Approach for Importance–Performance Analysis of Product Attributes From Online Reviews

Junegak Joung
Enterprise Systems Optimization Laboratory,
Department of Industrial and Enterprise Systems Engineering,
University of Illinois at Urbana-Champaign,
Urbana, IL 61801;
Department of Industrial Engineering,
Ulsan National Institute of Science and Technology,
Ulsan 44919, Republic of Korea
e-mails: junegak@illinois.edu;
june30@unist.ac.kr

Harrison M. Kim[1]
Enterprise Systems Optimization Laboratory,
Department of Industrial and Enterprise Systems Engineering,
University of Illinois at Urbana-Champaign,
Urbana, IL 61801
e-mail: hmkim@illinois.edu

The importance–performance analysis (IPA) is a widely used technique to guide strategic planning for the improvement of customer satisfaction. Compared with surveys, numerous online reviews can be easily collected at a lower cost. Online reviews provide a promising source for the IPA. This paper proposes an approach for conducting the IPA from online reviews for product design. Product attributes from online reviews are first identified by latent Dirichlet allocation. The performance of the identified attributes is subsequently estimated by the aspect-based sentiment analysis of IBM Watson. Finally, the importance of the identified attributes is estimated by evaluating the effect of sentiments of each product attribute on the overall rating using an explainable deep neural network. A Shapley additive explanation-based method is proposed to estimate the importance values of product attributes with a low variance by combining the effect of the input features from multiple optimal neural networks with a high performance. A case study of smartphones is presented to demonstrate the proposed approach. The performance and importance estimates of the proposed approach are compared with those of previous sentiment analysis and neural network-based method, and the results exhibit that the former can perform IPA more reliably. The proposed approach uses minimal manual operation and can support companies to take decisions rapidly and effectively, compared with survey-based methods.
[DOI: 10.1115/1.4049865]

Keywords: data-driven design, interpretable machine learning, neural network

1 Introduction

Importance–performance analysis (IPA) was first introduced by Martilla and James [1]. It identifies the product/service attributes that a company must focus on based on the importance and performance. IPA aids in the effective distribution of resources to maximize customer satisfaction. It has been used to guide strategic planning in various fields, such as tourism [2–4], e-governance [5], healthcare [6], and telecommunication [7]. The Kano model has been more widely used to identify customer needs in product design than the IPA, but the Kano model does not consider both the performance and importance of product/service attributes [8]. In the IPA, product/service attributes are categorized into four quadrants based on the levels of importance and performance (Fig. 1). These four quadrants provide the following strategic guidelines. The attributes in quadrant 1 (Q1) with the managerial guideline, "Keep up the good work," have high performance and importance. They indicate competitive advantages or major strengths. The attributes in quadrant 2 (Q2) with the managerial guideline, "Concentrate here," have a low performance but a high importance. These attributes require immediate action for improvement, being the major weaknesses. The attributes in quadrant 3 (Q3) with the managerial guideline, "Low priority," have low performance and importance. These attributes are minor weaknesses. The attributes in quadrant 4 (Q4) with the managerial guideline, "Possible overkill," have a high performance but a low importance. These attributes are minor strengths; therefore, attribute investment can be deployed elsewhere.

Online reviews can be considered as a promising source for IPA because they provide companies opportunities to receive customer feedback and improve the corresponding product attributes [9]. Compared with surveys, numerous online reviews can be easily collected at a lower cost. The textual content of online reviews includes a high level of detail regarding product usage experience by presenting information in verbal format [10]. Conducting IPA based on online reviews allows companies to make strategic decisions rapidly and effectively to improve the performance of next-generation products.

However, numerous studies have used questionnaire surveys to conduct IPA. For example, in one survey, obtaining 540 responses required nine days, excluding the time for the pilot studies and survey analysis [3]. These surveys are time-consuming and expensive. Investigations on conducting the IPA of product attributes based on online reviews are scarce. Instead, some studies have suggested methods to measure the performance and importance of product attributes separately. Machine learning-based sentiment classifiers [11,12], extraction rules with sentiment dictionaries [13–15], and aspect-based sentiment analysis of IBM Watson

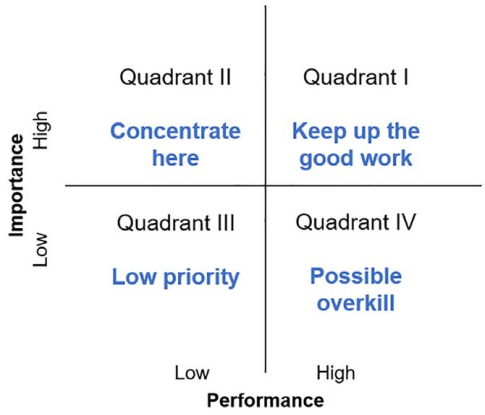

Fig. 1 IPA

[1]Corresponding author.
Contributed by the Design Automation Committee of ASME for publication in the Journal of Mechanical Design. Manuscript received July 21, 2020; final manuscript received December 22, 2020; published online February 11, 2021. Assoc. Editor: Mian Li.

[16] were used to infer the performance of product attributes from online reviews. Term frequency [16–18], regression [19], and neural networks [2] were employed to measure the importance of product attributes from online reviews.

The contributions of the present study are threefold. First, an approach with minimal manual operation is proposed to conduct IPA from the online reviews for product design. The total runtime of the case study, excluding product attribute identification with manual operation, was approximately 2 h on a PC with a 16GB RAM, Intel i7-8550U, and the Ubuntu operating system. This approach can aid companies to make strategic decisions more rapidly and effectively than surveys. Second, aspect-based sentiment analysis of IBM Watson is employed for estimating the performance of product attributes in comparison to existing sentiment analysis. The use of IBM Watson can reduce the time required for developing an aspect-based sentiment classifier and can ensure high accuracy. Third, a Shapley additive explanation (SHAP)-based method is proposed to estimate the importance of product attributes using an explainable deep neural network (DNN). It provides importance values with a low variance by combining multiple optimal neural network models with a high performance.

The remainder of this paper is organized as follows. Section 2 reviews the literature on customer requirements elicitation in product design and the performance and importance estimation of product attributes. Section 3 presents the proposed approach for IPA based on online reviews. Section 4 discusses a case study of smartphones to verify the proposed approach. Section 5 presents the proposed approach application and use of online reviews and an explainable DNN. Section 6 concludes the paper.

2 Literature Review

This section presents previous research on customer requirements elicitation in product design from online reviews and subsequently describes previous studies on the performance and importance estimation of product attributes in detail.

2.1 Customer Requirements Elicitation in Product Design. Identifying customer needs and preferences is highly important for a successful product design [20]. Previous research has determined performance and importance from online product reviews for deciding the design direction of next-generation products based on customer needs. Zimmermann et al. [12] proposed a framework to identify product features and their polarity. Zhang et al. [15] suggested an opinion mining extraction algorithm for identifying product features, their relationships, polarity, and opinion expression. Decker and Trusov [19] proposed a method to estimate the relative importance of product attributes and brand names on the overall rating of a product using negative binomial regression. Suryadi and Kim [14] proposed a method to identify the product features that influence sales ranking using multiple linear regression. A method using the term frequency was proposed to measure the relative importance of product attributes [16,18]. Jiang et al. [17] proposed a method to infer the future importance weights of product features using opinion mining and a fuzzy time series method.

Furthermore, numerous studies have attempted to identify customer needs and preferences from online product reviews. A method for identifying useful customer reviews from the perspective of a designer was presented using text mining [21–23]. A method for reducing the target design-feature range based on customer preferences was also proposed [24,25]. A method for identifying the Kano category of product attributes was presented using sentiment analysis [13,26]. Zhou et al. [11] proposed an approach to identify latent customer needs using case analogical reasoning from the sentiment analysis of online product reviews. Suryadi and Kim [27] proposed a method to automatically extract product usage context using machine learning. Wang et al. [28] presented a method to compare customer needs in two competitive products using latent Dirichlet allocation (LDA). El Dehaibi et al. [29] proposed a method to identify sustainable features using crowdsourced work.

However, to the best of the authors' knowledge, no studies have conducted IPA based on the estimation of both the importance and performance of product attributes from online reviews. The present study contributes to the product design literature by providing a method for conducting the IPA of product attributes from online reviews.

2.2 Previous Research on the Performance Estimation of Product Attributes. Previous studies have estimated the performance of product attributes by measuring their sentiment score based on online reviews. In a review of the pros and cons, the sentiment of product attributes was simply measured by their appearance in the pros and cons categories [19,30]. The sentiment score was considered positive for the pros and negative for the cons. Supervised machine learning was used to measure the sentiment of product attributes by learning sentiment patterns from labeled data. Zimmermann et al. [12] used semi-supervised sentiment learning to measure the polarity of product attributes. Zhou et al. [11] developed a support vector machine learning classifier for evaluating the performance of product attributes. Bi et al. [2] and Zhou et al. [26] used a sentence sentiment classifier based on machine learning. The performance of the product attributes was inferred from the sentiments of sentences containing product-related words. An extraction rule with sentiment dictionaries was employed to measure the performance of product attributes by identifying product-related words and their associated sentiment patterns [13–15,17]. This extraction rule with sentiment dictionaries first identifies product-related words in the forms of nouns and sentiment words in the forms of adjectives that modify a noun word. Subsequently, it evaluates the sentiment words based on a well-constructed sentiment word bank. Jeong et al. [16] used an aspect-based sentiment analysis of IBM Watson, which is a text mining technique that breaks down a text into aspects (i.e., attributes or components of a product or service) and estimates the sentiment score of the aspects.

However, previous studies to measure the performance of product attributes have the following limitations. The estimation of the sentiment of the product attributes as pros and cons is not applicable to natural language forms. Developing machine learning-based sentiment classifiers and extraction rules with sentiment dictionaries is expensive. Machine learning-based sentiment classifiers require labeled data to develop a model, and manual labeling is time-consuming. Extraction rules with sentiment dictionaries require manual rules to define various syntactic patterns that express emotions, which is time-consuming. Sentence sentiment classifiers cannot accurately measure the sentiment of a product attribute when a sentence has more than two product attributes. In contrast, the aspect-based sentiment analysis of IBM Watson is publicly available and applicable to numerous cases using predefined rules, although its specific rules are unknown [31].

Therefore, this study utilizes the aspect-based sentiment classifier of IBM Watson to estimate the performance of product attributes. Using IBM Watson to develop aspect-based sentiment classifiers saves time and ensures high accuracy because it trains classifiers using large-scale data [16].

2.3 Previous Research on the Importance Estimation of Product Attributes. Previous studies estimated the importance of product attributes using term frequency and regression from online reviews. The importance of product attributes was measured using the frequencies of product-related words. High-frequency product attributes and high-frequency, low-sentiment-score product attributes were considered to have high importance in Refs. [17,18], respectively. Suryadi and Kim [14] used multiple regression to determine the relationship between the sentiment of

a product attribute and its sales ranking. The coefficient of a product attribute in a regression model can be considered as an importance value. Previous studies newly defined the importance, but they did not provide theoretical evidence to support it.

Importance is classified into self-stated importance (i.e., relevance) and implicit importance (i.e., determinance) [32]. The direct measure of self-stated importance is achieved by asking customers about the satisfaction of each attribute. This presents general attribute-importance and aids in distinguishing between core and non-core attributes [33]. Implicit importance is a relatively flexible concept compared with self-stated importance and is measured by evaluating how each attribute affects overall satisfaction. This indicates case-based or situational attribute importance and describes the behavioral outcomes of customers for overall satisfaction [34]. Some studies have attempted to estimate the implicit importance of product attributes from online reviews. Decker and Trusov [19] used negative binomial regression based on sentiment scores and user ratings to estimate the implicit importance values in the review of pros and cons. Bi et al. [2] used a neural network for evaluating the implicit importance in natural language text. A shallow neural network (SNN) with a hidden layer was considered as the neural network architecture. The implicit importance was estimated by calculating the weights of the input and hidden layers and the hidden and output layers from the SNN.

Regression analysis assumes that the overall rating follows a Gaussian distribution and is a linear combination of the sentiment scores of the product attributes mentioned in the reviews. However, in online reviews, the overall ratings generally present a positively skewed distribution and can be a non-linear combination of the sentiment scores of the product attributes mentioned in the reviews. Therefore, regression analysis does not perform well compared with a neural network [2,32,35]. Furthermore, the SNN-based method cannot identify how input features affect each prediction since the neural network is described as a black-box model. The SNN-based method inferred the importance of input features from weights between layers in the training set, but the variability of the importance values is high because of the randomness of the training set and model. The SNN structure with a hidden layer is also unclear whether it is optimal between neural networks composed of various neurons and hidden layers. To overcome these limitations, the present study uses an explainable DNN technique to identify the effects of sentiments of product attributes on the overall rating. Implicit importance values of product attributes with a low variance are derived by combining the effect of product attributes from multiple optimal neural networks with a high accuracy.

3 Method

The overall process of performing the IPA of product attributes from online customer reviews is described here (Fig. 2). The online product reviews are the inputs, and the output is the IPA of the product attributes. The proposed approach comprises three important stages. It uses minimal manual operation, although a few stages require human involvement.

(1) Product attribute identification: Product attributes are identified using the LDA from online reviews. An automated method for keyword preprocessing is used prior to the LDA; human judgment is required to identify the product attributes from the LDA results.
(2) Performance estimation of each product attribute: After identifying the product attributes, the performance of each product attribute is estimated using the aspect-based sentiment analysis by IBM Watson. IBM Watson provides the sentiment intensities of product-related words; therefore, the performance of each product attribute is automatically measured.
(3) Importance estimation of each product attribute: The importance of each product attribute is estimated using an explainable DNN based on the performance estimation of each product attribute. Optimal neural networks are first designed to predict overall ratings based on the estimated sentiment scores of product attributes. The importance values are subsequently derived by combining deep SHAP values that present the influence of input variables on output variables as the explainable DNN technique from these neural networks. The importance estimation is automated, and initial parameter settings are necessary to determine the optimal neural networks.

3.1 Data Collection and Preprocessing. To perform the proposed approach, the collection of numerous reviews is required. A few reviews over a long period are likely to be biased in terms of representativeness. The customer reviews of a target product are collected from product review websites, including Amazon, eBay, and BestBuy. Web scraping can be used to automatically collect information such as title, review, date, and user rating from web pages. Duplicated reviews that appear more than once are removed, and non-English reviews are eliminated to refine the collected review. The emojis, emoticons, and newline characters, such as "U+1F600," "U+1F603," and "U+1F604," in each review are removed. The keywords and their sentiment intensities are extracted for the IPA of the collected reviews using IBM Watson natural language understanding (NLU). IBM Watson NLU automatically extracts keywords and their sentiment intensities by removing stop words, such as "and," "but," "how," and "what." Subsequently, the text preprocessing proceeds as follows: uppercase is converted to lowercase (e.g., "Screen" is transformed to "screen"), punctuation is eliminated (e.g., "high-end" is transformed to "high end"), and words are lemmatized (e.g., "batteries" is transformed to its root form "battery"). Consequently, each review is structured into keywords and their sentiment intensities.

3.2 Product Attribute Identification. LDA-based methods can be used for identifying the product attributes from online reviews [16,26,28,36]. The LDA is a powerful statistical topic model that summarizes numerous textual data by extracting the hidden topics [37]. It assumes that each review document is regarded as a mixture over a set of topic probabilities, and each topic is regarded as a mixture over a subsequent set of words.

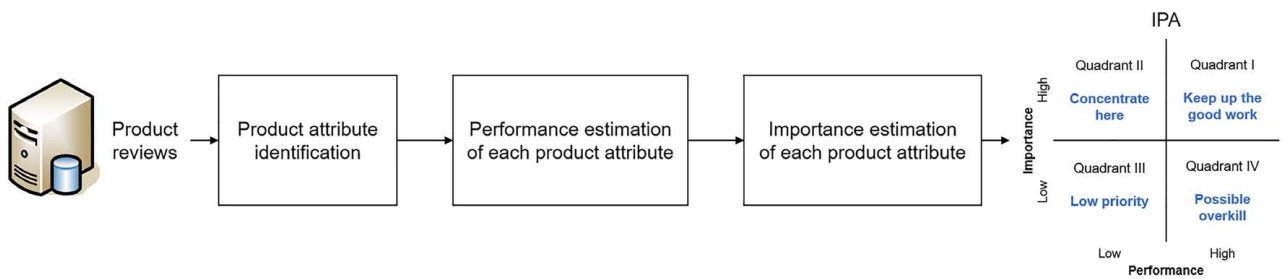

Fig. 2 Overall process of the proposed approach

The LDA model takes a review-keyword matrix as the input, and it yields a topic-keyword matrix as the output. Topic coherence [38] can be used to determine the number of topics in the LDA results. The LDA model with a higher topic coherence indicates a model best. Each topic is identified by interpreting the keywords in the topic. The label of each topic can be considered as an attribute of the product [16,26,28,36].

The LDA-based method by Joung and Kim [36] is used herein to identify product attributes from online reviews. This method is applied because it considers the forms of noun phrases as product attributes and reduces the manual effort by providing an automated method for keyword preprocessing in the LDA. The LDA-based method includes two steps after the extraction of keywords from the customer reviews.

Step 1: Noise keywords are automatically filtered out using product manuals. Subsequently, the product-related keywords are identified. For example, product-related keywords such as "camera," "screen," and "battery life" are identified as words, including nouns and noun phrases.

Step 2: Product-related keywords that are frequently mentioned together by customers are grouped into topics by the LDA, and the product attributes are identified by interpreting the top n keywords and typical reviews related to each topic. For example, a topic clustered with keywords such as "screen," "size," "display," and "glass" is named as a "screen" attribute of a product.

Additionally, the product-related keywords of each product attribute can be expanded by identifying synonyms through WordNet [39] or word embedding [40]. A more detailed process of the LDA-based method can be found in the study by Joung and Kim [36]. Some open-source libraries or software, such as the Gensim library of PYTHON [41] and the Stanford Topic Modeling Toolbox [42], can be used if the application of the LDA-based method by Joung and Kim [36] to product attribute identification is difficult. The keyword preprocessing in the LDA can be manually conducted here.

3.3 Performance Estimation of Each Product Attribute. After extracting the customer reviews that include the keywords of each product attribute, the aspect-based sentiment analysis of IBM Watson is used to estimate the performance of the identified product attributes. The aspect-based sentiment analysis provides the sentiment intensity of the keywords, which ranges from −1 to 1; −1 indicates a higher negative sentiment, and 1 a higher positive sentiment. M customer reviews are structured into the keyword sentiment intensity of the product attributes, A_i. If a review contains more than two keywords for the product attribute, the sentiment intensity of that product attribute is calculated by taking the average of the sentiment intensities of the keywords. For example, in the following review, "Screen is great. Size is too big for me," the sentiment intensities of "screen" and "size" are 0.97 and −0.68, respectively. Consequently, the sentiment intensity of the product attribute is their average, i.e., 0.145. However, quantifying emotional sentiment is challenging. For example, in IBM Watson, "convenient screen," "good screen," "fine screen," and "nice screen" are generally considered as very positive sentiments. However, they are strictly assigned to different sentiment intensity scores (i.e., 0.9 for "convenient," 0.95 for "good," 0.84 for "fine," and 0.96 for "nice"). To consider similar emotional expressions equally and perform the subsequent analysis, the keyword sentiment intensity is encoded into six labels using the following equation (Table 1):

Table 1 Transformed sentiment score for each product attribute in the customer reviews

Review	A_1	A_2	⋯	A_i	Overall rating
1		5	⋯		5
2	4	4	⋯		5
3		5	⋯	5	4
⋮	⋮	⋮	⋮	⋮	⋮
M			⋯		3

$$S_{im} = \begin{cases} 5 & \text{if } 0.6 < \text{Sentiment intensity} \le 1 \\ 4 & \text{if } 0.2 < \text{Sentiment intensity} \le 0.6 \\ 3 & \text{if } -0.2 \le \text{Sentiment intensity} \le 0.2 \\ 2 & \text{if } -0.6 \le \text{Sentiment intensity} < -0.2 \\ 1 & \text{if } -1 \le \text{Sentiment intensity} < -0.6 \\ 0 & \text{if Sentiment intensity is "missing value"} \end{cases} \quad (1)$$

The transformed labels indicate 5 (very positive), 4 (positive), 3 (normal), 2 (negative), and 1 (very negative). For example, "convenient screen," "good screen," "fine screen," and "nice screen" are generally assigned a score of 5. The above sentiment intensity ranges assigned to the six labels can be used for all types of products. However, to clearly identify very positive and negative sentiments, the ranges can be narrowed further. The performance of a product attribute, A_i, is calculated as follows:

$$\text{Perf}_i = \sum_{m=1}^{M} \frac{S_{im}}{R_i} \quad (2)$$

3.4 Importance Estimation of Each Product Attribute. An explainable DNN is used for estimating the importance of each product attribute. To exploit the explainable DNN, the transformed score values of the online reviews are used as the input features, whereas the overall rating corresponding to the reviews is the output variable (Table 1). Even though the overall ratings ranging from 1 to 5 can be used directly, they can be further categorized into two labels for higher performance when training the neural network. One- or two-star ratings are regarded as negative, and four- or five-star ratings are regarded as positive. Three-star ratings are classified into positive or negative depending on the pos/neg ratio and strength of the sentiment scores of the product attribute in the reviews.

Herein, in the explainable DNN, the SHAP-based method is proposed to derive the importance values of each product attribute with a low variance. The strategy for estimating the importance values of the product attributes involves initially constructing K optimal models from K training sets to solve the variance problem of the model constructed from a single training set with randomness. Subsequently, K importance values are calculated by measuring the influence of the input features based on SHAP method from the K optimal models and combined into one value. The importance value by the SHAP-based method is calculated as follows (Fig. 3):

(1) K training sets are prepared.
(2) By employing a genetic algorithm [43] with each training set, K optimal neural networks are designed.
(3) From each K optimal neural network, K importance values [44] of each product attribute are calculated based on the SHAP method.
(4) Based on the information fusion algorithm, the K importance values are combined [45].

3.4.1 Preparing K Training Sets. K-fold cross-validation is used to randomly partition the original sample into K equal-sized sub-samples [46]. A single sub-sample is used as a test set for

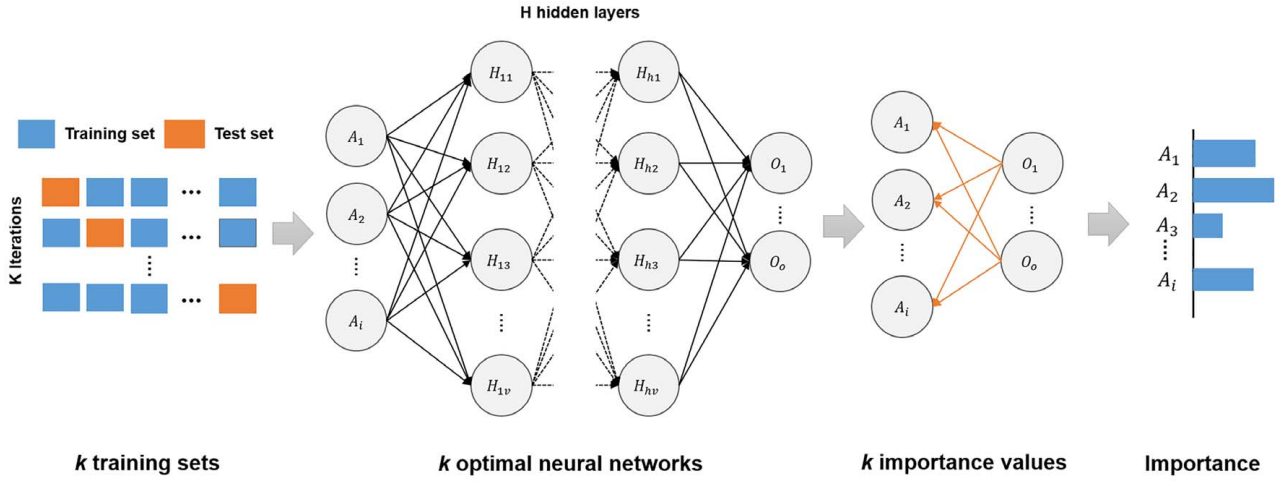

Fig. 3 Overall process of the SHAP-based method

measuring the performance of the model, and the remaining $K-1$ sub-samples are used as the training sets. A K-fold cross-validation can reduce the bias of the results from a single training set because all the observations are used for both training and testing. There is no strict rule to determine K; however, a tenfold cross-validation can generally be applied [47]. A small K can be considered, if the volume of the data is insufficient. The K can reduce the variance of the performance of the trained model when the test set is large.

3.4.2 Designing K Optimal Neural Networks. From each K training set, K optimal neural networks are designed to predict the overall rating based on the sentiment score of each product attribute. The feedforward neural network obtained by extending the previous SNN can be considered as the neural network architecture [2,32,35]. The feedforward neural network comprises an input layer, hidden layers, an output layer, and neuron units per layers. The number of neurons of the input and output layers depends on the number of variables in the data; however, the number of hidden layers and neurons per hidden layer need to be determined. In addition, an activation function that decides the activation between the neurons and an optimizer that decides the optimal weights between the neurons should be selected to train the neural network. "Rectified linear unit" (ReLU), "exponential linear unit" (ELU), and "Tanh" can be considered as activation functions for the DNN [48]. "Stochastic gradient descent" (SGD), "Adagrad," "Adadelta," "RMSProp," "Nadam," "Adam," and "Adamax" can be considered as the optimizers [49].

A genetic algorithm [43,50] is used to design K optimal feedforward neural networks based on the initial parameters, such as the number of hidden layers, neurons per hidden layer, activation function type, and optimizer type. The genetic algorithm is used to solve optimization or search problems depending on biological processes, such as life, reproduction, and death [51]. Biological processes include population, generation, selection based on the fitness score, crossover to generate new offspring, and random mutation of new offspring. After setting the range of initial parameters, the genetic algorithm to determine the optimal neural network proceeds as follows:

Step 1: The neural networks corresponding to the population at training set are constructed by randomly selecting initial parameters in a given range.
Step 2: The performance of each neural network is evaluated by the fitness function. For the fitness function, performance measures such as accuracy, precision, and recall at a test set can be considered.
Step 3: The neural networks corresponding to retention rate remain to breed children. The top neural networks are selected based on the fitness score, and a few non-top networks are randomly chosen. The selected neural networks become a part of the next generation.
Step 4: The initial parameters of the selected neural networks, such as the number of hidden layers, neurons per hidden layer, activation function type, and optimizer type, are considered as individual genes and bred through their combination.
Step 5: Some initial parameters for offspring are randomly mutated based on mutation chance.
Step 6: Step 2 is conducted until the termination condition is reached. The termination condition is determined by the number of generations.

After applying the genetic algorithm, the optimal model is the neural network with the highest fitness score, which assigns weights between the output and input layers.

The constructed optimal neural networks should not have an overfitting problem, that is, a high performance on the training set and not on the test set. Overfitting models reduce their generalizability. The importance values of the product attributes derived from these models may be the noise rather than the genuine value in the population.

3.4.3 Calculating K Importance Values Based on the SHAP Method From K Optimal Neural Networks. The SHAP method is used for calculating the K importance values of each product attribute from each K optimal neural network [44]. An explainable DNN technique, SHAP method, is a unified approach to interpret the DNN model predictions based on Shapley values. Given a specific prediction v, the Shapley values are calculated using a weighted sum of the influences of each feature over all possible orders of the features (Eq. (3)). The influence of each feature in each prediction is calculated by estimating the change in the prediction of the model when that specific feature is missing.

$$\phi_i(v) = \sum_{S \subseteq N: i \notin S} \frac{|S|!(|N| - |S| - 1)!}{|N|!}(v(S \cup i) - v(S)) = \sum_{S \subseteq N: i \notin S} \frac{1}{(|N| \text{choose} |S|)(|N| - |S|)}(v(S \cup i) - v(S)) \quad (3)$$

To estimate the Shapley values in the neural network model, the SHAP method infers the weights between the input and output layers by combining the influence of the features calculated for the small components of the neural network into that of the features for the whole neural network. The SHAP method has a solid theoretical foundation based on game theory and provides contrastive

Table 2 Deep SHAP values of each product attribute in the model

Review	A_1	A_2	⋯	A_i	Overall rating
1	SHAP_{11}	SHAP_{21}	⋯	SHAP_{i1}	5
2	SHAP_{12}	SHAP_{22}	⋯	SHAP_{i2}	5
3	SHAP_{13}	SHAP_{23}	⋯	SHAP_{i3}	4
⋮	⋮	⋮	⋮	⋮	⋮
M	SHAP_{1M}	SHAP_{2M}	⋯	SHAP_{iM}	3

explanations because the prediction is reasonably distributed between the feature values.

From each K optimal neural network, deep SHAP values that represent the influence of product attributes on the overall rating are calculated (Table 2). The larger the absolute value of deep SHAP in each review, the greater the effect on the overall rating. The importance of product attributes from kth optimal neural network is calculated as follows:

$$\text{Imp}_{ik} = \sum_{m=1}^{TR_k} \frac{|\text{SHAP}_{\text{imk}}|}{TR_k} \tag{4}$$

In K optimal neural networks, K importance values of each product attribute are estimated.

3.4.4 Combining K Importance Values. The information fusion algorithm is used to combine the K importance values of each product attribute [45]. Numerous studies have employed this algorithm for combining the effect of input variables in machine learning models. The fusion of multiple models in knowledge discovery provides reliable results [52]. The information fusion algorithm is formulated as

$$\hat{y}_{\text{fused}} = \sum_{k=1}^{K} w_k f_k(x) = w_1 f_1(x) + w_2 f_2(x) + \cdots + w_k f_k(x) \sum_{k=1}^{K} w_k \tag{5}$$

Based on Eq. (5), the importance of each product attribute is estimated by combining the K importance values in the optimal neural networks by the following equation:

$$\hat{\text{Imp}}_i = \sum_{k=1}^{K} w_k \text{imp}_{ik} \tag{6}$$

Finally, the importance value of each product attribute is normalized using the following equation:

$$\overline{\text{Imp}_i} = \frac{\hat{\text{Imp}}_i}{\sum_{i=1}^{I} \hat{\text{Imp}}_i}, \quad i = 1, 2, \ldots, I \tag{7}$$

The SHAP-based method provides the importance values of the product attributes with a low variance by combining multiple explainable neural network models derived from multiple training sets.

3.5 Importance–Performance Analysis. The IPA plot is drawn based on the estimated importance–performance of each product attribute (Fig. 1). The x- and y-axes represent the performance and importance, respectively. The IPA plots of the products of a target company or a target product model (i.e., high-end and mid-range models) can be drawn to provide insights into the product design.

The cross-hair of the IPA is determined using two types of methods: scale- [1] and data-centered methods [4,53]. The scale-centered method determines the cross-hair using the mid-point of the scale of the importance–performance. In contrast, the data-centered method determines the cross-hair using the means or medians of the scale of importance–performance. The data-centered method is frequently used because it provides a stronger distinguishing power between the attributes compared with to the scale-centered method [54]. The data-centered method is used herein for the cross-hair placement of the IPA. If the product attribute is located adjacent to the cross-hair, it can be considered to comprise two close quadrants.

4 Case Study

An IPA case study of smartphones was used to validate the proposed approach. Three IPA plots for all, high-end and mid-range smartphones were drawn from online customer reviews. The "all smartphones" category contains similarly sized high-end and mid-range products. The high-end products present a slightly better performance, but they share numerous common features with a phone released by a specific manufacturer.

4.1 Collecting Data and Preprocessing. Web scraper chrome extension (e.g., WebScraper.io) was used for collecting the customer reviews of verified purchases in the cell phone category of Amazon.com. After eliminating the overlapping and non-English reviews, 33,779 reviews of smartphones were obtained from April 2014 to September 2019. The reviews in the five years after the product was released in 2014 were selected because of their lack until 2013. The time trend of the collected reviews is shown in Fig. 4. The number of reviews of high-end and mid-range smartphones were 23,591 and 10,188, respectively. Each review was refined by stripping the emojis, emoticons, and newline characters.

From the collected reviews, 51,011 keywords and the corresponding sentiment intensities were extracted using the IBM Watson NLU. After text preprocessing, each review was structured into keywords and their sentiment intensities.

4.2 Identifying Product Attributes. After extracting 51,011 keywords, the LDA-based method of Joung and Kim [36] was used for identifying the product attributes from the online reviews. After filtering out noise keywords, 1226 product-related keywords were clustered in each topic by the LDA. The number of topics was selected as eight based on the maximum topic coherence value (0.711). These eight attributes were identified by interpreting the logical connections between the top-30 words and typical reviews (Table 3). Most topics were easy to identify without investigating the review comments. For example, the second topic was named "Screen" considering its top related keywords (e.g., "screen," "case," "size," "display," "protector," and "glass")." The third, fourth, fifth, sixth, seventh, and eighth topics were also named. The first topic, "Product check," was named by examining the typical reviews, which included the most probable topic keywords. The typical reviews contained "This product is worthless to me. The product arrived in good condition; however, the model number on the box and the sticker attached to the phone were different than the model numbers in the product specifications on the Amazon web site." This attribute was not directly related to the hardware of the smartphone but was considered because it represented initial condition or quality of the ordered product. Synonyms of the keywords were identified using word2vec. For example, synonyms of "camera" and "app," such as "cam" and "application," were considered in the "camera" and "app" attributes. The Gensim library of PYTHON [41] was used to conduct the LDA and to build word2vec. "Frequent keywords" were sorted in the descending order based on the probability of each attribute in the LDA. "Number of keywords" indicates the number of keywords, considering the synonyms of each product attribute. "Number of reviews" indicates the number of reviews that include the keywords of each product attribute. "Product check," which determines the first impressions of an ordered product, was the attribute most frequently mentioned by customers. Among the remaining seven attributes, the customers were

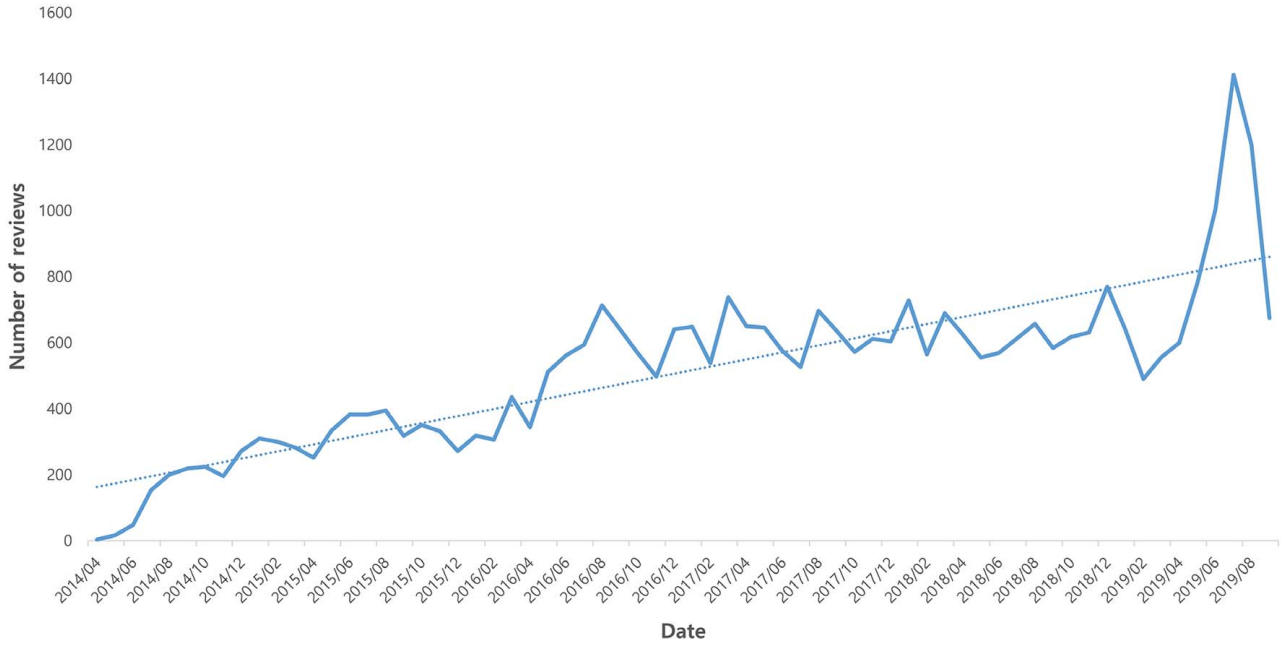

Fig. 4 The time trend of the data

Table 3 Eight product attributes from smartphones

Product attribute	Frequent keywords	Number of keywords	Number of reviews
Product check (A_1)	Product, problem, seller, box, device, condition, version, model, warranty, item, description, replacement, support, and return	26	6749
Screen (A_2)	Screen, case, size, display, protector, glass, cover, screen protector, pocket, and touch	21	5304
Camera (A_3)	Camera, quality, picture, video, photo, light, front, pic, resolution, and image	23	3612
App (A_4)	Apps, android, update, app, notification, email, application, mail, and file	26	2967
Communication (A_5)	Call, network, data, text, message, lte, internet, signal, voice, connection, contact, fi, and gps	25	3351
Battery (A_6)	Battery, life, battery life, charge, use, power, drain, battery drain, fast charging, battery charge, and battery power	17	3948
Card slot (A_7)	Card, sim, sim card, sd, slot, sd card, dual sim, memory card, pin, and microsd	16	2666
Accessory (A_8)	Charger, port, cable, accessory, plug, usb, earphone, wall, jack, microphone, assistant, and wireless charging	24	2261

Table 4 Sentiment score values of the online reviews for all, high-end and mid-range smartphones

Product		Review	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8	Overall rating
All	High-end	1	0	5	0	0	0	0	0	0	5
		2	0	0	5	0	0	0	0	0	5
		⋮	⋮	⋮	⋮	⋮	⋮				
		10,902	5	0	0	0	0	0	0	5	
	Mid-range	1	0	0	0	5	0	0	0	0	5
		2	4	4	4	3	4	3	3	0	5
		⋮	⋮	⋮	⋮	⋮	⋮				
		4191	0	0	0	0	0	5	0	0	5

more concerned about the "Screen," "Camera," "Communication," and "Battery." These attributes were mentioned more than 3000 times in all reviews.

4.3 Estimating Performance of Each Product Attribute. The sentiment score of each product attribute was measured by the sentiment intensity of the keywords identified in each product attribute. Among the original data of 33,779 reviews, reviews that did not include the product-related keywords and their sentiment intensities were excluded from the analysis:

R1: "Five Stars. Better than I expected flawless."
R2: "Five Stars. Perfect phone for me!"
R3: "She loves it! So far so good."
R4: "Works perfect. Nice for the price."
R5: "Very nice phone for the price, and can be used overseas… yesss…thank you."

Table 5 Performance of the eight product attributes for all, high-end and mid-range, smartphones

Product	$Perf_1$	$Perf_2$	$Perf_3$	$Perf_4$	$Perf_5$	$Perf_6$	$Perf_7$	$Perf_8$
All	2.994	2.788	3.753	2.687	2.177	2.687	2.786	2.293
High-end	3.016	2.679	3.716	2.660	2.088	2.522	2.691	2.276
Mid-range	2.914	3.051	3.821	2.742	2.347	3.166	2.975	2.375

Most of these reviews are short reviews and unsuitable for performing the three stages of the IPA. The IPA was conducted with 15,093 reviews, and the sentiment intensities of the keywords in the reviews were transformed into the score values of the six labels using Eq. (1) (Table 4). Based on the score values of the online reviews, the performance of all, high-end and mid-range, smartphones was estimated using Eq. (2) (Table 5). In Table 5, the performance of the product attributes of the high-end smartphones is generally lower than that of the mid-range smartphones. Customers were dissatisfied with the high-end products of the company compared with those of their competitors, whereas they were relatively satisfied with their mid-range products.

4.4 Estimating Importance of Each Product Attribute. To build the explainable DNN model, the input features were the transformed sentiment scores of the product attributes in the 15,093 reviews, whereas the output variable was the overall rating corresponding to the reviews. The accuracy of the neural network was approximately 60% when constructing the neural network using the five labeled ratings as the output variable. For the two-labeled ratings, the accuracy was approximately 80%. In this study, the output variable was transformed from the five labeled ratings into two-labeled ratings for estimating the importance values in the neural network with a high performance. The ratio of the positive and negative ratings was 6:4, which indicates a balanced class. If the ratio corresponds to an unbalanced class, such as 7:3 or 8:2, a high weight can be assigned to that class with a small number. According to Fig. 3, the importance of each product attribute on all, high-end and mid-range, smartphones were calculated.

A fivefold cross-validation (80% training set and 20% test set) was used based on the Pareto principle [55]. From the five training sets, five explainable DNNs were built, and the accuracy was used as the performance measure because of a balanced class [56]. For an imbalanced class with an extremely high class ratio on one side, the F-1 score can be used. The model with a high performance has a high weight for the importance estimation. The genetic algorithm was used to design five optimal neural networks with a high accuracy from the five training sets. The initial parameters of the genetic

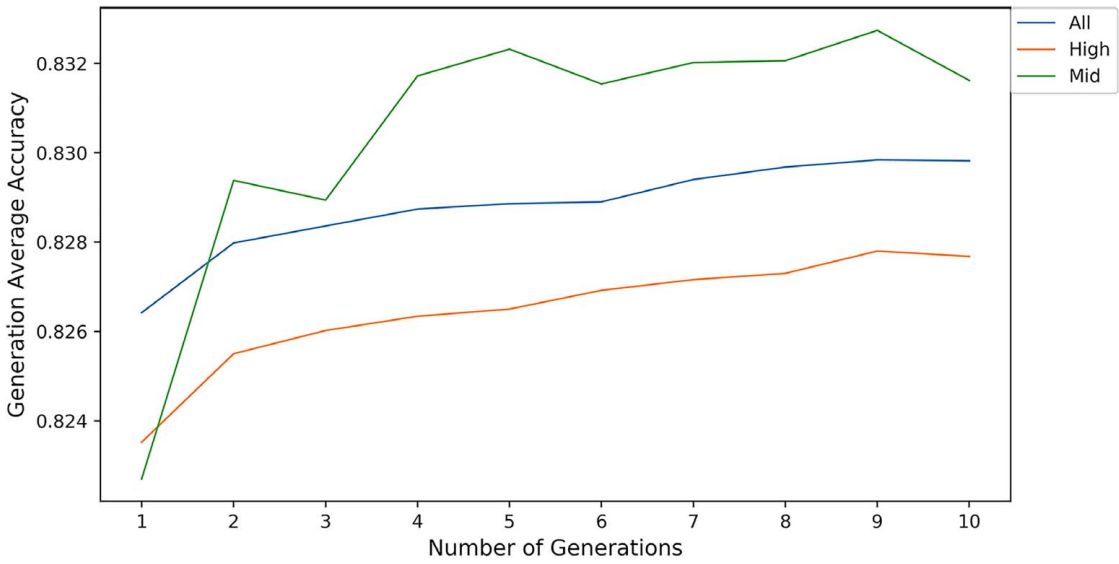

Fig. 5 Generation average accuracies by genetic algorithm

Table 6 Optimal neural networks for all, high-end and mid-range, smartphones

Product	Number of training sets	Number of layers	Number of neurons	Activation function	Optimizer	Test accuracy (training)
All	1	5	64	tanh	adam	0.844 (0.833)
	2	1	128	tanh	adam	0.827 (0.832)
	3	5	128	relu	nadam	0.828 (0.842)
	4	4	21	relu	nadam	0.845 (0.831)
	5	5	128	relu	rmsprop	0.832 (0.832)
High-end	1	2	128	tanh	adamax	0.830 (0.828)
	2	5	128	relu	adagrad	0.824 (0.838)
	3	5	31	elu	adadelta	0.839 (0.826)
	4	1	10	tanh	rmsprop	0.843 (0.824)
	5	4	128	relu	nadam	0.832 (0.828)
Mid-range	1	5	21	elu	nadam	0.844 (0.835)
	2	4	128	relu	adamax	0.868 (0.831)
	3	1	31	relu	adagrad	0.826 (0.835)
	4	5	64	elu	adam	0.833 (0.839)
	5	4	128	relu	adamax	0.839 (0.839)

Table 7 Importance values for all, high-end and mid-range, smartphones from the five optimal neural networks

Product	Number of training sets	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8
All	1	0.108	0.067	0.051	0.026	0.035	0.047	0.032	0.032
	2	0.106	0.069	0.057	0.029	0.032	0.050	0.039	0.039
	3	0.094	0.062	0.056	0.028	0.035	0.044	0.029	0.029
	4	0.085	0.066	0.056	0.032	0.036	0.051	0.035	0.035
	5	0.101	0.058	0.050	0.028	0.032	0.046	0.026	0.026
High-end	1	0.120	0.065	0.053	0.025	0.029	0.054	0.043	0.043
	2	0.106	0.063	0.042	0.023	0.027	0.049	0.036	0.036
	3	0.113	0.056	0.052	0.028	0.033	0.053	0.037	0.037
	4	0.116	0.057	0.067	0.023	0.034	0.060	0.038	0.038
	5	0.118	0.060	0.040	0.023	0.024	0.049	0.025	0.025
Mid-range	1	0.079	0.057	0.096	0.045	0.051	0.047	0.034	0.034
	2	0.063	0.066	0.099	0.035	0.030	0.050	0.032	0.032
	3	0.062	0.063	0.097	0.030	0.018	0.047	0.041	0.041
	4	0.067	0.068	0.084	0.040	0.042	0.043	0.033	0.033
	5	0.069	0.070	0.081	0.038	0.032	0.038	0.038	0.038

Table 8 Importance of the eight product attributes for all, high-end and mid-range, smartphones

Product	Imp_1	Imp_2	Imp_3	Imp_4	Imp_5	Imp_6	Imp_7	Imp_8
All	0.252	0.164	0.138	0.073	0.087	0.122	0.082	0.082
High-end	0.284	0.149	0.126	0.060	0.073	0.131	0.088	0.088
Mid-range	0.164	0.157	0.222	0.091	0.084	0.109	0.087	0.087

algorithm, such as population, generation, retention rate, mutation chance, number of hidden layers, neurons per hidden layer, an activation function type, and optimizer type, were determined by referring to previous studies [2,32,35,43]. The population and generation were determined as 50 and 10, respectively, to reduce the search time, because increasing the population and generation did not produce better results [43]. The retention rate and mutation chance were given as 0.4 and 0.25, respectively, because these settings help find mutant neural networks with a high accuracy while maintaining elitism. The number of hidden layers was considered as 1 [2,32,35], and 2, 3, 4, and 5 were additionally considered. The neurons per hidden layer were considered as 10 [32], 21 [2], 31 [35], and 64, and 128 was additionally considered. For the activation function type, "Tanh" was considered in previous studies, and "ReLU" and "ELU" were additionally considered. Various types of optimizers, "SGD," "Adagrad," "Adadelta," "RMSProp," "Nadam," "Adam," and "Adamax," were considered because the network optimizers were unclear in previous studies. The Keras library of PYTHON was used to build the neural network. The genetic algorithm results exhibited that the average accuracy with five test sets of each smartphone category continued to improve with the generations (Fig. 5). The optimal neural networks with the best accuracy were designed based on five training sets of each smartphone category without the overfitting problem (Table 6). The SHAP method[2] was used for calculating the effect of the eight product attributes on the overall rating obtained from five optimal neural networks. Based on the deep SHAP values from the five optimal neural networks, the importance of each product attribute on all, high-end and mid-range, smartphones was measured by Eqs. (4), (6), and (7) (Tables 7 and 8). The importance value of a product attribute estimated using the SHAP-based method correlated with the frequency of that attribute mentioned in the reviews because a frequently mentioned attribute is expected to affect the overall rating; however, it did not fit exactly. Based on the frequency, the top four product attributes were "product check," "screen," "battery," and "camera" (Table 3). In contrast, the top four product attributes affecting the overall ratings were "product check," "screen," "camera," and "battery" in the all smartphones category. The importance values derived by the SHAP-based method indicated the effect on the overall rating. The importance of "screen," "camera," and "battery" was also high in a previous study in which Chinese reviews were analyzed [57]. The importance of "app" was low in both the studies. In contrast, the importance of "communication" was low in this study, whereas it was higher than that of "screen" and "battery" in the previous study. This study and the previous study targeted the US and Chinese markets, respectively. These regional variations may have caused differences in the importance values of the product attributes.

4.5 Importance–Performance Analysis. The IPA plots of all, high-end and mid-range, smartphones were built based on the

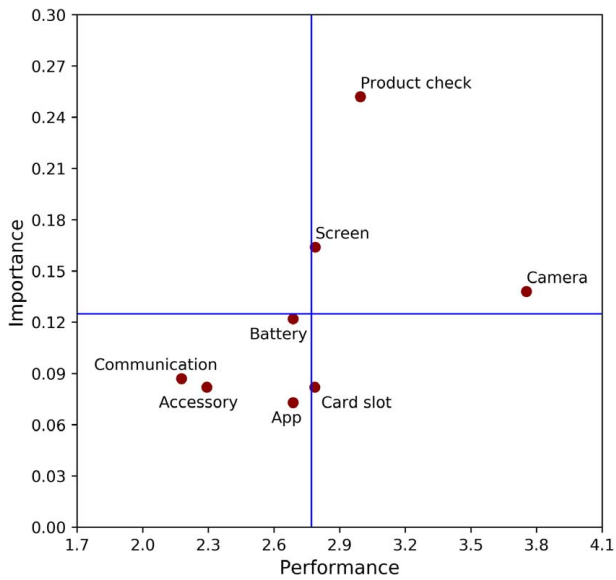

Fig. 6 IPA of the all smartphones category

[2]https://github.com/slundberg/shap

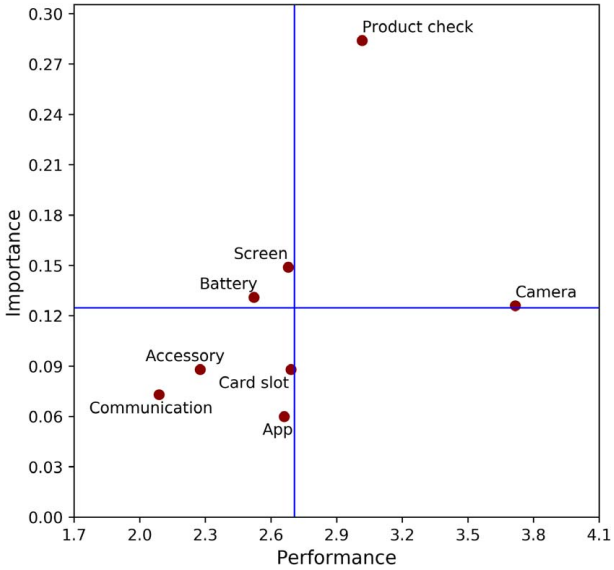

Fig. 7 IPA of high-end smartphones

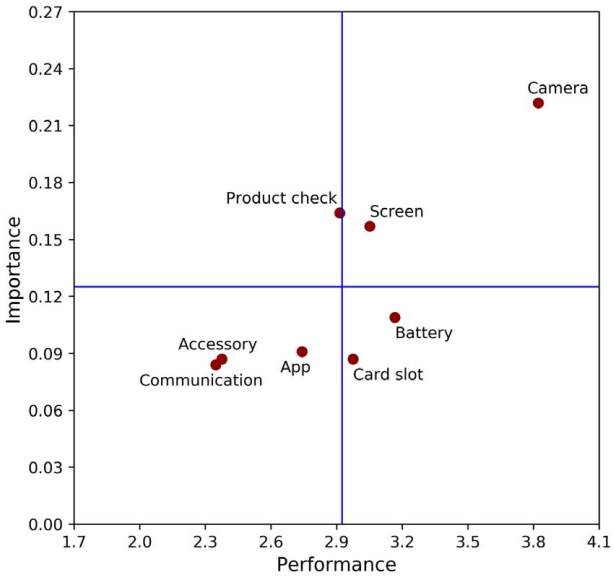

Fig. 8 IPA of mid-range smartphones

estimated performance and importance (Figs. 6–8). In the IPA plots of all smartphones, "product check" (A_1), "screen" (A_2), and "camera" (A_3) attributes were positioned in Q1. These attributes have both high performance and importance, which indicates the main merits and competitive advantages of the target smartphones. The advantages of these attributes must be maintained. There were no product attributes in Q2, and the all smartphones category did not have a major weakness. However, if the performance of "screen" (A_2) is lowered in the future, it can be considered in Q2. This attribute has a low performance and high importance; therefore, immediate investment and attention are expected. "App" (A_4), "communication" (A_5), "battery" (A_6), and "accessory" (A_8) attributes were positioned in Q3. These attributes have low performance and importance, which indicates low priority. "Card slot" (A_7) was located in Q4. This attribute has a high performance and low importance, which indicates that it is an overemphasized attribute. Therefore, companies may not consider investment and attention.

For high-end and mid-range smartphones, some product attributes were placed slightly different. "Screen" (A_2) was located in Q1 for the mid-range smartphones but in Q2 for high-end smartphones. "Screen" (A_2) is a main merit and competitive advantage for mid-range smartphones; however, immediate investment and attention are required to improve the performance of high-end smartphones. Customers may want to improve their screens because they frequently use high-end smartphones to play high-performance games and high-quality streaming services compared with mid-range smartphones. "Battery" (A_6) was also located in Q4 for mid-range smartphones but in Q2 for high-end smartphones. "Battery" (A_6) is an overemphasized attribute for mid-range smartphones; however, its performance for high-end smartphones must be improved. These results reflect that customers do not expect high performance from mid-range products, unlike high-end products. Attention and investment for the "battery" attribute placed in Q4 for mid-range smartphones need not be considered.

4.6 Validation of Performance and Importance Estimation. The performance and importance estimation of the proposed approach was validated for other cases. The approach was verified by comparing its results with those of existing methods. In the performance estimation, the aspect-based sentiment of IBM Watson was used to calculate the sentiment intensity of each product attribute in a case study. The aspect-based sentiment analysis of IBM Watson was compared with that of previous studies, such as sentence sentiment analysis [2,26] and dependency parser with SenticNet4 dictionary [14]. In the following example, the sentiment intensity of a "condition" word corresponding to the "product check" attribute was −0.574 (i.e., negative) according to IBM Watson and 0 (i.e., neutral) according to Vader, sentence sentiment analysis. The dependency parser can be measured as a positive value on identifying patterns such as "good condition."

Example: "*Be careful about the product you receive—check it thoroughly.* ***Product was received in good condition but immediately it became apparent that there was a battery or system issue with the product received.*** *After battery life (Which drained extremely fast) reached 20% or below it would automatically shut down and be unable to be turned on until plugged into a charger. Even in worse cases at 40% or 50% it might randomly shut down.*"

In this study, the aspect-based sentiment analysis of IBM Watson for the performance estimation exhibits that the sentiment intensity of the keywords can be measured in more patterns by considering the overall sentiment. The use of IBM Watson can also reduce the time to develop the keyword sentiment classifier [16].

According to Sec. 4.4, for the importance estimation, the SNN utilized in previous research was considered to determine the optimal neural networks. However, 80% (12/15) of the optimal neural networks were DNNs with more than two hidden layers in most cases for all, high-end and mid-range, smartphones (Table 6). These DNNs generally perform the prediction task better than the SNN [58]. This paper proposes the SHAP-based method to estimate the importance values of the input features from the optimal neural networks, including both the SNN and DNN, and achieving more than 80% accuracy in the optimal neural networks ensures the reliability of the importance estimation.

Furthermore, the SHAP-based method is compared with the SNN-based method to identify whether it derives constant importance values. The SNN-based method used a hidden unit between 10, 21, and 31, the "tanh" activation function, and an optimizer chosen among "SGD," "Adagrad," "Adadelta," "RMSProp," "Nadam," "Adam," and "Adamax" [2,32,35]. The SNN-based method was conducted repeatedly by randomly selecting the initial parameters of the existing study, such as the training set, hidden unit, and optimizer, in each trial (Fig. 9). The SHAP-based method was applied by sharing the same training set as the SNN-based method. The importance estimation by the SHAP-based method (Fig. 10) exhibited importance values with a lower variance compared with those with the SNN-based method.

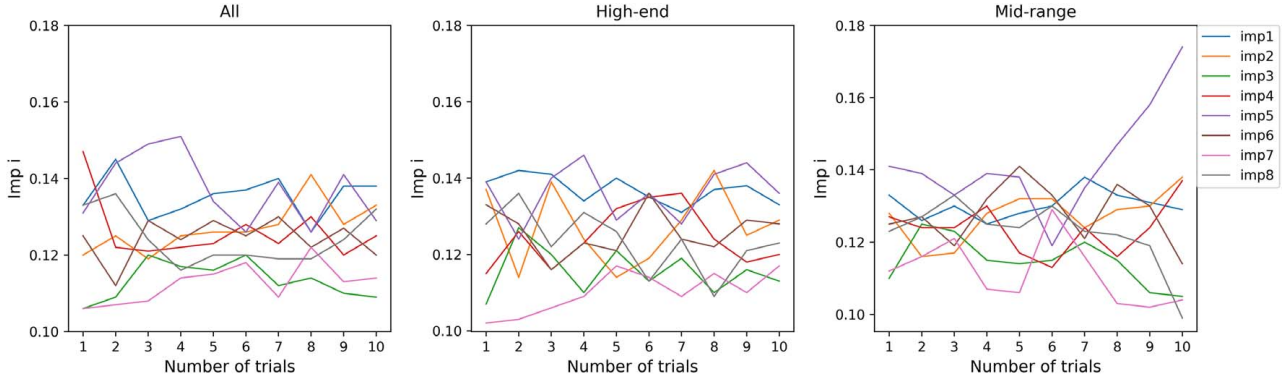

Fig. 9 Importance values by the SNN-based method

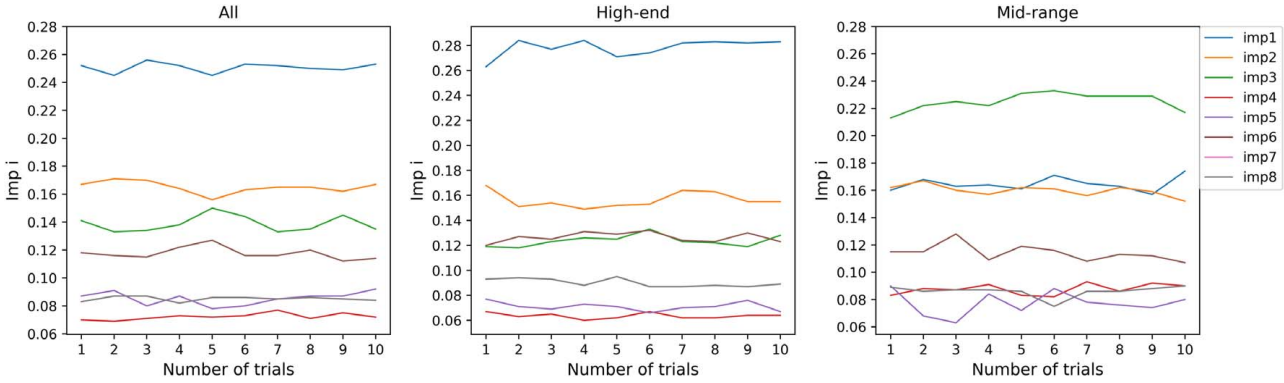

Fig. 10 Importance values by the SHAP-based method

Table 9 Top five product-related keywords of "screen" and "battery" attributes of high-end and mid-range smartphones

Product	Screen	Sentiment intensity	Battery	Sentiment intensity
High-end	Screen	−0.145	Battery	−0.171
	Case	−0.160	Life	0.110
	Button	−0.519	Battery life	0.038
	Size	0.424	Charge	−0.407
	Fingerprint	−0.019	Power	−0.163
Mid-range	Screen	0.026	Battery	0.182
	Case	−0.089	Life	0.491
	Size	0.606	Battery life	0.498
	Display	0.345	Charge	−0.122
	Fingerprint	−0.148	Power	−0.281

5 Discussion

This section discusses the following three aspects. First, the application of the proposed approach in product design is described. Second, to further analyze the strengths and weaknesses of the proposed approach, and the use of online product reviews as a source of the IPA is discussed. Finally, the use of the explainable DNN for the importance estimation is presented. A detailed discussion on each aspect is provided below.

5.1 Application of the Proposed Approach in Product Design. The proposed approach for conducting IPA can be used to identify customer needs in the early stages of product design. It provides an opportunity for comparing various product attributes by identifying the performance and importance of each product attribute from the perspective of the customers. The comparison of these product attributes can lead to effective resource allocation for enhancing competitiveness by providing customer-centric solutions, based on which the attributes of the product should be strengthened. The proposed approach can also be used to perform IPA by following various product groups, such as high-end and mid-range products of a company. The product segmentation can be evaluated from the perspective of the customers. In the case study, the IPA plots of the high-end and mid-range smartphones distinguished "screen" and "battery" from customer perceptions. Such a distinction helps to assess whether the products of a company are well-segmented based on the customer groups. The strategy of product segmentation may be considered as a failure if the IPA plots of the high-end and mid-range products do not differ.

According to Sec. 3.3, for performance estimation, the proposed approach can measure the sentiment intensities of the product-related keywords in each product attribute, which range from −1 (i.e., negative) to 1 (i.e., positive). The sentiment intensity helps to identify the response of a customer to the product feature. In the case study, the sentiment intensities of the top five product-related keywords for the "screen" and "battery" attributes of the high-end and mid-range smartphones, which can measure other product attributes but exhibit the most differences, were measured (Table 9). For both high-end and mid-range smartphones, "size" as a screen attribute was positively rated, whereas "button" was perceived as the most negative for high-end smartphones. The strengths of "size" need to be maintained in the next-generation products of both smartphone lines, and the weaknesses of "button" for high-end smartphones should be addressed. The top five product-related keywords of "battery" were generally negatively evaluated for high-end smartphones, compared with mid-range

smartphones. However, "battery" and "battery life" were positively recognized, and "charge" and "power" were negatively recognized for mid-range smartphones. For high-end smartphones, the capacity and charging method of the "battery attribute" need to be improved overall. For mid-range smartphones, the advantages of the "capacity" need to be maintained and the weaknesses of the "charging method" need to be addressed. Consequently, the sentiment values of product-related keywords can be used to identify the customer response to product features, and companies can redesign them to increase customer satisfaction.

5.2 Use of Online Product Reviews for the Importance–Performance Analysis. Online product reviews are used as the data source in the proposed approach to conduct IPA. Numerous studies have used online product reviews to identify customer needs and preferences [11,13,14,17,24,26,27]. These reviews are a significant information source for customer needs analysis, because customers participate actively when they write these reviews. A large volume of online reviews is also easier to obtain than surveys. This study requires overall ratings along with textual reviews for importance estimation. Online product reviews provide a large amount of labeled data; however, it is not easy to acquire labeled data for the development of a machine learning model in natural language processing.

However, the representativeness of online product reviews could be questioned if different products or brands are considered. For example, if numerous customers are not accustomed to buying products online or posting online reviews, the representativeness may not be ensured. Conversely, if numerous customers prefer buying some products online and posting online product reviews, online reviews of these products will be more representative. With the increase in online users, online reviews are increasing. IPA based on these reviews can be applied to increasing number of cases with the increase in the representativeness.

Furthermore, in an IPA survey, participant information can be easily obtained before the survey, whereas, in an online review, this information is relatively difficult to obtain. Using online reviews cannot measure the performance and importance of various customer segments according to demographic, geographic, behavioral, and psychographic segmentation. The results derived from the proposed approach indicate aggregated performance and importance of the reviewer group, compared to the performance and importance of various customer segments [13]. Considering the reviewer group as the representative of all the consumers may be biased. Despite these limitations, using online reviews for the IPA provides the benefits of obtaining and analyzing a large amount of data in a short time.

5.3 Use of the Explainable Deep Neural Network for the Importance Estimation. This study uses an explainable DNN to estimate the importance of each product attribute. A DNN is a powerful technique used to predict and assign weights to input features by identifying the non-linear relationships between the input and output variables. In this DNN, one of the explainable techniques, the SHAP method, was used to infer the weight of each input feature. The SHAP method provides a unique solution by considering various orders between the features, compared with other explainable DNN techniques such as local interpretable model-agnostic explanation and layer-wise relevance propagation [44]. In this study, the importance value estimated by the SHAP method in the DNN with high accuracy is reliable.

However, the proposed SHAP-based method for the importance estimation needs to build models from multiple training sets and measure importance values in various relationships of the input feature in each model. This estimation is more time-consuming than the SNN-based method. To reduce the required time, a genetic algorithm was used for the optimal neural network architecture design. This reduced the search time by half if the initial parameters were appropriately assigned, compared with the brute force search. Based on experiments, assigning the initial parameters of generation and population, which affect the search time, as 50 and 10, respectively, is recommended. However, if there is a significant change in the performance of the optimal model after increasing generation and population, they can be increased. In the case study, determining the optimal neural networks in the five training sets of the all smartphones category required 58.3 min using the genetic algorithm; brute force search required twice as much time. Although the optimal neural networks designed by the genetic algorithm are near the global optimum, they cannot be considered strictly as the global optimum [43]. Despite the limitation of the genetic algorithm, the SHAP-based method provides constant and reliable importance values, compared with the SNN-based method.

6 Conclusion and Future Work

This paper proposes an approach to perform the IPA of product attributes from online reviews. This is the first attempt to illustrate IPA plots from online product reviews for product design. This study first used an LDA-based method [36] to identify product attributes; this method automates the process of filtering out the keywords not related to the product in the keyword preprocessing of the LDA. Second, the aspect-based sentiment analysis of IBM Watson was used for estimating the performance of each product attribute. By the sentiment analysis of IBM Watson, the sentiment intensity of each product attribute was measured with a lower development cost and higher reliability compared with those of the existing methods, such as sentence sentiment analysis and dependency parser with sentiment dictionaries. Finally, the SHAP-based method was proposed to estimate the importance of each product attribute. In the SHAP-based method, the importance values were inferred from explainable DNNs with higher performance and explainability than an SNN-based method. It provides importance values with a low variance over several trials.

The limitations of this research will provide directions for further research. First, the popular convolutional neural networks, such as LeNet, AlexNet, and VGGNet, have been considered as the optimal neural network architecture [58]; however, they yielded lower performance than the feedforward neural network. Therefore, the proposed approach considered a feedforward neural network as the optimal neural network architecture. Future studies can explore convolutional neural network architectures with higher performance or other neural network architectures, because the input features are relatively sparse data. Second, the aspect-based sentiment analysis and explainable DNN techniques for the performance and importance estimation, respectively, are continually improving. Future studies may use a better estimation method while using the proposed approach. Third, the proposed approach cannot identify the importance and performance of the new features of newly released products in real-time. Future studies can achieve real-time monitoring by presenting a fully automated model for conducting IPA from online reviews. Finally, future research can improve the sampling level by linking user information on a review site to the company-owned information regarding that particular customer. This can be achieved by matching common information, such as user names and e-mail addresses. The combination with the customer information (e.g., region, age, gender, and usage experience) available with the company helps in conducting IPA in various customer groups.

Acknowledgment

This research was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2019R1I1A1 A01063298).

Conflict of Interest

There are no conflicts of interest.

Data Availability Statement

The authors attest that all data for this study are included in the paper.

Nomenclature

I = total number of product attributes
K = total number of training sets, test sets, and models
M = total number of customer reviews
N = set of all the features
S = all the feature subsets
w_k = normalized performance measure (i.e., accuracy value) of the kth prediction model
A_i = ith product attribute
R_i = total number of online reviews including the sentiments concerning product attribute A_i; sum (1 if $S_{im} > 0$; 0 otherwise)
S_{im} = sentiment score value of mth online review corresponding product attribute A_i
$\hat{y}_{\text{fused}}$ = fusion model
$f_k(x)$ = kth model
$v(S \cup i)$ = influence of the set of features with order and feature i in prediction v
$v(S)$ = influence of the set of features with order in prediction v
Imp_{ik} = importance value of product attribute A_i in kth neural network
$\hat{\text{Imp}}_i$ = importance value of product attribute A_i in fused models
$\overline{\text{Imp}_i}$ = normalized importance value of product attribute A_i
Perf_i = performance value of product attribute A_i
SHAP_{imk} = deep SHAP value of mth online review corresponding product attribute A_i in kth neural network
TR_k = total number of customer reviews in kth training set
$\phi_i(v)$ = the Shapley value of feature i in prediction v

References

[1] Martilla, J. A., and James, J. C., 1977, "Importance–Performance Analysis," J. Market., **41**(1), pp. 77–79.
[2] Bi, J.-W., Liu, Y., Fan, Z.-P., and Zhang, J., 2019, "Wisdom of Crowds: Conducting Importance-Performance Analysis (IPA) Through Online Reviews," Tourism Manage., **70**, pp. 460–478.
[3] Chu, R. K., and Choi, T., 2000, "An Importance-Performance Analysis of Hotel Selection Factors in the Hong Kong Hotel Industry: A Comparison of Business and Leisure Travellers," Tourism Manage., **21**(4), pp. 363–377.
[4] Deng, W., 2007, "Using a Revised Importance–Performance Analysis Approach: The Case of Taiwanese Hot Springs Tourism," Tourism Manage., **28**(5), pp. 1274–1284.
[5] Seng Wong, M., Hideki, N., and George, P., 2011, "The Use of Importance–Performance Analysis (IPA) in Evaluating Japan's E-Government Services," J. Theor. Appl. Electron. Commerce Res., **6**(2), pp. 17–30.
[6] Izadi, A., Jahani, Y., Rafiei, S., Masoud, A., and Vali, L., 2017, "Evaluating Health Service Quality: Using Importance Performance Analysis," Int. J. Health Care Qual. Assurance, **30**(7), pp. 656–663.
[7] Dahlgaard-Park, S. M., Pezeshki, V., Mousavi, A., and 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), pp. 82–92.
[8] MacDonald, E., Backsell, M., Gonzalez, R., and Papalambros, P., 2006, "The Kano Method's Imperfections, and Implications in Product Decision Theory," Proceedings of the 2006 International Design Research Symposium, Lisbon, Portugal, Nov. 1–4, pp. 1–12.
[9] Joung, J., Jung, K., Ko, S., and Kim, K., 2019, "Customer Complaints Analysis Using Text Mining and Outcome-Driven Innovation Method for Market-Oriented Product Development," Sustainability, **11**(1), p. 40.
[10] Ordenes, F. V., Theodoulidis, B., Burton, J., Gruber, T., and Zaki, M., 2014, "Analyzing Customer Experience Feedback Using Text Mining: A Linguistics-Based Approach," J. Service Res., **17**(3), pp. 278–295.
[11] Zhou, F., Jiao, R. J., and Linsey, J. S., 2015, "Latent Customer Needs Elicitation by Use Case Analogical Reasoning From Sentiment Analysis of Online Product Reviews," ASME J. Mech. Des., **137**(7), p. 071401.
[12] Zimmermann, M., Ntoutsi, E., and Spiliopoulou, M., 2015, "Discovering and Monitoring Product Features and the Opinions on Them With Opinstream," Neurocomputing, **150**, pp. 318–330.
[13] Hou, T., Yannou, B., Leroy, Y., and Poirson, E., 2019, "Mining Changes in User Expectation Over Time From Online Reviews," ASME J. Mech. Des., **141**(9), p. 091102.
[14] Suryadi, D., and Kim, H., 2018, "A Systematic Methodology Based on Word Embedding for Identifying the Relation Between Online Customer Reviews and Sales Rank," ASME J. Mech. Des., **140**(12), p. 121403.
[15] Zhang, H., Sekhari, A., Ouzrout, Y., and Bouras, A., 2016, "Jointly Identifying Opinion Mining Elements and Fuzzy Measurement of Opinion Intensity to Analyze Product Features," Eng. Appl. Artif. Intell., **47**, 122–139.
[16] Jeong, B., Yoon, J., and Lee, J.-M., 2019, "Social Media Mining for Product Planning: A Product Opportunity Mining Approach Based on Topic Modeling and Sentiment Analysis," Int. J. Inform. Manage., **48**, 280–290.
[17] Jiang, H., Kwong, C., and Yung, K., 2017, "Predicting Future Importance of Product Features Based on Online Customer Reviews," ASME J. Mech. Des., **139**(11), p. 111413.
[18] Rai, R., 2012, "Identifying Key Product Attributes and Their Importance Levels From Online Customer Reviews," ASME 2012 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, Chicago, IL, Aug. 12–15, pp. 533–540.
[19] Decker, R., and Trusov, M., 2010, "Estimating Aggregate Consumer Preferences From Online Product Reviews," Int. J. Res. Market., **27**(4), pp. 293–307.
[20] Chen, W., Conner, C., and Yannou, B., 2015, "User Needs and Preferences in Engineering Design," ASME J. Mech. Des., **137**(7), p. 070301.
[21] Wang, W., Li, Z., Tian, Z., Wang, J., and Cheng, M., 2018, "Extracting and Summarizing Affective Features and Responses From Online Product Descriptions and Reviews: A Kansei Text Mining Approach," Eng. Appl. Artif. Intell., **73**, pp. 149–162.
[22] Singh, A., and Tucker, C. S., 2017, "A Machine Learning Approach to Product Review Disambiguation Based on Function, Form and Behavior Classification," Decis. Support Syst., **97**, 81–91.
[23] Liu, Y., Jin, J., Ji, P., Harding, J. A., and Fung, R. Y., 2013, "Identifying Helpful Online Reviews: A Product Designer's Perspective," Comput. Aided Des., **45**(2), pp. 180–194.
[24] Chaklader, R., and Parkinson, M. B., 2017, "Data-Driven Sizing Specification Utilizing Consumer Text Reviews," ASME J. Mech. Des., **139**(11), p. 111406.
[25] Ferguson, T., Greene, M., Repetti, F., Lewis, K., and Behdad, S., 2015, "Combining Anthropometric Data and Consumer Review Content to Inform Design for Human Variability," ASME 2015 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, Boston, MA, Aug. 2–5.
[26] Zhou, F., Ayoub, J., Xu, Q., and Jessie Yang, X., 2020, "A Machine Learning Approach to Customer Needs Analysis for Product Ecosystems," ASME J. Mech. Des., **142**(1), p. 011101.
[27] Suryadi, D., and Kim, H. M., 2019, "A Data-Driven Approach to Product Usage Context Identification From Online Customer Reviews," ASME J. Mech. Des., **141**(12), p. 121104.
[28] Wang, W., Feng, Y., and Dai, W., 2018, "Topic Analysis of Online Reviews for Two Competitive Products Using Latent Dirichlet Allocation," Electron. Commer. Res. Appl., **29**, 142–156.
[29] El Dehaibi, N., Goodman, N. D., and MacDonald, E. F., 2019, "Extracting Customer Perceptions of Product Sustainability From Online Reviews," ASME J. Mech. Des., **141**(12), p. 121103.
[30] Wang, L., Youn, B., Azarm, S., and Kannan, P., 2011, "Customer-Driven Product Design Selection Using Web Based User-Generated Content," ASME 2011 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, Washington, DC, Aug. 28–31, pp. 405–419.
[31] Nasim, Z., and Haider, S., 2017, "Absa Toolkit: An Open Source Tool for Aspect Based Sentiment Analysis," Int. J. Artif. Intell. Tools, **26**(6), p. 1750023.
[32] Mikulić, J., and 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), pp. 5144–5153.
[33] Garver, M. S., 2003, "Best Practices in Identifying Customer-Driven Improvement Opportunities," Ind. Mark. Manage., **32**(6), pp. 455–466.
[34] Myers, J. H., and Alpert, M. I., 1977, "Semantic Confusion in Attitude Research: Salience Vs. Importance Vs. Determinance," ACR North Am. Adv., **4**, pp. 106–110.
[35] Deng, W.-J., Chen, W.-C., and Pei, W., 2008, "Back-Propagation Neural Network Based Importance–Performance Analysis for Determining Critical Service Attributes," Exp. Syst. Appl., **34**(2), pp. 1115–1125.
[36] Joung, J., and Kim, H. M., 2020, "Automated Keyword Filtering in LDA for Identifying Product Attributes From Online Reviews," ASME J. Mech. Des., pp. 1–10.
[37] Blei, D. M., Ng, A. Y., and Jordan, M. I., 2003, "Latent Dirichlet Allocation," J. Mach. Learn. Res., **3**(Jan.), pp. 993–1022.
[38] Mimno, D., Wallach, H. M., Talley, E., Leenders, M., and McCallum, A., 2011, "Optimizing Semantic Coherence in Topic Models," Proceedings of the Conference on Empirical Methods in Natural Language Processing, Edinburgh, Scotland, UK, July 27–31, pp. 262–272.
[39] Miller, G. A., 1995, "Wordnet: A Lexical Database for English," Commun. ACM, **38**(11), pp. 39–41.
[40] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J., "Distributed representations of words and phrases and their compositionality," Advances in

Neural Information Processing Systems 26 (NIPS 2013), Stateline, CA, Dec. 5–10, pp. 3111–3119.
[41] Rehurek, R., and Sojka, P., 2010, "Software Framework for Topic Modelling With Large Corpora," Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks, Valletta, Malta, May 22.
[42] Ramage, D., and Rosen, E.,2011, Stanford Topic Modeling Toolbox, http://nlp.stanford.edu/software/index.shtml.
[43] Miller, G. F., Todd, P. M., and Hegde, S. U., 1989, "Designing Neural Networks Using Genetic Algorithms." ICGA, Morgan Kaufmann, Palo Alto, CA, Vol. 89, pp. 379–384.
[44] Lundberg, S. M., and Lee, S. -I., "A Unified Approach to Interpreting Model Predictions," Advances in Neural Information Processing Systems 30 (NIPS 2017), Long Beach, CA, Dec. 4–9, pp. 4765–4774.
[45] Delen, D., Sharda, R., and Kumar, P., 2007, "Movie Forecast Guru: A Web-Based DSS for Hollywood Managers," Decis. Support Syst., **43**(4), pp. 1151–1170.
[46] Friedman, J., Hastie, T., and Tibshirani, R., 2001, *The Elements of Statistical Learning*, Vol. 1, Springer Series in Statistics, New York.
[47] McLachlan, G. J., Do, K. -A., and Ambroise, C., 2005, *Analyzing Microarray Gene Expression Data*, Vol. 422, John Wiley & Sons, Hoboken, NJ.
[48] Pedamonti, D., 2018, Comparison of Non-Linear Activation Functions for Deep Neural Networks on MNIST Classification Task. ArXiv180402763 Cs Stat, http://arxiv.org/abs/1804.02763
[49] Kingma, D. P., and Ba, J., 2014, "Adam: A method for stochastic optimization," 3rd International Conference on Learning Representations (ICLR 2015), San Diego, CA, May 7–9.
[50] Arifovic, J., and Gencay, R., 2001, "Using Genetic Algorithms to Select Architecture of a Feedforward Artificial Neural Network," Phys. A: Stat. Mech. Appl., **289**(3–4), pp. 574–594.
[51] Davis, L., 1991, *Handbook of Genetic Algorithms*, Van Nostrand Reinhold, New York.
[52] Batchelor, R., and Dua, P., 1995, "Forecaster Diversity and the Benefits of Combining Forecasts," Manage. Sci., **41**(1), pp. 68–75.
[53] Azzopardi, E., and Nash, R., 2013, "A Critical Evaluation of Importance–Performance Analysis," Tourism Manage., **35**, pp. 222–233.
[54] Eskildsen, J. K., and Kristensen, K., 2006, "Enhancing Importance–Performance Analysis," Int. J. Product. Perform. Manage., **55**(1), pp. 40–60.
[55] Box, G. E., and Meyer, R. D., 1986, "An Analysis for Unreplicated Fractional Factorials," Technometrics, **28**(1), pp. 11–18.
[56] Bekkar, M., Djemaa, H. K., and Alitouche, T. A., 2013, "Evaluation Measures for Models Assessment Over Imbalanced Data Sets," J. Inf. Eng. Appl., **3**(10), pp. 27–38.
[57] Bi, J.-W., Liu, Y., Fan, Z.-P., and Cambria, E., 2019, "Modelling Customer Satisfaction From Online Reviews Using Ensemble Neural Network and Effect-Based Kano Model," Int. J. Product. Res., **57**(22), pp. 7068–7088.
[58] Goodfellow, I., Bengio, Y., and Courville, A., 2016, *Deep Learning*, MIT Press, Cambridge, MA.