ICIC Express Letters
Part B: Applications
Volume **9**, Number **6**, June **2018**

ICIC International ©2018 ISSN 2185-2766
pp. **559–565**

TEXT MINING BASED ONLINE NEWS ANALYSIS ABOUT SMART FACTORY

Youn-Sub Jung and Tai-Woo Chang*

Department of Industrial and Management Engineering
Kyonggi University
154-42 Gwanggyosan-ro, Yeongtong-gu, Suwon, Gyeonggi 16227, Korea
*Corresponding author: keenbee@kgu.ac.kr

Received December 2017; accepted March 2018

Abstract. *Continuous advancement in information and communication technologies leads to wide range of changes in manufacturing field, and they are represented by the term, smart factory. Many people have been interested in them, and they are covered in news articles. In this study, text mining based analysis was conducted to uncover the meaningful trends of topics and issues about smart factory within the online news articles. We used association rule mining of keywords, and topic modeling for Korean online news corpus from 2014 to April 2017. We can find that the number of articles on the technology-oriented topic is increasing and policy-oriented articles have shown a decreasing tendency.*

Keywords: Smart factory, Text mining, Topic modeling, Latent Dirichlet allocation, Online news analysis

1. **Introduction.** Smart factory is a factory that utilizes information and communication technologies to collect and analyze data in real time, so that all situations within the factory are clearly visible, and can be controlled by itself. In manufacturing fields, the goal of both countries and companies is the implementation of a smart factory to strengthen their future competitiveness using Internet of Things (IoT), 3D printing, cyber-physical systems, and cloud computing technologies. As a result, numerous research papers and news articles have been produced. However, it is not easy to understand the industry trends and the trend of public opinion.

In this study, text mining techniques were used to analyze vast amounts of text data in online news articles related to smart factory or smart manufacturing. And the analysis needs to show the trend of public opinion to identify future directions for strengthening manufacturing competitiveness. To find the trends in news articles, keyword trend analysis was conducted using association rule mining (ARM) and topic trend analysis was conducted using latent Dirichlet allocation (LDA) which is commonly used in topic modeling. In this study, we analyzed keywords and topics in online news articles of 84 media related to smart factory.

The remainder of this paper is organized as follows. The related existing literature is reviewed in Section 2. We describe the research method in Section 3 and present the results of the study together with the visual data in Section 4. Section 5 states our conclusions.

2. **Related Works.** Plans for future construction related to smart factory are underway in Korea [1]. Chang and Yang analyzed the keywords that appeared in academic papers of Korean domestic journals related to smart factory using latent semantic analysis (LSA) method to find the research trends [2]. However, we hardly found academic papers which analyze the keywords that appeared in news articles related to smart factory. The objective of this study is to understand public opinion – which topics or events have been

mainly reported in news articles – and the trend of them about the smart factory in Korea.

There were various studies analyzing news articles by using text mining methods. For keyword analysis, many studies have used ARM analysis to identify the association rules of keywords in news articles. ARM is one of representative techniques used to find the correlations between set of items. In case of topic modeling, most studies about news articles were comparative analysis by news companies [3-5]. Whereas, Kim et al. presented headline click-based topic model, which is an extended model of LDA to find the effect of topical context on the click-value of words in headlines [6]. LDA is a widely used generative probabilistic topic modeling technique for collections of discrete data, such as text corpora. The basic idea is that documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over words [7]. As a similar previous research, Park analyzed articles of the Farmers Newspaper using topic modeling, and reviewed how the articles of the Farmers Newspaper have shown their position in the reform of a big agricultural cooperative, Nonghyup [8]. In this study, we have combined appropriate methodologies outlined in previous studies to understand trends in public opinion on smart factories.

3. **Research Methods.** Pre-processing such as removal of duplicated articles, deletion of unnecessary terms, standardization of terms, and generation of term-document matrix is conducted to articles before applying the algorithm. After performing ARM and LDA, we analyzed and visualized the results.

3.1. **Data collection.** We collected contents data of online news articles from a Korean portal site, Naver News (news.naver.com), which contained keywords of ‘smart factory’, ‘smart manufacturing’, ‘smart logistics’, ‘manufacturing innovation 3.0’, and ‘Industry 4.0’. Table 1 shows the number of articles collected by year. To analyze online news reported in Korea, we collected data from Naver News, which provides online news distribution services from most press companies in Korea. In June 2014, Korean government announced the “Manufacturing innovation 3.0” strategy for smart factory. Since then, articles about smart factory have started to appear in earnest. Therefore, we analyzed 15,595 online news data from 2014 to April 2017. A total of 16,248 articles were collected after duplicates had been removed.

Table 1. Number of online news articles related to smart factory

∼2013	***2014***	***2015***	***2016***	***2017 (∼April)***
653	1,294	5,459	5,421	3,421

3.2. **Data pre-processing.** After removing stop words, 442 unique meaningful terms were obtained for the analysis. Specifically, only in case of topic modeling, we also removed three words that occurred in over 60% of articles, ‘smart factory’, ‘technology’, and ‘company’, which affect the inference negatively. Those words with high probability in all the topics, do not help decompose the collection.

3.3. **Association rule mining (ARM).** In this paper, we detected high associative patterns between keywords in online news data. To find out the patterns, we used Apriori algorithm [9]. Apriori algorithm finds the association rule as follows. First, we extract the frequent items exceeding the minimum support by seeking support and constructing a candidate item set based on this. Then, the frequent item set is obtained by repeating the process of extracting the items satisfying the minimum support again. Finally, we find high association between keywords with support, confidence and lift values. Lift values greater than 1 indicate a positive correlation, values equal to 1 indicate zero correlation, and values less than 1 indicate a negative correlation. Lift indicates whether the appearance of the keyword is improved or decreased as compared with when the keyword is independent. For example, if we assume that there are two observed keywords X and Y, we can find values of support, confidence, and lift as follows. We can find P(X), number of articles including keyword X divided by number of total articles, which means the occurrence probability of keyword X.

support ($\mathbf{S}$) = $\mathrm{P(X \cap Y)}$
confidence ($\mathbf{C}$) = $\mathrm{P(Y|X) = P(X \cap Y)/P(X)}$
lift ($\mathbf{L}$) = $\mathrm{P(X \cap Y)/(P(X)P(Y))}$

3.4. **Latent Dirichlet allocation (LDA).** In this algorithm, posterior inference process is required to derive the topic of a document through the model. It is intractable to compute in general, and so a lot of approximate posterior inference algorithms for this model have been developed, including mean field variational methods [7], collapsed Gibbs sampling [10], and collapsed variational inference [11]. We used collapsed Gibbs sampling, as it provides a simple method for obtaining parameter estimates under Dirichlet priors and allows combination of estimates from several local maxima of the posterior distribution [12].

Finally, topical decomposition of the articles was the result of running this algorithm. Since this is unsupervised learning, we need to set the number of topics. To determine the number of topics, we measured the similarity between documents by using cosine similarity which is one of the most popular similarity measures applied to text documents [13]. When documents are represented as term vectors, the similarity of two documents is measured by the cosine of the angle between them.

4. **Research Results.** With ARM analysis, we detected high association rules from keywords. With LDA, we found topics in the news data, and identified trends from 2014-2017 by half year.

4.1. **Results of keyword analysis.** Before ARM analysis, we found the annual occurrence ratio of keywords in the entire articles (Table 2). We excluded ‘smart factory’, ‘company’, ‘technology’, and ‘Korea’ keywords that have been on the top list for over about 60% each year, because we considered they are obvious keywords.

Since 2014, the ratio of ‘government’ and ‘support’ keywords is decreasing annually. On the other hand, from 2015 to 2017, the top keywords are similar. And, the domestic term ‘manufacturing innovation 3.0’ is replaced with the globally used term ‘the 4th industrial revolution’ from 2015 to 2017. And the keywords of key technologies such as artificial intelligence (AI) and IoT were included in the article at a high rate.

In ARM, this study sets minimum support value to 0.1, and confidence value 0.7. Then, rules were ordered by lift value, and we found meaningful relationships. Table 3 shows high association rules with keywords X to Y, support, confidence, and lift value by year.

For example, in 2014, the support value, the probability that both ‘contents’ and ‘business’ appear at the same time, was low at 0.114, but the confidence value, which is the probability of finding ‘business’ in the article containing ‘contents’, was 0.892, which was high. In addition, these two keywords show that probability of being found at the same time is 5.127 times bigger than the one that they will be found by observing an article at random. The relation of other subject words can be interpreted in the same way.

In 2014, the terms of the preparation stage for smart factory such as business future, and launch of ‘creative economy’, which is one of Korean governmental policies, showed high association rule. From 2015 to 2017, the distribution of similar keywords has continued, such as support, production and business. In 2015, especially big data with IoT has a high association. In 2016, platform or automobile with R&D has high correlation. From

Table 2. Occurrence ratio of keywords by year

2014	*ratio*	*2015*	*ratio*	*2016*	*ratio*	*2017 (∼April)*	*ratio*
R&D	0.621	support	0.585	R&D	0.548	R&D	0.567
government	0.619	R&D	0.571	support	0.493	The 4th industrial revolution	0.518
manufacturing	0.579	production	0.476	production	0.469	production	0.469
support	0.546	government	0.461	business	0.423	support	0.468
production	0.538	business	0.435	domestic	0.334	business	0.396
innovation	0.495	manufacturing	0.415	government	0.331	AI	0.359
manufacturing innovation 3.0	0.488	cooperation	0.399	investment	0.329	IoT	0.357
strategy	0.443	investment	0.395	global	0.328	government	0.352
competitiveness	0.430	creative economy	0.356	market	0.327	domestic	0.329
creative economy	0.420	growth	0.355	competitiveness	0.327	growth	0.326
investment	0.406	manufacturing innovation 3.0	0.352	IoT	0.324	competitiveness	0.316
SME	0.319	SME	0.349	factory	0.313	investment	0.316

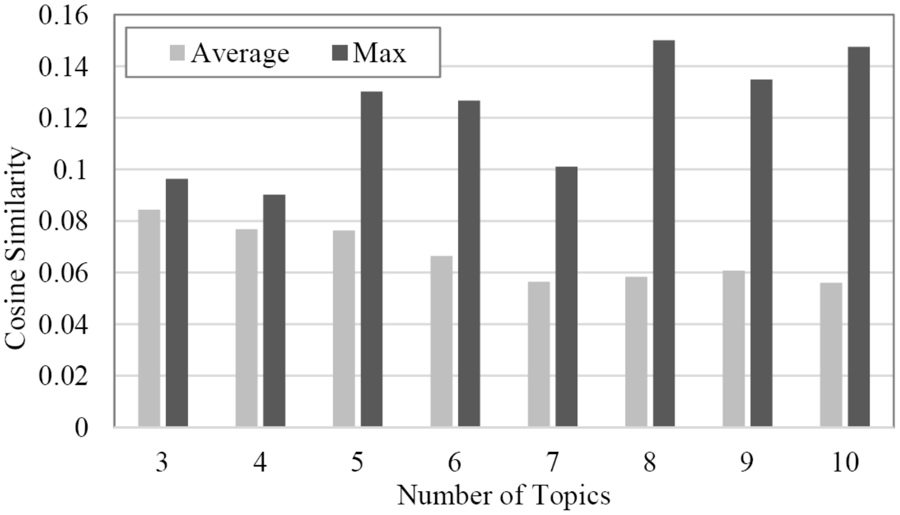

Figure 1. Cosine similarity by the number of topics (K) (LDA parameter: alpha = 0.01, eta = 0.001)

2016, there are growing interests in small and medium-sized enterprise (SME) support and the relationship between big company and SMEs. In 2017, we can see many high association rules related to the 4th industrial revolution.

4.2. **Results of topic modeling.** Before conducting a specific research, we created models by changing the number of topics, K from 3 to 10, and found cosine similarity value between them to determine K. Figure 1 shows average and maximum values of cosine similarity among the documents belonging to each topic by changing K.

Because the values of average and maximum similarity seem relatively low when K is 7, we conducted a detailed analysis about seven topics. Major keywords of each topic are shown in Table 4. We determined the name of seven topics to 'ICT or solution market',

Table 3. High association rules (**S**: support, **C**: confidence, **L**: lift) by year

2014					***2015***				
X	**Y**	**S**	**C**	**L**	**X**	**Y**	**S**	**C**	**L**
contents	business	0.114	0.892	5.127	big data	IoT	0.129	0.822	2.525
contents	field	0.113	0.890	3.695	venture	creative economy	0.152	0.872	2.449
contents	future	0.113	0.890	3.407	launch	creative economy	0.105	0.734	2.063
business	future	0.133	0.764	2.962	policy	government	0.154	0.777	1.683
industrial estate	creative economy	0.127	0.837	1.994	process	production	0.107	0.768	1.683
launch	creative economy	0.104	0.823	1.961	start-up	support	0.132	0.920	1.574
convergent	manufacturing innovation 3.0	0.114	0.840	1.723	customized	production	0.107	0.705	1.482
customized	support	0.138	0.873	1.600	idea	support	0.101	0.862	1.475
high tech	production	0.126	0.836	1.554	creative economy	support	0.300	0.841	1.440
2016					***2017 (∼April)***				
X	**Y**	**S**	**C**	**L**	**X**	**Y**	**S**	**C**	**L**
process	production	0.140	0.852	1.818	big company	SME	0.105	0.730	2.638
creative economy	support	0.241	0.771	1.644	regulation	government	0.103	0.754	2.143
Dept of industry	support	0.130	0.782	1.586	process	production	0.127	0.706	1.506
automation	production	0.119	0.716	1.527	automation	production	0.123	0.783	1.670
SME	support	0.206	0.744	1.509	SME	support	0.216	0.779	1.666
platform	R&D	0.134	0.715	1.305	response	the 4th industrial revolution	0.118	0.820	1.584
automobile	R&D	0.117	0.707	1.290	country	the 4th industrial revolution	0.113	0.808	1.561
convergence	technology	0.127	0.858	1.287	big company	support	0.104	0.724	1.548
performance	business	0.119	0.779	1.301	job	the 4th industrial revolution	0.134	0.8	1.545

'Commercialization policy', 'Economic or labor policy', 'Manufacturing IT convergence', 'Manufacturing innovation', 'Production automation', and 'Global business' which are selected by the algorithm.

We represented the ratio of reported news for each of the seven topics from semi-annual to 2014-2017 in Figure 2. The ratios of articles about 'Commercialization policy', and 'Economic or labor policy' have been decreased. This seems to be due to the policy change at the end of the regime. The increasing trend of the first policy topic by 2015 seems to be due to the establishment of 'creative economy innovation centers'. The ratios of 'Manufacturing IT convergence' and 'Production automation' topics continue to

TABLE 4. Major keywords of seven topics

Topics	***Major keywords***
ICT or solution market	IoT, AI, data, big data, 5G, cloud, security, service, analysis, digital, solution, platform, smart, connection, network, communication, real time, smart city, Intel, smartphone, business.
Commercialization policy	creative economy, support, venture, Samsung, start-up, president Park, launch, innovation center, smart logistics, Ministry of future, Gwangju city, Gumi city, Incheon city, Daegu city.
Economic or labor policy	regulation, reform, citizen, president Park, creative economy, investment, export, economic innovation, Chamber of commerce and industry, economy, new industry, labor market, job.
Manufacturing IT convergence	The 4th industrial revolution, industry 4.0, AI, Korea, manufacturing business, 3D printing, robot, U.S.A, digital, software, Japan, smart manufacturing, IoT, convergence, future.
Manufacturing innovation	SME support, co-growth, cooperation, Ministry of industry, big company, industrial revolution, industrial estate, production, supply, competitiveness, manufacturing innovation 3.0.
Production automation	POSCO, production, factory, automation data, machine tools, process, solution, efficiency, manufacturing, facility, product, smart, real time, mold, Siemens, quality, work.
Global business	China, semiconductor, two countries, growth, cooperation, Japan, global, Samsung, U.S.A, investment, market, sales, Korea, logistics, business, world, merger and acquisitions, Europe.

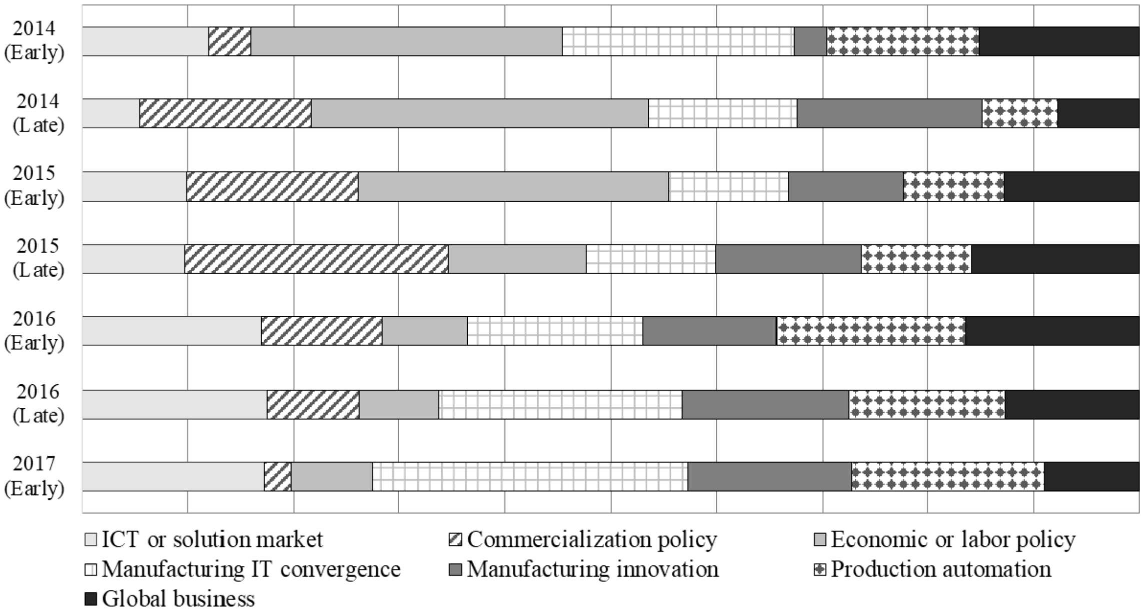

FIGURE 2. Changes in the ratio of online news articles on seven topics by half a year

increase from the end of 2014. This seems to be since many smart factory related IT and automation technologies have been realized, and the business interests are increasing.

5. **Conclusions.** To be competitive in the manufacturing industry, many companies and organizations are investing a lot of money and effort in building smart factory. Additionally, many people are interested in these changes recently, and the changes are covered

in news articles. This study analyzed keywords and topics of Korean online news articles from 2014 to April 2017 related to smart factory and revealed what kind of trends have been made about smart factory so far. We used ARM method for keyword analysis, and LDA method for topic analysis. By analyzing keywords and potential topics in the online news articles, we were able to grasp business and policy trends related to smart factory.

In keyword analysis, we can see that the occurrence of terms related to government policy has been reduced and the frequency of technology-related terms has increased. And it can be seen that the direction of smart manufacturing is centered on SMEs, because interests in support for SMEs and relationships between big company and SMEs are increasing with regard to smart factory from 2016. In topic modeling, we have seen a continued decline in the coverage rate of policy-related topics and an increase in coverage of topics related to manufacturing IT convergence and automation. At the same time, technologies such as 3D printing, big data and AI are increasingly emphasized. These trends show that smart factories are being implemented and the interest in related technologies is expanding.

It is necessary to complement the limitations of this study through validation of the subjective interpretation of the topics. In addition, comparative studies with media in developed countries, especially in the manufacturing industry, such as Germany and the United States of America, will be necessary.

Acknowledgement. This work was supported by Kyonggi University Research Grant 2017.

REFERENCES

[1] J. Lee and J. Park, Future direction of classification on smart factory related industry, *Proceedings on 2016 Spring Conference of Korean Institute of Industrial Engineers*, pp.1493-1515, 2016.

[2] T. Chang and H. Yang, Latent semantic analysis of research papers on smart factory, *ICIC Express Letters*, vol.11, no.4, pp.899-904, 2017.

[3] B. Kang, M. Song and W. Jho, A study on opinion mining of newspaper texts based on topic modeling, *Journal of the Korean Society for Library and Information Science*, vol.47, no.4, pp.315-334, 2013.

[4] J. An, K. Ahn and M. Song, Text mining driven content analysis of Ebola on news media and scientific publications, *Journal of the Korean Society for Library and Information Science*, vol.50, no.2, pp.289-307, 2016.

[5] Y. M. Park, *Newspaper Analysis on Discourse upon Korea Unification in South Korean Society*, Master Thesis, Yonsei University, 2015.

[6] J. H. Kim, A. Mantrach, A. Jaimes and A. Oh, How to compete online for news audience: Modeling words that attract clicks, *Proc. of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp.1645-1654, 2016.

[7] D. M. Blei, A. Y. Ng and M. I. Jordan, Latent Dirichlet allocation, *Journal of Machine Learning Research*, vol.3, pp.993-1022, 2003.

[8] S. H. Park, *Analysis of the Farmers Newspaper Article Using Topic Modeling: Target on the Topics Reform of Nonghyup*, Master Thesis, Chonbuk National University, 2017.

[9] R. Agrawal and R. Srikant, Fast algorithms for mining association rules, *VLBD*, vol.1215, pp.487-499, 1994.

[10] T. L. Griffiths and M. Steyvers, Find scientific topics, *Proc. of the National Academy of Sciences*, vol.101, no.suppl 1, pp.5228-5235, 2004.

[11] Y. W. Teh, K. Kurihara and M. I. Jordan, Hierarchical Dirichlet processes, *Journal of the American Statistical Association*, vol.101, no.476, pp.1566-1581, 2006.

[12] M. Rosen-Zvi, T. Griffiths, M. Steyvers and P. Smyth, The author-topic model for authors and documents, *Proc. of the 20th Conference on Uncertainty in Artificial Intelligence*, pp.487-494, 2004.

[13] H. Anna, Similarity measures for text document clustering, *Proc. of the 6th New Zealand Computer Science Research Student Conference (NZCSRSC2008)*, Christchurch, New Zealand, 2008.