Expert Systems With Applications 144 (2020) 113111

Contents lists available at ScienceDirect

Expert Systems With Applications

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

Unusual customer response identification and visualization based on text mining and anomaly detection

Seungwan Seo[a], Deokseong Seo[b], Myeongjun Jang[c], Jaeyun Jeong[c], Pilsung Kang[a,*]

[a] School of Industrial Management Engineering, Korea University, Seoul, South Korea
[b] KEPCO, Data Science Lab, South Korea
[c] Data Machine Intelligence Group, AI center, SK telecom, South Korea

ARTICLE INFO

Article history:
Received 27 January 2019
Revised 29 August 2019
Accepted 28 November 2019
Available online 29 November 2019

Keywords:
Voice of customers
Keyword network
Local outlier factor
TF-IDF

ABSTRACT

The Vehicle Dependability Study (VDS) is a survey study on customer satisfaction for vehicles that have been sold for three years. VDS data analytics plays an important role in the vehicle development process because it can contribute to enhancing the brand image and sales of an automobile company by properly reflecting customer requirements retrieved from the analysis results when developing the vehicle's next model. Conventional approaches to analyzing the voice of customers (VOC) data, such as VDS, have focused on finding the mainstream of customer responses, many of which are already known to the enterprise. However, detecting and visualizing notable opinions from a large amount of VOC data are important in responding to customer complaints. In this study, we propose a framework for identifying unusual but significant customer responses and frequently used words therein based on distributed document representation, local outlier factor, and TF–IDF methods. We also propose a procedure that can provide useful information to vehicle engineers by visualizing the main results of the framework. This unusual customer response detection and visualization framework can accelerate the efficiency and effectiveness of many VOC data analytics.

1. Introduction

The identification of customer satisfaction determinants is an important concern for service providers and engineers (Matzler & Sauerwein, 2002). This can be conducted through a voice of customer (VOC) Griffin and Hauser (1993) analysis, which aims at understanding the needs of customers and transforming them into key functional requirements (Peng, Sun, Revankar, & Li, 2012). However, some challenges remain in VOC analysis owing to the lack of automated systems that can provide engineers with customers' needs (Aguwa, Monplaisir, & Turgut, 2012). Moreover, it cannot support engineers with any more than customers' overall sentiments or issues already known. There have been two main research directions for VOC analysis. The first is the sentiment analysis-based VOC analysis, which evaluates the positive or negative attitudes of individual consumers towards the product or service (Abbasi, Chen, & Salem, 2008; Abrahams, Jiao, Wang, & Fan, 2012; Kujiraoka, Saitoh, & Ishizu, 2017; Pang, Lee et al., 2008; Torizuka, Oi, Saitoh, & Ishizu, 2018). The main advantage of sentiment classification-based VOC analysis is that it can quickly and automatically determine whether the consumer considers the product positively, but it is difficult to extract particular satisfaction/dissatisfaction factors. The second research direction is to attempt to extract representative responses that are provided by the majority of consumers (Aguwa et al., 2012; Carulli, Bordegoni, & Cugini, 2013; Peng et al., 2012; Zondag & Ferrin, 2014). In this approach, unstructured text data are transformed into structured quantitative data to which various conventional analytical techniques are applied. However, many of the findings based on this approach may not be new but are already known to the company. In addition, individual consumers' fine nuances cannot be fully delivered.

To overcome the limitations of existing VOC analysis approaches, we propose a framework that can extract unusual but noteworthy customer responses automatically. In addition, we design two types of visualization methods and a query-based customer response retrieval system. The former shows N-step keywords and key phrases of customers' comments and the latter pro-

[*] Corresponding author.
E-mail addresses: zach0206@korea.ac.kr (S. Seo), douglas.seo@kepco.co.kr (D. Seo), mj.jang@sktair.com (M. Jang), jayhey@sk.com (J. Jeong), pilsung_kang@korea.ac.kr (P. Kang).

https://doi.org/10.1016/j.eswa.2019.113111
0957-4174/© 2019 Elsevier Ltd. All rights reserved.

vides full information of customers‘ needs to help product/service engineers better understand and digest the findings. In contrast to the answers to multiple-choice questions, the main data we focus on in this study come from a short essay—i.e., a sentence or paragraph, written by the survey respondents. Hence, a text analytics system is required to discover and interpret useful information therein (Bross & Ehrig, 2013). The bag-of-words (BOW) method has been widely used to transform an unstructured text document into a structured data format—i.e., a fixed-sized vector—based on the frequency of words appearing in a specific document and the entire corpus. Although BOW representation is simple to compute and intuitive to understand, it has a significant limitation: it cannot preserve the context because the semantic relationship between words is lost during the transformation (Li, Mei, Kweon, & Hua, 2011; Tirilly, Claveau, & Gros, 2008; Wallach, 2006; Wu, Hoi, & Yu, 2010). To solve this problem, we use a paragraph vector (Le & Mikolov, 2014) trained with a neural network, which can transform a variable length document to a fixed-sized continuous vector. In the second step, we use the local outlier factor (LOF), which is a well-known machine learning-based anomaly detection algorithm, to identify unusual customer responses. By considering both distance and density information around a target instance, LOF can make more flexible decisions than other parametric distance-based anomaly detection models (Breunig, Kriegel, Ng, & Sander, 2000). In the third step, we retrieve significant keywords from the unusual customer responses identified in the second step using the term frequency–inverse document frequency (TF–IDF) method (Salton & Buckley, 1988). These keywords are used to visualize co-occurrence-based and distance-based keyword networks and to extract the related unusual customer responses to help engineers develop engineering design improvements to resolve the issues raised by the customers. To verify the proposed framework, we conduct an experiment based on Vehicle Dependability Study (VDS) (Power, 2018) data, which are a real VOC dataset for a global vehicle company from J.D. Power and Associates.

The rest of this paper is organized as follows. In Section 2, we describe the VDS dataset and briefly review related studies focusing on VOC data analysis. In Section 3, the analytic process of the proposed framework is demonstrated, and the experimental design is described in Section 4. Finally, we conclude our study with some future research directions in Section 5.

2. Background

2.1. Vehicle dependability study (VDS)

Every year, USA-based J.D. Power and Associates reports the results of VDS, which surveys customers who have bought a new car within three years. The VDS is introduced on the official website of J.D. Power and Associates[1] as follows:

“The Vehicle Dependability Study, now in its 28th year, examines problems experienced during the past 12 months by original owners of 3-year-old vehicles. The study determines overall dependability by examining the number of problems experienced per 100 vehicles (PP100), with a lower score reflecting higher quality. The 2017 study examines cars, trucks, minivans, and SUVs from the 2014 model year and covers 177 specific problems grouped into eight major vehicle categories: Exterior, Engine/Transmission, Audio/Communication/Entertainment/Navigation (ACEN), Interior, The Driving Experience, Features/Controls/Displays (FCD), Heating/Ventilation/Air Conditioning (HVAC), Seats.”

In 2015, J.D. Power and Associates introduced a new evaluation system (VDS3) that increased the number of emotional items from 22 (11%) to 44 (25%) to better understand customers’ feelings. This suggests that it is increasingly important for vehicle companies to satisfy not only functional requirements but also emotional needs by understanding their consumers’ feelings about their products and services.

2.2. Related works

2.2.1. Significant customer opinion detection

The purpose of analyzing VOC is to understand customers’ subjective opinions toward a product or service and thus improve it by appropriately responding to the needs revealed from the customer responses. (Aguwa et al., 2012; Yang, 2008). Aguwa et al. (2012) developed a new customer satisfaction ratio (CSR) indicator that quantifies customer responses. The CSR indicator is evaluated based on four components: n, the total number of customer responses; a, the total number of positive responses; x_i, the rating crisp value per issue; and WC, the warranty cost. Peng et al. (2012) suggested a clustering-based methodology for finding representative customer responses. In their method, stop words are first removed from the customer responses, and significant keywords are extracted based on the TF–IDF method. The knowledge-based transformation model (Li, Ding, Zhang, & Shao, 2008) is then applied to transform the preprocessed customer responses to key features. Because the key features of customer responses vary over time, some studies have attempted to visualize the trends of key features and identify major changes after clustering VOC data at various times (Last, Klein, & Kandel, 2001; Zhang & Zhou, 2004). Because the above studies used the TF–IDF method to quantify the customer responses, they had a high risk of losing semantic relationship between words in the responses. To make matters worse, they focused on finding the representative opinions of the majority of customers. However, minor and unusual opinions are often more helpful to identify and solve significant problems before they are recognized by the majority of customers. Although Sezgen, Mason, and Mayer (2019) extracted important factors from airline reviews preserving semantic information using latent semantic analysis, they provide neither raw key phrases nor whole sentences.

2.2.2. Prediction of minority opinion using classification

Contrary to the aforementioned studies focusing on finding main themes from VOC, there have been some attempts to discover minor opinions from VOC (Amrit, Paauw, Aly, & Lavric, 2017; Wang & Xu, 2018). Amrit et al. (2017) extracted the summarizing features for classification after some text preprocessing such as stop word elimination, stemming, and tokenizing. They then trained classification algorithms such as naive Bayes (Kononenko, 1993), random forest (Breiman, 2001), and support vector machine (Cortes & Vapnik, 1995) classifiers to discriminate the minority opinions from the majority opinions. Wang and Xu (2018) combined the quantified responses such as multiple choices with the qualitative response, such as short essay answers, transformed by latent Dirichlet allocation (LDA) (Blei, Ng, & Jordan, 2003). A deep neural network–based classifier was trained to detect minor responses. These studies are similar to our study in that they attempted to find unusual customer responses. However, the usage of VOC data analysis based on classification approach is limited because it requires many labeled opinions to train classification models. Although labeled customer responses are available, classification models would suffer from the class imbalance problem; majority opinions significantly outnumber minority opinions (Ali, Shamsuddin, & Ralescu, 2015). In addition, because the ultimate goal of the above studies is minority opinion identification, the more practically significant question cannot be answered: what complaint fac-

[1] http://www.jdpower.com/cars/articles/jd-power-studies/infographic-key-stats-and-findings-2017-us-vehicle-dependability.

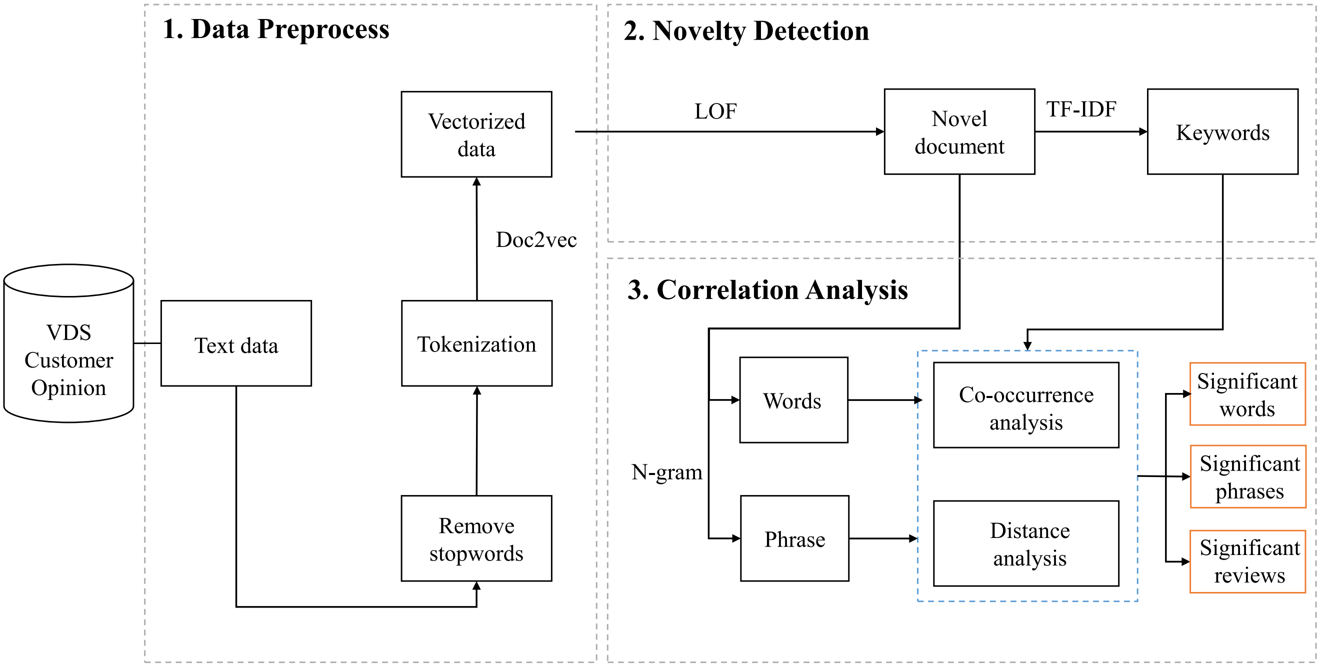


Fig. 1. Research framework.

tors or customer requirements are presented in the minority opinions?

Contrary to the previous studies, we attempt to identify significant customer opinions based not on classification algorithms but on an anomaly detection algorithm. By doing so, a large labeled dataset, which requires significant human labor costs, is no longer demanded, which in turn can be more practically applicable to various VOC datasets already possessed by companies. We then attempt to extract and visualize noticeable keywords and the related words/phrases. For engineers who want to read the full response including the keyword and its related words/phrases, the proposed framework can also provide the responses containing the query words or phrases. By reviewing the visualized keyword networks and the extracted full responses, engineers can deeply digest customer feedback and respond to it for the next product/service development or improvement process.

3. The proposed framework to identify and visualize unusual VOC

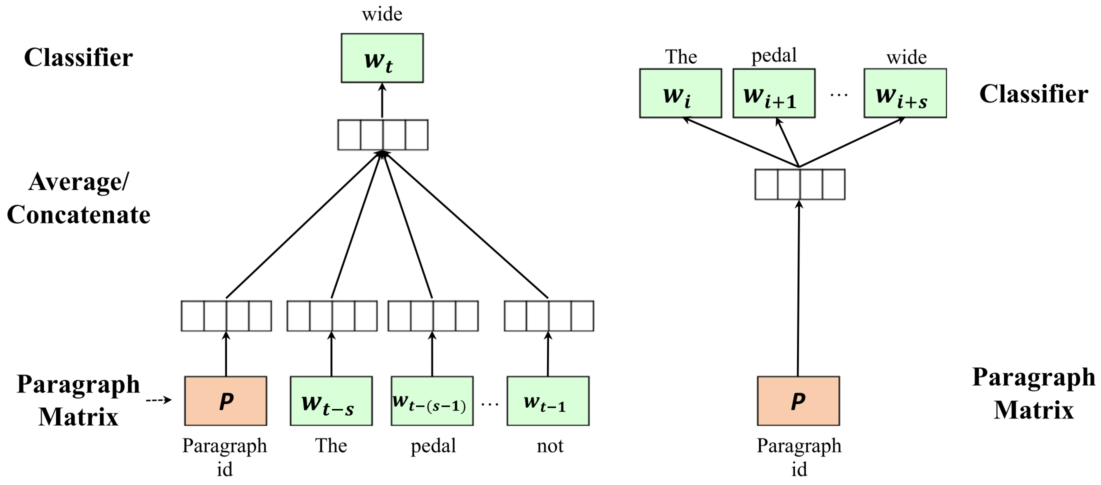
The purpose of this study is to discover unusual but noteworthy customer responses and extract significant words and phrases within them. Fig. 1 shows the overall framework of the proposed method. In the first phase, we transform unstructured customer responses to structured and quantified vectors. From the raw text data, we conduct text preprocessing including stop word removal and tokenizing. A neural network-based text embedding algorithm, named Doc2Vec, is then trained to vectorize each customer response (Le & Mikolov, 2014). In the second phase, the LOF is applied to identify responses that are unusual investigating. Based on TF–IDF analysis, significant keywords frequently used in the unusual customer responses but less frequently used in the majority of responses are extracted. In the last phase, closely related words and phrases to the extracted keywords are identified based on co-occurrence analysis and distance analysis in the embedding space. Once these related words and phrases are found, they are visualized by a network graph. In addition, they can be used as a search query to extract the full unusual customer responses in which the significant words or phrases are presented.

3.1. Paragraph vector

Preserving semantic relationship is important in understanding customer responses (Bansal & Srivastava, 2018; Zhang, Xu, Su, & Xu, 2015). Paragraph vector is a methodology to represent a variable length of text as a fixed-sized continuous vector. Although BOW representation can perform the same task, it loses semantic relations between words because it does not consider the word orders but simply counts their frequency. In contrast, paragraph vectors, are designed to preserve the semantic relationship by training the model using local neighborhood words. The simplest distributed representation for paragraphs is to use the average of the Word2Vec vectors of all words in the paragraph. However, Le and Mikolov (2014) showed that if the distributed representations of paragraphs are trained, they can better capture the contextual information, which, in turn, results in better performance in document processing tasks such as document classification. There are two main structures of the paragraph vector: paragraph vector with distributed memory (PV-DM) and paragraph vector with distributed BOW (PV-DBOW) as shown in Fig. 2 (Le & Mikolov, 2014).

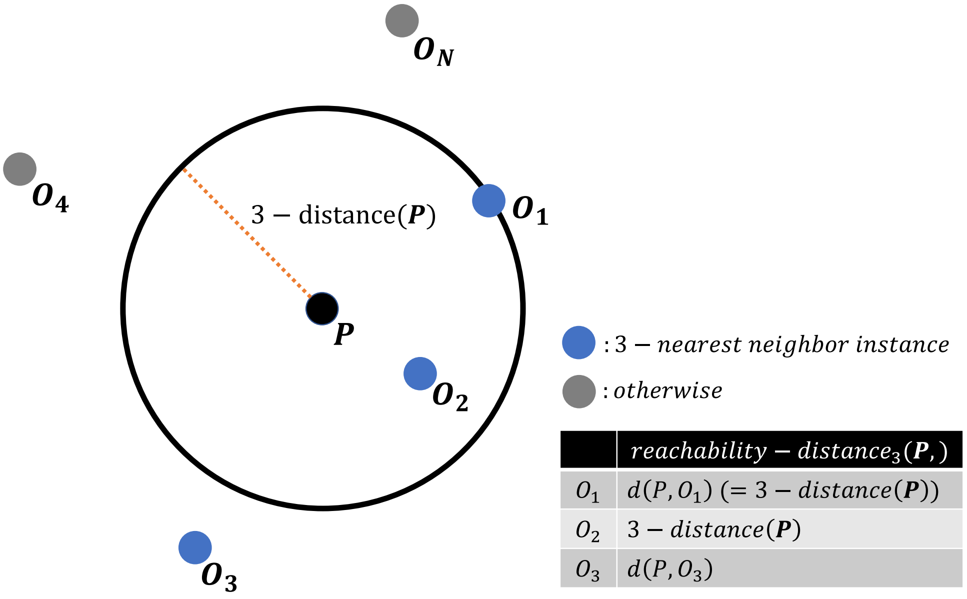
The PV-DM method is similar to the continuous bag-of-words (CBOW) for the Word2Vec model (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013). Le and Mikolov (2014) showed that PV-DM performed better than CBOW in various natural language processing tasks. The purpose of the PV-DM model is to predict the targeted word $\mathbf{w}_t$ using the previous s words $(\mathbf{w}_{t-s}, \mathbf{w}_{t-(s-1)}, \ldots, \mathbf{w}_{t-1})$ and the paragraph-token vector. s is the hyperparameter for the sliding window size. In PV-DBOW, conversely, the paragraph-token vector is the only input of the model, and s randomly selected sequential words $(\mathbf{w}_i, \mathbf{w}_{i+1}, \ldots, \mathbf{w}_{i+(s-1)})$ in the paragraph are used as the targets.

Although Le and Mikolov (2014) recommended using the concatenation of PV-DM and PV-DBOW in a complicated situation, they also showed that PV-DM performed well when used by itself in a less complicated case. In addition, we found that the results using PV-DM by itself were very similar to those obtained by using both PV-DM and PV-DBOW in our experiments. Hence, only



(a) PV-DM (b) PV-DBOW

Fig. 2. Two representative structures of paragraph vector.


O_N
O_4
$3-\text{distance}(\boldsymbol{P})$
O_1
$\boldsymbol{P}$
O_2
O_3
● : $3-$ *nearest neighbor instance*
● : *otherwise*

	$reachability - distance_3(\boldsymbol{P},)$
O_1	$d(P,O_1)$ $(= 3 - distance(\boldsymbol{P}))$
O_2	$3 - distance(\boldsymbol{P})$
O_3	$d(P,O_3)$

Fig. 3. Examples of reachability distance.

PV-DM was finally selected in this study to reduce computational costs.

3.2. Local outlier factor (LOF)

Local outlier factor (Breunig et al., 2000) is a well-known anomaly detection algorithm that takes both distance information and local density information to identify abnormal instances. Because it does not assume any distribution on normal instances and can generate arbitrary shapes of a normal class boundary, LOF has been successfully applied to various real-world anomaly detection tasks (Christou et al., 2011; Kang, Kim, & Kang, 2012; Lundström, Järpe, & Verikas, 2016). The first step of LOF is to define the k-distance of an object $\mathbf{p}$ as follows:

Definition 1. *k*-distance of an object **p**
For any positive integer k, the k-distance of object $\mathbf{p}$, denoted as k-distance($\mathbf{p}$), is defined as the distance $d(\mathbf{p}, \mathbf{o})$ between $\mathbf{p}$ and an object $\mathbf{o} \in$ D such that:

(1) for at least k objects $\mathbf{o}' \in D\backslash\{\mathbf{p}\}$ it holds that $d(\mathbf{p}, \mathbf{o}') \le d(\mathbf{p}, \mathbf{o})$,

(2) for at least $k-1$ objects $\mathbf{o}' \in D\backslash\{\mathbf{p}\}$ it holds that $d(\mathbf{p}, \mathbf{o}') < d(\mathbf{p}, \mathbf{o})$.

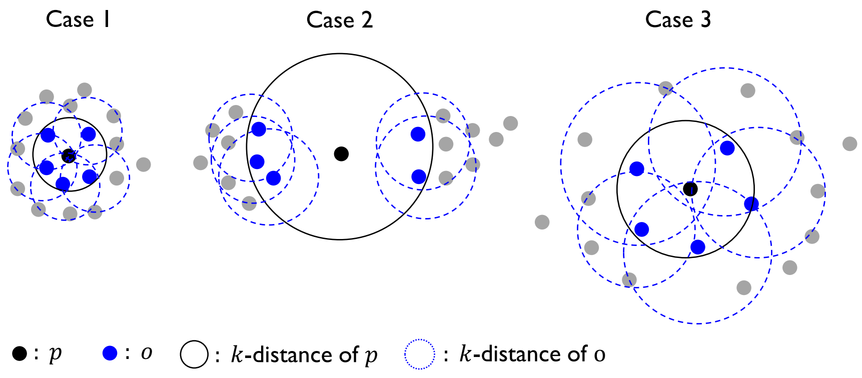
The k-distance of an object $\mathbf{p}$ is simply the distance to the k^{th} nearest neighbor instance considering ties. With the k-distance of an object $\mathbf{p}$, the k-distance neighborhood of an object $\mathbf{p}$ is defined as follows:

Definition 2. *k*-distance neighborhood of an object **p**
Given the k-distance of $\mathbf{p}$, the k-distance neighborhood of $\mathbf{p}$, denoted as $N_k(\mathbf{p})$, contains every object whose distance from $\mathbf{p}$ is not greater than the k-distance,

$$N_k(\mathbf{p}) = \{\mathbf{q} \in D\backslash\{\mathbf{p}\} | d(\mathbf{p}, \mathbf{q}) \le k\text{-distance}(\mathbf{p})\}. \tag{1}$$

These objects $\mathbf{q}$ are called the k-distance neighborhood of $\mathbf{p}$.

$N_k(\mathbf{p})$ contains a set of objects whose distances to $\mathbf{p}$ are shorter than or equal to the k-distance($\mathbf{p}$) around p. k is a user-specified hyperparameter often determined by the cross-validation process. The reachability distance is the larger value between the distance from the object $\mathbf{p}$ to $\mathbf{o}$ and the k-distance($\mathbf{p}$). Three different examples of the reachability distance are provided in Fig. 3.



Case	$lrd_k(p)$	$lrd_k(o)$	$LOF_k(p)$
Case 1	Large	Large	Small
Case 2	Small	Large	Large
Case 3	Small	Small	Small

Fig. 4. LOF score examples.

Definition 3. reachability distance of an object **o** from **p**

$$\text{reachability-distance}_k(p, o) = max(k\text{-distance}(\mathbf{p}), d(\mathbf{p}, \mathbf{o})). \quad (2)$$

Based on the reachability distance and $N_k(\mathbf{p})$, the k-distance neighborhood of **p**and the local reachability density of an object **p** are computed as follows:

Definition 4. local reachability density of an object **p**

$$lrd_k(\mathbf{p}) = \frac{|N_k(\mathbf{p})|}{\sum_{\mathbf{o}\in N_k(\mathbf{p})} \text{reachability-distance}_k(\mathbf{p}, \mathbf{o})}. \quad (3)$$

$lrd_k(\mathbf{p})$ computes the local density around the object **p**. If **p** is in the middle of a dense area, reachability-distance$_k$(**p, o**) becomes small, which in turn results in a large value of $lrd_k(\mathbf{p})$. Conversely, if **p** is in a sparse area, reachability-distance$_k$(**p, o**) becomes large, which in turn results in a small value of $lrd_k(\mathbf{p})$.

Finally, the anomaly score of an object **p** is computed as follows:

Definition 5. local outlier factor of an object p

$$LOF_k(\mathbf{p}) = \frac{\frac{1}{N_k(\mathbf{p})}\sum_{\mathbf{o}\in N_k(\mathbf{p})} lrd_k(\mathbf{o})}{lrd_k(\mathbf{p})} \quad (4)$$

The LOF score of objects **p** is inversely proportional to the local reachability density of **p** and is proportional to the local reachability density of the objects belonging to $N_k(\mathbf{p})$. Fig. 4 shows three examples of LOF score computation. All cases compute the LOF score of the black circle. In case 1, it is in the middle of a dense area, so both $lrd_k(\mathbf{p})$ and $lrd_k(\mathbf{o})$ have large values, which in turn results in a small LOF score. In case 3, the object is in the middle of a sparse area so that both $lrd_k(\mathbf{p})$ and $lrd_k(\mathbf{o})$ have small values, which in turn results in a small LOF score. Hence, the black objects in case 1 and case 3 are not considered as abnormal points. Conversely, in case 2, the black object is in the sparse region between two dense clusters. Hence, $lrd_k(\mathbf{p})$ is small but $lrd_k(\mathbf{o})$ is large, which in turn results in a large LOF score and is identified as an abnormal object. The only hyper-parameter of LOF is the number of local neighbors k, whose best value is very difficult to determine because the variant of an outlier cannot be precisely quantified (Ma, Ngan, & Liu, 2016). Hence, we determined it empirically based on the knowledge of domain experts. We provided some sampled unusual and normal customer responses extracted from different k values to expert engineers and asked them to choose the most appropriate one. As a result, k was set to 9 in our study.

3.3. Term frequency–inverse document frequency (TF–IDF)

Term frequency–inverse document frequency (Salton & Buckley, 1988) is the most widely adopted word weighting scheme in text mining. It computes how significant a word w is to a document d by combining two scores: term-frequency (TF), which is the frequency of w in d, and document frequency (DF), which is the number of documents in the corpus containing w regardless of its frequency. w is more important for d when its TF is large but DF is small. In other words, words frequently appearing in a small group of documents have higher TF–IDF scores than commonly used words such as prepositions and articles. The TF–IDF score of w for d is computed as follows:

$$\text{TF–IDF}(w, d, D) = f_{w,d} * log\frac{|D|}{f_{w,D}}, \quad (5)$$

where $f_{w,\,d}$ is the frequency of word w in document d, $|D|$ is the total number of documents in the corpus, and $f_{w,\,D}$ is the number of documents in the corpus containing the word w.

4. Experiments

4.1. Data exploration and preprocessing

In this study, we used the VDS dataset for a global vehicle manufacturing company. The surveys were conducted annually from 2014 to 2017. The questionnaire consists of multiple choices and short essays, the former of which were analyzed by various statistical analysis techniques, whereas the latter of which have not been fully utilized by the company because of the lack of text analytics expertise. As the real customer response data are not well organized in their initial state, we performed some data preprocessing, such as eliminating empty answers and duplicate responses. After preprocessing, the dataset consists of 82,059 responses. The number of responses for each vehicle type–functional category pair is summarized in Table 1.

4.2. Document embedding

The first step of the proposed framework is to transform the variable lengths of customer response texts to fixed-sized continuous vectors based on the paragraph vector method. To find the most suitable distributed representations for individual customer

Table 1
Number of customer responses for each vehicle type–functional category pair.

Vehicle type	Adjustments and controls	BlueLink Telematic System	Bluetooth	Navigation system	Ride quality	Seat belt	Seat material	Wind noise	Total responses
A	270	1	403	22	471	62	406	205	1840
B	139	210	495	743	275	83	103	114	2162
C	1288	276	3864	1909	3089	325	2093	1317	14,161
D	46	3	106	48	51	30	28	15	327
E	270	271	619	362	600	137	253	288	2800
F	83	60	171	541	70	11	122	51	1109
G	13	22	16	34	10	1	4	4	104
H	532	383	1454	3076	761	124	535	213	7,078
I	1777	1395	2520	3229	1351	322	1166	1313	13,073
J	2568	3515	7760	5090	3051	975	2938	1792	27,689
K	557	180	1490	1352	980	171	583	545	5858
L	294	885	1534	842	924	113	371	317	5280
M	43	0	143	182	55	9	84	61	577
Total	7880	7201	20,575	17,430	11,688	2363	8686	6235	82,058

Table 2
Paragraph vector examples in Adjustments and controls category.

Document index	dim_1	dim_2	$\cdots$	dim_{128}
12	-0.094	-0.073	$\cdots$	-0.003
26	0.087	-0.077	$\cdots$	0.098
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
7,880	0.143	0.01	$\cdots$	0.068

responses, we trained the PV-DM model from scratch based on the current database. To train the PV-DM model, we set the word window size = 5, train epoch = 100, and embedding dimension = 128. These hyper-parameters were optimized by a grid search by considering the training loss and the convergence rate. Once the PV-DM model is trained, each customer response is represented by a 128-dimensional vector as shown in Table 2.

4.3. Unusual document extraction

Once the customer responses are transformed to fixed-sized continuous vectors, LOF is applied to the embedded vectors to identify customer responses that are worth exploiting. In the VDS dataset, a significant number of responses are too short to be analyzed. "*Disappointed*," "*Its a whistle sound*," and "*It is a very stiff ride*" are some examples of very short and worthless comments for the following analysis. Hence, we remove the comments whose lengths are shorter than one-third of the total length distribution. The times taken for each category are shown in Table 3. Each cell in the last row represents the average time to execute LOF for 100 responses. In addition, to avoid anomaly score bias toward the response length when an unnormalized embedded vector is used—i.e., the longer the response, the higher the anomaly score—each response vector is normalized by its L_2 norm as follows:

$$|\mathbf{x}| = \sqrt{\sum_{i}^{n} x_i^2}, \tag{6}$$

where **x** is the embedded customer response vector.

Finally, the comments with the top 10% highest LOF scores are determined as unusual customer responses. Table 4 shows some response comments with the highest LOF scores for the three categories: Navigation system, Seat material, and Bluetooth.

4.4. Extracting keywords

Once the unusual response comments are determined, keywords for unusual comments are extracted as follows. Because the purpose of keyword extraction is to extract the words frequently used in the unusual responses but rarely used in the usual responses, we first extract the top 100 words with the highest TF–IDF scores from both the unusual comments (group A) and usual comments (group B). We then remove the overlapping words in both groups from group A to finally determine the keywords for the unusual response comments. Table 5 lists the keywords for each survey category.

4.5. Keyword-related word identification based on co-occurrence graph

To identify the important words highly related to the extracted keywords in the unusual customer responses, a word network is constructed based on the co-occurrence between the extracted keywords from the unusual customer responses and the other words in the whole corpus. For each keyword, the top five highly co-occurring words are defined as the first-degree neighborhood words. The second-degree neighborhood words of the keyword are then defined as the first-degree neighborhood words of the first-degree neighborhood words. An example of a word co-occurrence graph is shown in Fig. 5. The word "help" is one of the keywords for unusual customer responses in the Bluetooth category (Fig. 5 (c)), and it is represented by the yellow circle node. The keyword is connected to its first-degree neighborhood words, which are represented as gray square nodes, by red solid links: the thicker the link, the higher the co-occurrence. Similarly, the first-degree neighborhood words are connected to their first-degree neighborhood words, which are the second-degree neighborhood words of the original keyword and they are represented by green triangle nodes, connected by blue dashed links. The example of the Bluetooth category in Fig. 5 shows that the first-degree neighborhood words of the keyword "help" are "problem," "bluelink," "link," "blue," and "get," whereas its second-degree neighborhood words are "never," "remote," "app," "start," "car," "system," "dealership," and "would." Based on this keyword's co-occurrence network, it was found that some customers complained about malfunction of the Bluetooth app (bluelink) connection between their mobile phones and vehicles in the Bluetooth category.

4.6. Keyword-related phrase identification

The co-occurrence matrix for a keyword is a network of words related to the targeted keyword. In addition to individual significant words, it is worth finding customary expressions or phrases consisting of more than two words. Hence, we also generate a phrase graph to discover significant expressions related to the targeted keyword. To construct a phrase graph, bigrams (two consec-

Table 3
Average time (seconds) to compute the LOF scores for 100 customer responses for each category.

	Adjustments and controls	BlueLink Telematic System	Bluetooth	Navigation system	Ride quality	Seat belt	Seat material	Wind noise
Document volume	5274	4790	13,620	11,655	7810	1581	5843	4148
Time (second)	5.9026	4.7898	40.7533	38.1822	16.1882	0.4245	7.5141	3.2863

Table 4
Customer responses with the highest anomaly ranking (highest LOF scores) for the three categories. Numbers in the second column refer to the response's anomaly ranking and the total number of responses.

Response comments	Ranking
had to be reprogrammed to give expressway routes....stopped doing this w/o my making changes.	1/17,430
This problem has persisted the entire length I have owned the vehicle; however, it only happens on 1 out of 10 start ups so it is not something that causes much grief.	1/17,430
Also, bell comes at a specific place near my home which address is deleted.	1/17,430
Responses with the highest LOF scores for Navigation System	
Response comments	Ranking
I drove a Lexus for many years. The seats were soft cushioning and soft leather covered.	1/8,687
THIS IS MY SECOND HYUNDAI. SAME PROBLEM. ON 2012 VERA CRUZ.	1/8,687
I have not had any trouble with the seat covers but I do think that they should be treated with something before they are sold, that would make them easier to keep clean.	1/8,687
Responses with the highest LOF scores for Seat Material	
Response comments	Ranking
The app takes forever to load up, forever to send signal, and more often than not it's useless because my car is far from warmed up. Perhaps if I could set it to warm for longer than 10 min.	1/20,575
I ABSOLUTELY hate the fact that the hazards flash the entire time when remote starting the car. We live in NH, have had over 4 feet of snow this past month, temperatures haven't been above freezing in weeks and I still won't use the auto start because of the flashing lights. We will definitely be canceling the service because of this. And judging by all of the comments from other owners I read online, we are in the majority on this.	1/20,575
Whether connected to wiFi or mobile data, the app takes forever to log in (application to database coding flaws I'm sure) then, when in the app, when I press "lock" or "unlock" it takes a good 30 seconds to say "request sent" then, once the request is sent, it takes almost 3–5 minutes for the door to unlock! The app, by concept is brilliant, however the team working on the coding clearly is inexperience and doesn't know how to write scripts to compile and run efficiently!	1/20,575
Responses with the highest LOF scores for Bluetooth	

Table 5
Keywords that are more frequently used in unusual customer response comments for each category.

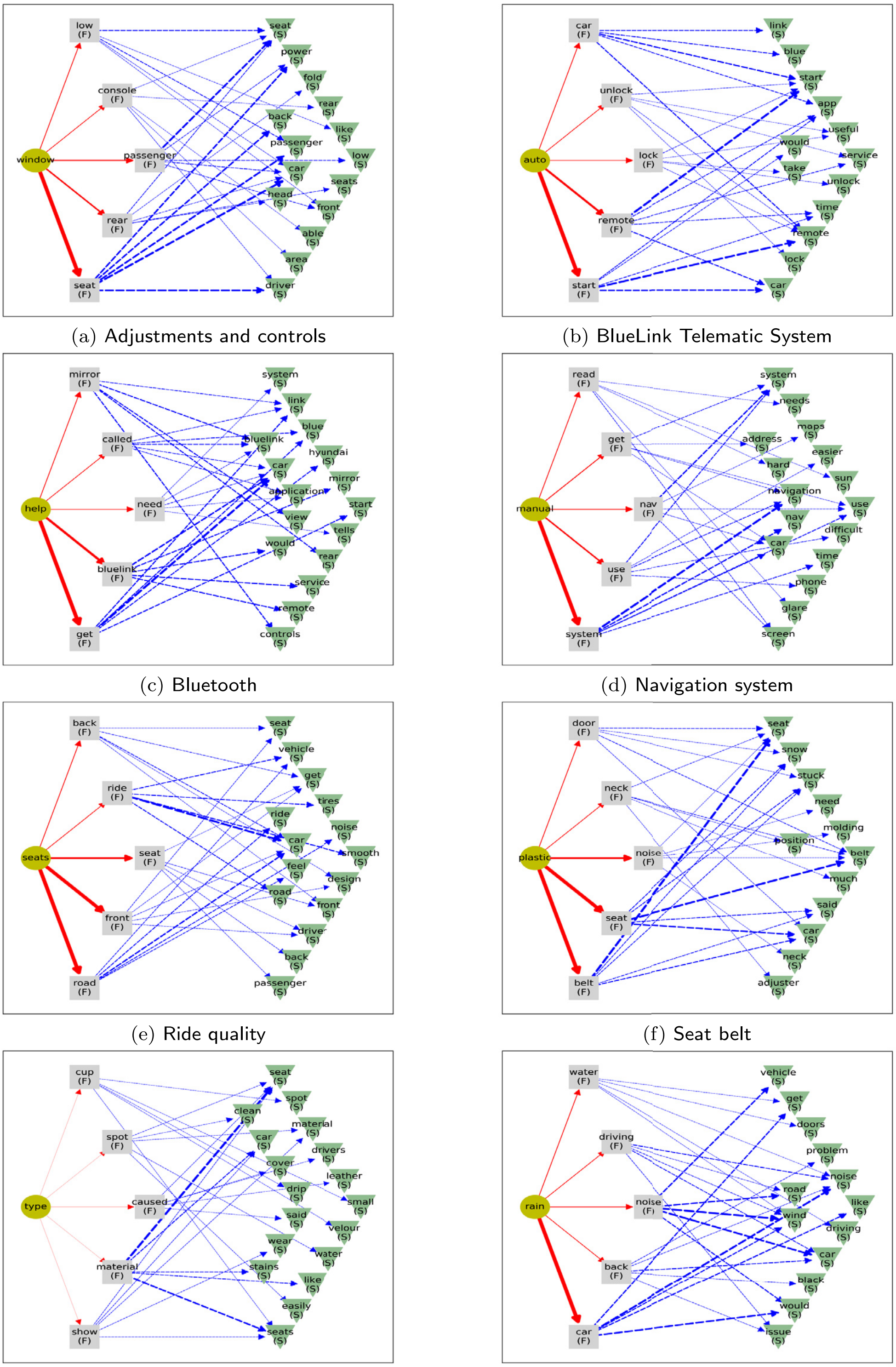
Category	Keywords
Adjustments and Controls	air, leather, window, cars, belt, setting, cooling, door, view
BlueLink Telematic System	money, information, area, month, right, auto, miles, oil, driving, user
Bluetooth	update, names, help
Navigation System	entering, iphone, dealership, locations, speed, manual
Ride Quality	engine, firm, genesis, body, driver, seats
Seat Belt	leg, extenders, thing, plastic, road, wheel, leather, arm, care, miles end, frame, years
Seat material	warranty, type, seating
Wind noise	bumper, seat, paint, water, trunk, mirror, rain, mileage

utive words), trigrams, and quadgrams are extracted from each response in the unusual response set. Among them, phrases composed of articles, conjunctions, pronouns, and special symbols are removed because they are not semantically noticeable expressions but only grammatically necessary. A phrase vector is then computed by the average of embedded word vectors constituting the phrase. Table 6 shows an example of phrase vector computation. Once the phrase vectors are embedded in the same space with the word vectors, the distances between the keyword and phrase vectors are computed to find semantically related phrases to the targeted keyword. Note that because the phrase vector is defined as the average of the word vectors constituting the phrase, the phrases including the keyword are obviously very closely located to the keyword. Hence, we ignore the phrases including the targeted keyword when finding the related phrases. Table 7 shows the inverse distance between the keyword "window" and the phrases in the Adjustment and controls category.

Table 6
An example of phrase vector calculation.

Uni-grams	d_1	d_2	d_3	$\cdots$	d_{128}
is	0.195	0.986	0.172	$\cdots$	0.653
not	0.389	0.868	0.251	$\cdots$	0.521
wide	0.399	0.931	0.235	$\cdots$	0.538
Mean(is, not, wide)	0.328	0.929	0.219	$\cdots$	0.571
		↓ Phrase vector			
N-grams	d_1	d_2	d_3	$\cdots$	d_{128}
is not wide	0.328	0.929	0.219	$\cdots$	0.571
The pedal	0.451	0.231	0.784	$\cdots$	0.612
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
Not wide	0.983	0.142	0.331	$\cdots$	0.972

In the Adjustments and controls category, the keyword "window" is closely related to the phrases "rear view," "rear view visibility," "block view driver," "rear headrests block view," and so on.



(a) Adjustments and controls

(b) BlueLink Telematic System

(c) Bluetooth

(d) Navigation system

(e) Ride quality

(f) Seat belt

(g) Seat material

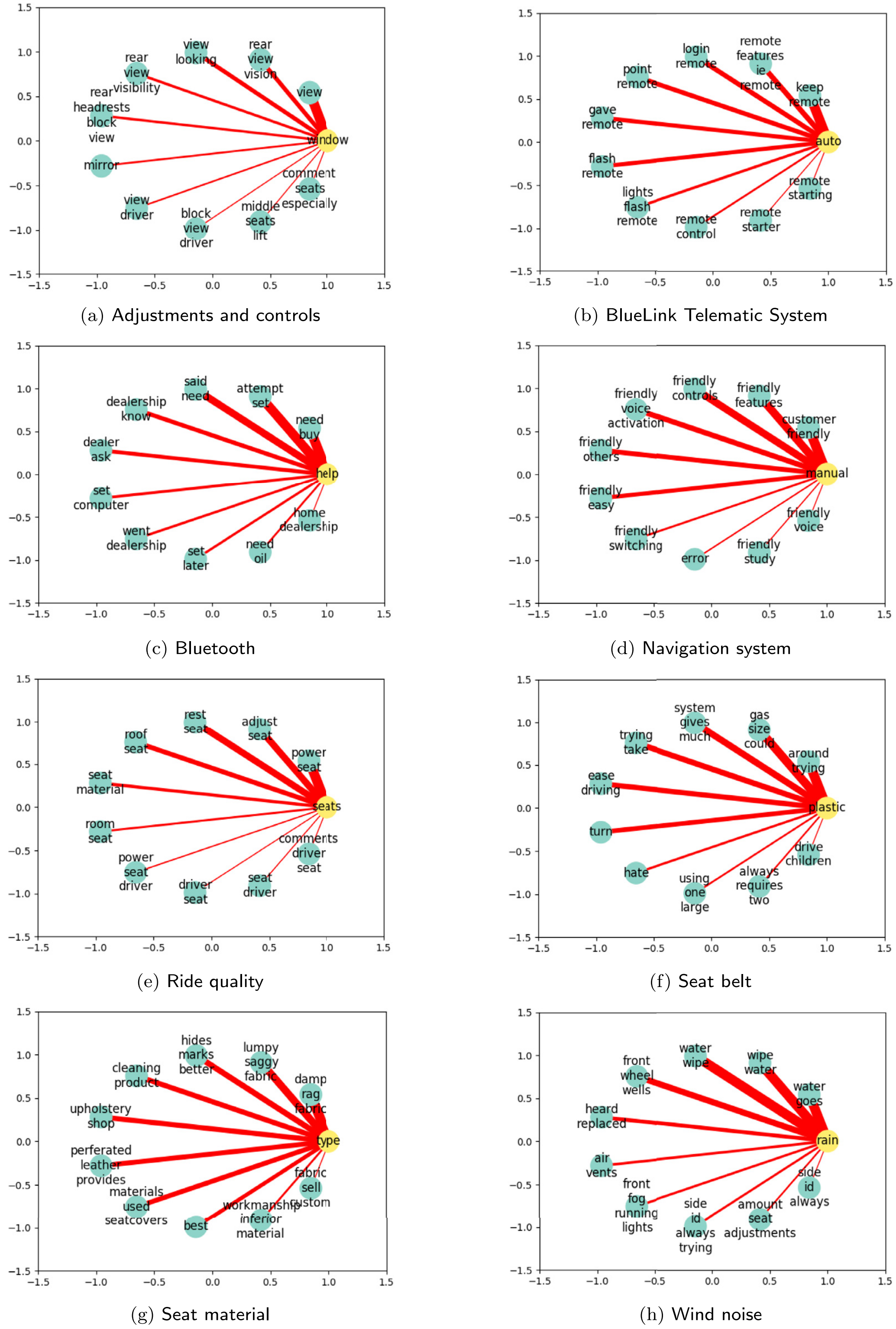
(h) Wind noise

Fig. 5. Co-occurrence graph examples.

Engineers can infer that there is a rear-view issue caused by the rear seat headrests. To visualize the key phrase identification results, a phrase graph is created by setting the top 10 phrases with the shortest Euclidean distance from the keyword as nodes and the inverse distance (similarity) as the edge weight: the thicker the weight, the more closely related the keyword. Examples of a phrase graph are shown in Fig. 6.

4.7. User query-based customer response extraction

Once the related words and phrases are identified, system users can acquire the full responses using the keyword and its related words or phrases. Tables 8 and 9 show some examples of the original customer responses with the keyword “window” and its related words or phrases. The only difference is that the keyword and its first- and second-order related words are used as the search query



(a) Adjustments and controls

(b) BlueLink Telematic System

(c) Bluetooth

(d) Navigation system

(e) Ride quality

(f) Seat belt

(g) Seat material

(h) Wind noise

Fig. 6. Co-occurrence graph examples.

Table 7
Related phrases for the keyword “window” in Adjustments and controls.

Phrase	Similarity	Phrase	Similarity
rear view vision	0.8393	view looking	0.8361
rear view visibility	0.7940	rear headrests block view	0.7895
view driver	0.7672	block view driver	0.7517
middle seats lift	0.7499	comment seats especially	0.7494

for Table 9, whereas a keyword and its related phrase are used as the search query for Table 8. Based on the word or phrase network shown in Figs. 5 and 6, system users can quickly and briefly understand the customer complaints, whereas more detailed nuances can be digested by reading the full responses as shown in Tables 8 and 9. Note that although Table 8 shows only one exam-

Table 8
Customer response examples including the keyword “window” in the Adjustment and controls category.

Keyword	Phrases	Responses
window	block view driver	Headrests won’t stay down, block view from driver’s seat. I HATE them and would remove them if I could.
window	view driver	The head restraints in back are too wide and block view for driver when looking out rear window.
window	rear view vision	The rear head restraints do not fold down easily to allow more driver vision through the rear view mirror.

Table 9
Responses including keyword “window” first-degree and second-degree words.

Query	Response
window, seats, back	The seats are very low, which makes visibility for kids in the passenger seat or back seats to see out of the side windows. This causes slot of carsickness and neck straining to see out the window. It s a little ridiculous esp for a vehicle meant for families with children. My kids hate it. My twelve year old has to sit on a cushion when in the passenger seat to see out.
window, rear, seats	I DISCOVERED A RIP IN THE CARPET JUST UNDER THE PASSENGER SEAT AND IT LOOKED LIKE IT WAS CUT WITH A KNIFE ON PURPOSE TO GET AT A GRAY PANEL UNDERNEATH THE CARPET. I REPAIRED IT WITH CONTACT CEMENT MYSELF BEFORE IT COULD BE A BIGGER PROBLEM. YOUR SHOP COULD NEVER FIGURE OUT HOW TO MAKE THE TRUNK COVER STAY ATTACHED TO THE BACK OF THE REAR SEATS AND JUST KEPT CHANGING THE MAT AND VELCRO THAT NEVER HELD FOR MORE THEN A DAY!. ALSO, WITHIN THE FIRST FEW MONTHS WE HAVE BEEN IN TWICE FOR A FAULTY WINDOW WASHER AND THE THIRD TIME THEY FINALLY REPLACED IT. I DON T TRUST THAT YOUR SHOP PEOPLE ARE VERY WELL TRAINED. WE WERE PROMISED A CARWASH EVERY TIME WE WAITED FOR REPAIRS BUT IT NEVER HAPPENED SO WE JUST LEFT EACH TIME NOT WANTING TO RETURN FOR ANYTHING. ALL AND ALL I BELIEVE WE WERE OVER CHARGED/TAXED WHEN WE BOUGHT THE CAR.

Table 10
The number of responses detected as unusual among sets of three fabricated responses.

	Adjustments and controls	BlueLink Telematic System	Bluetooth	Navigation system	Ride quality	Seat belt	Seat material	Wind noise	Total responses
Adjustments and controls	–	3	3	3	3	2	2	3	5,274
BlueLink Telematic System	3	–	3	3	3	3	3	3	4790
Bluetooth	2	3	–	3	3	3	3	3	13,620
Navigation system	3	3	3	–	3	3	3	2	11,655
Ride quality	3	3	3	3	–	2	3	2	7810
Seat belt	0	0	0	1	1	–	0	1	1581
Seat material	3	3	3	3	2	2	–	3	5843
Wind noise	3	0	0	2	1	1	1	–	4148

ple for each keyword–phrase pair, multiple responses can contain the keyword–phrase pair.

4.8. Validation of unusual customer response detection

Since explicit label information stating whether a given customer response is unusual or normal is absent, the use of classification-oriented performance measures such as accuracy, precision, recall, and area under receiver operating characteristic (AUROC) are not appropriate in validating our result (Goix, 2016). Hence, we validated the proposed method based on the two following indirect criteria: fabricated response insertion and statistical tests for average distance of each customer response from its neighbors, in the two groups being considered (unusual vs. common).

With regard to fabricated response insertion, we assumed that the responses in one category are highly likely to be unusual responses in other categories. Based on this assumption, for each category, we randomly selected three responses from every other category and added them to its customer response dataset. Hence, a total of 21 (3 responses from 7 categories) fabricated responses were created for each category and we checked whether they were identified as unusual for the target category by the proposed method. Table 10 shows the number of fabricated responses that were detected as unusual ones. Almost all fabricated responses were detected as unusual for all categories except *Seat belt* and *Wind noise*. These two categories have a relatively lower aggregate of customer responses than the other categories, and so some fabricated responses were missed since it might be difficult to generalize the boundary of common responses in their case.

With regard to the statistical test, we first computed the average distance of each customer response from its nine nearest neighbors (the same as k in the LOF algorithm). Since the purpose of Doc2Vec is to find an appropriate location for each document in a semantic space, the distance of a document from its neighbors is proportional to the extent to which they are semantically different from each other. Table 11 shows the average distances of both unusual responses (first row) and common responses (second row) from their neighbors, along with the corresponding standard deviations. The p-value of two-sample t-test is provided in the last row of the table. As we expected, unusual responses have greater average distances from their neighbors than common responses do from theirs. This difference is statistically supported at the significance level of 0.01 for all categories. Hence, we can conclude that the detected unusual responses are semantically more different from their neighbors than common responses are from theirs.

5. Conclusion

In this study, we proposed a VOC data analytics framework focusing on finding unusual customer responses from user-provided comments. We first transformed variable-length customer responses to fixed-size numerical vectors based on a neural network-

Table 11
The average distances of usual and unusual responses from their neighbors, along with the corresponding standard deviations. The *p*-value of statistical test is provided in the last row.

	Adjustments and controls	BlueLink Telematic System	Bluetooth	Navigation system	Ride quality	Seat belt	Seat material	Wind noise
Unusual responses	1.1957 (0.0063)	1.2088 (0.0067)	1.1934 (0.0124)	1.2069 (0.0042)	1.1821 (0.0088)	1.1896 (0.0209)	1.2262 (0.0009)	1.1636 (0.010)
Common responses	1.0574 (0.0120)	1.0759 (0.0394)	1.0441 (0.0326)	1.0789 (0.0231)	1.0229 (0.0673)	0.9875 (0.1093)	1.1424 (0.0144)	0.9417 (0.0381)
p-value	0.00001	0.00218	0.00036	0.00030	0.00804	0.01979	0.00004	0.00019

based paragraph vector model. LOF, a well-known anomaly detection algorithm, was then applied to identify unusual responses. To understand the customer requirements in the unusual responses, we extracted significant keywords based on TF–IDF analysis. For each keyword, related words with high co-occurrence and semantically related phrases were identified and visualized. Finally, we also provided the full response text when a system user provided a set of a keyword and related words/phrases to more deeply digest the nuances of individual responses.

Although the proposed framework was found to be effective to analyze the VOC dataset, there are some limitations of the current study, which lead us to future research directions. First, the effectiveness of the proposed framework would be enhanced if it could be quantitatively evaluated. To do so, a survey can be conducted for system users or product engineers on the helpfulness of the proposed framework compared with their ordinary job without such systems. Second, the proposed framework would be more meaningful if the results could be reflected during the product improvement or development process. This would help verify how many customer requirements identified by the proposed framework are actually reflected in the subsequent vehicle model.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Credit authorship contribution statement

Seungwan Seo: Methodology, Software, Validation, Writing - original draft, Visualization. **Deokseong Seo:** Software, Investigation, Resources, Data curation. **Myeongjun Jang:** Software, Methodology, Visualization. **Jaeyun Jeong:** Resources, Data curation. **Pilsung Kang:** Conceptualization, Writing - review & editing, Supervision, Funding acquisition.

Acknowlgedgments

This work was supported by the National Research Foundation of Korea (NRF) grants funded by the Korea government (MSIT) (No. NRF-2019R1F1A1060338) and Korea Electric Power Corporation (Grant number: R18XA05). Thanks to an anonymous reviewer for enabling quantitative validation of the proposed framework using fabrication data.

References

Abbasi, A., Chen, H., & Salem, A. (2008). Sentiment analysis in multiple languages: feature selection for opinion classification in web forums. *ACM Transactions on Information Systems (TOIS), 26*(3), 12.

Abrahams, A. S., Jiao, J., Wang, G. A., & Fan, W. (2012). Vehicle defect discovery from social media. *Decision Support Systems, 54*(1), 87–97.

Aguwa, C. C., Monplaisir, L., & Turgut, O. (2012). Voice of the customer: customer satisfaction ratio based analysis. *Expert Systems with Applications, 39*(11), 10112–10119.

Ali, A., Shamsuddin, S. M., & Ralescu, A. L. (2015). Classification with class imbalance problem: a review. *International Journal of Advances in Soft Computing Applications, 7*(3), 176–204.

Amrit, C., Paauw, T., Aly, R., & Lavric, M. (2017). Identifying child abuse through text mining and machine learning. *Expert Systems with Applications, 88*, 402–418.

Bansal, B., & Srivastava, S. (2018). Sentiment classification of online consumer reviews using word vector representations. *Procedia Computer Science, 132*, 1147–1153.

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

Breiman, L. (2001). Random forests. *Machine learning, 45*(1), 5–32.

Breunig, M. M., Kriegel, H.-P., Ng, R. T., & Sander, J. (2000). Lof: identifying density-based local outliers. In *Proceedings of the ACM SIGMOD record: 29* (pp. 93–104). ACM.

Bross, J., & Ehrig, H. (2013). Automatic construction of domain and aspect specific sentiment lexicons for customer review mining. In *Proceedings of the 22nd ACM international conference on conference on information & knowledge management* (pp. 1077–1086). ACM.

Carulli, M., Bordegoni, M., & Cugini, U. (2013). An approach for capturing the voice of the customer based on virtual prototyping. *Journal of Intelligent Manufacturing, 24*(5), 887–903.

Christou, I., Bakopoulos, M., Dimitriou, T., Amolochitis, E., Tsekeridou, S., & Dimitriadis, C. (2011). Detecting fraud in online games of chance and lotteries. *Expert Systems with Applications, 38*(10), 13158–13169. doi:10.1016/j.eswa.2011.04.124. http://www.sciencedirect.com/science/article/pii/S0957417411006518

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning, 20*(3), 273–297.

Goix, N. (2016). How to evaluate the quality of unsupervised anomaly detection algorithms?arXiv:1607.01152.

Griffin, A., & Hauser, J. R. (1993). The voice of the customer. *Marketing Science, 12*(1), 1–27.

Kang, B., Kim, D., & Kang, S.-H. (2012). Real-time business process monitoring method for prediction of abnormal termination using Knni-based LOF prediction. *Expert Systems with Applications, 39*(5), 6061–6068. doi:10.1016/j.eswa.2011.12.007. http://www.sciencedirect.com/science/article/pii/S0957417411016782

Kononenko, I. (1993). Inductive and Bayesian learning in medical diagnosis. *Applied Artificial Intelligence an International Journal, 7*(4), 317–337.

Kujiraoka, T., Saitoh, F., & Ishizu, S. (2017). Extraction of customer satisfaction topics regarding product delivery using non-negative matrix factorization. In *Proceedings of the IEEE international conference on industrial engineering and engineering management (IEEM)* (pp. 225–229). IEEE.

Last, M., Klein, Y., & Kandel, A. (2001). Knowledge discovery in time series databases. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 31*(1), 160–169.

Le, Q., & Mikolov, T. (2014). Distributed representations of sentences and documents. In *Proceedings of the international conference on machine learning* (pp. 1188–1196).

Li, T., Ding, C., Zhang, Y., & Shao, B. (2008). Knowledge transformation from word space to document space. In *Proceedings of the 31st annual international ACM SIGIR conference on research and development in information retrieval* (pp. 187–194). ACM.

Li, T., Mei, T., Kweon, I.-S., & Hua, X.-S. (2011). Contextual bag-of-words for visual categorization. *IEEE Transactions on Circuits and Systems for Video Technology, 21*(4), 381–392.

Lundström, J., Järpe, E., & Verikas, A. (2016). Detecting and exploring deviating behaviour of smart home residents. *Expert Systems with Applications, 55*, 429–440. doi:10.1016/j.eswa.2016.02.030. http://www.sciencedirect.com/science/article/pii/S0957417416300616

Ma, M. X., Ngan, H. Y., & Liu, W. (2016). Density-based outlier detection by local outlier factor on largescale traffic data. *Electronic Imaging, 2016*(14), 1–4.

Matzler, K., & Sauerwein, E. (2002). The factor structure of customer satisfaction: an empirical test of the importance grid and the penalty-reward-contrast analysis. *International Journal of Service Industry Management, 13*(4), 314–332.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems* (pp. 3111–3119).

Pang, B., Lee, L., et al. (2008). Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval, 2*(1–2), 1–135.

Peng, W., Sun, T., Revankar, S., & Li, T. (2012). Mining the “voice of the customer” for business prioritization. *ACM Transactions on Intelligent Systems and Technology (TIST), 3*(2), 38.

Power, J. D. (2018). *U.S. Vehicle Dependability Study (VDS)*. https://www.jdpower.com/business/resource/us-vehicle-dependability-study.

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management, 24*(5), 513–523.

Sezgen, E., Mason, K. J., & Mayer, R. (2019). Voice of airline passenger: a text mining approach to understand customer satisfaction. *Journal of Air Transport Management, 77*, 65–74.

Tirilly, P., Claveau, V., & Gros, P. (2008). Language modeling for bag-of-visual words image categorization. In *Proceedings of the international conference on content-based image and video retrieval* (pp. 249–258). ACM.

Torizuka, K., Oi, H., Saitoh, F., & Ishizu, S. (2018). Benefit segmentation of online customer reviews using random forest. In *Proceedings of the IEEE international conference on industrial engineering and engineering management (IEEM)* (pp. 487–491). IEEE.

Wallach, H. M. (2006). Topic modeling: beyond bag-of-words. In *Proceedings of the 23rd international conference on machine learning* (pp. 977–984). ACM.

Wang, Y., & Xu, W. (2018). Leveraging deep learning with lda-based text analytics to detect automobile insurance fraud. *Decision Support Systems, 105*, 87–95.

Wu, L., Hoi, S. C., & Yu, N. (2010). Semantics-preserving bag-of-words models and applications. *IEEE Transactions on Image Processing, 19*(7), 1908–1920.

Yang, K. (2008). *Voice of the Customer: Capture and Analysis (Six SIGMA Operational Methods)*. United States of America: The McGraw-Hill Companie. s.

Zhang, D., Xu, H., Su, Z., & Xu, Y. (2015). Chinese comments sentiment classification based on WORD2VEC and svmperf. *Expert Systems with Applications, 42*(4), 1857–1863.

Zhang, D., & Zhou, L. (2004). Discovering golden nuggets: data mining in financial application. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 34*(4), 513–522.

Zondag, M. M., & Ferrin, B. (2014). Finding the true voice of the customer in CPG supply chains: shopper-centric supply chain management. *Journal of Business Logistics, 35*(3), 268–274.