# A Data-Driven Methodology to Construct Customer Choice Sets Using Online Data and Customer Reviews

**Dedy Suryadi**
Enterprise Systems Optimization Laboratory,
Department of Industrial and Enterprise Systems Engineering,
University of Illinois at Urbana-Champaign,
Urbana, IL 61801;
Department of Industrial Engineering,
Parahyangan Catholic University,
Bandung 40141, Indonesia
e-mails: suryadi2@illinois.edu; dedy@unpar.ac.id

**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 recent development in engineering design has incorporated customer preferences by involving a choice model. In generating a choice model to produce a good quality estimate of parameters related to product attributes, a high-quality choice set is essential. However, the choice set data are often not available. This research proposes a methodology that utilizes online data and customer reviews to construct customer choice sets in the absence of both the actual choice set and the customer sociodemographic data. The methodology consists of three main parts, i.e., clustering the products based on their attributes, clustering the customers based on their reviews, and constructing the choice sets based on a sampling probability scenario that relies on product and customer clusters. The proposed scenario is called Normalized, which multiplies the product cluster and customer cluster fractions to obtain the probability sampling distribution. There are two utility functions proposed, i.e., a linear combination of product attributes only and a function that includes the interactions of product attributes and customer reviews. The methodology is implemented to a data set of laptops. The Normalized scenario performs significantly better than the baseline, Random, in predicting the test set data. Moreover, the inclusion of customer reviews into the utility function also significantly increases the predictive ability of the model. The research shows that using the product attribute data and customer reviews to construct choice sets generates choice models with higher predictive ability than randomly constructed choice sets.* [DOI: 10.1115/1.4044198]

*Keywords: design automation, choice set, choice model, online customer reviews*

## 1 Introduction

Customer preferences have become an integral part of decision-making in engineering design. Recent researches have emphasized the importance of including customer preferences to the decision-making process. Li and Azarm [1] apply conjoint analysis to incorporate customer preferences in selecting the best product design. Kumar et al. [2] use nested logit model to accommodate customer preferences in the proposed market-driven product family design methodology. Michalek et al. [3] utilize logit model to model product demand as a part of the product line design optimization. He et al. [4] propose a choice modeling framework for usage context-based design to quantify the impact of usage context toward customer choices. Morrow et al. [5] incorporate a consider-then-choose model into engineering design optimization.

In order to describe customer preferences, an essential component of choice models is choice set. It is defined as a set of product alternatives that are available to a customer [6], who will compare the alternatives before making the final choice [7]. As the choice model is explicitly expressed in terms of product attributes, as well as sociodemographic attributes of customers, high-quality choice set generates a reliable choice model that provides better quality of parameter estimates for the product attributes. Consequently, the choice model would support designers to make better design decisions with respect to customer preferences. Therefore, choice set is also an important factor that supports design decisions in the decision-based design framework [8].

Despite its importance, while the purchase data are generally available, the choice set data are rarely recorded. Wang and Chen [7] propose a method to learn from an existing choice set information in a data set to predict the missing choice set in another data set. In addition, the customer sociodemographic data also becomes a vital information to generate the prediction. Their findings confirm that the learned choice set results in better choice models than both universal and randomly sampled choice sets, in terms of log-likelihood and pseudo R-squared measures.

While the purpose of this paper is also constructing customer choice sets to create a better choice model, the main contribution of this paper is proposing a methodology to construct customer choice sets in the absence of both existing choice set and customer sociodemographic data. In the absence of both, the methodology proposes the usage of publicly available online data of product attributes and customer reviews from e-commerce websites. It becomes a promising alternative to conduct survey for collecting customer choice set data, which can be time consuming, labor intensive, and expensive [9]. The findings in Sec. 4 show that the usage of online data and customer reviews results in a better choice model compared with the model that uses randomly sampled choice sets. Furthermore, this paper contributes to linking online self-presentation—which will be discussed in Sec. 2, in the form of customer reviews, with choice modeling. It is achieved by clustering customers based on the reviews and subsequently utilizing the customer clusters to construct customer choice sets.

The paper is organized as follows. Section 2 discusses relevant researches related to the main topics in this paper. Section 3 elaborates the proposed methodology in constructing customer choice sets using online data and customer reviews, as well as the metric for performance evaluation. Section 4 presents the data and results for the case studies. Section 5 provides discussion of the

[1] Corresponding author.

Contributed by the Design Theory and Methodology Committee of ASME for publication in the Journal of Mechanical Design. Manuscript received October 24, 2018; final manuscript received June 19, 2019; published online July 9, 2019. Assoc. Editor: Jitesh H. Panchal.

findings and limitations of the proposed methodology. Finally, the paper is concluded in Sec. 6.

## 2 Literature Review

This section presents three main topics related to the paper. It starts with discussing the discrete choice analysis and the role of choice set in it, followed by the natural language processing (NLP) tools to analyze customer reviews and finally presents the findings from the studies of online self-presentation.

**2.1 Discrete Choice Analysis.** Discrete choice analysis models people's choices among a set of alternatives, i.e., a choice set. It is developed based on the assumption that people act to maximize the utility. The utility of alternative $j$ for person $n$ ($U_{nj}$) is formulated as the sum of the observable ($V_{nj}$) and unobservable ($\varepsilon_{nj}$) parts. Logit model assumes that the unobservable part is independently and identically distributed as extreme value. By using the assumption, the formula for the probability of person $n$ choosing alternative $j$ ($P_{nj}$) takes a closed form as shown in Eq. (1) [6]. In the formula, the observable part of the utility is further defined as a linear combination of the alternative's attribute vector $\boldsymbol{x}_{ni}$ and the parameter vector β. These parameters are estimated by fitting the model to the training data.

$$P_{ni} = \frac{e^{\mathbf{V}_{ni}}}{\sum_j e^{\mathbf{V}_{nj}}} = \frac{e^{\beta' \mathbf{x}_{ni}}}{\sum_j e^{\beta' \mathbf{x}_{nj}}} \tag{1}$$

The denominator of Eq. (1) refers to all alternatives in the choice set. Thus, the choice probability is directly related to the alternatives included in the choice set. It is often the case that the number of possible alternatives is very large, and thus the choice set is constructed randomly [10]. The usage of random choice sets is relatively common in the literature. For example, it is used in the study of warehouse location choice [11], vehicle choice [7], neighborhood selection [12], the benefits of improved water quality in the fishing site [13], and product esthetics [14]. In the context of this paper, random choice set also represents a contrast to the proposed choice set, which is constructed based on the probability distribution that utilizes the information from the online data and customer reviews. Therefore, the choice sets that are constructed randomly become the appropriate baseline for the performance evaluation in this paper.

Nevertheless, there are researches suggesting the nonrandom underlying process of constructing choice set. Gensch [15] proposes a two-stage disaggregate attribute choice model. The model follows a two-stage choice paradigm [16], in which customers filter the set of all feasible alternatives to generate a choice set of few alternatives and closely compares the few alternatives to select one of them. The model requires a survey data, in which customers are asked to rate and rank attributes in each alternative. Wang and Chen [7] proposes a methodology to identify product communities from an existing choice set data (J.D. Power Vehicle Survey) using Newman's modularity method and to obtain customer segmentation from customer sociodemographic profiles using the K-means clustering method. The results are used to predict the missing choice sets in another data set of similar products (National Household Travel Survey). In contrast to the aforementioned literature, this paper proposes a method that does not require survey data of product attribute rating and ranking, existing customer choice set, and customer sociodemographic profiles. Alternatively, in order to construct customer choice sets, the proposed method utilizes product attribute descriptions and customer reviews from a product's webpage.

**2.2 Natural Language Processing Tools.** One of the basic ways to classify words is using part of speech. Part of speech are classes of words that have similar function with respect to the words that occur nearby or to the affixes they take [17]. In this paper, the classes that are used to analyze product reviews are noun and adjective. Nouns are used to identify product feature words, and adjectives are used to identify sentiment words.

When the words form a sentence, a dependency tree can describe the structure of it by relating words in terms of binary semantic or syntactic relations [17]. Therefore, each link in the tree explains the relation between two words. The advantage of using the tree over the bag-of-words approach, i.e., treating a sentence as a linear sequence of words, is its ability to describe relations between words, regardless of the distance between the words [18]. Therefore, as in Refs. [19,20], this paper uses the dependency tree to identify the related product feature and sentiment words in a review sentence.

In order to automatically identify product feature words from free-format reviews, a word embedding technique is applied in this paper. This technique aims to learn high-quality vector representations of words [21]. The objective of the learning model is quantified as maximizing the average log-likelihood of a sequence of training words $w_1, w_2, \ldots, w_T$ [22,23], as shown in Eq. (2). In Eq. (2), $T$ is the number of words in the training data set and $C$ is the window size that defines the context words surrounding word $w_t$. Given the objective function, the learning process is performed via neural networks. The detailed derivation of the learning updates to finally output the vector representations of words is provided in Ref. [23].

$$\frac{1}{T}\sum_{t=1}^{T} E_t = -\frac{1}{T}\sum_{t=1}^{T} \log P(w_t | w_{t-C}, \ldots, w_{t-1}, w_{t+1}, \ldots, w_{t+C}) \tag{2}$$

Finally, in order to identify the sentiment polarity of a sentiment word, i.e., whether a word contains positive, neutral, or negative sentiment, SenticNet 4 dictionary [24] is used in this paper. The dictionary is built by linking the words to their primitives, e.g., "eat" to "ingest." The generalization is claimed to boost the accuracy of SenticNet 4 compared with that of the previous version, as well as with that of the state-of-the-art statistical sentiment analysis research.

**2.3 Online Self-Presentation.** The emergence of Internet has attracted researchers to study people's self-presentation in the online world. In one of the earliest studies, online personal homepages in Yahoo are successfully classified into one of the five self-presentation strategies that people use in real interpersonal settings, i.e., ingratiation, competence, intimidation, exemplification, and supplication [25]. In addition, it suggests that gender differences in the real interpersonal settings are reflected in the online homepages.

A more recent study shows that Facebook usage and observable information on a person's Facebook page are associated with personality traits [26]. It implies that real-life personalities are extended into online domain. Similarly, another research identifies that the difference in personal information disclosure is related to the difference in age groups of users [27]. Moreover, the amount of information disclosure also reflects the relationship status of a person.

In terms of people's writings, the word usage in blogs is related to the writer's personality to an extent, e.g., extraversion personality is significantly correlated with the use of positive emotion words [28]. From the study of tweets in Twitter, both semantic and linguistic style features are discovered to be useful to predict personality and profession with high accuracy [29]. It concludes that not only what people say but also how to say it also reveals information about a person's personality and profession.

Although the aforementioned studies were not specifically conducted toward customers who write online reviews, there are evidences that online self-presentation represents a person's real characteristics. In conclusion, since the same personality traits and social processes expressed in real life are also expressed in the online world, online interactions have become an extension to people's social lives in the real world [30].

## 3 Methodology

The proposed methodology is summarized in Fig. 1. It consists of three main parts, i.e., clustering the products, clustering the customers, and finally constructing customer choice sets based on the clustering results.

The proposed methodology relies on online data and customer reviews to cluster product and customers. Therefore, the methodology works best when all products in the data set are generally feasible to be purchased by any customer, such that the clustering generates a feasible result as well. However, if there is a hidden constraint—which is not explicitly available on the online data and customer reviews, which strongly restricts a particular customer to a particular subset of products, then the clustering may generate an infeasible result. For example, a customer who would like to purchase an in-car DVD player is strongly restricted to choose from a particular subset of in-car DVD players that is physically and technically compatible with the customer's car. In the online data and customer reviews of the in-car DVD players, however, the compatibility information may not be present. Without revealing the constraint, the product clustering might cluster two in-car DVD players in a cluster despite the fact that they are compatible with different types of car.

**3.1 Clustering Products.** In contrast to the existing research that builds product communities based on actual choice set data [7], the proposed methodology clusters the products based on their attributes. The product attributes are acquired from publicly available sources such as the web pages of products in an e-commerce website. Based on the product attributes, X-means clustering is performed. X-means clustering automatically obtains the best number of clusters by maximizing the Bayesian information criterion iteratively [31]. It is advantageous compared with the methods that require the number of clusters as an input, such as K-means clustering, because the true number of groups of products is not always known.

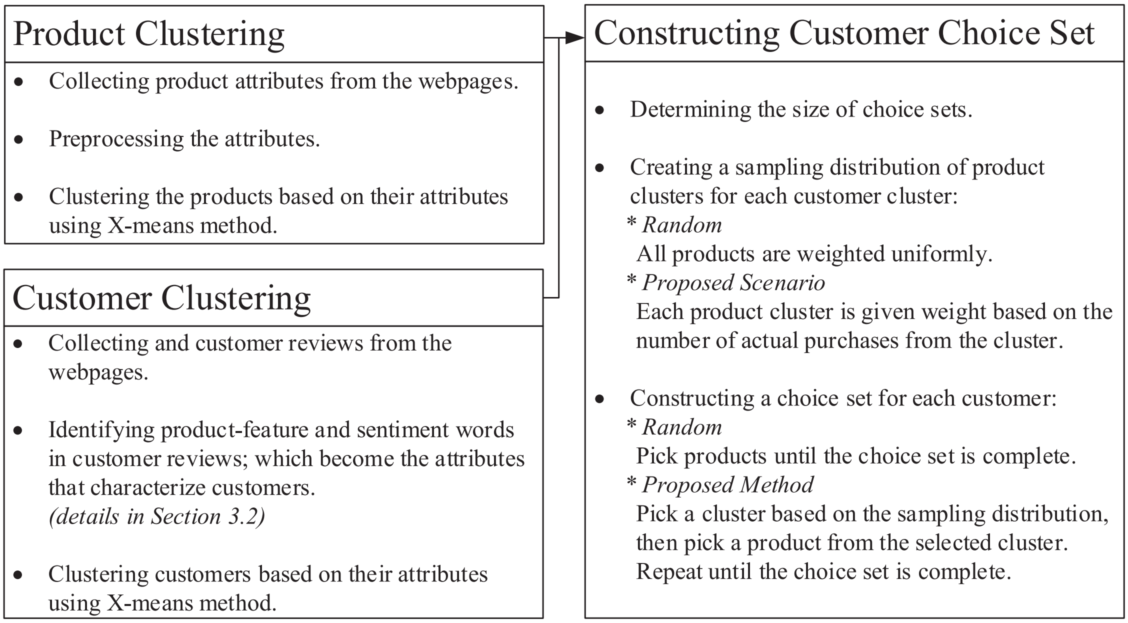
Compared with the product communities in Ref. [7], product clusters contain a less direct information about the actual customer choice sets. However, the information is proven valuable to construct customer choice sets. As demonstrated later in Sec. 4, the constructed choice sets create choice models that have higher predictive ability than the models that use randomly picked choice sets.

**3.2 Clustering Customers.** In contrast to the usage of socio-demographic data to cluster customers in Ref. [7], the proposed methodology utilizes online customer reviews to cluster customers based on the characteristics of their online self-presentations. More specifically, the customers are characterized by the product features that they discuss in the reviews, as well as the sentiment expressed toward those features, e.g., a group of customers who are satisfied with the laptop screen but dissatisfied with the laptop fan. Once each customer has been characterized by a vector that records the frequency of the customer mentioning each product feature word in the review, then all customers may be clustered using X-means clustering method.

There are four stages to identify product feature and sentiment words from customer reviews based on the methodology given in Ref. [20], as shown in Fig. 2. It is considered necessary to summarize each stage in this section, while the details are available in Ref. [20]. The first stage is preprocessing the review data. It involves cleaning the sentences from symbols, lemmatizing the sentences, parsing the sentences into dependency trees, and tagging each word in a sentence by its part of speech.

The second stage is automatically identifying and grouping product feature words that are discussed in the reviews. This stage is necessary because not all product features that are displayed in a product's web page are discussed in the reviews, and vice versa. Moreover, there are similar words that refer to the same product feature, e.g., "screen" and "display," such that they should be interpreted as the same product feature. In order to obtain the product feature words, a word embedding technique is used to embed the words into real vectors and X-means clustering is used to cluster the word vectors. In order to reflect a word's importance in the clustering process, each word is assigned a weight proportional to its tf.idf (term frequency, inverse document frequency). Based on the cluster centers, the words closest to each center become product feature word candidates. The word candidates with high similarity are combined into a single entity, e.g., "(web–Internet)."

At this point, further filtering is needed to remove irrelevant words from the product feature word candidates. The irrelevant words have a high tf.idf, yet are not related to the product itself, e.g., "son" or too specific for a particular brand, e.g., "ASUS." Those words are filtered out by a t-test that tests the words' average proportions in product manual documents. If the average proportions of those words are not significantly different from



Product Clustering
- Collecting product attributes from the webpages.
- Preprocessing the attributes.
- Clustering the products based on their attributes using X-means method.

Customer Clustering
- Collecting and customer reviews from the webpages.
- Identifying product-feature and sentiment words in customer reviews; which become the attributes that characterize customers. *(details in Section 3.2)*
- Clustering customers based on their attributes using X-means method.

Constructing Customer Choice Set
- Determining the size of choice sets.
- Creating a sampling distribution of product clusters for each customer cluster:
  * *Random*
  All products are weighted uniformly.
  * *Proposed Scenario*
  Each product cluster is given weight based on the number of actual purchases from the cluster.
- Constructing a choice set for each customer:
  * *Random*
  Pick products until the choice set is complete.
  * *Proposed Method*
  Pick a cluster based on the sampling distribution, then pick a product from the selected cluster. Repeat until the choice set is complete.

**Fig. 1 Proposed methodology for constructing customer choice sets**

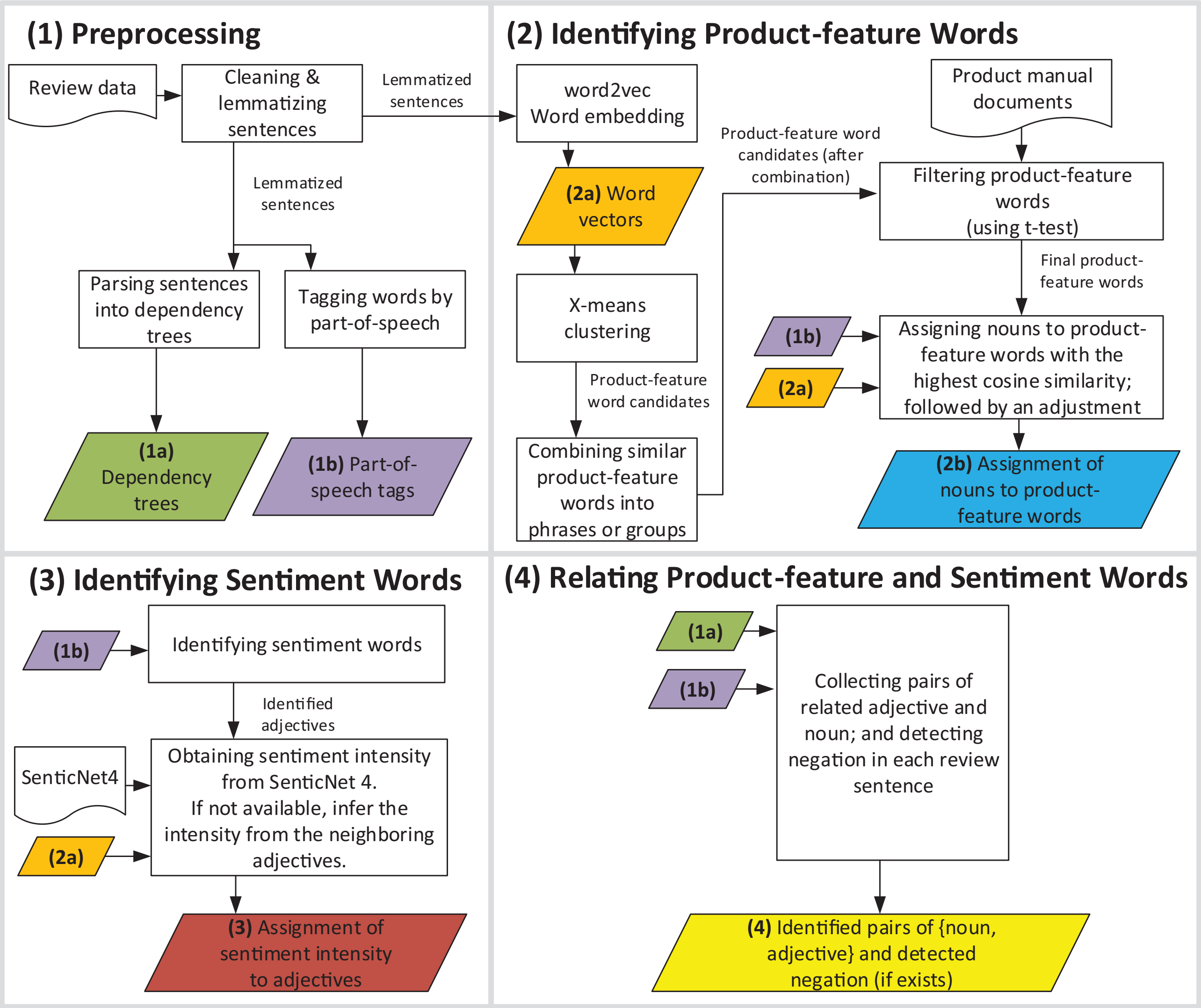

(1) Preprocessing
Review data
Cleaning & lemmatizing sentences
Lemmatized sentences
Lemmatized sentences
Parsing sentences into dependency trees
Tagging words by part-of-speech
(1a) Dependency trees
(1b) Part-of-speech tags
(2) Identifying Product-feature Words
word2vec Word embedding
(2a) Word vectors
X-means clustering
Product-feature word candidates
Combining similar product-feature words into phrases or groups
Product-feature word candidates (after combination)
Product manual documents
Filtering product-feature words (using t-test)
Final product-feature words
(1b)
(2a)
Assigning nouns to product-feature words with the highest cosine similarity; followed by an adjustment
(2b) Assignment of nouns to product-feature words
(3) Identifying Sentiment Words
(1b)
Identifying sentiment words
Identified adjectives
SenticNet4
(2a)
Obtaining sentiment intensity from SenticNet 4. If not available, infer the intensity from the neighboring adjectives.
(3) Assignment of sentiment intensity to adjectives
(4) Relating Product-feature and Sentiment Words
(1a)
(1b)
Collecting pairs of related adjective and noun; and detecting negation in each review sentence
(4) Identified pairs of {noun, adjective} and detected negation (if exists)

**Fig. 2 Proposed methodology for identifying product feature and sentiment words in customer reviews (Source: Ref. [20])**

zero, the words are considered irrelevant to the product and thus removed from the candidate list. In this research, α = 5% is used as the significance threshold. The remaining word candidates become the final product feature words. Finally, to group similar words that refer to the same product feature, all nouns are assigned to the product feature word that has the highest cosine similarity.

At the second stage, word embedding technique is chosen because it enables the quantification of distance between words, which is useful for grouping similar words. As for the clustering technique, X-means clustering is chosen because the number of product feature words that are discussed in the customer reviews is not known beforehand, unless the reviews have been manually annotated.

The third stage is identifying sentiment words in the customer review sentences. In this paper, the identification is done through a word's part-of-speech tag, i.e., an adjective is identified as a sentiment word. Afterward, the sentiment intensity of those words is obtained from a sentiment dictionary SenticNet 4 [24]. The intensity provides the polarity of a sentiment word, i.e., either positive or negative.

Finally, the last stage is relating the results from the previous stages, i.e., product feature (noun) and sentiment (adjective) words in a sentence. This stage is performed using a dependency tree approach because dependency tree may capture the related words regardless of the distance between them. It is advantageous compared with the adjacency-based approach, in which the relation is defined by a fixed window of adjacent words. A pair of adjective and noun is identified to have a relation if the noun is either the direct child or parent of the adjective. If an adjective has no nouns as the direct child or parent, it would move toward the root of the sentence. At each step of the move, it would collect the nouns that are now either its parent or child.

After the four stages are performed, each sentence in a customer review may be converted into a list of counts of product feature words and the corresponding sentiment polarities. The counts are then aggregated for all sentences in a customer review. As the result, each customer is now characterized by a list of counts and it becomes the basis to cluster customers using X-means clustering method, which is chosen because the true number of clusters of customers is not known beforehand.

**3.3 Constructing Customer Choice Set.** At this point, product clusters and customer clusters have been obtained from Secs. 3.1 and 3.2. Based on the clustering results, this section proposes the scenario to create a probability distribution for sampling the product clusters in order to construct customer choice sets. The reason for creating the probability distribution at cluster level is the absence of the actual choice set data, such that there is not enough confidence to build a probability distribution of products. Moreover, since the number of products is usually large, the

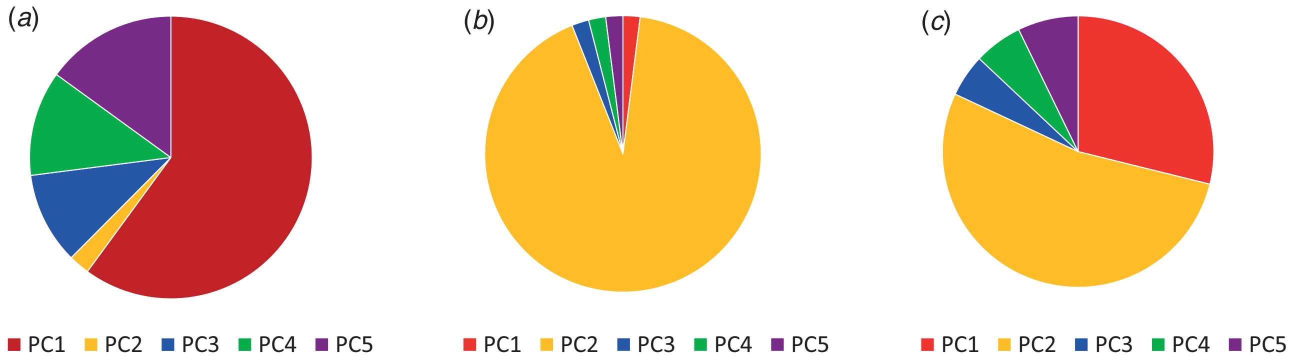


**Fig. 3 Illustration of creating the sampling probability based on the Normalized scenario: (*a*) number of products in product cluster, (*b*) number of purchases in customer cluster CC, and (c) normalized probability sampling for customers in CC**

inaccuracy of a probability distribution at product level is expected to be higher than it would be at cluster level.

The available data are the actual purchases made by customers and the product clusters. As illustrated in Fig. 3, the charts represent the size of each product cluster *PC* (Fig. 3(*a*)) and the number of purchases made by customers in a particular customer cluster *CC* (Fig. 3(*b*)). The information from both sources is combined to build the probability distribution of product clusters for each customer cluster.

The proposed scenario called Normalized assigns a probability value to a product cluster *PC* as a function of the product cluster size and the number of purchases from that product cluster, as defined in Eq. (3) and normalized using Eq. (4) such that the sum equals 1. The first term in Eq. (3) computes the proportion of products in product cluster *PC* ($I_{PC}$) to the total number of products in all $Q$ clusters. Similarly, the second term computes the proportion of purchased products in product cluster *PC* made by customers in customer cluster *CC* ($S_{PC,CC}$) to the total purchases of products in all $Q$ clusters made by customers in customer cluster *CC*. The multiplication of the two terms is denoted as $R_{CC}(PC)$, which is the unnormalized probability of a customer in customer cluster *CC* to choose a product from product cluster *PC*. In Eq. (4), the normalization results in $P_{CC}(PC)$, i.e., the probability of a customer in customer cluster *CC* to choose a product from product cluster *PC* to be included in the choice set, which is illustrated in Fig. 3(*c*). The performance of Normalized scenario is compared with Random scenario as the baseline. In Random scenario, a choice set is constructed by picking a set of items randomly.

$$R_{CC}(PC) = \frac{I_{PC}}{\sum_{\forall Q} I_Q} \cdot \frac{S_{PC,CC}}{\sum_{\forall Q} S_{Q,CC}} \tag{3}$$

$$P_{CC}(PC) = \frac{R_{CC}(PC)}{\sum_{\forall Q} R_{CC}(Q)} \tag{4}$$

Once the choice sets have been constructed for all customers, they become the inputs for the multinomial logit model. As shown in Eq. (1), each alternative $j$ in a customer's choice set contributes to the denominator of the choice probability formula. The contribution of each alternative is proportional to its utility. In order to define an alternative's utility, there are two functions used in this paper. The first function, shown in Eq. (5), defines the utility of alternative $j$ for customer $n$ ($V_{nj}$) as a linear combination of its attributes, i.e., the multiplication of the value of product attribute $k$ of alternative $j$ ($x_{jk}$) and the corresponding logit model parameter for product attribute $k$ ($\beta_k$).

$$V_{nj} = \sum_{k \in K} \beta_k x_{jk} \tag{5}$$

The second function, shown in Eq. (6), defines the utility of alternative $j$ for customer $n$ ($V_{nj}$) by adding an interaction term to the first function. The interaction involves a set of product attributes $K^{Rev}$ that are discussed in customer reviews. It is defined as the multiplication between product attribute $k' \in K^{Rev}$ of alternative $j$ ($x_{jk'}$) and its frequency of being discussed by customer $n$ in the review ($y_{njk'}$) either positively or negatively. Accordingly, the corresponding logit model parameter for the interaction term related to product attribute $k'$ is denoted as $\beta_{k'}^{Rev}$.

$$V_{nj} = \sum_{k \in K} \beta_k x_{jk} + \sum_{k' \in K^{Rev}} \beta_{k'}^{Rev} x_{jk'} y_{njk'} \tag{6}$$

### 3.4 Performance Evaluation.

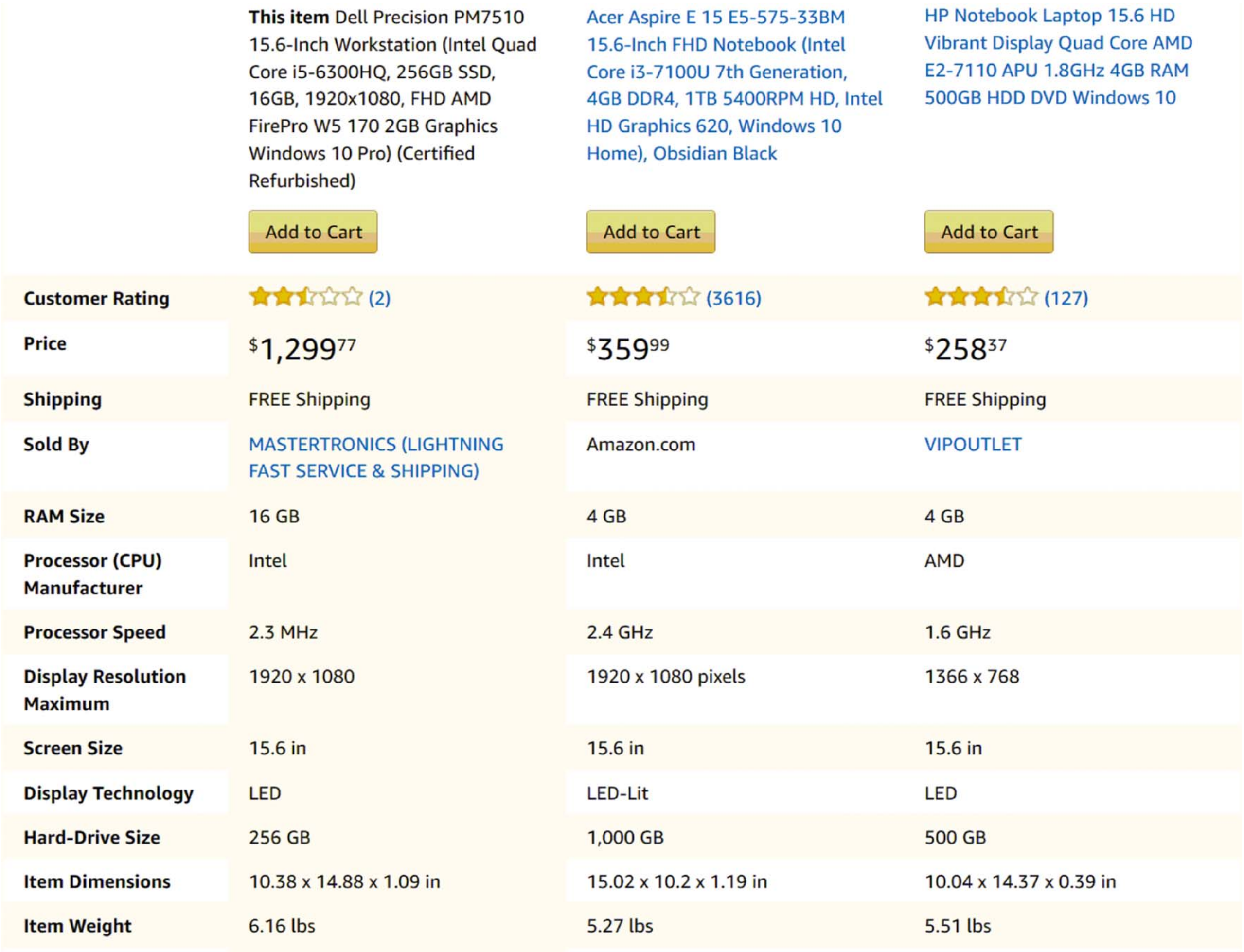
At this point, the choice sets have been constructed and the utility model has been defined. In order to evaluate the performance of different scenarios in constructing customer choice sets, a data set is divided into a training set and a test set. The training set is used to train the multinomial logit model that provides the estimates of the β parameters in the utility function by maximizing the likelihood of the training set. The estimates of the β parameters are subsequently applied to predict the probability of purchases in the test set. In the test set, the choice set for each customer contains all items that have been purchased by customers in both the training and test sets. Therefore, since it is different with the choice sets from either Random or Normalized scenarios, the test set becomes a fair assessment of the predictive ability of the scenario that is used in the training set.

The predictive ability is measured at the aggregate level, i.e., the market shares of products, instead of calculating the percentage of individual customers whose purchases are correctly predicted. In order to compare the predicted and actual probability distributions, Kullback–Leibler (*KL*) divergence in Eq. (7) is chosen as the metric. Kullback–Leibler (KL) divergence measures the difference between two distributions over the same event space [32], such that the higher KL divergence indicates more different distributions. The actual distribution may be represented by a vector of zeros for all items, except for item $j$ that customer $n$ purchases ($P_{nj}$) that has a value of 1. The prediction on the test set provides the probability of customer $n$ purchasing item $j$ ($Q_{nj}$). A good performance is indicated by the distribution of $Q$ being similar to $P$ and quantified by a low *KL* value.

$$KL = \sum_j P_{nj} \log \frac{P_{nj}}{Q_{nj}} \tag{7}$$

## 4 Data and Results

In this section, the implementation of the proposed methodology is presented. A data set of laptop products is collected from the website Amazon.com. The data set contains the attributes of 2631 laptops, which are utilized for clustering products. The data set also contains 46,194 verified reviews from customers who



| | This item Dell Precision PM7510 15.6-Inch Workstation (Intel Quad Core i5-6300HQ, 256GB SSD, 16GB, 1920x1080, FHD AMD FirePro W5 170 2GB Graphics Windows 10 Pro) (Certified Refurbished) | Acer Aspire E 15 E5-575-33BM 15.6-Inch FHD Notebook (Intel Core i3-7100U 7th Generation, 4GB DDR4, 1TB 5400RPM HD, Intel HD Graphics 620, Windows 10 Home), Obsidian Black | HP Notebook Laptop 15.6 HD Vibrant Display Quad Core AMD E2-7110 APU 1.8GHz 4GB RAM 500GB HDD DVD Windows 10 |
|---|---|---|---|
| | Add to Cart | Add to Cart | Add to Cart |
| **Customer Rating** | ★★½☆☆ (2) | ★★★½☆ (3616) | ★★★½☆ (127) |
| **Price** | $1,299.77 | $359.99 | $258.37 |
| **Shipping** | FREE Shipping | FREE Shipping | FREE Shipping |
| **Sold By** | MASTERTRONICS (LIGHTNING FAST SERVICE & SHIPPING) | Amazon.com | VIPOUTLET |
| **RAM Size** | 16 GB | 4 GB | 4 GB |
| **Processor (CPU) Manufacturer** | Intel | Intel | AMD |
| **Processor Speed** | 2.3 MHz | 2.4 GHz | 1.6 GHz |
| **Display Resolution Maximum** | 1920 x 1080 | 1920 x 1080 pixels | 1366 x 768 |
| **Screen Size** | 15.6 in | 15.6 in | 15.6 in |
| **Display Technology** | LED | LED-Lit | LED |
| **Hard-Drive Size** | 256 GB | 1,000 GB | 500 GB |
| **Item Dimensions** | 10.38 x 14.88 x 1.09 in | 15.02 x 10.2 x 1.19 in | 10.04 x 14.37 x 0.39 in |
| **Item Weight** | 6.16 lbs | 5.27 lbs | 5.51 lbs |

**Fig. 4 Snapshot of a similar item section**

purchased 84 different laptops. The reviews were posted between January 2015 and February 2017, and they are used for clustering customers. In constructing customer choice sets, the customer reviews of products of which the product attributes are inaccessible are excluded. Therefore, the proposed methodology is implemented to a data set of 39,000 customers and 62 products.

At the preprocessing stage for the customer review data, the lemmatizer from NLTK package [33] in PYTHON is used to lemmatize the sentences. The Stanford parser from NLTK package and PYSTANFORD-DEPENDENCIES package [34] in PYTHON are used to parse each sentence into a dependency tree, as well as tagging each word with its part of speech.

**4.1 Product Attributes Data and Product Clustering Result.** A product's attributes are collected from its Amazon webpage. For laptops, there is a section that compares similar laptops and lists their attributes, as shown in Fig. 4.[2] The product attribute information may also be obtained from a product's title and item description section. The attributes are preprocessed such that the unit within an attribute is consistent, e.g., all values in the processor speed attribute are converted to have a GHz unit. However, the value itself remains as it is, e.g., the processor speed of 2.3 MHz is converted into 0.0023 GHz because it is the information displayed and thus received by customers.

The product attributes are used to cluster the products. X-means clustering method is used for the purpose, and it is implemented via PYCLUSTERING package [35] in PYTHON. There are 25 product clusters obtained, and the top 8 clusters with the highest number of products are shown as the representatives in Table 1, with their corresponding center points. As expected, it shows that the more expensive laptops generally have higher specifications, as well as being physically larger and heavier.

**4.2 Customer Review Data and Customer Clustering Result.** Verified customer reviews are verified by Amazon as being written by customers who have purchased the product. The verification provides the information of the actual purchase made by a particular reviewer. Therefore, for the purpose of this paper, only verified customer reviews are considered. An example of such review is shown in Fig. 5.[3] The sentence is parsed into a dependency tree, as shown in Fig. 6.

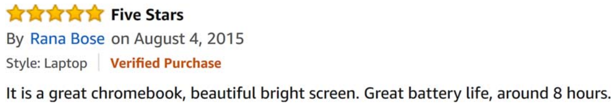
Product feature words are obtained by applying the word embedding GENSIM package [36] in PYTHON to obtain the vector representations of the words, then followed by X-means clustering to cluster the vectors. The words closest to the cluster centers are determined as the initial product feature words. After filtering and grouping similar words, the final product feature words are shown in Table 2. The result is obtained by setting the Word2vec parameters as follows: the dimension of the word embedding vector is 100, the window size is 2, the cutoff frequency is 8, hierarchical softmax is used, and the initial random seed is 0.

[2] https://www.amazon.com/dp/B01LZUPUG2

[3] https://www.amazon.com/gp/customer-reviews/R2LEZTBHDUVOZG/ie=UTF8&#x0026;ASIN=B00N99FXIS

Table 1 Center points of product clusters (laptop data set) with the largest number of products, sorted by price

| Product attribute | PC15 | PC22 | PC16 | PC14 | PC0 | PC2 | PC1 | PC6 |
|---|---|---|---|---|---|---|---|---|
| Price ($) | 1791.69 | 1425.62 | 886.62 | 821.87 | 811.75 | 506.98 | 489.02 | 288.36 |
| Processor speed (PS) (GHz) | 2.82 | 2.50 | 2.31 | 2.52 | 1.77 | 2.09 | 2.62 | 1.65 |
| Processor count (PC) | 3.30 | 2.59 | 2.02 | 2.43 | 2.23 | 1.95 | 2.32 | 2.15 |
| Memory (GB) | 20.84 | 14.71 | 9.33 | 10.45 | 9.62 | 5.54 | 8.21 | 3.39 |
| Hard disk (HD) (GB) | 613.76 | 603.42 | 301.82 | 711.63 | 506.68 | 241.49 | 746.44 | 111.65 |
| Ratio (Megapixel/in.) | 0.1333 | 0.4931 | 0.1531 | 0.1328 | 0.0001 | 0.0748 | 0.0672 | 0.0904 |
| Screen size (SS) (in.) | 15.55 | 14.82 | 13.56 | 15.58 | 14.47 | 14.01 | 15.53 | 11.60 |
| Volume (in.$^3$) | 187.53 | 204.94 | 142.28 | 173.44 | 181.54 | 228.75 | 193.37 | 135.26 |
| Weight (lb) | 4.70 | 4.35 | 3.20 | 4.81 | 3.62 | 3.87 | 21.48 | 2.64 |
| Operating system (OS) (1 = Windows) | 1.00 | 0.96 | 0.94 | 0.96 | 0.83 | 0.90 | 0.95 | 0.60 |
| Number of products | 109 | 138 | 260 | 309 | 684 | 126 | 391 | 164 |



★★★★★ Five Stars

By Rana Bose on August 4, 2015

Style: Laptop | Verified Purchase

It is a great chromebook, beautiful bright screen. Great battery life, around 8 hours.

Fig. 5 A customer review

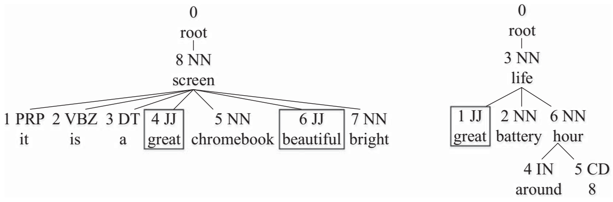


Fig. 6 The dependency trees of preprocessed sentences in the customer review shown in Fig. 5

In Fig. 6, there are three pairs of adjective and noun identified from the laptop review example, i.e., “great screen,” “beautiful screen,” and “great life.” The word “screen” corresponds to product feature word “screen-display,” and based on SenticNet 4, the polarity of “great” is positive; therefore, the first pair is translated into “(screen-display)+.” The second pair is translated into “(screen-display)+” as well due to the positive polarity of “beautiful.” Overall, the review in Fig. 5 can be converted into a list of counts: “(screen-display)+” = 2, “life+” = 1, and all the remaining pairs are 0.

Based on the counts of product feature words and the rating assigned to the reviews, customers may be clustered using X-means clustering method. In this case study, each customer is represented by a vector of 37 integers, i.e., 18 product feature words paired with both positive and negative sentiments and 1 customer rating. The clustering results in 30 clusters. The number of customers in each cluster as well as each cluster center’s rating value are shown in Fig. 7. The figure shows that the customer clusters capture the differences among customers, at least based on the cluster’s average rating.

The characteristics of each cluster may be analyzed further through the cluster’s center. Since cluster 14 in Fig. 7 has the highest number of customers, the cluster’s characteristics are analyzed here. The center of cluster 14 is a vector of size 37. Excluding the rating, the remaining 36 values of cluster 14’s center are plotted in Fig. 8, divided into 18 positive attributes on the left graph and 18 negative attributes on the right graph. The *Y*-axis of the graphs corresponds to the frequency of a product feature word and sentiment pair. The center of cluster 14 is compared with the average of all other clusters’ centers, as well as the average of the centers of all other clusters that have ratings of 4 and 5.

It can be observed from Fig. 8 that customers in cluster 14, whose average rating is 4.73, are generally satisfied customers who write reviews without frequently expressing explicit sentiment towards any product feature in particular. In contrast, compared with the overall average, customers who assigns rating 4 and 5 (excluding cluster 14) tend to specifically and frequently mention the product features along with their positive or negative sentiment toward them. The examples of original customer reviews, which do not specify explicit sentiment toward any product features, from customers in cluster 14 are shown in the first row of Table 3.[4,5] In contrast, the sentences from customers in cluster 0, which explicitly express negative sentiments toward disk space, computer, and battery life, are shown in the second row of Table 3.[6,7]

**4.3 Constructed Customer Choice Set Result.** In this paper, the choice sets are constructed with the choice set sizes of 3, 5, and 7. The numbers are chosen to be relatively small, referring to the previous research that constructs a choice set consisting of one purchased item and three predicted items [7], as well as the statement that the average choice set size is between 2 and 8 [16]. The varied choice sets are implemented to examine whether there are differences in the proposed methodology’s performance.

The first product in a choice set is the actual purchase by the customer, which is known from the customer’s verified review. The other products to complete the choice set are picked based on either Random or Normalized scenario, with no duplications allowed. An example of the constructed choice set for a customer is shown in Table 4. The first column indicates whether an item is purchased. The second column shows a product’s name, and the product’s attributes are shown in the following columns. After constructing choice sets for all customers, the utility of a product for a person may be computed using Eq. (5). In the formula, K is

Table 2 Product feature words obtained from the reviews

| Data set | Product feature words |
|---|---|
| Laptops (18 product feature words) | Apps, battery, cable, card, drive, fan, issue, laptop, life, network, office, performance, resolution quality, screen display, service, supervisor, track mouse, web–Internet |

[4] http://www.amazon.com/gp/customer-reviews/RE8QQH55NKO92/ref=cm_cr_arp_d_rvw_ttl?ie=UTF8&ASIN=B00MNOPS1C

[5] http://www.amazon.com/gp/customer-reviews/R31N88N7MGDRKY/ref=cm_cr_arp_d_rvw_ttl?ie=UTF8&ASIN=B00L49X8E6

[6] http://www.amazon.com/gp/customer-reviews/R3BAIKVU0E5TA3/ref=cm_cr_arp_d_rvw_ttl?ie=UTF8&ASIN=B00NSHLUBU

[7] http://www.amazon.com/gp/customer-reviews/R1CCJIS37WSSAX/ref=cm_cr_arp_d_rvw_ttl?ie=UTF8&ASIN=B00NGK98GS

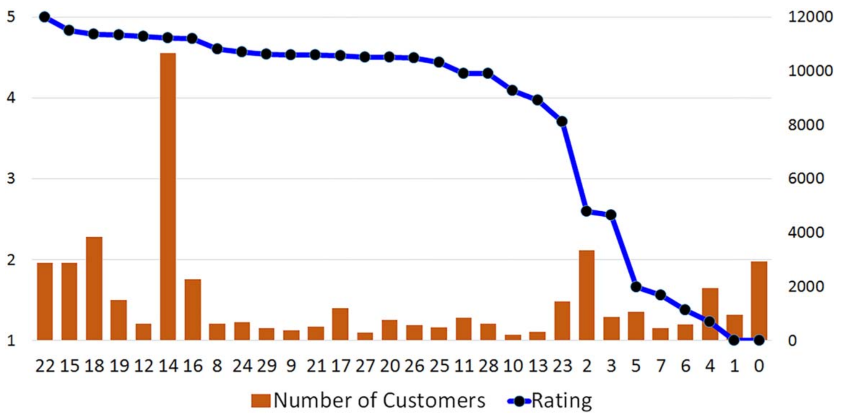


Fig. 7 Snapshot of customer clusters

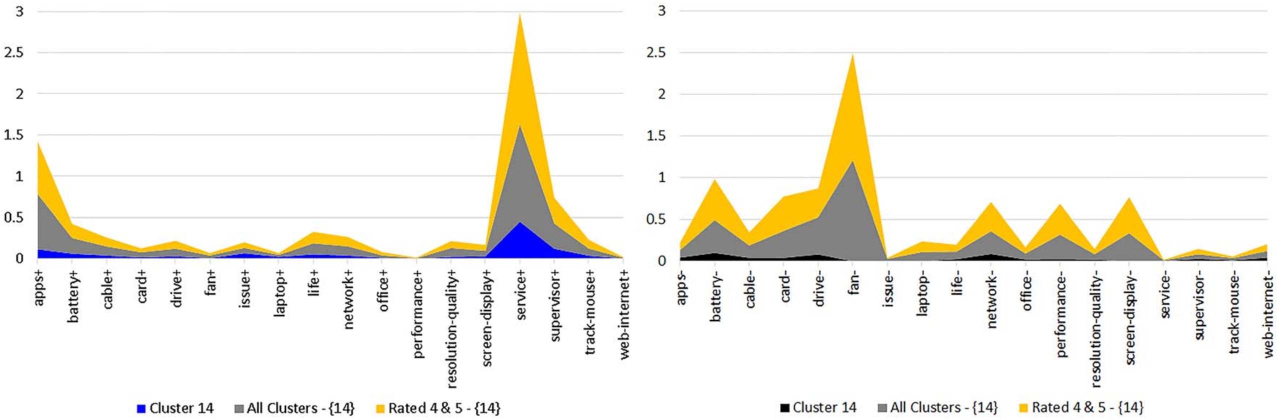


Fig. 8 Comparison between cluster 14 with the remaining clusters and with the clusters having 4 and 5 ratings, for positive (left graph) and negative (right graph) sentiments

Table 3 Comparison between selected sentences from customers in cluster 14 and cluster 0

| CC | Sentences |
|---|---|
| 14 | "i have being using it since arrival, the acer has not disappointed me and i am glad i sold my sell phone to help buy this"[a]<br>"this was purchased for our child in 7th grade, she is very pleased with it, it suits her purpose for school and recreational activities"[b] |
| 0 | "very *little* disk **space**, do not buy this laptop, absolutely *terrible* on **space**, not good for saving school work either"[c]<br>"it is a *slow* running **computer** with a *short* battery **life**"[d] |

[a] See Note 4.
[b] See Note 5.
[c] See Note 6.
[d] See Note 7.

the set of product attributes, and there are ten product attributes of laptops as listed in Table 4.

For computing the second utility function, defined in Eq. (6), each product attribute is matched with a product feature word from the reviews, based on the highest cosine similarity. If a match is found, then the product attribute is included into the set $K^{Rev}$. For example, the word "memory" (one of the product attributes) has the highest similarity with the product feature word "drive." Therefore $k' =$ "memory" is included in $K^{Rev}$. The value of $x_{jk'}$ is the memory (GB) of laptop $j$, and the value of $y_{njk'}$ is the frequency of customer $n$ discussing "memory" in the review, which is represented by the sum of the frequencies of "drive+" and "drive−." In the case that a product attribute does not match with any of the product feature words, then the attribute is excluded from $K^{Rev}$. The matching between product attributes and product feature words is summarized in Table 5.

Table 4 Example of a customer's constructed choice set

| Choice | Product | Price | PS | PC | Memory | HD | Ratio | SS | Volume | Weight | OS |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Yes | B00N99FXIS | 719.57 | 2.16 | 2 | 4 | 0 | 0.1559 | 13.3 | 80.44 | 2.95 | 0 |
| No | B071XSKHWV | 399.00 | 2.40 | 2 | 6 | 1000 | 0.0672 | 15.6 | 136.78 | 5.20 | 1 |
| No | B015P3SSD2 | 989.99 | 2.60 | 4 | 8 | 1000 | 0.1198 | 17.3 | 262.61 | 8.33 | 1 |
| No | B073R41NPW | 2099.99 | 2.80 | 4 | 32 | 1240 | 0 | 17.3 | 207.22 | 6.17 | 1 |
| No | B06WVGCQ8H | 719.00 | 2.50 | 2 | 12 | 1000 | 0 | 15.6 | 135.00 | 4.80 | 1 |

**Table 5 Product feature words with the highest cosine similarity to the product attribute words**

| Product attribute | Product feature word |
|---|---|
| Price | Performance |
| Processor speed (PS) | Performance |
| Processor count (PC) | Performance |
| Memory | Drive |
| Hard disk (HD) | Drive |
| Ratio | Resolution quality |
| Screen size (SS) | Resolution quality |
| Volume | Laptop |
| Weight | Laptop |
| Operating system (OS) | Apps |

**4.4 Performance Evaluation Result.** There are two sets of experiments presented in this subsection. The first set of experiments is used to compare different sampling probability scenarios, i.e., Random and Normalized, with the utility function that only considers product attributes, as shown in Eq. (5). In the Normalized scenario, the sampling procedure can be done with or without replacement. The with-replacement procedure means that a selected product cluster is returned to the sampling pool, such that it has a chance to be selected again. In both procedures, once a product cluster has been selected, an individual product is subsequently selected randomly from the selected product cluster.

Since probability sampling is involved in constructing customer choice sets, in order to avoid bias due to the random numbers, different starting random seeds are used to construct the choice sets for all customers. In the experiment, ten starting random seeds are used to create choice set data sets. Each data set is further divided into smaller data sets randomly. In the experiment, each data set is further divided into 10 smaller data sets containing 3900 customers each, such that finally there are 100 smaller data sets. Each of the smaller data sets of size 3900 becomes the input for training the multinomial logit model, which is implemented via PYLOGIT package [37] in PYTHON. The output of the multinomial logit model is a set of coefficients, which are the estimates of the β parameters in Eq. (5). The coefficients are subsequently applied to the test data set of size 35,100 to evaluate the predictive ability of the model.

The process is illustrated in Table 6. The table contains all items that are purchased by customers in the entire data set along with the values of their attributes, e.g., processor speed (PS), operating system (OS). The *Purchase* column contains the number of purchases of each item, while the *Purchase (test)* column excludes the purchases in the data that are used to train the multinomial logit model. Based on the *Purchase (test)* column, the fraction in the *Fraction (test)* column may be computed and thus represent the actual market share in the test set. Based on the utility function in Eq. (5), the utility may be computed for each item, as shown in the *Utility* column. Finally, the predicted probability of purchasing each item may be computed using Eq. (1), which may as well be interpreted as the predicted market share, as shown in the last column. The performance metric, KL divergence in Eq. (7), may then be computed from the columns *Fraction (test)* and *Predicted fraction*. The computation of KL divergence in this paper is implemented via SPACY package in PYTHON [38].

The performance comparison between choice models that use different choice set construction scenarios is presented in Table 7. Based on the average of KL divergence values in 100 samples, the Normalized scenario without replacement procedure is significantly better ($p$-value = 0.000) than the baseline, i.e., Random scenario, for choice set sizes of 5 and 7. The Normalized scenario with replacement procedure, however, is significantly worse than the baseline for all choice set sizes. Therefore, for the second set of experiments, the Normalized scenario with replacement procedure is excluded.

The second set of experiments is used to compare different sampling probability scenarios, i.e., Random and Normalized, with the utility function that includes the interaction between product attributes and the frequencies of the attributes being discussed in the customer review, as shown in Eq. (6). The process is illustrated in Table 8. The first difference with the illustration in Table 6 is the inclusion of the frequency of the product feature word that is related to a product attribute. For example, Table 5 shows that the attribute PS matches with the product feature word "performance." Customer $n1$ discusses it once in the review, while customer $n2$ does not discuss it at all; hence, the numbers 1 and 0 shown in the columns "*perf*"$_{n1}$ and "*perf*"$_{n2}$. These individual differences cause the utility of each item to differ for each individual, as illustrated by the columns $Utility_{n1}$ and $Utility_{n2}$.

The KL divergence may be computed for an individual by setting $P_{nj}$ in Eq. (7) equals 1 for item $j$ that is purchased by the individual and 0 for all other items. The total KL divergence of the test set is obtained by summing the KL divergence over all individuals. The comparison between scenarios are shown in Table 9. Similar to the result given in Table 7, the Normalized scenario is significantly better ($p$-value = 0.000) than the baseline, i.e., Random scenario, for choice set sizes of 5 and 7.

The estimates of β parameters for Eq. (6) that are obtained from the Random and Normalized scenarios are shown in Table 10. Those are the coefficients from the data sets of size 3900 that provides the best (lowest) KL divergence values, i.e., 143,107 (Random scenario with choice set size of 7) and 140,157 (Normalized scenario with choice set size of 7). There is no dramatic difference between scenarios, as the signs of the coefficients of significant variables ($p$-value < 0.05) are the same for both scenarios. The variable *Operating Systems* is significant in the Random scenario, but not in the Normalized. A possible explanation is that the Normalized scenario reflects the fact that customers have filtered out the laptops with different operating systems. Therefore, it is no longer significant to predict their choices. The variables such as *Memory*, *Hard Disk*, *Price*, and *Processor Count* have negative coefficients, which means that the increase of these variables is related to the decrease of the probability of being purchased. On the other hand, the increase of *Screen Size* and *Ratio* variables is related to the increase of the purchase probability. In order to address the issue of possible multicollinearity, it is assumed that the input values for all variables in the model are reasonable, e.g., a higher processor speed comes with a higher price as well, such that the choice model outputs the correct choice probability. Furthermore, the coefficients given in Table 10 should be used altogether to predict the choice probability in Eq. (1), instead of being interpreted individually.

**Table 6 Illustration of comparing true and predicted distributions of purchased item in the test set**

| No. | Item | PS | … | OS | Purchase | Purchase (test) | Fraction (test) | Utility | Predicted fraction |
|---|---|---|---|---|---|---|---|---|---|
| 0 | B00O65HZKS | 2.16 | … | 1 | 3985 | 3607 | 0.10276353 | 2.56388858 | 0.03111692 |
| 1 | B00NSHLTVG | 2.16 | … | 1 | 3985 | 3587 | 0.10219373 | 2.62021862 | 0.03292004 |
| 2 | B00O65HZIK | 2.16 | … | 1 | 3982 | 3560 | 0.10142450 | 2.94748784 | 0.04566585 |
| 3 | B00NSHLUBU | 2.16 | … | 1 | 3981 | 3569 | 0.10168091 | 2.46594490 | 0.02821371 |
| … | … | … | … | … | … | … | … | … | … |
|  |  |  |  |  | Σ = 39,000 | Σ = 35,100 | Σ = 1 |  | Σ = 1 |

**Table 7 K-L divergence summary of experiments with different choice set construction scenarios**

| Choice set size | Scenario | $N$ | Mean | SD | SE mean |
|---|---|---|---|---|---|
| 3 | Random | 100 | 0.66270 | 0.01660 | 0.00170 |
| | Normalized | 100 | 0.75360 | 0.02830 | 0.00280 |
| | Normalized-replaced | 100 | 0.80930 | 0.02040 | 0.00200 |
| 5 | Random | 100 | 0.65110 | 0.01520 | 0.00150 |
| | Normalized | 100 | 0.60730 | 0.01010 | 0.00100 |
| | Normalized-replaced | 100 | 0.73990 | 0.01260 | 0.00130 |
| 7 | Random | 100 | 0.64870 | 0.01540 | 0.00150 |
| | Normalized | 100 | 0.56178 | 0.00842 | 0.00084 |
| | Normalized-replaced | 100 | 0.70740 | 0.01590 | 0.00160 |

**Table 8 Illustration of the difference in individual utility values toward an item due to the inclusion of the interaction terms in the utility function**

| No. | Item | PS | … | “perf”$_{n1}$ | “perf”$_{n2}$ | Purchase | Purchase (test) | Fraction (test) | Utility$_{n1}$ | Utility$_{n2}$ |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | B00O65HZKS | 2.16 | … | 1 | 0 | 3985 | 3607 | 0.102763533 | $V_{n1,1}$ | $V_{n2,1}$ |
| 1 | B00NSHLTVG | 2.16 | … | 1 | 0 | 3985 | 3587 | 0.102193732 | $V_{n1,2}$ | $V_{n2,2}$ |
| 2 | B00O65HZIK | 2.16 | … | 1 | 0 | 3982 | 3560 | 0.101424501 | $V_{n1,3}$ | $V_{n2,3}$ |
| 3 | B00NSHLUBU | 2.16 | … | 1 | 0 | 3981 | 3569 | 0.101680912 | $V_{n1,4}$ | $V_{n2,4}$ |
| … | … | … | … | … | … | … | … | … | … | … |

Finally, a comparison is made between the best results from different utility functions, i.e., with or without interaction terms included in the function, as shown in Table 11. The KL divergence value for the Normalized scenario in Table 7 is converted by computing individual KL divergence first and then summing over all individual, such that the value becomes directly comparable with the Normalized scenario in Table 9. The comparison shows that the inclusion of the interaction terms results in a significantly lower (better) ($p$-value = 0.009) KL divergence.

## 5 Discussion

The proposed Normalized scenario shown in Fig. 3 is developed based on two types of information, i.e., the product clusters and the number of purchases within a customer cluster. The multiplication in Eq. (3) represents the combining of product and customer information. In the Normalized scenario, a product cluster $PC$ obtains a high probability only if it contains many products and customers in $CC$ purchase many products that belong to $PC$. It follows the assumptions that (1) when there is no additional information, a bigger cluster has a higher probability to be picked and (2) a cluster of products that is frequently purchased by similar type of customers has a higher probability to be included in those customers' choice sets. The second assumption is parallel to the idea of the neighborhood method in a recommender system [39]. The method may be used to, for example, recommend a movie to a person based on a set of movies that is highly rated by people who like similar types of movies.

Based on the comparison between Random and Normalized scenarios in Table 7, the Normalized scenario without replacement procedure benefits from using the information, i.e., achieving a significantly better predictive ability than the baseline Random scenario, which is indicated by the lower (better) KL divergence values for choice set sizes of 5 and 7. Furthermore, the information is proven valuable because if the information were worthless, then the KL divergence value would not be significantly different with using no information, which is appropriately represented by the Random scenario.

In both sets of experiments in Sec. 4, the Normalized scenario with the smallest choice set size, i.e., 3, performs worse than the Random scenario. It may be explained that the small choice set size prevents the training set from having enough variation in the selection of items for the choice sets. The small choice set size focuses the selection from the cluster with big probability portions, e.g., PC2 and PC1 in Fig. 3(*c*). The test set, however, requires the model to face a highly varied items, because the choice sets include all items in the data set. As the choice set size grows larger, e.g., 5 or 7, it allows the training set to construct choice set by focusing on the clusters with high probabilities, as well as having the opportunity to pick items from clusters with lower probabilities. As the result, the training set has a higher predictive ability on the test set. The similar explanation may be applied to the fact that the replacement procedure results in a significantly worse performance for all choice set sizes, as listed in Table 7. The replacement procedure allows a cluster to be chosen repeatedly during the sampling process. Therefore, a cluster with a high probability in the distribution is likely to dominate the constructed choice set in

**Table 9 K-L divergence summary of experiments with different choice set construction scenarios using utility function that includes interaction terms**

| Choice set size | Scenario | $N$ | Mean | SD | SE mean |
|---|---|---|---|---|---|
| 3 | Random | 100 | 144,546 | 791 | 79 |
| | Normalized | 100 | 147,598 | 1088 | 109 |
| 5 | Random | 100 | 144,035 | 434 | 43 |
| | Normalized | 100 | 142,391 | 356 | 36 |
| 7 | Random | 100 | 143,825 | 505 | 51 |
| | Normalized | 100 | 140,754 | 317 | 32 |

Table 10 Comparison of choice model coefficient estimates between Random and Normalized scenarios

| Variable | Coefficient (Random) | SE (Random) | $p$-Value (Random) | | Coefficient (Normalized) | SE (Normalized) | $p$-Value (Normalized) | |
|---|---|---|---|---|---|---|---|---|
| Processor speed (PS) | 0.0034311 | 0.0060412 | 5.70E–01 | | 0.0084258 | 0.0061247 | 1.69E–01 | |
| Memory | −0.1099155 | 0.0089848 | 2.06E–34 | * | −0.1663226 | 0.0105410 | 4.37E–56 | * |
| Ratio | 0.0000005 | 0.0000001 | 5.34E–05 | * | 0.0000071 | 0.0000003 | 1.83E–121 | * |
| Hard disk (HD) | −0.0026165 | 0.0001005 | 2.19E–149 | * | −0.0019658 | 0.0000940 | 4.84E–97 | * |
| Volume | −0.0000034 | 0.0000079 | 6.73E–01 | | 0.0000693 | 0.0000838 | 4.09E–01 | |
| Weight | 0.0000396 | 0.0005264 | 9.40E–01 | | 0.0000071 | 0.0006401 | 9.91E–01 | |
| Price | −0.0012553 | 0.0000800 | 1.69E–55 | * | −0.0014482 | 0.0000876 | 2.29E–61 | * |
| Processor count (PC) | −0.2334449 | 0.0230215 | 3.66E–24 | * | −0.2034307 | 0.0225258 | 1.70E–19 | * |
| Screen size (SS) | 0.0766159 | 0.0133647 | 9.88E–09 | * | 0.2428041 | 0.0151927 | 1.72E–57 | * |
| Operating system (OS) | −0.3529272 | 0.0527470 | 2.22E–11 | * | −0.0086996 | 0.0464558 | 8.51E–01 | |
| PS*“performance” | 0.1088088 | 0.0429116 | 1.12E–02 | * | 0.0078080 | 0.0132741 | 5.56E–01 | |
| Memory*“drive” | −0.1146116 | 0.0223388 | 2.89E–07 | * | −0.0596659 | 0.0183426 | 1.14E–03 | * |
| Ratio*“resolution–quality” | 0.0000023 | 0.0000003 | 2.14E–14 | * | 0.0000002 | 0.0000003 | 4.93E–01 | |
| HD*“drive” | 0.0004134 | 0.0001978 | 3.66E–02 | * | 0.0002314 | 0.0001793 | 1.97E–01 | |
| Volume*“laptop” | −0.0000007 | 0.0000054 | 8.98E–01 | | 0.0000866 | 0.0000480 | 7.12E–02 | |
| Weight*“laptop” | −0.0000097 | 0.0003648 | 9.79E–01 | | −0.0000340 | 0.0005104 | 9.47E–01 | |
| Price*“performance” | −0.0001753 | 0.0000945 | 6.35E–02 | | −0.0003485 | 0.0001016 | 6.01E–04 | * |
| PC*“performance” | 0.0125752 | 0.0304176 | 6.79E–01 | | 0.0130294 | 0.0296393 | 6.60E–01 | |
| SS*“resolution–quality” | −0.0338806 | 0.0177498 | 5.63E–02 | | −0.0049395 | 0.0201716 | 8.07E–01 | |
| OS*“apps” | 0.0265436 | 0.0535511 | 6.20E–01 | | −0.0156893 | 0.0431251 | 7.16E–01 | |

Table 11 Comparison of choice models based on the inclusion of interaction terms in the utility function

| Choice set size | Scenario | Interaction terms | $N$ | Mean | SD | SE mean |
|---|---|---|---|---|---|---|
| 7 | Normalized | Excluded (Eq. (5)) | 100 | 140,855 | 278 | 28 |
| 7 | Normalized | Included (Eq. (6)) | 100 | 140,754 | 317 | 32 |

the training set. As the result, the model performs worse when it is applied to predict the test set in the case study.

In the second set of experiments, the utility function in Eq. (6) is used. It is analogous to the function in a previous research [7], in which customer sociodemographic attributes (e.g., household income, number of children younger than 18 years, and fuel price at the vehicle purchase year) are included into the utility function by interacting them with product attributes (e.g., fuel_price * HEV_indicator). In this paper, since the customer sociodemographic attributes are not available, customer online reviews are utilized to represent the online self-presentation of customers. Specifically, both positive and negative comments from a customer toward particular product features are included in the model, as the comments may indicate the important product features for a customer. The important product features may subsequently be used to characterize customers. Table 11 shows that the explicit inclusion of customer reviews into the utility function results in significantly lower (better) KL divergence. The results reaffirm the importance of information from customer reviews in constructing choice models that have a better predictive ability. Moreover, it also shows that customer reviews, as a form of online self-presentation, reflect a person’s characteristics to an extent.

As for the limitations, the first limitation of the research comes from the inaccuracy of NLP tools, which are used to characterize customers based on their reviews. For example, it can be seen in Fig. 6 that the word “bright” is tagged as a noun (NN), instead of an adjective (JJ). The inaccuracy causes “bright screen” being excluded from the collected pairs of product feature word and sentiment polarity. The NLP tools with higher accuracy may be expected as there are more annotated data available, as well as due to the advancement of the research in the area. The other limitation is the inaccuracy of product feature words identification method, as discussed in Ref. [20]. It can be seen in Table 2 that irrelevant product feature words appear, e.g., “supervisor.” This limitation may be overcome by incorporating manual filtering toward the final product feature words, which may be performed by a product designer or an expert in the domain. The final limitation is the inability of the multinomial logit model to include nonexisting product attributes, although those attributes might have been mentioned by customers in their reviews as an expectation for a product’s improvement.

Finally, the challenge of the future research is to discover whether online data and customer reviews can replace actual choice set and sociodemographic data, when the latter data are absent. In the scope of this paper, the claim is that the online data and customer reviews contribute significantly toward constructing choice sets that generate choice models with higher predictive ability compared with constructing choice sets randomly. However, a complete data set that contains customers’ purchases, choice sets, sociodemographic data, and those customers’ reviews is required in order to answer the question of replacing the actual data with the online data.

## 6 Conclusions and Future Work

In the absence of the actual choice set and sociodemographic data, the publicly available online data of product attributes and customer reviews are valuable to construct customer choice sets. In the proposed Normalized scenario, the information is utilized to build a probability sampling for constructing customer choice sets.

In the case study, the constructed choice sets generate choice models with significantly higher predictive ability compared with the models that are created using Random scenario. Furthermore, the explicit inclusion of customer reviews to the utility function results in choice models with significantly higher predictive ability. Since the choice models with higher predictive ability provide more accurate parameter estimates of the product attribute variables, they become more useful to support designers in making engineering design decisions, especially by allowing designers to observe the change in demand with respect to the changes in the product attributes.

For the future works, more types of online self-presentation may be considered to characterize customers, e.g., past purchase history, review history, and reviewer rank. Also, different types of logit models may be used to extend the methodology for wider types of products. For example, nested logit might be appropriate for products with a nested structure, such as the aforementioned in-car DVD players.

## References

[1] Li, H., and Azarm, S., 2000, "Product Design Selection Under Uncertainty and With Competitive Advantage," ASME J. Mech. Des., **122**(4), pp. 411–418.

[2] Kumar, D., Chen, W., and Simpson, T. W., 2009, "A Market-Driven Approach to Product Family Design," Int. J. Prod. Res., **47**(1), pp. 71–104.

[3] Michalek, J., Ebbes, P., Adigüzel, F., Feinberg, F., and Papalambros, P., 2011, "Enhancing Marketing With Engineering: Optimal Product Line Design for Heterogeneous Markets," Int. J. Res. Market., **28**(3), pp. 1–12.

[4] He, L., Chen, W., Hoyle, C., and Yannou, B., 2012, "Choice Modeling for Usage Context-Based Design," ASME J. Mech. Des., **134**(3), p. 0310071.

[5] Morrow, W. R., Long, M., and MacDonald, E. F., 2014, "Market-System Design Optimization With Consider-Then-Choose Models," ASME J. Mech. Des., **136**(3), p. 0310031.

[6] Train, K. E., 2003, *Discrete Choice Methods With Simulation*, Cambridge University Press, Cambridge, UK.

[7] Wang, M., and Chen, W., 2015, "A Data-Driven Network Analysis Approach to Predicting Customer Choice Sets for Choice Modeling in Engineering Design," ASME J. Mech. Des., **137**(7), p. 0714101.

[8] Chen, W., Wassenaar, H. J., and Hoyle, C., 2013, *Decision-Based Design: Integrating Consumer Preferences Into Engineering Design*, Springer-Verlag, London.

[9] Decker, R., and Trusov, M., 2010, "Estimating Aggregate Consumer Preferences From Online Product Reviews," Int. J. Res. Market., **27**(4), pp. 293–307.

[10] McFadden, D., 1978, "Modeling the Choice of Residential Location," Transp. Res. Rec., (673), pp. 72–77.

[11] Kang, S., 2018, "Warehouse Location Choice: A Case Study in Los Angeles, CA," J. Transport Geogr.

[12] Ioannides, Y. M., and Zabel, J. E., 2008, "Interactions, Neighborhood Selection and Housing Demand," J. Urban Econ., **63**(1), pp. 229–252.

[13] Peters, T., Adamowicz, W. L., and Boxall, P. C., 1995, "Influence of Choice Set Considerations in Modeling the Benefits From Improved Water Quality," Water. Resour. Res., **31**(7), pp. 1781–1787.

[14] Valencia-Romero, A., and Lugo, J. E., 2017, "An Immersive Virtual Discrete Choice Experiment for Elicitation of Product Aesthetics Using Gestalt Principles," Des. Sci., **3**, p. e11.

[15] Gensch, D. H., 1987, "A Two-Stage Disaggregate Attribute Choice Model," Market. Sci., **6**(3), pp. 223–239.

[16] Shocker, A. D., Ben-Akiva, M., Boccara, B., and Nedungadi, P., 1991, "Consideration Set Influences on Consumer Decision-Making and Choice: Issues, Models, and Suggestions," Market. Lett., **2**(3), pp. 181–197.

[17] Jurafsky, D., and Martin, J. H., 2009, *Speech and Language Processing*, 2nd ed., Pearson Education Inc., Upper Saddle River, NJ.

[18] Levy, O., and Goldberg, Y., 2014, "ependency-Based Word Embeddings," Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Vol. 2: Short Papers), Association for Computational Linguistics, June 23–25, Baltimore, MD, pp. 302–308.

[19] Somprasertsri, G., 2010, "Mining Feature-Opinion in Online Customer Reviews for Opinion Summarization," J. Universal Comput. Sci., **16**(6), pp. 938–955.

[20] 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. 1214031.

[21] Mikolov, T., Chen, K., Corrado, G., and Dean, J., 2013, "Efficient Estimation of Word Representations in Vector Space," CoRR, abs/1301.3781, https://arxiv.org/abs/1301.3781, Accessed September 21, 2016.

[22] Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J., 2013, "Distributed Representations of Words and Phrases and Their Compositionality," CoRR, abs/1310.4546, http://arxiv.org/abs/1310.4546, Accessed September 18, 2016.

[23] Rong, X., 2014, "word2vec Parameter Learning Explained," CoRR, abs/1411.2738, http://arxiv.org/abs/1411.2738, Accessed August 8, 2018.

[24] Cambria, E., Poria, S., Bajpai, R., and Schuller, B. W., 2016, "Senticnet 4: A Semantic Resource for Sentiment Analysis Based on Conceptual Primitives," COLING 2016, 26th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, Dec. 11–16, Osaka, Japan, pp. 2666–2677.

[25] Dominick, J. R., 1999, "Who Do You Think You Are? Personal Home Pages and Self-Presentation on the World Wide Web," J. Mass Commun. Q., **76**(4), pp. 646–658.

[26] Gosling, S. D., Augustine, A. A., Vazire, S., Holtzman, N., and Gaddis, S., 2011, "Manifestations of Personality in Online Social Networks: Self-Reported Facebook-Related Behaviors and Observable Profile Information," Cyberpsychol., Behav. Soc. Network., **14**(9), pp. 483–488.

[27] Nosko, A., Wood, E., and Molema, S., 2010, "All About Me: Disclosure in Online Social Networking Profiles: The Case of Facebook," Comput. Hum. Behav., **26**(3), pp. 406–418.

[28] Li, J., and Chignell, M., 2010, "Birds of a Feather: How Personality Influences Blog Writing and Reading," Int. J. Hum. Comput. Stud., **68**(9), pp. 589–602.

[29] Wagner, C., Asur, S., and Hailpern, J., 2013, "Religious Politicians and Creative Photographers: Automatic User Categorization in Twitter," Proceedings of the 2013 International Conference on Social Computing, IEEE, Sept. 8–14, Alexandria, VA, pp. 303–310.

[30] Marriott, T. C., and Buchanan, T., 2014, "The True Self Online: Personality Correlates of Preference for Self-Expression Online, and Observer Ratings of Personality Online and Offline," Comput. Hum. Behav., **32**, pp. 171–177.

[31] Pelleg, D., and Moore, A. W., 2000, "X-Means: Extending k-Means With Efficient Estimation of the Number of Clusters," Proceedings of the Seventeenth International Conference on Machine Learning, ICML '00, Morgan Kaufmann Publishers Inc., San Francisco, CA, June 29–July 2, Stanford, CA, pp. 727–734.

[32] Bigi, B., 2003, "Using Kullback-Leibler Distance for Text Categorization," *Advances in Information Retrieval*, F. Sebastiani, ed., Springer, Berlin, Heidelberg, pp. 305–319.

[33] Bird, S., Klein, E., and Loper, E., 2009, *Natural Language Processing With Python*, 1st ed, O'Reilly Media, Inc., Sebastopol, CA.

[34] Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., and McClosky, D., 2014, "The Stanford CoreNLP Natural Language Processing Toolkit," Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics (ACL) System Demonstrations, June 23–24, Baltimore, MD, pp. 55–60.

[35] Novikov, A., 2018, Annoviko/Pyclustering: Pyclustering 0.8.1 Release, May.

[36] Řehůřek, 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, ELRA, Valletta, Malta, May 22, pp. 46–50. See also URL http://is.muni.cz/publication/884893/en

[37] Brathwaite, T., and Walker, J. L., 2018, "Asymmetric, Cosed-Form, Finite-Parameter Models of Multinomial Choice," J. Choice Modell., **29**, pp. 78–112.

[38] Jones, E., Oliphant, T., and Peterson, P., 2001, SciPy: Open Source Scientific Tools for Python, http://www.scipy.org/, Accessed January 11, 2019.

[39] Koren, Y., Bell, R., and Volinsky, C., 2009, "Matrix Factorization Techniques for Recommender Systems," Computer, **42**(8), pp. 30–37.