal of Information Management*, 44, 194–203.
- Piris, Y., & Gay, A. C. (2021). Customer satisfaction and natural language processing. *Journal of Business Research*, 124, 264–271.
- Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C., Ng, A., & Potts, C. (2013). Recursive deep models for semantic compositionality over a sentiment Treebank. *Empirical Methods in Natural Language Processing*, 120, 1631–1642.
- Sun, S., Fang, D. W., & Cao, J. (2020). Exploring the asymmetric influences of stop attributes on rider satisfaction with bus stops. *Travel Behaviour and Society*, 19, 162–169.
- Wu, X. L., & Liao, H. C. (2021). Modeling personalized cognition of customers in online shopping. *Omega*. <https://doi.org/10.1016/j.omega.2021.102471>
- Wu, X. L., & Liao, H. C. (2022). A Dempster-Shafer-theory-based entry screening mechanism for small and medium-sized enterprises under uncertainty. *Technological Forecasting and Social Change*. <https://doi.org/10.1016/j.techfore.2022.121719>
- Xu, S., Barbosa, S. E., & Hong, D. (2020). Bert feature based model for predicting the helpfulness scores of online customers reviews. *Future of Information and Communication Conference* (pp. 270–281). Cham: Springer.
- Yang, Z. L., Gao, Y., & Fu, X. L. (2021). A decision-making algorithm combining the aspect-based sentiment analysis and intuitionistic fuzzy-VIKOR for online hotel reservation. *Annals of Operations Research*. <https://doi.org/10.1007/s10479-021-04339-y>
- Yeap, J. A., Ignatius, J., & Ramayah, T. (2014). Determining consumers' most preferred eWOM platform for movie reviews: A fuzzy analytic hierarchy process approach. *Computers in Human Behavior*, 31, 250–258.
- Zha, Z. J., Yu, J. X., Tang, J. H., Wang, M., & Chua, T. S. (2013). Product aspect ranking and its applications. *IEEE Transactions on Knowledge and Data Engineering*, 26(5), 1211–1224.
- Zhang, C., Tian, Y. X., Fan, L. W., & Li, Y. H. (2020). Customized ranking for products through online reviews: A method incorporating prospect theory with an improved VIKOR. *Applied Intelligence*, 50, 1725–1744.
- Zhang, C. X., Xu, Z. S., Gou, X. J., & Chen, S. X. (2021). An online reviews-driven method for the prioritization of improvements in hotel services. *Tourism Management*. <https://doi.org/10.1016/j.tourman.2021.104382>
- Zhao, M., Zhang, C. X., Hu, Y. Q., Xu, Z. S., & Liu, H. (2021). Modelling consumer satisfaction based on online reviews using the improved Kano model from the perspective of risk attitude and aspiration. *Technological and Economic Development of Economy*, 27(3), 550–582.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.