

# **AN AUTOMATED METHOD TO CONDUCT IMPORTANCE-PERFORMANCE ANALYSIS OF PRODUCT ATTRIBUTES FROM ONLINE REVIEWS - AN EXTENSION WITH A CASE STUDY**

**Lin, Kangcheng;**  
**Kim, Harrison**

University of Illinois at Urbana-Champaign

## **ABSTRACT**

With the growth of online marketplaces and social media, product designers have been seeing an exponential growth of data available, which can serve as an extremely valuable source of information communicated from customers without geographical limitations. The data will reveal customers' preferences, which can be expensive and slow to obtain via traditional methods such as survey and questionnaires. While existing methods in the literature have been proposed to extract product information and make inference from online data, they have limitations, especially in providing reliable results and in dealing with data sparsity. Therefore, this paper proposes a method to conduct an Important-performance analysis from online reviews. The major steps of this method involve using latent Dirichlet allocation (LDA) to identify product attributes, using IBM Watson Natural Language Understanding tool to perform aspect-based sentiment analysis, and using XGBoost model to infer product attribute importance from the collected dataset. In our case study, we have collected over 150,000 text reviews of more than 3,000 laptops from Amazon.

**Keywords:** Big data, Semantic data processing, Market implications

## **Contact:**

Kim, Harrison  
University of Illinois at Urbana-Champaign  
Industrial and Enterprise Systems Engineering  
United States of America  
hmkim@uiuc.edu

**Cite this article:** Lin, K., Kim, H. (2021) 'An Automated Method to Conduct Importance-Performance Analysis of Product Attributes from Online Reviews - An Extension with a Case Study', in *Proceedings of the International Conference on Engineering Design (ICED21)*, Gothenburg, Sweden, 16-20 August 2021. DOI:10.1017/pds.2021.42

# 1 INTRODUCTION

One of the major reasons for a product to be successful is that it is designed to fulfill the needs of customers (Abbie Griffin, 1993). Product designers have long been eager to find a way to effectively comprehend the needs of different customers. One of the traditional methods includes inviting people to testing center and listening to their opinions via surveys or questionnaires. However, these methods are ineffective, unresponsive, and often biased due to temporal and geographical limitations. Nowadays, with the rise of online market places and social media, a considerable amount of attention has been attracted to the emerging voice of online customers, since these online platforms have soon become one of the major channels for sharing and storing large-scale first-hand opinions of customers. To tap into this valuable resource, an effective method is needed to process the large-scale dataset, to transform the gigantic amount of text reviews into formats that machine learning models can understand, and then to produce useful inferences for product designers. Therefore, this paper proposes a method, which serves as an extension of previous research (Joung and Kim, 2020) with additional case study, to conduct important-performance analysis (IPA) (John A. Martilla, 1977) from online reviews. The major steps of this method involve using latent Dirichlet allocation (LDA) (Blei et al., 2003) to identify product attributes, using IBM Watson Natural Language Understanding tool to perform aspect-based sentiment analysis, and using XGBoost model (Chen and Guestrin, 2016) to infer product attribute importance from the collected dataset. Furthermore, this method is intended to involve as little manual labor as possible. In our case study, we have collected over 150,000 text reviews of more than 3,000 laptops from Amazon.

IPA has been used to identify product attributes that product designers should focus on according to their associated importance and performance (John A. Martilla, 1977). This creates an effective system to distribute limited resources for customer satisfaction maximization. Based on their performance and importance levels, product attributes can be classified into four quadrants (Figure 1). According to each quadrant, product designers can plan different strategies. To elaborate, product attributes falling into Quadrant 1 (Q1) have high performance and high importance. These attributes represent the strength of studied products. Product attributes falling into Quadrant 2 (Q2) have low performance but high importance. These attributes require the utmost attention from product designers in order to maximize the customer satisfaction. Product attributes falling into Quadrant 3 (Q3) have low performance and low importance. These attributes should be at the bottom of product designers' priority list. Product attributes falling into Quadrant 4 (Q4) have high performance but low importance. These attributes may be the icing on the cake, but certainly not the top priority if resources are limited.

This paper is organized as follows. Section 2 presents a review of current literature in customer preference elicitation and product attribute extraction. Section 3 presents the proposed method of conducting IPA using online reviews. Section 4 demonstrates the validity of the proposed method using a case study of Amazon reviews of laptops. Section 5 concludes the paper with some thoughts about the direction of future studies.

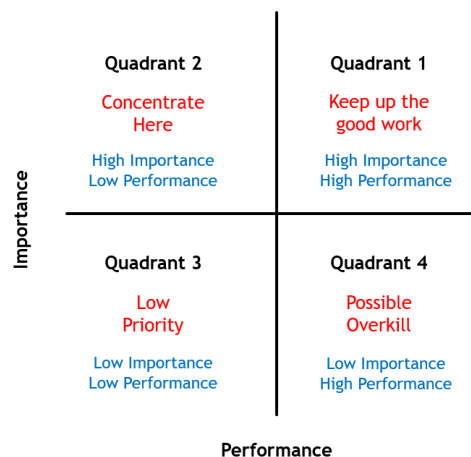


Figure 1. Importance performance analysis chart (Yuvina Tileng et al., 2013).

## 2 LITERATURE REVIEW

### 2.1 Customer Preference Elicitation

The tastes of customers are reflected from their interaction with product attributes (Chen et al., 2013a). Many traditional methods, such as survey, questionnaires, observations, have been proposed to study customers' interaction with product attributes (Zhou et al., 2012; Kinnear and Taylor, 1996). Unfortunately, these traditional methods are expensive to scale in size and contain inherent biases due to temporal and geographical limitations. Chen et al. (Chen et al., 2013b) proposed analytical discrete choice models to comprehend different customer tastes and predict their purchase decisions. Recently, more literatures have been using sentiment analysis to elicit customer preferences from online reviews. Zhang et al. (Zhang et al., 2016) proposed an opinion mining algorithm to capture the relationship of product attributes. Jiang et al. (Jiang et al., 2017) proposed a method to infer the future product feature importance using a fuzzy time series model. Zhou et al. (Zhou et al., 2015) proposed a method to identify latent customer needs with case analogical reasoning. Titov and McDonald (Ivan Titov, 2008) proposed a method to use latent Dirichlet allocation (LDA) to topic summary of product attributes. Suryadi and Kim (Suryadi and Kim, 2019) proposed a method to automatically obtain product usage context via machine learning models. Joung and Kim (Joung and Kim, 2020) proposed a method to apply information fusion-based SHapley Additive exPlanation (IFSHAP) in deep neural network to understand product attribute importance. Some of the traditional methods are easy to implement but suffer in accuracy and robustness against sparse datasets, whereas some of the recent methods, while able to achieve high accuracy, require intensive training and tuning to adjust to different datasets. XGBoost (Chen and Guestrin, 2016), on the other hand, has been famous for years for its effectiveness with sparsity and excellent accuracy in handling most kinds of datasets. Thus, this paper offers a flexible and effective method in understanding customer preference.

### 2.2 Product Attribute Extraction

A plethora of interests have been shown recently in product attribute extraction, as it serves to be one of the tasks of sentiment analysis of online product reviews. Ghani et al. (Ghani et al., 2006) proposed a text-mining method to extract attribute level pairs from textual product descriptions online. Hu and Liu (Hu and Liu, 2004) proposed a method to apply association rule mining (ARM) to extract frequent items in the product reviews as candidates for product attributes. Raju et al. (Raju et al., 2009) proposed an unsupervised method to extract product attributes from Amazon. Tucker and Kim (Tucker and Kim, 2011) proposed a method to detect statistical trend so as to classify product attributes based on customers' preferences. Wassenaar et al. (Wassenaar et al., 2005) proposed a discrete choice demand model to assess product profits. Stone and Choi (Stone and Choi, 2013) proposed a method to study validation and uncertainty analysis from user-generated content. Jeong et al. (Jeong et al., 2019) proposed an opportunity mining approach to uncover latent product topics discussed by customers in social media using topic modeling. While some of the traditional methods have worked well on small datasets and served as a constant source of inspiration for newer methods, they can hardly maintain their effectiveness on data of exponentially growing size. Furthermore, in order to develop a sophisticated model that is able to achieve state-of-the-art results on various forms of text, one needs to obtain and store the huge amount of information, and then possess the necessary expensive equipments (such as GPU and TPU) to train a model within reasonable time frame. The process can be extremely costly for an individual or even a small group of individuals to sustain. On the other hand, large companies, due to their financial capabilities and existing systems, are able to fully tap into the exponentially growing online data and to train complex models within a short amount of time. This paper uses IBM Watson Natural Language Understanding models to perform keyword extraction and aspect-based sentiment analysis, which are not only public available and free (for limited usage), but also ensure reliability and high accuracy, as the models are trained on a dataset of much larger scale.

## 3 METHODOLOGY

This paper designs a streamline process that converts text reviews into IPA of product attributes (Figure 2). Specifically, the input to the process is online product reviews in text, and the output is the IPA of product attributes. While some tuning is unavoidable, the proposed method is intended to be as

automatic as possible. The following subsections will provide a detailed elaboration of each step in the streamline process.

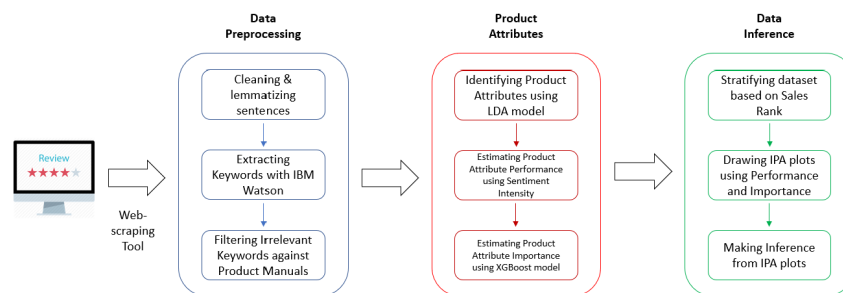


Figure 2. Streamline process.

### 3.1 Data Collection and Preprocessing

Product reviews are collected from online market places, such as Amazon and eBay. Then, information, such as product names, product prices, product sales rank, review title, review date, review content, and review rating, can be collected automatically and systematically using web-scraping tools readily available. For the scope of this paper, we conduct our analysis using English reviews from the USA only. After collection, we preprocess the text reviews by stripping off punctuations and converting all letters to lower-case (Denny and Spirling, 2018). Then, using IBM Watson Natural Language Understanding tool, we isolate nouns and noun phrases from the general text reviews, and obtain sentiment intensity corresponding to each noun and noun phrase.

However, the extracted nouns and noun phrases may sometimes be irrelevant to our product, or sometimes be too general and broad. For example, the following review, "my daughter was so happy with her laptop Christmas gift that she literally cried!", contains an irrelevant noun "daughter" and a too general noun "laptop". To ensure the validity and accuracy of our analysis, we need to filter these unwanted noun/noun phrases systematically. A solution this issue is to compare the collect nouns/noun phrases against product manuals or user guides of the corresponding product (Suryadi and Kim, 2019; Joung and Kim, 2020). Similar with product reviews, we preprocess each page of product manuals and isolate nouns/noun phrases using IBM Watson. Then, we perform a one-sample t-test (with  $\alpha = 0.10$ ) to identify the important nouns/noun phrases that appear frequently enough in each product manual. In particular, we compute the proportion of each noun/noun phrase. For example, if the total number of noun/noun phrases collected in a product manual is 1,000, and a particular noun appears 10 times, then its proportion is  $10/1000 = 1\%$ . Then, in the null hypothesis of the one-sample t-test, we assume that the proportion of a particular noun is equal to zero, whereas in the alternative hypothesis of the one-sample t-test, we assume that the proportion of that noun is strictly greater than zero. If a null hypothesis is rejected, then we will keep the noun/noun phrase and call it a product keyword.

Once a set of product keywords is collected, we will compare the extracted nouns/noun phrases of each review against the set. If a review does not contain any noun or noun phrase from the set of product keywords, the review will be dropped from the dataset. Ultimately, each text review will be turned into a list of tuples in the form of noun/noun phrase (called product keyword) and its corresponding sentiment intensity.

### 3.2 Product Attribute Identification

Once we obtain a list of product keywords, we use LDA to identify product attributes from them. LDA makes use of a generative statistical model that creates a list of common topics that categorize all product reviews (Blei et al., 2003). In particular, LDA assumes that each product review can be summarized by a set of topics in probability, and each topic in the set can be summarized by a set of (key)words in probability. In fact, each topic is summarized by the whole dictionary of keywords, each of which is assigned with a probability. Consequently, different topics will always contain the same set of keywords, but possibly with different probabilities assigned to each of the keywords. To more

accurately and succinctly comprehend each topic, we will select the top 30 keywords for each topic. In other words, each topic will be composed by a set of 30 keywords whose probabilities are among the largest, with respect to the particular topic, in the dictionary. Then, by looking into its composed keywords and into its typical reviews, we will interpret each topic by a label. These labels will be our product attributes. In the past, LDA has shown success in identifying product attributes from online reviews (Jeong et al., 2019; Zhou et al., 2019; Joung and Kim, 2020).

### 3.3 Product Attribute Performance Estimation

Recall from the data preprocessing step, we yield a list of tuples in the form of product keywords and their corresponding sentiment intensity. In particular, the sentiment intensity is obtained using the aspect-based sentiment analysis of IBM Watson, which will evaluate each keyword in a review in a range from  $-1$  to  $+1$ , such that the most negative sentiment will approach  $-1$  whereas the most positive sentiment will approach  $+1$ . Consequently, for each review, its associated tuples will generate performance estimation for product attributes. Specifically, for a review, we will look at each of its keywords and identify which product attribute (topic in LDA) it belongs to. If a keyword,  $w_i$ , belongs to a product attribute  $A_j$ , then the sentiment intensity  $s_i$  associated with the keyword  $w_i$  will be used to compute the performance of the product attribute,  $p_{ij}$ . If, in a particular review, we have multiple keywords,  $w_{i1}, w_{i2}, \dots, w_{in}$  belong to the same product attribute  $A_j$ , then the performance of  $A_j$  computed using this particular review will be  $p_{ij} = \frac{1}{n} \sum_{k=1}^n s_{ik}$ , where  $s_{ik}$  is the sentiment intensity associated with  $w_{ik}$  for  $k = 1, 2, 3, \dots, n$ . If a review does not contain any keyword related to a particular product attribute, then we will have the corresponding performance  $p_{ij} = 0$ . In addition, for the sake of better interpretability, we will discretize the performance scores in a way similar to 5-star ratings commonly used in reviews. This is because IBM Watson will provide similar yet different sentiment intensity scores to synonymous words (Joung and Kim, 2020). As a result, for each review, we will obtain a performance measure for each product attribute as follows:

$$p_{ij}^d = \begin{cases} 5, & \text{for } 0.8 < p_{ij} \leq 1 \\ 4, & \text{for } 0.6 < p_{ij} \leq 0.8 \\ 3, & \text{for } 0.4 < p_{ij} \leq 0.6 \\ 2, & \text{for } 0.2 < p_{ij} \leq 0.4 \\ 1, & \text{for } 0 \leq p_{ij} \leq 0.2 \\ -1, & \text{for } -0.2 \leq p_{ij} < 0 \\ -2, & \text{for } -0.4 \leq p_{ij} < -0.2 \\ -3, & \text{for } -0.6 \leq p_{ij} < -0.4 \\ -4, & \text{for } -0.8 \leq p_{ij} < -0.6 \\ -5, & \text{for } -1 \leq p_{ij} < -0.8 \\ 0, & \text{if the review contains no keyword associated with topic } j \end{cases}$$

where  $p_{ij}^d$  is the discrete performance score of product attribute  $A_j$  computed from the review  $i$ . Hence, we have transformed the enormous dataset of text reviews into a  $m \times n$  matrix  $(p_{ij}^d)_{i=1,2,\dots,m;j=1,2,\dots,n}$ , where  $m$  denotes the total number of reviews and  $n$  denotes the total number of product attributes identified from LDA in the previous subsection. This review-attribute matrix is an input format that can be understood by machine learning models. Lastly, we will compute the overall performance measure of product attribute  $A_j$  as follows:

$$\text{Performance}(A_j) = \frac{1}{\sum_{i=1}^m I_{ij}} \sum_{i=1}^m p_{ij}^d$$

where  $I_{ij}$  is an indicator function such that  $I_{ij} = 1$  if review  $i$  contains at least one keyword from  $A_j$  and  $I_{ij} = 0$  otherwise.

Table 1. A review-attribute matrix

Review No.	$A_1$	$A_2$	$A_3$	$A_4$	$A_5$	$A_6$	$A_7$	Label
1	0	0	0	5	0	0	0	Positive
2	5	0	0	0	0	5	5	Positive
3	0	-4	4	0	0	0	0	Negative
...	...	...	...	...	...	...	...	...
84046	-4	5	-4	0	5	-3	5	Positive

### 3.4 Product Attribute Importance Estimation

We will estimate the importance of each product attribute by training an XGBoost model (Chen and Guestrin, 2016) on the previous review-attribute matrix and computing the feature importance measures from the trained model. The nature of our dataset is mostly sparse; instead of touching on every single possible product attribute of a product, customers will most likely only mention product attributes (via product keywords) that are of their interests. Thus, the review-attribute matrix will contain a lot of 0's, which are used to denote missing performance measure. XGboost model is a scalable end-to-end tree boosting system that is able to achieve state-of-the-art results on sparse dataset (Chen and Guestrin, 2016).

Besides the input review-attribute matrix, which is usually denoted as  $X$ , we also need to decide the set of labels, which is usually denoted as  $y$ . In our study, we hope to investigate the relationship between product attribute and star ratings of each review. In Amazon, customers will provide a rating of 1-, 2-, 3-, 4-, 5- star alongside with their reviews. Consequently, by training a model using star ratings, we will be able to understand how each product attribute contributes in making a particular review 5-star or 1-star. Once again, for the sake of better interpretability, we combine 4- and 5-star ratings as one label (positive), 1- and 2-star ratings as one label (negative), and 3-star as one label (neutral).

In order to train an XGBoost model, we need to fine-tune some hyperparameters, such as learning rate, maximum depth, and number of estimators. To do so, we will apply k-fold cross-validation technique by partitioning the original dataset into  $k$  equal-size subset, training the model on the  $k - 1$  subsets, testing the model performance on the remaining 1 subset, and repeating the above process after permuting the subsets (Hastie et al., 2017). In our case, we set  $k = 5$ , which indicates 80% of the dataset will be used to train the model, whereas the remaining 20% of the dataset will be used to test the performance of the model. Then, in each cross validation, we will measure the performance of the model under a specific combination of hyperparameters. The set of hyperparameters that performs the best will be used in training the whole model. In Python, we may use the *GridsearchCV()* function in the *sklearn* package. Once the model is trained on the whole dataset, we will compute the product attribute importance using feature importance.

### 3.5 Importance-Performance Analysis

From previous subsections, we have obtained performance measure as well as importance measure of each product attribute. Then, we can draw an IPA plot by using performance measure as the x-axis and importance measure as the y-axis (Figure 1). Depending on which quadrant a particular product attribute falls into, product designers can gain insight into product design and plan different strategies to improve this product attribute.

## 4 CASE STUDY

To validate our method, we perform a case study using product reviews of laptops. In this case study, we are not only investigating the relationship between product attributes identified from product reviews, and star ratings, but also studying how different laptops are performing in different segmentation of sales rank.

### 4.1 Data Collection and Preprocessing

In this case study, we have collected 151,275 reviews, from 3,631 laptops at the (new) laptop category of Amazon. For the scope of this study, we focus only on English comments from the USA. In order to filter out irrelevant nouns/noun phrases extracted from reviews, we have collected 10 product manuals, whose

brands (models) include Acer, Alienware, Asus, HP, Lenovo, Samsung, Apple Macbook. From the product manuals, we have extracted 2,392 unique product keywords. In particular, we manually remove the word *computer* and brands/models associated with product manuals, because these words are either too general or providing little information related to product attributes. After comparing product reviews against the unique product keywords obtained from product manuals and preprocessing using IBM Watson Natural Language Understanding, we are left with 111,674 reviews, which include 1,144,576 tuples of keywords and corresponding sentiment intensity scores.

## 4.2 Product Attribute Identification

With 1,144,576 keywords, we have used the LDA model to identify product attributes from online reviews. By using coherence score as a measure of LDA model performance, we have extracted 7 topics from the corpus. For each topic, we label it with a product attribute by carefully interpreting the top-30 keywords and typical reviews related to the particular topic (Table 2). In the process, we have removed reviews whose keywords are not included in any of the top-30 keywords under each topic. In general, every single keyword should appear in every single topic generated by LDA model. However, since we are only considering the top-30 keywords, then some keywords will be "left out" as they are nowhere close to top-30 keywords in any of the topics. As a result, the number of reviews is finalized to be 84,046, with 1,102,849 product keywords.

Table 2. Product attributes and their associated keywords

Product Attribute	Top 10 Keywords	# Reviews
Utility	battery, life, hour, light, speaker, sound, charge, charger, brightness, audio	111,582
Hardware	screen, keyboard, touch, mouse, feature, side, touchpad, resolution, model, light	64,911
Software	window, power, key, system, purchase, button, program, software, web, office	58,673
Customer Service	time, product, problem, month, support, warranty, service, driver, box, customer	60,267
Work Compatibility	work, device, internet, update, app, tablet, need, connection, window, phone	43,783
Performance	use, machine, video, performance, issue, setting, pc, fan, mode, user	56,663
Drive and Connectivity	drive, ssd, port, usb, card, display, memory, storage, processor, space	55,777

## 4.3 Product Attribute Performance Estimation

In this part, we will use the sentiment intensity scores included in the remaining 84,046 tuples to estimate product attributes performance (Table 3). Moreover, we stratify the dataset according to their corresponding Amazon sales rank. Amazon sales rank is an indicator of how well a product is selling when compared to other products in the same category. A product's sales rank is assigned by a number, such that the smaller the number, the better the sales of the product (the top is number 1). In our final dataset, we have a list of sales rank ranging from number 9 to number 277,972. For our study, we will stratify the dataset by the median sales rank, which is number 12,492, so that those products whose sales ranks fall below 12,492 are considered "high", whereas those whose sales ranks go above 12,492 are considered "low". There are 1,756 laptops that do not have sales rank. According to Amazon, this is usually because the laptops are not sold enough. Thus, for this study, we will categorize them together with low sales rank laptops. As a result, we have 59,976 reviews under products with "high" sales rank, and 24,070 reviews under products with "low" sales rank.

Table 3. Product attribute performance estimation based on sales rank

Product	Perf( $A_1$ )	Perf( $A_2$ )	Perf( $A_3$ )	Perf( $A_4$ )	Perf( $A_5$ )	Perf( $A_6$ )	Perf( $A_7$ )
Full	0.381	0.049	0.292	-0.506	-0.050	0.330	0.034
High Sales Rank	0.513	0.165	0.459	-0.417	0.118	0.418	0.198
Low Sales Rank	0.066	-0.217	-0.129	-0.693	-0.484	0.128	-0.317

## 4.4 Product Attribute Importance Estimation

In order to train an XGBoost model, we input the review-attribute matrix generated in the previous subsection. The input matrix has a dimension  $84046 \times 7$ , and its entries  $p_{ij}^d \in$

$\{-5, -4, -3, -2, -1, 0, 1, 2, 3, 4, 5\}$ . In addition, to maximize the performance of the model, we have performed a grid search 5-fold cross validation on the hyperparameters, such as learning rate, maximum depth, and number of estimators, etc. The final model is able to achieve a 74.9% testing accuracy on the whole dataset, 75.8% testing accuracy on the dataset containing only the reviews from "high" sales rank products, and 72.1% testing accuracy on the dataset contain only the reviews from "low" sales rank products. For each of the final models, we will use its feature importance as a measure of product attribute importance estimation (Table 4).

Table 4. Product attribute importance estimation based on sales rank

Product	Imp( $A_1$ )	Imp( $A_2$ )	Imp( $A_3$ )	Imp( $A_4$ )	Imp( $A_5$ )	Imp( $A_6$ )	Imp( $A_7$ )
Full	0.197	0.116	0.135	0.300	0.084	0.085	0.083
High Sales Rank	0.192	0.119	0.134	0.278	0.092	0.095	0.089
Low Sales Rank	0.178	0.125	0.127	0.275	0.093	0.104	0.099

#### 4.5 Importance-Performance Analysis

Using the estimation of product attribute importance and performance, we are able to plot IPA plots using the full dataset, the dataset containing only high sales rank laptops, and the dataset containing only low sales rank laptops (Figure 3,4,5). In the IPA plot of the full dataset (Figure 3), we have observed that utility falls into Quadrant 1 (Q1), which means that on average the laptops in the collected dataset are doing well in terms of utility, which includes features such as battery life, working hour, speed of charging, etc. On the other hand, customer service seems to fall way into Quadrant 2 (Q2) on average. While this may not be the area that product designers are commonly working in, product manufacturers can use this information to build an edge over their competitors, as people purchasing laptops from Amazon do care about customer service, which includes time of delivery, warranty, customer support, repairs, etc. Hardware, work compatibility, and performance fall into Quadrant 3 (Q3). They are some of the product attributes that despite thier low performances, they appear not to be very crucial to customers' purchase decisions and feedback. Last but not least, drive and connectivity falls into Quadrant 4 (Q4). In general, laptops in the collected dataset seem to do quite well in this product attribute, which includes USB port, memory, ssd, etc. Nevertheless, according to the reviews, customers are not too concerned about this attribute.

There are some movement of the points on the IPA plots (Figure 4,5) of high sales rank dataset and low sales rank dataset. To begin with, we have observed that utility is consistently in the Q1, whereas customer service is consistently in the Q2. They seem to be on the polarized ends of the product attribute spectrum. Consequently, while product designers/manufacturers should maintain their standard in the utility product attribute, they may really want to investigate issues related to customer service. Hardware and performance move into Q4 in low sales rank dataset from Q3 in high sales rank dataset, whereas drive and connectivity goes in a reverse direction. These changes may define some characteristics of customers' expectation and needs when they are purchasing a laptop in the high sales rank or low sales rank.

### 5 CONCLUSION AND FUTURE WORK

This paper proposes a method to identify product attributes using latent Dirichlet allocation (LDA), investigate the processed data using XGBoost models, and gain insight using Importance-Performance Analysis (IPA). By using IBM Watson, we are able to perform the tasks of keyword extraction and sentiment analysis with much ease, reliability and accuracy. In addition, the use of XGBoost not only handles the issue of data sparsity, but also is able to achieve high accuracy with less effort and to adapt to different datasets.

This research has some limitations. Firstly, this research has only collected English comment from the USA. While English is the major language used in communicating ideas in North America, Spanish and French (in Canada) reviews should also be considered by product designers, as for some products, people from different ethnicities may have different needs. One of the challenges lies in clustering product keywords in other languages with those in English. Secondly, this research is only able to identify patterns in existing product attributes, but it is unable to excavate meaningful insight from unknown



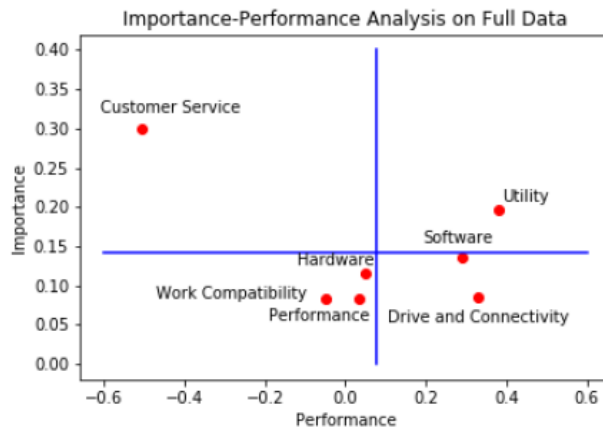


Figure 3. IPA on the full dataset.

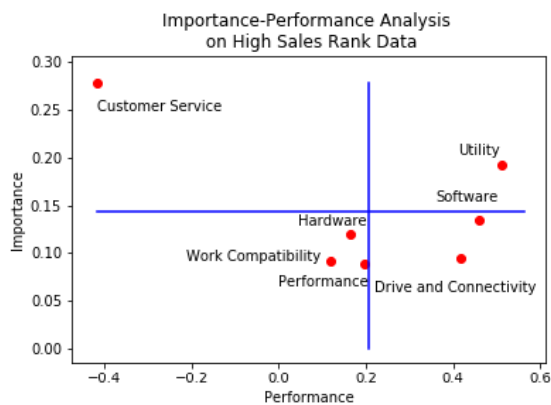


Figure 4. IPA on the high sales rank dataset.

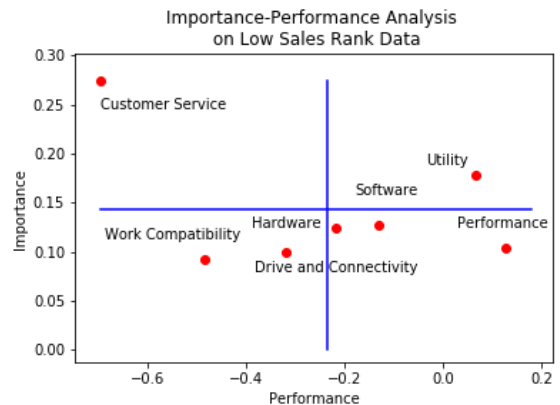


Figure 5. IPA on the low sales rank dataset.

product attributes. These unknown product attributes are usually brought up by lead users (Tuarob and Tucker, 2014), people who have special needs that are different from most of the people. Since the extracted keywords will be filtered against those from product manuals, and product manuals usually only contain existing product attributes, then unknown product attributes will be filtered out at the very beginning, resulting in a loss of valuable information and potential opportunity of innovation.

One of the directions of future work is to validate the veracity of collected online reviews. The proposed method is based on the assumption that all reviews are trustworthy, which is sometimes not true. It has been known that some companies would hire outsiders to fabricate fake reviews to either boost their own reputation or jeopardize their competitors'. While Amazon has a "Verified Purchase" label to indicate if a customer has actually purchased the product he or she is reviewing, it is still not robust enough to completely eradicate fake reviews. The existence of fake reviews will pollute the veracity of dataset, and thus invalidate the inference made out of it. In fact, it has been decades since researchers started investigating this problem (Jindal and Liu, 2008; Mukherjee et al., 2012). To enhance the robustness of this research, a system that can detect and filter fake reviews is needed. In addition, our proposed method has so far been only applied to the category of electronic products, but it has the potential to be extended to other domains, such as services and tourism. Hence, application studies in different domains will be conducted in future work.

## REFERENCES

- Abbie Griffin, J.R.H. (1993), "The voice of the customer", *Marketing Science*, Vol. 12.
- Blei, D., Ng, A. and Jordan, M. (2003), "Latent dirichlet allocation", *Journal of Machine Learning Research*, Vol. 3, pp. 993–1022.
- Chen, T. and Guestrin, C. (2016), "Xgboost: A scalable tree boosting system", pp. 785–794.
- Chen, W., Hoyle, C. and Wassenaar, H.J. (2013a), *Decision-Based Design Integrating Consumer Preferences into Engineering Design*, Springer.

- Chen, W., Hoyle, C. and Wassenaar, H.J. (2013b), *Decision-Based Design Integrating Consumer Preferences into Engineering Design*, Springer.
- Denny, M.J. and Spirling, A. (2018), “Text preprocessing for unsupervised learning: Why it matters, when it misleads, and what to do about it”, *Political Analysis*, Vol. 26, pp. 168–189.
- Ghani, R., Probst, K., Liu, Y., Krema, M. and Fano, A. (2006), “Text mining for product attribute extraction”, *ACM SIGKDD Explorations Newsletter*, Vol. 8, pp. 41–48.
- Hastie, T., Friedman, J. and Tibshirani, R. (2017), *The Elements of statistical learning: data mining, inference, and prediction*, Springer.
- Hu, M. and Liu, B. (2004), “Mining and summarizing customer reviews”, *Proceedings of the 2004 ACM SIGKDD international conference on Knowledge discovery and data mining - KDD 04*.
- Ivan Titov, R.M. (2008), “A joint model of text and aspect ratings for sentiment summarization”, .
- Jeong, B., Yoon, J. and Lee, J.M. (2019), “Social media mining for product planning: A product opportunity mining approach based on topic modeling and sentiment analysis”, *International Journal of Information Management*, Vol. 48, p. 280–290.
- Jiang, H., Kwong, C.K. and Yung, K.L. (2017), “Predicting future importance of product features based on online customer reviews”, *Journal of Mechanical Design*, Vol. 139, p. 111413.
- Jindal, N. and Liu, B. (2008), “Opinion spam and analysis”, in: *WSDM '08*.
- John A. Martilla, J.C.J. (1977), “Importance-performance analysis”, *Journal of Marketing*, Vol. 41.
- Joung, J. and Kim, H.M. (2020), “Importance-performance analysis of product attributes using explainable deep neural network from online reviews”, *Volume 11A: 46th Design Automation Conference (DAC)*.
- Kinney, T.C. and Taylor, J.R. (1996), *Marketing research: an applied approach*, McGraw-Hill.
- Mukherjee, A., Liu, B. and Glance, N. (2012), “Spotting fake reviewer groups in consumer reviews”, *WWW'12 - Proceedings of the 21st Annual Conference on World Wide Web*.
- Raju, S., Pingali, P. and Varma, V. (2009), “An unsupervised approach to product attribute extraction”, pp. 796–800.
- Stone, T. and Choi, S.K. (2013), “extracting consumer preference from user-generated content sources using classification”, Vol. Volume 3A: 39th Design Automation Conference of *International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*. V03AT03A031.
- Suryadi, D. and Kim, H. (2019), “Automatic identification of product usage contexts from online customer reviews”, *Proceedings of the Design Society: International Conference on Engineering Design*, Vol. 1 No. 1, p. 2507–2516.
- Tuarob, S. and Tucker, C.S. (2014), “Discovering next generation product innovations by identifying lead user preferences expressed through large scale social media data”, *Volume 1B: 34th Computers and Information in Engineering Conference*.
- Tucker, C.S. and Kim, H.M. (2011), “Trend mining for predictive product design”, *Journal of Mechanical Design*, Vol. 133 No. 11.
- Wassenaar, H., Chen, W., Cheng, J. and Sudjianto, A. (2005), “Enhancing discrete choice demand modeling for decision-based design”, *Journal of Mechanical Design - J MECH DESIGN*, Vol. 127.
- YuvinaTileng, M., Herry Utomo, W. and Latuperissa, R. (2013), “Analysis of service quality using servqual method and importance performance analysis (ipa) in population department, tomohon city”, *International Journal of Computer Applications*, Vol. 70, pp. 23–30.
- Zhang, H., Sekhari, A., Ouzrout, Y. and Bouras, A. (2016), “Jointly identifying opinion mining elements and fuzzy measurement of opinion intensity to analyze product features”, *Engineering Applications of Artificial Intelligence*, Vol. 47, p. 122–139.
- Zhou, F., Ayoub, J., Xu, Q. and Yang, X.J. (2019), “A machine learning approach to customer needs analysis for product ecosystems”, *Journal of Mechanical Design*, Vol. 142, p. 1.
- Zhou, F., Ji, Y. and Jiao, R.J. (2012), “Affective and cognitive design for mass personalization: status and prospect”, *Journal of Intelligent Manufacturing*, Vol. 24 No. 5, p. 1047–1069.
- 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”, *Journal of Mechanical Design*, Vol. 137 No. 7.