International Journal of Hospitality Management 118 (2024) 103684

Contents lists available at ScienceDirect

International Journal of Hospitality Management

journal homepage: www.elsevier.com/locate/ijhm

Determining directions of service quality management using online review mining with interpretable machine learning

Jongkyung Shin [a], Junegak Joung [b,*], Chiehyeon Lim [a,c,*]

[a] Graduate School of Artificial Intelligence, Ulsan National Institute of Science and Technology, Ulsan 44919, the Republic of Korea
[b] School of Interdisciplinary Industrial Studies, Hanyang University, Seoul 04763, the Republic of Korea
[c] Department of Industrial Engineering, Ulsan National Institute of Science and Technology, Ulsan 44919, the Republic of Korea

ARTICLE INFO

Keywords:
Service management
Feature importance
Interpretable machine learning
Explainable artificial intelligence
Customer reviews
Customer needs

ABSTRACT

Determining the importance values of service features is necessary to prioritize the points in service quality management and improvement. Existing studies have used linearly additive relationship models to estimate service feature importance, such as linear and logistic regression. This traditional approach is interpretable but often limited in terms of model fitness and prediction performance. Meanwhile, modern advanced machine learning models provide high fitness and performance but often lack interpretability. Thus, to achieve both reliable prediction and interpretation, we propose a systematic framework for estimating the importance of service features using online review mining with interpretable machine learning. An interpretable machine learning-based method is proposed to estimate the importance values of features by applying the shapley additive global importance metric to the highest-performance prediction model. We validate the superiority of our framework over existing methods through a case study on the global importance estimation of hotel service features in Singapore. To facilitate additional applications, we offer the implementation code of our work at https://github.com/JK-SHIN-PG/OnReviewServImprovement.

1. Introduction

Determining the importance values of service features is necessary to prioritize the points in service quality management and improvement (Bi et al., 2019b; Joung and Kim, 2021; Palese and Usai, 2018; Xu et al., 2017). Traditionally, the importance values of different features were estimated by statistically analyzing the relative effects of features on the overall customer satisfaction using linear additive relationship models, such as linear and logistic regression (Decker and Trusov, 2010; Farhadloo et al., 2016; Suryadi and Kim, 2018). Numerous studies have used this traditional approach of feature importance estimation for specific service fields, such as tourism (Deng, 2007) telecommunication (Pezeshki et al., 2009) and healthcare (Izadi et al., 2017). However, the relationship between customer perceptions of service features and customer satisfaction was found to be nonlinear (Deng et al., 2008; Matzler et al., 2003, 2004). Although linear regression using dummy variables can be alternatively used to accommodate a nonlinear relationship (Anderson and Mittal, 2000; Mittal et al., 1998; Ting and Chen, 2002), the prediction errors are high when real-world data do not follow a specific distribution or a simple functional relationship. Modern neural network (NN)-based models may yield a smaller prediction error than traditional models; however, the interpretability of this case is insufficient owing to the use of a complex mechanism of learning or assumptions in learning (Molnar, 2020). In short, there is a trade-off between interpretability and predictability in the importance estimation of service features.

A large amount of data is a prerequisite to estimate parameter values for reliable prediction and to calculate metric values for interpretation across many cases (Covert et al., 2020; Lundberg and Lee, 2017). Online service platforms offer such datasets. As customer-generated review records in online service platforms contain various "voices of the customer", the key resource for customer understanding and service quality management (Griffin and Hauser, 1993), online review mining is used as an efficient approach for service firms to transform a large number of raw voices of customers into rich insights on customers' perceptions of their services (Fernandes et al., 2022; Girardin et al., 2021; Lee and Huang, 2009; Ye et al., 2020). Moreover, online reviews can be collected and analyzed automatically and regularly, allowing firms to continuously capture evolving customer perceptions and preferences for service quality management and improvement (Culotta and

* Corresponding authors.
E-mail addresses: shinjk1156@unist.ac.kr (J. Shin), june30@hanyang.ac.kr (J. Joung), chlim@unist.ac.kr (C. Lim).

https://doi.org/10.1016/j.ijhm.2023.103684
Received 16 June 2023; Received in revised form 7 December 2023; Accepted 27 December 2023
Available online 13 January 2024

Cutler, 2016; Kim and Lim, 2021; Rese et al., 2014). As such, considerable number of studies have recently addressed methods for online review mining, such as feature identification through topic modeling (Alzate et al., 2022; Joung and Kim, 2020; Kim and Lim, 2021; Mejia et al., 2021; Wang et al., 2018), sentiment analysis (Bi et al., 2019a; Jeong et al., 2019; Nasiri and Shokouhyar, 2021; Wu and Chang, 2020; Zhang et al., 2023), prediction model development for service management (Decker and Trusov, 2010; Farhadloo et al., 2016; Nilashi et al., 2021, 2022).

However, as we aforementioned in the first paragraph, what firms actually need, but what the current literature of online review mining for service quality management has yet to address in depth, is to accurately estimate the extent to which focal service features affect customer satisfaction (Bi et al., 2019b; Luo and Tang, 2019; Wu et al., 2023), i.e., feature importance values, such that they can interpret it to improve the service eventually. Thus, to address this research gap and efficiently achieve both high prediction performance and interpretability, we propose a systematic framework for estimating the importance of service features reliably with interpretable machine learning (IML) using online customer review data. The proposed framework comprises (1) data collection and preprocessing, (2) feature identification and dataset preparation for prediction model development, and (3) global importance estimation of the service features using optimal prediction model. First, online customer reviews of the target services are collected from online platforms. Second, service features and sentiments from online reviews are identified using the latent Dirichlet allocation (LDA) and the valence aware dictionary and sentiment reasoner (VADER) sentiment analysis, respectively. A dataset with the identified features and determined sentiment scores is used to develop the highest-performance prediction models by applying modern advanced machine learning models, when considering the nonlinear relationships among the service features and the overall customer satisfaction. Finally, an IML-based method is proposed to estimate the importance values of features using the shapley additive global importance (SAGE) method, which considers various interactions among features. We validated the superiority of our framework over existing methods through a case study on the global importance estimation of hotel service features in Singapore. This case study focused on hotels with low customer satisfaction that must improve their services effectively and efficiently. Reviews with ratings less than 4 are associated with problematic services and customer dissatisfaction, making them significant in addressing service improvement (Zhu et al., 2021). Therefore, analyzing hotels with many of these reviews is beneficial for identifying problematic features of service quality. In line with this, our case study analyzed 32,044 reviews of hotels with overall ratings (i.e., average star ratings) of less than 4, and reliably estimated the importance of service features.

To the best of our knowledge, this study is the first academic attempt to estimate accurately the extent to which focal service features affect customer satisfaction using online review mining with IML. While more customer reviews become available in online service websites and platforms (Nilashi et al., 2021; Zhang et al., 2021), our work shows the great potential of IML as a new research methodology to use these review resources for service studies in academia and service management in practice. Both the methodological and practical contributions of this study were validated through a real-world application to hotel services in Singapore. The light gradient boosting machine (LGBM) with SAGE selected in the case study yielded reliable importance values for the service features with better prediction performance and higher interpretability than existing methods, and the total runtime for estimating the importance values from 32,044 reviews was approximately 3 h on a PC with Intel i9–9880 H CPU, 16 GB of RAM, and the Windows operating system. The outcome of our case study was validated and confirmed by the customer survey questionnaires used in a hotel reservation platform. Thus, with the proposed framework, researchers and managers can easily monitor, manage, and improve the quality of their focal services based on online review mining. For such users, we offer the implementation code of our work at https://github.com/JK-SHIN-PG/OnReviewServImprovement.

For further details, the remainder of this paper is organized as follows. Section 2 describes related studies on customer-oriented service quality management, importance estimation using online reviews, and IML techniques. Section 3 describes the proposed framework. Section 4 presents the real-world applications of the framework. Section 5 discusses the use of the proposed framework for service improvement and the strengths and weaknesses of our work with customer survey questionnaires. Finally, Section 6 provides concluding remarks and presents future research directions.

2. Related work

This section describes previous studies and limitations regarding service quality management and an importance estimation and presents different IML techniques.

2.1. Customer-oriented hospitality service quality management

Understanding customer perceptions of service quality is essential for improving service (Johnson and Gustafsson, 2006; Olsen et al., 2014). To understand customer perceptions of a service, Parasuraman et al., (1985, 1988) identified service quality dimensions through interviews and developed a service quality measurement method called SERVQUAL. Subsequent studies adapted SERVQUAL for various hospitality domains, including international tourist hotels (Chen, 2013), business hotels (Akbaba, 2006), and Airbnb accommodations (Ding et al., 2020). Improvements are frequently prioritized using feature importance estimation, which is traditionally performed through surveys and interviews (Albayrak, 2015; Cheng et al., 2012; Choi and Chu, 2001; Martilla and James, 1977; Mejia et al., 2022; Mohsin et al., 2019). However, these approaches are time-consuming for designing questionnaires and can easily become outdated, providing limited practical assistance in rapidly changing service sectors (Lee and Huang, 2009; Kim and Lim, 2021; Mejia et al., 2021). As another form of customer information, online customer reviews can serve as valuable real-time data sources for capturing and monitoring customer perceptions (Culotta and Cutler, 2016; Kim and Lim, 2021). Prior studies on hospitality services have analyzed reviews to identify the main features that customers focus on (Ding et al., 2020; Wang et al., 2018). After feature identification, review data are utilized to derive customer sentiments to analyze service performance (Chang et al., 2019; Qiao et al., 2022; Luo et al., 2021; Luo and Xu, 2021) and to measure feature importance (Kwon et al., 2020; Luo et al., 2021; Ye et al., 2020).

It is notable that in the above investigations, the processes of feature identification, sentiment analysis, and measurement of feature importance using review data were conducted in a segmented manner instead of simultaneously. However, what firms actually need in practice is to conduct these tasks all at once in a consistent manner. Based on this, recent studies have proposed an integrated framework that simultaneously performs the entire process using online customer reviews for service quality management and improvement (Nie et al., 2023; Pan et al., 2023). Despite this advancement, these frameworks remain limited, and related studies remain scarce. Therefore, this study proposes a more advanced integrated systematic framework than the above that utilizes online review mining with IML for realizing customer-oriented hospitality service quality management.

2.2. Feature importance estimation with machine learning

Numerous studies have used surveys to obtain data for the importance estimation, while a few studies have estimated the importance of service or product features from online reviews using linear and logistic regression, choice models, conjoint analysis, and a Bayesian method based on linear models (Decker and Trusov, 2010; Farhadloo et al.,

2016; Nilashi et al., 2021; Suryadi and Kim, 2018; Zhang et al., 2021). In these linear models, the magnitude of the coefficients of features was regarded as importance features with normalization. However, the methods employed in the previous studies for importance estimation are inadequate for determining the nonlinear relationships among the sentiments of service or product features and overall customer satisfaction with interpretability. Prior studies have assumed that the overall level of customer satisfaction is a linear combination of sentiments of service features, and that these features are independent (Izadi et al., 2017; Mejia et al., 2021). The linearity assumption between the sentiments of features and overall customer satisfaction has been demonstrated to be inaccurate in various studies (Deng et al., 2008; Matzler et al., 2004). Specifically, prior studies in the hospitality and tourism domains have revealed that service features and overall satisfaction have a nonlinear or asymmetric relationship (Caber et al., 2013; Lai and Hitchcock et al., 2016; Ramanathan and Ramanathan, 2016; Slevitch and Oh, 2010). Features can also interact according to the various order of features (Molnar, 2020). Although an ensemble NN-based method (Bi et al., 2019b) can identify nonlinearity, a lack of explaining the influence of features on overall customer satisfaction exists because the corresponding machine learning model involves a black-box learning mechanism. In case of complex interactions among features, their importance values obtained using this method have low reliability (Joung and Kim, 2021).

Therefore, this study aims to fill the above gap in the feature importance estimation literature by proposing a framework with IML to estimate the importance values of service features on overall customer satisfaction using online reviews. The proposed framework with IML is distinct from previous studies in that it determines the nonlinear relationships among the sentiments of service features and overall customer satisfaction, and that it estimates the importance of each feature in all possible interactions of the features.

2.3. IML techniques

IML techniques involve methods and models that provide understandable explanations to humans in prediction of machine learning systems (Molnar, 2020). Generally, IML can be classified into intrinsically interpretable models and model-agnostic interpretation methods (Du et al., 2019). Intrinsically interpretable models are human-understandable models owing to their simple structure. For example, a decision tree (DT), which is an intrinsically interpretable model, provides a tree for classifying instances; if the tree structure is short, humans can determine the classification process. In this tree structure, the importance of a feature can be calculated by summing the number of splits in the tree that contain the feature for classification, which is called Gini importance or mean decrease impurity (Louppe et al., 2013). Meanwhile, modern advanced machine learning models, such as random forest (RF), LGBM, and NN models, which have exhibited high classification performance in recent years, are not self-explanatory; therefore, model-agnostic interpretation methods are needed after building such models.

Model-agnostic interpretation methods refer to techniques for interpreting already built machine learning models. The importance of each feature in a machine learning model can be derived by various model-agnostic interpretation methods, such as feature ablation, permutation, shapley additive explanations (SHAP), and SAGE (Molnar, 2020). Feature ablation calculates the importance of each feature by estimating the decrease in the predictive performance of the model when each feature is removed. Features that significantly degrade performance when absent are considered highly important. However, this method does not consider the interactions between features, resulting in inaccurate feature importance estimation for correlated data. Permutation estimates the importance of each feature by calculating the influence on the prediction error of a model when shuffling each feature value (Breiman, 2001). Features that highly increase the prediction error of a model after shuffling have high importance. Although this method considers all interactions between features, it can yield biased estimates of feature importance using shuffled data that do not reflect realistic relationships. Particularly, when some features are correlated, the importance of the associated features can be underestimated by dividing the importance between the features. Additionally, this method lacks consistency because the feature importance may vary with different permutations. SHAP and SAGE calculate feature importance by measuring changes in the results of the model when a feature is included, instead of excluded, across all possible feature subsets, considering all interactions (Covert et al., 2020; Lundberg and Lee, 2017; Molnar, 2020). Compared with other model-agnostic interpretation methods, SHAP and SAGE have a theoretical foundation in game theory, which ensures that importance is unbiasedly distributed among the features (Covert et al., 2020; Lundberg and Lee, 2017). In addition, this method is robust for estimating feature importance, even for correlated data, because it considers interactions, and the results are consistent. The difference between SHAP and SAGE is that SHAP explains the effects of features on individual predictions (i.e., local interpretability), whereas SAGE calculates the importance of a feature to the prediction performance of a model in the entire dataset (i.e., global interpretability). SAGE is also more efficient in calculating the importance of a feature than SHAP (Covert et al., 2020).

Therefore, this study employs SAGE for global importance estimation, which efficiently provides a unique solution for interpreting high-performance machine learning models while considering the interactions among all possible features. This study is the first attempt to use IML to estimate the importance values of hotel service features.

3. Framework

The overall framework for estimating the importance values of service features from online reviews is presented in Fig. 1. Online reviews of a target service are the inputs, and the outputs are the importance values of the sentiments of the service features on the overall customer satisfaction. The proposed framework includes three stages: data collection and preprocessing, feature identification and dataset preparation for prediction model development, and global importance estimation using optimal prediction model. The IML-based method is proposed to estimate the importance of features based on SAGE.

3.1. Data collection and preprocessing

Web scraping is used to obtain online customer reviews from well-known review websites, such as Amazon, Yelp, and TripAdvisor. The contents and star ratings of the reviews are collected along with information such as title, date, and service category. Duplicate reviews in the collection are removed by checking the user ID, title, and content. The online reviews are structured into preprocessed words with a part-of-speech (POS) tagging and original sentences with emojis, emoticons, and punctuation. Text preprocessing with POS tagging proceeds as follows (Boyd-Graber et al., 2014). An uppercase is converted to a lowercase, and punctuation and stop words are eliminated. Words are lemmatized, and words that occur either very frequently or very rarely are removed.

3.2. Feature identification and dataset preparation for prediction model development

The features that customers frequently mention in the online reviews and their sentiments are determined to prepare a dataset for the global importance estimation. LDA is employed to identify service feature words because it has been used in numerous studies to obtain feature words from online reviews (Bi et al., 2019b; Joung and Kim, 2020; Wang et al., 2018). LDA is a powerful probabilistic topic model that summarizes massive textual data by finding hidden topics (Blei et al., 2003). In

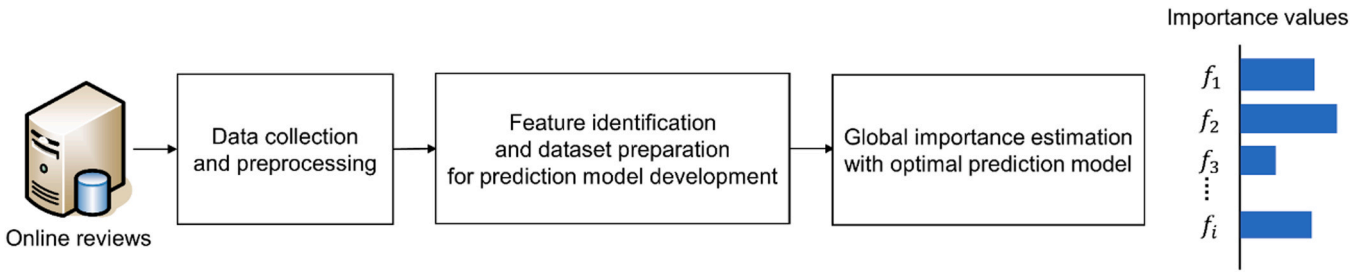

Fig. 1. Overall process of proposed framework.

LDA, it is assumed that each review is a mixture of a set of topic probabilities, and each topic is a mixture of a subsequent set of words. A review-noun matrix is generated as the input for LDA because service feature words are assumed to be nouns (Guo et al., 2009; Hu and Liu, 2004; Joung and Kim, 2021; Suryadi and Kim, 2018). Perplexity, topic dissimilarity, and topic coherence can be used to determine the number of topics in the LDA results. The output of LDA is a topic-noun matrix. Each topic is named by interpreting the nouns in the topic. The label of each topic can be regarded as a feature of a service (Bi et al., 2019b; Joung and Kim, 2021; Wang et al., 2018). Additionally, the nouns of each feature can be extended by considering synonyms via WordNet (Miller, 1995) or word embedding (Mikolov et al., 2013). Some open-source libraries or softwares, such as the Gensim library of Python and the Stanford Topic Modeling Toolbox, can be utilized for this task.

VADER sentiment analysis is used to estimate the sentiments of the service features, which are identified from the reviews. VADER sentiment analysis is an unsupervised machine learning model based on lexicons and rules to measure the sentiments of social media texts (Hutto and Gilbert, 2014). This model can be easily employed in other fields because manually labeling the training data is unnecessary. For this reason, previous hospitality service studies have utilized this model to derive the sentiment of reviews (Kostromitina et al., 2021; Luo et al., 2021). In this study, the VADER sentiment analysis yields the polarity intensity after extracting the original sentences containing the service features from the reviews as follows. First, the affective lexicons and their intensities in each sentence are assigned values, − 4 (extremely negative valence) and 4 (extremely positive valence), using well-established word banks and five generalizable heuristics. Second, the overall intensity is calculated by averaging all affective lexicon scores, followed by normalization to − 1 (extremely negative) and 1 (extremely positive). For example, in the sentence, "Overall this place is great for its location and prices," the polarity intensities of "location" and "price" features are 0.6249 each based on the VADER sentiment analysis. Finally, the polarity intensities of the service features in a review are calculated by averaging the polarity intensities of the sentences and including them in the review.

Let S_{im} denote the sentiment score of the i^{th} feature (f_i) in the m^{th} review, where $i = 1, 2, \ldots, I$ and $m = 1, 2, \ldots, M$. To prepare the dataset for the prediction model development, the polarity intensities of the identified service features are transformed into five labels using Eq. 1 (Table 1). We utilize these transformed data to avoid mislearning in the model owing to the variations in sentiment intensity arising from subtle differences in the customer expressions (Bi et al., 2019b; Liu et al., 2017). The range of polarity intensity is categorized into five labels, extending the tripartite classification of positive, neutral, and negative sentiments proposed by Hutto and Gilbert (2014). In their approach, sentiments are classified using threshold values set as − 0.05 and 0.05 (Hutto, 2022). To consider the differences in the polarity intensity of service features reflected in the customer expressions, we introduce two additional labels, "very positive" and "very negative," referring to Bi et al. (2019b). These labels are assigned values of "4" and "1," respectively. For instance, although both "Service is not bad." and "Service is excellent!" are positive reviews, the latter expresses a more positive sentiment. Accordingly, in the latter case, the "service" is assigned a score of 4 (very positive). To facilitate this, we divide the positive and negative polarity ranges into two and allocate "very positive" and "very negative" labels to each high-intensity polarity. Consequently, we set the thresholds at 0.525 and − 0.525. Concurrently, comments containing evident sentiments about service features significantly affect customer satisfaction. However, comments that describe a situation or place are typically written with neutral sentiments and can interfere with learning about the relationships between service features and overall satisfaction. Therefore, service features with neutral sentiment intensity are assigned a value of "0," to not affect the results of the model, similar to dealing with missing service features (Bi et al., 2019b). The star ratings are converted into 0 (i.e., negative: 1, 2, and 3 stars) and 1 (i.e., positive: 4 and 5 stars) because the performance of the prediction model is low when using five-star ratings (Joung and Kim, 2021, 2023). Based on Table 1, the sentiment scores of the determined service features are used as input variables and the star rating (i.e., overall customer satisfaction) is used as the output variable, to build the prediction model for global importance estimation.

$$S_{im} = \begin{cases} 4, \ if \ 0.525 \leq sentiment\ intensity \leq 1 \\ 3, \ if \ 0.05 \leq sentiment\ intensity < 0.525 \\ 0, \ if \ -0.05 < sentiment\ intensity < 0.05 \\ 2, \ if \ -0.525 < sentiment\ intensity \leq -0.05 \\ 1, \ if \ -1 \leq sentiment\ intensity \leq -0.525 \end{cases} \quad (1)$$

Table 1
Dataset for global importance estimation.

Review	f_1	f_2	⋯	f_I	Star rating
1		4	⋯		1
2	2	2	⋯		1
3		4	⋯	4	1
⋮	⋮	⋮	⋮	⋮	⋮
M			⋯		0

3.3. Global importance estimation using optimal prediction model

IML-based method is proposed to estimate the importance values of the identified service features on the star ratings. Let $D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_M, y_M)\}$, where $m = 1, 2, \ldots, M$, denote the dataset for building the prediction model (Table 1). Let $x_m = (S_{1m}, S_{2m}, \ldots, S_{Im})$ and y_m denote the sentiment scores of the service features and the corresponding star rating of the m^{th} review, respectively. The proposed method is described below.

3.3.1. Preparing K training and test sets

K-fold cross-validation is performed to randomly split review data *D* into *K* equal-sized training and test sets (Hastie et al., 2009). This method can solve the variance problem of a classifier that is derived from one randomly selected training set because all observations are used in both training and testing. *K*-1 sub-samples are used as the training sets, and the remaining single sub-sample is used as the test set to validate the classifier.

3.3.2. Building K optimal classifiers

From K training sets, K optimal classifiers are built to predict the star ratings based on the sentiment scores of service features. Various machine learning models, such as NN with a hidden layer, DT, RF, and LGBM, can be considered to construct optimal classifiers. A neural network with a hidden layer is selected as the neural network architecture owing to the available low-dimensional data (Bi et al., 2019b; Deng et al., 2008; Mikulić and Prebežac, 2012). For hyperparameter tuning, in each machine learning model, various search algorithms, such as grid search, random search, and genetic algorithms, can be used. To determine the prediction performance of a classifier on the test set, evaluation metrics, such as precision, recall, accuracy, and F-1 score, are utilized.

3.3.3. Estimating K SAGE values of service features on star ratings from K classifiers

The SAGE method is used to estimate the K importance values of service features from K optimal classifiers (Covert et al., 2020). SAGE is a unified approach for computing the global importance of a feature in the entire dataset based on game theory. Let $\phi_i(p)$ denote the SAGE value of feature i in the prediction performance of the classifier p. Let $|N|$ and $|C|$ denote set of all features and all feature subsets, respectively. Let $p(C \cup i)$ and $p(C)$ denote the contributions of the set of features with order and feature i and of features with order, respectively. The SAGE value satisfies the desirable properties, such as efficiency, symmetry, dummy, and consistency.

Property1 (Efficiency) $\sum_{i=1}^{I} \phi_i(p) = p(N) - p(\varnothing)$, SAGE value sum to the contribution of the set of all the features over an empty set.

Property2 (Symmetry) If $p(C \cup i) = p(C \cup j)$ for all C, then $\phi_i(p) = \phi_j(p)$.

Property3 (Dummy) If $p(C \cup i) = p(C)$ for all C, then $\phi_i(p) = 0$.

Property4 (Consistency) If $p(C \cup i) - p(C) \geq p'(C \cup i) - p'(C)$, then $\phi_i(p) \geq \phi_i(p')$.

The SAGE value of each feature is calculated by the weighted averaging of the marginal contributions of each feature over all possible orders of the features (Eq. 2). The contribution of each feature in the prediction model is calculated by the difference in prediction performance of the model when the feature is missing. For regression tasks, the difference is estimated using the Euclidean distance, and for classification tasks, the difference is estimated using the Kullback-Leibler (KL) divergence.

$$\phi_i(p) = \sum_{C \subseteq N : i \notin C} \frac{|C|!(|N| - |C| - 1)!}{|N|!} (p(C \cup i) - p(C)) \tag{2}$$

Let $SAGE_{ik}$, where $k = 1, 2, \ldots, K$, denote the SAGE value of feature i in the k^{th} classifier. The K SAGE values of each feature are calculated from the K optimal classifiers.

3.3.4. Combining K SAGE values

Let $\widehat{SAGE}_i$ denote the SAGE value of feature i in the fused models. Let $\overline{w_k}$ denote the normalized performance on the test set of the k^{th} classifier, which is calculated as $\overline{w_k} = \frac{w_k}{\sum_{k=1}^{K} w_k}$. $\widehat{SAGE}_i$ is calculated using Eq. 3.

$$\widehat{SAGE}_i = \sum_{k=1}^{K} \overline{w_k} \widehat{SAGE}_{ik} \tag{3}$$

Finally, the SAGE value of each feature is normalized using Eq. 4.

$$\overline{SAGE_i} = \frac{\widehat{SAGE}_i}{\sum_{i=1}^{I} \widehat{SAGE}_i}, i = 1, 2, \ldots, I \tag{4}$$

4. Case study of estimation of global importance of hotel service features to customer satisfaction

A case study of the estimation of global importance of hotel service features to customer satisfaction was conducted to validate the proposed framework. The hotel industry is selected because it provides customer segment information (e.g., hotel class and travel type) with numerous reviews.

4.1. Results

4.1.1. Collecting data and preprocessing

In Singapore, which is a popular tourist destination, customer reviews were collected from TripAdvisor for hotels with overall ratings (i.e., average star ratings) of less than 4 because hotels with low customer satisfaction require improving their services effectively and efficiently. We analyzed reviews with ratings ranging from 1 to 5 for these hotels. Reviews with ratings of less than 4 are important for identifying service improvement scenarios, as they highlight problematic services and indicate customer dissatisfaction. In addition, reviews with ratings of 4 and 5 are important because they frequently highlight the unique service features of hotels, aiding in the identification of service features. Moreover, these higher-rated reviews are pivotal for training the relationship between service features and customer satisfaction in developing prediction models. The ratio of the positive (i.e., 4 and 5) and negative (i.e., 1, 2, and 3) star ratings is balanced. This balanced ratio of star ratings was effective to identify the sentiments of the service features that affect the star ratings (Joung and Kim, 2021). After removing duplicate reviews, 32,044 English reviews were collected from January 2010 to December 2019 based on the dates of stay (Table 2). The customer reviews for hotels in 2020 were not considered because a different pattern of customer needs could arise due to Covid-19. The dataset included customer reviews for hotel classes, such as 2, 3, and 4 classes and travel types, such as solo, couple, friends, family, and business. "Ratio" represents the ratio of positive reviews (i.e., 4 and 5 star ratings) in all reviews. The most customer reviews were for class 4 hotels among the hotel classes and for couples among traveler types. The Selenium web crawler library of Python was used to collect the customer reviews from TripAdvisor. The NLTK package of Python was used to structure each review into preprocessed words with POS and original sentences.

4.1.2. Identifying hotel service features and preparing dataset for prediction model development

The hotel service features that customers mainly mention were identified using LDA. A 33,132·5933 review-noun matrix was prepared as the input matrix for the LDA, where the LDA parameters, alpha and beta, were 0.1 and 0.01, respectively (Wang et al., 2018). Nouns can be considered as parts of speech representing service features (Bi et al., 2019a; Jeong et al., 2019; Joung and Kim, 2021), and 5933 nouns were selected by identifying nouns appearing in service catalogs of hotels for which reviews were collected. The optimal number of topics was selected as nine based on the maximum topic coherence (0.54). A total of nine hotel service features were determined by identifying the logical connections among the top-30 words and typical reviews (Joung and Kim, 2020) (Table 3). In the Table 3, "Frequent words" are arranged in descending order according to each topic and the associated probability obtained from the LDA. "Number of words" represents the number of words related to each feature by identifying the synonyms using word2vec. "Number of reviews" represents the number of reviews containing the words related to each feature. Among the nine hotel service features, those mentioned more than 20,000 by customers were "location" and "service," and the features mentioned less than 10,000 were "view" and "internet." The Gensim library of Python was used for LDA and word2vec. Using Eq. 1 and Table 1, the sentiment scores of the hotel service features were determined from 0 to 4. The statistical analysis of the sentiment scores of the hotel service features is shown in

Table 2
Collected data information based on customer segments.

Information	All	Hotel classes			Travel types				
		Class2	Class3	Class4	Solo	Couple	Friends	Family	Business
Sample	32,044	4571	12,528	14,945	3318	9853	3593	7812	4892
Ratio	0.56	0.51	0.54	0.60	0.59	0.56	0.58	0.60	0.50

Table 3
Hotel service features in Singapore.

	Feature	Frequent word	# of words	# of reviews
f_1	Location	Location, ⋯	63	26,700
f_2	View	View, outlook, ⋯	15	6527
f_3	Breakfast	Breakfast, buffet, ⋯	24	13,484
f_4	Sleep quality	Bed, mattress, ⋯	20	10,707
f_5	Bathroom	Bathroom, toilet, ⋯	24	11,466
f_6	Service	Service, staff, ⋯	32	20,864
f_7	Check	Check, checkin, ⋯	19	12,651
f_8	Value	Value, price, ⋯	6	11,477
f_9	Internet	Internet, wifi, ⋯	32	6137

Fig. 2. In the hotel service features, there were more positive emotions (e.g., 3 and 4) than negative emotions (e.g., 1 and 2). The VADER library of Python[1] was used to determine the sentiment scores of the hotel service features.

4.1.3. Estimating global importance values of hotel service features using optimal prediction model

The importance values of the hotel service features to star ratings were estimated using the proposed IML-based method. They were measured from nine datasets based on the hotel class and traveler type. Five-fold cross-validation was performed on each dataset based on the Pareto principle (80% training set and 20% test set). Five classifiers were built from five training sets, and the prediction performance of the classifier was obtained using five test sets. Accuracy was used as the prediction performance metric due to balanced class (Bekkar et al., 2013). Machine learning models, i.e., NN with a hidden layer, DT, RF, and LGBM, were adopted to derive optimal classifiers with the best accuracy. After determining the range of hyperparameters from a preliminary experiment, the optimal classifiers were derived by a grid search for hyperparameter tuning (Table 4). On the nine datasets, LGBM presented the best performance among all models. The scikit-learn package of Python was used to implement the NN with a hidden layer, DT, and RF models. The lightgbm library of Python was used to implement the LGBM. SAGE[2] was used to estimate the importance values of the hotel service features in each optimal classifier. The five SAGE values of the hotel service features were combined into an importance value using Eqs. 3 and 4. The resulting importance values of the nine hotel service features in the nine datasets are listed in Table 5. Among the nine hotel features, "service" and "bathroom" are generally more important than the other features in the nine datasets. The importance values of the hotel service features in different customer segments are slightly different. Excluding "service" and "bathroom," the customers of class 2 hotels value "sleep quality" (0.107) and "value" (0.126) more than other hotel class customers, and business travelers value "location" (0.112) and "breakfast" (0.114) more than other travel-type customers.

4.2. Validation

To validate our work, the above-mentioned findings from the proposed IML-based framework are compared with those cases using a previous linear model based on a survey questionnaire dataset as the ground truth. The questionnaire dataset contains the service features defined by TripAdvisor and the explicit information of customer perceptions on the features, and therefore can be used for the validation of online review mining outcomes. The proposed and previous methods are evaluated in terms of the prediction performance of the classifier and the consistency of the results with both sources of online reviews and the survey questionnaire. With both sources, if the proposed method presents a higher prediction performance of the classifier and the obtained importance values of the hotel service features are more consistent than those of the previous method, the proposed method is valid. In addition to online reviews and star ratings, TripAdvisor encourages customers to provide ratings from 1 (very unsatisfied) to 5 (very satisfied) for six features (i.e., location, rooms, sleep quality, service, value, and cleanliness). Because rating these features is not obligatory to customers, there are no ratings for some features. The ratings of these six features, corresponding to the case study, were collected (Table 6), which can be considered as questionnaires directly answered by customers (Bi et al., 2019b). As shown in Table 6, the features identified from online reviews are consistent with the features provided by TripAdvisor.

The proposed IML-based method and the previous linear model were used to estimate the importance values of the six hotel service features in these questionnaires. The missing feature ratings in the questionnaires were filled in with a value of 0, following the same approach as for the review data, to ensure that they did not affect the outputs of the model (Bi et al., 2019b). A five-fold cross-validation was conducted to implement the proposed method, and machine learning models, i.e., NN, DT, RF, and LGBM, were considered to derive the optimal classifiers. The average accuracies of the NN, DT, RF, and LGBM were 0.851, 0.868, 0.872, and 0.878, respectively, on five test sets. The LGBM also had the best prediction performance when using questionnaires as data from the reviews. The SAGE method was used to estimate the importance values of the six hotel service features in each classifier. For each feature, five SAGE values were combined into an importance value using Eqs. 3 and 4. The logit model was conducted as the previous linear model due to categorical data. The importance values of the hotel service features obtained by the proposed IML-based method and previous logit model were estimated using both sources, and common features, i.e., "location", "rooms", "sleep quality", "service", and "value" were compared (Table 7). In the online reviews, "rooms" were considered by combining the importance of "bathrooms" and "views" which are closely related to "rooms". The accuracies achieved using Review-logit, Review-IML, Rating-logit, and Rating-IML were 0.6247, 0.7336, 0.8173, and 0.8776, respectively, on the test set using both sources. The LGBM classifier presented a higher prediction performance than the previous logit classifier because the linear logit model cannot determine the nonlinearity between the sentiments of features and the overall customer satisfaction. Moreover, cosine similarity is used to compare the importance values of the five hotel service features from two sources obtained by the proposed method and the previous method. The similarity from both sources by the proposed method was 0.942 and that by the previous method was 0.782. The proposed method yields a more consistent importance estimation than the previous method with both sources. Therefore, the proposed method is more reliable than the previous logit model because it exhibits the best prediction performance and high consistency with the validation dataset.

[1] https://github.com/cjhutto/vaderSentiment
[2] https://github.com/iancovert/sage

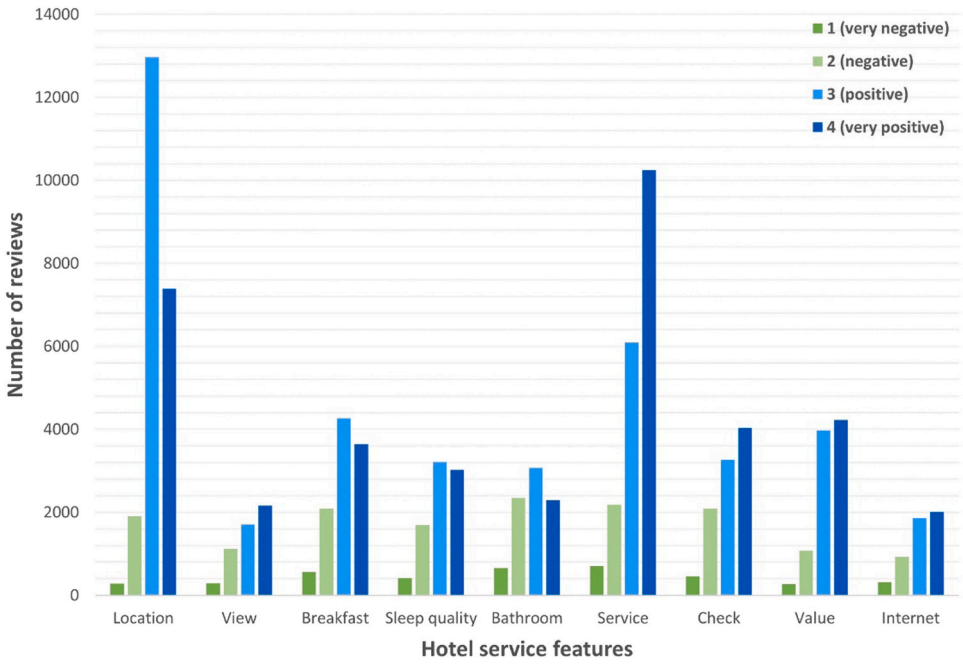

Fig. 2. Statistical results of sentiment scores of each feature in reviews of all hotels.

Table 4
Comparison of prediction performance of classifiers.

Dataset	Model (Training (Test))			
	NN	DT	RF	LGBM
All	0.707(0.715)	0.681(0.683)	0.699(0.704)	**0.734(0.735)**
Class 2	0.625(0.648)	0.650(0.650)	0.657(0.668)	**0.691(0.714)**
Class 3	0.661(0.685)	0.631(0.631)	0.669(0.690)	**0.714(0.720)**
Class 4	0.708(0.720)	0.706(0.710)	0.707(0.714)	**0.750(0.761)**
Solo	0.656(0.614)	0.665(0.666)	0.693(0.711)	**0.725(0.745)**
Couple	0.671(0.683)	0.641(0.641)	0.685(0.703)	**0.727(0.733)**
Friends	0.642(0.657)	0.691(0.694)	0.697(0.710)	**0.727(0.758)**
Family	0.666(0.685)	0.678(0.678)	0.695(0.706)	**0.742(0.750)**
Business	0.649(0.676)	0.660(0.669)	0.682(0.703)	**0.714(0.733)**

5. Discussion

Our work presents a new approach using online review mining with IML for importance estimation that is different from traditional linear model approaches. IML contributes to accurate modeling and analysis of the complex nonlinear relationship between service feature satisfaction and overall customer satisfaction and reliable estimation of importance by interpreting this relationship. In this section, we further describe how the proposed framework can be used for customer-oriented service quality measurement and improvement and discuss its application merits (Section 5.1). In addition, we discuss the strengths and weaknesses of importance estimation from online reviews and questionnaires (Section 5.2), highlighting the significance of this work in exploiting the cost efficiency of online reviews and in validating them using reliable questionnaires.

5.1. Further use of the proposed framework for service improvement

The proposed framework for the global importance estimation can be utilized for an importance-performance analysis (IPA), which is a widely used technique for understanding customer satisfaction and strategic planning for service improvement (Martilla and James, 1977). Our framework provides more accurate and reliable importance values for service features than traditional linear models. An IPA provides strategic

Table 6
Ratings data of six features of Singapore hotels.

Feature	5-point Likert scale					Total
	1	2	3	4	5	
Location	251	604	2882	5654	6129	15,520
Rooms	1234	1851	5261	4906	2132	15,384
Sleep quality	975	1222	3590	5335	3508	14,630
Service	1594	1772	5798	7663	5376	22,203
Value	1179	1577	4392	5432	3347	15,927
Cleanliness	819	1094	3386	6142	4510	15,951

Table 5
Importance values of nine hotel service features in nine datasets.

Dataset	Hotel service features								
	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8	f_9
All	0.083	0.058	0.086	0.097	**0.125**	**0.379**	0.068	0.083	0.021
Class 2	0.048	0.084	0.035	**0.107**	**0.127**	**0.380**	0.067	**0.126**	0.025
Class 3	0.070	0.038	**0.111**	0.098	**0.106**	**0.392**	0.067	0.099	0.019
Class 4	0.089	0.051	0.096	0.095	**0.132**	**0.359**	0.084	0.068	0.026
Solo	0.061	0.095	0.083	0.085	**0.127**	**0.342**	0.059	**0.110**	0.038
Couple	0.094	0.073	0.093	**0.121**	**0.130**	**0.326**	0.075	0.070	0.017
Friends	0.066	0.057	0.073	**0.105**	**0.146**	**0.336**	0.096	0.066	0.055
Family	0.070	0.048	0.081	**0.107**	0.091	**0.406**	0.090	0.092	0.015
Business	**0.112**	0.034	**0.114**	0.077	0.096	**0.362**	0.070	**0.109**	0.025

Table 7
Comparison of importance values and accuracy obtained by proposed IML-based method and previous logit model using two sources.

Source-Method	Feature					Accuracy
	Location	Rooms	Sleep quality	Service	Value	
Review-logit	0.227	0.204	0.009	0.328	0.087	0.625
Review-IML	0.083	0.183	0.097	0.379	0.083	0.734
Questionnaire-logit	0.005	0.089	0.042	0.769	0.066	0.817
Questionnaire-IML	0.039	0.243	0.117	0.300	0.187	0.878

guidelines in four quadrants by dividing the performance and importance of the features into two levels (Fig. 3). Quadrant 1 (Q1) provides "keep up the good work" guidelines. The features in Q1 are considered to be major strengths and competitive advantages because they achieve a high performance and importance. Quadrant 2 (Q2) provides "concentrate here." The features in Q2 are regarded as major weaknesses because they have a lower performance but higher importance. Quadrant 3 (Q3) provides "low priority" guidelines. The features in Q3 are regarded as minor weaknesses because of their low performance and importance. Quadrant 4 (Q4) provides "possible overkill" guidelines. The features in Q4 are considered as minor strengths because they have a high performance but low importance.

In the case study, the IPA of all hotels, class 2 hotels, and business travelers in Singapore, which show distinct differences, was drawn using the proposed framework (Fig. 3). In Section 4, the importance values of hotel service features are calculated, and the performance of each feature is calculated using $\sum_{m=1}^{M} S_{im}/R_i$, R_i = sum (1 if $S_{im} > 0$; 0, otherwise) (Bi et al., 2019b; Joung and Kim, 2021). The cross-hair of the IPA is determined based on the importance and performance values of hotel service features, and this method has high discriminative power between the features (Eskildsen and Kristensen, 2006). In the IPA of all hotels, "service" (f_6) was located in Q1. The competitive advantage of this feature must be maintained through sustained investment. "bathroom" (f_5) was located in Q2. Immediate investment and attention to features are required. "View" (f_2), "breakfast" (f_3), "sleep quality" (f_4), "check" (f_7), and "internet" (f_9) were located in Q3. These features have low priority in terms of service improvement. "Location" (f_1) and "value" (f_8) were located in Q4. Investment and attention to these features may not be considered overemphasized features.

In the IPA of class 2 hotels and business travelers, some features are arranged differently compared with the IPA of all hotels. "Value" was positioned in Q4 in the IPA of all hotels but moved to Q1 in the IPA of class 2 hotels. Customers of class 2 hotels may have considered "value" as their major strength because they are price sensitive. "Location" and "breakfast" were Q4 and Q3, respectively in the IPA of all hotels, but they moved to Q1 and Q2, respectively in the IPA of business travelers. Business travelers may have valued "location" and "breakfast" because the purpose of their visit is clear. Moreover, "bathroom" was Q2 in the IPA of all hotels but moved to Q3 in the IPA of business travelers. Unlike couple travelers who value the "bathroom" the most, business travelers showed a low priority regarding this category.

This analysis allows hotel managers to effectively allocate their resources to improve service operations and strengthen marketing strategies tailored to their hotel businesses. For example, class 2 hotels can focus on "value" to strengthen service competitiveness and highlight cost-effectiveness to customers for promotions. In addition, business hotels may invest more in improving location accessibility and breakfast service instead of improving the quality of bathrooms, a major feature perceived by "couples" Furthermore, they can strengthen their marketing by emphasizing "convenient locations" and "high-quality breakfast." Therefore, this analysis provides hotels insight into the specific needs and preferences of various customer groups and enables the customization of their business strategies.

Although the above analysis shows the utility of reliable estimation of service feature importance, traditional linear models may fail to capture the nonlinear relationship between service features and overall satisfaction, potentially overlooking some improvement points. With the proposed framework, which utilizes online reviews with IML, we can efficiently and effectively identify the overall improvement points of service. Additionally, we can conduct more detailed analyses, such as compare hotel classes or different types of travelers, without requiring time-consuming and expensive surveys. The establishment of such an analysis pipeline lays the foundation for continuous and systematic customer-oriented service monitoring by effectively utilizing automatically collected data. This will contribute significantly to identifying opportunities for service improvement and providing prompt responses. Therefore, we expect our proposed framework to efficiently reveal new service improvement points that may have been missed.

5.2. Comparison of sources for service feature importance estimation

Section 4 presented the use of online reviews and questionnaires as the two sources to estimate the importance values of hotel service features, whose strengths and weaknesses are discussed herein. The advantage of using questionnaires for importance estimation is that it is more reliable for deriving importance values than reviews because customers directly assign importance and satisfaction scores to

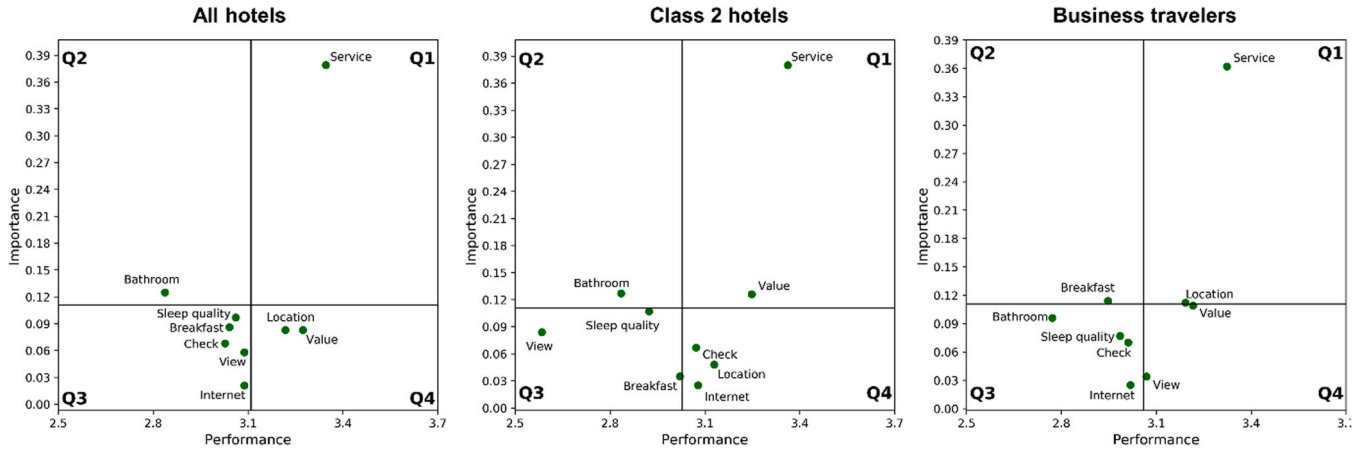

Fig. 3. IPA of the all hotels, class 2 hotels, and business travelers.

predefined features. Some reviews may not be completely accurate for determining the sentiment of features because of many reasons, including sarcastic or vague expressions. Meanwhile, collecting questionnaires at a low cost is difficult. Moreover, as questionnaires utilize predefined features, the importance values of new features that customers frequently mention in recent reviews cannot be estimated with questionnaires. For example, in the case study, the importance values of nine hotel service features were derived from online reviews and the set of features can be changed according to the changes around the hotels (e.g., Covid-19), whereas those of only six predefined hotel service features were fixed and cannot be changed flexibly.

Thus, we suggest using both online reviews and questionnaires as the complementary resources for customer-oriented service quality evaluation and improvement; the former resource has an efficiency advantage while the latter has high reliability. When there are no or few questionnaire data available, firms should utilize the review data only. Although most review websites provide textual reviews and star ratings, they do not provide ratings for the predefined features as the case of TripAdvisor. More questionnaire data can be collected if the websites and service managers encourage customers to provide reviews and ratings of predefined features. For example, in the case study of using TripAdvisor data, 32,044 reviews were collected, while 22,319 responses were collected from questionnaires. Both types of data should be used to automate and validate the online review mining with IML, and it is interesting that our work showed that the IML-based global importance estimations using both online reviews and questionnaires with the proposed framework were almost identical, as presented in Section 4.2.

6. Conclusion

This paper presents a framework based on IML to estimate the importance values of service features from online reviews. The proposed framework provides an opportunity for service managers to utilize online customer reviews to lead effective resource allocation for service improvement, based on which the service features that should be strengthened. Specifically, an IML-based method is proposed to estimate the importance of the sentiments of features on the overall customer satisfaction. In our case study, for estimating the feature importance values of hotel services, the proposed method using the LGBM classifier achieved a higher performance than a previous logit classifier. The importance values of features obtained by the proposed method using both online reviews and questionnaires are more consistent than those obtained by the previous logit model. These experimental results provide the theoretical support for the nonlinear assumptions required when estimating the relationship between service feature satisfaction and overall satisfaction in the hospitality domain. Moreover, these results demonstrate that our method contributes to enhancing the accuracy and reliability of the results of analytical tools that utilize importance, such as the IPA.

Although this work shows the great potential of IML as a new research methodology for social science and business research, this work has some limitations, which will provide directions for further research. First, the proposed framework is applied and validated for the hotel service context only, particularly Singapore hotel services. Future studies can test it on hotels in other regions, services, and goods. Second, machine learning models and their interpretation techniques are continually improving. Future studies may use other advanced models and interpretation techniques in the proposed framework.

CRediT authorship contribution statement

Chiehyeon Lim: Project administration, Supervision, Writing – review & editing. **Junegak Joung:** Conceptualization, Data curation, Funding acquisition, Investigation, Methodology, Supervision, Validation, Writing – original draft, Writing – review & editing. **Jongkyung Shin:** Data curation, Software, Writing – review & editing, Writing – original draft.

Declaration of Competing Interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: The authors report that financial support was provided by the institutions listed in the following Acknowledgements.

Acknowledgments

This research was supported by the research fund of Hanyang University (HY-202300000001721), the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2021R1I1A1A01044552, 2021R1I1A4A01049121, 2021S1A5A2A03065747), Institute of Information & Communications Technology Planning & Evaluation (IITP) grants funded by the Korean government (MSIT) (No. 2020-0-01336: Artificial Intelligence Graduate School Program - UNIST, No. 2021-0-02068, Artificial Intelligence Innovation Hub), and 2023 Project Fund (1.230019.01) of UNIST (Ulsan National Institute of Science & Technology).

References

Akbaba, A., 2006. Measuring service quality in the hotel industry: a study in a business hotel in turkey. Int. J. Hosp. Manag. 25 (2), 170–192.
Albayrak, T., 2015. Importance performance competitor analysis (ipca): a study of hospitality companies. Int. J. Hosp. Manag. 48, 135–142.
Alzate, M., Arce-Urriza, M., Cebollada, J., 2022. 'Mining the text of online consumer reviews to analyze brand image and brand positioning'. J. Retail. Consum. Serv. 67, 102989.
Anderson, E.W., Mittal, V., 2000. 'Strengthening the satisfaction-profit chain'. J. Serv. Res. 3 (2), 107–120.
Bekkar, M., Djemaa, H.K., Alitouche, T.A., 2013. 'Evaluation measures for models assessment over imbalanced data sets'. J. Inf. Eng. Appl. 3 (10).
Bi, J.W., Liu, Y., Fan, Z.-P., Cambria, E., 2019a. 'Modelling customer satisfaction from online reviews using ensemble neural network and effect-based kano model'. Int. J. Prod. Res. 57 (22), 7068–7088.
Bi, J.W., Liu, Y., Fan, Z.-P., Zhang, J., 2019b. 'Wisdom of crowds: Conducting importance-performance analysis (IPA) through online reviews'. Tour. Manag. 70, 460–478.
Blei, D.M., Ng, A.Y., Jordan, M.I., 2003. 'Latent dirichlet allocation'. J. Mach. Learn. Res. 3, 993–1022.
Boyd-Graber, J., Mimno, D., Newman, D., 2014. 'Care and feeding of topic models: Problems. Diagn., improvements', Handb. Mixed Membsh. Models their Appl., 225255
Breiman, L., 2001. 'Random forests'. Mach. Learn. 45 (1), 5–32.
Caber, M., Albayrak, T., Loiacono, E.T., 2013. The classification of extranet attributes in terms of their asymmetric influences on overall user satisfaction: an introduction to asymmetric impact-performance analysis. J. Travel Res. 52 (1), 106–116.
Chang, Y.-C., Ku, C.-H., Chen, C.-H., 2019. Social media analytics: Extracting and visualizing hilton hotel ratings and reviews from tripadvisor. Int. J. Inf. Manag. 48, 263–279.
Chen, W.J., 2013. Factors influencing internal service quality at international tourist hotels. Int. J. Hosp. Manag. 35, 152–160.
Cheng, C.C., Chen, C.T., Hsu, F.S., Hu, H.Y., 2012. Enhancing service quality improvement strategies of fine-dining restaurants: New insights from integrating a two-phase decision-making model of ipga and dematel analysis. Int. J. Hosp. Manag. 31 (4), 1155–1166.
Choi, T.Y., Chu, R., 2001. Determinants of hotel guests' satisfaction and repeat patronage in the hong kong hotel industry. Int. J. Hosp. Manag. 20 (3), 277–297.
Covert, I., Lundberg, S., Lee, S.-I., 2020. Understanding global feature contributions with additive importance measures. Adv. Neural Inf. Process. Syst. 33.
Culotta, A., Cutler, J., 2016. Mining brand perceptions from twitter social networks. Mark. Sci. 35 (3), 343–362.
Decker, R., Trusov, M., 2010. Estimating aggregate consumer preferences from online product reviews. Int. J. Res. Mark. 27 (4), 293–307.
Deng, W., 2007. Using a revised importance–performance analysis approach: The case of taiwanese hot springs tourism. Tour. Manag. 28 (5), 1274–1284.
Deng, W.J., Chen, W.C., Pei, W., 2008. Back-propagation neural network based importance–performance analysis for determining critical service attributes. Expert Syst. Appl. 34 (2), 1115–1125.

Ding, K., Choo, W.C., Ng, K.Y. & Ng, S.I. , 2020, 'Employing structural topic modelling to explore perceived service quality attributes in airbnb accommodation', International Journal of Hospitality Management 91, 102676.
Du, M., Liu, N., Hu, X., 2019. Techniques for interpretable machine learning. Commun. ACM 63 (1), 68–77.
Eskildsen, J.K., Kristensen, K., 2006. Enhancing importance- performance analysis. Int. J. Product. Perform. Manag.
Farhadloo, M., Patterson, R.A., Rolland, E., 2016. Modeling customer satisfaction from unstructured data using a bayesian approach. Decis. Support Syst. 90, 1–11.
Fernandes, S., Panda, R., Venkatesh, V., Swar, B.N., Shi, Y., 2022. Measuring the impact of online reviews on consumer purchase decisions– a scale development study. J. Retail. Consum. Serv. 68, 103066.
Girardin, F., Bezencon, V., Lunardo, R., 2021. Dealing with poor online ratings in the hospitality service industry: The mitigating power of corporate social responsibility activities. J. Retail. Consum. Serv. 63, 102676.
Griffin, A., Hauser, J.R., 1993. The voice of the customer. Mark. Sci. 12 (1), 1–27.
Guo, H., Zhu, H., Guo, Z., Zhang, X. & Su, Z. , 2009, Product feature categorization with multilevel latent semantic association, in 'Proceedings of the 18th ACM Conference on Information and Knowledge Management', pp. 1087–1096.
Hastie, T., Tibshirani, R., Friedman, J., 2009. The elements of statistical learning: data mining, inference, and prediction. Springer Science & Business Media.
Hu, M., Liu, B., 2004. Mining opinion features in customer reviews. 'AAAI' Vol. 4, 755–760.
Hutto, C. & Gilbert, E.2014, Vader: A parsimonious rule-based model for sentiment analysis of social media text, in 'Proceedings of the International AAAI Conference on Web and Social Media', Vol. 8.
Hutto, C.J. , 2022, vaderSentiment, GitHub, https://github.com/cjhutto/vaderSentiment.
Izadi, A., Jahani, Y., Rafiei, S., Masoud, A., Vali, L., 2017. Evaluating health service quality: using importance performance analysis. Int. J. Health Care Qual. Assur.
Jeong, B., Yoon, J., Lee, J.-M., 2019. Social media mining for product planning: a product opportunity mining approach based on topic modeling and sentiment analysis. Int. J. Inf. Manag. 48, 280–290.
Johnson, M., Gustafsson, A., 2006. Improving Customer Satisfaction, Loyalty and Profit: An Integrated Measurement and Management System. John Wiley & Sons.
Joung, J., Kim, H., 2023. Interpretable machine learning-based approach for customer segmentation for new product development from online product reviews. Int. J. Inf. Manag. 70, 102641.
Joung, J., Kim, H.M., 2020. Automated keyword filtering in latent Dirichlet allocation for identifying product attributes from online reviews. J. Mech. Des. 1–10.
Joung, J., Kim, H.M., 2021. Approach for importance-performance analysis of product attributes from online reviews. J. Mech. Des. 1–20.
Kim, J., Lim, C., 2021. Customer complaints monitoring with customer review data analytics: an integrated method of sentiment and statistical process control analyses. Adv. Eng. Inform. 49, 101304.
Kostromitina, M., Keller, D., Cavusoglu, M., Beloin, K., 2021. His lack of a mask ruined everything." Restaurant customer satisfaction during the covid-19 outbreak: An analysis of yelp review texts and star-ratings. Int. J. Hosp. Manag. 98, 103048.
Kwon, W., Lee, M., Back, K.J., 2020. Exploring the underlying factors of customer value in restaurants: a machine learning approach. Int. J. Hosp. Manag. 91, 102643.
Lai, I.K.W., Hitchcock, M., 2016. A comparison of service quality attributes for stand-alone and resort-based luxury hotels in Macau: 3-Dimensional importance-performance analysis. Tour. Manag. 55, 139–159.
Lee, Y.C., Huang, S.Y., 2009. A new fuzzy concept approach for kano's model. Expert Syst. Appl. 36 (3), 4479–4484.
Liu, Y., Bi, J.W., Fan, Z.P., 2017. A method for multi-class sentiment classification based on an improved one-vs-one (OVO) strategy and the support vector machine (SVM) algorithm. Inf. Sci. 394, 38–52.
Louppe, G., Wehenkel, L., Sutera, A., Geurts, P., 2013. Understanding variable importances in forests of randomized trees. Adv. Neural Inf. Process. Syst. 26.
Lundberg, S., Lee, S.-I., 2017. A unified approach to interpreting model predictions. arXiv Prepr. arXiv 1705, 07874.
Luo, J.M., Vu, H.Q., Li, G., Law, R., 2021. Understanding service attributes of robot hotels: a sentiment analysis of customer online reviews. Int. J. Hosp. Manag. 98, 103032.
Luo, Y., Tang, R.L., 2019. Understanding hidden dimensions in textual reviews on Airbnb: an application of modified latent aspect rating analysis (LARA). Int. J. Hosp. Manag. 80, 144–154.
Luo, Y., Xu, X., 2021. Comparative study of deep learning models for analyzing online restaurant reviews in the era of the COVID-19 pandemic. Int. J. Hosp. Manag. 94, 102849.
Martilla, J.A., James, J.C., 1977. Importance-performance analysis. J. Mark. 41 (1), 77–79.
Matzler, K., Sauerwein, E., Heischmidt, K., 2003. Importance-performance analysis revisited: the role of the factor structure of customer satisfaction. Serv. Ind. J. 23 (2), 112–129.
Matzler, K., Bailom, F., Hinterhuber, H.H., Renzl, B., Pichler, J., 2004. The asymmetric relationship between attribute-level performance and overall customer satisfaction: a reconsideration of the importance– performance analysis. Ind. Mark. Manag. 33 (4), 271–277.
Mejia, C., Bąk, M., Zientara, P., Orlowski, M., 2022. Importance-performance analysis of socially sustainable practices in us restaurants: A consumer perspective in the quasi-post-pandemic context. Int. J. Hosp. Manag. 103, 103209.
Mejia, J., Mankad, S., Gopal, A., 2021. Service quality using text mining: Measurement and consequences. Manuf. Serv. Oper. Manag. 23 (6), 1354–1372.
Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J., 2013. Distributed representations of words and phrases and their compositionality. arXiv Prepr. arXiv 1310, 4546.
Mikulić, J., Prebežac, D., 2012. Accounting for dynamics in attribute-importance and for competitor performance to enhance reliability of BPNN-based importance–performance analysis. Expert Syst. Appl. 39 (5), 5144–5153.
Miller, G.A., 1995. Wordnet: a lexical database for english. Commun. ACM 38 (11), 39–41.
Mittal, V., Ross Jr, W.T., Baldasare, P.M., 1998. The asymmetric impact of negative and positive attribute-level performance on overall satisfaction and repurchase intentions. J. Mark. 62 (1), 33–47.
Mohsin, A., Rodrigues, H., Brochado, A., 2019. Shine bright like a star: Hotel performance and guests' expectations based on star ratings. Int. J. Hosp. Manag. 83, 103–114.
Molnar, C. , 2020, Interpretable machine learning, Lulu. com.
Nasiri, M.S., Shokouhyar, S., 2021. Actual consumers' response to purchase refurbished smartphones: Exploring perceived value from product reviews in online retailing. J. Retail. Consum. Serv. 62, 102652.
Nie, R.X., Chin, K.S., Tian, Z.P., Wang, J.Q., Zhang, H.Y., 2023. Exploring dynamic effects on classifying service quality attributes under the impacts of COVID-19 with evidence from online reviews. Int. J. Contemp. Hosp. Manag. 35 (1), 159–185.
Nilashi, M., Abumalloh, R.A., Alghamdi, A., Minaei-Bidgoli, B., Alsulami, A.A., Thanoon, M., Asadi, S., Samad, S., 2021. What is the impact of service quality on customers' satisfaction during covid-19 outbreak? new findings from online reviews analysis. Telemat. Inform. 64, 101693.
Nilashi, M., Abumalloh, R.A., Minaei-Bidgoli, B., Zogaan, W.A., Alhargan, A., Mohd, S., Azhar, S.N.F.S., Asadi, S., Samad, S., 2022. Revealing travellers' satisfaction during covid-19 outbreak: moderating role of service quality. J. Retail. Consum. Serv. 64, 102783.
Olsen, L.L., Witell, L. & Gustafsson, A. , 2014, 'Turning customer satisfaction measurements into action', Journal of Service Management.
Palese, B., Usai, A., 2018. The relative importance of service quality dimensions in e-commerce experiences. Int. J. Inf. Manag. 40, 132–140.
Pan, M., Li, N., Law, R., Huang, X., Wong, I.A., Zhang, B., Li, L., 2023. Service attribute prioritization based on the marginal utility of attribute performance. Int. J. Hosp. Manag. 114, 103560.
Parasuraman, A., Zeithaml, V.A., Berry, L.L., 1985. A conceptual model of service quality and its implications for future research. J. Mark. 49 (4), 41–50.
Parasuraman, A., Zeithaml, V.A., Berry, L., 1988. Servqual: A multipleitem scale for measuring consumer perceptions of service quality. 1988 64 (1), 12–40.
Pezeshki, V., Mousavi, A., Grant, S., 2009. Importance-performance analysis of service attributes and its impact on decision making in the mobile telecommunication industry. Meas. Bus. Excell. 13 (1), 82–92.
Qiao, T., Shan, W., Zhang, M., Wei, Z., 2022. More than words: understanding how valence and content affect review value. Int. J. Hosp. Manag. 105, 103274.
Ramanathan, R., Ramanathan, U., 2016. A new rational IPA and application to cruise tourism. Ann. Tour. Res. 61, 264–267.
Rese, A., Schreiber, S., Baier, D., 2014. Technology acceptance modeling of augmented reality at the point of sale: can surveys be replaced by an analysis of online reviews? J. Retail. Consum. Serv. 21 (5), 869–876.
Slevitch, L., Oh, H., 2010. Asymmetric relationship between attribute performance and customer satisfaction: a new perspective. Int. J. Hosp. Manag. 29 (4), 559–569.
Suryadi, D., Kim, H., 2018. A systematic methodology based on word embedding for identifying the relation between online customer reviews and sales rank. J. Mech. Des. 140 (12).
Ting, S.C., Chen, C.N., 2002. The asymmetrical and non-linear effects of store quality attributes on customer satisfaction. Total Qual. Manag. 13 (4), 547–569.
Wang, W., Feng, Y., Dai, W., 2018. Topic analysis of online reviews for two competitive products using latent dirichlet allocation. Electron. Commer. Res. Appl. 29, 142–156.
Wu, J., Yang, T., Zhou, Z., Zhao, N., 2023. Consumers' affective needs matter: open innovation through mining luxury hotels' online reviews. Int. J. Hosp. Manag. 114, 103556.
Wu, J.J., Chang, S.T., 2020. Exploring customer sentiment regarding online retail services: a topic-based approach. J. Retail. Consum. Serv. 55, 102145.
Xu, X., Wang, X., Li, Y., Haghighi, M., 2017. Business intelligence in online customer textual reviews: Understanding consumer perceptions and influential factors. Int. J. Inf. Manag. 37 (6), 673–683.
Ye, F., Xia, Q., Zhang, M., Zhan, Y., Li, Y., 2020. Harvesting online reviews to identify the competitor set in a service business: evidence from the hotel industry. J. Serv. Res. 1094670520975143.
Zhang, C., Xu, Z., Gou, X., Chen, S., 2021. An online reviews-driven method for the prioritization of improvements in hotel services. Tour. Manag. 87, 104382.
Zhang, D., Shen, Z., Li, Y., 2023. Requirement analysis and service optimization of multiple category fresh products in online retailing using importance-kano analysis. J. Retail. Consum. Serv. 72, 103253, 29.
Zhu, J.J., Chang, Y.C., Ku, C.H., Li, S.Y., Chen, C.J., 2021. Online critical review classification in response strategy and service provider rating: Algorithms from heuristic processing, sentiment analysis to deep learning. J. Bus. Res. 129, 860–877.

Jongkyung Shin is currently a Ph.D. student in the Graduate School of Artificial Intelligence at UNIST. He obtained his B.S (2020) from the Department of Industrial Engineering in UNIST. His research interests include AI-based service system, data-driven service management, and machine learning operations.

Junegak Joung is currently an assistant professor in the School of Interdisciplinary Industrial Studies and the Graduate School of Industrial Data Engineering at Hanyang University. He received a BS and a Ph.D. from the Department of Industrial and Management Engineering at Pohang University of Science and Technology (POSTECH) in 2013 and 2018. His main research interests include user data mining, interpretable machine learning applications, and data-driven product/service quality management.

Chiehyeon Lim is an Associate Professor in the Department of Industrial Engineering and the Graduate School of Artificial Intelligence at UNIST. He obtained his B.S. and Ph.D. from the Department of Industrial and Management Engineering at POSTECH. As part of his postdoctoral experience, he served as an Assistant Project Scientist and Lecturer in the School of Engineering at UC Merced. His research interests include data-driven service management and knowledge discovery with data mining.