A Data-Driven Approach to Product Usage Context Identification From Online 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

This paper proposes a data-driven methodology to automatically identify product usage contexts from online customer reviews. Product usage context is one of the factors that affect product design, consumer behavior, and consumer satisfaction. The previous works identify the usage contexts using the survey-based method or subjectively determine them. The proposed methodology, on the other hand, uses machine learning and Natural Language Processing tools to identify and cluster usage contexts from a large volume of customer reviews. Furthermore, aspect sentiment analysis is applied to capture the sentiment toward a particular usage context in a sentence. The methodology is implemented to two data sets of products, i.e., laptop and tablet. The result shows that the methodology is able to capture relevant product usage contexts and cluster bigrams that refer to similar usage context. The aspect sentiment analysis enables the observation of a product's position with respect to its competitors for a particular usage context. For a product designer, the observation may indicate a requirement to improve the product. It may also indicate a possible market opportunity in a usage context in which most of the current products are perceived negatively by customers. Finally, it is shown that overall rating might not be a strong indicator for representing customer sentiment toward a particular usage context, due to the moderate linear correlation for most of the usage contexts in the case study.
[DOI: 10.1115/1.4044523]

Keywords: design methodology, usage context, Natural Language Processing, customer reviews

1 Introduction

Product usage context is one of the important factors that affect product design and beyond. Green et al. [1], in a study of the products that perform the primary function to broadcast light and allow mobility, concluded that the differences in product requirement design targets and customer needs may be convincingly explained according to the differences in product usage contexts. Green et al. [2], in a study of food boiling and mobile lighting products, conducted a survey that indicates different product preferences for different usage contexts. Due to its relevance toward product design, Kanis [3] stated that, instead of relying on assumed usage, discovering the actual usage context is indisputably critical.

The importance of product usage context extends beyond the product design. In an earlier research, Belk [4] showed an indication that consumer behavior is influenced by situational characteristics, including the task definition characteristic. Ram and Jung [5] showed statistically significant differences in consumer satisfaction among groups with different usage contexts, i.e., usage frequency, usage function, and usage situation. He et al. [6] argued that the reasons behind and the situations under which a product is being used, i.e., usage contexts, are essential to fully understand and model heterogeneous choice behavior. Related to choice behavior, Ratneshwar and Shocker [7] theorized that usage contexts act as environmental constraints that help define consumers' goals, such that they limit the nature of products that may be chosen to achieve those goals.

Considering the importance of product usage contexts, it is beneficial to understand product usage contexts. There are at least three benefits from understanding product usage contexts [1]:

(1) Facilitate and organize the customer needs gathering process more effectively.
(2) Improve the task of setting target design values, by taking usage contexts into consideration.
(3) Leverage the known to design for the unknown. The contextual understanding has been shown to improve the final designs, even when the design problems are outside of the designer's expertise [8].

In the literature, product usage context data are mostly collected through survey-based methods and the list of usage contexts has been predetermined, as discussed in Sec. 2.1. The main disadvantage of survey-based methods is that they may be expensive and time-consuming to conduct [9,10]. As an alternative, online customer reviews are the publicly available data that may be utilized for the purpose. Online customer reviews are mostly written based on the willingness of customers out of their own interests [10]. Customers intentionally and voluntarily invest time and energy into sharing their opinions in their reviews, such that a high level of authenticity may be expected [11]. It implies that, in terms of product usage contexts, the usage contexts that are mentioned in the reviews are of customers' true interests.

The massive volume of customer reviews, however, makes it virtually impossible to analyze the reviews manually. Therefore, this paper proposes a methodology to automatically identify product usage contexts from online customer reviews, using as little supervision as possible. In order to achieve that purpose, the data-driven methodology is supported by machine learning and Natural Language Processing tools.

[1]Corresponding author.

Contributed by the Design Theory and Methodology Committee of ASME for publication in the Journal of Mechanical Design. Manuscript received March 1, 2019; final manuscript received July 31, 2019; published online September 4, 2019. Assoc. Editor: Ying Liu.

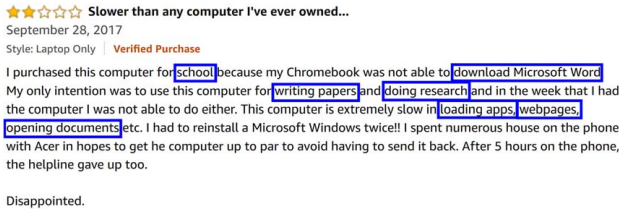

Fig. 1 An example of a customer review that perceives a laptop negatively in the usage context of writing papers

For customers and e-commerce websites, this paper contributes in proposing the possibility to allow customers to filter products based on their prioritized usage contexts. In the laptop category, up to May 2, 2019, both Amazon.com[2] and BestBuy.com[3] only offer the usage-related filtering by three general groups, i.e., Personal, Business, and Gaming. These three groups may not represent a customer's prioritized usage contexts well. Moreover, this paper shows that the overall rating may not always strongly correlate with the sentiment toward a particular usage context, i.e., a high overall rating may not guarantee that a product is good for a particular usage context.

For designers, this paper contributes in providing an insight of customer sentiment toward the usage contexts that the product is either intentionally or not intentionally designed for. For example, a review for a laptop that is marketed with the slogan "Better Everyday Computing" is shown in Fig. 1.[4] The rectangles in the figure highlight the usage contexts in that customer review. Based on the review, it is obvious that the customer is not satisfied with the laptop's performance in several usage contexts including writing papers. In other words, the reviewer's aspect sentiment toward the aspect of "writing papers" is negative. Considering the volume of the reviews for a product which is commonly in the order of thousands, the proposed methodology significantly helps designers to focus on several specific reviews regarding particular usage contexts, which may or may not have been previously realized by the designers.

The paper is organized as follows. Section 2 discusses the literature that is related to product usage contexts and recent literature in a data-driven approach to product design, especially for identifying usage contexts. Section 3 presents each stage of the methodology. Section 4 shows the results of implementing the methodology to the data sets of laptops and tablets. Section 5 discusses the performance of the proposed methodology, as well as providing further discussion on the paper's contributions for customers and designers. Finally, Sec. 6 concludes this paper.

2 Literature Review

This section presents the definitions of product usage contexts and the previous works in identifying the contexts. Section 2.2 specifically introduces data-driven approaches to product design, as well as discussing the comparison between this paper and relevant recent papers in identifying usage contexts from online reviews.

2.1 Product Usage Contexts. LaFleur [12] defined four environments in the design engineering framework, i.e., application, design, verification, and construction. The environment that is related to the product usage context is the application environment. Application environment is defined as the actual situation that a device encounters, including conditions, constraints, and actual tasks to perform. Ram and Jung [13] stated that the usage of a product may be examined from three perspectives, i.e., social interaction, experiential consumption, and functional utilization. The functional utilization perspective studies the usage of product attributes in different situations. In particular, for technological products such as personal computers, customers may use a combination of features or functions in order to enjoy usage variety in different applications, e.g., word-processing and computer games. The variety results from both the product attributes and the usage situations. Green et al. [1] defined product usage context as all factors relating to the situation in which a product may be used, including how the product is used (for what application). Finally, He et al. [6] defined product usage context as "all aspects describing the context of product use that vary under different use conditions and affect product performance and/or consumer preferences for the product attributes." Based on the definitions from the literature, *the product usage context in this paper includes the tasks or applications that a user performs using the product*.

There have been works in the literature that collect data regarding product usage contexts. In the study of the usage contexts of videocasette recorder (VCR), computer, microwave, and food processor, the data are collected from self-reported questionnaires and diaries [13]. Similarly, a field survey is conducted in order to study the usage context of VCR [5]. In the study of choice modeling for the usage context-based design, the usage context data are collected from the combination of surveying respondent and secondary data [6]. More recently, in the study of automatically identifying usage context using convolutional neural network, the data are collected from the accelerometer and gyroscope, which are embedded in the smartphones that are attached to the respondents [14]. Zhou et al. [15] utilize the usage contexts in order to elicit latent customer needs from customer reviews. However, the use case categories are subjectively predetermined (e.g., contextual events use case category includes "Seated," "On a trip," "Cooking," and "Working out"), instead of being identified from the customer reviews. As a consequence, it requires either an expert in the product domain or reading many customer reviews to create a reasonable set of use case categories. All other aforementioned works [5,6,13,14] also predetermine the usage contexts subjectively. *In contrast to the aforementioned works, this paper uses publicly available online customer reviews as the data to automatically identify product usage contexts*.

2.2 Usage Context Research in the Data-Driven Product Design Domain. The data-driven approach that does not rely on collecting data through conventional survey-based methods has become an emerging topic in the design domain. A number of recent publications propose methodologies to collect data efficiently, describing the relations among the concepts in the data and how to filter them. Lim and Tucker [16] developed a Bayesian-sampling-based methodology to identify the optimal search keyword combinations that maximize the veracity of the data acquired to make a valid conclusion. Shi et al. [17] proposed a text mining methodology that utilizes part-of-speech tags and collocations to build a network that relates the knowledge concepts in design and engineering. Zhang and Tran [18] proposed a helpfulness score to filter online customer reviews, and Zheng et al. [19] proposed a semi-supervised method to classify online customer reviews into high quality (useful) and low quality (spam or containing little information). Qi et al. [20] also filtered online customer reviews by predicting their helpfulness using five categories of features including linguistic features.

In the research that apply the data for design-related purposes, online customer reviews are used to measure the attractiveness of new product function candidates, relate product features and sales rank, and evaluate design alternatives. Zhang et al. [21] predicted the attractiveness of new product function candidates for a particular user by predicting the user's rating toward the new function. Suryadi and Kim [22] proposed a methodology to identify product features that are significantly related to sales rank. Chiu and Lin [23] evaluated design alternatives using online customer reviews.

[2] https://www.amazon.com/
[3] https://www.bestbuy.com/
[4] https://www.amazon.com/gp/customer-reviews/R3HWIC4CAWWVJ8?ASIN=B01K1IO3QW

Nevertheless, the idea of identifying usage contexts using online customer reviews has not been extensively explored. In fact, in the comprehensive review on recent advances in the data-driven product design [24], utilizing the Big Data to reveal product usage contexts is mentioned as one of the several crucial challenges and open problems in the product design domain.

In a recently published work, Yang et al. [25] addressed the challenge by proposing a faceted model of user experience. The model is illustrated in Fig. 2. Referring to the model, the usage context in this paper is represented by the sub-facet"Activities" in the Situation Facet. Despite the similarity in the attempt to identify the activities in the Situation Facet, there are at least three main differences between this paper and the work done by Yang et al. [25], i.e., the methodology, the level of generalization and automation, and the application of the obtained knowledge, as summarized in Table 1 which also includes the aforementioned work by Zhou et al. [15]. The differences are further elaborated below.

First, there are two differences in the methodology as follows:

(1) Yang et al. [25] identify product, situation, and sentiment facets separately. Those facets are subsequently combined without considering the relations between words in a sentence. Consequently, the result may be partially accurate, as shown by the examples from the customer reviews of a laptop below. In the examples, the Situation Facets are obtained from the result of the case study in Yang et al. [25] and the Product Facets are inferred from the same source.
   (a) Sentence: "i've had the laptop *for a day* - i'm pretty disappointed that the 450 g2 does not have a removable *battery*, and uses a different power supply plug than the 650 g1"
   Product Facet: battery; Situation Facet: for a day; User Sentiment State: negative
   Comment: The triple of product, situation, and sentiment may be interpreted as the battery lasting for a day and it is perceived negatively by the customer.
   (b) Sentence: "casual and hardcore gamers will find a lot to love here, and video or *photo* editing folks (who don't need to rely on a laptop screen for color accuracy) will feel *at home*"
   Product Facet: photo; Situation Facet: at home; User Sentiment State: positive
   Comment: The triple of product, situation, and sentiment may be interpreted as a positive experience of using a photo-related feature at home, although the term "at home" in this sentence has a different word sense.

   Therefore, to avoid the partially accurate results due to combining separately identified facets from a sentence, the proposed methodology in this paper attempts to identify the usage contexts along with their corresponding aspect sentiments. Based on the approach in this paper, the sentence in (b) will produce "photo editing" as a specific usage context, as opposed to just "photo" that may refer to different usage contexts (e.g., taking photo, storing photo, and photo editing) and therefore may require a designer to read the entire sentence in order to clarify it.

(2) Zhou et al. [15] and Yang et al. [25] inferred the sentiment at the sentence level. Consequently, the obtained sentiment may not actually refer to a particular usage context in the sentence. On the other hand, the state-of-the-art sentiment analysis has been performed at the aspect level, because an aspect is an integral part of an opinion. An opinion is defined as a quintuple of an entity, an aspect of the entity, the orientation of the opinion about the aspect, the opinion holder, and the time when the opinion is expressed [26]. In the context of customer reviews, aspects are defined as opinion targets, i.e., the specific features of a product or service that the reviewer likes or dislikes [27]. In this paper, an aspect is defined as the usage context of a product (i.e., the entity). Thus, aspect sentiment analysis is defined as a task to determine whether an opinion on an aspect is positive, neutral, or negative [26]. Identifying aspect sentiment is crucial because a sentence may express opposite polarities about different aspects of a product [27,28], as shown by the following

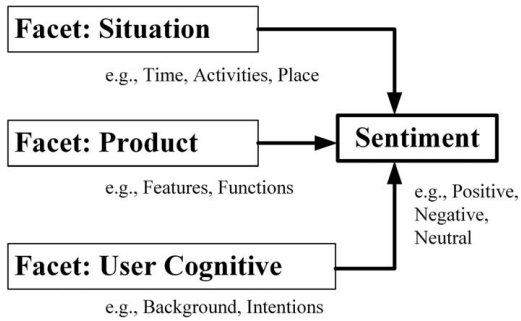

Fig. 2 A faceted model of user experience

Table 1 The summary of differences between the relevant recent works and this paper

Yang et al. (2019)	Zhou et al. (2015)	This paper
Method		
• Identify Product Facet, Situation Facet, and User Sentiment State separately.	• Determine the use cases subjectively.	• Identify the usage contexts along with their corresponding aspect sentiments.
• Sentiment analysis: sentence sentiment.	• Sentence sentiment.	• Aspect sentiment.
Level of Generalization or Automation		
• Situation Facet: sentences without product-feature or opinion words are discarded.	• Use cases are predetermined; not utilizing review sentences.	• Sentences without product-feature or opinion words are not discarded.
• Situation Facet: sentences are filtered using a model that requires manually labeled and domain-specific seeds.	• Use cases are predetermined; not utilizing review sentences.	• Sentence classifier are trained using the training set that is constructed based on domain-free grammatical rules.
• Sentiment: based on part-of-speech tagging, top *k* positive and negative words are chosen as seeds for sentiment analysis.	• Fuzzy Support Vector Machines are trained by lexicons of sentiment words.	• Aspect sentiment analysis model is trained by a big corpus.
• Clustering: local and global connection scores become the basis for clustering; the weights of each connection are subjective and the similarity measure to compute global connection is not mentioned.	• The clusters of use cases have been predetermined subjectively.	• Word vectors become the basis for clustering; they capture the meaning between phrases beyond the sameness of words.
• Clustering: predetermined *k* in k-nearest neighbors.	• Predetermined.	• X-means clustering determines *k* automatically.
Application of Obtained Knowledge		
• Building network of Product & Situation Facets to explain User Sentiment	• Identifying extraordinary use cases and the latent needs.	• Identifying a product's position in the market. • Filtering products by usage contexts.

sentence: "The voice of my Moto phone was unclear, but the camera was good" [26]. *Therefore, to obtain the corresponding sentiment toward the usage contexts, this paper applies the aspect sentiment analysis*.

Regarding the level of generalization and automation, it is argued here that a number of predetermined or subjective inputs in the methodologies proposed by Zhou et al. [15] and Yang et al. [25] may hinder their abilities to generalize to other domain of products, since it is dependent upon the subjective inputs from the experts in a particular domain. In Ref. [25], the subjective inputs are as follows:

(1) In identifying Situation Facet, the sentences without product-feature or opinion words (i.e., adjectives or verbs that contain sentiment) are discarded. On the other hand, the methodology in this paper does not rely on the existence of both product-feature and opinion words. In fact, the sentences without sentiments are still useful to inform product designers about customers' usage contexts, regardless of the existence of the sentiment. For example, the information may be utilized to obtain extraordinary usage contexts and identify lead users [15]. *Therefore, in this paper, the following sentences are not discarded and the usage contexts (in italic) are successfully identified*:
   (a) "this laptop will get the job done: *writing papers*, youtube videos in 720 (anything above 720 will have issues), gaming here and there (i can play league of legends with 30–60 fps), etc." (Note: no opinion words)
   (b) "this isn't for *hardcore gaming*" (Note: no product-feature and opinion words)
(2) In identifying Situation Facet, the sentences are filtered by a model that requires initial seeds that are manually labeled by annotators. Consequently, the annotators must be are adequately knowledgeable about the product. Furthermore, the procedure of selecting the initial seeds is not proposed. *On the other hand, this paper proposes a domain-free grammatical rules in Sec. 3.2 to construct the training set for the classifier to filter the sentences*.
(3) In identifying User Sentiment State, "based on the POS tagging, top k positive and negative terms from reviews are selected respectively as seed word lists" [25]. The approach is questionable because part-of-speech tags do not inform the sentiment of words. Moreover, the determination of k and the selected seed words may significantly affect the result. *In this paper, the aspect sentiment analysis is performed by an attention-based long short term memory network (LSTM) model* [29] *that is trained by a large corpus of similar electronic products and has been shown to perform better than or comparable to the other state-of-the-art models.*
(4) In obtaining the scores to cluster the Situation Facets, subjective weights are assigned to the local (containing same words) and global connection (appearing in similar reviews) scores. Furthermore, the similarity function to measure the similarity between reviews to calculate global connection score is not mentioned. *In this paper, the word vectors are used as the basis for clustering the usage contexts*. The word vectors are expected to capture the meaning of words beyond the sameness of words in phrases, because the phrases that contain the same word may not refer to similar usage contexts, e.g., "playing games" and "playing music." *Also, this paper utilizes X-means clustering to automatically obtain the number of clusters, as opposed to using k-nearest neighbor clustering [25] that requires the subjective determination of k, considering the fact that the true number of usage context clusters is unknown*.

Finally, the obtained knowledge in Ref. [15] is utilized to elicit latent needs. In Ref. [25], it is used to construct a network to explain the relations from Product and Situation Facets to User Sentiment State. In this paper, the applications are more practical, i.e., providing visualizations (boxplots) for designers to gauge their products' positions in the market and enabling customers to filter products based on the usage contexts of their interests. It is natural for customers to express their needs in terms of usage contexts. For example, in the research about customer-oriented product design, the inputs for a mountain bike frame design originate from customers in the imprecise linguistic forms of usage purposes (e.g., speedy, free style) and contexts (e.g., rainy, rough road) [30]. Therefore, the application of this paper should help customers naturally filter the products based on usage contexts (e.g., "video editing"), instead of based on specifications (e.g., "256 GB RAM").

Based on the similarity and differences between this paper and Ref. [25] in particular, the two methodologies should be able to complement each other. As shown in the examples above, there are sentences that might be informative for designers but they are not identified by the methodology in Ref. [25]. On the other hand, since the methodology in Ref. [25] attempts to identify broader Situation Facets including "Time" and "Place," their methodology may produce sentences with usage contexts that are not identified by the methodology in this paper. Since identifying usage contexts is an emerging topic in the data-driven product design domain, there are opportunities to combine, refine, and optimize the two methodologies.

3 Methodology

The proposed methodology consists of four stages, as shown in Fig. 3. Each stage of the methodology is discussed in the following subsections.

3.1 Preprocessing Review Sentences. A set of customer reviews is the input to this stage. Each customer review is parsed into a set of sentences, using full stops, question marks, and exclamation marks. The sentences are subsequently parsed into dependency trees. Also, the words in the sentences are represented by word embedding and tagged by their part-of-speech tags.

A dependency tree is a representation of grammatical dependencies between words in a sentence [31]. The dependency trees become the input to create a training set in Sec. 3.2. Word embedding is vectors of real numbers that represent words. The vectors are obtained from the technique, such as *word2vec*, that learns high-quality word vectors from data sets with a large number of words in the vocabulary [32]. The word vectors become the input to cluster words as well as to compute the similarity between words or phrases, in Sec. 3.4. Part-of-speech is classes of words that have similar function with respect to the adjacent words or the affixes they take [33]. The part-of-speech tags, along with the dependency trees, become the input to create the training set in Sec. 3.2.

3.2 Creating Training and Test Sets. A set of sentences from customer reviews, along with their dependency trees and part-of-speech tags of words, become the input to this stage. This stage creates labels for sentences, i.e., whether or not the sentences contain product usage contexts, based on several grammatical rules. Of all the sentences, there is generally a large fraction of sentences that cannot be labeled by the rules; due to the fact that the grammatical rules may not capture all grammatical variations of the sentences. Therefore, the labeled sentences become a training set to train the classifier in the next stage, which is used to classify the sentences that cannot be labeled by the grammatical rules.

The proposed grammatical rules are designed to be able to generalize to most types of product. Therefore, the rules are not designed to be highly elaborate. The examples of labeled sentences that are produced by the rules are presented and discussed in Sec. 5. The rules are as follows:

(1) *Rule 1 (for sentences that contain the word "usage")*: In the dependency tree, if the child of the word "usage" and

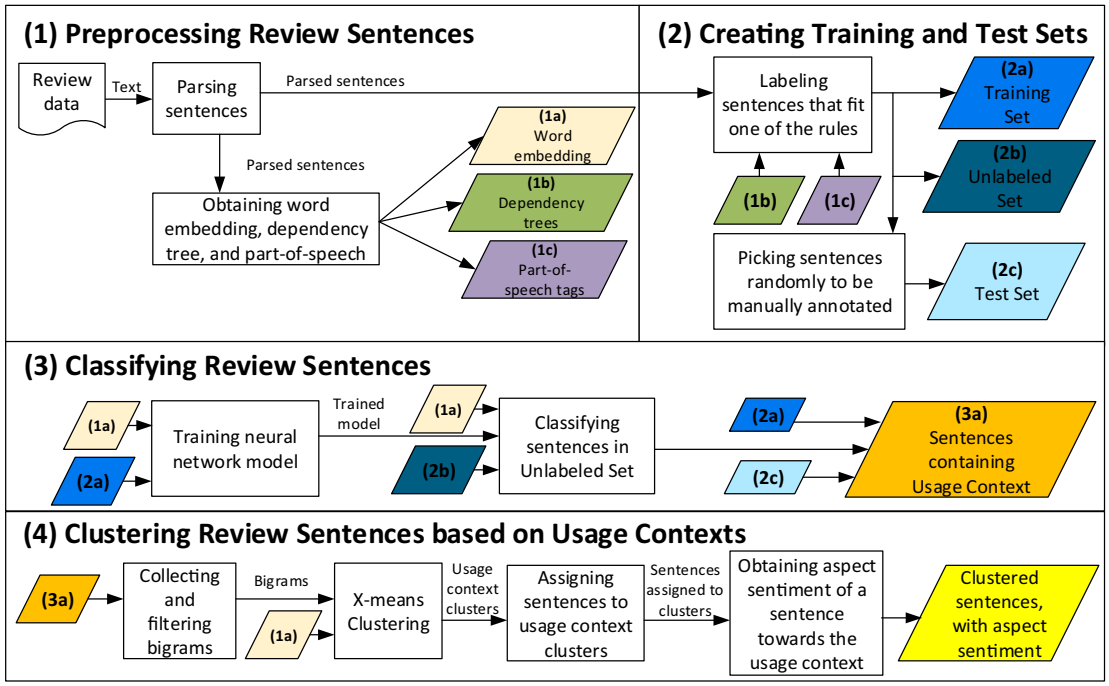

Fig. 3 Proposed methodology to automatically identify usage contexts and cluster review sentences based on the usage contexts

the child's descendants include a noun or a verb, then the sentence is labeled as a positive example. Otherwise, the sentence is temporarily labeled as a negative example.

This rule originates from the purpose of this paper, i.e., identifying usage context. Therefore, it is reasonable to collect sentences that contain the word "usage." Furthermore, the child of the word "usage" with noun or verb part-of-speech is assumed to indicate a specific task or activity, e.g., "*gaming* usage". Other parts-of-speech, such as an adjective, might not provide the specificity, e.g., "*heavy* usage".

(2) *Rule 2 (for sentences that contain the word "use," "used," "uses," or "using")*: If a sentence contains "use to" or "used to", it is temporarily labeled as a negative example because those phrases are frequently referred to either a routine or a past habit. If a sentence contains one of the aforementioned words, and the children of the word include the word "for" and a direct object, then it is labeled as a positive example. Otherwise, the sentence is labeled as a negative example.

This rule is created based on the fact that a usage context may be expressed with the verb "use" and its variants. In addition, the existence of a direct object becomes an indication that a sentence is likely to contain a usage context, such as in the following example where "it" is the direct object: "right now i just use *it* for internet browsing and Pandora."

(3) *Rule 3 (for sentences that do not contain the words that are queried by* Rule 1 *and* Rule 2*, but contain the word "for")*: If the child of the word "for" is tagged as a VBG (i.e., verb, gerund, or present participle [34]), then the sentence is labeled as a positive example. Otherwise, the sentence is labeled as a negative example.

This rule is created based on the fact that the preposition "for" is a function word that is used to indicate purpose, as explained by the entry in Merriam-Webster dictionary.[5] In the example of the entry, the purpose is stated with a VBG-tagged word as well, i.e., "a grant for *studying* medicine."

A sentence that does not meet a rule's condition remains unlabeled by the rule, e.g., a sentence that has no "usage" word in it is unlabeled by Rule 1. As for Rule 1 and Rule 2, if a sentence is temporarily labeled as negative by one rule but is unlabeled by the other, then it is given a final label as negative. If a sentence is labeled as negative by both Rule 1 and Rule 2, it is also given a final label as negative. Otherwise, its final label is positive, because a sentence might contain both "usage" and "use," but only one of them meets the rule. As for Rule 3, the given label is final. The sentences with their final labels form the training set. Finally, all sentences that do not pass any rule form the unlabeled set.

As for creating the test set to evaluate the classifier's performance, a set of sentences are randomly selected and excluded from the training and unlabeled sets. In order to maintain the similar distribution to the training set, it may be suggested to build at most half of the test set by randomly selecting sentences from the training set and build the other portion by randomly selecting sentences from the unlabeled set. The sentences from the training set are stripped from the labels assigned by the three rules above. In order to ensure the correct label of the sentences in the test set, the sentences are manually labeled by more than one annotator. The order of the sentences has been randomly shuffled before being annotated. Since the product usage context in this paper includes the tasks or applications that a user performs using the product, a question is used to guide the annotator, i.e., "Does the sentence tell the tasks or applications that a user performs using the product?" The sentences of which annotators agree on their labels form the test set.

3.3 Classifying Review Sentences. The training, test, and unlabeled sets that are created in the previous stage become the inputs to this stage. The main purpose of the classifier in this stage is to filter the sentences in the unlabeled set, which is generally larger than the other two sets, such that the sentences that contain usage contexts may be obtained without applying overly elaborate rules. Furthermore, in this paper, the classifier performance is also used as the basis to select the hyper-parameter values in word embedding. In Sec. 4, several sets of hyper-parameter values are applied to the case study and the best set of hyper-parameter values is selected based on the classifier performance.

[5]https://www.merriam-webster.com/dictionary/for

The classifier in this paper is a one-layer neural network. The inputs are the word embedding of words in a sentence. The squashing function in the output node is a sigmoid function, such that the output is between 0 and 1. The weights are trained using the training set, in which the sentences are labeled as either positive (1) or negative (0). Once the classifier has finished the learning process, the weights may be applied to any sentence and produce a value between 0 and 1. After applying a thresholding, as explained in Sec. 4.2, the sentences in the unlabeled set may accordingly be labeled as either positive or negative. The sentences that are classified as negative are excluded from the next stage of the methodology.

In order to assess the performance of a classifier, the metric Area under the Receiver Operating Characteristic (AUROC) is used. The receiver operating characteristic curve is obtained by plotting the true positive versus false-positive rates for all possible threshold values. The AUROC may then be interpreted as, given a positive and a negative example, the probability of the classifier to output a higher prediction value for the positive example [35]. Therefore, the larger AUROC value indicates the better classifier.

3.4 Clustering Review Sentences Based on Usage Contexts. The sentences that are classified as positive by the classifier become the input for this stage, along with the sentences that are labeled as positive in the training and test sets. The purpose of this stage is to cluster usage contexts, such that a sentence that contains a usage context may be clustered as well, and thus the proportion of usage contexts of a product may be obtained. Furthermore, this stage aims to reveal the sentiment in a sentence with respect to the usage context in the sentence, which is known as aspect sentiment. By obtaining the aspect sentiments, analysis on the aspect sentiment distribution among products and correlation between aspect sentiment and overall rating may be performed, as shown in Sec. 4.4.

In order to obtain usage contexts from the input sentences, bigrams are collected from the sentences. In this paper, the usage contexts are assumed to be bigrams. The assumption is taken because including unigrams is expected to return a set of words that contains too much noise, i.e., words that are irrelevant to usage contexts, e.g., "wondering." As a consequence, a one-word usage context is omitted, e.g., "writing" is omitted, but "writing paper" is included. Furthermore, the collected usage contexts should be specific enough such that they are informative for designers and useful for customers. In many cases, the specificity of a usage context may not be captured with a unigram. For example, "video" is not specific enough, as it may refer to the activity of watching video, editing video, etc.; "playing" is also not specific enough as it may refer to playing music, playing games, etc. Therefore, in order to obtain adequately specific usage contexts, this paper assumes the usage contexts as bigrams.

The collected bigrams are subsequently clustered using the X-means clustering method. It is chosen due to its ability to obtain the number of clusters automatically, by optimizing the Bayesian Information Criterion [36]. In the context of usage contexts, it is difficult to determine the correct number of clusters of usage contexts. For example, in the case of laptops, it is highly debatable whether or not "watching movie" and "watching youtube" should be in the same cluster of usage contexts. Therefore, X-means is considered suitable for the task at this stage. Moreover, it has been shown that X-means clustering performs better than a spherical K-means clustering in the case study of laptops [37].

As for the aspect sentiment, it is obtained by applying the attention-based LSTM method proposed by He et al. [29]. The method is chosen because it is a state-of-the-art aspect sentiment analysis method, which achieves a relatively comparable or even better performance than the other recent methods, including when it is applied to a data set of laptop customer reviews from Amazon.com.[6] The method outputs the sentiment of a sentence with respect to an aspect sentiment in three scores that sum up to 1, which correspond to positive, negative, and neutral sentiments. In this paper, the numbers are aggregated by subtracting the negative score from the positive score. Therefore, the range of the aggregated sentiment is [−1,1].

[6]See Note 2.

Table 2 The number of sentences in each data set

Data set	Training set (+)	Training set (−)	Test set (+)	Test set (−)	Unlabeled set
Laptop	15,578	100,058	72	465	1,028,573
Tablet	5188	18,520	51	166	318,448

4 Data and Results

This section starts with describing the data sets that are used to implement the proposed methodology in Sec. 3. The first subsection discusses the word embedding hyper-parameter value selection. It is followed by comparing the results from using and not using a sentence classifier. The third subsection presents the detailed results from clustering the identified usage contexts. Finally, the results from applying the aspect sentiment analysis to the usage contexts are presented.

Two data sets are used in this paper. The laptop data set contains 5419 laptops from the traditional laptops category in Amazon.com.[7] It also contains 218,570 customer reviews of those laptops up to Dec. 13, 2017. The tablet data set contains 373 tablets from BestBuy.com.[8] It also contains 134,219 customer reviews of those tablets that are posted between Nov. 4, 2014, and Oct. 8, 2018. The proportions of reviews with verified purchase label are 85.82% and 98.66% for laptop and tablet data sets, respectively. Therefore, most of the reviews are expected to be authentic because they are written by customers who have been verified to purchase the products.

When a classifier is used, as proposed in the methodology shown in Fig. 3, a training set is required to train the classifier's parameters and a test set is needed to assess the classifier's performance. Therefore, as explained in Sec. 3.2, the review sentences are divided into training set, unlabeled set, and test set. The number of sentences in each set for both laptop and tablet data sets is shown in Table 2.

4.1 Word Embedding Hyper-Parameter Value Selection Result. The performance of a classifier is affected by the word embedding. Therefore, this subsection shows the selection of the word embedding hyper-parameter values based on the classifier performance. The hyper-parameters that are included in the experiment are the dimension of a word vector (*size*), the window size (i.e., the maximum distance between the farthest context word and the predicted word) (*window*), and the minimum frequency for a word to be included in the embedding (*minCount*). The performance of a classifier is measured by the AUROC metric.

In this paper, the word embedding is implemented via `gensim` package in PYTHON [38]. The word embedding becomes the input for the classifier that consists of one layer and applies a sigmoid function as the squashing function in the output layer. The classifier is implemented via `keras` package in PYTHON. The classifier performance comparison for the selected word embedding hyper-parameter values is shown in Table 3. The highest AUROC value is denoted with an asterisk and the word embedding obtained from the corresponding hyper-parameter values is used in the later stages.

4.2 Sentence Classifier Result. This subsection shows the qualitative and quantitative comparisons between using a classifier

[7]See Note 2.
[8]See Note 3.

Table 3 The classifier performance comparison in laptop and tablet datasets

(Laptop) size	Window	minCount	AUROC		(Tablet) size	Window	minCount	AUROC	
25	2	5	0.8055		25	2	5	0.7816	
25	2	10	0.8389	*	25	2	10	0.8110	
25	3	5	0.8190		25	3	5	0.7958	
25	3	10	0.8158		25	3	10	0.8280	*
50	2	5	0.7722		50	2	5	0.7264	
50	2	10	0.7683		50	2	10	0.7570	
50	3	5	0.7756		50	3	5	0.7377	
50	3	10	0.7934		50	3	10	0.7323	

to classify the review sentences, as proposed in the methodology in Sec. 3 and not using a classifier. The comparison is made in order to justify the Classifying Review Sentences stage in the proposed methodology.

After obtaining the word embedding with the best hyperparameter values in Table 3, review sentences in the unlabeled set may be filtered by a classifier before entering the bigram clustering stage. First, the bigrams are collected by the CountVectorizer function of `sklearn` package in PYTHON [39]. The collected bigrams are then refined by removing the unlikely phrases, using the phraser function of `gensim` package in PYTHON [38]. The function is based on the pointwise mutual information (PMI) metric that calculates the probability of words in a phrase appearing together compared with the multiplication of the probabilities of each word appearing by itself [40]. A phrase with high PMI indicates that the phrase is likely a valid phrase.

Moreover, each word in the bigram must be either a noun or a verb and one of the words must end with "-ing." The reasoning behind that is as follows. The words that end with "-ing" are likely to be verbs, gerunds, or present participles; which reasonably describe activities. Other than a verb, a specific bigram often contains a noun as well, e.g., "typing *documents*," "reading *e-books*," or even a pair of nouns that describe usage contexts, e.g., "*web surfing*" and "*photo editing*." These filtering steps are performed in order to reduce noise in the collected bigrams and produce specific activity phrases. The final set of bigrams are clustered into usage context clusters using `pyclustering` package in PYTHON [41]. On the other hand, without using a classifier, bigrams are immediately collected from the review sentences, refined, and clustered.

Before going into the comparisons, it is worth noting that the classifier outputs a value between 0 and 1, due to the sigmoid as the squashing function in the output layer. However, in order to classify a sentence as containing usage context or not, a binary decision is required, i.e., 0 or 1. Therefore, a thresholding process is performed. The classifier is applied to the test set, and the best threshold is chosen such that the accuracy on the test set is the highest. In laptop data set, the best threshold is obtained at 0.214, resulting in 89.01% accuracy on the test set. In tablet data set, the best threshold is obtained at 0.323, resulting in 87.56% accuracy on the test set. The thresholding graphs are shown in Fig. 4, in which the *X*-axis shows the threshold and the *Y*-axis shows the accuracy on the test set. The classifier is applied to the unlabeled set, with the aforementioned thresholds, and yields 25,300 positive sentences in laptop data set and 25,556 sentences in tablet data set. Those numbers are 2.46% and 8.02% of the sentences in the unlabeled sets of laptop and tablet data sets, respectively. This reduction supports the methodology to obtain relevant bigrams, as the sentences that are unlikely to contain usage contexts have been filtered out.

A qualitative comparison is made by comparing the most frequent bigrams in the clusters. In laptop data set, the list of most frequent bigram in each cluster is as follows:

- With classifier (ten clusters): gaming rig/power saving (equally frequent), operating system, processes running, playing game, word processing, transferring files, web browsing, writing paper, video editing, and watching movie.
- Without classifier (15 clusters): viewing angles, operating system, web browsing, video editing, stopped working*, processing speed, docking station*, learning curve, shipping label*, star rating*, processing power, cooling pad*, selling point*, transferring files, and playing games.

In tablet data set, the list of most frequent bigram in each cluster is as follows:

- With classifier (four clusters): operating system, web browsing, watching movies, and reading books.
- Without classifier (six clusters): selling point*, operating system, photo editing, web browsing, watching movies, and reading books.

It can be seen that, without using a classifier, there are frequent-but-irrelevant bigrams in the clusters of usage contexts, which are denoted by an asterisk in the list. On the other hand, the clusters that are produced from the classified sentences are represented by bigrams that are relevant to usage contexts.

A quantitative comparison is made by comparing the average cosine distance between bigrams within a cluster and between most frequent bigrams of the clusters. The smaller distance between bigrams within a cluster shows more cohesiveness of the clusters, i.e., the bigrams within a cluster have the similar

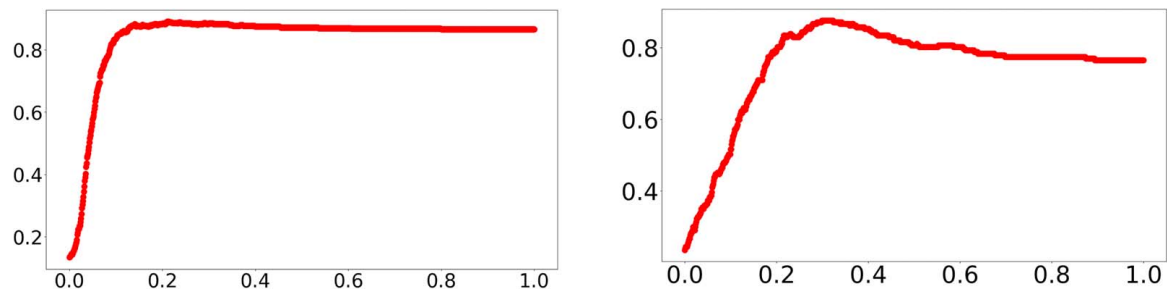

Fig. 4 Thresholding for classifier in laptop (left) and tablet (right) data sets

Table 4 The comparison based on average cosine distance between with and without using a classifier

Data set	Within cluster (with classifier)	Within cluster (without classifier)	Between most frequent bigrams (with classifier)	Between most frequent bigrams (without classifier)
Laptop	0.3806	0.5110	0.7638	0.9383
Tablet	0.3106	0.4569	0.7550	0.8194

meaning. The smaller distance between most frequent bigrams shows that the identified frequent usage contexts are more likely to refer to the same concept, e.g., the usage contexts of laptops. The comparison is shown in Table 4. It can be seen that a classifier produces more cohesive clusters, as shown by the lower average cosine distance values compared with without using a classifier, as well as more similar most frequent bigrams. The result holds for both data sets.

4.3 Usage Context Clustering Result. Section 4.2 justifies the usage of a classifier in the proposed methodology. This subsection further observes the obtained clusters of usage contexts. Once the bigrams have been clustered into usage context clusters, the customer review sentences may be assigned to the clusters based on the usage context bigrams that are contained in the sentence. The assignment produces the charts in Fig. 5, which show the proportions of usage contexts for both laptop and tablet data sets. Each fraction in Fig. 5 is labeled by the most frequent bigram in the cluster. It may be observed that there are fewer usage contexts identified from tablet data set.

Table 5 shows a sample of frequent bigrams in each cluster in laptop data set. The sample of frequent bigrams in tablet data set is shown in Table 6. The top bigrams may be used to analyze the qualitative performance of the clustering, which will be discussed in Sec. 5.

In order to measure the quantitative performance of the identified usage contexts in the sentences, the Precision metric is calculated. For both laptop and tablet data sets, 150 sentences are randomly selected to be independently labeled by two annotators. A sentence is labeled as "True" if the identified usage context is correct, i.e., providing the information about the task or application that is performed by the customer using the product, but otherwise "False". The sentences that two annotators agree on the labels are collected. The precision metric is then calculated as the ratio of the "True" sentences to the total number of sentences. The results are shown in Table 7. The precision of the laptop data set (almost 70%) is lower than the tablet data set (more than 87%) due to the fact that a laptop has more functions than a tablet, such that it has more usage contexts as well and thus it is more difficult to identify the correct usage contexts.

The performance metric Recall is not calculated here due to the difficulty in determining the true number of usage contexts. The Recall metric calculates the ratio of correctly identified usage contexts to the number of true usage contexts. The difficulty lies in the fact that the true number of usage contexts depends on the granularity of the contexts. For example, in the sentence in Fig. 1, "loading apps," "(loading) webpages," and "opening documents" may be considered as three different types of usage contexts. However, "loading apps" and "opening documents" may reasonably be merged into one context, i.e., "loading apps", because "opening documents" may be interpreted as loading an app to open a document as well. The more complicated difficulty is when one activity may be considered as a subset of the other. For example, in the context of customer review

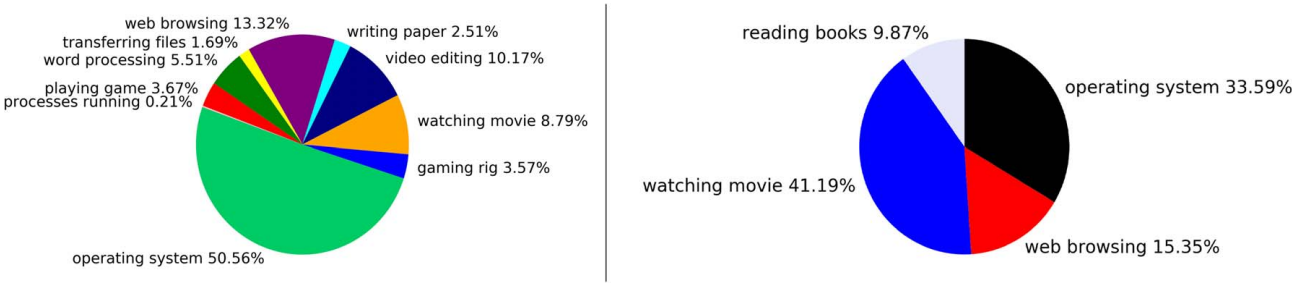

Fig. 5 The proportion of customer reviews in each usage context cluster in laptop (left) and tablet (right) data sets

Table 5 The sample of bigrams in each cluster in laptop data set sorted by descending frequency

Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4
operating system	gaming rig	playing game	processes running	transferring files
stopped working	power saving	demanding game	loading webpages	computing tasks
viewing angle	computing power	streaming media	handles multitasking	demanding tasks
stop working	computing needs	playing minecraft	—	loading pages
processing power	hardcore gaming	playing fallout	—	tried uninstalling
learning curve	engineering student	playing overwatch	—	demanding applications
Cluster 5	**Cluster 6**	**Cluster 7**	**Cluster 8**	**Cluster 9**
word processing	writing paper	web browsing	watching movie	video editing
internet surfing	reading reviews	web surfing	watching video	photo editing
surfing internet	writing document	internet browsing	watching netflix	streaming video
document processing	typing papers	checking e-mail	watching youtube	video streaming
—	writing essays	browsing internet	streaming movie	document editing
—	reading text	surfing web	playing music	editing photo

Table 6 The sample of bigrams in each cluster in tablet data set sorted by descending frequency

Cluster 0	Cluster 1	Cluster 2	Cluster 3
operating system	web browsing	watching movie	reading books
learning curve	web surfing	playing games	reading ebooks
processing speed	checking e-mail	watching videos	reading magazines
stopped working	internet browsing	watching netflix	reading articles
processing power	surfing web	video editing	reading glasses
photo editing	surfing internet	movie watching	reading comics

Table 7 Precision of the identified usage contexts

Data set	True	False	Precision (%)
Laptop	85	37	69.67
Tablet	118	17	87.41

in Fig. 1, "writing papers" is a subset of "doing research." It is not clear whether or not those two usage contexts should be counted separately. Suppose the methodology identifies "writing papers" but not "doing research." It is unclear whether it should be counted as a miss. In fact, "doing research" is an unspecific term that may include various activities such as web browsing, watching video, running simulation, etc., such that it is arguably acceptable to either identify it as a usage context or not.

4.4 Aspect Sentiment Analysis Result. Aspect sentiment analysis is performed for the most frequent bigram in each usage context cluster, as the representation of the cluster. First, the aspect sentiment analysis is used to show the distribution of sentiment toward a particular usage context among all products in the data set. Therefore, product A, for example, may compare its relative position to product B based on average customer sentiment with respect to "video editing" usage context. Furthermore, product A may obtain its relative position among all products in the data set with respect to that usage context. For laptop and tablet data sets, the aspect sentiment distributions are shown in Figs. 6 and 7, respectively. It may be observed that the customers in tablet data sets are generally more positive toward the product in all usage contexts.

Moreover, in order to examine whether or not there is a strong linear correlation between aspect sentiment towards a particular usage context and the overall rating, the boxplots in Figs. 8 and 9 are created. For each usage context, a plot that consists of four boxplots is created. The *X*-axis corresponds to the ranges of aspect sentiment of [−1, 0.5), [−0.5, 0), [0, 0.5), and [0.5, 1]. The *Y*-axis is the overall rating of a product. The interpretation of the plots may be made as follows. For example, in Fig. 8, for the usage context "playing game," the laptops that have average sentiment toward that context in the range of [−1,0.5) are the laptops whose overall rating median is around 4. There is an outlier laptop in that group, whose overall rating is below 2. The boxplots, along with the correlation coefficient values, demonstrate that there is a

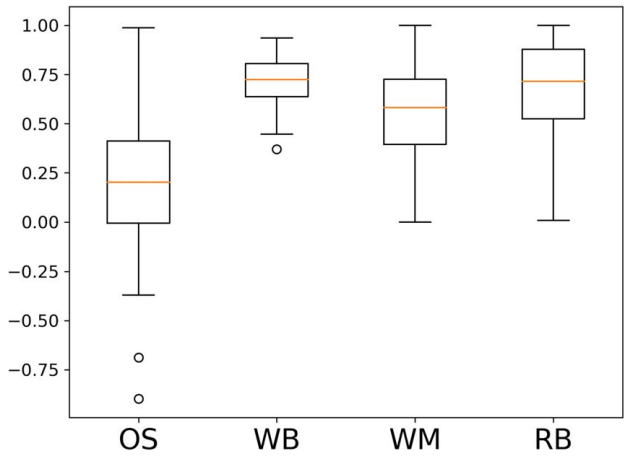

Fig. 7 Boxplots of the aspect sentiment related to each usage context in tablet data set, where OS, operating system; WB, web browsing; WM, watching movie; and RB, reading books

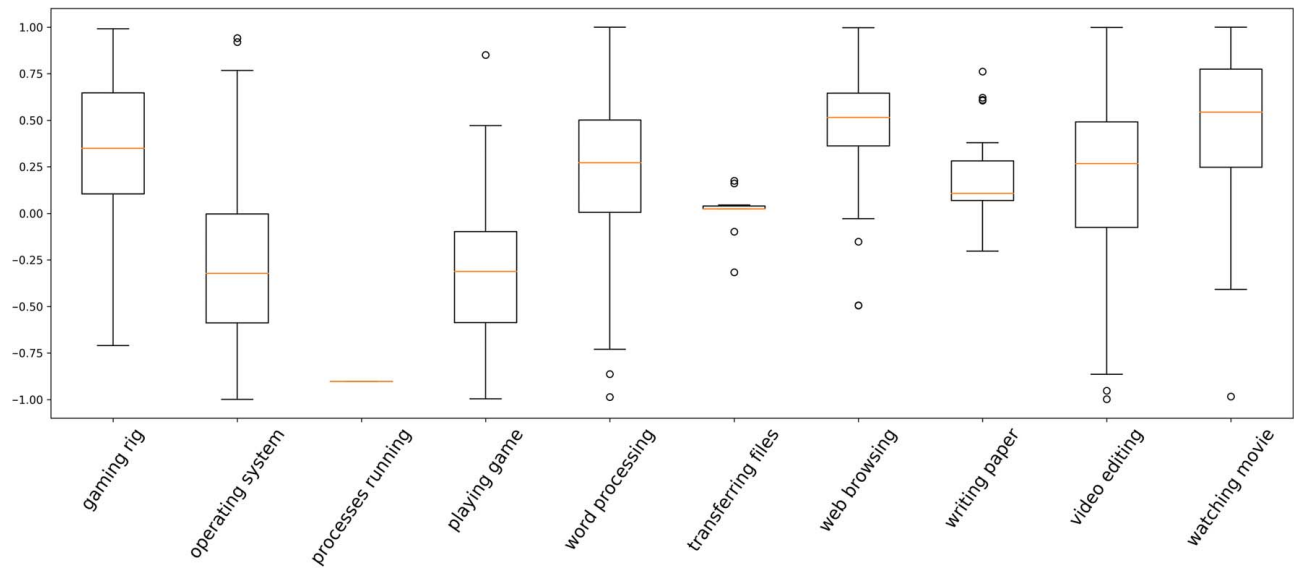

Fig. 6 Boxplots of the aspect sentiment related to each usage context in laptop data set

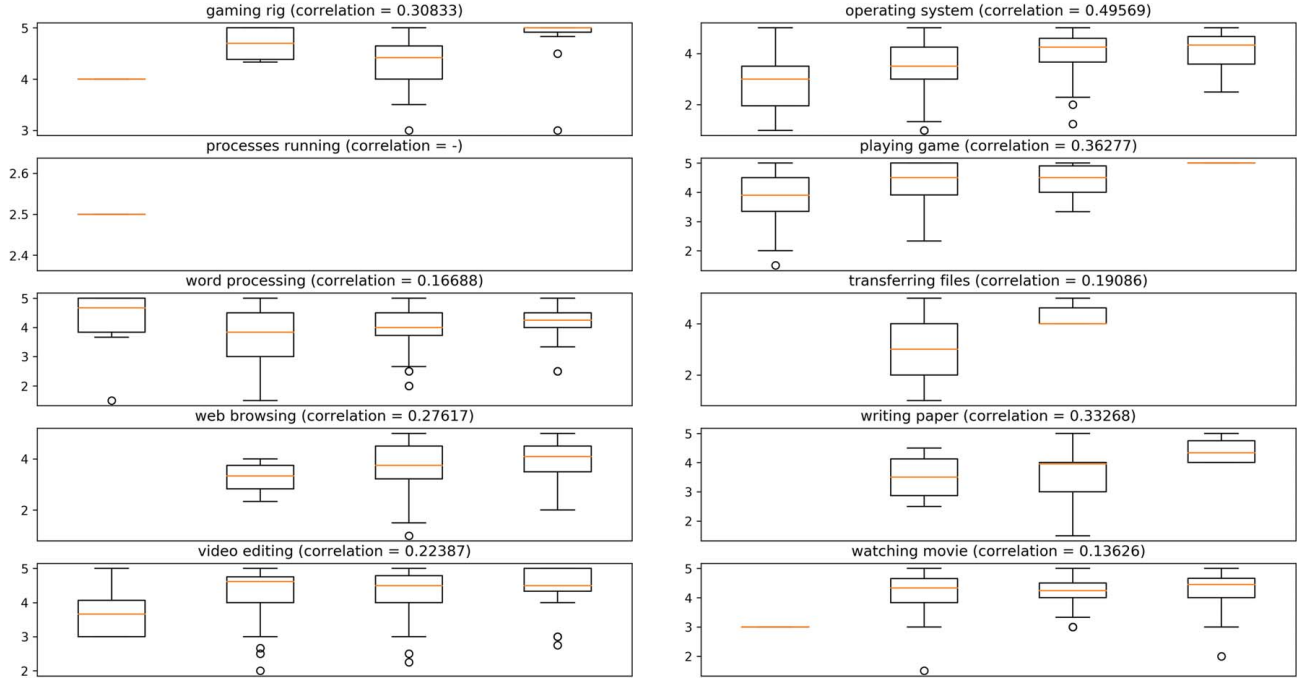

Fig. 8 Boxplots of the aspect sentiment related to each use case in laptop data set

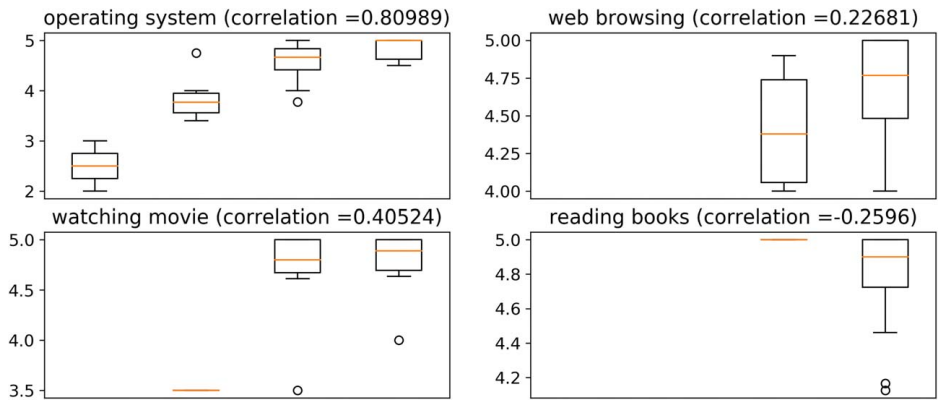

Fig. 9 Boxplots of the aspect sentiment related to each use case in tablet data set

weak-to-moderate positive correlation between aspect sentiment and overall rating for most of the usage contexts. In other words, it is shown here that the higher overall rating of a product does not strongly correlate to a higher sentiment toward a particular usage context of the product.

5 Discussion

This section first discusses the proposed methodology's performance based on the results in Sec. 4. It is followed by the subsections related to the contributions of the methodology for customers and product designers.

5.1 Methodology's Performance. The grammatical rules that are used to label sentences generally produce correct labels. The examples of positive and negative sentences that are produced by the rules in laptop data set are presented as follows, in which the sentences are retained in their original writings including the grammatical and typographical errors:

- *Rule 1 (positive)*: "it's slim and lightweight, not too fast, not recomended formulti tasking or complex software, but is ok for *everyday* **usage** *like web browsing, email, word processor, etc*"
- *Rule 1 (negative, because the child of the word "usage", i.e., "moderate", is neither a noun nor a verb)*: "battery life is perfect, it lasts 12 h as stated in the description and up to 10 h on moderate **usage**"
- *Rule 2 (positive)*: "if you plan to **use** the *laptop* **for** more than *browsing* and *watching movies* then you might consider it a waste of $ 250"
- *Rule 2 (negative, because the children of the word "use" do not include a direct object and the word "for")*: "it automatically selects which one to **use** based on what you are doing to conserve battery power"
- *Rule 3 (positive)*: "the non-glare finish is much better than the glossy displays gracing many other notebooks, and the ips display has dramatic superiority **for** *viewing angle accuracy*"
- *Rule 3 (negative, because the part-of-speech tag of the child of the word "for" is not VBG)*: "i was lucky enough to get mine **for** 1000 as it was mis-marked at the base exchange where i bought it"

The rules also produce false-positive examples, such as "i **use** it pretty much everyday **for** *extended periods* of *time* and only

charge a few times a week" (Rule 2), and false negative examples, such as "i bought it mainly **for** work, they have desk-tops there but we have to log - in with our clock # and they can watch your every move and put you on the corporate i/t watchlist if you transgress, omg google images, he looked up what" (Rule 3). In the false-positive example, "extended periods (of) time" is identified as a usage context. Meanwhile, in the false negative example, "work" is not identified as a usage context. A set of more elaborate rules may produce fewer false examples. Nevertheless, the rules in the proposed methodology are intentionally designed to be not too detailed, such that the rules may generalize to other product domains and there is as little subjective supervision as possible in the methodology.

As for the qualitative performance of the clustering, Table 5 displays a reasonable result in clustering "gaming rig" and "hardcore gaming" together with "computing power" in Cluster 1, which indicates the need for computing power. Meanwhile, there is another cluster, i.e., Cluster 2, that groups usage contexts that are similar to "playing game," which may be interpreted as requiring less computing power. This result indicates that the proposed method is able to capture the meaning behind the bigrams, instead of simply capturing bigrams that contain the same word. This argument is further supported by the separation of "playing music" in Cluster 8 and "playing game" in Cluster 2, although both bigrams contain the word "playing." The proposed methodology is not totally accurate, obviously, because there are some bigrams that seem to fit better in another cluster, e.g., "streaming media" in Cluster 2 is intuitively more compatible with the terms in Cluster 8.

When a product can be considered as a subset of the other in terms of functionality, such as tablet to laptop, the methodology obtains less number of usage contexts as well. It can be seen that there are four clusters of usage contexts in tablet data set in Table 6, compared with 12 in Table 5 for laptop data set. This result qualitatively justifies the ability of the methodology in obtaining usage contexts from customer reviews.

Aside from the limitations that arise from the inaccuracy of machine learning and Natural Language Processing tools, the proposed methodology is unable to weigh the usage contexts that are mentioned by one customer. The weights should capture the finer level of the importance of different usage contexts for a customer. For example, a customer might comment positively on the wide viewing angle for a laptop, but the customer does not find it important, because the customer's main usage context is streaming music. Therefore, it would be appropriate to weigh the customer's positiveness accordingly.

5.2 Contribution for Customers. The proposed methodology may benefit customers in a way as follows. Suppose a customer compares two laptops as shown in Table 8, along with their average aspect sentiments with respect to "watching movie" and "video editing" usage contexts. The methodology allows customers to notice that, while both laptops have similar average sentiment for "watching movie," laptop 7b has a significantly higher average sentiment for "video editing." Therefore, if the customer considers both usage contexts as important, the comparison may cause laptop 7b to be preferable for the customer. Under the current filtering options in Amazon.com[9] and BestBuy.com,[10] it is difficult for customer to filter and compare laptops by these criteria.

The sample of review sentences for both laptops in Table 8 with respect to both usage contexts are shown in Table 9. The review sentences are presented to qualitatively justify the sentiment scores. In Table 9, it may be observed that both laptops receive sentences with positive sentiment towards "watching movie" usage context. For "video editing" usage context, laptop 7a has been mostly described as being capable for light video editing. On the other hand, laptop 7b has been positively described as being suitable for video editing, except for the fourth sentence that complains about the nonexistence of a set of numeric keys on the keyboard. Therefore, as shown in Table 8, laptop 7b has a higher average sentiment value than laptop 7a for "video editing" usage context.

[9]See Note 2.
[10]See Note 3.

Table 8 Example of two laptops with their average aspect sentiments for two usage contexts

Product	Context: watching movie	Context: video editing
Laptop 7a (B005CWJB5G)	0.8283	0.0661
Laptop 7b (B007474DSM)	0.8239	0.5699

5.3 Contribution for Designers. For the designers, Figs. 6 and 7 may be used to identify the opportunity in the market. In the case of laptops, the improved products may be targeted for the usage contexts of playing game and operating system. Those are the usage contexts for which most of the laptops are perceived negatively by the customers. In the case of tablets, there is also an opportunity to improve the operating system in order to stand out from the competitors.

Moreover, in a more detailed level, designers may examine the extracted sentences from the customer reviews with respect to a particular usage context. For example, the products that have the highest and the lowest average aspect sentiment with respect to the usage context of "writing papers" are shown in Table 10, along with the corresponding review sentences and the average aspect sentiments.

Taking laptop 9b as an example, the designers of laptop 9b may want to improve their product, since it currently has the lowest sentiment with respect to "writing papers" usage context compared with all other laptops in the data set. The improvement becomes essential if laptop 9b targets customers who frequently write papers on laptops. While the sentence may not offer the complete problem description by itself, the designer of laptop 9b may carefully examine the entire review from this particular customer as shown in Fig. 1. The review reveals that the customer experiences the need to reinstall the operating system, although in fact the laptop has been equipped with Windows 10 Home. Also, the customer perceives the laptop as extremely slow in performing basic functions. The result has therefore significantly narrowed down the number of customer reviews that a product designer needs to focus on.

Addressing the importance of obtaining actual, as opposed to assumed, usage context [3], a pie chart is created in order to show the usage contexts of gaming laptops, i.e., the laptops that contain the words "Gamer," "Gaming," "Alienware," and "MSI" in their names. The latter two terms are the brands of gaming laptops. The chart in Fig. 10 shows that gaming laptops obviously have larger proportions in the usage contexts of "gaming rig" and "playing game" compared those in the overall laptop data set in Fig. 5. It may be seen that the proportions of several other usage contexts are not negligible. Therefore, the gaming laptop designers should not assume that customers only use the laptops for gaming purpose, especially since there have been negative sentiments toward these other usage contexts. Furthermore, the negative sentiment may include a suggestion for improvement, as shown by the following sentence: "-there is *no dedicated pgup/pgdn key* on the razer, kind of annoying during **web browsing**," which is written toward "web browsing" usage context. The improvement might be beneficial in order to attract people who frequently use gaming laptops for web browsing as well. By noticing the usage contexts that might have not been previously considered as important, designers might formulate design improvements in order to attract either the targeted or new customers.

Table 9 The sample of review sentences for two laptops in Table 8 with respect to the corresponding usage contexts

7a	• battery life is amazing , i get 4 hours of youtube watching or **watching movies** on a plane • in addition , i'll admit that i kind of use it as a "portable dvd player" **watching movies** in bed when i'm too lazy to head over to my mac pro	• i only do very minor music and **video editing** • **video editing** is functional as well and would likely work for most casual users , but massive projects simply wouldn't be possible on this machine for a multitude of reasons ranging from storage space to video card , screen size , processor etc • if you have a main computer , the air is a very good addition but if you only need an all in one computer like me to take with you , store data , watch videos , do picture and **video editing** and plan to use for a very long time , then the macbook pro is a better choice and for me is more sturdy • if you were a graphic designer or video editor , this might not have enough power for you , but i've done a bit of **video editing** in imovie on this , and it worked really well , and didn't slow down at all
7b	• i travel a lot and **watching movies** / tv shows while on flights with the retina display is amazing • why i , personally , chose the retina mbp : - i use my machine for **watching movies** , work in grad school , internet , and a potpurri of other things	• i mainly use it for **video editing** , photo editing , household management and other personal tasks • if you use your mac for picture editing or for **video editing** its a good use of your money • those who don't require this performance might want to look at other macs , but if you run graphic - intense programs , do **video editing** , watch a lot of media via your computer , apple has really delivered • the problem is that this laptop has a great screen , runs fast and therefore should be great for **video editing** on the road , but i use the 10 key portion of a keyboard when video editing , and this laptop does not have one • with intel's latest turboboost - capable ivy bridge processors , up to 16gb of sdram and a potential of 768gb of the fastest flash storage available , the retina macbook pro can easily accomplish simple tasks such as video playback , as well as the more complex - such as hardware - intensive **video editing**

Table 10 The example of review sentences from the products with the most positive and negative average aspect sentiment for a particular usage context

Product	Sentence	Average
Laptop 9a	"i'm liking windows 7, and the computer comes with ms works which gives you as much as most need for **writing papers**	0.76262
(B0030INLSW)	or doing spreadsheets" "overall i'm satisfied : i have a huge screen for studying and **writing papers**, the keyboard is a great design allowing for comfortable typing with responsive keys, and appropriately clicky buttons"	
Laptop 9b	"it works fine for me, someone just using it for college and **writing papers** but i wouldn't buy it again"	−0.20224
(B01K1IO3QW)	"my only intention was to use this computer for **writing papers** and doing research and in the week that i had the computer i was not able to do either"	

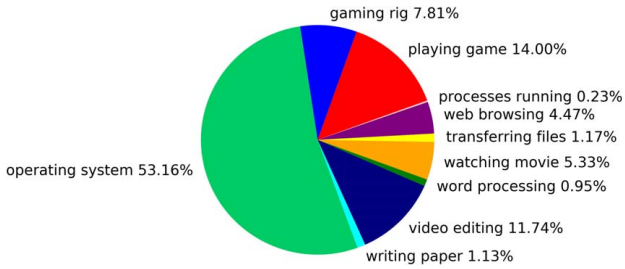

Fig. 10 The proportion of customer reviews in each usage context cluster in gaming laptops

6 Conclusion and Future Work

A data-driven methodology has been proposed to automatically identify product usage contexts from online customer reviews. The theoretical contributions of this paper are: (1) proposing grammatical rules, which are not specific to a particular product domain, to create a data set for training a sentence classifier (Sec. 3.2), (2) proposing a sentence classifier to obtain sentences that contain usage contexts (Sec. 3.3), and (3) identifying usage contexts from customer review sentences, as well as obtaining their corresponding aspect sentiments; even when the sentences may not contain either product-feature or sentiment words.

When the identified product usage contexts are complemented with aspect sentiment analysis, the interpretation of the results may be beneficial in several ways. For designers, the results may be used to evaluate the position of a product with respect to its competitors in different usage contexts, which enables the identification of product improvements and market opportunities. For customers, the results provides opportunity to filter products based on the sentiment toward their prioritized usage contexts. It is also shown that the overall rating is not strongly correlated with the sentiment toward individual usage contexts.

To address the strength of utilizing online reviews for identifying product usage context, the main benefits are the amount of data and its availability. Data may be obtained, analyzed, and interpreted in a time period that is faster than the required time to design, obtain approval, and conduct a survey-based method. Consequently, companies may be able to make faster decisions in many aspects, e.g., changing the advertisement strategy after learning about customers' usage contexts, deciding to improve the next generation product's performance in a particular usage context, etc. On the other hand, sentences in online reviews may not always provide detailed usage contexts. In contrast to survey-based methods, there are vague usage contexts that cannot be easily clarified or verified. For example, when a review states "doing research," it is hard to clarify the type of research activities. Also, when a review states that "photo editing is slow," it is hard to verify whether the laptop is actually incapable of performing the task or, for example, the user has not installed the software correctly. Moreover, the survey-based method may provide a higher granularity of the result. For example, the survey-based methods might reveal that people who complain about "photo editing" are mostly graphic designers.

For future work, the limitation of the proposed methodology in identifying usage contexts by using the filter of bigram and "-ing" suffix (Stage 4) may be improved. Other units of words (unigrams, trigrams, etc.), other patterns of bigrams, and the dependency relations of words in a sentence (e.g., the words that are connected by conjunctions) may also be considered as the bases to identify usage contexts that are currently not discovered by the proposed methodology. Obviously, there may be specific challenges in applying those filters; for example, the challenge with unigrams would be inferring the activity from a vague sentence that does not implicitly mention the activity, e.g., "good for video on youtube

every day"—which the activity is more likely to be inferred as watching video, instead of editing video.

Furthermore, for the future work, the main challenge would be extending the product usage context identification to identifying extraordinary usage contexts. Extraordinary usage contexts are important, since they are related to lead users, i.e., the customers that use a product in an extraordinary context such that they reveal latent needs that are crucial for product innovation [15]. The challenge lies in the fact that the frequency of these extraordinary contexts is generally very low. Therefore, it is challenging to identify them among a massive number of irrelevant terms that appear with low frequency as well. The usage context identification may also become the basis to construct a cross-product choice set in choice modeling, since a choice set may be formed by different types of items that serve the same usage intent [42]. Therefore, the proposed method may contribute to construct, for example, the set of devices (both laptops and tablets) that are compatible for the usage context of "web browsing."

References

[1] Green, M. G., Rajan, P. K. P., and Wood, K. L., 2004, "Product Usage Context: Improving Customer Needs Gathering and Design Target Setting," ASME 2004 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, Salt Lake City, UT, Sept. 28–Oct. 2, pp. 393–403.

[2] Green, M. G., Tan, J., Linsey, J. S., Seepersad, C. C., and Wood, K. L., 2005, "Effects of Product Usage Context on Consumer Product Preferences," ASME 2005 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, Long Beach, CA, Sept. 24–28, pp. 171–185.

[3] Kanis, H., 1998, "Usage Centred Research for Everyday Product Design," Appl. Ergonomics, **29**(1), pp. 75–82.

[4] Belk, R. W., 1975, "Situational Variables and Consumer Behavior," J. Consumer Res., **2**(3), pp. 157–164.

[5] Ram, S., and Jung, H.-S., 1991, "How Product Usage Influences Consumer Satisfaction," Mark. Lett., **2**(4), pp. 403–411.

[6] He, L., Chen, W., Hoyle, C., and Yannou, B., 2012, "Choice Modeling for Usage Context-Based Design," ASME J. Mech. Des., **134**(3), p. 031007.

[7] Ratneshwar, S., and Shocker, A. D., 1991, "Substitution in Use and the Role of Usage Context in Product Category Structures," J. Mark. Res., **28**(3), pp. 281–295.

[8] Green, M. G., Linsey, J. S., Seepersad, C. C., Wood, K. L., and Jensen, D. J., 2006, "Frontier Design: A Product Usage Context Method," ASME 2006 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, Philadelphia, PA, Sept. 10–13, pp. 99–113.

[9] Lim, S., and Tucker, C., 2017, "Mitigating Online Product Rating Biases Through the Discovery of Optimistic, Pessimistic, and Realistic Reviewers," J. Mech. Des. – Trans. ASME, **139**(11), p. 1114091.

[10] Jiang, H., Kwong, C. K., and Yung, K. L., 2017, "Predicting Future Importance of Product Features Based on Online Customer Reviews," ASME J. Mech. Des., **139**(11), p. 111413.

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

[12] LaFleur, R. S., 1992, "Principal Engineering Design Questions," Res. Eng. Des., **4**(2), pp. 89–100.

[13] Ram, S., and Jung, H.-S., 1990, "The Conceptualization and Measurement of Product Usage," J. Acad. Mark. Sci., **18**(1), pp. 67–76.

[14] Ghosh, D., Olewnik, A., and Lewis, K., 2017, "Application of Feature-Learning Methods Toward Product Usage Context Identification and Comfort Prediction," ASME J. Comput. Inf. Sci. Eng., **18**(1), p. 011004.

[15] Zhou, F., Jiao, R. J., and Linsey, J. S., 2015, "Latent Customer Needs Elicitation by Use Case Analogical Reasoning From Sentiment Analysis of Online Product Reviews," ASME J. Mech. Des., **137**(7), p. 071401.

[16] Lim, S., and Tucker, C. S., 2016, "A Bayesian Sampling Method for Product Feature Extraction From Large-Scale Textual Data," ASME J. Mech. Des., **138**(6), p. 061403.

[17] Shi, F., Chen, L., Han, J., and Childs, P., 2017, "A Data-Driven Text Mining and Semantic Network Analysis for Design Information Retrieval," ASME J. Mech. Des., **139**(11), p. 111402.

[18] Zhang, R., and Tran, T., 2011, "An Information Gain-Based Approach for Recommending Useful Product Reviews," Knowledge Inf. Syst., **26**(3), pp. 419–434.

[19] Zheng, X., Zhu, S., and Lin, Z., 2013, "Capturing the Essence of Word-of-Mouth for Social Commerce: Assessing the Quality of Online E-commerce Reviews by a Semi-Supervised Approach," Decision Support Syst., **56**, pp. 211–222.

[20] Qi, J., Zhang, Z., Jeon, S., and Zhou, Y., 2016, "Mining Customer Requirements From Online Reviews: A Product Improvement Perspective," Inf. Manage., **53**(8), pp. 951–963.

[21] Zhang, Z., Liu, L., Wei, W., Tao, F., Li, T., and Liu, A., 2017, "A Systematic Function Recommendation Process for Data-Driven Product and Service Design," ASME J. Mech. Des., **139**(11), p. 111404.

[22] Suryadi, D., and Kim, H., 2018, "A Systematic Methodology Based on Word Embedding for Identifying the Relation Between Online Customer Reviews and Sales Rank," ASME J. Mech. Des., **140**(12), p. 121403.

[23] Chiu, M.-C., and Lin, K.-Z., 2018, "Utilizing Text Mining and Kansei Engineering to Support Data-Driven Design Automation At Conceptual Design Stage," Adv. Eng. Inf., **38**, pp. 826–839.

[24] Jin, J., Liu, Y., Ji, P., and Kwong, C. K., 2018, "Review on Recent Advances in Information Mining From Big Consumer Opinion Data for Product Design," ASME J. Comput. Inf. Sci. Eng., **19**(1), p. 010801.

[25] Yang, B., Liu, Y., Liang, Y., and Tang, M., 2019, "Exploiting User Experience From Online Customer Reviews for Product Design," Int. J. Inf. Manage., **46**, pp. 173–186.

[26] Liu, B., and Zhang, L., 2012, *A Survey of Opinion Mining and Sentiment Analysis*, Springer US, Boston, MA, pp. 415–463.

[27] Cambria, E., Poria, S., Gelbukh, A., and Thelwall, M., 2017, "Sentiment Analysis is a Big Suitcase," IEEE Intell. Syst., **32**(6), pp. 74–80.

[28] He, R., Lee, W. S., Ng, H. T., and Dahlmeier, D., 2018, "Effective Attention Modeling for Aspect-Level Sentiment Classification," Proceedings of the 27th International Conference on Computational Linguistics, Association for Computational Linguistics, Santa Fe, NM, Aug. 20–26, pp. 1121–1131.

[29] He, R., Lee, W. S., Ng, H. T., and Dahlmeier, D., 2018, "Exploiting Document Knowledge for Aspect-Level Sentiment Classification," Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Association for Computational Linguistics, Melbourne, Australia, July 15–20, pp. 579–585.

[30] Gologlu, C., and Mizrak, C., 2011, "An Integrated Fuzzy Logic Approach to Customer-Oriented Product Design," J. Eng. Des., **22**(2), pp. 113–127.

[31] Culotta, A., and Sorensen, J., 2004, "Dependency Tree Kernels for Relation Extraction," Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics, ACL '04, Association for Computational Linguistics, Barcelona, Spain, July 21–26, pp. 423–429.

[32] 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.

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

[34] Santorini, B., 1990, "Part-Of-Speech Tagging Guidelines for the Penn Treebank Project (3rd revision)," University of Pennsylvania, Philadelphia, PA, Technical Report No. MS-CIS-90-47.

[35] Pepe, M. S., 2000, "Receiver Operating Characteristic Methodology," J. Am. Stat. Assoc., **95**(449), pp. 308–311.

[36] 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., Stanford, CA, June 29–July 2, pp. 727–734.

[37] Suryadi, D., and Kim, H., 2019, "Automatic Identification of Product Usage Contexts From Online Customer Reviews," Proceedings of the International Conference on Engineering Design, ICED, Delft, The Netherlands, Aug. 5–8, pp. 2507–2516.

[38] 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.

[39] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E., 2011, "Scikit-Learn: Machine Learning in Python," J. Mach. Learn. Res., **12**, pp. 2825–2830.

[40] 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

[41] Novikov, A., 2018, "Annoviko/Pyclustering: Pyclustering 0.8.1 Release," May. https://dx.doi.org/10.5281/zenodo.1254845

[42] 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," Mark. Lett., **2**(3), pp. 181–197.