JKSCI

한국컴퓨터정보학회논문지
Journal of The Korea Society of Computer and Information
Vol. 25 No. 8, pp. 181-188, August 2020
https://doi.org/10.9708/jksci.2020.25.08.181

Impact of Word Embedding Methods on Performance of Sentiment Analysis with Machine Learning Techniques

Hoyeon Park*, Kyoung-jae Kim**

*Ph.D. Candidate, Dept. of MIS, Graduate School, Dongguk University, Seoul, Korea
**Professor, Dept. of MIS, Business School, Dongguk University, Seoul, Korea

[Abstract]

In this study, we propose a comparative study to confirm the impact of various word embedding techniques on the performance of sentiment analysis. Sentiment analysis is one of opinion mining techniques to identify and extract subjective information from text using natural language processing and can be used to classify the sentiment of product reviews or comments. Since sentiment can be classified as either positive or negative, it can be considered one of the general classification problems. For sentiment analysis, the text must be converted into a language that can be recognized by a computer. Therefore, text such as a word or document is transformed into a vector in natural language processing called word embedding. Various techniques, such as Bag of Words, TF-IDF, and Word2Vec are used as word embedding techniques. Until now, there have not been many studies on word embedding techniques suitable for emotional analysis. In this study, among various word embedding techniques, Bag of Words, TF-IDF, and Word2Vec are used to compare and analyze the performance of movie review sentiment analysis. The research data set for this study is the IMDB data set, which is widely used in text mining. As a result, it was found that the performance of TF-IDF and Bag of Words was superior to that of Word2Vec and TF-IDF performed better than Bag of Words, but the difference was not very significant.

▸ **Key words**: sentiment analysis, Bag of words, TF-IDF, Word2Vec, machine learning

[요 약]

본 연구에서는 다양한 워드 임베딩 기법이 감성분석의 성과에 미치는 영향을 확인하기 위한 비교연구를 제안한다. 감성분석은 자연어 처리를 사용하여 텍스트 문서에서 주관적인 정보를 식별하고 추출하는 오피니언 마이닝 기법 중 하나이며, 상품평이나 댓글의 감성을 분류하는데 사용될 수 있다. 감성은 긍정적이거나 부정적인 것으로 분류될 수 있기 때문에 일반적인 분류문제 중 하나로 생각할 수 있으며, 이의 분류를 위해서는 텍스트를 컴퓨터가 인식할 수 있는 언어로 변환하여야 한다. 따라서 단어나 문서와 같은 텍스트를 자연어 처리에서 벡터로 변형하여 진행하는데 이를 워드 임베딩이라고 한다. 워드 임베딩 기법은 Bag of Words, TF-IDF, Word2Vec 등 다양한 기법이 사용되고 있는데 지금까지 감성분석에 적합한 워드 임베딩 기법에 대한 연구는 많이 진행되지 않았다. 본 연구에서는 영화 리뷰의 감성분석을 위해 다양한 워드 임베딩 기법 중 Bag of Words, TF-IDF, Word2Vec을 사용하여 그 성과를 비교 분석한다. 분석에 사용할 연구용 데이터 셋은 텍스트 마이닝에서 많이 활용되고 있는 IMDB 데이터 셋을 사용하였다. 분석 결과, TF-IDF와 Bag of Words의 성과가 Word2Vec보다 우수한 것으로 나타났으며 TF-IDF는 Bag of Words보다 성과가 우수하였으나 그 차이가 매우 크지는 않았다.

▸ **주제어**: 감성분석, Bag of words, TF-IDF, Word2Vec, 기계 학습

• First Author: Hoyeon Park, Corresponding Author: Kyoung-jae Kim
*Hoyeon Park (hoyeonpark@dongguk.edu), Dept. of MIS, Graduate School, Dongguk University_Seoul
**Kyoung-jae Kim (kjkim@dongguk.edu), Dept. of MIS, Business School, Dongguk University_Seoul
• Received: 2020. 07. 28, Revised: 2020. 08. 11, Accepted: 2020. 08. 11.

http://www.ksci.re.kr pISSN:1598-849X | eISSN:2383-9945

I. Introduction

Data mining is used in various fields, among them, natural language processing (NLP) developed late, unlike image classification and speech recognition. For the NLP, words are used for the analysis, and words in a sentence are sequential data different from image or audio data. Also, NLP uses a different approach. Sentences can be defined as the probability with the discrete data, and the relationship is affected by the form of words around. In this case, it is defined as a discrete random variable that provides a probability value for the sample. Sample probability one into the terms of the Bag of Word. However, Word2Vec able to express them using the order information of the sentences. Due to the use of time or sequential information, Word2Vec is considered a method of sequential modeling.

In this study, the known word embedding techniques such as TF-IDF, Bag of Words and the relatively recently used Word2Vec are used to perform sentiment analysis and compare the results. Many studies have been undertaken on word embedding techniques, but few papers have studied the impact of each method in sentiment analysis. Therefore, in this study, we intend to apply Bag of Words, TF-IDF, and Word2Vec to product review sentiment analysis and explore the results. Within the sentiment analysis, classification of sentiment is performed using representative machine learning techniques such as Naive Bayes, Support Vector Machines, Random Forests, XGBoost, and XGBoost.

Besides, the movie dataset in this study uses IMDB, because its movie reviews are used by people worldwide. Therefore, we will proceed with the sentiment analysis using word embedding methods to recognize the positive and negative review reactions. Each word embedding method has advantages and disadvantages, so exploratory research is needed to find a technique suitable for sentiment analysis.

The structure of this study is organized as follows. Related work on sentiment analysis and the word embedding method is discussed in Section II. In Section III, we presented the evaluation of experimental results. Finally, we conclude our work in Section IV.

II. Prior Research

1. Sentiment Analysis

Sentiment analysis is a technique that can generally identify sentiment within the text, and there are (1) sentence concepts (2) document concepts (3) aspect-based concepts[1][2][3]. In the case of sentence concepts, sentiment analysis polarizes text depend on the NLP, which is the process of translating words. When representing table mapping in NLP, the sentences were converted one emotions into positive or negative categories[4]. The idea of document classification proceeds using the similarity to the sentence of a given document[5]. Aspect-based concepts are extract and process according to object feature or property extraction[6]. The recent sentiment analysis is based on the results of the categorization of texts. It is essential to define and extract features. In this process, the classification algorithm was mainly used by NB(Naive Bayes), SVM(support vector machine), ANN(artificial neural network), and kNN in many studies. Recently, deep learning has been applied to sentiment analysis and proving its high accuracy.

2. Word Embedding

Word embedding is a method of converting words into vectors that can be computed. Bag of Words (BoW) does not proceed with the grammar and word order of the text document, but the lexical list produces a vector based on the entire corpus. Each lexical list is represented by a numeric vector and can be efficiently modeled for data scalability, classification[7], and text processing. Meanwhile,

Term Frequency-Inverse Document Frequency (TF-IDF) is a method that uses the relative term frequency for each document and consists of two separate calculations. TF represents the frequency of words in a document, and IDF defines another document-specific frequency within the same word. Words with high TF-IDF weights are more important than words with low TF-IDF weights for the analysis. Word embedding has been upgraded in text mining research with the advent of Word2Vec. Word2Vec is the algorithm proposed firstly by the Google team, led by Mikolov[8]. Word2Vec understands words that are frequently located around words through a distributed hypothesis of words. Word2Vec's learning method is based on the assumption that words assume coexistence information as context. If the document is sufficient, Word2Vec learning can also be used to learn semantic relationships for abstract words.

Word2Vec is based on the following assumptions. First, the meanings of words can be expressed as distribution around words. Second, the meanings of words are encoded in word vectors[9,10]. Third, Word2Vec is known to be able to represent each word as a low dimensional vector so that new sentences can be integrated quickly and easily, and words can be added to the vocabulary list[11]. Forth, the weight of Word2Vec depends on the sequence or word position, not the frequency in the same context. After the weight of Word2Vec is calculated, the similarity between the two words can be estimated[12,13]. The method of Word2Vec is divided into CBOW (continuous bag of words) and Skip-gram. The two methods are shown in the following Fig.1.

CBOW constructs a model by predicting surrounding words as a central word, and Skip-gram builds a model to estimate surrounding words as a central word. The commonality between CBOW and Skip-gram is based on the assumption that similar words appear more similar vector values. Because Word2Vec's window contains word location information, it learns word embedding using the surrounding words based on a specific word.

3. Sentiment Analysis based on Machine Learning Techniques

In sentiment analysis, text data is analyzed by dividing the text into positives and negatives using predictive classification. The positives and negatives must be extracted from the review to the classifier for training the classifier on the lexical list. Trained classifiers are used to determine the classification of test data with positives and negatives[14,15].

Many prior studies used machine learning techniques including NB, SVM, Maximum Entropy (ME), k-NN, ANNs for sentiment analysis. Table 1 shows the studies comparing the results of sentiment analysis based on machine learning techniques. The accuracy of Table 1 is the highest performed classifier among the compared classifiers. The results show that SVM and NB outperform the other classifiers.

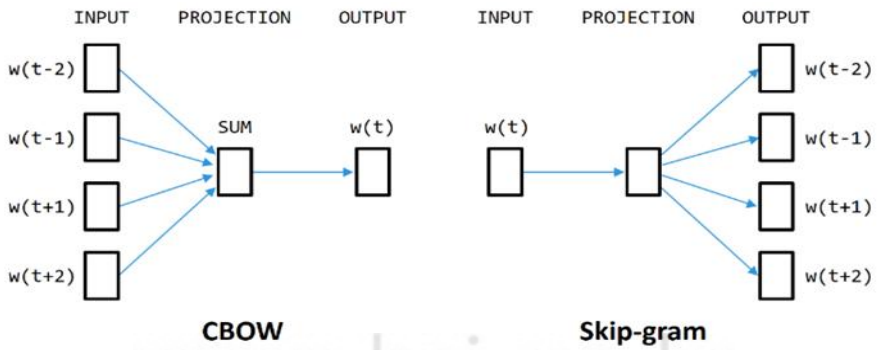

Fig. 1. CBOW and Skip-gram

Table 1. Brief comparison of sentiment analysis based on machine learning techniques (Dataset : IMDB)

Reference #	Techniques	Accuracy
[4]	NB, ME, SVM	SVM (82.9%)
[16]	NB, SVM	SVM (82.9%)
[17]	NB, k-NN	NB (82.84%)

III. Experiments and Results

This section presents the experimental design and results. The research dataset is IMDB (imdb.com) movie review data provided by Keras, which was conducted on 30,000 datasets already classified as positive and negative. The specification of the experimental environment of this study is tested on Intel Core i5-8250 CPU 3.4GHz, 16GB DDR4 2400MHz.

The experiment in this study was performed using Python, and the Python library was executed using Scikit-learn, NumPy, and SciPy math libraries. Scikit-learn is a free software machine learning library for the Python programming language. It offers a variety of classification, regression, and clustering algorithms including SVM, random forests, and gradient boosting. It is designed to interoperate with the Python numerical and scientific libraries NumPy and SciPy.

1. Experimental Process

This study intends to classify sentiments by applying machine learning techniques to the vector of word embedding. First, IMDB text data is divided into positives and negatives. Next, the training and the test datasets were composed for classifying the sentiments. As a result, training dataset is 70% (21,000), and test dataset is composed of 30% (9,000) of the whole dataset. Labeling is made by classifying with training and test set vector. Word embedding is prepared by using BoW, TF-IDF, and Word2Vec. In addition to NB and SVM, which have been commonly used for sentiments analysis, the ensemble technique, random forests (RF), gradient boosting, and XGBoost, which are widely used in classification problems, is additionally used. The experimental process is as shown in Fig.2.

2. Data Preprocessing

The IMDB dataset is data that is labeled “1” for positive and “0” for negative based on review data. In general, proper preprocessing is required to apply machine learning techniques to text data. It follows procedures such as review, removing punctuation, tokenization, removing stopwords, and lemmatization.

Fig.3 visualizes t-SNE(t-stochastic neighbor embedding) using Word2Vec to compare text preprocessing. t-SNE expresses high dimensional data as a two-dimensional map by learning two-dimensional embedding vectors that preserve neighbor structures between data represented by high-dimensional vectors. t-SNE shows more stable

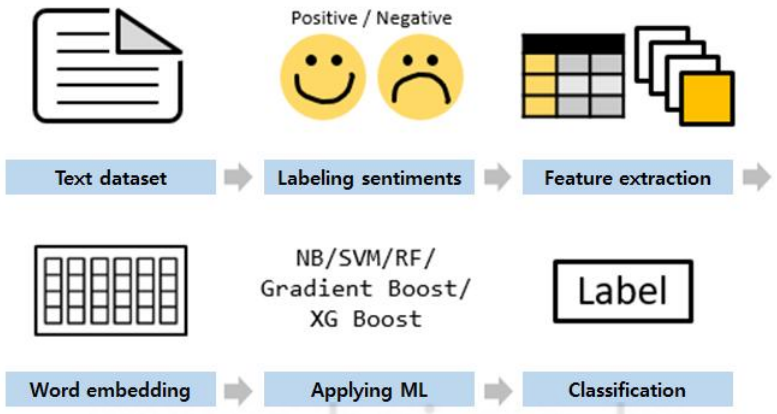

Fig. 2. Experimental process

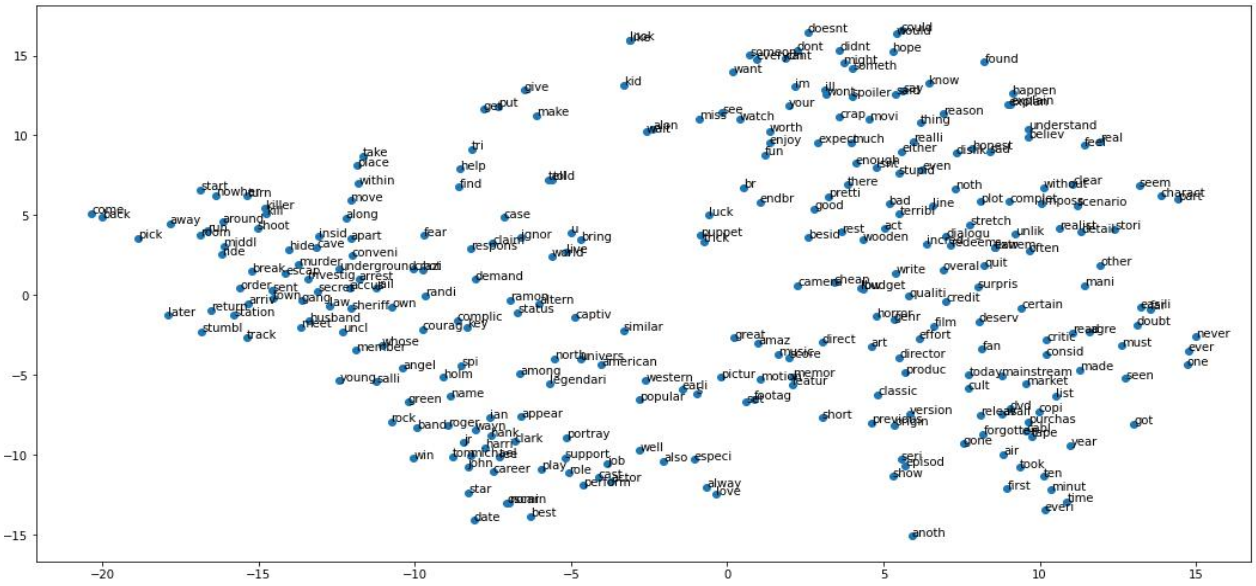

Fig. 3. Mapping word embedding with Word2Vec

embedding learning results than other algorithms for vector visualization. This is because t-SNE converts the distance between data into stochastic probability and uses it for embedding. And this stochastic probability is controlled by perplexity.

3. Evaluation Metrics

The results of the proposed models were based on the popular evaluation metrics such as Precision, Accuracy, Recall, and F1-score.

4. Experimental Results

In this study, we experimented with NB, SVM, RF, XGboost, and gradient boosting as comparative models. The performance is compared based on the accuracy measurement according to the word embedding method of BoW, TF-IDF, and Word2Vec.

Table 2 shows results using BoW, and Table 3 uses TF-IDF. In word embedding of both models, SVM showed the highest accuracy for the test dataset. SVM is 87.21% in BoW and 89.42% in TF-IDF. The lowest accuracy is NB (76.70%) for BoW and RF (80.91%) for TF-IDF.

Table 2. Experimental results of BoW (%)

Model	Precision	Recall	F1-score	Accuracy
NB	77.03	77.08	77.06	76.70
SVM	84.44	89.00	86.66	87.21
RF	82.55	82.47	82.51	82.22
Gradient Boosting	84.16	89.16	86.59	86.76
XGBoost	84.83	82.49	83.65	83.22
Average	82.60	84.04	83.29	83.22

Table 3. Experimental results of TF-IDF (%)

Model	Precision	Recall	F1-score	Accuracy
NB	82.43	82.36	82.39	82.16
SVM	91.21	87.77	89.46	89.42
RF	84.51	78.59	81.44	80.91
Gradient Boosting	84.71	87.15	85.91	85.89
XGBoost	84.62	81.85	83.21	82.96
Average	85.50	83.54	84.48	84.27

Table 4 presents the results of Word2Vec. It shows BoW and TF-IDF outperformed Word2Vec. The accuracy of Word2Vec was the lowest at NB (74.10%). Except for NB, the accuracies of Word2Vec were close to the average (81.23%).

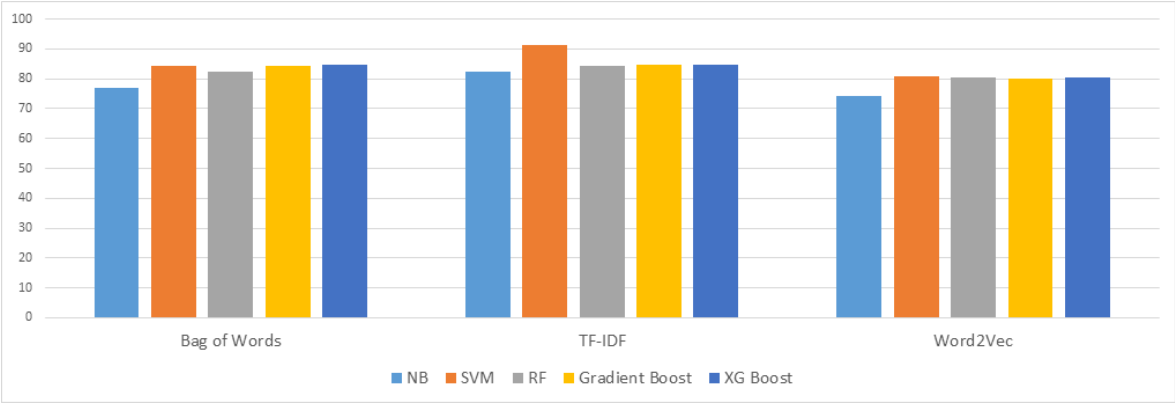

Fig. 4. The bar chart of the performances for each model

Overall, the average accuracy was 82.23% for BoW, 84.27% for TF-IDF, and 79.8% for Word2Vec. The following Fig.4 shows the bar chart of the performances for each model.

Table 4. Experimental results of Word2Vec (%)

Model	Precision	Recall	F1-score	Accuracy
NB	74.09	72.83	73.45	74.10
SVM	80.88	81.50	81.19	81.42
RF	80.31	82.54	81.41	81.46
Gradient Boosting	80.03	81.10	80.56	80.74
XGBoost	80.50	81.75	81.12	81.28
Average	79.16	79.94	79.55	79.80

In this study, we implement a two-sample tests for proportions to validate the differences in the average accuracies of three embedding methods. Table 5 shows TF-IDF outperforms BoW at 5% and Word2Vec at 1% statistical significance level. And BoW performs better than Word2Vec at 1% statistical significance level. The results show the differences in the average accuracies are statistically significant to each other.

Table 5. Results of statistical significance test (Z-score)

Model	TF-IDF	BoW
BoW	1.9091*	
Word2Vec	7.8109**	5.9096**

(*: significant at 5% level, **: significant at 1% level)

IV. Conclusions

The purpose of this study is to explore for a word embedding method suitable for sentiment analysis that classifies positive and negative. Many studies have been conducted to explore a machine learning technique ideal for the classification of sentiments, but there have been insufficient studies to find a suitable word embedding method for sentiment analysis.

To this end, BoW, TF-IDF, and Word2Vec were compared by word embedding method, and NB, SVM, RF, Gradient boosting, and XGBoost were used to classify sentiments in this study. As a result, TF-IDF(84.27%) was the best, and Word2Vec(79.8%) was the lowest. This result indicated that vector modeling in word embedding of sentiment analysis is more suitable for machine learning than sequential modeling. Many studies show that Word2Vec is ideal for deep learning than machine learning. However, the reason we studied in this paper was that Word2Vec was flexible for words, so we experimented with increasing the similarity.

This study has business significance because it has confirmed the word embedding technique suitable for sentiment analysis by confirming the impact of various word embedding techniques in the sentiment analysis for product reviews, which are widely used in the recent business research area. There has been a lot of interest in sentiment analysis for product reviews, but by confirming

the difference in performance according to the choice of word embedding method, it is possible to perform more accurate sentiment analysis, which is believed to contribute to the advancement of marketing analytics. Besides, academic research has been conducted with a lot of interest in the classification technique of sentiment analysis, but there is a contribution that has made an academic contribution that can improve the results of sentiment analysis by analyzing the impact of the difference in word embedding techniques.

In this study, the IMDB dataset, which is a representative movie review dataset, was used for analysis, but only one dataset was used, so there may be concerns about generalization of the results. In future research, the possibility of generalization of the results of this study can be confirmed by additionally analyzing various movie review datasets and product review datasets in other fields.

Regarding classification, although representative classification techniques were used in this study, there is a limitation in that the deep learning technique was not used. The deep learning technique, which has received much attention recently, is expected to show good results in solving classification problems in sentiment analysis, and this will be one of our future research topics.

Lastly, among the word embedding techniques, there is a limitation in that the BERT(Bidirectional Encoder Representations from Transformers) technique, which has received a lot of interest, was not compared. This should be supplemented in future studies.

ACKNOWLEDGEMENT

This work was supported by the Dongguk University Research Fund of 2019.

REFERENCES

[1] T. A. Rana and Y.-N. Cheah, "Aspect extraction in sentiment analysis: comparative analysis and survey," Artificial Intelligence Review, vol. 46, no. 4, pp. 459-483, Feb. 2016.

[2] Q. T. Ain, M. Ali, A. Riaz, A. Noureen, M. Kamran, B. Hayat, and A. Rehman, "Sentiment analysis using deep learning techniques: a review," International Journal of Advanced Computer Science and Applications, vol. 8, no. 6, pp. 424-433, Jun. 2017.

[3] A. Abdi, S. M. Shamsuddin, S. Hasan, and J. Piran, "Deep learning-based sentiment classification of evaluative text based on Multi-feature fusion," Information Processing & Management, vol. 56, no. 4, pp. 1245-1259, Jul. 2019.

[4] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up? Sentiment classification using machine learning techniques." in Proc. of EMNLP 2002, pp. 79-86, Jul. 2002.

[5] F. H. Khan, U. Qamar, and S. Bashir, "SentiMI: Introducing point-wise mutual information with SentiWordNet to improve sentiment polarity detection," Applied Soft Computing, vol. 39, pp. 140-153, Feb. 2016.

[6] F. Tang, L. Fu, B. Yao, and W. Xu, "Aspect based fine-grained sentiment analysis for online reviews," Information Sciences, vol. 488, pp. 190-204, Jul. 2019.

[7] C. Bhadane, H. Dalal, and H. Doshi, "Sentiment analysis: Measuring opinions," Procedia Computer Science, vol. 45, no. 0, pp. 808-814, Mar. 2015.

[8] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in Advances in Neural Information Processing Systems, pp. 3111-3119, 2013.

[9] W. J. Kim, D. H. Kim and H. W. Jang, "Semantic extension search for documents using the Word2vec," Journal of the Korea Contents Association, vol. 16, no. 10, pp. 687-692, Oct. 2016.

[10] D. K. Sung, and Y. S. Jeong, "Political opinion mining from article comments using deep learning," Journal of The Korea Society of Computer and Information, vol. 23, no. 1, pp. 9-15, Jan. 2018.

[11] T. Lee, K. Kim, J. Lee, and S. Lee, "An efficient BotNet detection scheme exploiting Word2Vec and accelerated hierarchical density-based clustering," Journal of Internet Computing and Services, vol. 20, no. 6, pp. 11-20, Dec. 2019.

[12] E. H. Kim, "A deeping learning-based article and paragraph-level classification," Journal of the Korea Society of Computer and Information, vol. 23, no. 11, pp. 31-41, Nov. 2018.

[13] J. Park, H. Kim, H. G. Kim, T. K. Ahn, and H. Yi, "Structuring of unstructured SNS messages on rail services using deep learning techniques," Journal of The Korea Society of Computer

and Information, vol. 23, no. 7, pp. 19-26, Jul. 2018.
[14] S. M. Liu and J.-H. Chen, "A multi-label classification based approach for sentiment classification," Expert Systems with Applications, vol. 42, no. 3, pp. 1083-1093, Feb. 2015.
[15] G. Gautam and D. Yadav, "Sentiment analysis of twitter data using machine learning approaches and semantic analysis," in Proc. of IC3, IEEE, pp. 437-442, Aug. 2014.
[16] J. Read, "Using emoticons to reduce dependency in machine learning techniques for sentiment classification," in Proceedings of the ACL Student Research Workshop, pp. 43-48, Jun. 2005.
[17] L. Dey, S. Chakraborty, A. Biswas, B. Bose, and S. Tiwari, "Sentiment analysis of review datasets using Naive Bayes and k-nn classifier," International Journal of Information Engineering and Electronic Business, vol. 8, no. 4, pp. 54-62, Jul. 2016.

Authors

Hoyeon Park received the B.S. degree in Computer Sciences, M.B.A. degree in Management Information Systems(MIS) from Dongguk University, Korea. She is currently a Ph.D candidate in the Department of MIS, Dongguk University. She is interested in deep learning, big data analytics, and text mining.

Kyoung-jae Kim received the B.B.A. degree from Chungang University, and M.E. and Ph.D. degrees in Management Engineering from KAIST, Korea. Dr. Kim joined the faculty of the Department of MIS at Dongguk University, Seoul, Korea, in 2003. He is currently a Professor in the Department of MIS, Dongguk University. He is interested in business analytics, customer relationship management, recommender systems, and big data analytics.