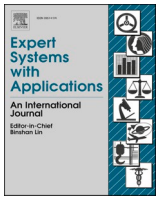
Expert Systems With Applications 232 (2023) 120767

Contents lists available at ScienceDirect

Expert Systems With Applications

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



Combined Text-Mining/DEA method for measuring level of customer satisfaction from online reviews

Jaehun Park

Department of Business Administration, Changwon National University, 20 Changwondaehak-ro Uichang-gu, Changwon 51140, Republic of Korea

ARTICLE INFO

Keywords:
Customer satisfaction
Online text reviews
Perception evaluation
Sentiment analysis
Data envelopment analysis

ABSTRACT

Determining the best way to measure level of customer satisfaction (LCS) with service quality and its determinants has been a matter of concern for both practitioners and researchers. This is especially so, as LCS can be used to retain customers, sell products and services, improve the quality and value of offers and ensure more efficient and economical operating. The present study proposes a new LCS evaluation methodology that combines data envelopment analysis (DEA) with text mining to analyze online textual reviews. The proposed methodology identifies, from online reviews using a term-frequency–inverse document frequency (TF-IDF) algorithm, multiple satisfaction metrics that significantly affect customers' service experience, quantifies them by sentiment analysis, and evaluates, by a DEA model, service providers' LCS with respect to those metrics. To illustrate the efficacy and applicability of the proposed approach, an empirical case study applying it to the world's top 20 airlines in 2020 was conducted. This study demonstrates how the DEA model can be effectively utilized for evaluation of LCS from online textual review data by combining it with the TF-IDF text-mining technique.

1. Introduction

Customer satisfaction (CS) is one of the most important sources of competitive advantage in service industries, and service quality has an important influence on CS (Martín-Cejas, 2006). The marketing literature suggests that the long-term success of a firm is clearly based on its ability to rapidly respond to changing customer needs and preferences (Webster Jr, 1992). A key motivation for the increasing emphasis on CS is that higher CS can lead to a stronger competitive position and higher market share and profitability for firms (Yang et al., 2009). During the past two decades, determining the best way to measure the level of CS (LCS) and its determinants with respect to the quality of service provided has been a matter of concern for both practitioners and researchers, since higher CS levels can be leveraged to retain customers, sell products and services, improve the quality and value of offers, and ensure more efficient and economical operating. The LCS in many service industries might be measured in various ways for different customers, as different customers have different perceptions of the same service quality.

Many customers nowadays refer to various websites and social-media outlets when preparing consumption, and those communication channels become popular as increasing numbers of users share their own experiences in online reviews. Online reviews, which can share information on various products with many customers, are typically known to reflect experienced CS and to have a stronger influence on purchasing decisions for consumers than offline information (Yoon & Ha, 2010). In general, customers can use online reviews to support their purchase decisions, and service providers can understand customers' current perceptions and utilize the knowledge of their preferences in product improvement, marketing, and customer relationship management (Somprasertsri & Lalitrojwong, 2010). More specifically, according to Gremler et al. (2001), Peterson and Merino (2003), and Hennig-Thurau et al. (2004), positive reviews, such as compliments or explanations of useful points, have a positive effect on CS, which subsequently lead to purchases, while negative reviews, such as dissatisfaction with the product, have a negative effect on purchases. Based on the above argument, CS in this study was judged based on the principle that "the higher the positive opinions from online reviews, the higher the CS." As online review platforms have become increasingly widespread in recent years, customer-generated content in the form of online reviews has explosively proliferated (Mariani & Borghi, 2018). For instance, the number of online reviews on *TripAdvisor* increased from 200 million in 2014 to about 859 million in 2019. Online reviews have generated a paradigm shift in the way that consumers share their experiences as to

E-mail address: pjh3479@changwon.ac.kr.

https://doi.org/10.1016/j.eswa.2023.120767
Received 7 October 2021; Received in revised form 6 June 2023; Accepted 6 June 2023
Available online 12 June 2023

the positive and negative aspects of a given service online; such reviews are used as important information for new customers in making purchase decisions as well as for business managers seeking to understand current customer perceptions of their service provision. Archak et al. (2011) and Ghose & Ipeirotis (2011) have shown strong evidence that online reviews affect consumer purchasing decisions and are an important means of analyzing customers' requirements for, and evaluating their satisfaction with, product or service-quality.

Despite the increasing role that online reviews play in representing more effective and realistic customer opinions, and the empirical evidence that social media rating is a more significant predictor than the traditional customer, many studies still utilize offline data obtained from customer interviews or questionnaire-surveys (i.e. SERVQUAL/SERVPERF) in measuring LCS for service providers. However, interview- or questionnaire survey-based CS measurement has two major drawbacks in that 1) the capacity to collect various opinions is limited due to data sampling and pre-defined question contents by the investigator (i.e. the sample size and survey methods must be selected to ensure it is representative of customers), and 2) it cannot quickly identify the real-time requirements of consumers, given the significant amount of time required to gather and analyze opinions (Gordon, 2008). Recently, online reviews have been attracting attention as fundamental data sources for measurement of service providers' LCS, because they can collect multifaceted information on products from various customers in a short time and at a relatively low cost. Above all, online reviews, as they are written voluntarily by actual customers, contain less distortion as to customers' practical service experience (Gan et al., 2016). Although some studies have utilized the service rating score (a quantitative score measured by customers on a 5- or 10-point scale for each predefined service factor) for CS evaluation, its use for CS evaluation has several limitations. The service rating score does not include the various and specific feelings, expressions, and opinions of customers that are often expressed in online reviews, and consequently, it may not give the discriminative power to the LCS between service providers. The service rating score from the TripAdvisor website was surveyed in five-star hotels in Korea to examine this. 18 hotels among them have the same service rating score of 4.5 points (5-point scale), so it was not possible to strictly distinguish the LCS difference between these hotels with only the service rating scores. In addition, many online travel agencies such as TripAdvisor predefine key service attributes (e.g., lodging charges, location, staff kindness, facilities, etc.) for their own service rating survey or evaluation, but these service attributes may not reflect the subjective attributes that customers consider important for service satisfaction evaluation. The motivation of this study stems from the need for a method to evaluate the level of discriminatory CS by reflecting the realistic and diverse opinions of customers.

The aim of this paper is to propose a new LCS evaluation methodology that combines a data envelopment analysis (DEA) model with a text-mining technique in order to analyze online review text data. More specifically, the proposed methodology uses a term-frequency–inverse document frequency (TF-IDF) algorithm to identify, from online reviews, multiple satisfaction metrics that significantly affect customers' service experience, quantifies them using sentiment analysis, and, finally, evaluates the LCS of the service providers with respect to multiple satisfaction metrics using the DEA model. To demonstrate the efficacy and applicability of the proposed approach, an empirical case study applying it to international aviation was conducted. The main contribution of this study is its demonstration of how the DEA model can be effectively utilized with text mining (i.e., TF-IDF algorithm and sentiment analysis) to evaluate LCS from online textual review data. Although a few studies have utilized online reviews for CS-related purposes such as determinants identification of CS from online reviews (Hsu, 2008; Li et al., 2013; Zhao et al., 2019; Baydogan & Alatas, 2019) and CS analysis (Ba & Johansson, 2008; Ramanathan et al., 2017; Sari et al., 2019), they did not provide any systematic decision-making method for evaluation of LCS among service providers. The approach proposed herein is, to our best knowledge, the first of its kind.

The rest of this study is organized as follows. Section 2 discusses the contributions of this study in the context of the previous achievements in the literature. Section 3 introduces the component methodologies of the proposed methodology. Section 4 provides various analysis results based on examination of real-world data using the proposed methodology. Section 5 discusses the implication of this study. Section 6 draws conclusions.

2. Literature review

Online reviews, which can share information about various products with many customers, have a stronger influence on purchasing decisions for consumers than offline information (Yoon & Ha, 2010), and it is useful for service providers to utilize the knowledge of customer's current perceptions and preferences in product improvement, marketing, and customer relationship management (Somprasertsri & Lalitrojwong, 2010). In the literature on online review analysis, most of the studies focus on the improvement of text-extracting techniques to derive information from in more effective and efficient ways (Zhan et al., 2009; Yi & Niblack, 2005; Lee & Bradlow, 2011). A few studies have been carried out on subjectivity analysis and sentiment analysis (Pang et al., 2002; Arora et al., 2009) of online customer reviews. Also, some researchers have focused on the investigation of the relationship between product review content, conceptual cues, and review helpfulness (Ghose & Ipeirotis, 2011). Additionally, other studies have utilized sentiment analysis for new product development in the cosmetic industry (Haddara et al., 2019) and for measurement of quality satisfaction with mobile services (Kang & Park, 2014).

Customer satisfaction (CS) is defined as "a person's feelings of pleasure or disappointment that results from comparing a product's perceived performance or outcome with his/her expectations" (Kotler & Keller, 2009). CS is the leading criterion for determining the quality that is actually delivered to customers through the product/service and by the accompanying servicing. In short, CS is important to all commercial firms owing to its influence on repeat purchases and word-of-mouth (WOM) recommendations, both positive and negative (Abubakar & Mavondo, 2014). The best known and most widely applied technique is the SERVQUAL method (Parasuraman et al., 1988). The SERVQUAL method introduced the concept of CS as a function of expectations (what customers expect from the service) and perceptions (what customers receive), and defined 5 service quality dimensions, namely tangibles, reliability, responsiveness, assurance, and empathy, along with 22 items for measurement of service quality. To the existing literature on CS, a number of both national and international indexes based on customer perceptions and expectations have been introduced. Some of those CS indexes were designed according to the determinants or configuration of a satisfaction index such as the Swedish Customer Satisfaction Barometer (SCSB), the American Customer Satisfaction Index (ACSI), the Norwegian Customer Satisfaction Barometer (NCSB), and the European Customer Satisfaction Index (ECSI) (Bayraktar et al., 2012). All of the other models are based on the same concepts, but differ from the originals regarding the variables considered and the cause-and-effect relationships applied. The models from which these indexes are derived have a very complex structure. More recently, indexes based on discrete choice models and random utility theory such as Service Quality Index (SQI) have been introduced. SQI is calculated according to the utility function of a choice alternative representing a service (Hensher & Prioni, 2002). For SQI calculation, the user makes a choice between the habitual service, which is described by the user by assigning a value to each service aspect, and hypothetical services, which are defined through Stated Preferences (SP) techniques by varying the level of quality of aspects characterizing the service.

There have been several studies entailing CS analysis or measurement of LCS utilizing online reviews. Hsu (2008) proposed, for online CS purposes, an index adapted from the American Customer Satisfaction

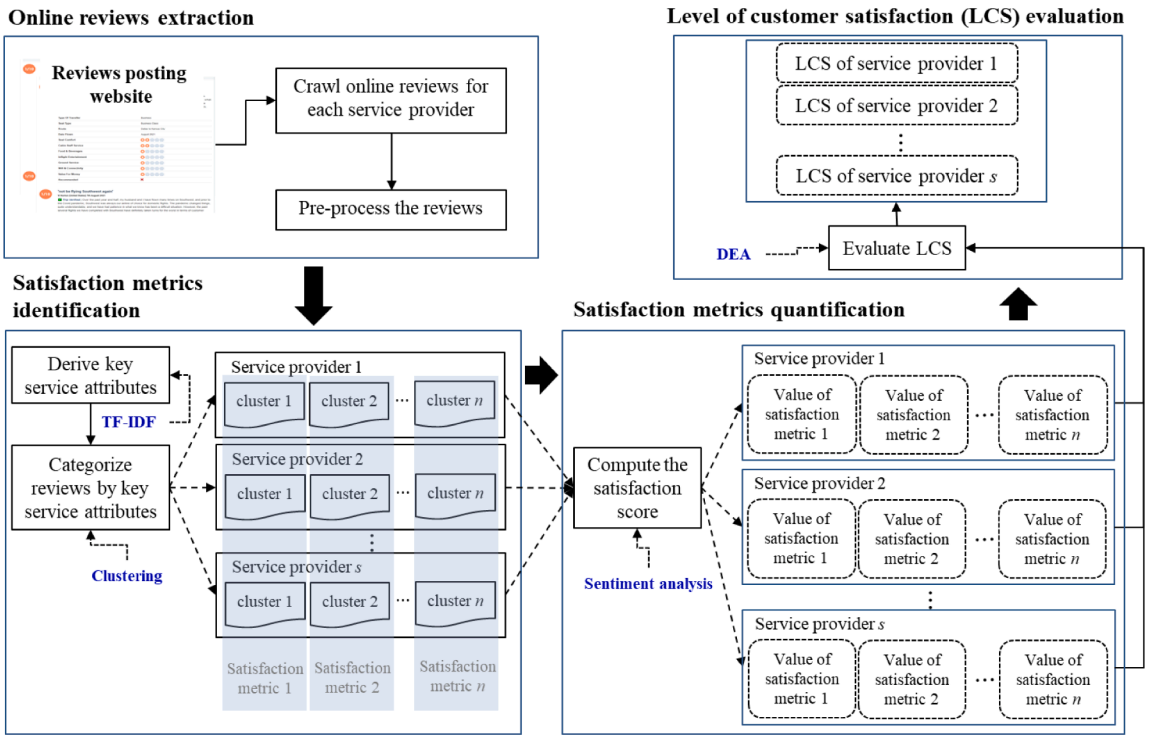


Fig. 1. Proposed methodology.

Index (ACSI). His approach allowed the online retailer to understand the specific factors that significantly influence overall CS by reading causal relationships on a strategic management map in an electronic-customer satisfaction index model. Li et al. (2013) identified determinants of CS in the hospitality industry through an analysis of online reviews. They demonstrated that transportation convenience, food and beverage management, convenience to tourist destinations and value for money are factors that customer booking both luxury and budget hotels consider important. Zhao et al. (2019) predicted overall CS using the technical attributes of online reviews and customers' involvement in the review community. They found that a higher level of subjectivity and readability and a longer length of textual review lead to lower overall CS, that a higher level of diversity and sentiment polarity of textual review leads to higher overall CS, and that customers' review involvement positively influences their overall satisfaction. Baydogan and Alatas (2019) identified CS in comments made on products or services using various Natural Language Processing and Machine learning methods. Ba and Johansson (2008) proposed viewing the interface between online buyers and sellers through the lens of service management to identify possible determinants of online CS. They found that as the electronic service delivery system process improves, a customer's perception of the website's ease of use increases, leading to increased service value and perceived control over the process, which increases CS. Also, they found that the technological capabilities embedded in website processes are an important factor in determining service quality and ultimately online CS. Ramanathan et al. (2017) employed a survey questionnaire method to elicit opinions on retail CS based on social media reviews, service operations and marketing efforts. They found that social media reviews dramatically impact CS, and their empirical analysis identified the significant and positive effects of service operations on CS levels. Sari et al. (2019) analyzed CS sentiment regarding online transporatation in Indonesia using public opinion as expressed on Twitter. They used the Naive Bayes method to classify positive, neutral and negative sentiments from tweets that had been published on twitter by customers.

To sum up this literature review, most studies related to CS estimation from online reviews have focused on automatic processes and/or algorithms for extraction of online reviews from online websites. Although a few studies have conducted determinants identification or polarity analysis of CS from online reviews using sentiment analysis, they did not deal with evaluation of CS among service providers from the multiple service dimension perspectives. Although the DEA model has been utilized to evaluate CS with service providers for the purpose of object weight assignment to reflect the relative significances of decision-making units (Mariani & Visani, 2019), customers' opinions had been collected not by online review analysis but by questionnaire survey. Questionnaire-survey-based data collection, as noted above, has some drawbacks in that it is limited in collecting various opinions, and it cannot quickly identify the real-time requirements of consumers (Gordon, 2008). Main difference between the relevant previous studies and the present one is that the latter focuses on provision of a coherent analysis construct for evaluation of LCS among service providers that derives customer perceptions from online reviews through a combination text-mining/DEA method.

3. Proposed methodology

Fig. 1 shows the proposed methodology, which is realized in the sequence of online review extraction, satisfaction metrics identification, satisfaction metrics quantification, and LCS evaluation. The online review extraction, satisfaction metrics identification, and satisfaction metrics quantification parts are the preliminary stage, while the CS evaluation part forms the backbone of the methodology.

The online review extraction part (Section 3.1) extracts customer reviews for each service provider from posting websites and cleans them for more highly accurate analysis. These reviews are used as the data source for LCS evaluation. As mentioned, the proposed methodology applies the DEA model in LCS evaluation, and DEA evaluates the performance of decision-making units according to input and output factors (i.e., performance measures). In DEA, input and output factors are generally determined by evaluators or decision makers. In our case, however, they are determined by deriving key service attributes from

the reviews. The satisfaction metrics identification part (Section 3.2) corresponds to the selection of input and output factors in DEA. The key service attributes, which represent the specific CS indicators, are derived by applying text-mining techniques such as term-frequency–inverse document frequency (TF-IDF). The reviews of each service provider are categorized by a clustering algorithm according to the key service attributes. We note that each cluster corresponds to a satisfaction metric (the same meaning as performance measures in DEA), which are used as the output factors in the DEA model; consequently, the number of clusters determines the number of satisfaction metrics. To quantify the satisfaction metrics, the satisfaction metrics quantification part (Section 3.3) computes the satisfaction score for each cluster through the sentiment analysis, which score is applied as the satisfaction metric value. The LCS evaluation part (Section 3.4) evaluates the LCS of each service provider through the DEA model by processing the satisfaction metrics as output factors.

3.1. Online review extraction

The customer reviews for each service provider are extracted from online review websites. Most online reviews are textual. Xiang et al. (2015) noted that in order to secure the validity of the text-data analysis results, it is necessary to refine the text-data by removing unnecessary words and synchronizing similar-meaning words. Ramadan et al. (2019) argued that words with the same meaning have to be unified into one word to increase the accuracy of the analysis. Thus, the extracted online reviews are pre-processed by removing numerous types of stop-word morphemes, such as special characters (e.g., %,!, @, *, #, etc.) and unclarified numbers, and by making consistent same-but-different expressions with errata (e.g., “checked-out,” “check out,” and “checking-out”) into the same expression (e.g., “check-out”). In general, a single review contains a number of sentences even on the same subject, and a single sentence may include multiple opinions even of the same entities. Thus, for a more fine-grained view of the different opinions as well as for derivation of the various feelings from reviews, we break down the reviews to the sentence level (Karamibekr & Ghorbani, 2013). Sentences are identified by recognizing characters and tabs (e.g.,!,?, ‘…’, ~, etc.).

3.2. Satisfaction metrics identification

The refined reviews are categorized into several clusters according to the key service attributes. The service attributes, which are evaluation indicators of CS, may vary depending on the type of service. For example, according to Sainaghi et al. (2019), CS with the service quality of a hotel is addressed by service attributes such as ‘service products’ (e.g., cleanliness, comfort, tangibility, and room amenities), ‘staff’ (e.g., reliability, responsiveness, assurance, and empathy), and ‘hotel traits’ (e.g., facilities, ambiance, certification, size, and location). In other words, ‘service products’, ‘staff’, and ‘hotel traits’ can be the key service attributes that determine the satisfaction with a hotel’s service quality.

The well-known TF-IDF algorithm is used to discover a number of key service attributes from reviews. The TF-IDF provides the statistical information required for evaluating the importance of words based on frequency, both document- and text-wise (Salton & Buckley, 1988), and the TF-IDF result shows that words having high weightage scores have a high frequency of occurrence in all reviews. From the TF-IDF result, the words are ranked in descending order from those having the highest weightage to those having the lowest weightage. In general, since the key service attribute is the object or outcome of service satisfaction, the words that are emotional expressions (e.g., good, happy, wonderful, or unpleasant) are excluded. We apply the concept of Pareto analysis to derive the key service attributes from the words having the higher weightages. A Pareto analysis helps to identify the top portion of causes that need to be addressed to resolve the majority of problems. It ranks the data classifications in descending order from the highest frequency of occurrence to the lowest. The ‘vital few’ items occupy a substantial portion (80 percent) of the cumulative percentage of occurrences, and the ‘useful many’ occupy only the remaining 20 percent of occurrences. From the words derived from the Pareto analysis, the words with similar meanings are grouped as key service attributes. For example, words with similar meanings such as ‘food’, ‘drink’, ‘dinner’, ‘breakfast’, and other subsidiary meal services are grouped and are determined as a key service attribute. Note that this means that one key service attribute may include a plurality of words with similar meanings.

Then, the reviews of each service provider are classified based on the key service attributes. For a specific illustration of the clustering process, let k, i, j be each index of key service attributes, sentences, and service providers, respectively. The clustering procedure searches to determine whether the set of words having a similar meaning to that of the k^{th} key service attribute exist in the i^{th} sentence of reviews for the j^{th} service provider. If a similar-meaning word grouped into the k^{th} key service attribute matches one of the words in the i^{th} sentence of reviews for the j^{th} service provider, the sentence is clustered into a cluster, G_{kj}. This procedure iterates over all service providers. Consequentially, each service provider has as many clusters as the number of key service attributes, and each cluster consists of a set of sentences including words grouped into the corresponding key service attribute. For example, assume that the words ‘food’, ‘drink’, ‘dinner’, and ‘breakfast’ are grouped into one key service attribute named ‘Food’, and assume also that there is a review consisting of the following 3 sentences: ‘Breakfast was fantastic’, ‘And, the staff were friendly’, ‘However, I feel the price is too high’. Since the sentence ‘Breakfast was fantastic’ contains the word ‘breakfast’, it is classified as one group corresponding to the key service attribute named ‘Food’.

3.3. Satisfaction metrics quantification

As noted, in evaluation of LCS by DEA, the clusters correspond to the satisfaction metrics. In order to quantify the satisfaction metrics, we compute the satisfaction score in each cluster through polarity identification applying sentiment analysis and use it as the value of satisfaction metrics. The sentiment analysis is defined as the task of finding opinions, such as perceptions and emotions, of authors about specific entities (Gundecha & Liu, 2012). With the constant increase of information in terms of opinions, emotions and feelings, sentiment analysis has attracted more and more interest from the scientific community (Liang & Dai, 2013). Note that the present study utilized, among the various sentiment analysis methods, a sentiment-lexicon-based sentiment analysis algorithm that determines the polarity (positive, negative or neutrality) of each sentiment expression by comparing it with the sentiment-lexicon resource. This method is the most commonly applied to the various domains, owing to its relatively simplicity and ease of implementation (Ding et al., 2008; Feldman, 2013). The sentiment-lexicon-based algorithm is, indeed, very simple. It detects the numbers of positive and negative words in each sentence by comparing them with the positive and negative lexicons in the sentiment-lexicon resources, and then determines the polarity of sentence; as a result, all sentences in each cluster are determined as either positive, negative, or neutral.

For a specific explanation of polarity identification by sentiment analysis, let L^p and L^n be a set of positive and negative lexicons in the sentiment-lexicon resources, respectively, and let S_i be the i^{th} sentence in a review. Each sentence is determined as positive, negative, or neutral by $S_i^* = w_i^p - w_i^n$, where w_i^p is the number of positive words found in S_i compared with L^p, and w_i^n is the number of negative words found in S_i compared with L^n. If $S_i^* > 0$, S_i is regarded as positive, and in the opposite case, it is regarded as negative; if equal (neither greater nor less than), S_i is neutral. To aid understanding, consider the following example review: “*My friend took me to the restaurant. The facilities were good. But the food was terrible*”. The review is split into three sentences by periods. Assume that sets of {‘*good*’} and {‘*terrible*’} are counted as positive and negative lexicons in the sentiment-lexicon resource, respectively. The first

Table 1
Top 20 airlines ranked by *SKYTRAX*.

No.	Airline	Number of reviews	Number of sentences
1	Qata*	1,777	14,916
2	Sing*	1,339	10,989
3	Nipp*	540	4,340
4	Cath*	1,414	12,474
5	Emir*	2,132	19,591
6	Eva_*	621	4,933
7	Hain*	417	2,934
8	Qant*	1,592	13,555
9	Luth*	1,718	13,645
10	Thai*	925	7,298
11	Japa*	366	3,014
12	Garu*	876	6,164
13	Swis*	835	6,864
14	Chin*	1,925	13,674
15	Aust*	582	4,596
16	AirN*	714	5,907
17	Bang*	388	2,643
18	KLM*	1,202	9,944
19	Brit*	2,234	20,372
20	AirA*	780	5,335
Sum		**22,377**	**183,188**

sentence is regarded as neutral ($S_1^* = 0$), the second sentence is regarded as positive ($S_2^* > 0$), and the third sentence is regarded as negative ($S_3^* < 0$), respectively. As a result, this review contains one positive sentence, one negative sentence, and one neutral sentence.

Based on the polarity analysis results, the satisfaction score is calculated as $Sc_{rj} = N_{rj}^p / \left(N_{rj}^n + N_{rj}^{ne}\right)$, where N_{rj}^p, N_{rj}^n, and N_{rj}^{ne} are the aggregation of the number of sentences regarded as positive, negative and neutral in the r^{th} cluster of the j^{th} service provider, respectively. A higher satisfaction score indicates a higher proportion of positive opinions among all, and thus represents a higher CS in the r^{th} satisfaction metric of the j^{th} service provider. In particular, when the satisfaction score is greater than 1, the positive opinions in the r^{th} satisfaction metric are superior to the sum of the rest (negative and neutral); it also represents that the CS in the r^{th} satisfaction metric is relatively higher than dissatisfaction or other opinions.

3.4. Levels of customer satisfaction (CS) evaluation

The proposed methodology considers multiple satisfaction metrics in evaluating LCS; thus it can be considered to be a multi-criteria decision-making (MCDM) problem. One key issue in MCDM is how to aggregate multiple metrics into a single measure in a proper manner by choosing a set of reasonable weights on multiple metrics. The DEA approach provides a way to systematically choose weights on multiple metrics where optimal weights are determined by solving mathematical (typically linear) programs. A DEA run determines a performance score for a decision-making unit (DMU), and the DEA can rank DMUs according to their performance scores. Basically, DMUs in DEA correspond to multiple alternatives in MCDM; input and output factors in DEA correspond to multiple performance metrics in MCDM, and the notion of performance in DEA corresponds to that of convex performance of MCDM (Bouyssou, 1999; Ramanathan, 2006). When DEA is used as an MCDM technique, it can be called multi-factor performance-measurement model.

Note that all satisfaction scores for satisfaction metrics are assumed to be positively related to LCS; that is, the larger the level of a service provider in terms of these satisfaction scores, the more likely it is to have a higher LCS. This assumption renders a DEA model as an output-maximizing multiplier with multiple outputs and without consideration of inputs. Thus, the proposed methodology uses the following model (1) by taking the service providers as DMUs and considering only the values of satisfaction metrics as outputs:

$$\begin{aligned} &LCS_k = Max \sum_{r=1}^{s} u_{rk} Sc_{rk} \\ &s.t. \sum_{r=1}^{s} u_{rk} Sc_{rj} \le 1,\ j = 1, \cdots, n \\ &u_{rk} \ge 0,\ \forall r \end{aligned} \tag{1}$$

where Sc_{rk} is the value of he r^{th} satisfaction metric of the evaluated service provider k, and u_{rk} is the weight given to the Sc_{rk}. The weighted additive function, LCS_k, aggregates the service scores of multiple satisfaction metrics, and its optimal value is used as the LCS for an evaluated k^{th} service provider. The function is maximized under the condition that the weighted sum of the satisfaction scores for each service provider, computed using the same set of weights, should be less than or equal to 1. Model (1) is similar to the reduced form of an output-maximizing multiplier CCR (Charnes, Cooper and Rhodes) model with multiple outputs and without consideration of inputs. Through evaluation by model (1), all service providers are classified into two groups based on their optimal scores. If a service provider is given an optimal score of 1, it is classified as the highest-LCS service provider. Otherwise, it is classified as a relatively lower-LCS service provider that needs to improve its customer service satisfaction.

4. Experimentation

As a demonstration of the proposed methodology, we apply it to real online reviews. The online reviews were extracted for the world's top 20 airlines, which had been selected as the best airlines for 2020 by *SKYTRAX* (https://www.airlinequality.com). Information on the title and content in the review comments was trawled and stored in a temporary database. We utilized the R programming language to extract the online reviews and conduct the text mining techniques. The process of crawling from an online website and building a database from the reviews thus obtained was performed using the '*rvest*' package (https://cran.r-project.org/web/packages/rvest/index.html) provided by the R programming language. The R programming language is an open-source software developed by the R Development Core Team of the R Foundation for statistical calculations (see (R Core Team, 2016) for more details on R programming).

A total of 22,377 reviews were extracted for the 20 airlines. Preprocessing was performed on the reviews by removing special characters, unclarified numbers, making consistent same-but-different expressions with errata into one, and identifying sentences by recognizing characters and tabs. After cleaning the reviews, a total of 183,188 sentences were extracted as shown in Table 1.

When applying the TF-IDF algorithm to all of the refined reviews, the weightage values were calculated for a total of 987 words. After removing emotional and concise expressions, there was a total of 412 words. The distribution of the 412 words in terms of the percentages of TF-IDF scores in descending order, obtained through Pareto analysis, is presented in Fig. 2. Eighteen (18) words were selected based on the cumulative probability of about 80% (i.e., the summation of their TF-IDF scores was 81% of the total score). The four most distinctive key service attributes were chosen by categorizing the words with similar meanings into the same group. The key service attributes were named 'Services', 'Foods', 'Seats', and 'Entertainments'. The key service attribute, 'Services,' included a number of similar variants, such as service (TF-IDF score of 763), crew, cabin, staff, and other subsidiary services. The variants for 'Foods' were food (745), meal, drink, dinner, tea, stack, and other eating-related names. Several monetized representations were discovered for 'Seats,' including seat (678), space, bag, and luggage. 'Entertainments' represented entertainment (376), screen, wifi, and other subsidiary entertainment.

To verify the results of the categorization of key service attributes, we compared them with the results of a topic analysis by utilizing the

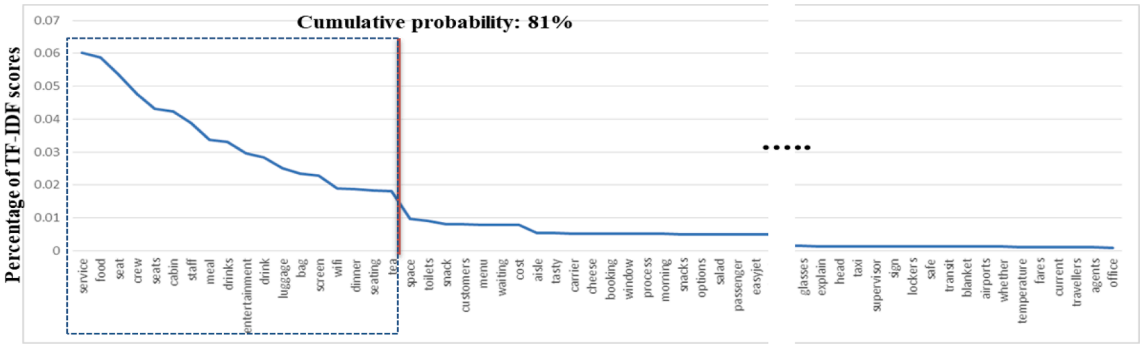
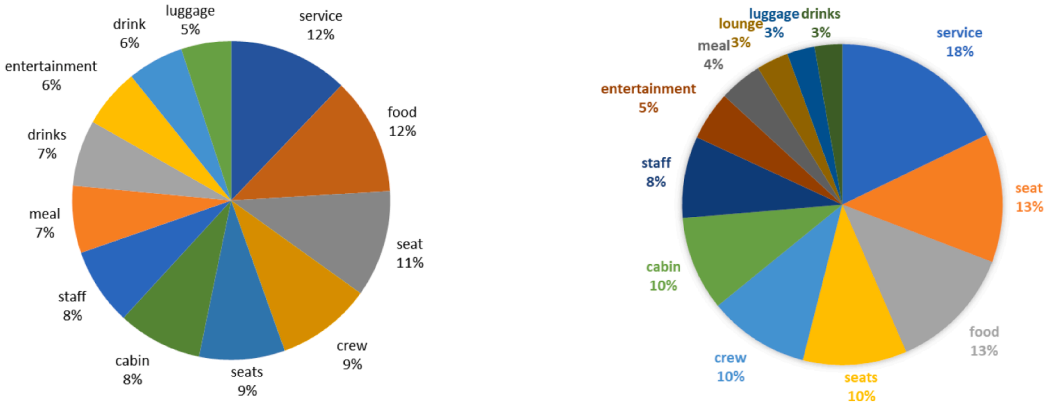


Fig. 2. Percentages of TF-IDF scores for discovered 412 words.



(a) Results from TF-IDF (b) Results from LDA analysis

Fig. 3. Comparison of results for categorized key service attribute words by TF-IDF and LDA analysis.

Table 2
Numbers of sentences categorized into four key service attributes.

No	Airline	Numbers of sentences			
		Services	Foods	Seats	Entertainments
1	Qata*	2,964	1,904	1,828	711
2	Sing*	2,263	1,624	1,373	629
3	Nipp*	811	648	579	202
4	Cath*	2,386	1,785	1,634	520
5	Emir*	4,072	2,489	2,200	924
6	Eva_*	901	720	611	226
7	Hain*	640	386	306	174
8	Qant*	2,730	1,763	1,507	572
9	Luth*	2,446	1,514	1,769	593
10	Thai*	1,538	1,107	941	361
11	Japa*	593	453	384	151
12	Garu*	1,443	879	590	299
13	Swis*	1,236	743	941	263
14	Chin*	2,881	1,855	1,458	815
15	Aust*	850	591	616	163
16	AirN*	1,054	651	874	266
17	Bang*	514	360	204	58
18	KLM*	1,764	1,060	1,388	372
19	Brit*	3,758	2,476	2,798	631
20	AirA*	756	323	618	76
Mean		**35,600**	**23,331**	**22,619**	**8,006**

Table 3
Satisfaction scores corresponding to four service attributes for 20 airlines.

No	Airline	Satisfaction score			
		Services	Foods	Seats	Entertainments
1	Qata*	0.88	1.46	1.51	2.53
2	Sing*	0.91	1.55	1.81	2.57
3	Nipp*	0.73	1.46	1.44	1.54
4	Cath*	0.92	1.60	1.61	2.41
5	Emir*	0.84	1.41	1.19	2.05
6	Eva_*	0.80	1.37	1.40	2.95
7	Hain*	0.90	1.47	1.38	2.09
8	Qant*	0.96	1.65	1.39	2.91
9	Luth*	0.98	1.67	1.36	2.59
10	Thai*	1.02	1.53	1.64	2.24
11	Japa*	0.80	1.72	1.86	2.33
12	Garu*	0.92	2.02	1.32	2.68
13	Swis*	1.03	1.60	1.44	3.30
14	Chin*	0.90	1.61	1.55	2.62
15	Aust*	0.91	1.65	1.24	2.60
16	AirN*	0.94	1.80	1.27	2.08
17	Bang*	0.80	1.58	1.18	1.19
18	KLM*	0.95	1.77	1.49	2.57
19	Brit*	1.15	1.98	1.46	3.38
20	AirA*	0.93	1.54	1.23	2.00
Mean		**0.91**	**1.62**	**1.44**	**2.43**

LDA (Latent Dirichlet Allocation) method. Fig. 3 shows a comparison of the 12 words of higher weightage by TF-IDF (presented in Fig. 2) with the top 12 words of higher per-topic-per-word probability (called β (beta)) as computed by LDA. Refer to Appendix A for details on the LDA analysis process and results. In particular, since the purpose was to compare the derived key service words with each other, we focused on computing the cumulative β of words into the whole topics rather than interpreting the topic in the LDA analysis. As shown in Fig. 3, the percentage rankings of words from TF-IDF and LDA were somewhat different, but the distinctive words derived by categorizing the words

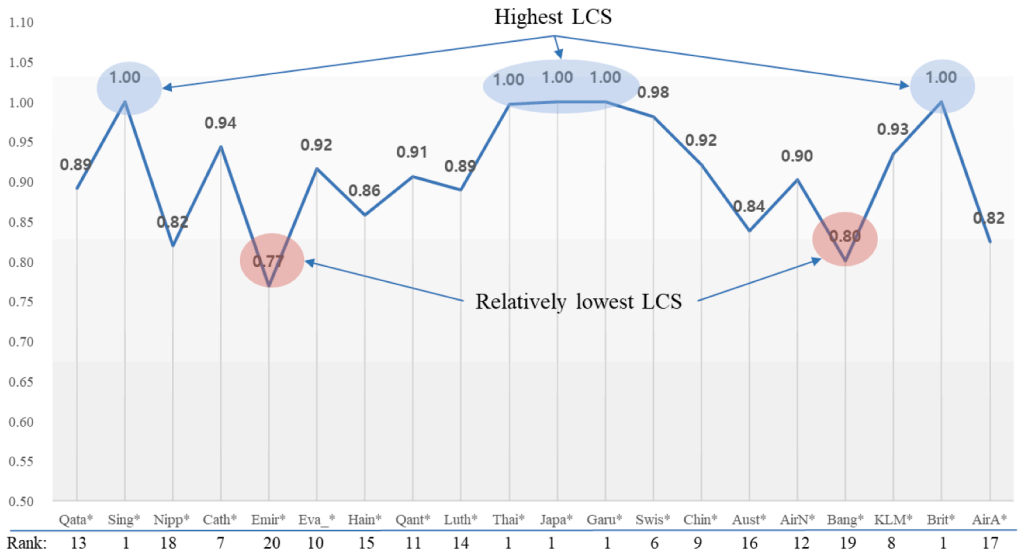


Fig. 4. LCS results for 20 airlines.

Table 4
Service improvement guidelines.

No	Airline	Required improvement (satisfaction score)				Required improvement (positive sentence)				Target set (No)
		Services	Foods	Seats	Entertainments	Services	Foods	Seats	Entertainments	
1	Qata*	0.11	0.24	0.18	0.31	174	186	131	62	2, 1
2	Sing*	–	–	–	–	–	–	–	–	–
3	Nipp*	0.16	0.32	0.32	1.06	75	84	76	84	2, 11, 1
4	Cath*	0.06	0.10	0.10	0.39	75	69	63	59	2, 11, 1
5	Emir*	0.25	0.47	0.35	1.14	552	486	351	345	2, 19
6	Eva_*	0.30	0.53	0.13	0.27	150	161	33	15	2, 19
7	Hain*	0.15	0.33	0.23	0.95	51	51	29	53	2, 19
8	Qant*	0.14	0.24	0.14	0.30	195	160	88	44	2, 19
9	Luth*	0.13	0.22	0.17	0.63	161	125	127	104	2, 11, 19
10	Thai*	–	0.22	–	0.71	–	96	–	79	2, 19
11	Japa*	–	–	–	–	–	–	–	–	–
12	Garu*	–	–	–	–	–	–	–	–	–
13	Swis*	0.11	0.37	0.03	0.07	67	105	12	4	2, 19
14	Chin*	0.08	0.13	0.13	0.23	121	92	74	52	2, 11, 19
15	Aust*	0.23	0.32	0.24	0.73	102	71	66	33	11, 19
16	AirN*	0.13	0.19	0.14	1.04	71	44	54	89	12, 19
17	Bang*	0.34	0.39	0.29	2.15	97	54	27	56	11, 19
18	KLM*	0.08	0.13	0.11	0.46	72	50	61	48	11, 19
19	Brit*	–	–	–	–	–	–	–	–	–
20	AirA*	0.20	0.41	0.26	1.31	78	52	72	33	2, 19

with similar meanings into the same group were similar to each other. In addition, we assessed the overlap of the top 40 words from TF-IDF and LDA analysis. The Jaccard coefficient was used to test the degree of overlap between the TF-IDF and LDA analyses. $N(A_{tf-idf})$ representing the set of words derived from TF-IDF and $N(A_{lda})$ representing the set of words derived from LDA, we calculated the Jaccard coefficient as Model (2) and obtained a Jaccard coefficient of 0.84. The higher the Jaccard coefficient's value, the higher the degree of overlap between the two sets of words.

$$JC = \frac{\left|N(A_{tf-idf} \cap A_{lda})\right|}{\left|N(A_{tf-idf} \cup A_{lda})\right|} \quad (2)$$

For the categorizing of the reviews according to the key service attributes, a pattern-matching function, namely "match" in the R programming language (https://stat.ethz.ch/R-manual/R-devel/library/base/html/match.html), was applied. The sentences in the reviews for the 20 airlines were categorized and arranged by the key service attributes, as presented in Table 2. The mean of sentences in 'Services' clusters (i.e., service attribute) was higher than in the other three clusters, and it was observed that consumers had more opinions on service from cabin crews and staff than from the other service attributes.

This study used Senticnet 5.0 (https://sentic.net/), commonly employed for semantic and polarity analysis of text, for the sentiment-lexicon resources. The satisfaction scores corresponding to the four clusters are summarized in Table 3. It was observed that the satisfaction scores were high in the order of 'Entertainments', 'Foods', 'Seats', and 'Services', on average. Specifically, the customers' positive opinions on 'Foods', 'Seats', and 'Entertainments', which had mean satisfaction scores higher than 1, were more than the sum of the dissatisfaction and neutrality opinions. On the other hand, the customers' positive opinions on 'Services' were lower than the sum of the other opinions (negative and neutrality). It can be interpreted that the airline manager needs to make more efforts toward improving in-flight and ground services in order to increase the CS with this service attribute.

Next, the satisfaction scores of the four clusters were used as outputs in model (1). The LCS evaluation results for the 20 airlines are plotted in

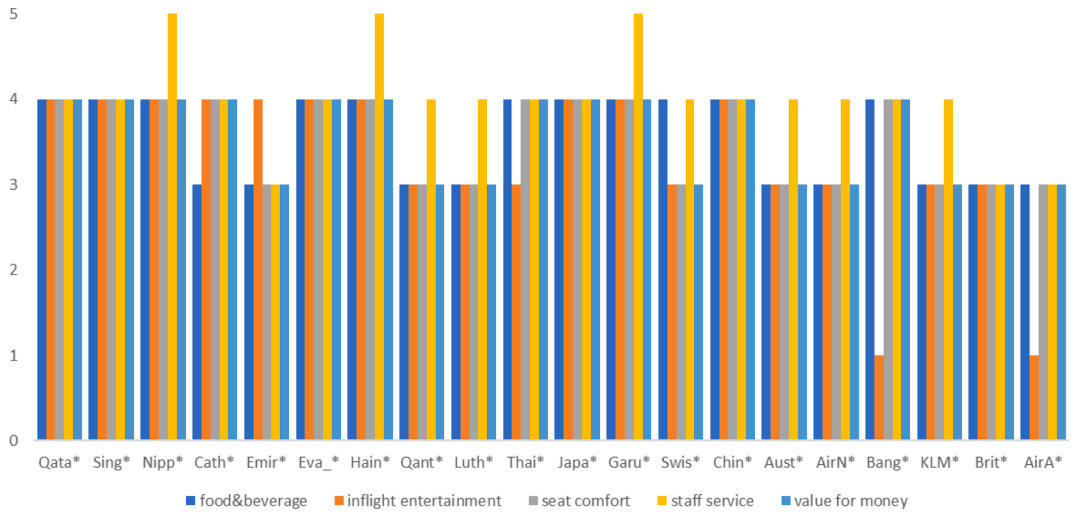
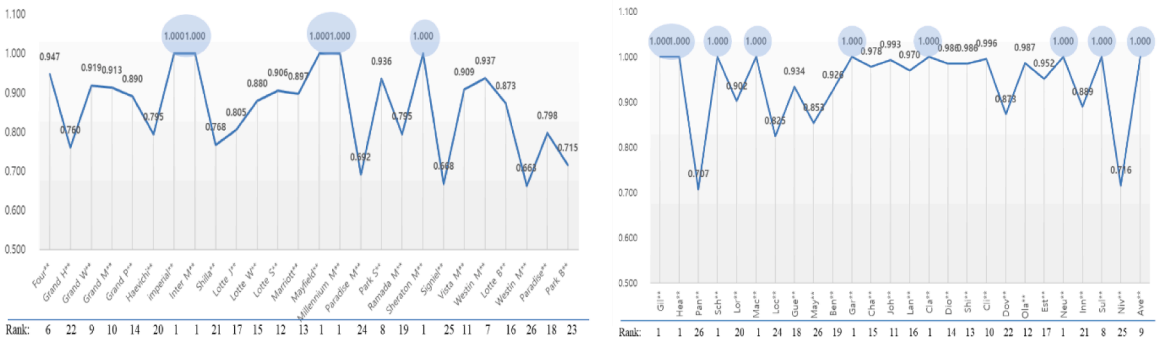


Fig. 5. Results of service ratings for 20 airlines from the website, *arilinequality.com*.



(a) 26 5-star hotels in South Korea (b) 26 cosmetic brands worldwide

Fig. 6. LCS results of 26 hotels and cosmetic brands.

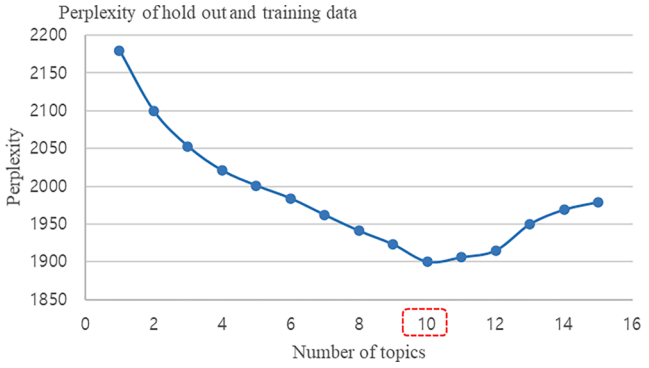


Fig. A.1. Result of Perplexity.

Fig. 4. Five of the airlines, namely Sing*, Thai*, Japa*, Garu*, and Brit*, were evaluated as having the highest LCS in terms of the four service attributes. The remaining 15 airlines are classified as relatively lower-LCS airlines needing to improve their service satisfaction. Specifically, the airlines Emir* and Bang* were evaluated to have relatively lower LCS, because their satisfaction scores were lower than the mean in all four service attributes, as shown in Table 3. The airline with the highest LCS is likely to be considered a best-practices airline, and would have a high chance of becoming a service-improvement target for airlines with lower LCS.

Conventional DEA can provide service improvement guidelines for the relatively lower-LCS airlines to enable them to improve their CS. Table 4 shows the service improvement (benchmarking) targets, guidelines and number of resources that need to be improved for the relatively lower-LCS airlines to be able to reach their improvement target. Regarding airline Emit*, whose LCS was 0.77, a service improvement target set composed of two airlines, Sing* and KLM*, was assigned. Accordingly, it needs to improve its CS by 0.25, 0.47, 0.35, and 1.14 in the four service attributes, respectively, to reach its target. The required amount of improvement for Emit* was highest in the service attribute 'entertainments', followed by 'Foods', 'Seats', and 'Services'. Note that the required amount of improvement is a ratio value, not a quantitative one. In terms of required positive sentences and quantitative values, Emit* should increase the number of positive opinions by at least 552, 486, 351, and 345 (sentences) in order to improve its satisfaction scores by 0.25, 0.47, 0.35, and 1.14 in the service attributes of 'Services', 'Foods', 'Seats', and 'Entertainments', while maintaining the current non-positive opinions.

The website *arilinequality.com* provides a service rating survey system on a 5-point scale in 0.5-point increments to enable customers to record their satisfaction with each of the predetermined five service attributes including 'food&beverage', 'inflight entertainment', 'seat comfort', 'staff service', and 'value for money. These are similar to the service attributes estimated in the present study, except for 'value for money'. Even though *airlinequality.com* further suggests 'value for money' as a service attribute, it was found that the number of related sentences comprising words such as 'money', 'value', 'price', 'pricing', and 'value of money' was very low (2,015 sentences) relative to the

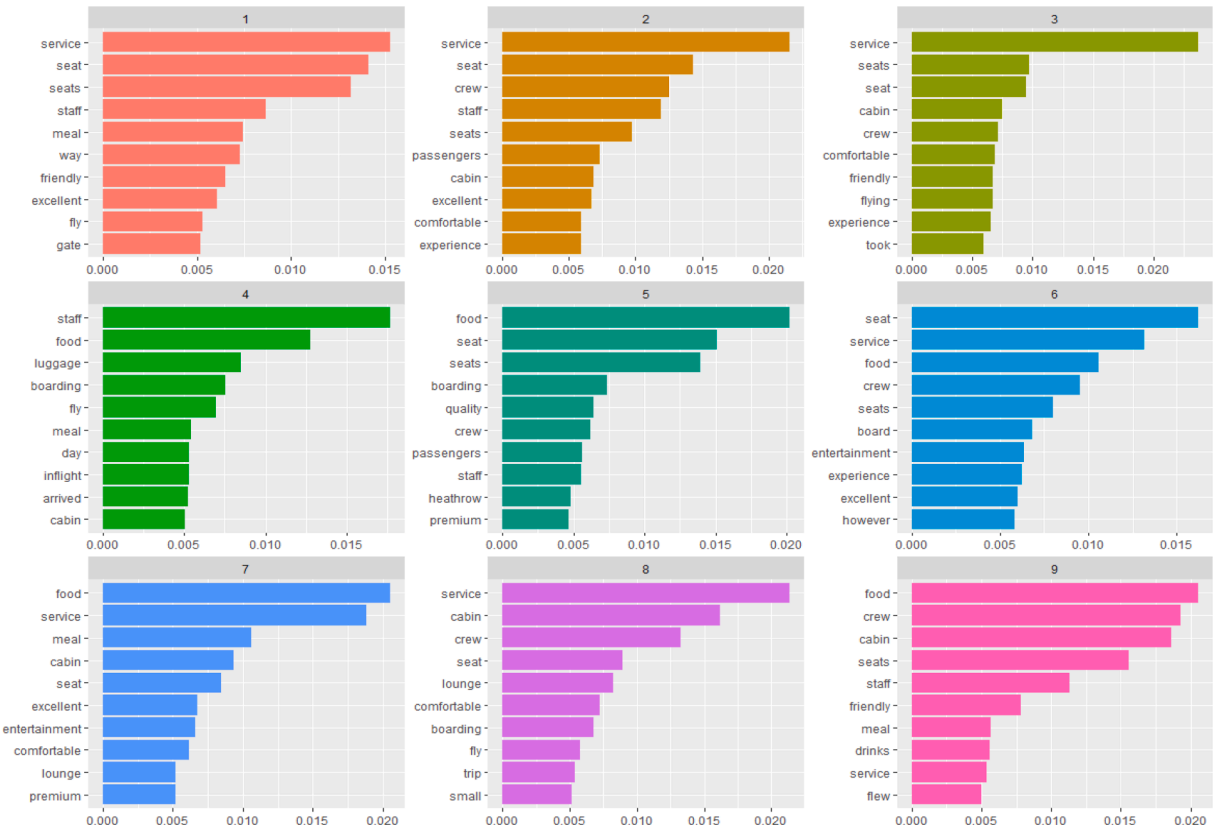


Fig. A.2. Top 10 words for each topic when number of topics is 9.

others (e.g. 35,600 sentences for 'Services', 23,331 sentences for 'Foods', 22,619 sentences for 'Seats', and 8,006 sentences for 'Entertainments,' as indicated in Table 2). In the present study, LCS was estimated additionally for 'value for money', and the result was compared with those in Fig. 2. Although there was a slight difference in the LCS scores for Aust* and Qant*, the correlation coefficient was estimated to be 0.896, indicating that both LCS scores had considerable similarity over the airlines. From this comparison, we could say that the minor difference in the number of service attributes had an insignificant influence on the estimation of LCS. What is more important, however, is that the service rating score was rated quite similar across all airlines in Fig. 5, which did not distinguish the differences in the positive feedback from customers. The results of the service rating survey system revealed the necessity of a more discriminating customer feedback rating for airlines. This issue also was raised in a previous study (Fernández and Bedia, 2004), which questioned the usefulness of star-ratings as an indicator of service quality, and revealed that the service rating survey system is in fact not very useful as an evaluation indicator of service quality.

To validate the generalizability of the proposed framework, it was applied to the hotel and cosmetic domains, where customer review analysis is considered important, and LCS of hotels and cosmetic brands was evaluated. A total of 21,014 reviews were extracted from tripadvisor.com and ebooking.com websites for 26 randomly selected 5-star hotels in South Korea, and a total of 8,357 reviews were extracted from MakeupAlley for 26 cosmetic brands worldwide, between March 2021 and February 2023. The five most distinctive key service attributes for hotels were identified as facilities, locations, lodging charges, meals, and guest service. Similarly, and the five most distinctive key service attributes for cosmetic brands were identified as moisturizing ability, color, fragrance, formulation, and absorption power. Each of their LCS results is shown in Fig. 6. Through the results presented in Fig. 6, it can be indirectly inferred that the proposed framework may also be applicable and useful in other domains..

5. Implications and discussion

The case study showed that the proposed text-mining/DEA combination method can be effectively utilized to determine LCS from online reviews. Above all, the proposed approach provided for a multi-service dimension-based LCS evaluation, showing that more sophisticated evaluation and analysis than the service rating score provided by many websites is indeed possible. Through experimentation, we identified that the proposed method discriminates the LCS more strictly in contrast to the star-rating survey system of *airlinequality.com* website was rated quite similar across all airlines. And we found that the satisfaction scores were high in the order of 'Entertainments', 'Foods', 'Seats', and 'Services' on average, and that customers' positive opinions on 'Foods', 'Seats', and 'Entertainments' were more than the sum of the other opinions, indicating that airline managers need to make more efforts to improve in-flight and ground services to increase CS. In addition, even though the *arilnequality.com* website suggested value for money as a key service attribute, we found that customers have few expressions of satisfaction or dissatisfaction with this attribute, indicating that the website managers need to discuss whether to continue to include the attribute value for money as one of the key service attributes.

The proposed approach can be applied to various service practices, though it is important to select proper DEA and sentiment analysis approaches. This study employed only the output-oriented CCR model with multiple outputs and a constant input along with sentiment-lexicon-resource-based sentiment analysis to evaluate LCS from online reviews; there are, however, various DEA models and sentiment analysis approaches available. Selecting a proper DEA model and sentiment analysis approach according to the specific purpose of the analysis and the domain properties may enhance the applicability and diagnostic power of the proposed approach.

Also, the following two issues need to be discussed with regard to selection of the appropriate DEA model. First, the weight to be assigned to the four service attributes is not absolutely reflected in evaluating convinced LCS for all airlines. The main advantage of a general DEA is that it does not require information on the weights assigned to the inputs and outputs. However, this advantage leads to distorted results via assignment of extremely favorable weights to specific inputs or outputs for a particular DMU to achieve higher efficiency result. The weights to inputs and outputs are chosen in a manner that assigns a best set of weights to each DMU to obtain a maximized efficiency score. However, if commonly accepted views on the weights of inputs and outputs are taken into account, the weight flexibility in DEA leads to unrealistic efficiency scores (Allen et al., 1997). Adopting weights-restriction DEA models such as assurance regions (AR) would be expected to produce more realistic LCS evaluation results by considering the preference of weight assignments according to the various service industries (Thompson et al., 1986). Second, the efficiency score derived by the general DEA model is limited to a maximum of 1, and thus, the general DEA models do not allow for ranking of efficient DMUs. Some full-ranking techniques have been developed, such as cross-efficiency (Sexton et al., 1986) and super-efficiency (Andersen & Petersen, 1993). Utilizing these models can support more realistic competitiveness evaluation by enhancing discriminant power.

In addition, sentiment analysis can be classified into three categories according to the process learning, which are the machine-learning approach, the lexicon-based approach, and the hybrid approach combining both machine learning and lexicon-based approaches as mentioned before. Notably, the lexicon-based approach is divided into two parts, a dictionary-based approach and a corpus-based approach, which means that positive-ratio results can vary depending on the type of sentiment analysis approach applied (Mukwazvure & Supreethi, 2015). Thus, calculation of more sophisticated and reliable positive-ratio results by sentiment analysis needs to be considered in any discussion of the above-mentioned approaches.

6. Conclusions

This study derived a new data envelopment analysis (DEA)-based level of customer satisfaction (LCS) evaluation methodology based on the exploitation of online review text data. The proposed methodology quantifies LCS by the term-frequency–inverse document frequency (TF-IDF) algorithm, a text-mining technique, according to multiple service attributes that significantly affect customers' service experience; subsequently, the LCS of the service providers is evaluated by the DEA method. To illustrate the efficacy of the proposed approach, an empirical case study applying it to the top 20 global airlines was conducted, the results of which indicated how it can effectively utilize online reviews for LCS evaluation. In addition, we showed the applicability of the proposed approach in examining a benchmark reference set to provide service improvement guidelines for the relative lower-LCS airlines.

The proposed approach is expected to replace the questionnaire survey method that has been widely applied by service-provider managers for analysis of customer satisfaction (CS) and its improvement. Furthermore, we hope that it can be used as a source of fundamental data applicable to efforts to improve both airline service competitiveness and provision of systematic services quality. Despite its valuable contributions, this study also has several drawbacks. First, sentiment scoring is very sensitive depending on the accuracy of the sentiment-lexicon resource applied. Although this study applied the general sentiment-lexicon resource widely utilized in other studies, a more highly accurate sentiment dictionary will be needed for more sophisticated sentiment analysis. Second, this study focused on the LCS evaluation from online reviews analysis and did not consider a detailed content analysis to deduce the reason of customer's negative reviews. However, as the results of the LCS evaluation must ultimately lead to the improvement of the service level of the service providers, a detailed content analysis is essential. Both of these issues will be addressed in upcoming research.

CRediT authorship contribution statement

Jaehun Park: Conceptualization, Methodology, Software, Investigation, Writing – original draft, Writing – review & editing, Supervision.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The data that has been used is confidential.

Acknowledgements

This work was supported by National Research Foundation grant (NRF- 2021R1F1A1052422) of Republic of Korea.

Appendix A

LDA is a particularly popular method for fitting a topic model; it treats each document as a mixture of topics, and each topic as a mixture of words. LDA is used to obtain the distribution of the keywords for each topic and the distribution of the topics in the text. In general, it needs to determine the number of topics for a more reliable analysis result in applying the topic analysis. If the number of topics is too many, there is a high possibility that excessively subdivided topics will appear, whereas if the number of topics is too small, the interpretation of the topic may become ambiguous. Among the studies on the determination of the number of topics (Cao et al., 2009; Zhao et al., 2015), this study utilized the Perplexity method, which has been widely utilized in many studies. The topic modeling will be reliable when the number of topics is determined at the lower point of the Perplexity index. We set the number of topics as 10, according to the lowest value in the Perplexity index shown in Fig. A.1. The reason the number of topics was somewhat small is that the focus of this case study was limited to customer opinions of aviation services. Fig. A.2 shows the sort-out results of the top 10 words for each topic. By removing the words related to emotional expressions, the topics could be categorized into services of staff, comfort or convenience of seats, and quality of food and entertainment.

References

Abubakar, B., & Mavondo, F. (2014). Tourism destinations: Antecedents to customer satisfaction and positive word-of-mouth. *Journal of Hospitality Marketing & Management, 23*(8), 833–864. https://doi.org/10.1080/19368623.2013.796865

Allen, R., Athanassopoulos, A., Dyson, R., & Thanassoulis, E. (1997). Weights restrictions and value judgements in data envelopment analysis: Evolution, development and future directions. *Annals of Operations Research, 73*, 13–34. https://doi.org/10.1023/a:1018968909638

Andersen, P., & Petersen, N. C. (1993). A procedure for ranking efficient units in data envelopment analysis. *Management Science, 39*(10), 1261–1264. https://doi.org/10.1287/mnsc.39.10.1261

Archak, N., Ghose, A., & Ipeirotis, P. G. (2011). Deriving the pricing power of product features by mining consumer reviews. *Management Science, 57*(8), 1485–1509. https://doi.org/10.1287/mnsc.1110.1370

Arora, S., Joshi, M., & Rosé, C. (2009, June). Identifying types of claims in online customer reviews. In *Proceedings of human language technologies: The 2009 annual conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers* (pp. 37-40).

Ba, S., & Johansson, W. C. (2008). An exploratory study of the impact of e-service process on online customer satisfaction. *Production and Operations Management, 17*(1), 107–119. https://doi.org/10.3401/poms.1070.0006

Baydogan, C., & Alatas, B. (2019). Detection of customer satisfaction on unbalanced and multi-class data using machine learning algorithms. In *2019 1st international*

informatics and software engineering conference (UBMYK). 10.1109/ubmyk48245.2019.8965631.
Bayraktar, E., Tatoglu, E., Turkyilmaz, A., Delen, D., & Zaim, S. (2012). Measuring the efficiency of customer satisfaction and loyalty for mobile phone brands with DEA. *Expert Systems with Applications, 39*(1), 99–106. https://doi.org/10.1016/j.eswa.2011.06.041
Bouyssou, D. (1999). Using DEA as a tool for MCDM: Some remarks. *The Journal of the Operational Research Society, 50*(9), 974. https://doi.org/10.2307/3010194
Cao, J., Xia, T., Li, J., Zhang, Y., & Tang, S. (2009). A density-based method for adaptive LDA model selection. *Neurocomputing, 72*(7–9), 1775–1781. https://doi.org/10.1016/j.neucom.2008.06.011
Ding, X., Liu, B., & Yu, P. S. (2008, February). A holistic lexicon-based approach to opinion mining. In *Proceedings of the 2008 international conference on web search and data mining* (pp. 231-240).
Fernández, M. C. L., & Bedia, A. M. S. (2004). Is the hotel classification system a good indicator of hotel quality?: An application in Spain. *Tourism Management, 25*(6), 771–775. https://doi.org/10.1016/s0261-5177(04)00133-5
Feldman, R. (2013). Techniques and applications for sentiment analysis. *Communications of the ACM, 56*(4), 82–89. https://doi.org/10.1145/2436256.2436274
Gan, Q., Ferns, B. H., Yu, Y., & Jin, L. (2016). A text mining and multidimensional sentiment analysis of online restaurant reviews. *Journal of Quality Assurance in Hospitality & Tourism, 18*(4), 465–492. https://doi.org/10.1080/1528008x.2016.1250243
Ghose, A., & Ipeirotis, P. G. (2011). Estimating the helpfulness and economic impact of product reviews: Mining text and reviewer characteristics. *IEEE Transactions on Knowledge and Data Engineering, 23*(10), 1498–1512. https://doi.org/10.1109/tkde.2010.188
Gordon, A. (2008). *Future savvy: Identifying trends to make better decisions, manage uncertainty, and profit from change*. New York, NY: American Management Association.
Gremler, D. D., Gwinner, K. P., & Brown, S. W. (2001). Generating positive word-of-mouth communication through customer-employee relationships. *International Journal of Service Industry Management, 12*(1), 44–59. https://doi.org/10.1108/09564230110382763
Gundecha, P., & Liu, H. (2012). Mining social media: A brief introduction. *TutORials in Operations Research, 2012*, 1–17. https://doi.org/10.1287/educ.1120.0105
Haddara, M., Hsieh, J., Fagerstrøm, A., Eriksson, N., & Sigurðsson, V. (2019). Exploring customer online reviews for new product development: The case of identifying reinforcers in the cosmetic industry. *Managerial and Decision Economics, 41*(2), 250–273. https://doi.org/10.1002/mde.3078
Hennig-Thurau, T., Gwinner, K. P., Walsh, G., & Gremler, D. D. (2004). Electronic word-of-mouth via consumer-opinion platforms: What motivates consumers to articulate themselves on the Internet? *Journal of Interactive Marketing, 18*(1), 38–52. https://doi.org/10.1002/dir.10073
Hensher, D. A., & Prioni, P. (2002). A service quality index for area-wide contract performance assessment. *Journal of Transport Economics and Policy (JTEP), 36*(1), 93–113.
Hsu, S. (2008). Developing an index for online customer satisfaction: Adaptation of American Customer Satisfaction Index. *Expert Systems with Applications, 34*(4), 3033–3042. https://doi.org/10.1016/j.eswa.2007.06.036
Karamibekr, M., & Ghorbani, A. A. (2013). Sentence subjectivity analysis in social domains. In *2013 IEEE/WIC/ACM international joint conferences on web intelligence (WI) and intelligent agent technologies (IAT)*. 10.1109/wi-iat.2013.39.
Kang, D., & Park, Y. (2014). Review-based measurement of customer satisfaction in mobile service: Sentiment analysis and VIKOR approach. *Expert Systems with Applications, 41*(4), 1041–1050. https://doi.org/10.1016/j.eswa.2013.07.101
Kotler, P., & Keller, K. (2009). *Marketing management 13th ed.* Upper Saddle River: NJ: Pearson Education.
Lee, T. Y., & Bradlow, E. T. (2011). Automated marketing research using online customer reviews. *Journal of Marketing Research, 48*(5), 881–894. https://doi.org/10.1509/jmkr.48.5.881
Li, H., Ye, Q., & Law, R. (2013). Determinants of customer satisfaction in the hotel industry: An application of online review analysis. *Asia Pacific Journal of Tourism Research, 18*(7), 784–802. https://doi.org/10.1080/10941665.2012.708351
Liang, P. W., & Dai, B. R. (2013). Opinion mining on social media data. In *2013 IEEE 14th International Conference on Mobile Data Management* (pp. 91–96). Milan, Italy: IEEE. https://doi.org/10.1109/mdm.2013.73.
Mariani, M. M., & Borghi, M. (2018). Effects of the Booking.com rating system: Bringing hotel class into the picture. *Tourism Management, 66*, 47–52. https://doi.org/10.1016/j.tourman.2017.11.006
Mariani, M. M., & Visani, F. (2019). Embedding eWOM into efficiency DEA modelling: An application to the hospitality sector. *International Journal of Hospitality Management, 80*, 1–12. https://doi.org/10.1016/j.ijhm.2019.01.002
Martín-Cejas, R. (2006). Tourism service quality begins at the airport. *Tourism Management, 27*(5), 874–877. https://doi.org/10.1016/j.tourman.2005.05.005
Mukwazvure, A., & Supreethi, K. (2015). A hybrid approach to sentiment analysis of news comments. In *2015 4th international conference on reliability, Infocom technologies and optimization (ICRITO) (Trends and Future Directions)*. 10.1109/icrito.2015.7359282.
Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up? Sentiment classification using machine learning techniques. *arXiv preprint cs/0205070*.
Parasuraman, A., Zeithaml, V. A., & Berry, L. (1988). SERVQUAL: A multiple-item scale for measuring consumer perceptions of service quality. *1988*, *64*(1), 12-40.
Peterson, R. A., & Merino, M. C. (2003). Consumer information search behavior and the internet. *Psychology and Marketing, 20*(2), 99–121. https://doi.org/10.1002/mar.10062
R Core Team. (2016). R: A language and environment for statistical computing. (R Foundation for Statistical Computing) Retrieved 12 20, 2019, from http://www.R-project.org/.
Ramadan, S., Ibrahim Baqapuri, H., Roecher, E., & Mathiak, K. (2019). Process mining of logged gaming behavior. In *2019 international conference on process mining (ICPM)*. 10.1109/icpm.2019.00019.
Ramanathan, R. (2006). Data envelopment analysis for weight derivation and aggregation in the analytic hierarchy process. *Computers & Operations Research, 33*(5), 1289–1307. https://doi.org/10.1016/j.cor.2004.09.020
Ramanathan, U., Subramanian, N., & Parrott, G. (2017). Role of social media in retail network operations and marketing to enhance customer satisfaction. *International Journal of Operations & Production Management, 37*(1), 105–123. https://doi.org/10.1108/ijopm-03-2015-0153
Sainaghi, R., Phillips, P., Baggio, R., & Mauri, A. (2019). Hotel performance: Rigor and relevant research topics. *International Journal of Hospitality Management, 78*, 13–26. https://doi.org/10.1016/j.ijhm.2018.11.008
Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management, 24*(5), 513–523. https://doi.org/10.1016/0306-4573(88)90021-0
Sari, E. Y., Wierfi, A. D., & Setyanto, A. (2019). Sentiment analysis of customer satisfaction on transportation network company using Naive Bayes Classifier. In *2019 international conference on computer engineering, network, and intelligent multimedia (CENIM)*. 10.1109/cenim48368.2019.8973262.
Sexton, T. R., Silkman, R. H., & Hogan, A. J. (1986). Data envelopment analysis: Critique and extensions. *New Directions for Program Evaluation, 1986*(32), 73–105. https://doi.org/10.1002/ev.1441
Somprasertsri, G., & Lalitrojwong, P. (2010). Mining feature-opinion in online customer reviews for opinion summarization. *Journal of Universal Computer Science, 16*, 938–955.
Thompson, R. G., Singleton, F. D., Thrall, R. M., & Smith, B. A. (1986). Comparative site evaluations for locating a high-energy physics lab in Texas. *Interfaces, 16*(6), 35–49. https://doi.org/10.1287/inte.16.6.35
Webster, F. E., Jr. (1992). The changing role of marketing in the corporation. *Journal of Marketing, 56*(4), 1–17. https://doi.org/10.2307/1251983
Xiang, Z., Schwartz, Z., Gerdes, J. H., & Uysal, M. (2015). What can big data and text analytics tell us about hotel guest experience and satisfaction? *International Journal of Hospitality Management, 44*, 120–130. https://doi.org/10.1016/j.ijhm.2014.10.013
Yang, C. C., Jou, Y. T., & Cheng, L. Y. (2009). Using integrated quality assessment for hotel service quality. *Quality & Quantity, 45*(2), 349–364. https://doi.org/10.1007/s11135-009-9301-4
Yi, J., & Niblack, W. (2005). Sentiment mining in WebFountain. In *21st international conference on data engineering (ICDE'05)*. 10.1109/icde.2005.132.
Yoon, Y., & Ha, D. (2010). The effects of source credibility to internet WOM communication on brand equity: Focused on customers of deluxe hotel. *Journal of The Korean Academic Society of Hospitality Administration, 19*(4), 81–97. In Korean.
Zhan, J., Loh, H. T., & Liu, Y. (2009). Gather customer concerns from online product reviews – A text summarization approach. *Expert Systems with Applications, 36*(2), 2107–2115. https://doi.org/10.1016/j.eswa.2007.12.039
Zhao, W., Chen, J. J., Perkins, R., Liu, Z., Ge, W., Ding, Y., & Zou, W. (2015). A heuristic approach to determine an appropriate number of topics in topic modeling. *BMC Bioinformatics, 16*(S13). https://doi.org/10.1186/1471-2105-16-s13-s8
Zhao, Y., Xu, X., & Wang, M. (2019). Predicting overall customer satisfaction: Big data evidence from hotel online textual reviews. *International Journal of Hospitality Management, 76*, 111–121. https://doi.org/10.1016/j.ijhm.2018.03.017