Asia Pacific Journal of Information Systems
Vol. 29 No. 4 (December 2019), 838-855

ISSN 2288-5404 (Print) / ISSN 2288-6818 (Online)
https://doi.org/10.14329/apjis.2019.29.4.838

Special Issue: Data Analytics in Artificial Intelligence Era

Multidimensional Analysis of Consumers' Opinions from Online Product Reviews

Taewook Kim[a,*], Dong Sung Kim[b], Donghyun Kim[c], Jong Woo Kim[d,**]

[a] *M.S. Student, Philosophy & Computer Science, Hong Kong University of Science and Technology, Hong Kong SAR*
[b] *Postdoctoral researcher, Business Administration at the School of Business, Hanyang University, Korea*
[c] *Engineer, LG Electronics Company, Korea*
[d] *Professor, School of Business, Hanyang University, Korea*

ABSTRACT

Online product reviews are a vital source for companies in that they contain consumers' opinions of products. The earlier methods of opinion mining, which involve drawing semantic information from text, have been mostly applied in one dimension. This is not sufficient in itself to elicit reviewers' comprehensive views on products. In this paper, we propose a novel approach in opinion mining by projecting online consumers' reviews in a multidimensional framework to improve review interpretation of products. First of all, we set up a new framework consisting of six dimensions based on a marketing management theory. To calculate the distances of review sentences and each dimension, we embed words in reviews utilizing Google's pre-trained word2vector model. We classified each sentence of the reviews into the respective dimensions of our new framework. After the classification, we measured the sentiment degrees for each sentence. The results were plotted using a radar graph in which the axes are the dimensions of the framework. We tested the strategy on Amazon product reviews of the iPhone and Galaxy smartphone series with a total of around 21,000 sentences. The results showed that the radar graphs visually reflected several issues associated with the products. The proposed method is not for specific product categories. It can be generally applied for opinion mining on reviews of any product category.

Keywords: Multidimensional Analysis, Opinion Mining, Product Reviews, Sentiment Analysis

Ⅰ. Introduction

The masses of social media users provide a multitude of opinions. The subjects of these opinions vary across a wide spectrum, ranging from certain political issues, to governments, to specific products or even personal chats. Opinion mining extracts value from these texts through the use of semantic analysis. Such

*The work was done during the author's undergraduate years at Hanyang University.

**Corresponding Author. E-mail: kjw@hanyang.ac.kr Tel: 82222201067

mining identifies a sentiment from a given context like whether the text is positive or negative, and, if it is positive, how positive it is. Opinion mining is actively applied in many different types of documents such as those with large numbers of random comments, product reviews, articles and so on (Duan et al., 2013; Giachanou and Crestani, 2016; Iosifidis and Ntoutsi, 2017; Paul et al., 2017; Zhai et al., 2011). For example, we can even extract public opinion from comments which arbitrary internet users left on portal sites or online journals (Duan et al., 2013; Zhai et al., 2011).

In this paper, we focus on the use of opinion mining in product reviews. Companies are eager to utilize product reviews to refine the raw data from social media (Hu et al., 2006). As product reviews on social media indicate what people have in mind, they are an invaluable resource for companies to act upon. Although there are already various opinion mining approaches to extract value from these online reviews, they largely only measure the results in a single dimension. However, this is insufficient to fully represent reality, as consumers rarely consider only a single aspect of products. For example, one may appreciate the design and quality of a certain product while she or he might not feel the same about its price. To address this problem, we present a multidimensional analysis of consumers' opinions from online product reviews.

We designed a six-dimensional framework based on an existing marketing management theory (Kotler and Levy, 1969; Kotler et al., 2012). We specified criteria for each dimension to categorize the sentences of reviews. We mainly utilized the semantic distances between these criteria, the rules for each dimension, and the sentences of the product reviews. We vectorized all the words to measure the similarity between text datasets. Once each sentence was categorized, we conducted sentiment analysis for the sentences. We visualized those results in radar graphs by using numerical outcomes of sentimental analysis. We illustrated the whole procedure empirically with Amazon product reviews for the iPhone series (6s, SE, and 7) and Galaxy series (S6, S6 edge, and S7). We have plotted boxplots to show the distributions of sentences of reviews in each dimension. Also, we presented the reviews in a radar graph with the six axes of a multidimensional framework. Finally, we evaluated our proposed approach by comparing the F1 score of sentence classification result with the baseline, which is a manual classification done by trained participants.

The new approach is designed to be universally applicable to other product reviews. With our multidimensional approach, companies can obtain practical information which was not easy to glean from earlier methods. For example, the variations in customers' opinion by product generation can be identified with this new approach. Companies can also easily recognize the gap between their goals and the real market situation. Additionally, companies can obtain insights on their competitors' strategies and market trends by comparing the results of products to those of their competitors. This provides companies with clearer information to help them determine the direction they should take.

Ⅱ. Related Work

2.1. Market Sensing

Product are important for the success of a company in any given market (Kotler and Levy, 1969; Kotler et al., 2012). To successfully launch their products, companies need to be good at market sensing. Market

sensing is the process of gathering market intelligence and acting upon that information (Kotler et al., 2012). This is considered to be a critical aspect that is necessary for a company to survive in a competitive business environment (Day, 1994; Kim and Shin, 2015; Kim et al., 2017).

Before the concept of market sensing arose, there was a marketing mix called the 4Ps - Product, Place, Promotion, and Price (McCarthy et al., 1979). The marketing mix included the "product" as one of its components. This strategy suggests that the product actually plays a key role in business management. Although it is crucial that companies have clear pictures of their products, it is also highly important that companies know how their consumers view their products (Kotler et al., 2012). Therefore, there have been many studies covering marketing and market sensing (Kordupleski, 2003; Piercy, 2016).

Philip Kotler has pointed out that classifying products by their forms is the starting point for broadening the concept of marketing (Kotler and Levy, 1969). After that, he even suggested a framework that involved Levels of Products, which helps to identify distinct points of products (Kotler et al., 2012). Further, James Adams specified the multiple dimensions of a product to optimize its quality (Adams, 2012).

2.2. Opinion Mining

More opinions are accumulated as more people make use of social media. These accumulated opinions on social media make it easy to conduct market sensing. Opinion mining from social media data is one of the most vigorous market analyzing activities. Opinion mining research has been applied to review data (Hu and Liu, 2004; Koo et al., 2015; Lu et al., 2011; Wang et al., 2016; Zhang et al., 2006; Zhu et al., 2010). Since customers' reviews are available online, opinion mining is widely studied in multiple domains such as mobile products, tourism, and so on (Koo et al., 2015; Wang et al., 2016; Zhang et al., 2006; Zhu et al., 2010). Zhang utilized sentiment analysis to predict the utility of product reviews. For example, the utility scoring of product reviews (Zhang and Varadarajan, 2006) has been studied based on Amazon product reviews. There are more studies on improving review interpretation through the use of linguistic rules (Lu et al., 2011). They applied theories from linguistics to opinion mining, particularly on product reviews. They built an opinion observer that works based on the linguistic rules they designed.

On the other hand, capturing human emotions from text data has been actively studied. One of the original ways to determine emotions from a given sentence is based on lexicons. For example, a certain set of words is labeled as a sentiment lexicon depending on their semantic polarities, i.e., whether they are positive or negative (Liu, 2010). To improve the accuracy of sentiment analysis, Hu and Liu expanded the sentiment lexicon set by as much as 6,800 words (Hu and Liu 2004; Liu et al., 2005) using WordNet, a famous English word database (Oram, 1998). More research and approaches on sentiment analysis are organized in (Liu, 2012).

However, as Hutto and Gilbert pointed out, it would be more useful if the sentiment intensity could be measured (Hutto and Gibert, 2014). SentiWordNet and VADER (Valence Aware Dictionary for sEntiment Reasoning) libraries make it easy to deal with deeper sentiment analysis of given text (Agarwal et al., 2016; Baccianella et al., 2010; Hutto and Gilbert, 2014). Clayton and Gilbert proposed a rule-based VADER sentiment analyzing toolkit (Hutto and Gibert, 2014). This provides a sentiment degree with-

in the range of [-1.0, 1.0] from a given sentence (Hutto and Gibert, 2014). Since they built a set of rules based on Twitter's dataset, the VADER approach even catches emojis such as ":D" by giving positive values. Thus, the VADER sentiment analysis is highly optimized to measure the social media text data.

2.3. Understanding Natural Language

There have been many studies to improve our understanding of natural language. Fellbaum et al. designed and introduced WordNet, which is an interconnected network structure of semantically relevant words and concepts (Fellbaum et al., 2005). However, Ma et al. noted that the WordNet domain does not appropriately reflect human cognition of words and concepts (Ma et al., 2012).

Another method is word embedding. Mikolove et al. introduced a novel approach for understanding natural language (Mikolov et al., 2013a). In this approach, each word is vectorized as a multiple dimensional matrix using a computational method. Word embedding provides many benefits for performing research with text data (Lebret and Collobert, 2013). Most of all, we can make simple calculations with the vectors (Garten et al., 2015; Mikolov et al., 2013b). For example, we can measure semantic similarities between words by simply computing cosine similarities between two vectors. The gensim library[1] makes it possible to conduct these calculations. Recently, Google provided a word2vector model. It has been pre-trained with the Google news corpus, and it spans 300-dimensional vector space.[2] We decided to utilize Google's pre-trained model for word embedding our online product review data considering it already spans a vector space based on a huge amount of data, i.e., hundred billions of words from Google news.

Ⅲ. Proposed Method

The overall structure of the proposed approach is illustrated in <Figure 1>. Once the reviews of a specific product are secured, we distributed the sentences of the reviews into the six-dimensional framework. To correctly map the sentences on a dimension, we measured the semantic similarities using word embedding.

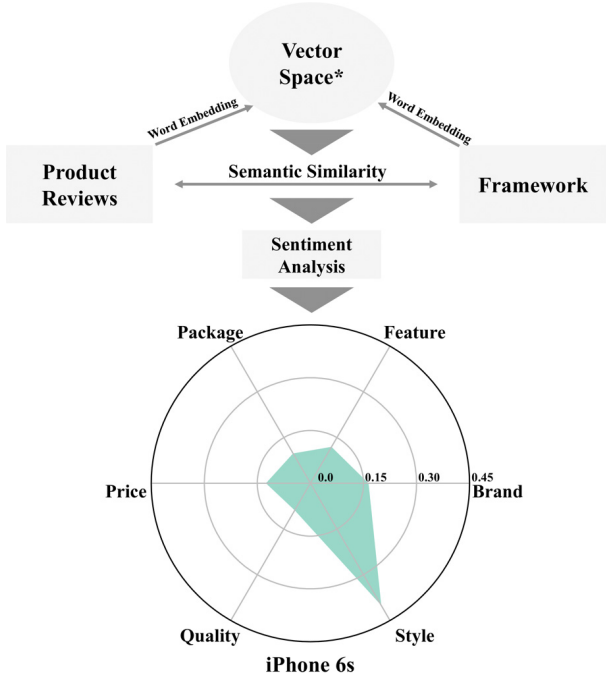
After classifying every sentence, we conducted sentiment analysis to obtain sentiment scores of each sentence. Utilizing the sentiment scores of sentences for each dimension, we obtained a representative score for each dimension. With those six scores, we determined the differences in customers' opinions using six categories. Finally, we plotted a radar graph as shown in <Figure 1>[3].

3.1. Multidimensional Framework Design

We designed a multidimensional framework to implement the top-down approach. We constructed our framework by referring to the most well-known theory for product analysis in marketing management, the Three Levels of a Product (Kotler and Levy, 1969; Kotler et al., 2012). Considering the definitions in the theory, the criteria for each dimension are defined as follows:

Quality: Quality is the ability of a product to perform its functions; it includes the product's overall

1) https://radimrehurek.com/gensim/models/word2vec.html
2) https://code.google.com/archive/p/word2vec/
3) The vector space is spanned by Google News dataset.



<Figure 1> Overview of the Proposed Method

durability, reliability, precision, ease of operation and repair, and other valued attributes.

Feature: Features are a competitive tool for differentiating the company's product from competitors' products.

Style: Style simply describes the appearance of a product. It does not necessarily make the product perform better.

Brand: A brand is a name, term, sign, symbol, or design, or a combination of these, that identifies the maker of a product or service.

Package: The package may include the product's primary container. A secondary package that is thrown away when the product is about to be used, and the shipping package necessary to store, identify, and ship the product.

Price: The sum or amount of money or its equivalent for which anything is bought, sold or offered for sale.

3.2. Word Embedding

To classify entire sentences of reviews into six dimensions, we utilized Google News word2vector pre-trained model[4] for word embedding. Since we do not have a large dataset for building a language representation model, it is better to use an existing model that has already been trained by the Google news dataset. The Google News word2vector was constructed using a 1.6 billion word dataset, and its usefulness was verified in previous research (Mikolov et al., 2013a). This model is frequently used when a sematic relationship among words needs to be represented via vector spaces (Kim, 2014; Lilleberg et al., 2015; Swoboda et al., 2016). This pre-trained model for word embedding is replaceable if a better language representation model could be established

4) https://code.google.com/archive/p/word2vec/

for the future studies.

To compute the semantic similarity between words, we employed Google's word2vector pre-trained model. Each word can be vectorized using the model, meaning that we can measure the semantic similarities by computing cosine similarities between vectors. We removed stop words from the review dataset before word embedding. That is to say, we simply embedded our dataset into a given pre-trained model to compute semantic similarity, which is required for subsequent classification tasks.

3.3. Sentence Classification

In accordance with the criteria of each dimension, we can map each sentence of a product review on a certain dimension. That is, a sentence consists of words. Each word of a given sentence has different semantic distances with the criteria of dimension. In other words, a sentence always has a word showing the highest similarity with a certain dimension among the six dimensions. The sentence is thus categorized into that dimension.

To be more concrete, let a review R be a set with n number of sentences, given as $\{sent_1, sent_2, sent_3, ..., sent_n\}$, where the k-th sentence $sent_k$ is composed of m number of words as $\{w_{k1}, w_{k2}, w_{k3}, ..., w_{km}\}$. For example, suppose there is a review R: *"the battery is terrible! wasted money."* Then, R has two sentences. Here, $sent_1$ is *"the battery is terrible!"* and $sent_2$ is *"wasted money."* Additionally, $sent_1$ is the set of words *{the, battery, is, terrible}*, while $sent_2$ is the set *{wasted, money}*. In short, the relation between w, $sent$, and R can be ordered as in Equation(1).

$$w_{kl} \in sent_k \subset R \quad (1)$$

Meanwhile, a set D has six dimensions $\{d_1, d_2, d_3, ..., d_6\}$ and each dimension $d_i \in D$ has a set with j number of words (terms) t_{ij} as $\{t_{i1}, t_{i2}, t_{i3}, ..., t_{ij}\}$. More concretely, D is the set: *{Quality, Feature, Style, ..., Price}*. *Price* is denoted by d_6, and has associated to it another set of words *{The, sum, or, amount, ..., sale}*. The relations between t_{ij}, d_i, and D can be described as in Equation(2).

$$t_{ij} \in d_i \subset D \quad (2)$$

Then, we can define the function *cat(x),* which returns the corresponding dimension d_i of x. Thus, given that x is the kth sentence $sent_k$ of R, the $cat(sent_k)$ could be described as in Equation (3), where the function $sim(w, d_{ij})$ returns a value within [0.0, 1.0]. Since the return value is calculated by using the cosine distance between two vectors of a word w and a term $t_{ij} \in d_i$, it can be used to determine the semantic similarity between them.

$$\text{cat}(sent_k) = \arg max \{sim(w, t_{ij}) | w \in sent_k\} \quad (3)$$

3.4. Sentiment Analysis

Once each sentence classification process has been completed, we finally conducted sentiment analysis for each sentence. We utilized the VADER (Valence Aware Dictionary for sEntiment Reasoning) sentiment analysis library, since it is optimized for text data on the internet (Hutto and Gibert, 2014). Among the things the VADER method provides, we used the compound return value, which is in the range of [-1.0, 1.0].

In the case of the example *"The camera works perfect!"*, the verb *work* directs its sentence to the *Quality* dimension. The positive adjective *perfect* brings a considerably high value of 0.6114 points

<Table 1> The Number of Data Set Crawled on January 9th, 2018

Product	Reviews	Sentences	Words
iPhone 6s	2,775	6,175	73,556
iPhone SE	1,259	3,048	39,503
iPhone 7	598	1,496	20,599
iPhone Total	4,632	10,719	133,658
Galaxy S6	1,954	5,655	80,466
Galaxy S6 edge	1,137	2,526	34,933
Galaxy S7	984	2,850	40,839
Galaxy Total	4,075	11,031	156,238

from VADER sentimental analysis. In short, the product has been evaluated with 0.6114 points in its *Quality* dimension.

$$\text{score}(sent_k) = VADER(sent_k) \quad (4)$$

To get the sentiment degree of a sentence, we apply VADER sentiment analysis library (Hutto and Gilbert, 2014). As described in Equation 4, the *score( )* returns a sentiment degree of its argument, which is a sentence($sent_k$) in this case.

Once we get all the sentiment scores of all sentences, we need to get the aggregated score for each dimension. The below Equation 5 shows how we compute the aggregated score for each dimension; d_i.

$$\text{score}(d_i) = Average_{\{sent_k | cat(sent_k)=d_i\}} \{score(sent_k)\} \quad (5)$$

As Equation 5 indicates, the *score(d_i)* is dependent on the *score($sent_k$)* in the appropriate dimension. Eventually, reviews of a product would have six values.

3.5. Data Visualization

Finally, we summarized the results in box and radar plots to present the information more effectively. There are a number of classified sentences for dimensions. Every sentence holds different sentiment degrees. Thus, the distribution of sentences on each dimension also varies.

A product would have six representative values for six dimensions. We used them to plot the radar graphs. The radar graphs present the multidimensional consumers' opinions on products all at once. Meanwhile, the box plot shows the distribution of sentiment degrees for each dimension. The plots presents both max and min values and the skewness of the data distribution.

Ⅳ. Experimental Setting

4.1. Data Collection

We first constructed a dataset by collecting product reviews for iPhone (6s, SE, and 7) and Galaxy (S6, S6 edge, and S7) from Amazon[5], which is one of the most famous online shopping platforms in the world. The reviews were crawled using Python libraries (BeautifulSoup[6] and selenium[7]) on January

5) https://www.amazon.com/
6) https://pypi.python.org/pypi/bs4
7) http://www.seleniumhq.org/projects/webdriver/

9, 2018. Since there could be some unreliable reviews, we filtered them out to secure the data reliability by selecting *Verified purchase only* after we conducted the data crawl.

<Table 1> shows the number of entire datasets. It describes both the number of reviews and the number of sentences and words for each product. As shown, there are obviously fewer reviews for recent products. This is because the physical period of recent products is relatively shorter than those of old products.

4.2. Baseline Setting

4.2.1. Participants

We recruited 6 volunteers including three females from a local university (Average age: 22.17 with SD=1.72). Before conducting the evaluation task, we held a training session to let participants fully understand our framework because it is critical to classify sentences. Further, two participants were iPhone users, while the rest were Galaxy users.

4.2.2. Task Design

After the training session, we introduced the overall procedure of the experiment to participants. First, using a questionnaire form, each participant was asked to classify review sentences into six dimensions without the intervention of other participants. The questionnaire included 400 sentences each for iPhone and Galaxy. In fact, the questionnaire included stratified sampled sentences based on the sentence distributions of six products (three iPhones and three Galaxies). Participants were asked to continually check the definition of dimensions so that they could clarify the differences between dimensions. After filling out the questionnaire, the participants had a discussion session. In the session, they shared each other's answers and explained why they assigned a sentence to a specific dimension. A consensus on the classification results was achieved using this process.

4.2.3. Performance measures

We evaluate the performance with the *F1-score*. It is a measure that considers both *recall* and *precision*. *Recall* is a detected the number of truths among the number of existing truths (i.e., it is a rating of how many valid answers are detected). On the other hand, the *Precision* reflects the number of truths among the number of selected truths (i.e., how many correct answers are selected by a classifier).

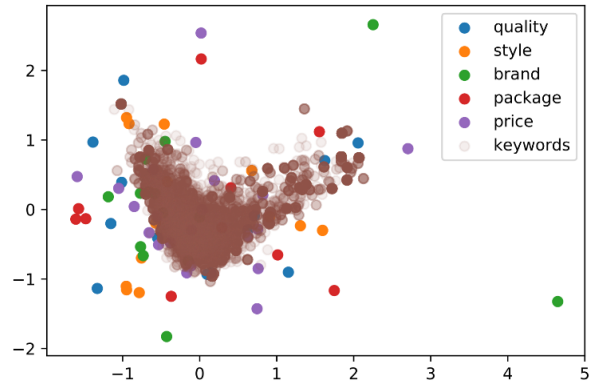
To get the overall results, we have summarized iPhone and Galaxy results. Also, we counted detected truths, existing truths and selected truths. The equations used to assess recall and precision are as follows:

$$Recall = DetectedTruths(a)/ExistingTruths(b), \quad (6)$$

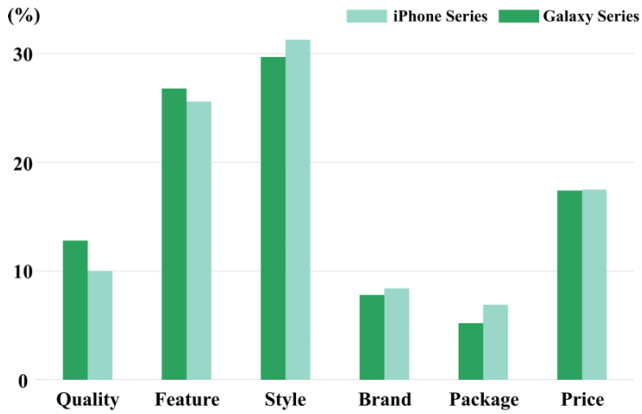
$$Precision = DetectedTruths(a)/SelectedTruths(c), \quad (7)$$

where the a is the number of *DetectedTruths* for both iPhone and Galaxy, b is the number of the *ExistingTruths* of both iPhone and Galaxy, and c is the number of the *SelectedTruths*. Using precision and recall values, we can acquire an *F1-score* using the following formula.

$$F1\text{-}score = 2 * Recall * Precision/(Recall + Precision) \quad (8)$$



<Figure 2> Scatter Plot of Embedded Results in Vector Space



<Figure 3> The Ratio of Classified Sentences for Each Dimension by Product: iPhone and Galaxy

Ⅴ. Results

5.1. Words Distribution

We embedded words from product reviews and criteria of dimensions to Google's pre-trained vector space. Each word was vectorized, and the results were visualized as a scatter plot in <Figure 2>. It shows the distribution of embedded words, which are from iPhone SE reviews. Since the vector coordinates involve 300 dimensions, we conducted dimensionality reduction through the use of principal component analysis to plot the high dimensional vectors in two dimensions (Jolliffe, 2002; Nie et al., 2014).

5.2. Sentence Classification

We classified every sentence into the best corresponding dimension based on the semantic distances between words of product reviews and framework criteria. <Table 2> shows the number of sentences mapped on each dimension. The ratio of those is also identified in <Table 2>. We categorized a total of about 21,000 sentences. We screened out sentences with neutral sentiments (i.e., in which the sentiment

<Table 2> The results of Sentence Classification for Each Product

Product	Quality	Feature	Style	Brand	Package	Price	Total Sentences
iPhone 6s	411 (9.6%)	1,088 (25.4%)	1,401 (32.8%)	347 (8.1%)	289 (6.7%)	732 (17.1%)	4,268 (100%)
iPhone SE	220 (10.8%)	525 (25.8%)	621 (30.5%)	160 (7.8%)	138 (6.7%)	367 (18.0%)	2,031 (100%)
iPhone 7	103 (10.1%)	260 (25.7%)	272 (26.9%)	108 (10.6%)	81 (8.0%)	187 (18.4%)	1,011 (100%)
iPhone Total	734 (10.0%)	1,873 (25.6%)	2,294 (31.3%)	615 (8.4%)	508 (6.9%)	1,286 (17.5%)	7,310 (100%)
Galaxy S6	516 (13.5%)	975 (25.6%)	1,126 (29.6%)	298 (7.8%)	194 (5.1%)	694 (18.2%)	3,803 (100%)
Galaxy S6 edge	208 (11.9%)	494 (28.4%)	540 (31.0%)	140 (8.0%)	74 (4.2%)	282 (16.2%)	1,738 (100%)
Galaxy S7	236 (12.2%)	537 (27.9%)	551 (28.7%)	145 (7.5%)	121 (6.3%)	329 (17.1%)	1,919 (100%)
Galaxy Total	960 (12.8%)	2,006 (26.8%)	2,217 (29.7%)	583 (7.8%)	389 (5.2%)	1,305 (17.4%)	7,460 (100%)

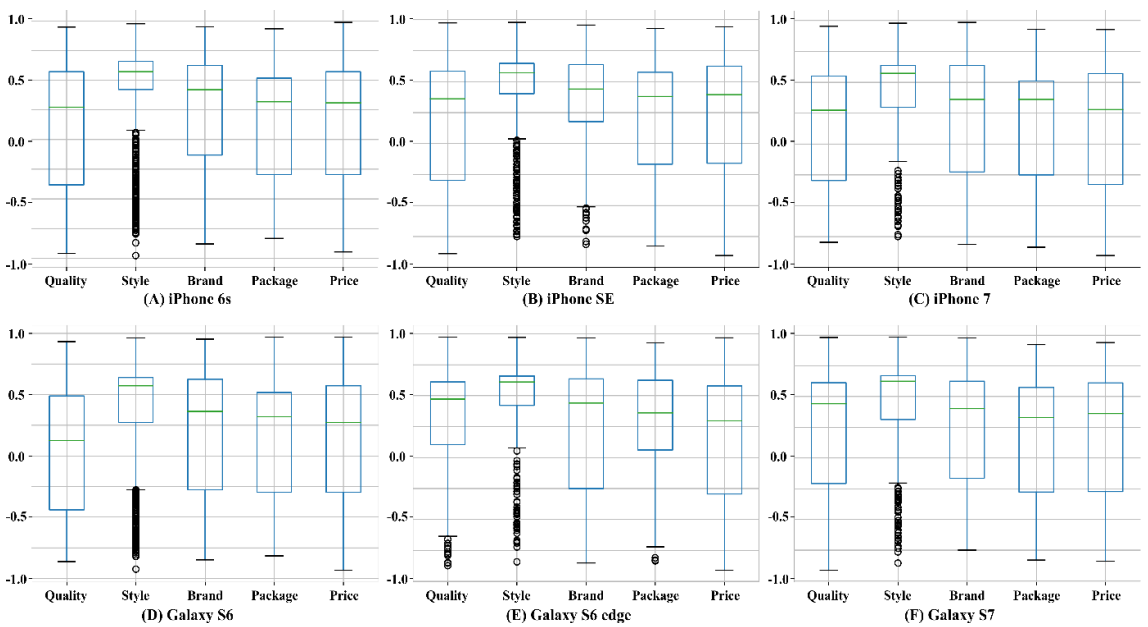
score was 0.0, 68% overall or 14.7 thousand out of 21.7 thousand).

Interestingly, both iPhone and Galaxy have a quite similar ratio of dimensions as displayed in <Figure 3>. *Feature* and *Style* are high for both iPhone and Galaxy. This implies that consumers would like to talk about feature and style rather than other topics. Obviously, the *Price* holds the next highest ratio in both iPhone and Galaxy product reviews.

5.3. Sentiment Data Distribution

As shown in <Table 2>, each dimension has an associated- number of mapped sentences. Since we already screened out neutral sentiment sentences, every sentence in each dimension holds various sentiment degrees between [-1.0, 1.0]. The median also varies. <Figure 4> displays the distribution of sentiment scores for each product review.

In particular, the interquartile range, which is the



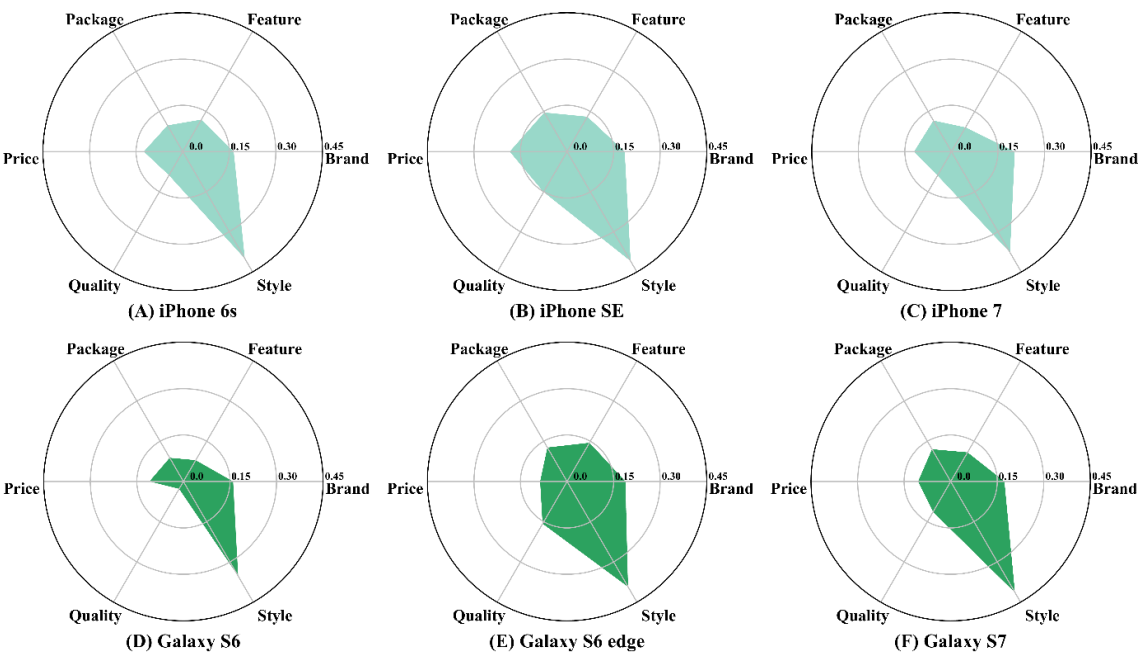
<Figure 4> Box Plots of Sentiment Scores by Product

length between Q1 and Q3, is comparably narrow in the *Style* dimension for every product. The median of *Style* is the highest, as shown in This implies that most of the sentences in *Style* scored higher than those of sentences from any other dimensions. In other words, consumers provided highly positive opinions in terms of *Style* aspects on products. We also found that all the median values were higher than 1.1. This means that the distribution is left-skewed. There are more positive sentences than negative sentences. That is, the range of negative sentiment scores is broader than that of positive sentiment scores. Although the median values could theoretically be within the range [-1.0, 1.0], there is a practical reason why the median is not significantly lower than 0.0. This is because companies invest their resources based on logical thinking. Therefore, they would not launch a product about which consumers are likely to be negative.

Another thing which stands out from the box plots is that the *Brand* category of iPhone SE is narrower than that of other iPhone series. So is the *Quality* category of Galaxy S6 edge, which became more left-skewed with a long tail comparing to that of other Galaxy series. In other words, the bottom lines of each box (which is Q1) are located higher on the graph in the case of iPhone SE and Galaxy S6 edge. This indicates that the iPhone SE successfully achieved consumers' satisfaction in terms of its *Brand*. Likewise, the Galaxy S6 edge made consumers satisfied with its *Quality*.

5.4. Radar Plots

As shown in <Figure 5>, consumers' positive opinions are higher for iPhone SE in the iPhone series, as shown by the growth in the size of the (B) for each new generation of the iPhone. In the case of



<Figure 5> The Radar Plots for Each Product

the Galaxy series, plot (E) takes up a larger space than that of the other Galaxy series. In both cases, the *Quality* dimension has remarkably expanded compared to that of other series. We also note consumers' positive sentiment on its *Price*, while the that of iPhone 6s and 7 are comparably small. This actually reflects the reality that Apple launched the iPhone SE as a distributional product, which is a medium and low price type. However, they show the same or less positive opinions on the *Price* of Galaxy S6 edge than that of Galaxy S6 and S7.

The *Brand* dimension continuously expands in the iPhone series. This is quite interesting because no matter how the *Quality* or other dimensions are, the name 'iPhone' in the *Brand* category barely shrinks in the radar plots regardless of iPhone generation. Thus, iPhone users may show higher loyalty towards their products than Galaxy users.

The Galaxy S6 has been criticized a lot especially on its *Quality* dimension, as is shown in <Figure 5> (D), where it almost reaches 0.0. Indeed, the Galaxy S6 officially has some issues related to battery durability and multitasking performance. This explicitly appears in sentences such as "This phones battery is absolutely terrible.", "Something is wrong with the battery.", "It would overheat, freeze and shut off.", and "It constantly freezes up and shuts off for no reason."

5.5. Evaluation Analysis

We evaluated the sentence classification accuracy. We randomly selected 400 sentences from each iPhone and Galaxy series, for a total of 800 sample sentences. Six trained participants manually classified 800 sentences. During the classification, those participants actively discussed the results to correctly categorize each given sentence. To examine the accuracy,

<Table 3> The F1 Classification Accuracy of iPhone Sentences

	Recall	Precision	F1 score
Quality	0.13	0.73	0.22
Feature	0.30	0.03	0.05
Style	0.32	0.05	0.09
Brand	0.11	0.18	0.14
Package	0.16	0.39	0.23
Price	0.76	0.31	0.44

<Table 4> The F1 Classification Accuracy of Galaxy Sentences

	Recall	Precision	F1 score
Quality	0.14	0.69	0.23
Feature	0.53	0.17	0.26
Style	0.47	0.13	0.20
Brand	0.09	0.06	0.07
Package	0.19	0.30	0.29
Price	0.70	0.19	0.30

we calculated the F1 score along with the recall and precision scores for each category. Moreover, categories that showed a big difference between Recall and Precision provided important implications about classification.

<Table 3> describes the evaluation results of the iPhone series, while <Table 4> displays the evaluation results of Galaxy series. We can see that the F1 scores of both series show few differences. This signifies that the sentence classification is not dependent on any specific product. Except for *Feature*, the disparities between F1 scores are around 0.1. This confirms that the classification can be implemented for different products to produce similar results.

<Table 5> shows the overall results for both iPhone and Galaxy. The precision of the *Price* section is comparably lower than its recall, because there are some sentences such as "The phone would not read or accept a sim card." and "Do not waste your money on it." In the first sentence, the semantic distance of the word "card" itself is obviously close to the *Price* criteria, even though "sim card" does not mean the same thing as "card." Further, the second sentence itself implies that he or she does not like the product. However, the keyword "money" is semantically close to the *Price* category. These kinds of keywords lead to incorrect categorizing of dimensions, causing a low precision score.

On the other hand, the precision is much greater than recall in the case of the *Quality* dimension on <Table 5>. Participants classified most of the sentences in *Quality* dimension. There are some sentences such as: "very good, my son love it very much." and "So far, so good." They simply express consumers' overall satisfaction on products. However, during the debates, participants agreed to classify such sentences in the *Quality* dimension. That is why recall is lower (due to a large number of elements that are selected as *Quality* by participants). Consequently, sentences classified as *Quality* were accurately detected, leading to a high precision score.

Ⅵ. Discussion

Considering the core of our approach is designed referring to a generic marketing management framework, it is universally applicable to other product reviews. That is, the six-dimensions considerably affect consumers' decisions (Kotler and Levy, 1969; Kotler et al., 2012). Additionally, this is highly constructive because it can draw more comprehensive information from social media data, and because it provides a source that allows companies to refer to for more insights if needed.

The proposed approach is particularly useful to

<Table 5> The F1 Classification Accuracy of Total Sentences

	Recall	Precision	F1 score
Quality	0.13	0.70	0.22
Feature	0.48	0.10	0.17
Style	0.41	0.09	0.15
Brand	0.10	0.12	0.11
Package	0.19	0.35	0.25
Price	0.73	0.25	0.37

conduct market sensing, especially when a company tries to launch new products. After launching a product, it is important for a company to verify that the product is meeting their original intended goals in the market. Hence, it is important for them to be able to gather accurate useful information related to the product and how the market is reacting to it. This multidimensional approach of opinion mining from social media can help these companies collect the required information. Additionally, this suggested approach makes it possible for companies to access substantial information that cannot be found solely from numbers on the financial statements.

Companies can even generate a report on their competitors' products since the opinions are public on social media. This allows a comparison between similar products from different brands and provides insight to what strategy competitors could be pursuing. However, only the company itself could judge the results. Referring to the example above, as new iPhone models are introduced into the series, consumers' inclination to Style and Feature dimensions drastically expand according to our empirical results. This response could have been what Apple intended. Conversely, in the case of the Galaxy series, we cannot easily discern any particular pattern from the radar graph. This could indicate that Samsung might be testing diverse variables to determine how to maximize their profit in the smartphone market. In short, the presented approach in this paper provides companies with a way to obtain insightful information to proceed with their next business decisions.

However, there are still constraints associated with this approach. Most of all, the multidimensional framework in the proposed approach is highly dependent on the existing theory itself. Considering that some theories sometimes might be totally or partially wrong, the suggested approach always takes the risk of opinion mining with erroneous conceptual dimensions. Additionally, we could not filter out biased reviews. Although we crawled verified Amazon reviews, they cannot be 100% free from accuracy problems as long as they are still from social media. Lastly, sentence classification has been done in a disjointed manner. However, there are sentences which would not be disjointly distributed such as "this wonderful pricing and quality." Though it holds keywords from both *Quality* and *Price*, it has been classified on the *Quality* dimension, because the word 'quality' satisfies an exact 100% matching with the criteria, while the word 'pricing' matches slightly lower than 100%.

Thus, further research could include the following: the reliability of the reviews, joint classification for each sentence, and a study of framework itself. We could expect solid results with more reliable data. Among billions of data on social media, there could be thousands of garbage datasets. Before conducting the multidimensional analysis, it is highly desirable to screen out the garbage. Also, since the suggested method cannot clearly classify when a given sentence contains more than two topics, it would be much better to apply multidimensional analysis that is able to split each topic in natural language processing. Lastly, although there are vast theories and frameworks which address this world, the ever changing nature of human behavior and discovery of new information might render these frameworks outdated or obsolete.

Ⅶ. Conclusion

In this paper, we have proposed a multidimensional approach of opinion mining from online prod-

uct reviews. By projecting product reviews into a six-dimensional framework, our novel approach provides richer information than the earlier methods do. To classify the sentences of reviews, we utilized the word2vector pre-trained model. We conducted sentiment analysis after distributing all sentences into dimensions. Based on the sentiment scores of the sentences, we visualized the reviews in a radar graph with six axes. The whole process has been illustrated with Amazon product reviews of iPhone and Galaxy series. The feasibility of the proposed approach was verified through experiments. The proposed approach has advantages to compare previous multidimensional analysis approaches of product opinion mining because it is based on universal dimensions of products rather than product-specific dimensions.

Acknowledgements

This work was supported by an Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korean government (MSIT) [2016-0-00562(R0124-16-0002), Emotional Intelligence Technology to Infer Human Emotion and Carry on Dialogue Accordingly].

<References>

[1] Adams, J. L. (2012). *Good products, bad products: Essential elements to achieving superior quality*. McGraw-Hill.

[2] Agarwal, A., Sharma, V., Sikka, G., and Dhir, R. (2016). Opinion mining of news headlines using SentiWordNet. In *Symposium on Colossal Data Analysis and Networking*, IEEE, 1-5.

[3] Baccianella, S., Esuli, A., and Sebastiani, F. (2010). Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the Seventh conference on International Language Resources and Evaluation* (LREC'10), 2200-2204.

[4] Day, G. S. (1994). The capabilities of market-driven organizations. *Journal of marketing*, *58*(4), 37-52.

[5] Duan, W., Cao, Q., Yu, Y., and Levy, S. (2013). Mining online user-generated content: using sentiment analysis technique to study hotel service quality. In *46th Hawaii International Conference on System Sciences*, IEEE, 3119-3128.

[6] Fellbaum, C. (2005). WordNet and wordnets. In: Brown, Keith et al. (eds.), *Encyclopedia of Language and Linguistics*, Second Edition, Oxford: Elsevier, 665-670.

[7] Garten, J., Sagae, K., Ustun, V., & Dehghani, M. (2015). Combining distributed vector representations for words. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, 95-101.

[8] Giachanou, A., and Crestani, F. (2016). Opinion retrieval in Twitter: Is proximity effective? In *Proceedings of the 31st Annual ACM Symposium on Applied Computing*, ACM, 1146-1151.

[9] Hu, M., and Liu, B. (2004). Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM, 168-177.

[10] Hu, N., Pavlou, P. A., and Zhang, J. (2006). Can online reviews reveal a product's true quality?: Empirical findings and analytical modeling of online word-of-mouth communication. In *Proceedings of the 7th ACM conference on Electronic commerce*, ACM, 324-330.

[11] Hutto, C. J., and Gilbert, E. (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth international AAAI conference on weblogs and social media*, 216-225.

[12] Iosifidis, V., and Ntoutsi, E. (2017). Large scale sentiment learning with limited labels. In *Proceedings of the 23rd ACM SIGKDD international conference*

on knowledge discovery and data mining, ACM, 1823-1832.

[13] Jolliffe, I. T. (2002). Principal components in regression analysis. *Principal Component analysis*, 167-198.

[14] Kim, J. B., and Shin, S. I. (2015). An empirical study on the interaction effects between the customer reviews and the customer incentives towards the product sales at the online retail store. *Asia Pacific Journal of Information Systems, 25*(4), 763-783.

[15] Kim, Y., Moon, H. S., and Kim, J. K. (2017) Analyzing the effect of electronic word of mouth on low involvement products. *Asia Pacific Journal of Information Systems, 27*(3), 139-155.

[16] Kim, Y. (2014). Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*.

[17] Koo, C., Shin, S., Hlee, S., Moon, D., and Chung, N. (2015) Online tourism review: Three phases for successful destination relationships. *Asia Pacific Journal of Information Systems, 25*(4), 746-762.

[18] Kotler, P., and Levy, S. J. (1969). Broadening the concept of marketing. *Journal of marketing, 33*(1), 10-15.

[19] Kotler, P., Keller, K. L., Ancarani, F., and Costabile, M. (2012). *Marketing management 12/e*. Pearson.

[20] Kordupleski, R. (2003). *Mastering customer value management: The art and science of creating competitive advantage*. Customer Value Management I.

[21] Lebret, R., and Collobert, R. (2013). Word emdeddings through hellinger PCA. *arXiv preprint arXiv:1312.5542*.

[22] Lilleberg, J., Zhu, Y., and Zhang, Y. (2015). Support vector machines and word2vec for text classification with semantic features. In *2015 IEEE 14th International Conference on Cognitive Informatics & Cognitive Computing*, 136-140.

[23] Liu, B. (2010). Sentiment analysis and subjectivity. *Handbook of natural language processing*, 627-666.

[24] Liu, B. (2012). *Sentiment analysis and opinion mining*. Morgan & Claypool Publishers.

[25] Liu, B., Hu, M., and Cheng, J. (2005). Opinion observer: Analyzing and comparing opinions on the web. In *Proceedings of the 14th international conference on World Wide Web*, ACM, 342-351.

[26] Lu, B., Ott, M., Cardie, C., and Tsou, B. K. (2011). Multi-aspect sentiment analysis with topic models. In *2011 IEEE 11th international conference on data mining workshops*, IEEE, 81-88.

[27] Ma, X., Fellbaum, C. (2012). Rethinking WordNet's domains. In *Proceedings of Global WordNet Conference*, 173-180.

[28] McCarthy, E. J., Shapiro, S. J., and Perreault, W. D. (1979). *Basic marketing*. Irwin-Dorsey.

[29] Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

[30] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013b). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, 3111-3119.

[31] Nie, F., Yuan, J., and Huang, H. (2014). Optimal mean robust principal component analysis. In *International conference on machine learning*, ICML'14, 1062-1070.

[32] Oram, P. (1998). *Wordnet: An electronic lexical database*. Christiane fellbaum.

[33] Paul, D., Li, F., Teja, M. K., Yu, X., and Frost, R. (2017). Compass: Spatio temporal sentiment analysis of us election what twitter says!. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, ACM, 1585-1594.

[34] Piercy, N. F. (2016). *Market-led strategic change: Transforming the process of going to market*. Routledge.

[35] Swoboda, T., Hemmje, M., Dascalu, M., and Trausan-Matu, S. (2016). Combining taxonomies using word2vec. In *Proceedings of the 2016 ACM Symposium on Document Engineering*, 131-134.

[36] Wang, Y., Aguirre-Urreta, M., and Song, J. (2016). Investigating the value of information in mobile commerce: A text mining approach. *Asia Pacific Journal of Information System, 26*(4), 577-592.

[37] Zhai, Z., Liu, B., Xu, H., and Jia, P. (2011). Clustering

product features for opinion mining. In *Proceedings of the fourth ACM international conference on Web search and data mining*, ACM, 347-354.

[38] Zhang, Z., and Varadarajan, B. (2006, November). Utility scoring of product reviews. In *Proceedings of the 15th ACM international conference on Information and knowledge management* (pp. 51-57). ACM.

[39] Zhu, F., and Zhang, X. (2010). Impact of online consumer reviews on sales: The moderating role of product and consumer characteristics. *Journal of Marketing, 74*(2), 133-148.

◆ About the Authors ◆

Taewook Kim

Master of Philosophy student in the department of Computer Science and Engineering at the Hong Kong University of Science and Technology (HKUST), Hong Kong SAR. He received dual degree; B.S. in Business Administration and B.S. in Computer Science and Engineering from Hanyang University, Seoul, Korea. His current research interests include Human-Computer Interaction, Social Computing, and Affective Computing.

Dong Sung Kim

Postdoctoral researcher of the Department of Business Administration at the School of Business, Hanyang University, Seoul, Korea. He received B.S. degree from the Department of Management Information System at Hyupsung University. He received his M.S. and Ph.D. degrees, respectively, from the Department of Business Administration at Hanyang University, Seoul, Korea. His current research interests include the application of machine learning techniques, opinion mining, and social network analysis.

Donghyun Kim

Engineer of LG Electronics Company, Pyeongtaek, Korea. He has a bachelor's degree in Information Systems from Hanyang University. His research interests include Web application, Data mining and Machine learning.

Jong Woo Kim

Jong Woo Kim is a professor at the School of Business, Hanyang University, Seoul, Korea. He received B.S. degree from the Department of Mathematics at Seoul National University, Seoul, Korea. He received his M.S. and Ph.D. degrees, respectively, from the Department of Management Science, and the Department of Industrial Management at Korea Institute of Science and Technology (KAIST), Korea. His current research interests include intelligent information systems, data mining and machine learning applications, text mining application, social network analysis, collaborative systems, and e-commerce recommendation systems. His papers have been published in *Expert Systems with Applications, Cyberpsychology Behavior and Social Networking, Computers in Human Behavior, Information Systems Frontiers, International Journal of Electronic Commerce, Electronic Commerce Research, Mathematical and Computer Modeling, Journal of Intelligent Information Systems,* and other journals.

Submitted: April 30, 2019; 1st Revision: September 17, 2019; Accepted: September 25, 2019