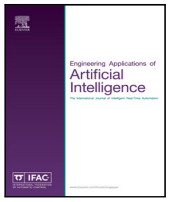
Engineering Applications of Artificial Intelligence 104 (2021) 104356

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

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



Customer sentiment analysis with more sensibility

Sunghong Park [a], Junhee Cho [b], Kanghee Park [c], Hyunjung Shin [a,b,*]

[a] Department of Artificial Intelligence, Ajou University, 206 Worldcup-ro, Yeongtong-gu, Suwon 16499, South Korea
[b] Department of Industrial Engineering, Ajou University, 206 Worldcup-ro, Yeongtong-gu, Suwon 16499, South Korea
[c] Division of Data Analysis, Korea Institute of Science and Technology Information, 66 Hoegi-ro, Dongdaemun-gu, Seoul 02456, South Korea

ARTICLE INFO

Keywords:
Customer review analysis
Sentiment analysis
Semi-supervised learning
Word embedding
Topic modeling

ABSTRACT

Customers' evaluations on products can be derived by analyzing online reviews using machine learning. Positive or negative responses can be sensed by words they write in reviews, and topics they compliment or complain about can be grasped by clustering reviews. Combination of those results is regarded as the customers' sentiment analysis. When reviews are given as free-form text without scores, general-purpose dictionaries are used to recognize sentiment words. However, customers do not only use standard words to express their emotions, but they also use non-grammatical language such as internet jargon. Unfortunately, existing methods cannot capture those sentiment words. Moreover, combination of sentiment words with customer topics simply represents frequencies and does not indicate detailed evaluation patterns. In this study, we propose a customer sentiment analysis method consisting of sentiment propagation and customer review analysis. It works more sensibly by expanding sentiment words from dictionary to those varying words as mentioned above. To implement this, semi-supervised learning is employed to a word graph that is constructed by a word embedding algorithm. Using this more sensible word graph, customer review analysis is conducted. Reviews are grouped into major complaint topics. Meanwhile, an index for customer dissatisfaction is designed by composition of 'controversy' and 'complaint'. The former stands for 'coverage of dissatisfaction' while the latter indicates 'degree of dissatisfaction'. The proposed method was applied to 3,11,550 reviews across five automobiles from ten internet communities. Case study illustrates which parts of automobiles lead to customer dissatisfaction, and therefore where investment and examination are required.

1. Introduction

Customers write online reviews regarding their experience with purchasing and using a product (Chung and Tseng, 2012; Hou et al., 2019). From the quality of a product to its follow-up services, customers evaluate all parts of a product in detail (Moghaddam and Ester, 2012). As the numbers of Internet communities and review websites have increased, a vast amount of online reviews has been accumulated in various domains. Hence, it has become easier to identify customers' evaluations by analyzing reviews through text mining methods. For this purpose, sentiment analysis and topic modeling are typically utilized. Sentiment analysis classifies words into positive or negative by focusing on emotions in the text (Liu and Zhang, 2012), and topic modeling clusters customer reviews into major issues using probabilistic models. Those results are combined and then regarded as customers' sentiment analysis. To implement this process accurately and efficiently, each method has the important consideration. In sentiment analysis, the construction of a lexicon is the most important task in that the lexicon serves as the criteria for the words' sentiment classification (Tang et al., 2009). Existing methods for constructing lexicons are divided into two categories according to the types of reviews; scored reviews and free-form reviews.

In case of sentiment analysis for scored reviews, each review has a rating with scores or stars which indicate the level of customer evaluation. By utilizing score information as labeled data, various supervised-learning algorithms are applied to sentiment analysis for scored reviews. Those analyses classify words' sentiments more accurately and efficiently than human baselines in various domains, such as movies (Chaovalit and Zhou, 2005; Pang et al., 2002) and automobiles (Gamon et al., 2005). However, the performance of existing methods is not guaranteed when they are applied to free-form reviews that are not in the form of scores or Likert scales. This limitation is even more important because the amount of data in free-form reviews is much more than that of scored reviews. In the case of sentiment analysis for free-form reviews, the words' sentiment classification is generally performed by referring to a general-purpose sentiment dictionary. During the algorithm learning process, the cited dictionary is utilized as label information. Many sources are typically used, including

* Corresponding author at: Department of Industrial Engineering, Ajou University, 206 Worldcup-ro, Yeongtong-gu, Suwon 16499, South Korea.
E-mail addresses: pshong513@ajou.ac.kr (S. Park), whwnsgml48@ajou.ac.kr (J. Cho), can17@kisti.re.kr (K. Park), shin@ajou.ac.kr (H. Shin).
URL: http://www.alphaminers.net (H. Shin).

https://doi.org/10.1016/j.engappai.2021.104356
Received 9 November 2020; Received in revised form 27 May 2021; Accepted 16 June 2021
Available online 7 July 2021
0952-1976/

SentiWordNet (Baccianella et al., 2010) and the Dictionary of Affect in Language (DAL) (Whissell, 1989). Therefore, as they are not limited to the existence of scores in reviews, the existing methods used for free-form reviews have more flexible ranges of analysis (Yi et al., 2003).

However, sentiment analysis of free-form reviews has less sensible aspects. Due to the application of already constructed sentiment dictionaries, the domain properties of reviews cannot be sufficiently reflected during the words' sentiment classification. Consider the words with positive and negative sentiments that often appear in ratings of customer reviews. For example, as shown in Fig. 1, suppose that seven words are given: 'terrific', 'good', 'interesting', 'acceptable', tolerable', 'bad', and 'terrible'. In this example, 'terrific' and 'good' are clearly positive words, while 'bad' and 'terrible' are clearly negative words. However, 'interesting', 'acceptable', and 'tolerable' are somewhat intermediate, and their sentiments may differ depending on context. To evaluate a student's report, if a teacher says 'interesting', it may mean 'it is (almost) excellent' or it may stand for 'it is only average or below average (but I do not want to hurt you).'. To determine the true meaning of a teacher's use of 'interesting' (https://dictionary.cambridge.org/dictionary/english/interesting), one possible method involves looking into what teacher has said in the past when evaluating students' work. Consequently, implicit sentiments of words for positive and negative may be different domain-specifically and contextually.

There is another less sensible aspect in topic modeling. Topic modeling clusters reviews into customers' major issues using probabilistic models. Then, sentiment words are applied to the customers' topics. However, this combination only represents simple frequencies, and does not capture detailed patterns of customers' evaluation. Most free-form reviews are written when customers face some inconvenience or have questions about a product (Hennig-Thurau et al., 2004; Sen and Lerman, 2007; Sparks et al., 2016). Therefore, a company assesses customers' reviews to observe their needs and responses to the company's product. By some reviews, a part of a product can be fixed, replaced, or renewed. And consequently, it leads to quality improvement of the entire product. It may be efficient to raise customers' satisfaction if the part is preferentially investigated and improved which is related to highly controversial and strongly complained topics in reviews. Note that in reviews, people seldom directly use the exact name of the part in question because they do not know what it is or because they simply omit it. Consequently, major topics are firstly exploited from review documents, and later they are related to the corresponding parts of a product.

In this study, we propose a machine learning-based customer sentiment analysis method with more sensibility. As shown in Fig. 1, the proposed method includes two processes: sentiment propagation and customer review analysis. First, sentiment propagation expands and refines words' sentiments reflecting their contextual usages in real domains. Through this process, words in reviews are classified into positive or negative sentiments. Next, we perform customer review analysis based on sentiment propagation. As a result, three indices are derived: controversy, complaint, and dissatisfaction. These indices focus on negative feedback, taking into account that voice of customer (VOC) represents more complaints than praise (Fornell and Wernerfelt, 1987; Pyon et al., 2010). In particular, the property of VOC appears more readily apparent in the online communities, and as a result, most opinions on the product appear negatively in the form of texts about malfunctions, complaints, and questions (Cho et al., 2002; Coussement and Van den Poel, 2008). The proposed indices are designed to well reflect these properties by indicating the intensity and range of how much more negative feedbacks stand out than positive feedbacks. Controversy indicates "how often are the topics raised?". If a certain topic has a high degree of controversy, it means that the topic is concerned by most reviewers or people. Complaint indicates "how severely is the topic complained about?". If a certain topic is strongly negatively complained about relative to other topics in review documents, it has to be seriously regarded, since customers' dissatisfaction is mostly likely to be incurred by it. The combination of controversy and complaint can be used as a measure indicating the degree of negative customer evaluation over products or services; this composite index is denoted as dissatisfaction. These indices can inform the priority for all parts of a product and the customer evaluation patterns.

2. Proposed method

2.1. Sentiment propagation

2.1.1. Word graph construction

A graph is a useful method for representing relationships between words. It is relatively simple to represent a graph of the similarity between words in the lexicographic sense. However, in a certain context of real-world, if two words are frequently used and replaceable, there may be similarities between them. To reflect the contextual semantic meaning of a word in similarity measurement, the meaning of the word in the context must first be inferred from a large amount of text data. To do this, Word2Vec (Mikolov et al., 2013a) can be used to vectorize words and numerically measure similarities between words. In this word graph, words that are similar not only 'lexicographically' but also 'contextually' are closely connected.

Word2Vec: Every word is quantified as a vector with the skip-gram model of Word2Vec, which is a neural network-based word embedding method. In the skip-gram model, embedding utilizes the context of words and the surrounding words (Mikolov et al., 2013b). First, as shown in Fig. 2(a), each word is vectorized by one-hot-encoding, and a neural network with one hidden layer h is constructed. The number of units in the hidden layer d indicates the number of dimensions to be vectorized. Then, the ith word w_i is placed in the input layer, and the surrounding $2k$ words from w_{i-k} to w_{i+k} are placed in the output layer, where k is the window size as a model hyperparameter. The weight parameter $P_{(|W|,d)}$ connects each word of the input layer and the hidden layer, and this parameter is shared during the learning process. After the learning, $P_{(|W|,d)}$ is derived as the result of Word2Vec. Fig. 2(b) shows embedded vectors of words in the toy example in Fig. 1. In this figure, vectors have similar patterns when the words involved are contextually similar.

Word graph: From the embedded word vectors, a word graph $G = (W, S)$ is constructed. The node set W corresponds to words and the edge set S corresponds to the similarity. The similarity is calculated by the Gaussian function in (1), where s_{ij} indicates the similarity between word vectors w_i and w_j, respectively corresponding to the ith word and the jth word. The similarity has the range from 0 to 1, where a larger value indicates more similarity in meaning and a higher frequency of co-occurrence. Therefore, the word graph quantifies domain characteristics in that similar words are more strongly connected.

$$s_{ij} = \begin{cases} exp(-\left\|w_i - w_j\right\|^2 / \sigma^2) & if\ i \sim j \\ 0 & otherwise \end{cases} \tag{1}$$

2.1.2. Sentiment label propagation

Next, words are classified into positive or negative that are represented as sentiment label $+1$ or -1, respectively. When the sentiment is propagated from words to words through the edges of a graph, words with ambiguous sentiment can be clearly classified. For this task, it is useful to employ graph-based semi-supervised learning. Semi-Supervised Learning (SSL) utilizes the similarity between data in the input space to determine the class label for each point of data. A well-known trait of SSL is that the algorithm can be applicable even with very small labeled data sets. This is because of its smoothness or cluster assumption, which describes that the predicted labels of nodes are assimilated with those of labeled ones nearby. This fits well with our situation, where sentiment labels of words should be similar if they have similar purpose and usage, in other words, if they are contextually

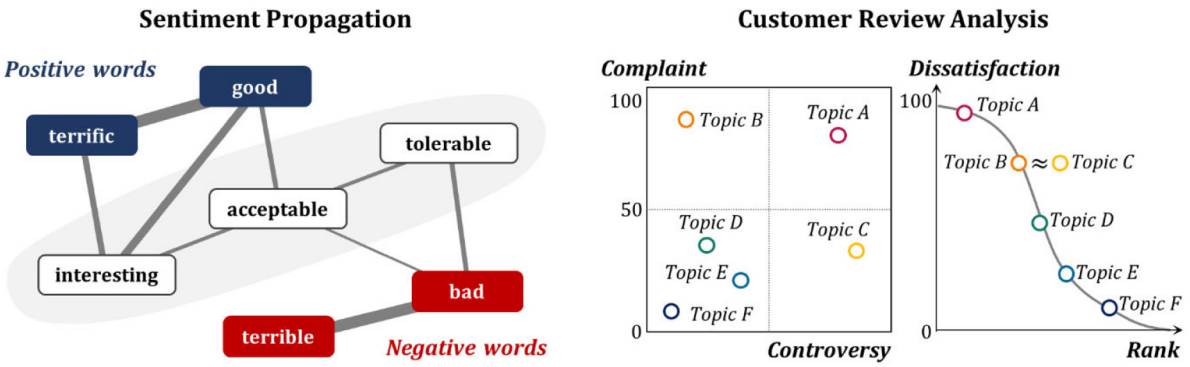
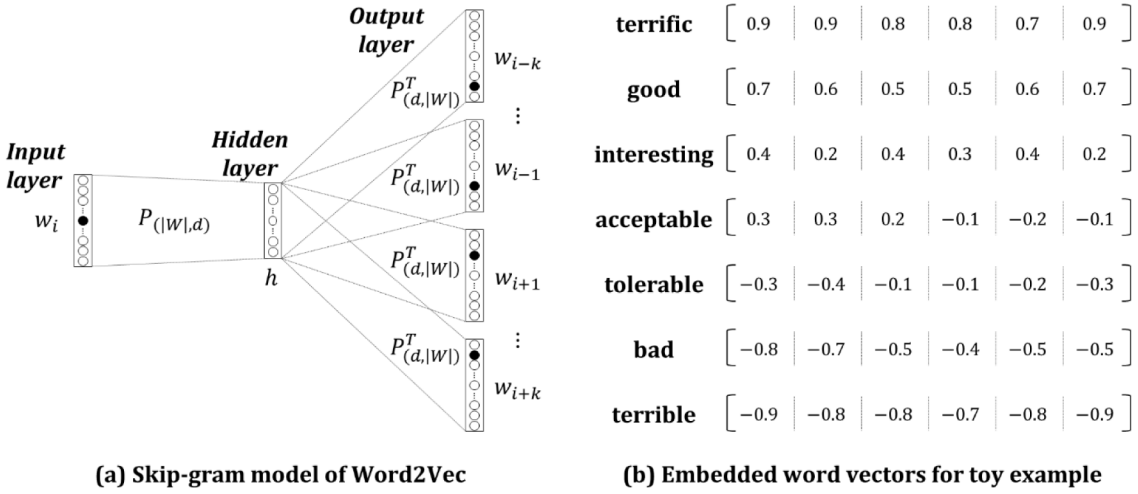


Fig. 1. Schematic description of the proposed method.



(a) Skip-gram model of Word2Vec **(b) Embedded word vectors for toy example**

Fig. 2. Skip-gram model of Word2Vec and its toy example.

similar. The process of label selection and propagation for unlabeled data is described in detail below.

Label selection: By citing a general-purpose sentiment dictionary, the ground truth label is selected as a small number of words. The label set has a clear definition and consistency in domain changes, and it is thereby used as the criteria for sentiment classification of other words. In this study, we cite the Korean Sentiment Analysis Corpus (KOSAC) (Shin et al., 2012) in which three researchers manually analyzed newspaper articles for the classification and the validation of words' sentiments as positive or negative. With denoting a word set of KOSAC by L, the label set is defined by the intersection of L and W. Therefore, node set W is divided into a labeled subset W^l and an unlabeled subset W^u. Then, Y^l and Y^u are defined as sets of numerical values of W^l and W^u, respectively, by denoting positive words as +1, negative words as −1, and unlabeled words as 0.

Graph-based semi-supervised learning: Words' sentiments are derived by applying graph-based semi-supervised learning (GSSL) (Bengio et al., 2006; Chapelle et al., 2006; Kim et al., 2019; Lee et al., 2018; Shin et al., 2013; Zhu and Ghahramani, 2002). With the given label set Y^l, the sentiment labels of words $f = \{f_1, f_2, \ldots, f_{|W|}\}^T$ are determined by minimizing the following quadratic objective functional (Bengio et al., 2006):

$$\min_f \ (f-Y)^T(f-Y) + \mu f^T L f \tag{2}$$

where L is the graph Laplacian defined as $D - S$, with $D = diag\left(d_i\right)$ and $d_i = \sum_j s_{ij}$. In (2), the first term is the loss for consistency with actual labels for labeled nodes, while the second term is the smoothness for consistency with the geometry of the data. The parameter μ is for a trade-off between the two terms. Because (1) is a convex problem,

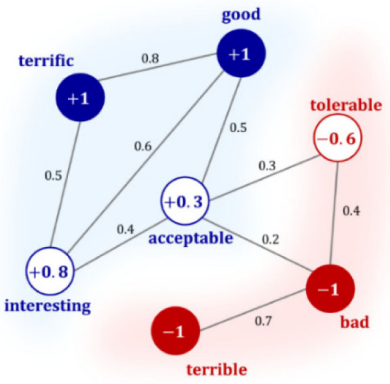


Fig. 3. Toy example for graph-based semi-supervised learning.

the analytical solution is easily calculated by its partial derivative with respect to f:

$$f = (I + \mu L)^{-1} Y \tag{3}$$

where I is the identity matrix. As a result, the unlabeled words W^u are classified into positive or negative sentiments when they are connected with stronger and larger numbers of edges to the labeled word with $Y^l = +1$ (or $Y^l = -1$). This means that similar words are classified into the same label.

Fig. 3 shows the result of a toy example with the application of GSSL. In the graph of the seven words, 'terrific' and 'good' are labeled as positive (+1) while 'terrible' and 'bad' are labeled as negative (−1). Every edge is calculated by (1) based on the embedded vectors in Fig. 2(b). As a result of sentiment label propagation, the unlabeled words, 'interesting' and 'tolerable', are classified with the same sentiment as their labeled neighbor words. In the case of 'acceptable', it roughly seems to be located in the neutral position. Consequently, the edges of 'acceptable' are more strongly connected with positive words, so the word is classified as positive.

2.2. Customer review analysis

Customer review analysis, following the result of sentiment label propagation, is performed by mutually comparing negative feedbacks with positives on whole parts of a product. Each part of a product is derived as a topic, and the sentiments of the words are applied. As a result, three indices, *controversy, complaint*, and *dissatisfaction*, are derived to further highlight negative feedbacks. Controversy indicates "how often are the topics raised?". If a certain topic has a high degree of controversy, it means that the topic is discussed by most reviewers or people. Complaint indicates "how severely is the topic complained about?". If a certain topic is strongly negatively complained about relative to other topics in review documents, it has to be seriously regarded, since customers' dissatisfaction is likely to be incurred by it. The combination of controversy and complaint can be used as a measure indicating the degree of negative customer evaluations on products or services; this composite index is denoted as dissatisfaction.

Controversy: To derive each part of a product as a topic, we apply latent dirichlet allocation (LDA), which is the most popular topic modeling algorithm (Blei et al., 2010, 2003). As a result, LDA derives n topics, each of which is composed of m keywords. We denote t_i as the ith topic, $k_i = \{k_i^1, k_i^2, \ldots, k_i^m\}$ as the keyword set of t_i, and c_i as the summation of frequencies for all keywords in t_i. Hence, c_i indicates "how often is the t_i raised?" Next, we calculate $controversy_i$ using (4), which denotes the controversy of t_i, with a sigmoid function for standardizing the frequency. Such post-processing methods prevent the controversy from shifting when a topic has substantially higher or lower frequency than others.

$$Controversy_i = s\left(\frac{c_i - \mu_f}{\sigma_f}\right), \quad s(x) = \frac{1}{1+e^{-x}} \tag{4}$$

where μ_f and σ_f are the average and the standard deviation of all frequencies $\{c_1, c_2, \ldots, c_n\}$, respectively.

Complaint: Customers' evaluation of each topic of a product is derived by measuring the co-occurrence by applying each keyword to sentiment words. For positive evaluation, p_i is the co-occurrence between k_i with W^p, and for negative evaluation, n_i is the co-occurrence between k_i with W^n. Then, we define the negative bias of sentiment for t_i as b_i. When the bias is high, it indicates that t_i is negatively evaluated more than other topics. The complaint of t_i, $Complaint_i$, can be calculated as (5) with standardization of all biases $\{b_1, b_2, \ldots, b_n\}$.

$$Complaint_i = s\left(\frac{b_i - \mu_b}{\sigma_b}\right), \quad b_i = \frac{n_i}{p_i} \tag{5}$$

where μ_b and σ_b are the average and the standard deviation of all biases $\{b_1, b_2, \ldots, b_n\}$, respectively.

Dissatisfaction: The *controversy* and the *complaint* are combined to measure the degree of negative customer evaluation. We denote this composite index as *dissatisfaction*. Higher *dissatisfaction* indicates a higher degree of customer dissatisfaction and a larger number of customers. *Dissatisfaction* for t_i is derived by multiplying $controversy_i$ and $complaint_i$ as follows:

$$Dissatisfaction_i = Controversy_i \times Complaint_i \tag{6}$$

Table 1
Data description.

Automobile	Number of data	Percentage
Compact sedan	93,354	30
Midsize sedan	133,434	43
Executive sedan	12,212	4
Luxurious sedan	32,573	10
Large SUV	39,977	13

If many topics have similar *dissatisfaction* values, a second investigation may be required. Therefore, we additionally derive the *controversy-complaint* quadrant as a supplementary indicator. It dissects the values of *dissatisfaction* and scatters the topics onto the quadrant, similar to importance-performance analysis (Martilla and James, 1977). This allows decision makers to prioritize the complaint topics and identify which topics are frequently raised by customers. This makes it possible to perform intuitively comparative analysis and clearly distinguish customer evaluation patterns (Chen and Ann, 2016; Lee, 2015; Yin et al., 2016).

3. Results and discussion

3.1. Data

In this study, reviews for automobiles were analyzed. Five automobiles from a Korean motor company in 2017 were selected: a compact sedan, a mid-size sedan, an executive sedan, a luxury sedan, and a large crossover SUV. For each automobile, the top two Internet communities were selected as data sources according to the number of members. In total, 311,550 reviews were collected via the web crawling of 10 internet communities, and Table 1 shows a data description. Due to the privacy concern, the sources of the internet communities cannot be disclosed; however, they may be provided on demand with automobile information masked.

3.2. Results for sentiment propagation

3.2.1. The graph of sentiment words

First, 62,486 words were embedded into 100-dimensional vectors with the skip-gram model of Word2Vec. Then, a word graph was constructed of 10,390 words that appeared in more than 100 reviews. Among them, sentiments of the label set were made up of 930 positive words and 1,030 negative words referring to KOSAC.

From the 1,960 labeled words, sentiments were propagated to the remaining 8,430 unlabeled words. As a result, 10,390 whole words were classified into 4,930 (47.4%) positive words and 5,460 (52.6%) negative words. Fig. 4 shows the subset of the word graph that compares the sentiment of words before and after sentiment propagation. Fig. 4(a) shows the original word graph, where only 18.9% of words have sentiments and the rest are unlabeled, whereas Fig. 4(b) shows the resulting word graph where every word takes on either a positive or negative sentiment. The extension of the word set of sentiments allows us to elucidate customers' hidden sentiments in free-form reviews by assigning sentiments to ambiguous or vague words.

3.2.2. Comparison results

To validate sentiment propagation, experiments were performed for comparison with the results achieved by artificial neural networks (ANN) (Abraham, 2005; Bishop, 1995) and support vector machines (SVM) (Schölkopf et al., 2002; Shin and Cho, 2006; Suykens and Vandewalle, 1999). As a performance measure, the area under receiving operating characteristic curve (AUC) was applied to 1,960 labeled words of five-fold cross validation with 100 times repetitions. The values of the parameters of each algorithm were adjusted for the best results. In GSSL, the density of the word graph was adjusted using the ε-nearest neighbor (εNN) and the k-nearest neighbor (kNN) methods. The

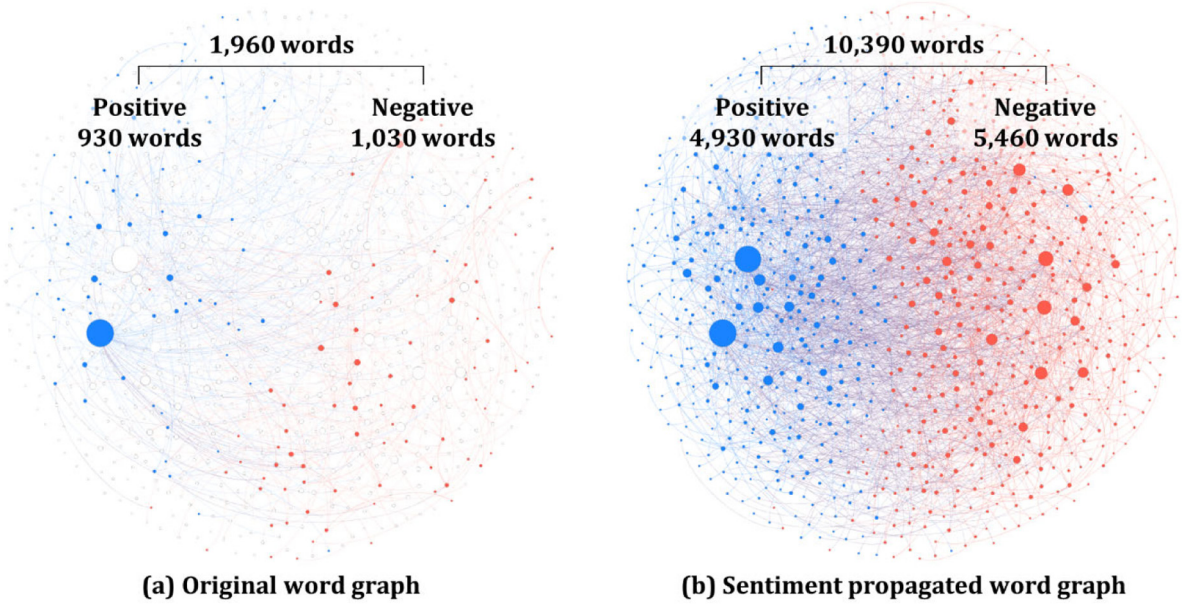


Fig. 4. Comparison for sentiment information in subsets of word graphs.

εNN connects nodes to edges where the similarity is larger than ε, and the kNN connects only the most similar k nodes to edges. In ANN, we set a hidden layer, then adjusted the number of hidden nodes. In SVM, we adjusted the kinds of kernels. Linear, polynomial, and Radial Basis Function (RBF) kernel were compared, and especially, there were three polynomial kernels subdivided by different degrees. For the validity of sentiment propagation, we performed additional experiments with two benchmark Korean review datasets: *Movie* (https://movie.naver.com/) and *Game* (https://store.steampowered.com/) including 200,000 and 100,000 short reviews, respectively. Table 2 shows the AUC results with the various parameters of each algorithm. As a result, GSSL showed the best performance from all datasets. In our dataset (*Automobile*), GSSL showed 0.981 AUC with the kNN method at k = 20. By contrast, ANN showed 0.724 AUC with 90 hidden nodes and SVM showed 0.974 AUC with the RBF kernel. In the same manner, proposed method performs well for two benchmark datasets. Through comparison experiments, it is shown that the proposed method performs highly accurate sentiment classifications.

In addition, we examined words' sentiment between domains. Firstly, the automobile dataset was divided into five subsets according to each automobile. After that, we had seven datasets from different domains: fives were homogeneous and the rest of them were heterogeneous. Next, sentiment label propagation was performed, and common words which are included in every domain were picked out. Finally, those sentiments were compared by correlation analysis and the results are shown in Fig. 5. It was remarkable that five automobile domains highly correlated each other. On the other hand, heterogeneous domains, movie and game dataset, showed different words sentiment. As the purpose of the proposed method was sentiment classification that reflects the domain characteristics well, the experimental results indicates that the purpose has been well achieved.

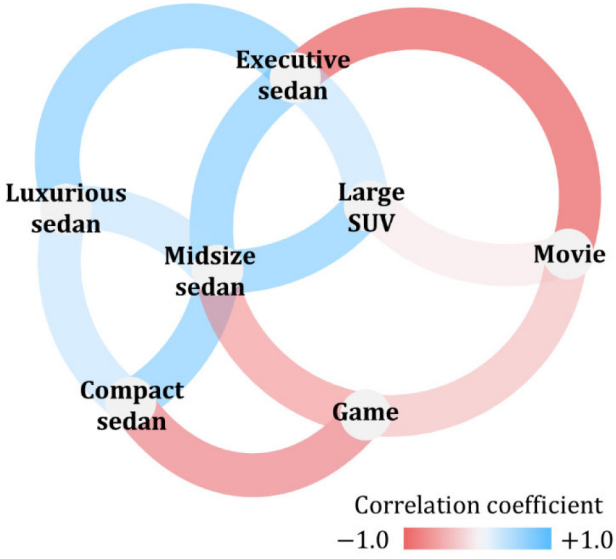


Fig. 5. Correlation graph for comparison of word's sentiment between domains.

3.2.3. Case study

We further examined case examples for the results of sentiment propagation to identify if the proposed method reflected the domain properties of the reviews. The case examples considered were selected from the types of Internet jargon that are frequently used in automobile reviews. The examples are shown in Fig. 6 along with actual reviews. As a positive case, we selected "force", which implies in Korean "a feeling of strong impression" and is used to describe appearance. Through actual reviews in Fig. 6(a), it can be seen that "force" in the automobile domain is used to describe the color of a body or the wheels. "Force" depicts strong and masculine images that emerge in black color. In addition, "luxurious" and "force" are often referred to together with a darker color along the lines of gray. It can be seen that "force" is used to positively describe the "powerful, masculine and luxurious" image that is evoked when car bodies or wheels are colored black or dark gray.

Next, *"kwak"* is selected as an example of a negative word, as shown in Fig. 6(b). According to a Korean dictionary, *"Kwak"* has two meanings: onomatopoeia and a mimetic word. First, as an onomatopoeia, *"kwak"* represents sounds made by relatively weak crashes. In the automobile domain, *"kwak"* is typically mentioned along with accidents from a "door crack" or a speed bump. A "door crack" accident refers to crashes of doors between adjacent cars that arise from drivers getting in and out of their cars. In Korea, "door crack" accidents frequently occur due to narrow parking spaces. Meanwhile, a speed bump accident describes situations caused by speed bumps such as scratches and dents

Table 2
Performance comparison for ANN, SVM, and GSSL.

Algorithm	Parameters		Accuracy performance (AUC) for datasets		
			Automobile	Movie	Game
ANN	Number of hidden nodes	10	0.597	0.546	0.565
		30	0.646	0.563	0.610
		50	0.677	0.611	0.638
		70	0.706	0.643	0.679
		90	0.724	0.683	0.729
SVM	Kernel	Linear	0.615	0.555	0.564
		Poly. 2	0.675	0.587	0.683
		Poly. 3	0.728	0.639	0.712
		Poly. 4	0.794	0.766	0.834
		RBF	0.974	0.905	0.912
GSSL	ϵNN	0.1	0.975	0.908	0.927
		0.3	0.977	0.912	0.931
		0.5	0.979	0.906	0.929
		0.7	0.976	0.919	0.930
		0.9	0.973	0.911	0.932
	kNN	5	0.979	0.930	0.965
		10	0.980	**0.934**	0.970
		20	**0.981**	0.932	**0.974**
		50	0.978	0.931	0.964
		100	0.979	0.930	0.966

in an underbody or a front body. Even though drivers speed down for speed bumps, accidents frequently occur because of the number and size of speed bumps in Korea in particular. Therefore, *"kwak"* is used by drivers to describe negative situations when asked about problems with automobiles, accident handling, and repairs. Next, as a mimetic word, *"kwak"* represents sudden and aggressive movements. In the automobile domain, it is often used to describe abnormal operating conditions on brakes. If a braking error occurs, deceleration is rough and not smooth. In this situation, *"kwak"* is used with a negative meaning in expressions such as "the stop is not smooth" or "the brake step is too rough and stiff".

Through these case examples, it is identified that the proposed sentiment propagation considers various features, such as characteristics of Internet jargon, automobile expressions, and specific situations in Korea. Therefore, the main purpose of sentiment propagation, which is to reflect the domain properties of reviews for sentiment classification, is successfully achieved here.

3.3. Results for customer review analysis

3.3.1. Topic modeling

LDA was applied to reviews for topic modeling by setting the number of topics to 10 and the number of features to 100 for each topic. The results covered the overall functions of automobiles, and Table 3 summarizes these. There were topics for fundamental functions, such as "powertrain & transmission", "driving performance", and "tire", and there were also topics for problems and controls that occur in automobiles, such as "noise", "engine maintenance", and "maintenance & repair". In addition, the results included topics for interior functions, such as "cooling, heating & ventilation" and "electric device & accessory", as well as topics for exterior appearance, such as "body & painting" and "light".

Table 3
Results for topic modeling.

Powertrain & transmission				
brake accelerator	starting gear	engine automatic	shifting neutral	speed parking
Driving performance				
drive mode	run expressway	mileage city-drive	long-distance sport	distance echo
Engine maintenance				
engine fuel	oil knocking	engine oil synthesis	replacement filter	change diesel
Maintenance & repair				
maintenance faulty	repair defect	check leak	service garage	shop reservation
Cooling, heating & ventilation				
interior temperature	air air conditioner	wind heater	ventilation seat warmer	smell Humidity
Electric device & accessory				
audio embedment	speaker genuine	Bluetooth connection	camera linkage	GPS navigator upgrade
Noise				
noise sound	din wind noise	racket underbody	stress cause	nervousness symptom
Tire				
tire puncture	wheel wear	front wheel air pressure	rear wheel ride comfort	spring drive
Body & painting				
glass film wax	oil film coating	underbody gloss	corrosion cover	remover dust
Light				
halogen light	bulb headlamp	taillight headlight	fog genuine	lamp change

3.3.2. Controversy, complaint, and dissatisfaction

Controversy and complaint were calculated by applying sentiment words to the topic modeling results, and therefrom, dissatisfaction was finally derived. The results for those indices are described in Fig. 7, and the related figures are summarized in Table 4.

The topic with the highest dissatisfaction was "noise". "Noise" can be understood as the strongest complaint in the majority of cases, because it ranked as the second controversy and the first complaint, with a large gap to the second complaint. This may be attributed to the fact that noises both act as the starting point of problem recognition and lead to anxiety about safety. Among the top-10 features of "noise" presented in Table 4, it is shown that "noise, din, and sound" encompass all kinds of sounds from all parts of an automobile, such as the side windows and the powertrain through "wind noise" and "underbody". In addition, customers recognized noise as the "symptom" and the "cause" of problems, and they felt a great sense of anxiety with "nervousness" and "stress". Following "noise",

(a) Positive word: *"force"*

> The black color also has ***force***.
>
> In recent trend, the dark gray is luxurious with ***force*** and the light gray is common.
>
> The glazed dark color for wheels of the 4WD model would be more fascinating with ***force***.

(b) Negative word: *"kwak"*

> ***Onomatopoeia***
>
> I had the "door crack" accident. The next car owner opened the door with ***"kwak"*** sound.
>
> As soon as passing the speed bump, I heard the ***"kwak"*** sound from the underbody.
>
> ***Mimetic word***
>
> The brake was strange, and car stopped in ***"kwak"*** manner.
>
> Although I gently stepped on the brake, it activated in ***"kwak"*** manner.

Fig. 6. Case studies for sentiment propagation.

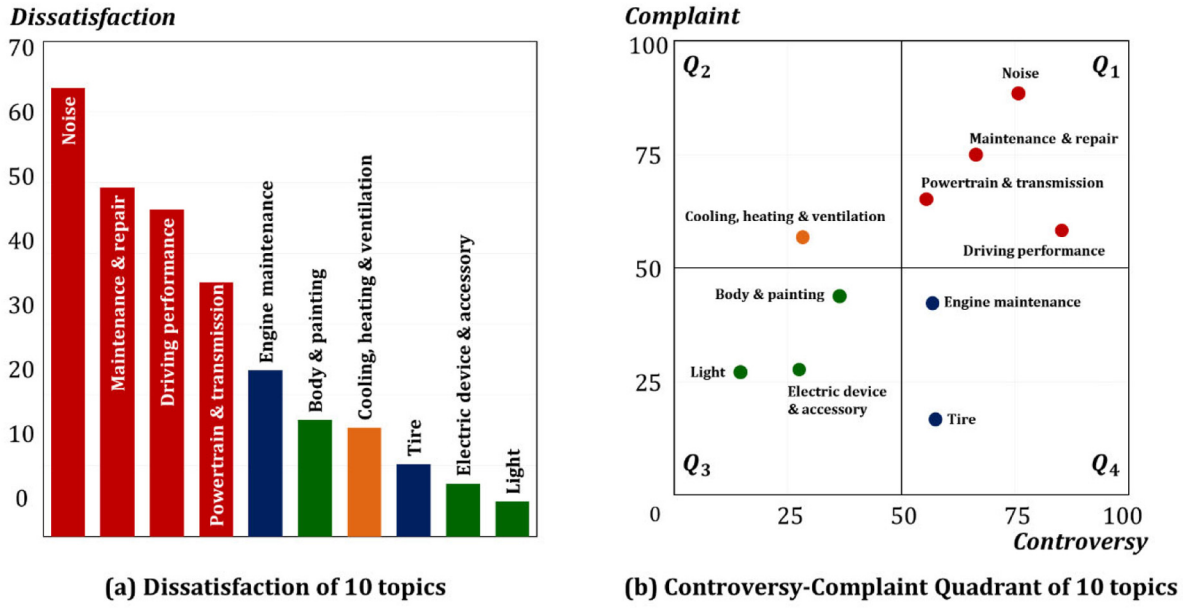


Fig. 7. Results for Customer Review Analysis of 10 topics.

Table 4
Values of customer dissatisfaction Index for 10 topics.

Topic	Dissatisfaction	Controversy	Complaint
Noise	63	73	87
Maintenance & repair	49	69	72
Driving performance	46	84	55
Powertrain & transmission	36	54	67
Engine maintenance	23	55	43
Body & painting	16	35	47
Cooling, heating & ventilation	15	30	52
Tire	10	56	18
Electric device & accessory	8	26	29
Light	5	18	28

"maintenance & repair", "driving performance", and "powertrain & transmission" were ranked from second to fourth, respectively, and all of them are included in Q_1. The first quadrant, Q_1, shows terms that are high in both controversy and complaint and therefore represent a "strong complaint of majority". Among these, "maintenance & repair" and "driving performance" were almost the same in dissatisfaction, but "maintenance & repair" was higher in complaint while "driving performance" was higher in controversy. Therefore, even if some topics share dissatisfaction and quadrant, decision-making can be performed according to the difference between controversy and complaint. Next, "engine maintenance" ranked as the fifth dissatisfaction with high controversy but low complaint, and thereby, "engine maintenance" was included in Q_4. As a result, by utilizing dissatisfaction, we could summarize the results of customer review analysis in a practical way and comprehensively understand the customer evaluation pattern in terms of the controversy-complaint quadrant.

3.3.3. Enrichment study across automobile types

For the enrichment study on dissatisfaction, the degree of negative evaluation was compared. Fig. 8(a) shows the result of this study. The overall average of bias was 1.22 in (6), indicating that customers generally expressed negative opinions more often than positive opinions.

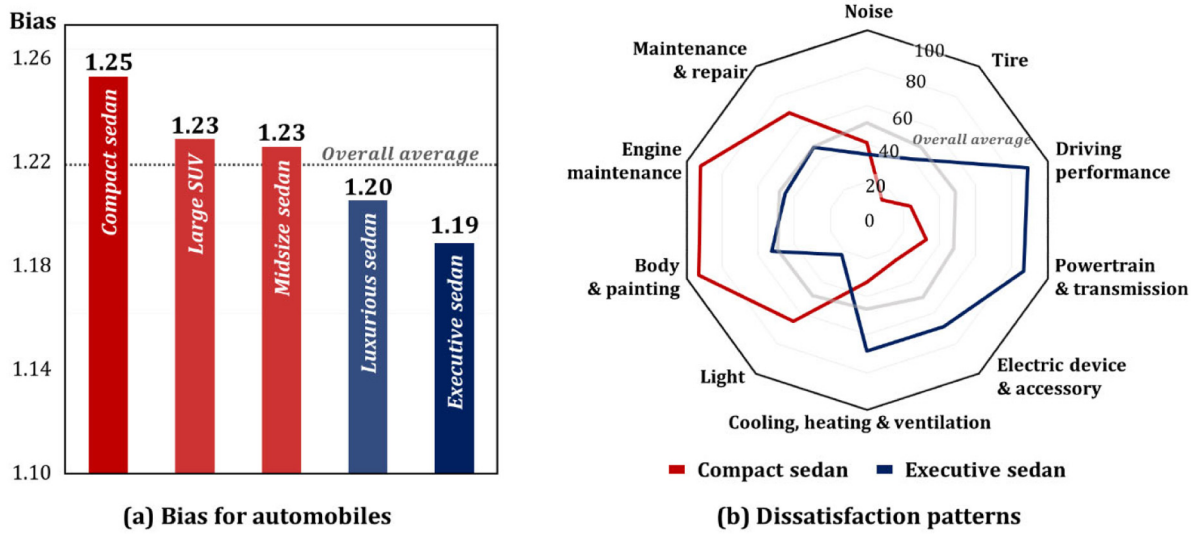


Fig. 8. Customer review patterns for automobiles.

The most negatively evaluated model was the compact sedan, while the least negatively evaluated model was the executive sedan.

Moreover, for two distinguishing automobiles, the compact sedan and the exclusive sedan, customer review analysis of the proposed method was applied to 93,354 and 12,212 reviews, respectively. Finally, each dissatisfaction distribution for the same topic is shown in Fig. 8(b). As a result, it indicates that the noticeable topics of the compact sedan were "engine maintenance" while those of the executive sedan were "powertrain & transmission". These topics showed clear differences that are caused by defects in each automobile.

GDI engine defects The compact sedan had an engine defect. There were cases of cylinder deformation and scratching due to durability issues from the mounted GDI engine. Therefore, it can be seen that "engine maintenance" is highly related issues to diesel, knocking, and engine oil.

Transmission defects: The executive sedan had a transmission defect in which a gear was fixed to the fifth and remained unchanged, even when driving and stopping, despite updating the software.

Examination and investment on quality control are required for the engine part in the case of the compact sedan and for the transmission part in the case of the executive sedan. The results of the enrichment study exemplify the practical use of the proposed method, as well as its potential expansion to other products.

4. Conclusion

In this study, we proposed a sentiment propagation for customer review analysis. The proposed sentiment propagation increased the domain adaptability of the existing sentiment analysis by expanding sentiment words both contextually and in a domain-specific manner. To implement this, semi-supervised learning was employed to the word graph that was constructed by word embedding algorithms on online product review. The case study supported that sentiment propagation enriches the sentiment dictionary by expanding its coverage to non-grammatical Internet language, industry-specific jargon, local dialect, and so on. On the other hand, the proposed analysis of customer evaluation involved postprocessing for the sentiment analysis results. The indicator of *dissatisfaction*, which consisted of *controversy* and *complaint*, aimed to determine the priorities of complaints about products and the evaluation patterns of customers. In our case of automobile online reviews, "noise" ranked the highest, followed in order by "maintenance & repair", "driving performance", and "powertrain & transmission". Those complaint topics represented 'strong complaints of the majority of customers', and therefore require urgent repair and quality improvement.

The proposed method has some limitations that provide opportunities for future work. First, the proposed method should be extended to other language domains along with translating Korean to English. As there are many sources of words' sentiments such as SentiWordNet (Baccianella et al., 2010), words could be subdivided into more classes according to the intensity of sentiment, such as strong positive or negative, weak positive or negative, and neutral words that do not belong to either. The classification performance, in addition, could be further enhanced by increased label information. Next, the proposed method should be applied to other industry domains. Aside from automobile reviews, there are many products or services with large amounts of free-form reviews. From various reviews of industry domain, the proposed method could compare words' sentiments and analyze customer evaluation patterns.

CRediT authorship contribution statement

Sunghong Park: Conceptualization, Methodology, Data curation, Formal analysis, Resources, Validation, Visualization, Writing - original draft, Writing - review & editing. **Junhee Cho:** Conceptualization, Methodology, Data curation, Resources, Validation. **Kanghee Park:** Conceptualization, Formal analysis, Resources. **Hyunjung Shin:** Conceptualization, Methodology, Formal analysis, Writing - original draft, Writing - review & editing, Supervision.

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.

Acknowledgments

HJS would like to gratefully acknowledge the support from the National Research Foundation of Korea (NRF) grant funded by the Korea government (MOE) (no. 2018R1D1A1B07043524), Institute for Information and communications Technology Promotion (IITP) grant funded by the Korea government (MSIT) (no. 2018-0-00440, ICT-based Crime Risk Prediction and Response Platform Development for Early

Awareness of Risk Situation), the BK21 FOUR program of the National Research Foundation of Korea funded by the Ministry of Education (NRF5199991014091), and the Ajou University research fund. KHP would like to gratefully acknowledge the support from the Research Program (no. K20L03C05S01) at Korea Institute of Science and Technology Information (KISTI).

References

Abraham, A., 2005. Artificial neural networks. In: Handbook of Measuring System Design.

Baccianella, S., Esuli, A., Sebastiani, F., 2010. Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. In: Lrec. pp. 2200–2204.

Bengio, Y., Delalleau, O., Le Roux, N., 2006. 11 Label Propagation and Quadratic Criterion.

Bishop, C.M., 1995. Neural Networks for Pattern Recognition. Oxford university press.

Blei, D., Carin, L., Dunson, D., 2010. Probabilistic topic models: A focus on graphical model design and applications to document and image analysis. IEEE Signal Process. Mag. 27 (55).

Blei, D.M., Ng, A.Y., Jordan, M.I., 2003. Latent dirichlet allocation. J. Mach. Learn. Res. 3, 993–1022.

Chaovalit, P., Zhou, L., 2005. Movie review mining: A comparison between supervised and unsupervised classification approaches. In: Proceedings of the 38th Annual Hawaii International Conference on System Sciences. IEEE, p. 112c.

Chapelle, O., Schölkopf, B., Zien, A., 2006. Semi-Supervised Learning, vol. 2. MIT Press, Cambridge;
Cortes, C., Mohri, M., 2014. Domain adaptation and sample bias correction theory and algorithm for regression. Theoret. Comput. Sci. 519, 103126.

Chen, C.-M., Ann, B.-Y., 2016. Efficiencies vs. importance-performance analysis for the leading smartphone brands of Apple, Samsung and HTC. Total Qual. Manag. Bus. Excell. 27, 227–249.

Cho, Y., Im, I., Hiltz, R., Fjermestad, J., 2002. An analysis of online customer complaints: implications for web complaint management. In: Proceedings of the 35th Annual Hawaii International Conference on System Sciences. IEEE, pp. 2308–2317.

Chung, W., Tseng, T.-L.B., 2012. Discovering business intelligence from online product reviews: A rule-induction framework. Expert Syst. Appl. 39, 11870–11879.

Coussement, K., Van den Poel, D., 2008. Improving customer complaint management by automatic email classification using linguistic style features as predictors. Decis. Support Syst. 44, 870–882.

Fornell, C., Wernerfelt, B., 1987. Defensive marketing strategy by customer complaint management: a theoretical analysis. J. Mar. Res. 24, 337–346.

Gamon, M., Aue, A., Corston-Oliver, S., Ringger, E., 2005. Pulse: Mining customer opinions from free text. In: International Symposium on Intelligent Data Analysis. Springer, pp. 121–132.

Hennig-Thurau, T., Gwinner, K.P., Walsh, G., Gremler, D.D., 2004. Electronic word-of-mouth via consumer-opinion platforms: what motivates consumers to articulate themselves on the internet?. J. Interact. Mark. 18, 38–52.

Hou, T., Yannou, B., Leroy, Y., Poirson, E., 2019. Mining customer product reviews for product development: A summarization process. Expert Syst. Appl. 132, 141–150.

Kim, M., Lee, Shin, H., 2019. Semi-supervised learning for hierarchically structured networks. Pattern Recognit..

Lee, H.-S., 2015. Measurement of visitors' satisfaction with public zoos in Korea using importance-performance analysis. Tour. Manag. 47, 251–260.

Lee, D.-g., Lee, S., Kim, M., Shin, H., 2018. Historical inference based on semi-supervised learning. Expert Syst. Appl. 106, 121–131.

Liu, B., Zhang, L., 2012. A survey of opinion mining and sentiment analysis. In: Mining Text Data. Springer, pp. 415–463.

Martilla, J.A., James, J.C., 1977. Importance-performance analysis. J. Mark. 41, 77–79.

Mikolov, T., Chen, K., Corrado, G., Dean, J.J.a.p.a., 2013a. Efficient Estimation of Word Representations in Vector Space.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J., 2013b. Distributed representations of words and phrases and their compositionality. Adv. Neural Inf. Process. Syst. 3111–3119.

Moghaddam, S., Ester, M., 2012. Aspect-based opinion mining from product reviews. In: Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, p. 1184.

Pang, B., Lee, L., Vaithyanathan, S., 2002. Thumbs up?: sentiment classification using machine learning techniques. In: Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing, vol. 10. Association for Computational Linguistics, pp. 79–86.

Pyon, C.U., Woo, J.Y., Park, S.C., 2010. Intelligent service quality management system based on analysis and forecast of VOC. Expert Syst. Appl. 37, 1056–1064.

Schölkopf, B., Smola, A.J., Bach, F., 2002. Learning with Kernels: Support Vector Machines, Regularization, Optimization and beyond. MIT press.

Sen, S., Lerman, D., 2007. Why are you telling me this? An examination into negative consumer reviews on the web. J. Interact. Mark. 21, 76–94.

Shin, H., Cho, S., 2006. Response modeling with support vector machines. Expert Syst. Appl. 30, 746–760.

Shin, H., Hou, T., Park, K., Park, C.-K., Choi, S., 2013. Prediction of movement direction in crude oil prices based on semi-supervised learning. Decis. Support Syst. 55, 348–358.

Shin, H., Kim, M., Jang, H., Cattle, A., 2012. Annotation scheme for constructing sentiment corpus in Korean. In: Proceedings of the 26th Pacific Asia Conference on Language, Information and Computation. pp. 181-190.

Sparks, B.A., So, K.K.F., Bradley, G.L., 2016. Responding to negative online reviews: The effects of hotel responses on customer inferences of trust and concern. Tour. Manag. 53, 74–85.

Suykens, J.A., Vandewalle, J., 1999. Least squares support vector machine classifiers. Neural Process. Lett. 9, 293–300.

Tang, H., Tan, S., Cheng, X., 2009. A survey on sentiment detection of reviews. Expert Syst. Appl. 36, 10760–10773.

Whissell, C.M., 1989. The dictionary of affect in language. In: The Measurement of Emotions. Elsevier, pp. 113–131.

Yi, J., Nasukawa, T., Bunescu, R., Niblack, W., 2003. Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques. In: Third IEEE International Conference on Data Mining. IEEE, pp. 427–434.

Yin, S.-Y., Huang, K.-K., Shieh, J.-I., Liu, Y.-H., Wu, H.-H., 2016. Telehealth services evaluation: a combination of SERVQUAL model and importance-performance analysis. Qual. Quant. 50, 751–766.

Zhu, X., Ghahramani, Z., 2002. Learning from Labeled and Unlabeled Data with Label Propagation.