Computer Science and Information Systems 16(2):359–380 https://doi.org/10.2298/CSIS180703006P

Product Reputation Mining: Bring Informative Review Summaries to Producers and Consumers

Zhehua Piao[1], Sang-Min Park[2], Byung-Won On[2], Gyu Sang Choi[3], and Myong-Soon Park[1]

[1] Department of Computer Science and Engineering, Korea University,
145 Anam-ro, Seongbuk-gu, Seoul 02841, South Korea
huanmie2199@gmail.com, myongsp@korea.ac.kr

[2] Department of Software Convergence Engineering, Kunsan National University,
558 Daehak-ro, Gunsan-si, Jeollabuk-do 54150, South Korea
{b1162, bwon}@kunsan.ac.kr

[3] Department of Information and Communication Engineering, Yeungnam University,
280 Daehak-ro, Gyeongsan-si, Gyeongbuk-do 38541, South Korea
castchoi@ynu.ac.kr

Abstract. Product reputation mining systems can help customers make their buying decision about a product of interest. In addition, it will be helpful to investigate the preferences of recently released products made by enterprises. Unlike the conventional manual survey, it will give us quick survey results on a low cost budget. In this article, we propose a novel product reputation mining approach based on three dimensional points of view that are word, sentence, and aspect–levels. Given a target product, the aspect–level method assigns the sentences of a review document to the desired aspects. The sentence–level method is a graph-based model for quantifying the importance of sentences. The word–level method computes both importance and sentiment orientation of words. Aggregating these scores, the proposed approach measures the reputation tendency and preferred intensity and selects top-k informative review documents about the product. To validate the proposed method, we experimented with review documents relevant with K5 in Kia motors. Our experimental results show that our method is more helpful than the existing lexicon–based approach in the empirical and statistical studies.

Keywords: product reputation mining, opinion mining, sentiment analysis, sentiment lexicon construction

1. Introduction

Data analysis strategies and technologies are widely used in the recent marketing research area. In order to investigate the preferences of recently released products, enterprises need a state–of–the–art strategy to accurately grasp the public's taste for the product by automatically collecting and analyzing various types of data on the Web. In the manual survey, if company executives want to know how consumers think of their brand–new product, the employees in the marketing department will conduct a survey via email and phone. However, the recent survey response rate is low because modern people do not have enough time to response sincerely to the questionnaire and recent consumers are mainly interested in customized products. Unlike this conventional process, product reputation systems that collect various online data and predict the accurate survey results can provide quick results on a low cost budget. Meanwhile, potential customers may save their time in making their buying decision if they are served by product reputation mining systems. Until now, to purchase a brand-new product like Hyundai Sonata, a customer is likely to spend a lot of time in gathering helpful information on the Web. He/she first attempts to search for Hyundai Sonata and carefully go over relevant web pages one by one. Even though he/she strives to figure out which model is better, it is difficult to take the clear point due to a lot of information and advertising. As above, the product reputation mining systems can provide many benefits to both producers and consumers.

In this section, we briefly define the product reputation mining system as:

- In the first problem, given a product of interest, it automatically measures the reputation tendency (sentiment orientation – positive or negative) and level (the intensity of the sentiment orientation) of various aspects (e.g., price, design, and service) of the product. We assume that all aspects of the target product are given in advance.
- In the second problem, for each aspect of the product, it selects the top–k documents including the most *informative* reviews in the corpus of review documents. We assume that all review documents irrelevant with the target product are already filtered out before this problem. In addition, we will carefully define how informative a review document is in the next section.

Through the proposed product reputation mining system, both companies and customers can easily know the public's preference (as positive or negative) and the preferred intensity (Level $1 \sim 5$) of a product. They can also know the detailed points with respect to each aspect of the product. For the details, please see Section 3.

To address the product reputation mining problem, in this work, we propose a novel three-dimensional reputation mining approach that consists of aspect, sentence, and word–level methods.

As shown in Figure 1, the aspect–level method is the aspect classification model based on SVM, Random Forest, and FNN to assign the sentences of review documents to the desired aspects. The sentence–level method is a graph-based model for quantifying the importance of each sentence in review documents. The word–level method computes the importance of a word and measures the sentiment score of the word based on Korean sentiment lexicon. Finally, aggregating the scores of the aspect, sentence, and word–level methods, our method measures the reputation tendency and level in each aspect of the target product. In addition, all review documents in each aspect are rearranged by the aggregated scores and then top-k review documents with the highest aggregated scores are selected as the informative documents.

Our experimental results show that the accuracy of the aspect–level method is at least 0.852. In the existing lexicon–based approach, F_1–scores of the positive and negative sentences are 0.678 and 0.688, while those are 0.758 and 0.795 in the proposed method. These results mean that the proposed method improves about 12% and 5% in the positive and negative sentences. In our case study for K5 in Kia motors, we observed top-k reviews retrieved by the proposed method and finally concluded that most of the review documents are informative. We will discuss the experimental results In Section 4 and conducted a user study for the results retrieved by the proposed method and performed statistical tests. Through the significance tests, it turns out that our method is statistically better than the existing method.

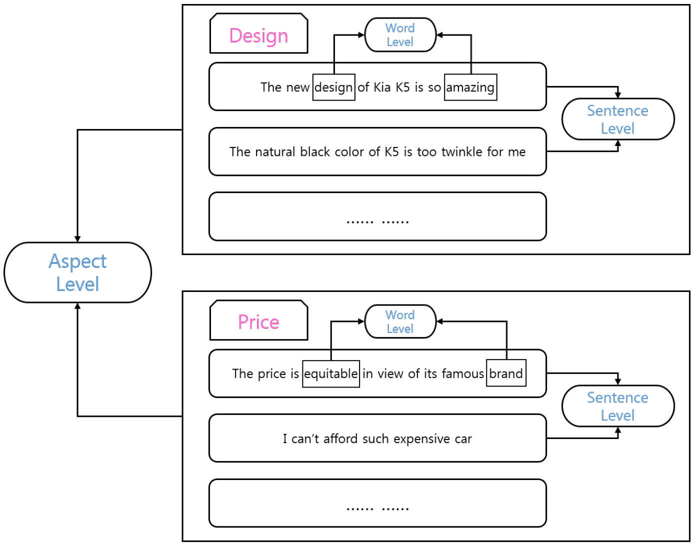

Design
Word Level
The new design of Kia K5 is so amazing
Sentence Level
The natural black color of K5 is too twinkle for me
....
Aspect Level
Price
Word Level
The price is equitable in view of its famous brand
Sentence Level
I can't afford such expensive car
....

Fig. 1. Aspect, Sentence, and Word–level based product reputation mining approach

The contributions of our work are as follows:

- To address the product reputation mining problem, we propose a novel three–dimensional reputation mining approach that consists of aspect, sentence, and word–level methods. We show the detailed algorithms of (1) measuring the reputation tendency and level about a target product and (2) selecting top-k informative review documents. These results will help customers make their buying decision and companies get to know the public's preference about the product in detail. We also constructed an elaborate Korean sentiment lexicon to determine the sentiment orientation of words.
- Our experimental results show that the proposed method is effective to address the product reputation mining problem. Compared to the existing lexicon–based approach, it improves up to 12% F_1– score. In addition, our statistical verification shows that the proposed method will be helpful for both company employees and customers. Consequently, these results indicate that it is beneficial to develop a web–based system based on the proposed method of aggregating three dimensional sentiment scores.
- According to our intensive literature survey, the key point of our method is to quantify the reputation of the product based on three dimensions (i.e., aspect, sentence, and word–levels). To the best of our knowledge, this is the first study to tackle the product reputation mining problem.

The remainder of this article is organized as follows: In section 2, we introduce previous main machine learning and lexicon–based approaches related to this work. In particular, we discuss the novelty of our method, in addition to the difference between previous studies and our work. In section 3, we deal with the formal problem definition. Then, we describe our product reputation mining approach in detail in section 4. Next, we explain the experimental set-up and discuss the experimental results in Section 5. Finally, we conclude our work and mention the future research direction in Section 6.

2. Literature Review

In 2000, Resnick et al. presetned several challenging issues and solution overview of product reputation systems that collect, distribute, and aggregate feedback about past consumers' behaviour so that these systems help people make their buying decision based on public history of particular sellers. By showing real cases of eBay's auction site, Bizrate's survey forum, and iExchange's product review site, the authors also stated main requirements of the product reputation systems. In particular, they focused on gathering reliable feedback in the reputation systems. For the detail, please see [18].

In 2008, Hwang and Ko proposed a Korean sentiment analysis method of labelling a document to either positive or negative and of classifying a sentence to either subjective or objective [9]. In the method, the authors made a Korean sentiment lexicon in which a Korean word is translated to the corresponding English word to obtain the polarity of the word. In addition, the sentiment analysis method classifies documents and sentences based on Support Vector Machines (SVM). However, their method does not consider aspects that are recently considered to be important in the product reputation mining problem.

Given a particular product, Jin and Ho presented a machine learning technique based on lexicalized Hidden Markov Models (HMMs) [10]. Specifically, their method extracts subjective sentences related to the target product from review documents and then labels each sentence to either positive or negative class. In particular, they trained the lexicalized HMMs with linguistic features including part-of-speech, phrases' internal formation patterns, and contextual clues surrounding words and phrases. Applying such linguistic features to HMM is different from previous approaches that address the product reputation mining problem. However, there is still room to improve the accuracy of the existing methods and to identify more informative review documents than the entire documents. Our proposed method shows the better result of correctly classifying the sentiment orientation of a target product based on aggregating the reputation scores measured by aspect, sentence, and word–level algorithms. For the detail, please refer to the experimental result section.

Steinberger et al. proposed a new approach that semi-automatically creates sentiment dictionaries in several languages [21]. They first made sentiment dictionaries for two source languages and then automatically translated them to the third languages in which a word is likely to be similar to that of the source languages. Through the third languages, the target dictionaries can be more corrected and further extended. In the experiment, they validated such a triangulation hypothesis by comparing triangulated lists to non–triangulated machine–translated word lists.

Khose and Dakhode proposed a product reputation analysis system that consists of five steps [12]. In the first step, review documents are collected and pre-processed for the next step. After target and opinion words are extracted, an object relation graph is formed by detecting the relation between them. In the third step, the weight of a node called confidence is computed based on a simple random walk model. In addition, to determine the sentiment scores of the target word and its opinion words, they used SentiWordNet that is a popular sentiment lexicon in opinion mining area. Each target word is represented as a vector of the target word's confidence and the sentiment scores of the opinion words associated with the target word. The reputation score of the target word is finally calculated as the product of the summation of the vectors related to the target word. Unlike our method, they consider only the reputation score of a target product in word–level.

In 2016, rather than classifying the porality (i.e., positive or negative) of words, Canales et al proposed a bootstrapping method that labels an emotional corpus automatically [4]. Based on NRC Word–Emotion Association Lexicon, they created the seed set and then expended the initial seed by means of similarity metrics. Furthermore, to distinguish between emotion categories in a fine–grained lexicon with 28 emotion categories, the authors in [26] proposed an approach of labelling primary and secondary words to one of emotion categories, where the primary words are used for detecting synonyms or other semantic words associated with each category, while the secondary words are used to mine the contextual relation between words. However, these methods focus mainly on constructing a fine-grained emotion lexicon which is considerably different from the production reputation mining problem that we present in this article. In addition, Ko presented a general method for creating an emotional word dictionary containing a semantic weight matrix and a semantic classification matrix [13]. Based on clustering synonymous relations and frequencies, he showed the detailed process of collecting a classification and weight matrix that can be used as the ontology and linked data of emotion.

Meanwhile, detecting emotion from text documents is a non-trivial task because of the limitation of human annotation. To tackle this problem, the authors in [25] utilized emoji as self–annotation of twitter users' emotional status. They believed that emoji is a good emotion indicator presenting a faithful representation of a user's emotional status but their approach is too limited to use other text documents rather than tweet mentions.

Sentiment analysis is generally categorized to two groups. One is machine learning approach and the other is lexicon-based approach. Although the lexicon-based approach has been used in wide applications, it does not work well to determine the sentiment orientation of tweets. This is because each tweet document is limited to only 140 characters and its sentences are not written according to the grammar. [19] presented SentiCircles, a lexicon-based approach for Twitters, that is based on the co-occurrence patterns of words in different tweet documents to update the prior degree and polarity in sentiment lexicons accordingly.

To improve the accuracy of the machine learning approach, Long Short Term Memory (LSTM) as one of Recurrent Neural Networks (RNN) deep learning models is widely used to classify a text document to either positive or negative class. LSTM is trained with a large number of training set containing a large number of pairs of the text document and sentiment polarity manually annotated by human experts. Then, given a text document in the test set, LSTM automatically determines the sentiment orientation of the text document. Teng et al. proposed a hybrid approach that is the trade-off between the context-sensitive method using LSTM and the lexicon-based method using the list of sentiment words [22]. This approach is not one of the product reputation mining algorithms but an advanced method for improving the conventional sentiment document classification methods. Similarly, [23] presented a hybrid approach that measures numerical numbers in multiple dimensions (i.e., valence-arousal space) by extracting/abstracting the locality information within each sentence based on Convolutional Neural Networks (CNN) and updating the context weights by means of long–distance dependency cross sentences based on LSTM.

Nowadays, online travel forums and social network sites are popular for sharing travel information. Review summaries for hotels automatically generated from many reviews in the sites can help travelers choose their preferred hotel during the trip. For (opinion) mining from online review documents about a target hotel, Hu et al. proposed a summarization method that finds top–k sentences using k–medoids clustering algorithm that removes sentences irrelevant with the target hotel [8]. They also proposed additional feature set that includes author reliability, review time, review usefulness, and conflicting opinions which are not considered in the previous review summarization methods. Although Hu et al.'s method is similar to our proposed method in that it selects top–k relevant sentences from review documents, there are main differences between them. While Hu's method first clusters text documents by contexutal information and then selects only top–k relevant sentences, our method computes the reputation score of a target product using word, sentence, and aspect–level methods and identifies top–k relevant but yet informative sentences. In addition, we make use of various learning models such as SVM, Random Forest, and even deep neural networks, whereas Hu's method is based on only k–medoids clustering method.

3. Problem definition

Table 1. An example of the first solution method

Aspect	Reputation tendency	Reputation level
Design	Positiveness	Level 2
Performance	Positiveness	Level 3
Price	Negativeness	Level 1
Quality	Positiveness	Level 5
Service	Positiveness	Level 5

In this section, we define the *product reputation mining* problem as two sub-problems. In the first problem, given a product of interest as input (e.g., a particular car e like K5 made by Hyundai and KIA motors), the goal of the product reputation mining method is to *automatically* measure the reputation tendency and level per aspect. For instance, *design*, *performance*, *price*, *quality*, and *service* may be the main aspects that many consumers often consider importantly when they are about to purchase their brand-new car. Table 1 shows the outcome of the product reputation mining method. Let us assume that design, performance, price, quality, and service are given in advance as the main aspects of evaluating general vehicles. Actually several domain experts recommended the five aspects to us. For each aspect, the product reputation mining method will label the reputation tendency to either positiveness or negativeness. The reputation level indicates the intensity of the reputation tendency. In our context, there are five levels in positiveness, neutrality, and five levels in negativeness. The five levels are specified to Level 1 $\sim$ Level 5. The strongest positiveness (negativeness) is Level 5, while the weakest positiveness (negativeness) is Level 1.

Next, to tackle the second problem, the product reputation mining method finds top–k documents including both *relevant* and *informative* reviews in the corpus. The top–k documents seldom contain meaningless advertisement and exaggeration, spam/fake reviews, and even text content which is not directly related to e. Here are two examples that we collected in the most popular web site with many reviews regarding vehicles in Korea. The following document is considered to be both relevant and informative.

> Title: Kia's next-generation K5 does not change but actually changed everything
> Author: Charisma4097
>
> The interior was completely obscured and difficult to identify, but the overall shape could be guessed. Once the change is expected to be quite large. Unlike the existing design that surrounds the driver's seat, it is likely to change to a horizontal feeling that stretches from the driver's seat to the next seat. The door design is completely different from that of the existing K5. Steeply raised buttons and knobs are flat. Sheets are very similar to those in the new Sonata. It is characterized by large and wide, with the middle vertical line. However, I am not sure that I will use this sheet similar to the new Sonata.

On the other hand, the following document shows the typical document that does not help consumers at all.

> Title: Someone who takes the new K5, What about the breaks?
> Author: Mr. Oral
>
> While I am getting ready to change from SM5 to K5, I wonder if there are too much talk about the bad breaks. I also wonder what K5 Turbo JBL sound is like. Once I heard from Mark Levinson audio in Lexus, it was so cool.

In practice, the sentiment analysis is the most important step in the product reputation mining problem. The sentiment analysis is generally categorized to two approaches [16]. One is the machine learning approach and the other is the lexicon–based approach that is also divided to dictionary–based and corpus–based approaches. Nowadays, even though main deep learning models such as CNN and RNN used for sentiment analysis have shown better results, the lexicon–based approach is still important. In a particular domain, the accuracy of the lexicon–based approach is much higher than machine learning approaches. Thus, sentiment analysis based on sentiment lexicons has been widely used in practical applications. In addition, the lexicon–based approach has no need for complex environment setting like GPU or long pre–training time before the learning model is used [1]. More importantly, for greater accuracy, the existing deep learning models need large–scale training data set. In fact, it is non–trivial to obtain the large–scale training data because human annotators need to label the classes manually. To avoid this problem, state–of–the–art researches are being carried out to pseudo–generate the large–scale training data using sentiment lexicons. Due to these reasons, the lexicon–based approach is still the important methodology in sentiment analysis. In this study, we focus only on the improvement of the existing lexicon–based approaches [7].

4. Main Proposal

To address the product reputation mining problem, we first consider the three dimensional coordinate system in which x-axis, y-axis, and z-axis indicate the aspect matching score, sentence importance score, and word sentiment score of a product of interest, respectively. In our problem, given a particular product, the three various scores are first measured by our proposed aspect classification method, sentence weight estimation method, and word sentiment scoring method, and then are aggregated to its total sentiment score.

Now we briefly summarize the key concept of the proposed three methods – the aspect classification method, sentence weight estimation method, and word sentiment scoring method. We will also describe the detailed algorithms in the following subsections. In our aspect classification method, we assume that five aspects of cars such as design, performance, price, quality, and service are given in advance. The aspects were manually decided by several experts in the automobile domain. Given a review document as input, it is divided to a set of sentences. Each sentence is automatically classified to one of the five aspects.

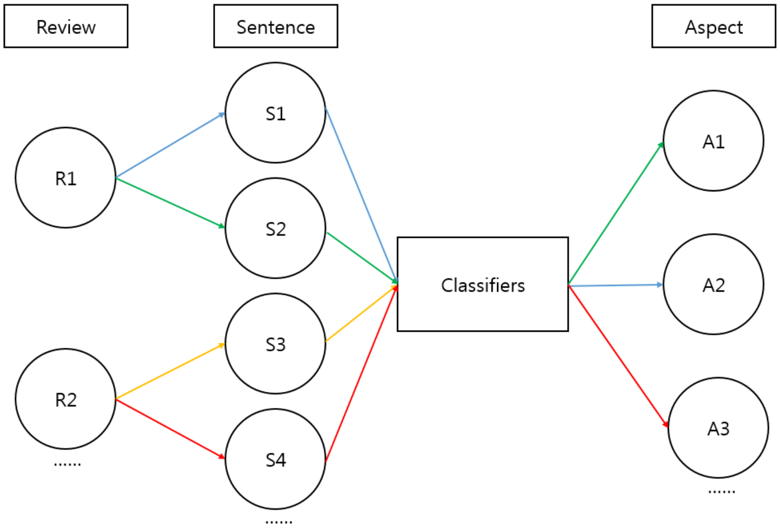


Fig. 2. Diagram of our aspect matching method

Figure 2 shows the diagram of the proposed aspect matching method and the detailed algorithm is depicted in Algorithm 1. The train set is a set of pairs like (sentence, one of five aspects (i.e., price, performance, design, service, and quality)) that is stored as a list of nodes, each of which contains a sentence and an aspect, in the main and secondary storage. To train a learning model and to conduct the test step, we use Support Vector Machine (SVM)[5], Random Forest[3], and Feed-forward Neural Network (FNN)[15]. SVM was the best classification method before deep learning models are employed actively. Random Forest often provides high accuracy because it is the best ensemble method. Unlike the conventional classification methods, it is known that FNN works effectively because of the deep neural network with multi hidden layers when we attempt to cope with the non-linear classification problem. Since each model has its pros and cons, the three learning models are used to assign sentences to the desired aspects.

Subsequently, to identify important sentences in a review document, we propose a graph-based model for estimating sentence weight values. After the review document is segmented to a set of sentences, each sentence is represented as a vertex in a graph G. The link weight between two nodes n_1 and n_2 is the similarity value $sim()$ between the

Algorithm 1 Aspect Matching Method

Require: The whole review data set about a given product DS;
Ensure: Several sub corpus with the corresponding to the labels;

```
Use classifiers to classify sentences in DS:
if sentences $\in$ classifier 1 then
    Align sentences to sub corpus about label 1;
else
    Send to classifier about label 2 to be classified;
    if sentences $\in$ classifier 2 then
        Align sentences to sub corpus about label 2;
    else
        Send to classifier about label 3;
                        ......
        if sentences $\in$ classifier n then
            Align sentences to sub corpus about label n;
        else
            These sentences are irrelevant;
        end if
        ......
    end if
end if
```

sentences corresponding to n_1 and n_2. If $sim(n_1, n_2) < \theta$ (a certain threshold value), the link between n_1 and n_2 is removed in G. To measure the connection strength between n_1 and n_2, our method is to compute the probability value to reach n_2 from n_1 via random walks over G, where for each step, random walkers visit neighbouring nodes with a certain probability. This graph-based method is based on the underlying assumption that more important sentences are likely to receive more links from other sentences. We will discuss the similarity and graph–based probability equations in Section 4.1.

Next, in word–level, we focus on both importance and sentiment orientation of words in a review document. To quantify the importance of a word, we use Term Frequency / Inverse Document Frequency (TF/IDF) that is widely used in the information retrieval community. Through TF/IDF metric, a word w is considered to be important if w appears many times in a document, while w seldom appears in the entire corpus. To compute the sentiment degree value of w, we constructed a sentiment lexicon for Korean language with assistance from Korean linguists. As illustrated in Figure 3, the Korean sentiment lexicon consists of the list of positive and negative words, incrementer and decrementer, flip words, and conjunction words. Based on the sentiment lexicon, we propose a word–level sentiment scoring method that computes the final score by merging the TF/IDF and sentiment scores of w. For the detail, we will discuss the detailed algorithm in Section 4.1.

Finally, after the above three methods are performed, each sentence is assigned to `(word--level score, sentence--importance score)` called *sentence reputation score (v)*. For each aspect a, the total score v_p of all positive sentences related to a are calculated. In the same way, the total score v_n of all negative sentences related to a are calculated as well. Then, the reputation tendency and level are approximated based on v_p and v_n. Furthermore, the document reputation score v_d is computed by $\Sigma_{i=1}^{k} v_i$, where k

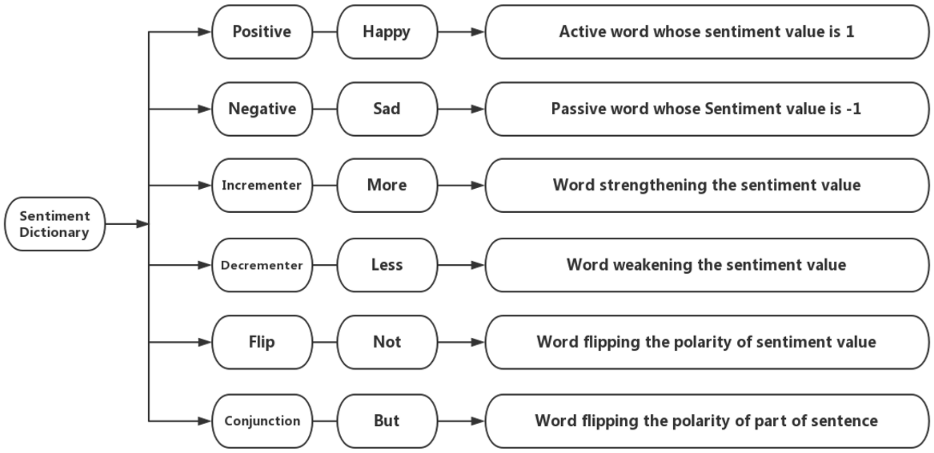

Sentiment Dictionary
Positive
Happy
Active word whose sentiment value is 1
Negative
Sad
Passive word whose Sentiment value is -1
Incrementer
More
Word strengthening the sentiment value
Decrementer
Less
Word weakening the sentiment value
Flip
Not
Word flipping the polarity of sentiment value
Conjunction
But
Word flipping the polarity of part of sentence

Fig. 3. A Korean sentiment lexicon

is the number of sentences in the document. In the final step, all review documents in the corpus are rearranged by v_d and then top-k review documents are chosen as informative ones.

Suppose that n is the number of sentences in the collection of reviews as input. Because we propose a FNN-based aspect matching model and compare it to the existing learning models such as SVM and Random Forest, we focus merely on computing the time complexity of FNN. Please refer to Table 3 in the paper. According to the table, each word of the sentence is converted to a 100-dimensional word embedding vector. If each sentence is composed of ten words, the dimension number of the input vector is 1,000. If the number of words is below ten, we put zero values to empty dimensions (as a padding approach). We develop the FNN model with five hidden layers (H1, ..., H5) that contain 1,000, 800, 600, 400, and 200 units, respectively. The input layers has 1,000 units and the output layer has 5 units (# of aspects). The number of the weight parameters between the input layer and H1 is 1,000∗1,000 and the number of biases between them is 1,000. Similarly, the number of the weight parameters between H1 and H2 is 1,000∗800 and the number of biases between them is 800. As a result, the total number of parameters is (1,000∗1,000+1,000) + (1,000∗800+800) + (800∗600+600) + (600∗400+400) + (400∗200+200) + (200∗5+5) = 2,604,005. This means that at least 2,604,005 memory spaces are required in both train and test sets. Since FNN model finds optimal parameters through forward and backward propagation, the time complexity is dominant to the number of computing the parameters between the input layer and H1. In other words, in case of the number of units in the input layer is n, it takes $O(n * n + n) = O(n^2)$.

4.1. Aspect Classification Models

Sentence Importance Estimation Method In this section, we present the similarity and graph-based probability equations by which the importance of each sentence in review documents is quantified. The similarity equation (sim) between two sentences s_i and s_j is defined as:

$$sim(s_i, s_j) = \frac{|\{w_k | w_k \in s_i, w_k \in s_j\}|}{log(|s_i|) + log(|s_j|)} \tag{1}$$

Eq. (18) means that the numerator is the number of the overlapped words between two sentences and the denominator is the length of the two sentences mainly used to normalize the similarity score. This proposed similarity measure is reasonable to see how similar the two sentences are and the meaning of the proposed similarity method is close to Jaccard similarity measure that is simple but yet high-accurate so is widely used in real applications. In Eq. (18), $|s_i|$ and $|s_j|$ are the numbers of words in s_i and s_j and the numerator indicates the number of the words appearing in both s_i and s_j.

To measure the importance of each node in a graph, where a node stands for a sentence, we refer to as:

$$\upsilon_i = (1 - \gamma)\frac{1}{n} + \gamma \Sigma_{j \in degree(i)} \frac{sim(i, j)}{\Sigma_{k \in degree(i) \wedge k \neq j} sim(i, k)} \upsilon_j \tag{2}$$

Eq. (19) is proposed to identify the global sentiment score of each sentence. It works based on random walks, where the local sentiment score of a sentence (e.g., x) is propagated to neighbor sentences (e.g., y and z) with its probability values (similarity between x and y, similarity between x and z). For example, suppose that xs sentiment score is 0.9; the weight between x and y is 0.7; and the weight between x and z is 0.3. In this case, a random walker visits to y from x at a probability of 0.7, while it also visits to z from x at a probability of 0.3. We believe that the equation makes sense to find each sentences optimal score by considering both the importance and sentiment of all sentences in the collection. In Eq. (19), i, j, and k are nodes and $degree(i)$ means a set of the neighbouring nodes of i. n is # of nodes in the graph and γ is the weight value of each equation term. Starting at node i, random walks continue to visit the neighbouring nodes until they arrive at all nodes in the graph to compute the probability value of i which is denoted by $P(i)$. The bigger $P(i)$ is, the higher the importance of i is. This is, if one node is pointed by important nodes in the graph, it may also be an important node. In Eq. (19), the first term $\frac{1}{n}$ needs because a random walker jumps to another node chosen at random with the equal probability whenever it meets terminal nodes in the graph. To quantify the weight of each sentence in a given corpus, the similarity between two sentences is computed and stored as a square matrix, where each row(column) means a sentence. In addition, the sentiment scores of all sentences are stored in a vector. The matrix-vector multiplication is performed iteratively until the values of the vector are converged. Suppose that n is the number of sentences in the collection of reviews as input. A n by n matrix and a n-dimensional vector are created, where the matrix contains the similarity values between two sentences and the vector means the sentiment score of each sentence. The space complexity is $O(n^2 + n) = O(n^2)$. The algorithm is peformed iteratively until there in no difference between the previous and current values in the vector. If we consider k to be the average number of iterations, the algorithm does the matrix-vector multiplication by k times. In each matrix-vector multiplication, the total number of the multiplications is n*n and the total number of the additions is n-1. For n rows in the matrix, the multiplications and additions are needed so $O(n(n * n + (n - 1))) = O(n^3)$. As a result, the time complexity is $O(k * n^3)$. Algorithm 2 shows the detailed algorithm of quantifying the importance of sentences.

Algorithm 2 Sentence–level Method

```
Require: One sub corpus processed by Algorithm 2: C;
Ensure: Reputation score for each sentence: $srs$;

Each sentence in the corpus gains a $tr$ by Eq. (19);
for sentence in C do
    $srs = ssc * tr$;
end for
```

Word Sentiment Scoring Method This method consists of two terms of the equation. For each word w, one is to measure the importance of w and the other is to measure the sentiment score of w. To quantify the importance of w, we use TF/IDF metric, where TF is Term Frequency meaning the frequency of a word within a document. For example, if a word ‘obama’ appears three times in a document that has 100 words, TF(‘obama’)= $\frac{3}{100}$ = 0.03. On the other hand, IDF is Inverse Document Frequency, indicating that a word is more important if it is unique in the corpus. Suppose that we have 10 million documents in the corpus. If ‘obama’ appears in only 1,000 documents, then IDF(‘obama’)=$\frac{10,000,000}{1,000}$=4. As a result, TF/IDF(obama)=0.03 $\times$ 4=0.12. In this way, the weight value of each word is quantitatively computed using TF/IDF which is between 0 and 1. If the weight value of the word is close to 1, then it means that the word is very important. On the other hand, if the weight value is low, the corresponding word is trivial. Such a word may be ‘a’, ‘the’, ‘in’, and so on. This weight value of each word captures the importance of the word. Meanwhile, to compute the sentiment score of w, we constructed and used our own Korean sentiment lexicon as shown in Fig-

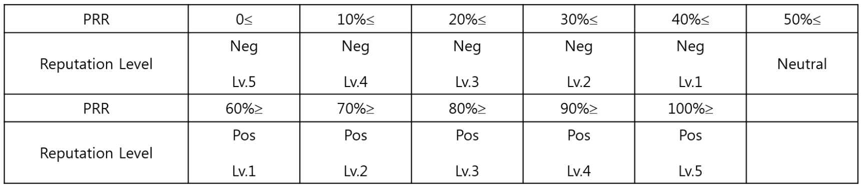
Algorithm 3 Word–level Method

```
Require: One sub corpus after aspect classification: C;
Ensure: Reputation score for each sentence: $ssc$;

Each word in the corpus gains a $ti$ value by TF/IDF;
for sentence in C do
    def sentence_rs(sentence_tokens, pw, nw, $ssc$):
    if not sentence_tokens then
        return $ssc$;
    else
        cw = sentence_tokens[0];
        Gain the $wss$ of cw from Rules;
        $ssc = ssc + wss * ti$;
        if nw $\in$ Conjunction dictionary then
            $ssc = ssc * (-1)$;
        end if
        return sentence_rs(sentence_token[1:], cw, nw, $ssc$)
    end if
end for
```

ure 3. In particular, the sentiment dictionary contains positive and negative words, incrementer/decrementer, flip words, and conjunction words. The final word–level score(w) = TF/IDF(w) × sentiment–score(w). Each sentence is tokenized to words and then is stored as a list of pairs like (Sentence ID, [$word_1$, $word_2$, $word_3$, ...]). In addition, HashMap is used, where the key is (Sentence ID, $word_i$) and the value is (TF/IDF and sentiment scores of $word_i$). Suppose that n is the number of sentences in the collection of reviews and each sentence contains m words on average. In the first step, all TF/IDF and sentiment scores are computed by n∗m times. In the second step, Algorithm 3 is performed by n∗m times. Thus, the time complexity is $O(n * m)$. Meanwhile, to store a list that contains the pairs of (Sentence ID, Words), it needs n∗m+n spaces, where n∗m means the total number of words in sentences and n means the number of the sentences identifiers. We also need a HashMap, where each key needs 2 for storing a sentence ID and each word, and each value needs 2 for storing TF/IDF and sentiment scores. Thus, the HashMap needs $O(n * m(2 + 2)) = O(n * m)$ spaces. As a result, the total space complexity is $O((n * m + n) + (n * m)) = O(n * m)$. Algorithm 3 describes the detailed procedure.

Estimation of Reputation Tendency and Level As the final result, each sentence is labelled to sentence reputation score (υ)=(`word--level score, sentence--importance score`). For each aspect a, the total score υ_p of all positive sentences related to a are calculated and then the final reputation score is estimated based on $\frac{\upsilon_p}{|\upsilon_p|+|\upsilon_n|}$. In the same way, the total score υ_n of all positive sentences related to a are also calculated and then the final reputation score is estimated based on $\frac{\upsilon_n}{|\upsilon_p|+|\upsilon_n|}$. Finally, the reputation scores are transformed to the relevant reputation tendency and level based on the index table in Figure 4.



PRR	0≤	10%≤	20%≤	30%≤	40%≤	50%≤
Reputation Level	Neg Lv.5	Neg Lv.4	Neg Lv.3	Neg Lv.2	Neg Lv.1	Neutral
PRR	60%≥	70%≥	80%≥	90%≥	100%≥	
Reputation Level	Pos Lv.1	Pos Lv.2	Pos Lv.3	Pos Lv.4	Pos Lv.5	

Fig. 4. Reputation tendency and level index

To find informative review documents, the document reputation score υ_d is computed by $\Sigma_{i=1}^{k}\upsilon_i$, where k is the number of sentences in the document. In the final step, all review documents in the corpus are rearranged by υ_d and then top-k review documents are chosen as the informative documents.

5. Experimental Validation

5.1. Experimental Set–up

In the previous section, we described the detailed algorithms of the proposed approach for computing the reputation tendency and level and selecting top–k informative sentences about a target product. Now we introduce the process of evaluating the proposed method, comparing to a straightforward lexicon–based approach as the baseline method with online reviews about K5 in Kia motors. We collected 1,585 review documents in Bobaedream, the most popular web sties related to car reviews. In the pre–processing step, we replaced all words by lower–case letters after removing images, moving pictures, and advertising texts. Then, we removed stop words [24] in the all documents and converted derived words to root forms through a stemming software [17]. After the pre–processing step, we collected 1,562 review sentences. To make the gold standard set (solution set), four human annotators subjectively labelled the aspect of each sentence to one of five aspects (design, performance, price, quality, and service) and conflicting sentences are decided by a majority vote. In the same way, they manually classified all sentences to a particular sentiment orientation (positive, neutral, and negative). For example, given a sentence "The front design with Raff is very good," the sentence orientation is positive and the aspect label is design. Figure 5 shows the brief characteristics of the data set.

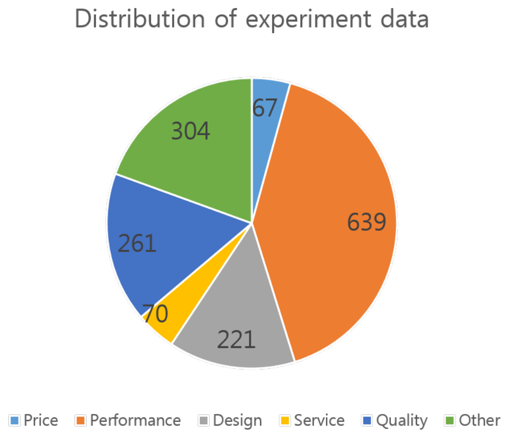


Fig. 5. Distribution of the review documents across five aspects

To select the discriminative features of input vectors, we first computed TF/IDF values of all words in the data set, and then used top–k words with the highest TF/IDF values as the feature set. For example, # of the words in the feature set is 1,000. In our repetitive experiments, we carefully investigated the results of all methods for all possible cases to find the optimal number of the features in the data set. Finally, after making feature vectors based on the feature set in the data set, we converted the feature vectors to the input vectors, which is the input of the models used in our experiment, using a popular word embedding method such as Word2Vec [20].

We implemented the aspect matching method based on FNN deep learning model in Python and TensorFlow [6]. The experimental set–up of the method used in our experiments is summarized in Table 2. Through our intensive experiment, we found the optimal values of the hyper parameters that are suitable in our problem. For the initial values of weight parameters, we used the truncated normal method [14]. As an activation function, ReLU was used in the entire layers except the output layer in which the activation function was SoftMax function. We also made use of cross entropy as loss function. To improve the accuracy of the models, we used dropout and regularization techniques in addition to Adam optimizer for carrying out backward propagation of errors. After completing the implementation of the deep learning model, we attempted to find the best dropout and learning rates. To validate the effectiveness of the aspect matching method, we compared the results of SVM [11], Random Forest [2], and FNN. The number of classes in the data set is 6. Through cross-validation in the training step, all sentences were divided into five run sets. Each model had been first trained with the four run sets and then classified each sentence in the rest set to one of the six aspects. Changing the order of the run sets, we performed the train and test steps five times, and measured the average accuracy, precision, recall, and F_1–score of the models. Each model was in standalone executed in a high-performance workstation server with Intel Xeon 3.6GHz CPU with eight cores, 24GB RAM, 2TB HDD, and TITAN-X GPU with 3,072 CUDA cores, 12GB RAM, and 7Gbps memory clock.

For the evaluation metric, we used accuracy, precision, recall, F_1–score measures that have been widely used in IR community. To measure the precision and recall values of a classification model, we first consider a confusion matrix of classes $M_{i,j}$, where each row of the confusion matrix represents predicted class, while each column represents actual class. n is the number of classes. True positive, False positive, and False negative in each class are represented as Eq. (3).

$$\begin{array}{l} \text{True positivie}_i = M_{i,i} \\ \text{False positivie}_i = \sum_{k=1}^{n} M_{i,k} | k \neq i \\ \text{False negative}_i = \sum_{k=1}^{n} M_{k,i} | k \neq i \end{array} \quad (3)$$

Based on Eq. (20), the precision, recall, and F_1–score (Harmonic mean between precision and recall) are defined as:

$$\begin{array}{l} \text{Precision} = \sum_{k=1}^{n} \frac{\text{True positivie}_i}{\text{True positivie}_i + \text{False positivie}_i} \\ \text{Recall} = \sum_{k=1}^{n} \frac{\text{True positivie}_i}{\text{True positivie}_i + \text{False negative}_i} \\ F_1\text{–score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\{\text{Precision} + \text{Recall}\}} \end{array} \quad (4)$$

Table 2. Experimental set–up for the used models

Methods	Experimental set-up
SVM	Through many experiments, the optimal trade-off value between training error and margin was selected in each data set
Random Forest	Through Many experiments, the optimal # of trees in the forest & max depth of the tree were selected in each data set
FNN	Batch size=50, Adam optimizer(learning rate=0.01), dropout rate=0.5, 5 hidden layers H_1, H_2, H_3, H_4, and H_5 – H_1 contains 1,000 units; and H_2 contains 800 units;5 hidden layers H_1, H_2, H_3, H_4, and H_5 – H_1 contains 1,000 units; and H_2 contains 800 units; H_3 contains 600 units; H_4 contains 400 units; H_5 contains 200 units

5.2. Experimental Results

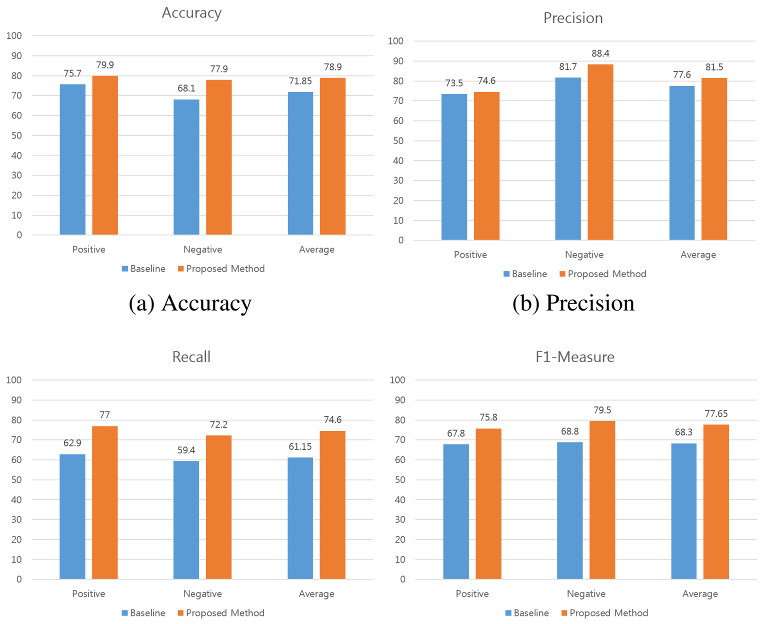
Table 3. Accuracy of three aspect matching models based on SVM, Random Forest, and FNN

Aspect	Price	Performance	Design	Service	Quality
FNN	95.7	85.2	93.9	94.6	86.4
SVM	97.3	73.4	89.6	96.1	84.1
Random Forest	96.2	70.6	88.6	95.6	84.2

Result of Aspect Matching Method Table 3 summarizes the average accuracy scores of the three aspect matching models based on SVM, Random Forest, and FNN. By and large, the average accuracy values are high for all aspects. For example, the accuracy of the performance aspect is at least 70.6% in Random Forest. In the price aspect, the accuracy of SVM is up to 97.3%. In three aspects such as performance, design, and service, FNN outperforms both SVM and Random Forest. Interestingly, we observed that the deep learning model like FNN is better than the conventional learning models such as SVM and Random Forest in the aspects including many sentences. In contrast, the price and service aspects have the small number of sentences. In these aspects, SVM is better than the deep learning model. However, the gap of the accuracies in the different learning models is not large. In the data set, a relatively large number of sentences are related to the performance and quality aspects. In general, many sentences in such aspects are often ambiguous because they may be semantically interpreted to other aspect. Thus, developing more intelligent aspect matching models is still challenging and there is room to improve the accuracy of the best learning models.

Sentiment Analysis of the Proposed Method Figure 6 shows the average accuracy, precision, recall, and F_1–scores of the proposed method, comparing to the baseline method that is the typical lexicon–based approach in the sentiment analysis. To find the preference for a particular product, the baseline approach collects (1) review posts, which are related to the product, from several product review web sites; (2) extracts sentences in the collection after the pre–processing step such as stemming and removal of stop words is performed; (3) classifies the polarity (either positive or negative sense) of each sentence based on a sentiment lexicon; and (4) estimates the positive and negative ratios of the product by dividing the total numbers of the positive and negative sentences by the total number of the sentences in the collection. Furthermore, the baseline approach automatically finds important sentences including the positive and negative meaning to/against the product.

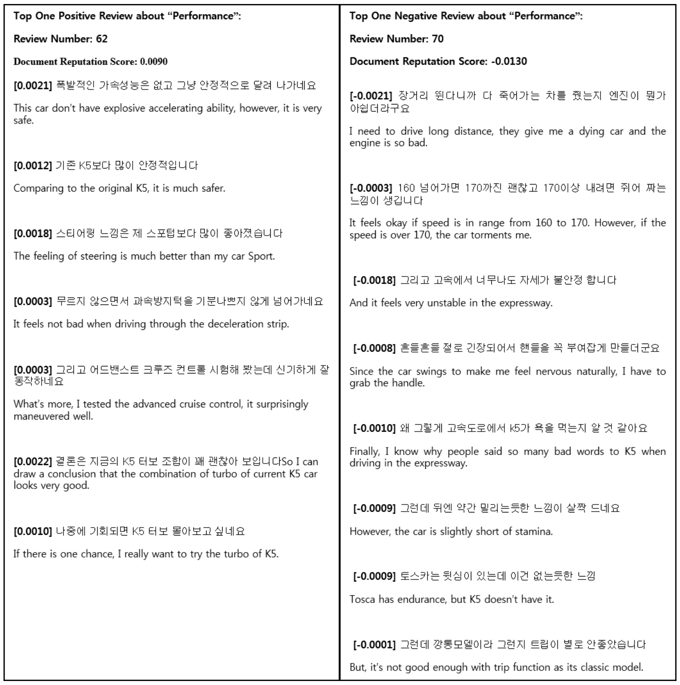
As a motivated example, given a product like Hyundai Sonata, customers often want to see the summary note including what positive points are and what negative points are in the 'car design' aspect. They also want to gain more useful information regarding other aspects such as 'car quality,' 'car performance,' and 'car service.' Such an information will enable customers to make good choice when they attempt to purchase their brand–new cars. In addition, car makers will be able to figure out the public's preferences and positive/negative points for new models on market. In the near future, the weak points of the models will be improved by the sentiment analysis. For this, the baseline approach computes the sentiment score of each sentence and then selects top–k sentences with the highest positive and negative scores. In the figures, the experimental results show that the proposed method outperforms the baseline method in all evaluation metrics. For instance, the average accuracy scores of the baseline method are 75.7% and 68.1% in positive and negative sentences, while those of the proposed method are 79.9% and 77.9% in positive and negative sentences. This indicates that the proposed method improves about 5% and 14% accuracies, compared to the baseline method. Similarly, the average F_1–scores of the baseline method are 67.8% and 68.8% in positive and negative sentences, while those of the proposed method are 75.8% and 79.5% in positive and negative sentences. This implies that the proposed method improves about 12% and 16% F_1–scores, compared to the baseline method. The main reason why the proposed method outperforms the baseline method is that three dimensions (word, sentence, and aspect–levels) are considered to find the reputation tendency and level. In addition, the proposed word–level method considers both importance and sentiment orientation of words, while the baseline method focuses only on measuring the sentiment orientation of words. Another reason is because the proposed method aggregates additional information about the importance of sentences in order to determine the reputation tendency and level. Besides, through the aspect matching method, because most sentences are first categorised to the right aspect, the rest methods have a little chance to get confused to estimate the sentiment scores of the sentences.



(a) Accuracy (b) Precision

(c) Recall (d) F_1–score

Fig. 6. Comparison of the proposed method to the existing lexicon–based approach



Top One Positive Review about "Performance":

Review Number: 62

Document Reputation Score: 0.0090

[0.0021] 폭발적인 가속성능은 없고 그냥 안정적으로 달려 나가네요

This car don't have explosive accelerating ability, however, it is very safe.

[0.0012] 기존 K5보다 많이 안정적입니다

Comparing to the original K5, it is much safer.

[0.0018] 스티어링 느낌은 제 스포텝보다 많이 좋아졌습니다

The feeling of steering is much better than my car Sport.

[0.0003] 무르지 않으면서 과속방지턱을 기분나쁘지 않게 넘어가네요

It feels not bad when driving through the deceleration strip.

[0.0003] 그리고 어드밴스트 크루즈 컨트롤 시험해 봤는데 신기하게 잘 동작하네요

What's more, I tested the advanced cruise control, it surprisingly maneuvered well.

[0.0022] 결론은 지금의 K5 터보 조합이 꽤 괜찮아 보입니다So I can draw a conclusion that the combination of turbo of current K5 car looks very good.

[0.0010] 나중에 기회되면 K5 터보 몰아보고 싶네요

If there is one chance, I really want to try the turbo of K5.

Top One Negative Review about "Performance":

Review Number: 70

Document Reputation Score: -0.0130

[-0.0021] 장거리 뛴다니까 다 죽어가는 차를 줬는지 엔진이 뭔가 아쉽더라구요

I need to drive long distance, they give me a dying car and the engine is so bad.

[-0.0003] 160 넘어가면 170까진 괜찮고 170이상 내려면 쥐어 짜는 느낌이 생깁니다

It feels okay if speed is in range from 160 to 170. However, if the speed is over 170, the car torments me.

[-0.0018] 그리고 고속에서 너무나도 자세가 불안정 합니다

And it feels very unstable in the expressway.

[-0.0008] 흔들흔들 절로 긴장되어서 핸들을 꼭 부여잡게 만들더군요

Since the car swings to make me feel nervous naturally, I have to grab the handle.

[-0.0010] 왜 그렇게 고속도로에서 k5가 욕을 먹는지 알 것 같아요

Finally, I know why people said so many bad words to K5 when driving in the expressway.

[-0.0009] 그런데 뒤엔 약간 밀리는듯한 느낌이 살짝 드네요

However, the car is slightly short of stamina.

[-0.0009] 토스카는 뒷심이 있는데 이건 없는듯한 느낌

Tosca has endurance, but K5 doesn't have it.

[-0.0001] 그런데 깡통모델이라 그런지 트립이 별로 안좋았습니다

But, it's not good enough with trip function as its classic model.

Fig. 7. Top–1 positive and negative review documents

A Case Study of Top–k informative review documents For each aspect, both enterprise executives and customers would like to know the summary of the detailed reviews. If they go over the review summary, they can know the reasons why customers really like the product and what inconvenient points exist to be improved. The proposed method provides top–k documents of the most informative reviews. To validate whether top–k informative reviews are really useful for producers and consumers, we conducted a case study of K5 in Kia motors.

The left figure in Figure 7 shows the top-1 document of positive reviews in the aspect of performance. The identifier of the review document is 62 and the document reputation score is 0.009 that is the sum of the scores of the six sentences in the document. Each sentence also shows the reputation score estimated by our proposed method. For instance, 0.0021 is the reputation score of the first sentence – "This car don't have explosive accelerating ability, however, it is very safe." The top–1 review document contains positive but yet informative meanings. Similarly, the right figure in Figure 7 shows the top–1 document of negative reviews in the same aspect. The top-1 review document contains negative but yet informative meanings. These results clearly show that the top-1 review documents are considerably informative. These review documents will help both producers and consumers figure out the detailed pros and cons of the product that they really want to know in the marketing research.

A User Study and Statistical Verification To validate the effectiveness of our proposed method, we first had interviewed with 30 volunteers who had nothing to do with the authors in this article and are willing to respond to this survey. For each aspect, each interviewee took a look at five sentences chosen at random which are related to the reputation level generated by the proposed method. The interviewee chose one of (i) agree, (ii) disagree, and (iii) N/A to see how much he/she agrees to the results. Figure 8 illustrates the survey results of the six aspects. Y-axis indicates the ratios of agree, disagree, and N/A answers from all interviewees. In the figure, it is obvious that the majority of interviewees agreed to the reputation level, especially in the aspects of design, performance, and service, while it seems that more people disagreed to the reputation level in the quality aspect.

Aspect	PRR	Reputation Level
Price	49.6%	Negative Level 1
Performance	50.4%	Positive Level 1
Design	63.2%	Positive Level 2
Service	20.8%	Negative Level 3
Quality	41.6%	Negative Level 3

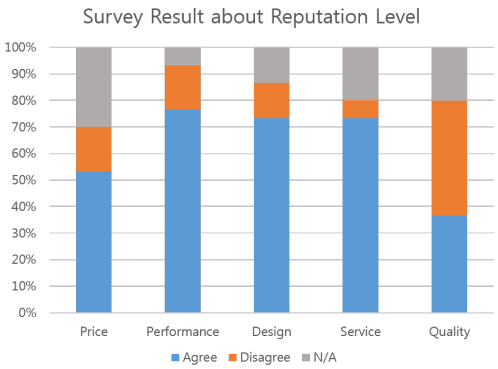
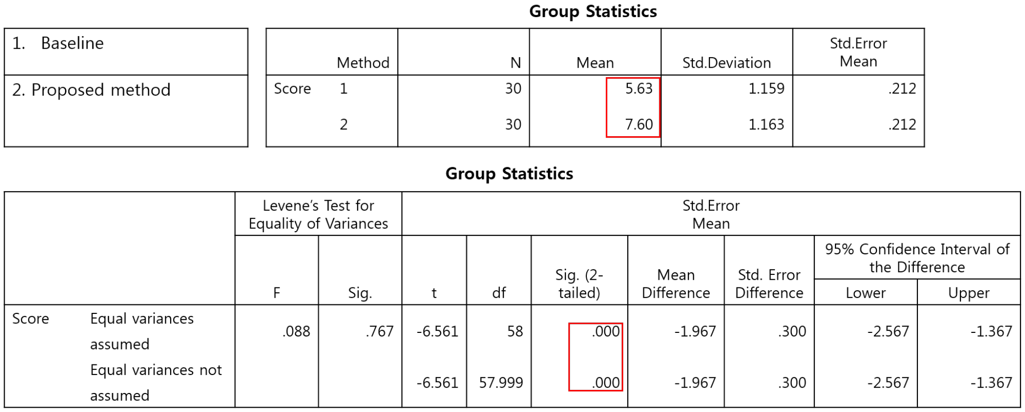


Fig. 8. Results of user study

In addition, we conducted additional survey for top–k informative documents of reviews. In the performance aspect, we prepared top–1 review documents retrieved by the baseline method and the proposed method and showed them to 30 interviewees who gave a score in range from 1 to 5 to each selected document to see how informative it is. We conducted the significance test using IBM-SPSS Statistics 21 and Figure 9 shows the statistical results. We compared the proposed method to the baseline method. When the significant level is 0.05, the null hypothesis H_0 is no statistical difference between the two methods and the alternative hypothesis H_1 is the significant difference between them. According to our Levene's test and t–test results, H_1 is accepted, indicating that the proposed method is statistically different from the baseline method because the p–value is extremely close to 0 and smaller than the significance level. In addition, the interviewees thought that the proposed method is better because the mean score of the proposed method is higher than the baseline method.

6. Concluding Remarks and Future Work

In this work, we propose a novel method of determining the reputation tendency and level and selecting top–k informative review documents about a particular product. This



1. Baseline
2. Proposed method

Group Statistics

	Method	N	Mean	Std.Deviation	Std.Error Mean
Score	1	30	5.63	1.159	.212
	2	30	7.60	1.163	.212

Group Statistics

		Levene's Test for Equality of Variances		Std.Error Mean						
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
									Lower	Upper
Score	Equal variances assumed	.088	.767	-6.561	58	.000	-1.967	.300	-2.567	-1.367
	Equal variances not assumed			-6.561	57.999	.000	-1.967	.300	-2.567	-1.367

Fig. 9. Statistical test results

product reputation mining approach can help both producers and consumers understand the product well. Unlike the existing lexicon–based approach, our proposed method is based on three dimensional points of word–level, sentence–level, and aspect–level views. In each level, the sentiment orientation of the product is quantified in addition to the consideration of the importance of words and sentences. In addition, the aspect matching process can be helpful in measuring the sentiment orientation of the product. To the best of our knowledge, our method is new, compared to the existing lexicon–based approach. Our experiment results show the the proposed method outperforms the baseline method and we also validated the proposed method through user study and statistical verification tasks.

For our future work, we have a plan to develop a web–based prototype system for the demonstration. We will also apply our method to other domains like smart phones and cosmetic products. Finally, we will propose an automatic method of mining main aspects about a particular product.

Acknowledgments. This work was supported by the National Research of Korea (NRF) grant funded by the Korea government (MSIT) (No. NRF-2019R1F1A1060752) for Byung-Won On, and by the Ministry of Trade, Industry Energy (MOTIE, Korea) under Industrial Technology Innovation Program, No. 10063130 and MSIT(Ministry of Science and ICT), Korea, under the ITRC(Information Technology Research Center) support program (IITP-2019-2016-0-00313) supervised by the IITP(Institute for Information communications Technology Promotion) for Gyu Sang Choi.

References

1. Bhonde, R., Bhagwat, B., Ingulkar, S., Pandc, A.: Sentiment analysis algorithms based on dictionary approach. International Journal of Emerging Engineering Research and Technology 3(1), 51–55 (2015)
2. Blondel, M.: Random forest classifier. In: http://scikit-learn.org/.../sklearn.ensemble.RandomForestClassifier.html (2017)
3. Breiman, L.: Random forests. Machine Learning 45(1), 5–32 (2001)

4. Canales, L., Strapparava, C., Boldrini, E., Martinez-Barco, P.: A bootstrapping technique to annotate emotional corpora automatically. In: Proceedings of IEEE International Conference on Data Science and Advanced Analytics (DSAA 2016), Montreal, Canada. IEEE (October 17–19, 2016)
5. Cortes, C., Vapnik, V.: Support-vector networks. Machine Learning 20(3), 273–297 (1995)
6. Google: Tensorflow. In: https://www.tensorflow.org/ (2018)
7. Grimmer, J., Stewart, M.B., Alvarez, M.: Text as data: The promise and pitfalls of automatic content analysis methods for political texts. Political Analysis 21(3), 267–297 (2013)
8. Hu, Y., Chen, Y., Chou, H.: Opinion mining from online hotel reviews – a text summarization approach. Information Processing & Management 53(2), 436–449 (2017)
9. Hwang, J., Ko, Y.: A korean sentence and document sentiment classification system using sentiment features. Korean Institute of Information Scientists and Engineers 14(3), 336–340 (2008)
10. Jin, W., Hung, H.: A novel lexicalized hmm-based learning framework for web opinion mining. In: Proceedings of the 26th International Conference on Machine Learning (ICML 2009), Montreal, Canada. ICML (June 14–18, 2009)
11. Joachims, T.: Support vector machine. In: https://www.cs.cornell.edu/people/tj/svm_light/ (2014)
12. Khose, N., Dakhode, V.: Product reputation analysis system based on partial supervised word alignment model. International Journal of Science and Research 5(8), 169–173 (2016)
13. Ko, M.: Semantic classification and weight matrices derived from the creation of emotional word dictionary for semantic computing. In: Proceedings of Emotion and Sentiment Analysis Workshop (ESA 2016), Portoroz, Slovenia (May 23, 2016)
14. LeCun, Y., Bottou, L., Orr, G.B., Muller, K.R.: Efficient backprop. In: Proceeding Neural Networks: Tricks of the Trade, this book is an outgrowth of a 1996 NIPS workshop. pp. 9–50. NIPS (1996)
15. Leshno, M., Vladimir Ya, L., Pinkus, A., Schocken, S.: Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. Neural Networks 6(6), 861–867 (1993)
16. Medhat, W., Hassan, A., Korashy, H.: Sentiment analysis algorithms and applications: A survey. Ain Shams Engineering Journal 5(4), 1093–1113 (2014)
17. Porter, M.: The porter stemming algorithm. In: https://tartarus.org/martin/PorterStemmer/index.html (2006)
18. Resnick, P., Kuwabara, K., Zeckhauser, R., Friedman, E.: Reputation systems. Communication of the ACM 43(12), 45–48 (2000)
19. Saif, H., He, Y., Fernandez, M., Alani, H.: Contextual semantics for sentiment analysis of twitter. Information Processing & Management 52(1), 5–19 (2016)
20. Skymind: Deeplearning4j. In: https://deeplearning4j.org/word2vec (2017)
21. Steinberger, J., Ebrahim, M., Ehrmann, M., Hurriyetoglu, A., Kabadjov, M., Lenkova, P., Steinberger, R., Tanev, H., Vazquez, S., Zavarella, V.: Creating sentiment dictionaries via triangulation. Decision Support Systems 53(4), 689–694 (2012)
22. Teng, Z., Vo, D., Zhang, Y.: Context-sensitive lexicon features for neural sentiment analysis. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP 2016), Austin, Texas (November 1–5, 2016)
23. Wang, J., Yu, L., Lai, K., Zhang, X.: Dimensional sentiment analysis using a regional cnn-lstm model. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016), Berlin, Germany (August 7–12, 2016)
24. Wikipedia: Stop_words. In: https://en.wikipedia.org/wiki/Stop_words (2018)
25. Wood, I., Ruder, S.: Emoji as emotion tags for tweets. In: Proceedings of Emotion and Sentiment Analysis Workshop (ESA 2016), Portoroz, Slovenia (May 23, 2016)
26. Yan, J., Turtle, H.: Emocues–28: Extracting words from emotion cues for a fine-grained emotion lexicon. In: Proceedings of Emotion and Sentiment Analysis Workshop (ESA 2016), Portoroz, Slovenia (May 23, 2016)

Zhehua Piao received his Master degree in Department of Computer Science and Engineering, Korea University, Seoul, South Korea. His recent research interests are around Data Mining and Machine Learning, mainly working on Product Reputation Mining, Opinion Mining, Sentiment Analysis and Sentiment Lexicon Construction.

Sang-Min Park is currently attending the Master program in Department of Software Convergence Engineering, Kunsan National University, Gunsan-si, Jeollabuk-do, Korea. His recent research interests are around Machine Learning and Data Mining, mainly working on AI-based Text Mining, Opinion Mining and Korean Sentiment Lexicon Construction.

Byung-Won On received his PhD degree in Department of Computer Science and Engineering, Pennsylvania State University at University Park, PA, USA in 2007. Then, he worked as a full-time researcher in University of British Columbia, Advanced Digital Sciences Center, and Advanced Institutes of Convergence Technology for almost seven years. Since 2014, he has been a faculty member in Department of Software Convergence Engineering, Kunsan National University, Gunsan-si, Jeollabuk-do, Korea. His recent research interests are around Data Mining and Databases, mainly working on AI-based Text Mining and Big Data Management Technologies. He is the corresponding author and can be contacted at: bwon@kunsan.ac.kr

Gyu Sang Choi received his PhD in Computer Science and Engineering from Pennsylvania State University. He was a research staff member at the Samsung Advanced Institute of Technology (SAIT) for Samsung Electronics from 2006 to 2009. Since 2009, he has been with Yeungnam University, where he is currently an associate professor. He is now working on non-volatile memory and storage systems, whereas his earlier research mainly focused on improving the performance of clusters. He is a member of ACM and IEEE. He is the corresponding author and can be contacted at: castchoi@ynu.ac.kr

Myong-Soon Park is Professor of Department of Computer Science and Engineering, Korea University, Seoul, South Korea. He received his BSc in Electronics Engineering from Seoul National University, an MSc in Electrical Engineering from the University of Utah in 1982, and a PhD in Electrical and Computer Engineering from the University of Iowa in 1985. He was an assistant professor at Marquette University from 1985 to 1987.1 and at Postech from 1987.2 to 1988.2. Since 1988.3 he has been an assistant, associate and full professor at Korea University until now. Professor Park was the chair of the SIG on parallel processing of KIISE (1997-2000) and has been on program committees for various international conferences. His research interests include sensor networks, internet computing, parallel and distributed systems, and mobile computing.

Received: July 3, 2018; Accepted: May 20, 2019.