

# Comparison of Sentiment Analysis from Large Twitter Datasets by Naïve Bayes and Natural Language Processing Methods

Bong-Hyun Back<sup>1</sup> and Il-Kyu Ha<sup>2\*</sup> , Member, KIICE

<sup>1</sup>Department of Computer Engineering, Yeungnam University, Gyeongsan 38541, Korea

<sup>2</sup>Department of Computer Engineering, Kyungil University, Gyeongsan 38428, Korea

## Abstract

Recently, effort to obtain various information from the vast amount of social network services (SNS) big data generated in daily life has expanded. SNS big data comprise sentences classified as unstructured data, which complicates data processing. As the amount of processing increases, a rapid processing technique is required to extract valuable information from SNS big data. We herein propose a system that can extract human sentiment information from vast amounts of SNS unstructured big data using the naïve Bayes algorithm and natural language processing (NLP). Furthermore, we analyze the effectiveness of the proposed method through various experiments. Based on sentiment accuracy analysis, experimental results showed that the machine learning method using the naïve Bayes algorithm afforded a 63.5% accuracy, which was lower than that yielded by the NLP method. However, based on data processing speed analysis, the machine learning method by the naïve Bayes algorithm demonstrated a processing performance that was approximately 5.4 times higher than that by the NLP method.

**Index Terms:** Big data processing, Machine learning, Naïve Bayes algorithm, Sentiment analysis, SNS big data

## I. INTRODUCTION

Recently, the number of users of social network services (SNS) has increased owing to the explosive growth of mobile devices, and the amount of data generated on SNS has increased correspondingly. SNS is widely used for social relations and friendship; however, recently, it has been increasingly used for the secondary purpose of gathering and analyzing large datasets on SNS and obtaining various pieces of information [1-3].

The data on SNS include content related to opinions being expressed in various fields such as economy, society, and culture. Therefore, by analyzing the data on SNS, information regarding various flows and opinions on topics such as society, economy, and politics can be extracted.

In recent years, as interest in big-data processing has

increased, studies have been conducted for collecting and storing big data stably and processing data more efficiently using limited computing resources [4-9]. In addition, studies are being conducted to improve the performance of big-data processing using machine learning methods. These studies suggest methods to more accurately process big data by adding a machine learning algorithm to the big-data-processing algorithm.

However, fewer studies are available regarding sentiment analysis by machine learning and natural language processing (NLP).

Therefore, we herein propose an effective data pattern analysis method that can extract sentiment information such as positive, negative, and neutral by analyzing patterns in SNS big data. In particular, we propose pattern analysis methods based on the naïve Bayes algorithm and NLP, and


Received 20 August 2019, Revised 05 November 2019, Accepted 06 November 2019

\*Corresponding Author Il-Kyu Ha (E-mail: [ikha@kiu.kr](mailto:ikha@kiu.kr), Tel: +82-53-600-5564)

Department of Computer Engineering, Kyungil University, Gyeongsan 38428, Republic of Korea.

**Open Access** <https://doi.org/10.6109/jicce.2019.17.4.239>

print ISSN: 2234-8255 online ISSN: 2234-8883

 This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<http://creativecommons.org/licenses/by-nc/3.0/>) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

Copyright © The Korea Institute of Information and Communication Engineering

we compare the feature points of these two methods.

This paper is organized as follows: In Chapter 2, we analyze related studies including those on big data processing. In Chapter 3, we describe the proposed pattern analyzer. In Chapter 4, we analyze the performance of the proposed system to evaluate its effectiveness. Finally, Chapter 5 concludes this study.

## II. RELATED STUDIES

In this section, we examine and analyze studies related to big data, including distribution processing and machine learning. The significance of this study is discussed in the context of prior studies.

First, to extract various pieces of information from a large quantity of SNS data, a technique for a stable collection and accumulation of SNS data is required, and methods for processing information according to a specific purpose are necessitated. Hence, the following research was conducted. Ha et al. [4] studied the method for storing SNS data reliably in Hadoop-based distributed systems and for separating sentiment information from various users. Gu et al. [5] argued that three factors were important for big data processing services in geographically distributed data centers: task assignment, data placement, and data movement. Furthermore, they studied how these three factors could be optimized. Ji et al. [6] revealed key issues such as distributed file systems, non-structural and semistructured data storage and open-source cloud platforms in the cloud computing environment and introduced optimization strategies for MapReduce, a big data processing model.

Zhu et al. [7] emphasized that technology can increase the use frequency of data collected when processing sensing data from the Internet of things (IoT) environment. Furthermore, they demonstrated the importance of cloud computing in the processing of large datasets collected from an IoT environment. Garlasu et al. [8] explained the importance of storage capacity and processing power in a big data processing environment and suggested grid computing as a distributed method that could solve the problem above. In addition, they described the key components of a grid computing system. Tan et al. [9] recognized SNS as informative and executable according to networking technology and analyzed the characteristics of each group. Finally, they explained that leveraging the social network paradigm could solve big data processing challenges. In addition, we discovered various studies [10-13] regarding big data processing system construction and big data processing methods. Among them, a few studies [14-19] have reported the processing of big data in social networks.

A technique for processing big data more efficiently is machine learning. Some of the studies regarding this topic

are discussed as follows.

Qiu et al. [20] described recent trends in the studies of machine learning for big data processing. State-of-the-art techniques such as deep learning, distributed and parallel learning, and representation learning have been described. Suthanharan et al. [21] discussed problems and challenges in managing big data networking technologies. To better explain the analysis in this study, the characteristics of big data were explained by three C factors, i.e., complexity, continuity, and cardinality.

Jarrah et al. [22] studied effective machine learning methods for processing big data. They explored data modeling methods and analyzed the efficiencies of the model and algorithm. Landset et al. [23] classified tools for machine learning as processing engines, machine learning frameworks, and learning algorithms and analyzed their association. Furthermore, machine learning frameworks such as Mahout, MLlib, H<sub>2</sub>O, and Samoa have been examined in parallel. Xing et al. [24] analyzed and compared implementation engines such as MapReduce, Spark, Flink, Storm, and H<sub>2</sub>O in the Hadoop ecosystem, a typical machine learning architecture. In addition, machine learning libraries and frameworks such as Mahout, MLlib, and Samoa have been examined. Chen et al. [25] proposed an algorithm for disease prediction based on machine learning for healthcare big data and demonstrated the effectiveness of the proposed algorithm experimentally. In addition, we discovered studies [26-28] that suggested a big data processing method using various machine learning techniques.

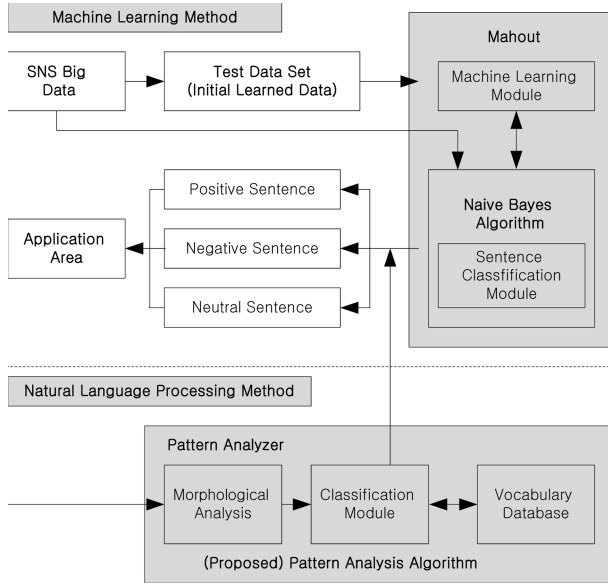
Compared with existing studies, this study exhibits the following characteristics. First, the proposed system extracts sentiment information that can be used for various purposes from SNS big data. Such sentiment information may be applied to various fields such as advertising, safety, and politics. Next, sentiment information is extracted by machine learning using the naïve Bayes algorithm and NLP. The feature points of the two methods are extracted through comparison.

## III. EMOTIONAL PATTERN ANALYSIS SYSTEM

In this section, we introduce the proposed system. The core building blocks of the system are explained, and the functional roles of the components are discussed.

We propose two types of pattern analysis: pattern analysis by machine learning and that by NLP. The former is a type of the data classification based on the Mahout machine learning module, while the latter analyzes patterns by performing morphological analysis on each sentence and referring to a data dictionary.

Fig. 1 shows the structure of the proposed sentiment analysis system. The proposed system is divided into two parts.



**Fig. 1.** Structure of the proposed sentiment analysis system.

The upper part of the figure shows the structure of the sentiment analysis system by machine learning, while the lower part shows that by the NLP method. In the machine learning method, the proposed system comprises various components including an initial test dataset, Mahout machine learning module, and sentence classification module. Initially learned data are generated by human judgment for the test dataset. The initial learned data are processed by Mahout's machine learning library and the sentiment is analyzed by the sentence classification module of the naïve Bayes algorithm. The classification module classifies the SNS sentences into three types of sentences: positive, negative, and neutral.

In the NLP method, SNS big data are sent to a pattern analyzer to extract information.

The pattern analyzer executes the following steps sequentially to process the input data and output the result. In the first step, sentences containing various information are received as input from SNS big data. Morpheme analysis is then conducted on the input sentences. In the next step, the pattern analyzer performs classification according to whether a word in a pattern type exists in a sentence. That is, sentences are classified into either positive, negative, or neutral sentences by referring to the data dictionary. Finally, the pattern classification result is output.

Algorithm 1 shows the procedure for pattern analysis using the natural-language processing method.

The input sentences are divided into meaningful words through morphological analysis. In addition, positive and negative counts are calculated by referring to the data dictionary.

“CP” means the number of positive words and “CN” the number of negative words. The total count “CT” can be

obtained by (1). “DV” is a decision value for sentiment decision, and a value of “0” is used in this study.

$$\text{Total Count (CT)} = \text{CP} * (+1) + \text{CN} * (-1). \quad (1)$$

The positive and negative counts are combined to yield a CT that is then compared with a threshold value (DV) to determine the character of the sentence.

#### Algorithm 1 Proposed Pattern Analysis Algorithm

1: **Inputs:**

2: SS: a sentence containing second group word

3: ST: a sentence containing third group word

3: DL- learned data for the keyword

4: **Outputs:**

5: RP: Positive sentence

6: RN: Negative sentence

7: RE: Neutral sentence

8:  $S = SS + ST$

9:  $S = \{S_1, S_2, \dots, S_k\}$  – sentences

10:  $S_i = \{W_1, W_2, \dots, W_k\}$  – syntactic word (morpheme)

11: CP: Positive Count (number of positive words)

12: CN: Negative Count (number of negative words)

13: CT: Total count ( $=CP*(+1)+CN*(-1)$ )

14: DV: Decision value for sentiment decision

15: for each  $S_i$

16: morphological analysis to  $S_i$

17: generate morpheme-specific words  $\{W_1, W_2, \dots, W_k\}$

18: for each  $W_i$

19: calculate CP // referring to the dictionary

20: calculate CN // referring to the dictionary

21: end

22: compute CT //  $CP*(+1)+CN*(-1)$

22: if  $CT == DV$  then

23: classify  $S_i$  to RE

24: else if  $CT > DV$  then

25: classify  $S_i$  to RP

26: else

27: classify  $S_i$  to RN

28: end

30: end

## IV. PERFORMANCE ANALYSIS OF THE PROPOSED SYSTEM

In this section, we analyze the performance of the proposed system. First, we construct a distributed processing system for the proposed method to analyze performance and conduct various experiments using SNS data as sample data in the system. Results from various experiments are then derived and analyzed. We study the performance level according to the analysis and demonstrate the effectiveness

of the proposed system.

The proposed system was implemented as described in Section 3. A Hadoop-based distributed system was used. Fig. 2 shows the architecture of the system. Three servers were built for a distributed processing, and a console PC with a Windows operating system was built for various operations including that of the server.

In Fig. 2, Servers 1, 2, and 3 are deployed as a distributed processing system based on Hadoop. We used MongoDB as a database management system and CentOS as the operating system. Mahout 0.8 was used as a library for machine learning. Mahout is a machine learning library that can be distributed and parallelized in big data processing. It provides various machine learning algorithms in the form of libraries. In this study, naïve Bayes classifiers were used in the library.

Servers 1, 2, and 3 use Intel Xeon CPUs running CentOS 6.7 with 128 GB or 64 GB of memory. Hadoop 2.7.3 was used for the distributed processing, MongoDB as the database, and JAVA as the processing language. A console PC was based on the Windows operating system was used as the console.

Experiments for the performance analysis were conducted in the environment shown in Table 1. In this experiment, we conducted two types of experiments: pattern analysis of unstructured data and that of data processing speed. For the pattern analysis experiments, five major keywords which are the most recent issue were selected for Twitter data, and 1,000 pieces of tweet sentences related to each keyword were used as experimental data. The five keywords used were as follows: Galaxy8, Kimjeongeun, Moonjaein, Thaad, and Trump, and the experiment was conducted over five days. For the data processing speed experiment, 10,000 sen-

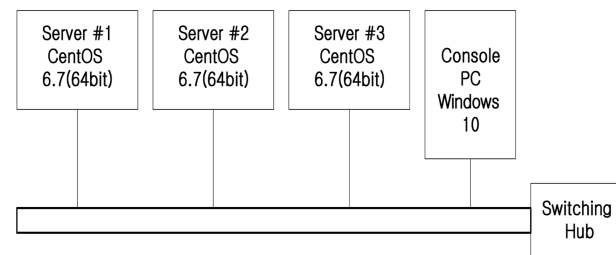


Fig. 2. Distributed processing system for experiment.

Table 1. Environment of the target search experiment

Item	Environment
Experimental data	Pattern Analysis: 1,000 pieces of Twitter sentences for five major keywords (total 5,000 sentences)
	Data Processing Speed: 10,000 pieces of tweet sentences for five major keywords (total 50,000 sentences)
Keywords	Galaxy8, Kimjeongeun, Moonjaein, Thaad, Trump (Korean language)
Experimental date	Dec 17, 2018–Dec 24, 2018

tences were used for each of the five key keywords. Thus, the overall experimental data comprised approximately 50,000 sentences.

Herein, we analyze the accuracy of the analysis of unstructured data by machine learning using naïve Bayes by keyword. The experiment was conducted to measure the consistency between the classified data and the correct answer set following the use of pattern analysis on the data originally classified via machine learning. The accuracy for keyword “Galaxy8” was the highest at 69.09%, and that for keyword “Trump” was the lowest at 55.49%. The experimental results are shown in Fig. 3, and the accuracy was derived in accordance with (2), where  $Num^{mat}$  represents the number of data points that match the correct answer set and  $Num^{def}$  the number of data points classified as definite data.

$$\text{Accuracy of Analysis} = \frac{Num^{mat}}{Num^{def}} \times 100. \quad (2)$$

Experiments to determine the accuracy of unstructured data handling via NLP were performed and the results are analyzed. The aim of this experiment was to measure the consistency between the classified data and the set of correct answers when pattern analysis was performed on the data classified using the general NLP, not involving machine learning. The accuracy was for keyword “Galaxy8” was the highest at 80.00%, and that of keyword “Thaad” was 64.25%. The accuracy was derived in accordance with (2).

Fig. 4 shows the accuracy of pattern analysis by machine learning for each keyword, while Fig. 5 shows that by NLP. In the pattern analysis method by machine learning using the naïve Bayes algorithm, the accuracy of the “neutral” opinion was the highest for each keyword, and the pattern analysis method by NLP exhibited the highest accuracy for the “negative” opinion for each keyword.

Fig. 6 shows the results of the pattern analysis accuracy tests. As shown, the accuracy in the pattern analysis test of unstructured data by machine learning using naïve Bayes is

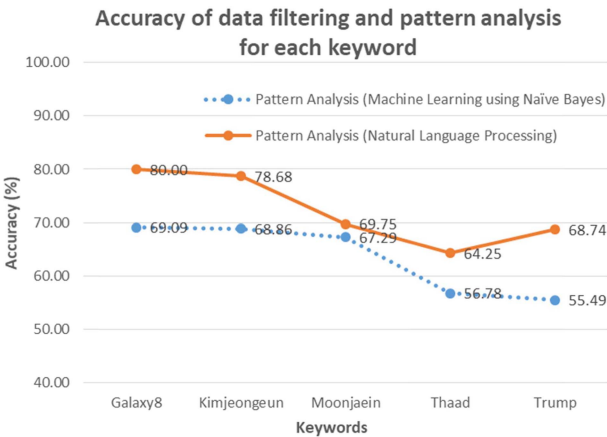
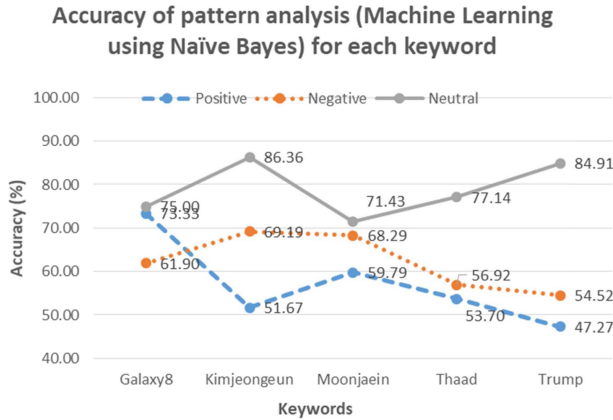
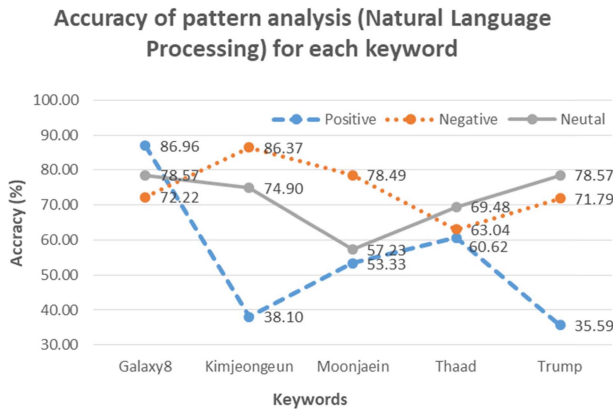


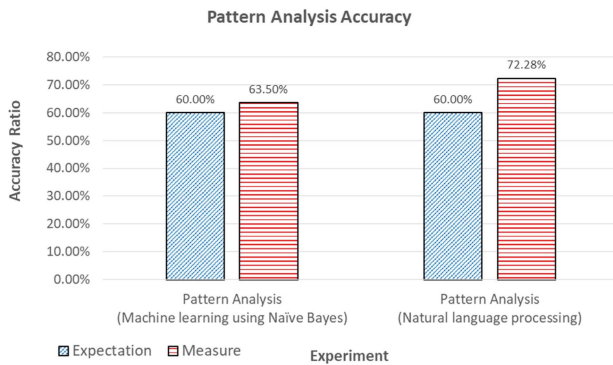
Fig. 3. Accuracy of pattern analysis for each keyword.



**Fig. 4.** Accuracy of pattern analysis (machine Learning using naïve Bayes) for each keyword.



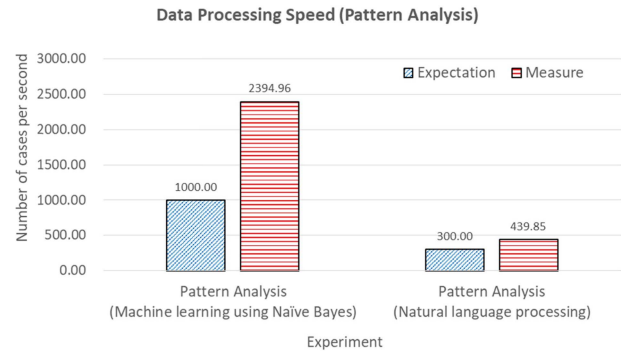
**Fig. 5.** Accuracy of pattern analysis (NLP) for each keyword.



**Fig. 6.** Accuracy of pattern analysis.

63.50%, which is higher than the expected value of 60%. Meanwhile, that by NLP is 72.28%, which is higher than the expected value of 60%.

Therefore, it is clear that the accuracy of pattern analysis by machine learning using naïve Bayes (63.50%) is lower than that by NLP (72.28%). However, the performance of the pattern analysis by the machine learning method using



**Fig. 7.** Pattern analysis speed for unstructured data.

naïve Bayes is excellent in terms of the data processing speed, as will be described in the next section.

Herein, we analyze the speed of pattern analysis by machine learning using naïve Bayes and NLP by keyword. The aim of this experiment was to measure the processing speed when pattern analysis was performed using the proposed machine learning or NLP methods. For the data processing speed experiment, 10,000 sentences were used for each of the five key keywords. Thus, the overall experimental data comprised approximately 50,000 sentences. The experimental results are shown in Fig. 8, and the speed was assessed in accordance with the (3), where  $Num^{proc}$  represents the number of processed data points and  $Time^{analysis}$  the data-analysis time frame in seconds.

$$\text{Speed of analysis} = \frac{Num^{proc}}{Time^{analysis}}. \quad (3)$$

Fig. 7 shows the experimental results for the speed of pattern analysis for unstructured data. The proposed machine learning pattern analysis method handles 2394.96 cases per second, which is higher than the expected 1000 cases per second; the NLP method handles 439.85 cases per second, which is higher than the expected 300 cases per second.

It is clear that the accuracy of pattern analysis by machine learning using naïve Bayes is lower than that by NLP; however, it is superior to pattern analysis by NLP in terms of data processing speed.

Additionally, the machine learning method using naïve Bayes is less accurate than NLP, but the processing speed is significantly better.

The reasons can be explained as follows. In the NLP method, a large amount of processing time is required because a database in which positive, negative, and neutral data are stored must be accessed every time to determine the positive, negative, and neutral of each word separated through morphological analysis. On the contrary, in the machine learning method using Naïve Bayes, because a modeler composed of learned data is constructed and sentiment judgment is immediately made by the modeler, the pro-

cessing time is relatively fast.

Therefore, we can conclude that the machine learning method by the naïve Bayes algorithm demonstrates better processing capacity in terms of processing speed than the NLP method. Furthermore, it is clear that pattern analysis by the machine learning method using naïve Bayes is more efficient in big data processing environment that processes large amount of data quickly.

## V. CONCLUSIONS

In this study, we proposed and implemented a system that could extract human sentiment information from vast amounts of SNS unstructured big data using the naïve Bayes algorithm and NLP.

To evaluate the performance of pattern analysis in the proposed system, several experiments were conducted.

Concerning the accuracy experiment, the accuracy of the pattern analysis showed different results depending on the analysis method. The accuracy of pattern analysis by machine learning was 63.50%, and that by NLP was 72.28% on average.

The accuracy of pattern analysis by machine learning using naïve Bayes was lower than that by NLP. However, based on the data processing speed experiment, the speed of pattern analysis by machine learning using naïve Bayes was approximately 5.45 times faster than that by NLP.

Therefore, the pattern analysis method by machine learning using naïve Bayes was less accurate than that by NLP; however, it may be advantageous in a big data processing environment where a large amount of data must be processed quickly.

The contribution of this study can be summarized as follows.

First, effective data pattern analysis methods that could extract sentiment information such as positive, negative, and neutral by analyzing patterns in unstructured SNS big data were proposed. In particular, pattern analysis methods by machine learning using naïve Bayes and natural language processing were proposed.

Additionally, the effectiveness and efficiency of the two methods were compared experimentally.

In our opinion, the results of this study are applicable for the efficient processing of SNS big data and for obtaining emotional information from SNS big data.

## ACKNOWLEDGMENTS

This research was supported by the Basic Science Research Program of the National Research Foundation of Korea (NRF) funded by the Ministry of Education

(2017R1D1A1B03029895). The results of were based a study on the “Leaders in INdustry-university Cooperation+ (LINC+)” Project, supported by the Ministry of Education (MOE) and the National Research Foundation of Korea (NRF).

## REFERENCES

- [1] X. Wu, X. Zhu, G. Wu, and W. Ding, “Data mining with big data,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 1, pp. 97-107, 2014. DOI: 10.1109/TKDE.2013.109.
- [2] C. Cheng and C. Zhang, “Data-intensive applications, challenges, techniques and technologies: A survey on Big Data,” *Information Sciences*, vol. 275, pp. 314-347, 2014. DOI: 10.1016/j.ins.2014.01.015.
- [3] K. Riesen and H. Bunke, “IAM graph database repository for graph-based pattern recognition and machine learning,” *Lecture Notes in Computer Science*, vol. 5342, pp. 287-297, 2008. DOI: 10.1007/978-3-540-89689-0\_33.
- [4] I. Ha, B. Bak, and B. Ahn, “MapReduce functions to analyze sentiment information from social big data,” *International Journal of Distributed Sensor Networks*, vol. 11, no. 6, pp. 1-11, 2015. DOI: 10.1155/2015/417502.
- [5] L. Gu, D. Zeng, P. Li, and S. Guo, “Cost minimization for big data processing in geo-distributed data centers,” *IEEE Transactions on Emerging Topics in Computing*, vol. 2, no. 3, pp. 314-323, 2013. DOI: 10.1109/TETC.2014.2310456.
- [6] C. Ji, Y. Ki, W. Qiu, U. Awada, and K. Li, “Big data processing in cloud computing environments,” in *Proceeding of 2012 International Symposium on Pervasive Systems, Algorithms and Networks*, San Marcos, TX, USA, pp.17-23, 2012. DOI: 10.1109/I-SPAN.2012.9.
- [7] T. Zhu, S. Xiao, Q. Zhang, Y. Gu, P. Yi, and Y. Li, “Emergent technologies in big data sensing: a survey,” *International Journal of Distributed Sensor Networks*, vol. 11, no. 10, pp. 1-3, 2015. DOI: 10.1155/2015/902982.
- [8] D. Garlasu, V. Sandulescu, I. Halcu, G. Neculoiu, and V. Marinescu, “A big data implementation based on grid computing,” in *Proceeding of the 2013 11th RoEduNet International Conference*, Sinaia, Romania, pp. 1-4, 2013. DOI: 10.1109/RoEduNet.2013.6511732.
- [9] W. Tan, M. Blake, I. Saleh, and S. Dustdar, “Social-network-sourced big data analytics,” *IEEE Internet Computing*, vol. 17, no. 5, pp. 62-69, 2013. DOI: 10.1109/MIC.2013.100.
- [10] W. Lizhe, M. Yan, Y. Jining, C. Victor, and Z. Albert, “pipsCloud: high performance cloud computing for remote sensing big data management and processing,” *Future Generation Computer Systems*, vol. 78, no. 1, pp. 353-368, 2016. DOI: 10.1016/j.future.2016.06.009.
- [11] B. Philip, T. Yong, E. Edward, R. William, N. Alex, N. Carl, C. Michael, P. Clinton, and C. Bridget, “Big data in cryoEM: automated collection, processing and accessibility of EM data,” *Current Opinion in Microbiology*, vol. 43, pp. 1-8, 2018. DOI: 10.1016/j.mib.2017.10.005.
- [12] O. Ahmed, B. Fatima, L. Ayoub, and B. Samir, “Big data technologies: A survey,” *Journal of King Saud University - Computer and Information Sciences*, vol. 30, no. 4, pp. 431-448, 2017. DOI: 10.1016/j.jksuci.2017.06.001.
- [13] Z. Qingchen, Y. Laurence, C. Zhikui, L. Peng, “A survey on deep learning for big data,” *Information Fusion*, vol. 42, pp. 146-157, 2017. DOI: 10.1016/j.inffus.2017.10.006.
- [14] B. Gema, J. Jason, and C. David, “Social big data: Recent achievements and new challenges,” *Information Fusion*, vol. 28, pp. 45-59, 2016.



DOI: 10.1016/j.inffus.2015.08.005.

- [15] K. Avita, W. Mohammad, and R. Goudar, "Big data: issues, challenges, tools and good practices," in *Proceeding of the 2013 Sixth International Conference on Contemporary Computing (IC3)*, Noida, India, pp. 404-409, 2013. DOI: 10.1109/IC3.2013.6612229.
- [16] G. Gerard, H. Martine, and P. Alex, "Big data and management," *Academy of Management Journal*, vol. 57, no. 2, pp. 321-326, 2014. DOI: 10.5465/amj.2014.4002.
- [17] L. Carson, J. Fan, P. Tik, and C. Paul, "Big data analytics of social network data: who cares most about you on Facebook?," *Big Data*, vol. 27, pp. 1-15, 2017. DOI: 10.1007/978-3-319-60255-4\_1.
- [18] B. Desamparados and D. Josep, "Big data sources and methods for social and economic analyses," *Technological Forecasting and Social Change*, vol. 130, pp.99-113, 2018. DOI: 10.1016/j.techfore.2017.07.027.
- [19] H. Amir and C. Erik, "Semi-supervised learning for big social data analysis," *Neurocomputing*, vol. 275, no. 31, pp. 1662-1673, 2017. DOI: 10.1016/j.neucom.2017.10.010.
- [20] J. Qiu, Q. Wu, G. Ding, Y. Xu, and S. Feng, "A survey of machine learning for big data processing," *EURASIP Journal on Advances in Signal Processing*, vol. 2016, pp. 1-16, 2016. DOI: 10.1186/s13634-016-0355-x.
- [21] S. Suthanharan, "Big data classification: problems and challenges in network intrusion prediction with machine learning," *Performance Evaluation Review*, vol. 41, no. 4, pp. 70-73, 2014. DOI: 10.1145/2627534.2627557.
- [22] O. Jarrah, P. Yoo, S. Muhaidat, G. Karagiannidis, and K. Taha, "Efficient machine learning for big data: A review," *Big Data Research*, vol. 2, no. 3, pp. 87-93, 2015. DOI: 10.1016/j.bdr.2015.04.001.
- [23] S. Landset, T. Khoshgoftaar, A. Richter, and T. Hasanin, "A survey of open source tools for machine learning with big data in the Hadoop ecosystem," *Journal of Big Data*, vol. 2, no. 24, pp. 1-36, 2015. DOI: 10.1186/s40537-015-0032-1.
- [24] E. Xing, Q. Ho, W. Dai, J. Kim, and Y. Yu, "Petuum: a new platform for distributed machine learning on big data," *IEEE Transactions on Big Data*, vol. 1, no. 2, pp. 49-67, 2015. DOI: 10.1109/TBDATA.2015.2472014.
- [25] M. Chen, Y. Hao, K. Hwang, L. Wang, and L. Wang, "Disease prediction by machine learning over big data from healthcare communities," *IEEE Access*, vol. 5, pp. 8869-8879, 2017. DOI: 10.1109/ACCESS.2017.2694446.
- [26] M. Gunasekaran, V. Vijayakumar, R. Varatharajan, K. Priyan S. Revathi, and H. Ching-Hsien, "Machine learning based big data processing framework for cancer diagnosis using hidden Markov model and GM clustering," *Wireless Personal Communications*, vol. 102, no. 3, pp. 2099-2116, 2018. DOI: 10.1007/s11277-017-5044-z.
- [27] W. Xiaofei, Z. Yuhua, L. Victor, G. Nadra, and J. Tianpeng, "D2D big data: content deliveries over wireless device-to-device sharing in large-scale mobile networks," *IEEE Wireless Communications*, vol. 25, no. 1, pp. 32-38, 2018. DOI: 10.1109/MWC.2018.1700215.
- [28] Z. Zhenhua, H. Qing, G. Jing, and N. Ming, "A deep learning approach for detecting traffic accidents from social media data," *Transportation Research Part C: Emerging Technologies*, vol. 86, pp. 580-596, 2017. DOI: 10.1016/j.trc.2017.11.027.



### Bong-Hyun Back

He received his Ph.D. degree in computer engineering from Yeungnam University, Korea, in 2014. He was a systems manager at the SecuAvail Corporation of Japan. He is currently the CEO of Argos Co. Ltd. and an adjunct professor at Yeungnam University. His research interests include data mining, machine learning, big data processing, and social network analysis.



### Il-Kyu Ha

He received his Ph.D. degree in computer engineering from Yeungnam University, Korea, in 2003. He is currently an assistant professor in the computer engineering department at Kyungil University, Korea. He was a software developer at the Financial Supervisory Service (FSS) of Korea, and a senior researcher at the center for innovation of engineering education at Yeungnam University. His research interests include big data processing, machine learning, sensor networks, body area networks, flying ad-hoc networks, and unmanned aerial vehicles.