International Journal of Information Management 70 (2023) 102641

Contents lists available at ScienceDirect

International Journal of Information Management

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

Research article

# Interpretable machine learning-based approach for customer segmentation for new product development from online product reviews

Junegak Joung [a], Harrison Kim [b,*]

[a] School of Interdisciplinary Industrial Studies, Hanyang University, Seoul 04763, Republic of Korea
[b] Department of Industrial and Enterprise Systems Engineering, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA

ARTICLE INFO

*Keywords:*
Customer segmentation
Explainable artificial intelligence
Text mining
Online review mining
Product planning

ABSTRACT

For new product development, previous segmentation methods based on demographic, psychographic, and purchase behavior information cannot identify a group of customers with unsatisfied needs. Moreover, segmentation is limited to sales promotions in marketing. Although needs-based segmentation considering customer sentiments on product features can be conducted to develop a new product concept, it cannot identify commonalities among customers owing to their diverse preferences. Therefore, this paper proposes an interpretable machine learning-based approach for customer segmentation for new product development based on the importance of product features from online product reviews. The technical challenges of determining the importance of product features in each review are identifying and interpreting the nonlinear relations between satisfaction with product features and overall customer satisfaction. In this study, interpretable machine learning is used to identify these nonlinear relations with high performance and transparency. A case study on a wearable device is conducted to validate the proposed approach. Customer segmentation using the proposed approach based on importance is compared with that employing a previous approach based on sentiments. The results show that the proposed approach presents a higher clustering performance than the previous approach and offers opportunities to identify new product concepts.

## 1. Introduction

With the rapid development of information and communication technology, the Internet is being used by numerous users daily to search and share information. This increase in Internet usage has led to the rapid growth of online traveler community TripAdvisor, e-commerce company Amazon, and app stores, which have collected numerous customer reviews for various products and services (Angelopoulos et al., 2021; Zuo et al., 2022. These customer reviews can influence the purchases of other customers by the word-of-mouth effect and offer companies an opportunity to obtain customer opinions (Chang et al., 2019; Xu et al., 2017. Company managers can analyze customer reviews to improve the quality of their products and services or adopt measures to proactively manage the word-of-mouth effect. Online customer reviews are advantageous over surveys in that numerous reviews can be collected in a short time for analysis. Moreover, online reviews provide richer information about product experiences and customer preferences than surveys because customers actively write reviews (Joung et al., 2021.

However, studies conducting customer segmentation based on online reviews are scarce. Most previous studies have used customer transaction databases, including demographic, psychographic, and purchase behavior information, to realize customer segmentation, instead of using online reviews (Cheng & Chen, 2009; Sarvari et al., 2016. Such customer segmentation is mainly conducted by executives to identify customers for targeting marketing and advertising programs. However, these previous segmentation methods have limitations in providing managers with customer segment information on product features that customer groups prefer or find unsatisfactory (Ulwick, 2005. Consequently, previous segmentation methods cannot be utilized for new product development. By contrast, the advantage of conducting customer segmentation using online reviews is that customer groups with unsatisfied needs can be rapidly identified based on customer preferences. Needs-based customer segmentation provides efficient and effective opportunities for new product development by identifying the unsatisfied needs of customer groups. Consequently, companies can strategically design various products to meet the unsatisfied needs of each customer segment.

* Corresponding author.
*E-mail addresses:* june30@hanyang.ac.kr (J. Joung), hmkim@illinois.edu (H. Kim).

https://doi.org/10.1016/j.ijinfomgt.2023.102641
Received 12 July 2021; Received in revised form 20 January 2023; Accepted 10 February 2023
Available online 24 February 2023
0268-4012/

However, studies conducting customer segmentation based on online reviews have limitations in customer clustering because of the free form of online reviews and the variety of customer preferences. Previous studies have used the sentiments on product features to achieve customer segmentation from online reviews (Ahani et al., 2019; Wu, 2011; Zhang et al., 2018. Identifying common needs among customers is fundamentally difficult because of customer heterogeneity (Kamakura et al., 1996. Determining common needs based on varied sentiments of customers for product features is difficult and intangible (Ulwick, 2005. The importance of product features that customers value can be used, instead of various sentiments, by estimating the extent to which the satisfaction with each product feature affects overall customer satisfaction (Garver, 2003; Mikulić & Prebežac, 2012; Myers & Alpert, 1977. Importance integrates the positive and negative responses of customers into one dimension. Characterizing customers based on the importance of product features will better identify their common needs than customer characterization using various sentiments.

The technical challenges of determining the importance of product features in online reviews are twofold. The first challenge is expressing the importance of product features at the individual review level, instead of at the overall review level. Previous linear models estimate the importance of product features on an entire dataset using the coefficients of the independent variables. However, each customer must be expressed as an importance vector for segmentation. The second challenge is to identify and interpret the nonlinear relationship between product feature and overall customer satisfaction. Numerous studies have found this relationship to be nonlinear (Deng et al., 2008; Matzler et al., 2004; Mikulić & Prebežac, 2012. Although recent machine learning models can identify these nonlinear relations, they cannot interpret them because of their black-box nature.

Therefore, to realize needs-based customer segmentation for new product development, this study developed an interpretable machine learning (IML)-based approach for customer segmentation based on the importance of product features from online product reviews. IML provides a tool to explain the identified nonlinear relations between independent and dependent variables by machine learning. Using IML, the influence of an independent variable on the dependent variable was elucidated at the individual level. Employing the IML technique enabled identification of these nonlinear relations at the individual level with a high performance and interpretability. This academic study is the first to realize needs-based customer segmentation using IML from online product reviews. With this approach, potential markets and new product concepts are better identified using the developed customer clustering method based on importance than previous clustering based on sentiments.

The developed approach comprises (1) data collection and preprocessing, (2) determination of the importance values of product features on star ratings, and (3) customer segmentation based on the importance values of product features. First, online product reviews are obtained by web scraping. Second, the importance values of product features are determined in each review using IML. After identifying the product features and their sentiments from customer reviews, a machine learning classifier is developed to predict star ratings (i.e., overall customer satisfaction) based on the sentiment values of the product features. Subsequently, the importance of the product features on star ratings is derived in each review using Shapley additive explanations (SHAP) as the IML technique. Finally, customer clustering is performed to identify customer segments based on customer reviews, which are expressed as vectors of the importance of product features. For new product development, unsatisfied customer segments are explored from the clustering results by identifying the customer groups with high importance and low satisfaction with specific product features. To validate the developed approach, a case study on a wearable device was conducted, and the results were compared to those of a previous approach based on the sentiments on product features. The case study results showed that the developed approach is better than a previous method in identifying commonalities among customers and offers potential markets for new product development.

The remainder of this paper is organized as follows. Section 2 presents the literature on the evolution of customer segmentation, customer segmentation using online reviews, nonlinear relationship between the satisfaction with product features and overall customer satisfaction, and IML. Section 3 describes the developed approach for customer segmentation to achieve new product development from online product reviews. Section 4 presents a case study on a fitness tracker. Section 5 discusses the theoretical contributions, practical implications, and future research directions of the developed approach. Finally, Section 6 concludes the study.

## 2. Literature review

This study contributes to customer segmentation research by establishing a method for customer segmentation based on online reviews using IML. In this section, first, previous studies and their limitations are described, and subsequently, the technical challenges and techniques to solve them using IML are introduced.

### 2.1. Evolution of customer segmentation

The increase in available customer information with the development of information technology led to customer segmentation based on demographic, psychographic, and purchase behavior and needs information (Nasiopoulos et al., 2015; Ulwick, 2005. Demographic characteristics such as gender, location, and age were initially used for customer segmentation because of the easy collection of such information. With the installation of transaction databases and access to more specific customer profiles, psychographics and purchase behavior, such as recency, frequency, and monetary, in purchase histories began to be used for customer segmentation. With the development of clustering techniques, needs-based segmentation is being conducted to identify product features that are attractive or unsatisfactory to customers.

However, earlier segmentation methods based on demographics, psychographics, and purchase behavior fail to identify customer groups with unsatisfied needs (Ulwick, 2005. Although such methods can be used for sales promotions in marketing, they cannot provide practical guidelines for new product development. Needs-based segmentation provides clustering results, including diverse customer groups based on the type of need. However, these clustering results are frequently difficult and intangible because of the variety of customer preferences.

Therefore, this study aimed to fill a gap in customer segmentation research by developing a method for needs-based segmentation based on importance values of product features that can better determine the commonalities among customers than previous techniques from online product reviews.

### 2.2. Customer segmentation based on online reviews

A few studies have developed methods to perform customer segmentation based on customer preferences from online reviews. In these methods, each review is typically characterized by the importance or sentiment values of the product or service features, assuming that a customer review can be considered as a customer opinion. Subsequently, the customer reviews are clustered to realize customer segmentation based on the voice of the customer (VoC) vector, expressed as importance or sentiment. The frequency of a product or service feature as a surrogate of its importance is used to characterize the customer reviews (Greco & Polli, 2020; Jiang et al., 2015; Park & Lee, 2011. After identifying the product or service features frequently mentioned in the reviews, each review is expressed as a vector depending on the occurrence of feature words. Moreover, five-point Likert ratings of sentiments of predefined product or service features are used to characterize each customer review (Ahani et al., 2019; Wu, 2011; Zhang et al., 2018. The sentiments and frequencies of the product or service features and their star ratings are also used to vectorize each customer review (Suryadi & Kim, 2019). Based on these VoC vectors, K-means clustering, X-means clustering, latent class analysis-based clustering, and self-organizing maps are used to identify the customer segmentation. Finally, the customer reviews are clustered by estimating the similarity among customers based on the VoC vector.

However, previously developed methods have limitations in terms of the surrogate and the clustering performance. The frequency of a product or service feature is insufficient as a surrogate for importance. The product or service features frequently mentioned by customers are not always important. A high frequency of a product or service feature implies a high probability of its importance. However, the ranking of high-frequency features does not match that of their influence on the overall customer satisfaction (Joung & Kim, 2021a). Theoretical evidence on the relation between frequency and importance is also lacking (Mikulić & Prebežac, 2012; Myers & Alpert, 1977). Moreover, because the sentiments for product or service features vary depending on the degree of positivity and negativity of the reviewers, finding commonality among customers based on sentiment vectors is difficult. In addition, considering sentiments, frequency, and star ratings makes it highly challenging to derive customer segments that share the same responses to a product.

To overcome these limitations, this study developed an approach for customer segmentation based on the importance of the sentiments of product features on star ratings (i.e., overall customer satisfaction) obtained from online reviews. Theoretical proof of the improved importance estimation of overall customer satisfaction compared to using frequency as a surrogate for importance was sufficiently shown (Garver, 2003; Mikulić & Prebežac, 2012; Myers & Alpert, 1977). Clustering of VoC vectors based on importance was better in identifying the commonality among customers than that based on various sentiments.

### *2.3. Nonlinear relationship between satisfaction of product features and overall customer satisfaction*

Many studies have determined that the relationship between satisfaction with product features and overall customer satisfaction is asymmetric and nonlinear (Bi et al., 2019; Deng et al., 2008; Joung & Kim, 2021c; Matzler et al., 2004; Mikulić & Prebežac, 2012). Equal amounts of change in satisfaction with product features produce different amounts of change in overall satisfaction. Typically, the Kano model classifies product features into attractive, performance, must-be, reverse, and indifferent based on the effect of satisfaction with a product feature on overall satisfaction (Kano, 1984). For example, dissatisfaction with attractive product features has little effect on overall satisfaction; however, their incorporation has a significant influence on it. In contrast, dissatisfaction with must-be product features has a major effect on overall satisfaction, whereas their inclusion has little effect. The asymmetric effects of product feature satisfaction on overall satisfaction are also supported by empirical studies (Matzler et al., 2004; Slevitch & Oh, 2010).

Therefore, the challenges in estimating importance are identifying and interpreting the nonlinear relations among the sentiments of product features and overall customer satisfaction (Deng et al., 2008; Matzler et al., 2004; Mikulić & Prebežac, 2012). Although machine learning can capture these nonlinear relations, it cannot explain how satisfaction with a product feature affects overall customer satisfaction; this is owing to its black-box nature. Moreover, the transparency and performance of machine learning models for identifying nonlinear patterns show a trade-off relationship. The approach developed in this study uses IML to solve these problems, as discussed in the following subsection.

### *2.4. Interpretable machine learning*

IML is defined as the extraction of significant knowledge concerning the relations learned by a machine learning model (Murdoch et al., 2019). Recent machine learning models, such as artificial neural network (ANN), random forests (RF), light gradient boosting machine (LGBM), and xgboost models, present high prediction performance in various data analyses; however, they lack explainability of the prediction results. IML provides understandable explanations by interpreting a black-box machine learning model. IML is classified into intrinsically interpretable models and post-hoc interpretation methods based on intrinsic properties (Molnar, 2020). Intrinsically interpretable models are machine learning models that are explainable because of their simple structures, such as linear, short decision tree (DT), decision rule, and naïve Bayes classifier models. Post-hoc interpretation methods are techniques that interpret machine learning models after their construction. Compared with intrinsically interpretable models, the advantage of post-hoc interpretation methods is their freedom to use multiple machine learning models, because a model and its interpretation are independent. Post-hoc interpretation methods are preferred for interpreting machine learning models owing to their excellent performance compared with intrinsically interpretable models.

Post-hoc interpretation methods are categorized into local and global interpretable methods based on their scope (Covert et al., 2020). Local interpretable methods address questions such as why a model yields a certain prediction, whereas global interpretable methods answer queries such as how a trained model predicts. The former and the latter describe the influence of each feature on individual predictions and the behavior of the model across an entire dataset, respectively.

Therefore, this study used the SHAP method, which is a local interpretable method, to explain the nonlinear relations between the sentiments on product features and overall customer satisfaction in each review (Lundberg & Lee, 2017). The SHAP method provides contrasting explanations of interpreting machine learning models by considering the interactions between all possible features. Section 3 describes the process of estimating the importance of product features in each review based on the SHAP method.

## 3. Method

The overall customer segmentation process using online product reviews is shown in Fig. 1. Online reviews of a target product are the inputs, and the customer segments based on the importance values of the product features are the outputs. The importance values of the product features in each review are calculated by estimating how much positive and negative comments for the product feature affect the star ratings (i.e., overall customer satisfaction). The developed approach comprises three stages: (1) data collection and preprocessing, (2) determination of the importance values of product features on star ratings, and (3) customer segmentation based on the importance values of product features. The developed approach is automated to provide a range for hyperparameter tuning in Stages 2 and 3.

### *3.1. Data collection and preprocessing*

In the developed approach, numerous online reviews over a short period are required to ensure representativeness of the customers. The target for customer segmentation can be several product models in a particular product segment or a well-known model. Web harvesting is used to collect online reviews from e-commerce websites such as Amazon, eBay, and BestBuy. Information such as product names, authors, dates, titles, contents, and star ratings are automatically obtained from the reviews. Because each review is assumed to be a customer opinion, duplicate reviews are removed by examining identical entries of the author, title, and content. The reviews are structured into preprocessed words using part-of-speech (POS) to identify the product features. After the reviews are tokenized, text preprocessing with POS tagging is performed (Boyd-Graber et al., 2014). Specifically, uppercases are converted into lowercases, stop words (e.g., I, you, what, and is) and

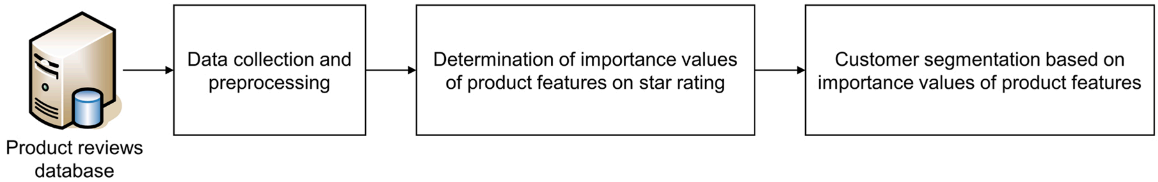


**Fig. 1.** Overall process of proposed approach.

punctuation (e.g., !"#$%&?@) are eliminated, and words are lemmatized (e.g., “steps” is converted to its root form “step”). The reviews are also structured as original sentences with emoticons and emojis to estimate the sentiments of the identified features.

### *3.2. Determination of importance values of product features on star ratings*

The importance values of the product features are determined in each review by estimating the influence of their sentiments on star ratings (i.e., overall customer satisfaction). These importance values reflect the behavioral outcomes of the customers for overall satisfaction (Myers & Alpert, 1977. If the overall satisfaction is positive, the importance of satisfactory product features is high, whereas if it is negative, the importance of unsatisfactory ones is high. The process of determining the importance values of the product features in each review is as follows. First, product feature words expressed by the customers are identified from the reviews, and the words corresponding to a particular product feature are grouped using a word-embedding-based method (Suryadi & Kim, 2018. A previous study manually set hyperparameters for word embedding (Suryadi & Kim, 2018; however, in this study, the optimal result among the combinations of multiple hyperparameters is searched. Second, the sentiments of the identified product features in each review are estimated using a valence-aware dictionary and sentiment reasoner (VADER) sentiment analysis (Hutto & Gilbert, 2014. Third, machine learning classifiers with the best prediction performance are developed using recent machine learning algorithms to predict star ratings from the sentiments of the product features. Previous studies used an ANN as the optimal classifier for prediction (Bi et al., 2019; Deng et al., 2008; Geng & Chu, 2012; Joung & Kim, 2021a; Mikulić & Prebežac, 2012. However, in this study, state-of-the-art machine learning algorithms including ANNs were compared for the best prediction performance. Finally, the importance of the sentiments on the product features on the star rating is derived for each review using the developed classifiers based on the SHAP method (Lundberg & Lee, 2017.

#### *3.2.1. Identifying product feature words*

Latent Dirichlet allocation (LDA)- (Blei et al., 2003; Joung & Kim, 2021b or word embedding- (Suryadi & Kim, 2018-based methods can be used to identify and group product feature words. LDA requires manual effort to interpret the product features in topics (Joung & Kim (2021b) (2018); Suryadi &. A word-embedding-based method is used because it can automatically identify and group product feature words for given hyperparameters for a word-embedding vector. Because product feature words have been assumed to be nouns in many previous studies (Abulaish et al. (2009); Guo et al. (2009); Hu & Liu (2004); Joung & Kim (2021a); Suryadi & Kim (2018), in this study, nouns were considered as candidates for product feature words. A word-embedding-based method comprises the following four steps:

1. Word2vec is used as the word-embedding technique to vectorize the nouns in the reviews (Mikolov et al., 2013. Word2vec represents the words occurring in reviews as vectors based on the context (i.e., surrounding words). The dimensions, window sizes, and cutoff frequencies of the words are set as hyperparameters to utilize Word2vec.
2. Affinity propagation (AP) clustering is used to group very similar nouns based on the constructed noun vectors (Frey & Dueck, 2007. Different from clustering algorithms such as hierarchical clustering, K-means clustering, and spectral clustering, AP clustering automatically determines the number of clusters. Each cluster is named by the nouns with the highest term frequencies within it. These nouns are regarded as words that represent the cluster. However, some clusters contain nouns (e.g., today, thinking, stars, friends, and husbands) that are unrelated to product features.
3. The clusters unrelated to product features are filtered using product manuals. Product description sections of the product manuals provide customer terms related to the important product features (Joung & Kim, 2021b. After removing the noise words mainly occurring in the customer reviews, such as the product name, product, process, and nouns representing time (e.g., hour, day, week, month, and year) (Jeong et al., 2019, the clusters related to the product features are identified. This is achieved by examining whether the nouns representing each cluster appear in the product description section.
4. Refinement is performed to improve cluster cohesion and separation. Cluster cohesion, which represents the distance between the nouns in a cluster, is improved by removing the nouns whose similarity to the noun representing the cluster is lower than the threshold in each cluster. Cluster separation, which represents the distance between clusters, is improved by combining clusters if the similarity between the nouns representing them is higher than the threshold. For example, if the similarity between “step” and “calorie” is higher than the threshold, these two clusters are merged into “step_calorie”. Refinement increases the similarity between the nouns in each cluster and avoids redundancy between the clusters.

#### *3.2.2. Estimating sentiments of product features*

The developed method includes a VADER sentiment analysis to estimate the sentiments of the identified product features (Hutto & Gilbert, 2014. This analysis, which is an unsupervised method, was selected because it is readily available in other domains without requiring manual labeling of the training data. A VADER sentiment analysis is particularly performed for estimating the sentiments of social media texts by considering emoticons (e.g.,:) and:D) and frequently used slangs (e.g., “nah” and “sux”). It scores sentiments from − 1 (very negative) to 1 (very positive). For example, “good screen” was estimated to have a sentiment value of 0.4404, “good screen!!!” of 0.5826, and “good screen: D” of 0.7865. A VADER sentiment analysis estimates sentiments by averaging the affective lexicons in a sentence, including product feature words, based on well-established word banks and predefined heuristic rules. For example, in the sentence, “I was surprised that I began having problems with the screen after three months and now it has stopped working completely.”, the “screen” feature had a sentiment score of − 0.4019. In each review, the sentiment of a product feature is calculated by averaging the sentiments of the sentences that mention it.

#### *3.2.3. Developing machine learning classifiers with best prediction performance*

Various machine learning algorithms—DT, RF, LGBM, xgboost, catboost, and ANN—are considered to obtain the optimal classifiers. The classifiers are developed to predict star ratings based on the sentiment scores of product features. In the construction of the machine learning classifiers, the input variables are the sentiment scores of the product features (i.e., $pf_1, pf_2, \ldots, pf_I$ in the 1, 2, …, $M$ review). The output variables are the labels of the positive and negative star ratings corresponding to the reviews (Fig. 2). The star ratings are transformed into two labels (i.e., negative and positive) because the predictive power of the classifiers is higher when predicting two labels than when predicting five-star ratings (Joung & Kim, 2021a. One- and two-star ratings are considered negative labels, and four- and five-star ratings are considered positive labels. Three-star ratings are classified as positive or negative based on the sentiment of the corresponding review using VADER sentiment analysis.

Using the above dataset, a nested cross-validation is conducted to optimize the hyperparameters of the classifiers and derive the best classifiers from the entire dataset. The optimization is aimed to overcome the bias in performance evaluation depending on the configuration of the training, validation, and test sets (Cawley & Talbot, 2010. The nested cross-validation constitutes inner and outer loops of the entire dataset, and a $K$-fold cross-validation is conducted in each loop (Fig. 3). In the outer loop, the entire dataset is first randomly partitioned into $K$ equal-sized training and test sets. Subsequently, each training set of the outer loop in the inner loop is randomly divided into $K'$ equal-sized training and validation sets. $K'$-1 sub-samples are used as the training sets to train a classifier, and the remaining single sub-sample is used as the validation set to optimize the hyperparameters of the classifier. The hyperparameters for constructing the machine learning classifier are optimized by performing a random or grid search after determining the hyperparameter range in a preliminary experiment. The performance of the classifier is evaluated using the test set corresponding to the training set of the outer loop. The prediction performance of each classifier is calculated using the f1-score, which considers both precision and recall in the test set. The f1-score is calculated as follows:

$$F1-score = 2 \times \frac{Precision \times Recall}{Precision + Recall} = \frac{2TP}{2TP + FP + FN} \tag{1}$$

where TP, FP, and FN denote true positive, false positive, and false negative, respectively.

Consequently, $K \times K'$ optimal machine learning classifiers with the best f1-score are obtained from the $K \times K'$ training sets.

#### 3.2.4. Determining importance values of product features on star ratings

The SHAP method is used to determine the influence of the sentiment scores of the product features on the star ratings (i.e., importance values) in each test set using the $K'$ optimal machine learning classifiers. The SHAP method is selected because it provides unique solutions by considering all interactions between input features based on game theory. The SHAP values of the product features are calculated as follows:

$$\phi_{pf_i}(v) = \sum_{S \subseteq F: i \notin S} \frac{|S|!(|F| - |S| - 1)!}{|F|!}(v(S \cup pf_i) - v(S)) \tag{2}$$

where $\phi_{pf_i}(v)$ denotes the Shapley value of the $i^{th}$ product feature ($pf_i$) in prediction $v$. $S$ and $F$ represent all feature subsets and sets of all features, respectively. $|S|$ and $|F|$ are the sizes of $S$ and $F$, respectively. $v(S \cup pf_i)$ denotes the contribution of the set of features with order and feature $i$. $v(S)$ represents the contribution of a set of features with an order.

The contribution of a product feature to each prediction is calculated by estimating the change in the prediction in the absence of that specific feature. The absolute SHAP value of product feature $i$ indicates its importance on the star rating in each review from the $k'th$ classifier if its sentiment score is nonzero (Eq. (3)). Consequently, each review is expressed as an importance vector (i.e., $\overline{Imp_{im}}$) by the weighted sum of the $K'$ importance values of the corresponding review based on the prediction performance of the $K'$ classifiers (Eq. (4)) (Table 1). The importance values represent the effects of the customer sentiments for the product features on the star ratings.

$$Imp_{imk'} = |SHAP_{imk'}| \tag{3}$$

$$\begin{aligned} \overline{Imp}_{im} &= \sum_{k'=1}^{K'} \overline{w}_{k'} Imp_{imk'} \\ \overline{w}_{k'} &= \frac{w_{k'}}{\sum_{k'=1}^{K'} w_{k'}} \end{aligned} \tag{4}$$

### 3.3. Customer segmentation based on importance values of product features

In the developed method, AP clustering is performed to group the customer reviews based on the importance values of the product features because it automatically determines the number of clusters. Each cluster $c$ is characterized by the average importance of the product features in cluster (i.e., $Imp_{1c}, Imp_{2c}, \ldots, Imp_{Ic}$). Customer clustering based on the importance values of product features provides companies market segmentation based on the commonalities among their customers who prefer specific product features. The characteristics of each segmented market can be explored by averaging the sentiments for these product features (i.e., $Sat_{1c}, Sat_{2c}, \ldots, Sat_{Ic}$) or the star ratings of each customer group. For example, a customer group that values "screen" expresses a negative sentiment for it on average is dissatisfied with the "screen" feature of the product. If this group also has an average star rating lower than 2, it indicates that the group is dissatisfied with the product. Based on the information on the importance of and satisfaction with specific product features in these market segments, companies can design a new product for each segment by enhancing those features that each customer group prefers.

| | Input variables | | | | Output variable |
|---|---|---|---|---|---|
| | $pf_1$ | $pf_2$ | … | $pf_I$ | Label |
| 1 | 0.296 | 0 | … | 0 | Positive |
| 2 | 0 | -0.679 | … | 0 | Negative |
| 3 | 0 | 0 | … | 0 | Negative |
| … | … | … | … | … | Positive |
| $M$ | 0 | 0 | … | 0.273 | Positive |

**Fig. 2.** Example of dataset for constructing machine learning classifiers.

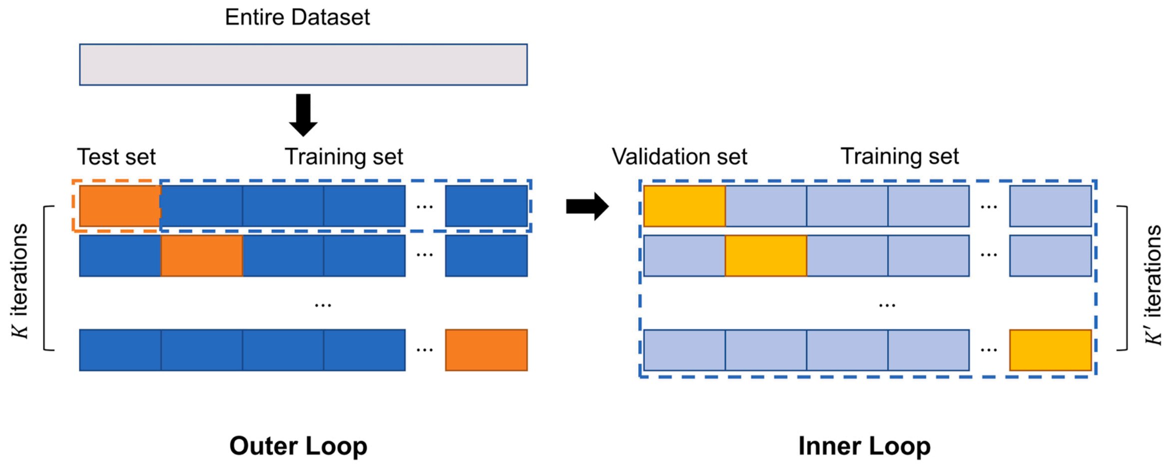


**Fig. 3.** Nested cross-validation.

**Table 1**
Example of importance values of product features in reviews.

| Review | $pf_1$ | $pf_2$ | … | $pf_i$ |
|---|---|---|---|---|
| 1 | $\overline{Imp}_{11}$ | 0 | … | 0 |
| 2 | 0 | $\overline{Imp}_{22}$ | … | 0 |
| 3 | 0 | 0 | … | 0 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| M | $\overline{Imp}_{1M}$ | $\overline{Imp}_{2M}$ | … | $\overline{Imp}_{iM}$ |

More specifically, the opportunity score of each customer segment for new product development is calculated based on the importance of and satisfaction with product feature $i$ that each customer group $c$ prefers (Eq. (5)) (Ulwick, 2002. The importance and satisfaction scores of each customer group are transformed into a scale of 0–10 by min–max scaling (Jeong et al., 2019. The opportunity score ranges from 0 to 20. Customer segments with opportunity scores of 10 or higher are considered potential markets because the product features that these customer groups prefer are important to overall customer satisfaction but currently do not satisfy the customer group needs. These customer segments become potential customers to fill unsatisfied needs if a product feature that they value is enhanced in the next-generation products. Based on this opportunity score, companies can quantitatively identify opportunities from customer segmentation for new product development.

$$Opportunity_c = Imp_{ic} + Max(Imp_{ic} - Sat_{ic}, 0) \quad (5)$$

## 4. Case study

A case study using Fitbit Charge 3 was conducted to validate the proposed approach. Fitbit Charge 3 was chosen because it is a well-known model for which sufficient reviews can be obtained in a short period. Customer segmentation of Fitbit Charge 3 was conducted based on online product reviews. Fitbit Charge 3 is a motion-measuring device that provides functions such as step, clock, floors, heart rate, sleep stages, swim, customizable band, and GPS.

### 4.1. Data collection and preprocessing

Customer reviews on Fitbit Charge 3 were collected from verified purchases on Amazon using WebScraper.io. Overlapping reviews were removed by identifying the same author, title, and contents, and unwritten reviews in the United States were removed by identifying the country at the beginning of the date entry, such as "Reviewed in Canada on May 20, 2019." The analysis period for customer segmentation was selected from October 2018 to March 2020 because Fitbit Charge 3 was released in October 2018, and the next version, Fitbit Charge 4, was released at the end of March 2020. During this period, customers purchased Fitbit Charge 3, the latest model in the Fitbit Charge series. Each review was structured into preprocessed words with POS and original sentences to identify the product feature words and perform sentiment analysis, as described in Section 4.2. The NLTK package in Python is used for the structure. After removing the reviews that did not contain product feature words or sentiments, as described in Section 4.2, a total of 5045 reviews were analyzed. The reviews for the analysis included comments on product features and their sentiments, e.g.,

1. "I was surprised that I began having problems with the ***screen*** after 3 months and now it has stopped working completely. It didn't last a year:("
2. "First day I've had it fails to ***sync*** and now it doesn't work at all"
3. "I love the challenge option to compete with friends and the ***sleep*** and ***heart rate*** monitor features as well"

### 4.2. Determining importance values of product features on star ratings

The importance values of the product features in each review were determined using the process described in Section 3.2. First, the product feature words in the reviews were identified and grouped by clustering, filtering, and refinement based on Word2vec. The hyperparameters of Word2vec—dimensions, window sizes, and cutoff frequencies of the words—were determined to be 50, 4, and 10, respectively, based on the lowest Davies–Bouldin index (DBI) value. This hyperparameter tuning with the lowest DBI value yielded the optimal clustering result (Table 2) (Davies & Bouldin, 1979. The number of clusters was 71, and the name of each cluster was defined by the noun that appeared most frequently in it. Python gensim package was used to construct Word2vec, and scikit-learn package in Python was employed to perform AP clustering.

**Table 2**
Clustering performance based on combination of Word2vec hyperparameters.

| Dimensions | Window size | Cutoff | DBI index |
|---|---|---|---|
| 50 | 4 | 5 | 2.18 |
| 50 | 4 | 10 | 2.04 |
| 50 | 5 | 5 | 2.15 |
| 50 | 5 | 10 | 2.05 |
| 100 | 4 | 5 | 2.18 |
| 100 | 4 | 10 | 2.07 |
| 100 | 5 | 5 | 2.15 |
| 100 | 5 | 10 | 2.10 |

After clustering, the nouns unrelated to the product features were filtered by examining whether these nouns representing each cluster were present in the Fitbit Charge 3 product manual. Clusters including "price," "money," "cost," and "value" that did not exist in the product manual were also considered because in previous studies, price sensitivity was adopted as an element for customer segmentation (Bolton & Myers, 2003; Masiero & Nicolau, 2012; Petrick, 2005; Stangl et al., 2020. After filtering, refinement was performed to improve the clustering cohesion and separation by applying a similarity threshold of 0.5. The average cosine similarity between the words within each cluster improved from 0.481 to 0.593. Fifteen product features were identified from the online reviews (Table 3).

Second, the sentiments of the identified product features were estimated by performing VADER sentiment analysis (Fig. 4). The VADER library of Python[1] was used to estimate the sentiments of sentences containing the product feature words. In Fig. 4, the x-axis represents the product features, and the y-axis represents the number of positive and negative reviews including the product features. If the sentiment score of the product features in each review is greater than 0.05 or lesser than − 0.05, the review is considered positive or negative (Hutto & Gilbert, 2014. Overall, the number of positive reviews is greater than that of negative reviews, and for "screen," "sync," "update," and "charger," the number of negative reviews are similar or greater than that of positive reviews.

Third, machine learning classifiers were developed to predict positive and negative star rating labels based on the sentiment scores of the 15 product features. The ratio of the positive to negative labels was 6:4. After a five-fold cross-validation in each loop (e.g., outer loop: $K = 5$, inner loop: $K' = 5$), 25 machine learning classifiers were developed to predict the 5 test sets (Table 4). The f1-score was obtained as the prediction performance metric for the five test sets. Modern machine learning algorithms—DT, RF, LGBM, xgboost, catboost, and ANN—were used to obtain the best classifiers by performing a grid search for hyperparameter tuning. Xgboost exhibited the highest f1-score among the machine learning algorithms. The scikit-learn package in Python was used to execute the DT, RF, and ANN models. The LGBM,[2] xgboost,[3] and catboost libraries[4] in Python were used to implement the corresponding algorithms.

Finally, the importance values of the product features in each review on the star ratings were estimated using the tree SHAP[5] from five test sets (Eqs. (2), (3), and (4) (Table 1). This method was selected because xgboost, which presented the best performance, is a tree-based model. False predictions were considered in the importance estimation because they were logical results that predicted negative labels when the sentiments of the product features were negative. In each review, the absolute value of SHAP exhibited the importance value of the star rating well by considering the influence of each feature based on feature interaction. For example, in a negative review, where the sentiment score of "sync" was − 0.4019 and the sentiment scores of the remaining features were zero, the importance value of "sync" on the star rating was 2.092. In a positive review where the sentiment scores of "heart_rate" and "notification" were 0.3888 and 0.6937, respectively, and the sentiment scores of the remaining features were zero, the importance values of "heart_rate" and "notification" were 0.8404 and 0.5311, respectively. The importance value of "heart_rate" was higher than that of "notification" because the sentiment score of the former was more influential than that of the latter in the predictions for all reviews.

**Table 3**
Product features of Fitbit Charge 3.

| Number | Product features | Frequent words | # of words | # of reviews |
|---|---|---|---|---|
| $pf_1$ | step_calorie | step, track, activity, exercise, calorie | 47 | 2816 |
| $pf_2$ | app_data | app, data, information, info, option | 19 | 1632 |
| $pf_3$ | screen | screen, face, display, clock, button | 10 | 1545 |
| $pf_4$ | sleep | sleep | 6 | 1391 |
| $pf_5$ | sync | sync, blutooth, iphone | 26 | 1156 |
| $pf_6$ | heart_rate | heart, rate, hr | 13 | 1102 |
| $pf_7$ | money_price | money, price, quality | 11 | 973 |
| $pf_8$ | notification | notification, text, call, message, alarm | 19 | 873 |
| $pf_9$ | band | band, size, design | 22 | 839 |
| $pf_{10}$ | battery_life | battery, life | 3 | 798 |
| $pf_{11}$ | support | support, service | 6 | 704 |
| $pf_{12}$ | water | water, waterproof, swim, shower | 11 | 701 |
| $pf_{13}$ | wrist | wrist, hand, arm | 6 | 583 |
| $pf_{14}$ | update | update, software | 3 | 552 |
| $pf_{15}$ | charger | charger | 7 | 273 |

### 4.3. Customer segmentation based on importance values of product features

Customer segments were identified by performing AP clustering based on the 15 importance values of the product features. The average importance values of the 15 product features and star ratings for the entire reviews were calculated to interpret the customer clustering results. These were as follows: $pf_1$ ("step_calorie") = 0.166, $pf_2$ ("app_data") = 0.072, $pf_3$ ("screen") = 0.111, $pf_4$ ("sleep") = 0.075, $pf_5$ ("sync") = 0.063, $pf_6$ ("heart_rate") = 0.061, $pf_7$ ("money_price") = 0.105, $pf_8$ ("notification") = 0.035, $pf_9$ ("band") = 0.06, $pf_{10}$ ("battery_life") = 0.063, $pf_{11}$ ("support") = 0.091, $pf_{12}$ ("water") = 0.045, $pf_{13}$ ("wrist") = 0.017, $pf_{14}$ ("update") = 0.021, $pf_{15}$ ("charger") = 0.015, and average star ratings = 3.32. The number of customer clusters was 115; the largest cluster included 474 reviews, whereas the smallest cluster included 7 reviews. The top customer clusters containing more than 100 reviews were 12, and the average importance values of the product features were calculated by rounding to the third decimal place for each top cluster (Table 5). Among the top customer clusters cluster 1 was related to $pf_1$ ("step_calorie"), the most important in all reviews Customer cluster 1 was generally satisfied with the product because its average star rating exceeded 3. Among the top customer clusters, the clusters related to $pf_7$ ("money_price") and $pf_3$ ("screen"), which were the second and third most important in all reviews, were clusters 2 and 3. These customer clusters were generally dissatisfied with the product, because their average star ratings were less than 3. Most of the top customer clusters had customer segments that valued each product feature; however, customer segments that valued $pf_{13}$ ("wrist"), $pf_{14}$ ("update"), and $pf_{15}$ ("charger") were not observed in the top clusters. The sizes of customer groups that valued these product features was considered small.

For new product development from the customer segmentation, opportunity scores were calculated based on the importance of and satisfaction with the product features preferred by each customer group (Table 6). The customer segments with an opportunity score of 10 or more that could become potential markets were clusters 2, 3, 5, 7, and 9. Product features significant to the overall customer satisfaction but presently unsatisfactory were "money_price," "screen," "support," "sync," and "battery_life." Therefore, a company manufacturing Fitbit can plan to develop various next-generation products for each customer cluster based on a divide-and-conquer strategy. For example, the company could strategically develop a cheaper version of the product, a

[1] https://github.com/cjhutto/vaderSentiment
[2] https://github.com/microsoft/LightGBM
[3] https://github.com/dmlc/xgboost
[4] https://github.com/catboost/catboost
[5] https://github.com/slundberg/shap

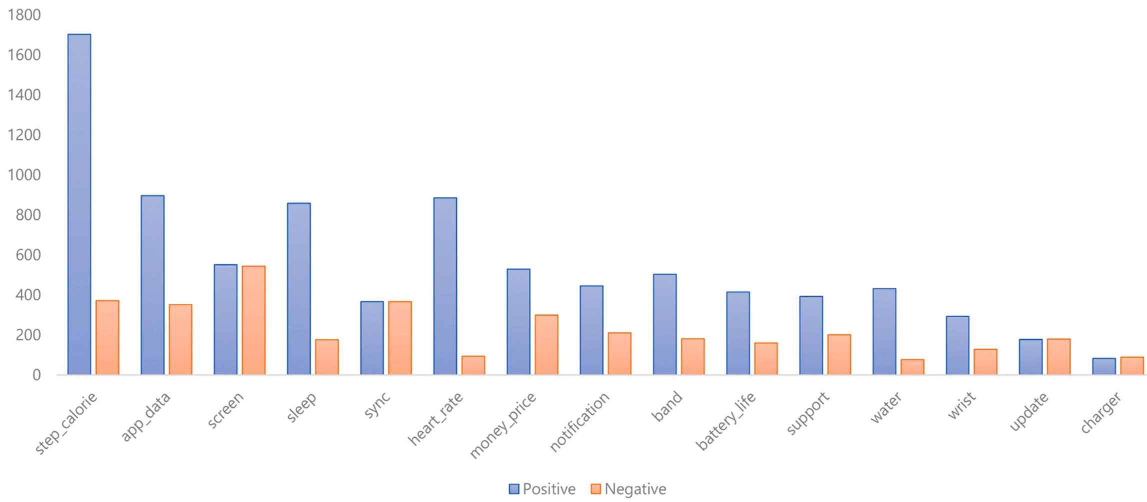


**Fig. 4.** Number of positive and negative reviews of 15 product features.

**Table 4**
Comparison of prediction performance of different machine learning algorithms.

| Outer loop | Inner loop | DT | RF | LGBM | Xgboost | Catboost | ANN |
|---|---|---|---|---|---|---|---|
| $k = 1$ | $k' = 1$ | 0.776 | 0.799 | 0.812 | 0.828 | 0.822 | 0.821 |
| | $k' = 2$ | 0.789 | 0.797 | 0.804 | 0.826 | 0.814 | 0.821 |
| | $k' = 3$ | 0.769 | 0.797 | 0.813 | 0.829 | 0.825 | 0.820 |
| | $k' = 4$ | 0.776 | 0.796 | 0.815 | 0.824 | 0.823 | 0.828 |
| | $k' = 5$ | 0.774 | 0.793 | 0.806 | 0.827 | 0.824 | 0.822 |
| $k = 2$ | $k' = 1$ | 0.781 | 0.808 | 0.843 | 0.843 | 0.842 | 0.828 |
| | $k' = 2$ | 0.776 | 0.808 | 0.844 | 0.848 | 0.846 | 0.822 |
| | $k' = 3$ | 0.781 | 0.822 | 0.843 | 0.840 | 0.846 | 0.820 |
| | $k' = 4$ | 0.779 | 0.814 | 0.840 | 0.844 | 0.844 | 0.813 |
| | $k' = 5$ | 0.773 | 0.811 | 0.842 | 0.837 | 0.843 | 0.821 |
| $k = 3$ | $k' = 1$ | 0.792 | 0.816 | 0.842 | 0.850 | 0.843 | 0.833 |
| | $k' = 2$ | 0.791 | 0.819 | 0.843 | 0.846 | 0.848 | 0.837 |
| | $k' = 3$ | 0.791 | 0.819 | 0.842 | 0.845 | 0.852 | 0.847 |
| | $k' = 4$ | 0.789 | 0.824 | 0.840 | 0.854 | 0.840 | 0.836 |
| | $k' = 5$ | 0.791 | 0.825 | 0.835 | 0.856 | 0.849 | 0.845 |
| $k = 4$ | $k' = 1$ | 0.766 | 0.806 | 0.818 | 0.842 | 0.834 | 0.826 |
| | $k' = 2$ | 0.766 | 0.804 | 0.817 | 0.842 | 0.828 | 0.835 |
| | $k' = 3$ | 0.766 | 0.804 | 0.819 | 0.839 | 0.824 | 0.822 |
| | $k' = 4$ | 0.772 | 0.817 | 0.820 | 0.845 | 0.824 | 0.824 |
| | $k' = 5$ | 0.766 | 0.812 | 0.820 | 0.839 | 0.824 | 0.830 |
| $k = 5$ | $k' = 1$ | 0.779 | 0.814 | 0.833 | 0.854 | 0.849 | 0.829 |
| | $k' = 2$ | 0.777 | 0.814 | 0.833 | 0.844 | 0.846 | 0.824 |
| | $k' = 3$ | 0.780 | 0.808 | 0.837 | 0.843 | 0.851 | 0.829 |
| | $k' = 4$ | 0.776 | 0.817 | 0.834 | 0.853 | 0.849 | 0.834 |
| | $k' = 5$ | 0.777 | 0.816 | 0.835 | 0.852 | 0.845 | 0.833 |
| Avg. | | 0.778 | 0.810 | 0.829 | **0.842** | 0.837 | 0.828 |

product with an improved screen size or resolution, improved customer support, solve sync problems, and improve battery life for each segment.

### 4.4. Validation

Customer segmentation using the proposed approach based on the importance values of product features was validated by comparing it with a previous approach based on sentiments of features (Ahani et al., 2019; Suryadi & Kim, 2019; Wu, 2011; Zhang et al., 2018. In the case study, based on the sentiments of the 15 product features (Fig. 2), customer segmentation was performed using a previous approach by applying the same AP clustering used in our proposed approach. The top customer clusters containing more than 100 reviews were characterized by averaging the sentiment values of all product features. The clustering

**Table 6**
Opportunity scores of top customer clusters.

| Cluster | Value feature | Importance | Satisfaction | Opportunity |
|---|---|---|---|---|
| 1 | step_calorie ($pf_1$) | 5.28 | 6.21 | 5.28 |
| 2 | money_price ($pf_7$) | 9.05 | 4.93 | **13.17** |
| 3 | screen ($pf_3$) | 7.02 | 3.69 | **10.35** |
| 4 | band ($pf_9$) | 6.18 | 5.7 | 6.66 |
| 5 | support ($pf_{11}$) | 10 | 4.63 | **15.37** |
| 6 | app_data ($pf_2$) | 4.18 | 5 | 4.18 |
| 7 | sync ($pf_5$) | 6.32 | 2.45 | **10.19** |
| 8 | sleep ($pf_4$) | 4.48 | 7.8 | 4.48 |
| 9 | battery_life ($pf_{10}$) | 7.58 | 3.37 | **11.79** |
| 10 | water ($pf_{12}$) | 6.33 | 6.94 | 6.33 |
| 11 | heart_rate ($pf_6$) | 4.01 | 8.09 | 4.01 |
| 12 | notification ($pf_8$) | 3.89 | 5.52 | 3.89 |

**Table 5**
Top customer clusters for Fitbit Charge 3 using proposed approach.

| Number | # of reviews | $pf_1$ | $pf_2$ | $pf_3$ | $pf_4$ | $pf_5$ | $pf_6$ | $pf_7$ | $pf_8$ | $pf_9$ | $pf_{10}$ | $pf_{11}$ | $pf_{12}$ | $pf_{13}$ | $pf_{14}$ | $pf_{15}$ | Ratings |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 474 | **0.83** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.01 | 0 | 0 | 0 | 4.27 |
| 2 | 302 | 0.01 | 0.02 | 0.01 | 0.01 | 0.01 | 0.01 | **1.43** | 0.02 | 0 | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 | 2.57 |
| 3 | 295 | 0.01 | 0.01 | **1.11** | 0.01 | 0 | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 | 0 | 0.01 | 0.01 | 0 | 0.01 | 2.41 |
| 4 | 185 | 0 | 0.01 | 0.01 | 0 | 0 | 0 | 0 | 0.02 | **0.97** | 0 | 0 | 0.01 | 0.02 | 0.02 | 0.01 | 4.02 |
| 5 | 179 | 0.02 | 0.02 | 0.02 | 0.02 | 0.01 | 0.01 | 0.02 | 0.06 | 0 | 0.01 | **1.58** | 0.01 | 0 | 0 | 0 | 1.93 |
| 6 | 161 | 0 | **0.66** | 0.01 | 0 | 0 | 0 | 0 | 0.01 | 0.01 | 0 | 0 | 0 | 0 | 0 | 0.01 | 3.71 |
| 7 | 153 | 0.02 | 0.03 | 0.01 | 0 | **1** | 0.02 | 0.02 | 0.02 | 0 | 0 | 0 | 0 | 0 | 0.01 | 0 | 2.40 |
| 8 | 144 | 0.04 | 0.01 | 0.01 | **0.71** | 0 | 0.01 | 0 | 0.01 | 0 | 0.01 | 0 | 0.01 | 0.01 | 0.01 | 0.01 | 4.17 |
| 9 | 136 | 0.02 | 0 | 0.02 | 0 | 0.01 | 0.01 | 0.02 | 0.01 | 0 | **1.19** | 0 | 0 | 0.02 | 0.01 | 0 | 3.80 |
| 10 | 119 | 0.02 | 0 | 0.01 | 0.02 | 0 | 0 | 0 | 0.01 | 0.02 | 0.01 | 0 | **1** | 0.01 | 0 | 0.01 | 4.29 |
| 11 | 114 | 0.01 | 0.01 | 0.01 | 0 | 0 | **0.63** | 0.01 | 0 | 0.01 | 0 | 0.01 | 0 | 0 | 0 | 0 | 3.39 |
| 12 | 106 | 0.01 | 0.01 | 0.02 | 0.02 | 0.01 | 0.01 | 0 | **0.61** | 0.01 | 0 | 0 | 0 | 0.01 | 0.01 | 0 | 3.63 |

results obtained by the proposed and previous approaches were compared in terms of clustering performance (Table 7). The number of clusters with more than 100 reviews and fewer than 10 reviews obtained using the proposed approach was greater than and less than those obtained using the previous approach, respectively. The proposed approach had a smaller DBI value (Davies & Bouldin, 1979 than that of the previous approach. These results show that importance-value-based customer clustering using the proposed approach achieves better customer segmentation than sentiment-value-based customer clustering using the previous approach.

Moreover, the top customer clusters derived from the previous and proposed approaches were compared (Tables 5 and 8). Top customer clusters 2–5 characterized by the sentiments on $pf_1$ ("step_calorie"), $pf_3$ ("screen"), and $pf_7$ ("money_price") using the previous approach were consistent with top customer clusters 1, 2, and 3 characterized by the importance of $pf_1$ ("step_calorie"), $pf_3$ ("screen"), and $pf_7$ ("money_price") using the proposed approach. However, customer cluster 1, which contained the most reviews based on the previous approach, could not be characterized based on the sentiments of the product features. In comparison, the proposed approach characterized the 12 top customer clusters well based on the importance of product features. Customer segmentation using the previous approach was performed based on various positive and negative responses to the features. Consequently, the customers in the case study were not grouped well, as demonstrated by a larger DBI value than that by the proposed approach. In contrast, customer segmentation by the developed approach characterized the customers using the importance values of the product features, allowing the customers to be grouped well, as shown by a smaller DBI value than that of the previous approach.

## 5. Discussion

This study developed an approach to realize customer segmentation for new product development from online product reviews. The develped approach involves customer clustering using IML to represent each customer review in terms of the importance values of product features to realize overall customer satisfaction. This study makes theoretical and practical contributions to customer segmentation, data-driven information systems (IS), and information management (IM) research (Davison, 2022; Kane, 2022; Kar & Dwivedi, 2020; Struijk et al., 2022. In this section, the relevance of the proposed approach is discussed from both the theoretical and practical perspectives. In addition, limitations of the developed approach and future research directions are presented.

### 5.1. Theoretical contributions

This study contributes to customer segmentation research by developing a method for needs-based segmentation based on the importance values of product features to achieve overall satisfaction. Customer segmentation research widely used the recency, frequency, monetary value (RFM) model based on purchase history (Chan, 2008; Güçdemir & Selim, 2015; Namvar et al., 2011; Rahim et al., 2021; Sarvari et al., 2016. Customers are grouped based on how recently they have made a purchase, how often they buy it, and the amount spent. After determining the types of customers in each customer group, target marketing

**Table 7**
Comparison of customer segmentation for Fitbit Charge 3 using previous and proposed approaches.

| Method | Number of clusters | Number of clusters more than 100 reviews | Number of clusters less than 10 reviews | DBI value |
|---|---|---|---|---|
| Previous approach | 302 | 6 | 187 | 2.04 |
| Proposed approach | 115 | 12 | 7 | 1.26 |

**Table 8**
Top customer clusters for Fitbit Charge 3 obtained using previous approach.

| Number | # of reviews | $pf_1$ | $pf_2$ | $pf_3$ | $pf_4$ | $pf_5$ | $pf_6$ | $pf_7$ | $pf_8$ | $pf_9$ | $pf_{10}$ | $pf_{11}$ | $pf_{12}$ | $pf_{13}$ | $pf_{14}$ | $pf_{15}$ | Ratings |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 323 | − 0.01 | 0 | 0 | − 0.01 | − 0.01 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2.71 |
| 2 | 119 | **0.39** | 0 | − 0.01 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.01 | 0 | 0.01 | 0 | 4.45 |
| 3 | 111 | − 0.01 | 0 | **− 0.24** | − 0.01 | − 0.01 | 0.01 | − 0.01 | − 0.01 | − 0.01 | 0 | 0 | − 0.01 | − 0.01 | 0 | − 0.01 | 1.82 |
| 4 | 109 | **0.62** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.01 | − 0.01 | 0 | 0 | 0 | 0 | 4.61 |
| 5 | 109 | 0.01 | − 0.01 | − 0.01 | 0.01 | − 0.02 | 0 | **0.37** | 0 | 0 | − 0.01 | 0 | 0.01 | 0.01 | 0 | 0 | 2.74 |
| 6 | 100 | **0.19** | 0.01 | 0 | 0.02 | 0 | 0.01 | 0.01 | 0.01 | 0 | 0 | 0.01 | 0 | 0.01 | 0 | 0 | 3.75 |

is strategically performed. However, RFM model-based customer segmentation is limited in new product concept development because it cannot identify product features with which each customer group specifically prefers or is dissatisfied. Therefore, the previous segmentation based on the RFM model is useful for the promotion of target segments.

In contrast, for needs-based customer segmentation that is directly related to developing new product concepts, most previous studies have characterized customers based on their preferences for product features in various contexts of product usage and delivery (Greengrove, 2002; Wilson-Jeanselme & Reynolds, 2006. Customer preferences are estimated based on sentiments or importance values for product features from online reviews (Park & Lee, 2011; Suryadi & Kim, 2019. However, needs-based customer segmentation has no consensus on the selection of the importance and sentiment values of product features to characterize customers. Identifying common needs among customers based on sentiments for product features is difficult because of the diversity in customer opinion expression according to the degree of satisfaction or dissatisfaction with product features. Importance estimation based on the effect of product feature satisfaction on the overall customer satisfaction for customer characterization provides theoretical evidence (Garver, 2003; Mikulić & Prebežac, 2012; Myers & Alpert, 1977. For importance estimation, the IML technique is used to identify and explain the nonlinear relationship between the satisfaction of product features and the overall customer satisfaction. It provides explanatory values for understanding how product feature satisfaction affects the overall customer satisfaction in each instance. In this study, using the SHAP method, which is an IML technique, each customer is characterized as a vector of importance-based explanatory values.

Briefly, this research contributes to customer segmentation research by providing an IML-based approach to realize customer segmentation for new product development from online product reviews. Using the proposed approach, opportunities for developing new product concepts were identified based on clusters with high importance and low satisfaction with specific product features (Table 6). It also provides data-driven evidence that customer characterization by the importance of product features clearly identifies commonalities among customers rather than by the sentiments of product features (Table 7). This study differs from other studies in that it solves the technical challenges of identifying and interpreting nonlinear relations to characterize each customer based on the importance of product features using IML.

Moreover, the IML technique used to explain the black-box machine learning model in this study also has potential for other data-driven IS and IM research applications. Numerous data-driven IS and IM studies have used machine learning to improve managers' decision-making in various fields (Georgiadou et al., 2020. However, in such studies, constructing a machine learning model with high accuracy and explainability is important for interpreting social phenomena (Dwivedi et al., 2021. In this study, IML was used to explain the influence of the independent variable on the dependent variable in each prediction using a machine learning model. Owing to IML, the influence of product features in each review on predicting overall customer satisfaction was measured. Finally, based on the importance of these product features, a new customer segmentation method was developed. This study lays the foundation for applying IML to data-driven IS and IM studies by demonstrating the advancement of customer segmentation theory using IML. The IML technique can be employed to explain the nonlinear relationships between independent and dependent variables in various prediction models for data-driven IS and IM research.

### 5.2. Practical implications

The proposed approach provides potential markets and new product concepts to product development and design managers for customer-driven product development. Using the proposed approach, product development and design managers can explore potential markets by identifying customer segments with high importance and low satisfaction with specific product features. They can also determine a new product concept for potential customers by enhancing the specific product features of the value of each customer segment. The customer groups that valued the price, screen, customer support, sync, and battery life of wearable devices were identified as potential markets (Table 6). A new product concept can be developed by improving the price, screen, customer support, sync, and battery life of wearable devices. Based on customer segmentation using the proposed approach, a company can strategically plan various new products to meet each unsatisfactory customer segment.

Moreover, the proposed approach includes the processes of identifying product features and estimating the sentiment and importance values of the identified product features for customer segmentation from online product reviews. Each process provides information on customer needs through product development, and design managers based on online reviews. In identifying product features, product development and design managers can identify product features in which customers are primarily interested. Fifteen product features of a wearable device that customers mainly mentioned were identified (Table 3). In estimating the sentiments of identified product features, product development and design managers can understand the strengths and weaknesses of current product features. The majority of product features can be considered strengths, because the number of positive reviews is greater than that of negative reviews. However, screen, syn, update, and charger can be regarded as weaknesses, owing to the relatively large number of negative reviews (Fig. 4). When deriving the importance of product features, product development and design managers can identify the relatively important product features from all customers. In the case study, the important values of step and calorie measurement, screen, and price were as high as 0.1 from the entire review.

Finally, customer segmentation by the proposed approach is performed based on online reviews. Product development and design managers can conduct customer segmentation by collecting large volumes of online reviews rapidly and cost-free. Because customer needs can change rapidly over time and with the environment, as occurred during Covid-19, product development and design managers can discover new opportunities for product development by exploring changing customer segments in online reviews. In the presented case study, the total runtime of the proposed approach is approximately 1 h 30 min on a PC with 16 GB RAM, Intel i9–9880 H, and Windows 10. Thus, an automated customer review analysis can identify customer needs faster than a survey can.

### 5.3. Limitations and future research

This study has some limitations that will provide directions for further research. First, the proposed approach uses the importance values of the product features for customer characterization. More demographic, psychographic, and purchase behavior information (Zuo et al., 2022 can be considered in future studies combined with the importance of product features to identify various customer segments with unsatisfied needs. The combination of these types of information provides companies with a more complete segmentation strategy for product positioning and target marketing. Second, the proposed approach is verified by performing a case study on a wearable device. Using the proposed approach, customer segments with unsatisfied needs for wearable devices were identified. The proposed approach is effective for a product type that can be designed to satisfy unsatisfactory customer groups. However, the effectiveness of the proposed approach is limited to a product type that has already satisfied customer needs and has no new customer requirements. In the product type that does not require functionally improved product development, the proposed approach is limited to target marketing. In future, case studies on various product types can be performed to improve the versatility of the proposed approach. Third, the proposed approach identified customer segments with high importance and low satisfaction with specific product features. Targeting customer segments with unsatisfied needs is useful in determining the initial product concept in the product development process; however, it cannot directly guide the specifications and functions of a new product. Future studies should conduct in-depth research on customer segments with unsatisfactory needs to determine the specifications of a new product. Finally, the proposed approach conducted AP clustering using the scikit-learn package of Python. However, if the number of customer reviews exceeds 50,000, clustering using this package may become extremely slow. Future studies can enhance computer specifications or apply additional techniques to solve these speed and big data problems.

## 6. Conclusions

In this study, an IML-based approach is developed to achieve customer segmentation from online product reviews. First, customer reviews of the target product were collected, followed by text preprocessing. Second, the importance of the product features in each review was determined by estimating the influence of the sentiment of the product features on star ratings based on the IML technique. The use of the IML technique in this study enabled identifying and interpreting the nonlinear relation between the satisfaction of product features and the overall customer satisfaction at the individual level, which previous models failed to understand. Based on this study, the IML technique has the potential to be used to interpret various black box prediction models in information management research. Finally, the unsatisfied customer segments are identified by AP clustering and an opportunity score based on the importance and satisfaction of the product features. In a case study, a comparison of the previous approach and the proposed method showed that the proposed importance-based customer clustering has a higher performance than the previous sentiment-based approach. The proposed approach provides managers with opportunities to develop customized products based on unsatisfactory product features, which are important for overall satisfaction. The proposed approach was automated to the maximum extent by assigning hyperparameters and thresholds to each stage. This automation can reduce the time and labor required for continuous customer segmentation over time compared with those required for a survey.

## CRediT authorship contribution statement

**Junegak Joung**: Conceptualization, Methodology, Software, Data curation, Writing – original draft, Visualization, Investigation, Validation. **Harrison Kim**: Supervision, Writing – review & editing.

## Acknowledgments

This research was supported by the research fund of Hanyang University (HY-202200000001129), the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-2021S1A5A8065184).

## References

Abulaish, M., Doja, M. N., Ahmad, T., et al., 2009.Feature and opinion mining for customer review summarization, in: International Conference on Pattern Recognition and Machine Intelligence, Springer. pp. 219–224.

Ahani, A., Nilashi, M., Ibrahim, O., Sanzogni, L., & Weaven, S. (2019). Market segmentation and travel choice prediction in spa hotels through tripadvisor's online reviews. *International Journal of Hospitality Management, 80*, 52–77.

Angelopoulos, S., Brown, M., McAuley, D., Merali, Y., Mortier, R., & Price, D. (2021). Stewardship of personal data on social networking sites. *International Journal of Information Management, 56*, Article 102208.

Bi, J. W., Liu, Y., Fan, Z. P., & Cambria, E. (2019). Modelling customer satisfaction from online reviews using ensemble neural network and effect-based kano model. *International Journal of Production Research, 57*, 7068–7088.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *The Journal of Machine Learning Research, 3*, 993–1022.

Bolton, R. N., & Myers, M. B. (2003). Price-based global market segmentation for services. *Journal of Marketing, 67*, 108–128.

Boyd-Graber, J., Mimno, D., & Newman, D. (2014). Care and feeding of topic models: Problems, diagnostics, and improvements. *Handbook of Mixed Membership Models and Their Applications*, Article 225255.

Cawley, G. C., & Talbot, N. L. (2010). On over-fitting in model selection and subsequent selection bias in performance evaluation. *The Journal of Machine Learning Research, 11*, 2079–2107.

Chan, C. C. H. (2008). Intelligent value-based customer segmentation method for campaign management: A case study of automobile retailer. *Expert Systems with Applications, 34*, 2754–2762.

Chang, Y. C., Ku, C. H., & Chen, C. H. (2019). Social media analytics: Extracting and visualizing hilton hotel ratings and reviews from tripadvisor. *International Journal of Information Management, 48*, 263–279.

Cheng, C. H., & Chen, Y. S. (2009). Classifying the segmentation of customer value via rfm model and rs theory. *Expert Systems with Applications, 36*, 4176–4184.

Covert, I., Lundberg, S., & Lee, S. I. (2020). Understanding global feature contributions with additive importance measures. *Advances in Neural Information Processing Systems*, 33.

Davies, D. L., & Bouldin, D. W. (1979). A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 224–227.

Davison, R. M., 2022.Impact and implications for practice.

Deng, W. J., Chen, W. C., & Pei, W. (2008). Back-propagation neural network based importance-performance analysis for determining critical service attributes. *Expert Systems with Applications, 34*, 1115–1125.

Dwivedi, Y. K., Hughes, L., Ismagilova, E., Aarts, G., Coombs, C., Crick, T., Duan, Y., Dwivedi, R., Edwards, J., Eirug, A., et al. (2021). Artificial intelligence (ai): Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy. *International Journal of Information Management, 57*, Article 101994.

Frey, B. J., & Dueck, D. (2007). Clustering by passing messages between data points. *Science, 315*, 972–976.

Garver, M. S. (2003). Best practices in identifying customer-driven improvement opportunities. *Industrial Marketing Management, 32*, 455–466.

Geng, X., & Chu, X. (2012). A new importance-performance analysis approach for customer satisfaction evaluation supporting pss design. *Expert Systems with Applications, 39*, 1492–1502.

Georgiadou, E., Angelopoulos, S., & Drake, H. (2020). Big data analytics and international negotiations: Sentiment analysis of brexit negotiating outcomes. *International Journal of Information Management, 51*, Article 102048.

Greco, F., & Polli, A. (2020). Emotional text mining: Customer profiling in brand management. *International Journal of Information Management, 51*, Article 101934.

Greengrove, K. (2002). Needs-based segmentation: Principles and practice. *International Journal of Market Research, 44*, 1–16.

Güçdemir, H., & Selim, H. (2015). Integrating multi-criteria decision making and clustering for business customer segmentation. *Industrial Management & Data Systems*.

Guo, H., Zhu, H., Guo, Z., Zhang, X., Su, Z., 2009.Product feature categorization with multilevel latent semantic association, in: Proceedings of the 18th ACM conference on information and knowledge management, pp. 1087–1096.

Hu, M., Liu, B., 2004.Mining opinion features in customer reviews, in: AAAI, pp. 755–760.

Hutto, C., Gilbert, E., 2014.Vader: A parsimonious rule-based model for sentiment analysis of social media text, in: Proceedings of the International AAAI Conference on Web and Social Media.

Jeong, B., Yoon, J., & Lee, J. M. (2019). Social media mining for product planning: A product opportunity mining approach based on topic modeling and sentiment analysis. *International Journal of Information Management, 48*, 280–290.

Jiang, S., Cai, S., Olle, G. O., & Qin, Z. (2015). Durable product review mining for customer segmentation. *Kybernetes*.

Joung, J., & Kim, H. M. (2021a). Approach for importance-performance analysis of product attributes from online reviews. *Journal of Mechanical Design, 143*, Article 081705.

Joung, J., & Kim, H. M. (2021b). Automated keyword filtering in latent dirichlet allocation for identifying product attributes from online reviews. *Journal of Mechanical Design, 143*, Article 084501.

Joung, J., & Kim, H. M. (2021c). Explainable neural network-based approach to kano categorisation of product features from online reviews. *International Journal of Production Research*, 1–21.

Joung, J., Kim, K. H., & Kim, K. (2021). Data-driven approach to dual service failure monitoring from negative online reviews: Managerial perspective. *SAGE Open, 11*, Article 2158244020988249.

Kamakura, W. A., Kim, B. D., & Lee, J. (1996). Modeling preference and structural heterogeneity in consumer choice. *Marketing Science, 15*, 152–172.

Kane, G. C. (2022). How to write an "a" paper. *Journal of the Association for Information Systems, 23*, 1071–1079.

Kano, N. (1984). Attractive quality and must-be quality. *Hinshitsu (Quality, The Journal of Japanese Society for Quality Control), 14*, 39–48.

Kar, A. K., & Dwivedi, Y. K. (2020). Theory building with big data-driven research-moving away from the "what" towards the "why. *International Journal of Information Management, 54*, Article 102205.

Lundberg, S., Lee, S. I., 2017.A unified approach to interpreting model predictions. arXiv preprint arXiv:1705.07874.

Masiero, L., & Nicolau, J. L. (2012). Tourism market segmentation based on price sensitivity: Finding similar price preferences on tourism activities. *Journal of Travel Research, 51*, 426–435.

Matzler, K., Bailom, F., Hinterhuber, H. H., Renzl, B., & Pichler, J. (2004). The asymmetric relationship between attribute-level performance and overall customer satisfaction: a reconsideration of the importance-performance analysis. *Industrial Marketing Management, 33*, 271–277.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J., 2013.Distributed representations of words and phrases and their compositionality. arXiv preprint arXiv:1310.4546.

Mikulić, J., & Prebežac, D. (2012). Accounting for dynamics in attribute-importance and for competitor performance to enhance reliability of bpnn-based importance-performance analysis. *Expert Systems with Applications, 39*, 5144–5153.

Molnar, C., 2020.Interpretable machine learning. Lulu. com.

Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R., & Yu, B. (2019). Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences, 116*, 22071–22080.

Myers, J. H., & Alpert, M. I. (1977). Semantic confusion in attitude research: Salience vs. importance vs. determinance. *ACR North American Advances*.

Namvar, M., Khakabimamaghani, S., & Gholamian, M. R. (2011). An approach to optimised customer segmentation and profiling using rfm, ltv, and demographic features. *International Journal of Electronic Customer Relationship Management, 5*, 220–235.

Nasiopoulos, D. K., Sakas, D. P., Vlachos, D., Mavrogianni, A., 2015.Modeling of market segmentation for new it product development, in: AIP conference proceedings, American Institute of Physics. pp. 51–59.

Park, Y., & Lee, S. (2011). How to design and utilize online customer center to support new product concept generation. *Expert Systems with Applications, 38*, 10638–10647.

Petrick, J. F. (2005). Segmenting cruise passengers with price sensitivity. *Tourism Management, 26*, 753–762.

Rahim, M. A., Mushafiq, M., Khan, S., & Arain, Z. A. (2021). Rfm-based repurchase behavior for customer classification and segmentation. *Journal of Retailing and Consumer Services, 61*, Article 102566.

Sarvari, P. A., Ustundag, A., & Takci, H. (2016). Performance evaluation of different customer segmentation approaches based on rfm and demographics analysis. *Kybernetes*.

Slevitch, L., & Oh, H. (2010). Asymmetric relationship between attribute performance and customer satisfaction: A new perspective. *International Journal of Hospitality Management, 29*, 559–569.

Stangl, B., Prayag, G., & Polster, L. (2020). Segmenting visitors' motivation, price perceptions, willingness to pay and price sensitivity in a collaborative destination marketing effort. *Current Issues in Tourism, 23*, 2666–2682.

Struijk, M., Ou, C. X., Davison, R. M., & Angelopoulos, S. (2022). Putting the is back into is research. *Information Systems Journal, 32*, 1–4.

Suryadi, D., & Kim, H. (2018). A systematic methodology based on word embedding for identifying the relation between online customer reviews and sales rank. *Journal of Mechanical Design*, 140.

Suryadi, D., & Kim, H. M. (2019). A data-driven methodology to construct customer choice sets using online data and customer reviews. *Journal of Mechanical Design*, 141.

Ulwick, A. (2005). *What customers want*. volume 8. NY: McGraw-Hill Professional Publishing New York.

Ulwick, A. W. (2002). Turn customer input into innovation. *Harvard Business Review, 80*, 91–97.

Wilson-Jeanselme, M., & Reynolds, J. (2006). The advantages of preference-based segmentation: An investigation of online grocery retailing. *Journal of Targeting, Measurement and Analysis for Marketing, 14*, 297–308.

Wu, C. (2011). Deriving collective intelligence from reviews on the social web using a supervised learning approach. *Expert Systems with Applications, 38*, 13149–13157.

Xu, X., Wang, X., Li, Y., & Haghighi, M. (2017). Business intelligence in online customer textual reviews: Understanding consumer perceptions and influential factors. *International Journal of Information Management, 37*, 673–683.

Zhang, C., Zhang, H., & Wang, J. (2018). Personalized restaurant recommendation method combining group correlations and customer preferences. *Information Sciences, 454*, 128–143.

Zuo, M., Angelopoulos, S., Liang, Z., & Ou, C. X. (2022). Blazing the trail: Considering browsing path dependence in online service response strategy. *Information Systems Frontiers*, 1–15.

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

Harrison Kim is currently a professor in the Department of Industrial and Enterprise Systems Engineering at the University of Illinois at Urbana-Champaign. He received BS (1995) and MS (1997) from KAIST and Ph.D. (2001) from the University of Michigan. His main research interests include user-centred sustainable product design; energy systems engineering; product design analytics; renewable energy and vehicle electrification; multi-scale, multidisciplinary optimization; green product portfolio design and manufacturing.