Research Paper Informatization Policy Vol. 27, No.2 (2020), pp.066-083
https://doi.org/10.22693/NIAIP.2020.27.2.066

Big Data Analysis of Busan Civil Affairs Using the LDA Topic Modeling Technique*

Park, Ju-Seop** · Lee, Sae-Mi***

Abstract Local issues that occur in cities typically garner great attention from the public. While local governments strive to resolve these issues, it is often difficult to effectively eliminate them all, which leads to complaints. In tackling these issues, it is imperative for local governments to use big data to identify the nature of complaints, and proactively provide solutions. This study applies the LDA topic modeling technique to research and analyze trends and patterns in complaints filed online. To this end, 9,625 cases of online complaints submitted to the city of Busan from 2015 to 2017 were analyzed, and 20 topics were identified. From these topics, key topics were singled out, and through analysis of quarterly weighting trends, four "hot" topics(Bus stops, Taxi drivers, Praises, and Administrative handling) and four "cold" topics(CCTV installation, Bus routes, Park facilities including parking, and Festivities issues) were highlighted. The study conducted big data analysis for the identification of trends and patterns in civil affairs and makes an academic impact by encouraging follow-up research. Moreover, the text mining technique used for complaint analysis can be used for other projects requiring big data processing.

Keywords : big data analysis, online filings, local administration, LDA topic modeling, text mining

LDA 토픽모델링 기법을 활용한 부산시 민원 빅데이터 분석*

박 주 섭** · 이 새 미***

요 약 시민들은 도시 내 발생되고 있는 지역문제에 대해 큰 관심을 가지고 있다. 지방정부는 이러한 지역문제들을 해결하기 위해 노력하고 있지만 시민들의 생활 불편을 줄여주기는 쉽지 않고 이로 인한 시민들의 불만은 민원으로 이어지고 있다. 이를 해소할 수 있는 대안으로 빅데이터 활용을 통해 민원의 특성을 파악하고, 시민들에게 선제적 편의성을 제공하기 위한 노력이 절실하다. 본 논문에서는 LDA 토픽모델링 기법을 활용하여 전자민원의 동향 분석에 관한 연구를 실시한다. 이를 위해 2015~2017년 9,625건의 부산시 전자민원을 대상으로 20개의 민원토픽을 추출하였다. 도출된 민원토픽을 통해 핵심민원을 파악하고, 분기별 비중 추이 분석을 통하여 4개의 Hot 민원(버스정차, 택시기사, 칭찬, 민원처리)과 4개의 Cold 민원(cctv설치, 버스노선, 공원주차장, 축제 불만)을 도출하였다. 본 연구는 민원동향을 파악하기 위해 빅데이터 분석 방법을 제시하였고, 후속 연구를 유발하였다는 학문적 기여도가 있다. 또한 민원분석을 위해 사용한 텍스트마이닝 기법은 빅데이터 처리가 필요한 다른 행정업무에도 활용될 수 있다.

주제어 : 빅데이터 분석, 전자민원, 지방행정, LDA 토픽모델링, 텍스트마이닝

Received Feb 18, 2020; Revised Mar 20, 2020; Accepted Apr 1, 2020

* This work was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-2018S1A3A2075240)

** First author, Full-time Researcher, Smart Governance Research Center, Dong-A University(juseop60@naver.com)
*** Corresponding author, Full-time Researcher, Smart Governance Research Center, Dong-A University(emailme6@naver.com)

Ⅰ. Introduction

Businesses and nations use big data in an effort to become more competitive, and this is also the case in civil affairs analysis. There are a growing number of cases where data are being analyzed and used for the improvement of administrative services and public safety(Yoon, 2013). In the case of Korea, the number of occasions where big data were mobilized at the government or public institution level surged from just 12 in 2013 to 447 by 2017. In terms of sectors, public administration led with 236 cases, mostly in the analysis of public complaints and opinions along with that of posts on the homepage(Ministry of the Interior and Safety, 2018).

Recent developments, such as enhanced public awareness of administrative matters, better logical thinking through higher education, and easier access to a wider variety of information through the use of smart devices mean that the previously light demand for public services is becoming more complex and varied (Park, 2016; Shin, 2009). Government agencies are trying hard to resolve such public complaints in a timely manner. For example, we can cite the introduction of service centers to facilitate communication between the public and agencies, interim progress reports on complaints, establish a mileage system to shorten the time taken to handle cases, and provide incentives to enhance services. However, there are limits to fully satisfying public demands through such means.

Thus, many municipalities resort to actively using big data(Yang, 2018; Kang, 2019). The district of Bupyeong in Incheon of Korea, for example, was selected as one of the "Best 7" for the second half of 2018 by the National Information Resources Service of the Ministry of the Interior and Safety for analysis of big data. In the district, staff use big data platforms to analyze data and indicate the locations on an online map where grievances have been raised to submit to relevant departments. Of note, the analysis is used to indicate road repair locations, parking and traffic issues and possible solutions, and illegal banners and rubbish dumping, along with possible solutions and improvements(Kim, 2018b). As can be seen above, big data analysis can be used to monitor changes in public needs, forecast administrative burden, and proactively provide solutions to the public, lowering the likelihood of additional complaints and raising the efficacy of public services.

Text mining is a representative technique in big data analysis, in which structured or unstructured data are subjected to natural language processing, with a view to extracting meaningful information. Stages of text mining consist of gathering of raw data, information processing where keywords are extracted, information extraction where an algorithm is used to identify keywords, and finally information analysis where keywords are prioritized(Park et al., 2014). Topic modeling is a method of text mining that employs a probability model to determine the presence of certain topics in each document and to ascertain and classify common topics, as well as to estimate the distribution of topics and dispersion of words per topic.

A key methodology in topic modelling is the Latent Dirichlet Allocation(LDA) technique. LDA is mainly used in Library Sciences, Sociology, Industrial Engineering, Technical Management, and other areas requiring pattern analysis and research(Kim & Jang, 2016; Lee, et al., 2017; Park, et al., 2017b; Seol, et al., 2018).

The objective of this study is to identify topics in civil affairs and subsequently analyze trends and patterns therein. For this purpose, online filings in Busan are empirically analyzed using LDA topic modelling. The thesis layout is as follows: online filings and LDA topic modeling are explained in section 2; results of pattern analysis and policy implication is offered in section 3; and section 4 presents the conclusion.

Ⅱ. Literature Review

1. Analysis of Online Filings Using Big Data

In the past, to file a complaint, a member of the public had to either visit the administrative office in person or use the telephone or postal service, unlike now where many complaints are submitted online via PC or mobile phone. Anything submitted via the internet is considered an online filing.

Examples of big data usage to research online filings can be seen both within Korea and abroad. In Korea, Won and Yoo(2016) collected online data in the city of Jinju over the past 10 years, identified categories, extracted location data, and used them to analyze patterns in spatial distribution and trends, as well as using the ARIMA model to forecast the next two years of complaint frequency. Son and Kim(2017) simplified the administrative process by collecting online data by region, using statistical analysis, classifying and extracting topics, categorizing them by association and department, and finally making the findings graphic. Cho(2016) in looking to effectively manage complaints related to official pricing in real estate, used text mining to analyze keyword frequency and social media to identify mutual relationships between keywords. Finally, Suh, et al.(2010) applied text and data mining techniques to forecast the occurrence of petitions on national newspaper websites.

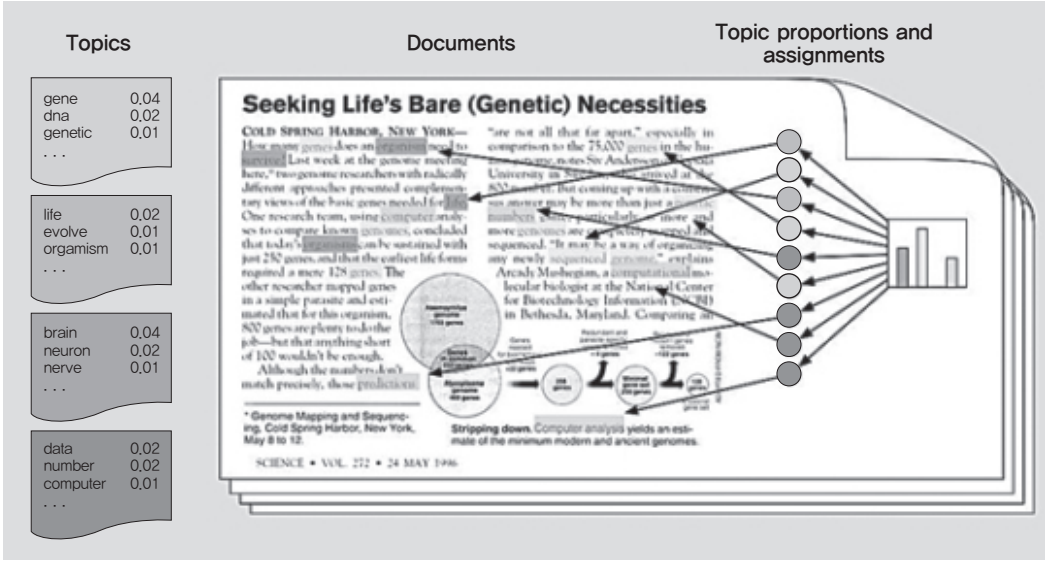
In international research, Evangelopoulos, et al.(2012) applied LSA(Latent Semantic Analysis) topic modelling on the 902 SMS messages sent to US President Barack Obama in 2009 by African citizens to identify and categorize key issues raised by Africans. Stylios, et al.(2010) employed text and data mining techniques to look into issues in government policy decisions, and in online communication of public opinion. Despite online complaints having become a popular means of political activism, Hagen(2018) emphasized the lack of effective ways to analyze them and came up with a framework for using LDA topic modelling on online filing data for training and verification.

2. Research Analysis Using Topic Modeling

Topic modelling is a method of identifying topics in complex documents by looking at certain word occurrences and grouping them by most representative subject(Alghamdi & Alfalqi,

2015; Lee, 2016; Yu, 2017). Depending on how the relationship between words is analyzed, it is called latent semantic analysis(LSA), probabilistic LSA(pLSA), or latent dirichlet allocation(LDA). While LSA and pLSA are prone to overfitting during the inference process, LDA is more useful for identifying topics relevant to actual data attributes(Park, et al., 2017a). LDA topic modelling is one type of probability statistic that can show the likelihood of a certain topic's appearance in each of the documents in a corpus, or collection of documents (Blei, 2012). A document can contain multiple topics, and each topic is expressed as a set of words. By applying time series analysis on the weighting of certain topics within each document, trend analysis can be performed. For example, if the document is submitted online, and the topic is subtopics, a classification of morphemes within the filings is carried out. Subsequently, once the number of topics to be extracted and basic parameters are set and analyzed, the proportion of each subtopic in the document and words within the subtopic is attained. Topics can be set by assigning those words with a high probability of occurrence, and by looking at patterns in online filings, it is possible to predict which topics are on the rise and which are on the wane.

〈Fig. 1〉 illustrates the basic concept of LDA topic modelling. LDA topic modelling shows the probability of a keyword appearing within



source : Probabilistic topic models(Blei, 2012)

〈그림 1〉 LDA 토픽모델링 개념

〈Fig. 1〉 Concept of LDA Topic Modelling

a topic. Looking at the probability of keywords in the third topic on the top left of Figure 1, we have brain 0.04, neuron 0.02, and nerve 0.01. We can guess that the topic has to do with "brain nervous system." To the extreme right of Figure 1, we see there are many words related to sticks, the middle stick in particular. Accordingly, the main subject of the document is likely to be the middle stick. The next relevant topic is the first stick on the left, followed by the third stick on the left. The weighting of a topic is expressed in decimals, by adding up the topics results in 1.

In looking at prior research using topic modeling, they can be grouped into trend analysis and research of specific fields using academic papers and patent abstracts(Abuhay, et al., 2018; Yang, et al., 2018; Kim & Chen, 2018), or into topic analysis based on texts(Park, et al., 2017c; Lee, et al., 2017; Mika, et al., 2018; Jacobi, et al, 2015).

In research and trend analysis, Yoon and Suh(2018) applied topic modeling and ego network analysis on 2,690 cases posted on Scopus to obtain research trends on smart healthcare. Park, et al.(2017b) carried out research on trends in science & technology by using LDA topic modeling on Artificial Intelligence(AI) abstracts in US patent papers.

In terms of topic analysis, Kim and Yun(2016) used topic modeling on reviews of hotel service in the Seoul area, extracting keywords and carrying out topic analysis by time period to identify perceptions of customer service and key issues. Yoon & Yoon(2017) used topic modeling on news articles to analyze trends in disaster and safety management.

3. Theoretical Framework

Topic modelling is a type of probability generative model that is widely used in computer sciences, with a particular focus on text mining and information retrieval (Liu, et al., 2016). Topic modeling evolved from Latent Semantic Analysis(LSA), presented by Deerwester, et al.(1990) to Hofmann's(2001) Probabilistic Latent Semantic Analysis(PLSA), which incorporated probability calculations, and eventually by Blei, et al.(2003) to LDA, a more complete probability generative model. Based on Bayesian statistics, LDA is a generative probabilistic model, and the LDA algorithm presumes that topic probability follows Dirichlet Distribution. The algorithm sees that a document contains a topic's probable distribution and calculates the probability of a word in the document belonging to a particular topic. In other words, it is an algorithm that collects specific topics in a document, analyzes the component words for probability, and extracts topic keywords from the results. The basic logic of the LDA algorithm is that each individual text carries a probability distribution towards a topic group, and that the author completes the document by selecting appropriate words according to the distribution.

To calculate probability distribution, Bayesian probability theory, which is widely used in machine learning, is applied. Topic modeling is based on the assumption that words are dependent on each other (Dirichlet Distribution), and post-probability is inferred according to word-generating conditions, which is expressed in the probability graph model as in. Accordingly,

on the basis of the probabilistic approach as a theoretical framework(Mannila, 2000), this study is carried out using LDA topic modeling based on Bayesian probability distribution.

Accordingly, the following questions have been posed for the research on trends and patterns in online filings.

- Research question 1: What are the main topics in online filings in Busan?
- Research question 2: Which topics account for the largest proportion of online filings in Busan?
- Research question 3: Which topics show a steady increase (hot) or decline (cold) in the number of filings?

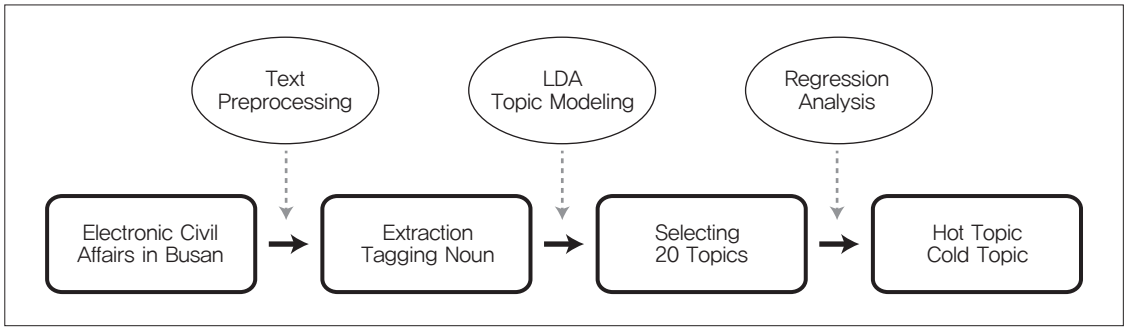
Ⅲ. Methodology

1. Overview

Topic modeling is a method of identifying document topics using keywords, with LDA as a representative technique. LDA topic modeling is used to determine how keywords are allocated to topics, and to model probability on how much of each document to include under a topic(Blei, 2012). LDA topic modeling is used to classify topics. To identify trends in civil affairs, 9,625 online filings over three years from 2015 to 2017 in Busan, Korea were taken and subjected to pre-processing. Through stemming, only nouns were extracted, and using LDA topic modeling 20 topics and theta values were derived. Using the theta values, topic weightings by quarterly period were calculated. Regression analysis on the data identified topic trends. 〈Fig. 2〉 shows the research process to analyze trends in online civil affairs on the basis of the literature review, as outlined in section Ⅱ.

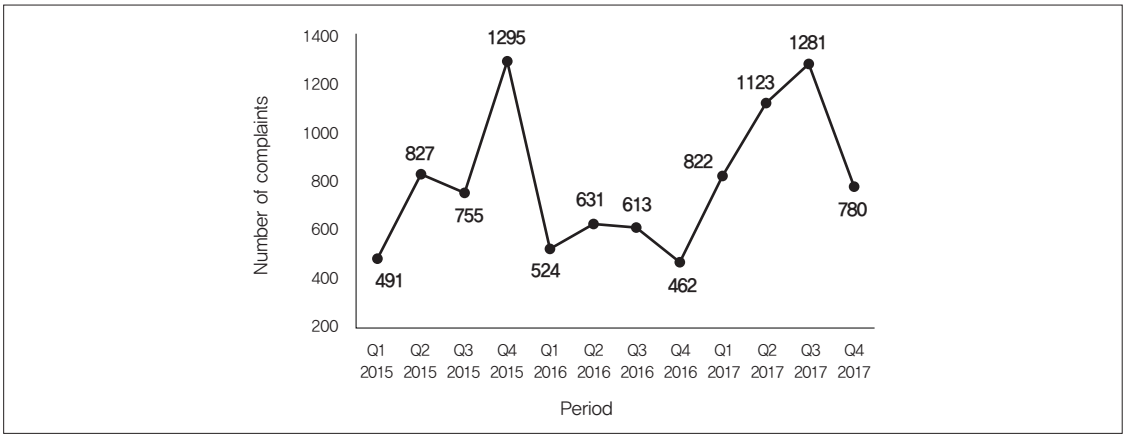
2. Data Collection and Pre-processing

Grouping by topic on a quarterly basis was carried out on three years of online filings in Busan, totaling 9,625 cases between 2015 and 2017. 〈Fig. 3〉 shows the number of online filings by quarter. For pre-processing, Komoran-a stemming tool for the Korean language-was used to extract nouns only and save them as one text file.



〈그림 2〉 전자민원 분석 절차

〈Fig. 2〉 Research Process on Online Filings



〈그림 3〉 부산시 전자민원 데이터 분기별 추이
〈Fig. 3〉 Online Filing Data by Quarter for Busan

3. Analysis of Trends in Online Filings

1) Sorting of Filings by Topic

To classify online filings by topic, the LDA topic modeling algorithm was applied to the pre-processed text file. We chose topic modeling as the analysis method because it is useful, as an unsupervised learning method, for identifying topics in a given document even without prior knowledge of the document's content(van der Meer, 2016). Futhermore, it shows the weightings of extracted topics for each document(DiMaggio, et al., 2013).

The validity of an unsupervised method is not as simple or as clear-cut as that of a supervised method(van der Meer, 2016). The internal and external validity of topic modeling is still being researched by multiple scholars (Chang, et al., 2009; DiMaggio, et al., 2013; Hagen, 2018; Jacobi, et al., 2015; Ramirez, et al., 2012).[1)]

External validity refers to the extent to which a given result can be generalized to a different time, environment or individual. To measure the external validity of LDA topic modeling, it is possible to make a comparison between the classification results on the same set of data using LDA topic modeling and those of subjects who have been restricted from exposure to LDA topic analysis (Hagen, 2018). However, this measure of validity remains under research, and it is still too early to discern an automated and objective method of measuring validity. LDA used for analysis in this study, is the most widely used topic modeling algorithm(Hu, et al., 2014; Shi, et al., 2016), and due to its outstanding generalization properties and scalability, it has generated great success in text mining(Cheng, et al., 2014).

1) To measure the internal validity of LDA topic modelling, a researcher must manually inspect the topic model and make the correct decision based on the words belonging to each topic, as well as the documents representing a particular topic(Jacobi et al., 2015).

At this stage, the researcher may decide the optimal number of topics and the number of recurrences for the most effective analysis(Song & Kim, 2013). For this study, the number of topics was set at 20, parameter α, β at the default value, and the number of iterations at 1,000. 〈Table 1〉 lists the 20 topics classified for online filings in Busan. Of the 20 topics, 6 relate to transportation, including [T1] Bus stops, [T4] Bus routes, [T5] Taxi drivers, [T10] Subway issues, [T12] Bus lanes, and [T19] Bus frequency. The 14 non-transportation topics include [T2] Contagious diseases, [T3] CCTV installation, [T6] Praises, [T7] Myeongji new town, [T8] Elderly welfare, [T9] Desalination, [T11] Public values, [T13] Administrative handling, [T14] New airport, [T15] Construction work, [T16] Illegal activities, [T17] Park facilities, [T18] New stay, and [T20] Festivities issues.

2) Weightings by Topic

〈Table 2〉 shows the proportional weightings of topics from 2015 to 2017, illustrating trends in online complaints in Busan. By topic, the biggest

〈표 1〉 부산시 전자민원 주제별 분류
〈Table 1〉 Classification by Topic

[T1] Bus stop	[T2] Contagious disease	[T3] CCTV installation	[T4] Bus route	[T5] Taxi drivers
Bus Bus stop Driver Passenger Stop	Child Hospital Patient MERS Open	Install safety CCTV Resolution Fire	Route Operated Section Bus Via	Taxi Driver Fare Credit card Phone
[T6] Praises	[T7] Myeongji new town	[T8] Elderly welfare	[T9] Desalination	[T10] Subway issues
Friendly Appreciation Staff Phone Praise	Myeongji Apartment New town International Gangseo-gu	Support Care Project Budget Elderly	Water Gijang Address Site Desalination	Card Disabled Usage Female Subway
[T11] Public awareness	[T12] Bus lanes	[T13] Administrative handling	[T14] New airport	[T15] Construction
Public Care Society Awareness Country	Lane Road Traffic Enforcement Traffic lane	Response Complaints Content Representative Handling	Airport Busan Public Market Evaluation	Works Noise Site Rubbish Housing
[T16] Illegal activities	[T17] Park facilities	[T18] New stay	[T19] Bus frequency	[T20] Festivities issues
Management Law Facilities Regulation Permit	Park Parking lot Bicycle Facility Management	New stay Apartment Resident Green space Grounds	Bus Time Frequency Usage Interval	Haeundae Tourism Event Festival Tourists

〈표 2〉 부산시 전자민원 주제별 비중도
〈Table 2〉 Weightings by Topic

Topics	Weight	Rank	Topics	Weight	Rank
T[1]	9.06%	1	T[11]	1.83%	20
T[2]	5.61%	6	T[12]	7.35%	3
T[3]	2.86%	19	T[13]	5.34%	8
T[4]	5.50%	7	T[14]	3.16%	18
T[5]	6.58%	4	T[15]	6.06%	5
T[6]	5.05%	9	T[16]	3.44%	17
T[7]	3.56%	16	T[17]	4.83%	10
T[8]	3.63%	15	T[18]	4.45%	13
T[9]	4.60%	12	T[19]	8.19%	2
T[10]	4.26%	14	T[20]	4.63%	11

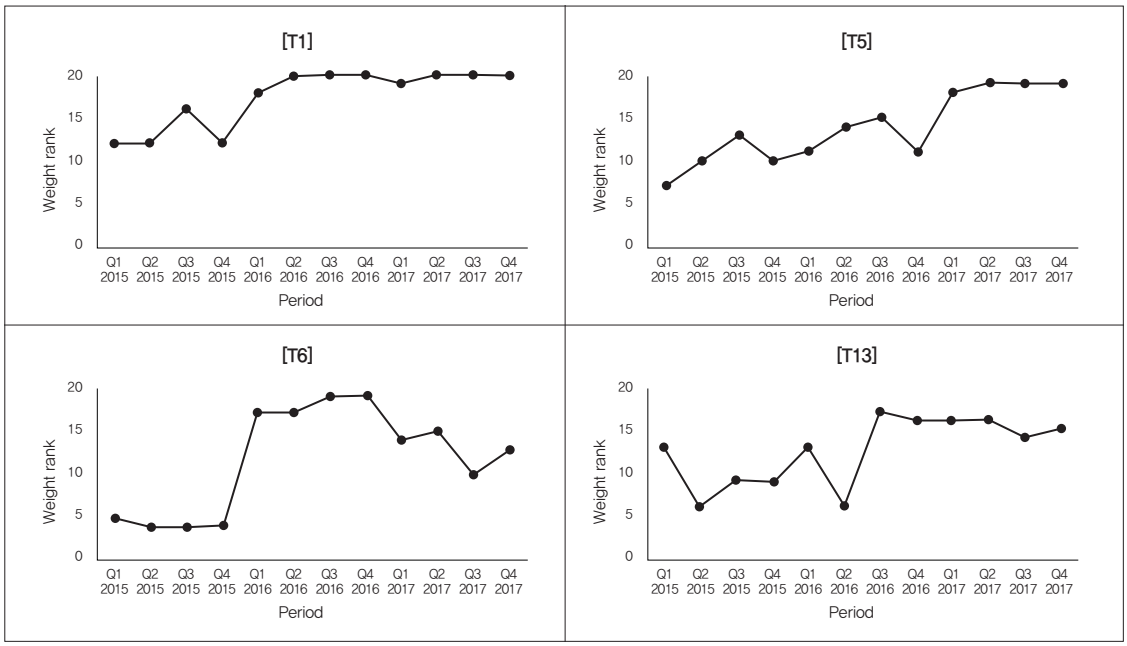
〈표 3〉 20개 민원토픽 회귀분석 결과
〈Table 3〉 Regression Analysis on 20 Topics

Topic	Regression coefficient	P-value	Hot/Cold	Topic	Regression coefficient	P-value	Hot/Cold
[T1]	0.830	0.001	Hot	[T11]	0.007	0.982	X
[T2]	-0.208	0.517	X	[T12]	-0.021	0.948	X
[T3]	-0.718	0.008	Cold	[T13]	0.642	0.024	Hot
[T4]	-0.718	0.009	Cold	[T14]	-0.326	0.301	X
[T5]	0.896	0.000	Hot	[T15]	-0.421	0.173	X
[T6]	0.614	0.034	Hot	[T16]	0.200	0.533	X
[T7]	-0.499	0.099	X	[T17]	-0.708	0.010	Cold
[T8]	0.524	0.081	X	[T18]	0.496	0.101	X
[T9]	-0.132	0.682	X	[T19]	-0.279	0.379	X
[T10]	-0.007	0.983	X	[T20]	-0.873	0.000	Cold

weighting is for [T1] Bus stop, followed by [T19] Bus frequency, [T12] Bus lanes, and [T5] Taxi drivers. This clearly shows that transportation-related topics account for the largest portion of online filings. On the other hand, [T11] Public awareness had the lowest weighting, along with [T3] CCTV installation, [T14] New airport, and [T7] Myeongji new town.

3) Grouping into Hot/Cold

After sorting Busan's online filings by quarter, each individual by quarter, each individual topic's weightings was analyzed. By comparing weightings to previous quarters, "hot" topics which show rising trends and "cold" topics which show a decline were identified(Kim, et al., 2017a, 2017b). In other words, the distinction between

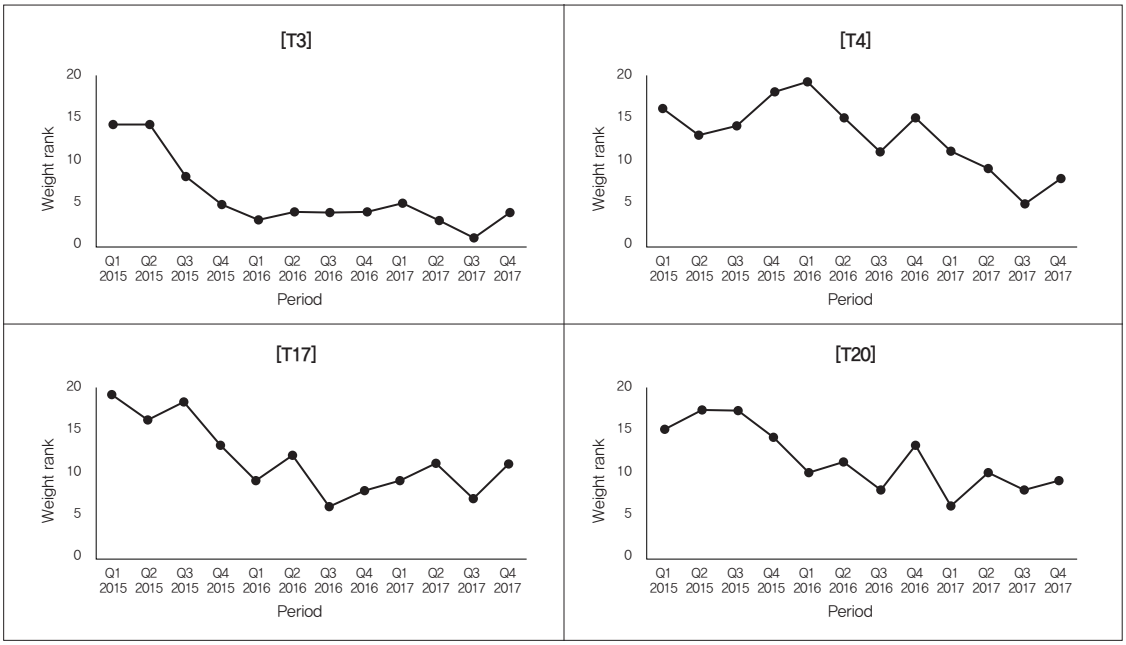


〈그림 4〉 핫 민원 토픽
〈Fig. 4〉 Hot Topic

hot and cold topics in this study is to see if the online filings of a particular topic increase or decrease over time. The regression coefficient from regression analysis was used to determine the trend for each topic. Simple regression analysis was performed with easily analyzed quarters as the independent variable and each topic's weighting as the dependent variable. With 0.05 as the significance level for regression analysis, a positive value denoted a hot topic, while a negative value was classified as cold.

This resulted in four hot and four cold topics. In 〈Table 3〉, an "X" under hot/cold means that the topic's probability of significance (p) exceeded the level of significance during regression analysis, rendering it meaningless. The four hot topics are: [T1] Bus stop, [T5] Taxi drivers, [T6] Praises, and [T13] Complaints handling(〈Fig. 5〉). The four cold topics are: [T3] CCTV installation, [T4] Bus routes, [T17] Park facilities(parking), and [T20] Festivities issues (〈Fig. 6〉). The others, including [T2] Contagious diseases, [T7] Myeongji new town, [T8] Elderly welfare, [T9] Desalination, [T10] Subway issues, [T11] Public awareness, [T12] Bus lanes, [T14] New airport, [T15] Construction, [T16] Illegal activities, [T18] New stay, and [T19] Bus frequency, were difficult to determine.

Bus stop, which is a hot topic, consisted of issues such as buses not coming to a complete stop, taking off while disembarking(leading to the risk of injury or accident), or failure to stop at designated stops. Taxi driver issues included denial of service, overcharging, non-acceptance



〈그림 5〉 콜드 민원 토픽
〈Fig. 5〉 Cold Topic

of credit cards, and unfriendliness. Praises(hot) are compliments to specific civil servants on their work, and Administrative handling(hot) is related to complaints regarding administrative processing. From this, it appears that Busan should educate bus and taxi drivers on customer service and make efforts to lower complaints over transport, while at the same time enforce stricter penalties for violations.

Ⅳ. Discussions and Policy Implications

It is impossible to read and analyze hundreds, even thousands of text documents, not to mention that this approach is extremely inefficient in terms of cost. To overcome this inefficiency, a text mining technique, one of the methods used in big data analysis, was used to easily identify topics from a large quantity of text data. Additionally, by using topic modeling, a form of unsupervised machine learning, the main keywords can be extracted on the basis of the co-occurrence of terms within a document, and topics can be identified, followed by time-series analysis for quick and accurate understanding of the overall characteristics of online filings as submitted by the public. Analyzing the trends and patterns of online filings can bring beneficial value to government policy makers, civil service administrators, and all relevant personnel including members of the public.

The following findings correspond to our three research questions to identify trends in online filings in Busan.

Firstly, 20 topics[2] have been identified by applying the LDA topic modeling technique to online filings in Busan.

Secondly, the topics were grouped into transportation-related and non-transportation related. Transportation-related consisted of six topics, including Bus stop, Bus routes, and Taxi drivers, while non-transportation related included 14 topics, such as Contagious diseases, CCTV installation, and Praises. Transportation issues occupied a relatively large share of weightings; Bus stop was highest, followed by Bus frequencies, Bus lanes, and Taxi drivers.

Thirdly, through regression analysis, four hot topics: Bus stops, Taxi drivers, Praises, and Administrative handling were identified. Complaints related to public bus services include rude drivers, failure to stop at designated bus stops, non-compliance with bus schedules, denied boarding, taking off while boarding or disembarking, and unsafe driving. Bus stop is a hot topic and addresses the issue of possible injury or accident due to sudden movements or failure to stop. A manual listing common complaints and instructions for addressing them should be produced and used to improve safety and service training for drivers. For drivers to naturally improve their driving skills, incidents of sudden acceleration and stopping should be digitally logged and used in evaluating service. We also suggest following the example of Daegu and providing customized education for drivers who repeatedly engage in dangerous driving (Kim, 2018a).

Complaints against Taxi drivers included unfriendliness, denied boarding, overcharging, non-acceptance of credit cards, unscheduled stops, and dangerous driving. Taxi drivers acting rudely towards customers is a longstanding and chronic complaint. Deteriorating working conditions due to a decline in passengers and excess capacity has naturally led to rudeness towards customers. The reality is that the perception of rudeness is subjective, making it difficult to establish a set of criteria for judgment and sanctions. This makes it necessary for the taxi industry itself to come up with a "self-cleansing" plan. It may be necessary to benchmark Seoul taxi drivers and their voluntary campaign to refund fares in case of rudeness. According to the above, should a passenger contact the taxi company with a complaint and identify the unfriendly driver along with an explanation, industry guidelines state part or all of the fare will be refunded. Should it be determined that the driver was rude, the driver in question is not only required to come up with the refund amount, but is also required to attend courtesy training and other ad hoc education sessions (Kwak, 2016). Focused oversight of the taxi industry by Seoul City following the introduction of a refund policy led to a reduction in the number of complaints, where taxi companies previously accounted for up to 65% of total taxi complaints. Seoul City has announced that by

2) Bus stop, Contagious disease, CCTV installation, Bus route, Taxi drivers, Praises, Myeonji new town, Elderly welfare, Desalination, Subway issues, Public awareness, Bus lanes, Administrative handling, New airport, Construction, Illegal activities, Park facilities, New stay, Bus frequency, Festivities issues

the first half of 2017, the number of taxi-related complaints fell by 33.5% from the first half of 2014, and in the case of taxi companies, by 40.1%(Kim, 2017). The hot topics Praises and Administrative service show two sides of the same coin. Praise for those staff who provide satisfactory service to the public and training for those who are not courteous are the parallel required strategies. Requiring consideration is a systematic approach including integration of platforms to reduce the number of complaints on administrative handling. In Britain, 2,000 government departments and public institutions are integrated under one website(with a single domain) and provide unified service and policy information from a single government portal(Kim, et al., 2015).

Lastly, the four cold topics consist of [T3] CCTV installation, [T4] Bus routes, [T17] Parking facilities, and [T20] Festivities issues. CCTV installation is related to improving resolution on CCTVs in new towns. Normally, CCTV is installed to prevent fires and for safety, including crime prevention. A CCTV consists of a camera and a DVR(Digital Video Recorder) for storing images from the camera. DVR prices can range from just a few thousand won to several hundred thousand won, often triggering disagreement between those supplying the CCTV and those using it. Particularly in the case of new apartments in new towns, the minimum required specification should be written into law, thereby lowering complaint incidents. This is a recurring topic, and as can be seen in Figure 6, the number of complaints appear early on during a period, only declining after half the period is over. Another cold topic is related to complaints on bus routes. Busan must continuously monitor route efficacy and make efforts to improve routes which face many complaints or are used inefficiently. Fortunately, this complaint shows a rising trend until the middle of the period but gradually declines from then on. Park facilities is related to complaints on the usage of facilities in the park, including the gymnasium and indoor soccer court. What is required here is to gather data on the overall situation regarding public sports facilities, and with continuous monitoring build a strategy to manage facilities in keeping with changes in their life cycles(Kim, 2016). Festivities issues is a cold topic and is related to complaints regarding water sports and the drone festival. The two cold topics of Park facilities and Festivities, as can be seen in Figure 6, show a rise at the beginning and a subsequent gradual decline, implying a decrease in public complaints.

V. Conclusions

With advances in communication, filing of public complaints can take many forms, irrespective of time and space. Popularization of social media, in particular, has necessitated the era of big data. Committed effort to reduce the number of complaints is required by applying big data analysis to raw data, and by grouping and identifying the nature of complaints.

The government has recently introduced an electronic system for dealing with public concerns and is receiving filings online in real time. However, processing this huge volume of

filings and dealing with actual public concerns is very costly and time-consuming. Currently, the office responsible for dealing with public complaints is engaged in directing concerns to relevant departments. The officer in charge reads individual complaints, decides which department is best suited to deal with them, and directs the complaints accordingly. Processing the filings takes a lot of time, which makes the system inefficient. In an attempt to solve this issue and increase productivity, a study was conducted applying text mining techniques to the process for online filings for Busan, in order to identify trends and patterns.

The current study has several applications. Firstly, by identifying the patterns and trends of public concerns, it is possible to anticipate complaints that arise regularly with regards to both time and location, allowing for a more effective response. For example, illegal parking, unlawful banners, damaged roadworks and fly tipping occur on a regular basis, and by identifying and analyzing when and where these incidents tend to happen, they can be dealt with pre-emptively. Secondly, the current study can be used as a basis for preparing a manual addressing frequently occurring issues, which will improve the quality of public services provided. Thirdly, the use of analysis through methods presented in the study prevents any subjective intervention by the public administrator, allowing for the objective analysis of online filing data, which can reduce time spent on categorizing types of public concerns.

The objective of this study is to analyze patterns and trends in online filings using text mining, a big data analysis technique. Public citizens worldwide have grievances on various issues including education, housing, transport, and welfare, and these complaints can be submitted through various platforms, including personal visits, written correspondence, telephone calls, or online via government websites. With the recent proliferation of Internet usage and the more active use of online spaces for voicing opinions, the filing of grievances online is increasing compared to other channels. Despite this increase, there is a lack of research on effective analysis of text data-based online filings, which is a form of unstructured data(Hagen, et al., 2016; Mergel, et al., 2016). The text mining technique used in this research can be applied to analyze the grievances of citizens from any part of the world, and through its application to identify topics, we can expect more effective processing of complaints and an increase in public satisfaction.

Moreover, through the use of LDA topic modeling, along with time-series regression analysis, we can pre-emptively deal with seasonal or annual complaints, leading to more effective resolution of public inconveniences. Also, as follow-up research, the text mining technique used in this study can be applied to automatically sort complaints and allocate them to relevant departments in the instance they are filed.

Academic contribution and practical usefulness of the study are as follows:

Firstly, trends and patterns in civil affairs in Busan were identified through literature review and theoretical basis for policy agenda established. The result of analysis on civil affairs can, in practice, be used by local governments in

setting actual policy agenda.

Secondly, raw data were processed into quantifiable data for the purpose of analyzing trends and patterns in civil affairs and setting a platform for forecasting. By identifying trends and patterns in civil affairs using big data, a platform is established whereby timely and cost-effective analysis is made possible.

Thirdly, big data analysis was used to identify trends and patterns in civil affairs and to set the stage for follow-up research. The text mining technique used in the analysis can be applied to other administrative tasks requiring big data processing.

Limitations of the study and future challenges are as follows:

Firstly, for effective analysis of trends and for forecasting, further study is required on readily converting raw data into quantifiable data. Secondly, although the study is limited to analyzing data from Busan only, further study is required whereby policy agenda can be identified using a wide range of data available online.

■ References

Abuhay, T. M., Nigatie, Y. G. & Kovalchuk, S. V. (2018). "Towards Predicting Trend of Scientific Research Topics Using Topic Modeling." *Procedia Computer Science, 136*, 304-310.

Alghamdi, R. & Alfalqi, K. A. (2015). "Survey of Topic Modeling in Text Mining." *International Journal of Advanced Computer Science and Application, s6*(1), 147-153.

Blei, D. M. (2012). "Probabilistic Topic Models." *Communications of the ACM, 55*(4), 77-84.

Blei, D. M., Ng, A. Y. & Jordan, M. (2003). "Latent Dirichlet Allocation." *Journal of machine Learning research, 3*(Jan), 993-1022.

Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J. L. & Blei, D. M. (2009). "Reading Tea Leaves: How Humans Interpret Topic Models." *Advances in neural information processing systems, 22*, 288-296.

Cheng, X., Yan, X., Lan, Y. & Guo, J. (2014). "BTM: Topic Modeling Over Short Texts." *IEEE Transactions on Knowledge and Data Engineering, 26*(12), 2928-2941.

Cho, T. I. (2016). "Spatiotemporal Characteristics Analysis of Complaints on Officially Assessed Land Price by Big Data Mining." Doctoral Thesis, Department of Civil and Environmental Engineering, Incheon University.

{조태인 (2016). 〈빅데이터 마이닝에 의한 공시지가 민원의 시공간적 특성 분석〉. 인천대학교 일반대학원 박사 학위논문.}

Deerwester, S., Dumais, S., Landauer, T., Furnas, G. & Harshman, R. (1990). "Indexing by Latent Semantic Analysis." *Journal of the American Society for Information Science, 41*(6), 391-407.

DiMaggio, P., Nag, M. & Blei, D. (2013). "Exploiting Affinities Between Topic Modeling and the Sociological Perspective on Culture: Application to Newspaper Coverage of U.S. Government Arts Funding." *Poetics, 41*(6), 570-606.

Evangelopoulos, N. & Visinescu, L. (2012). "Text-mining the voice of the people." *Communications of the ACM, 55*(2), 62-69.

Hagen, L. (2018). "Content Analysis of E-petitions with Topic Modeling: How to Train and Evaluate LDA Models?" *Information Processing & Management, 54*(6), 1292-1307.

Hagen, L., Harrison, T. M., Uzuner, Ö., May, W., Fake, T. & Katragadda, S. E. (2016). "Petition Popularity: Do Linguistic and Semantic Factors Matter?" *Government Information Quarterly, 33*(4), 783-795.

Hofmann, T. (2001). "Unsupervised Learning by Probabilistic Latent Semantic Analysis." *Machine Learning, 42*(1-2), 177-196.

Hu, Y., Boyd-Graber, J., Satinoff, B. & Smith, A. (2014). "Interactive Topic Modeling." *Machine Learning, 95*(3), 423-469.

Jacobi, C., Atteveldt, W. V. & Welbers, K. (2015). "Quantitative Analysis of Large Amounts of Journalistic Texts Using Topic Modelling." *Digital Journalism, 4*(1), 89-106.

Jang, B. M. (2015). "Analysis of Public Big Data for Promoting Benefits of Community Residents." Master's Thesis. Kyungpook National University.

{장병문 (2016). 〈공공부문 빅데이터를 이용한 민원편익 분석〉. 경북대학교 산업대학원 박사학위논문.}

Kang, K. J. (2019). "Uijeongbu City, Big Data Analysis Project Completion Report Meeting Held." *The Financial News*. January 21.

{강근주 (2019). "'빅데이터 활용' 의정부시 부서 간 협업체계 구축." 〈파이낸셜뉴스〉. 1월 21일.}

Kim, G. & Yun, H. (2016). "Topic Modeling Approach to Understand Changes in Customer Perceptions on Hotel Services in Seoul." *Journal of Korea Service Management Society, 17*(3), 217-231.

{김건·윤혜정 (2016). 토픽모델링을 활용한 서울지역 호텔서비스에 대한 고객인식의 변화 분석. 〈서비스경영학회지〉, 17권 3호, 217-231.}

Kim, H. W. (2017). "Seoul City, Unfavorable Rate Refund, Civil Service Regulation, 40% Reduction in Corporate Taxi Complaints." *Dongyang News Agency*, August 13.

{김혁원 (2017). "서울시, 불친절요금환불·민원총량제... 법인택시 민원 40%감축." 〈동양뉴스통신〉. 8월 18일.}

Korea Data Agency. (2017). *2017 Data Industry White Paper*. Seoul: Korea Data Agency.

{한국데이터산업진흥원 (2017). 〈2017 데이터 산업 백서〉. 서울: 한국데이터산업진흥원.}

Kim, C. S., Choi, S. J. & Kwahk, K. Y. (2017a). "Investigation of Research Trends in Information Systems Domain Using Topic Modeling and Time Series Regression Analysis." *Journal of Digital Contents Society, 18*(6), 25-39.

{김창식·최수정·곽기영 (2017a). 토픽모델링과 시계열회귀분석을 활용한 정보시스템분야 연구동향 분석. 〈한국디지털콘텐츠학회논문지〉, 18권 6호, 1143-1150.}2017a, b로 나누는 이유?

Kim, C. S., Kwahk, K. Y. & Yoon, H. J. (2017b). "An Analysis of Research Trends in Tourism Studies: Applying Topic Modeling and Time Series Regression Analysis." *Journal of Tourism and Leisure Research, 29*(12), 25-39.

{김창식·곽기영·윤혜진 (2017b). 관광분야 연구동향 분석: 토픽모델링과 시계열분석을 중심으로. 〈관광레저연구〉, 29권 12호, 25-39.}

Kim, J. H., & Chen, W. (2018). "Research Topic Analysis in Engineering Management Using a Latent Dirichlet Allocation Model." *Journal of Industrial Integration and Management, 3*(4), 1850016.

Kim, K. W. (2018a). "Daegu City Bus Passenger's Biggest Complaint is 'Unkind Bus Driver'." *Maeil Shinmun*, October 30.

{김근우 (2018a). "'대구 시내버스 탑승객 최고 불만은 '불친절한 버스기사'." 〈매일신문〉. 10월 30일.}

Korea Institute of Sports Science (2016). *Improvement plan of public sports facility management*. Seoul: Korea Institute of Sports Science.

{한국스포츠정책과학원 (2016). 〈공공체육시설 관리운영 개선방안〉. 서울: 한국스포츠정책과학원.}

Kim, S. K. & Jang, S. Y. (2016). "A Study on the Research Trends in Domestic Industrial and Management Engineering Using Topic Modeling." *Journal of the Korea Management Engineers Society, 21*(3), 71-95.

{김상겸·장성용 (2016). 토픽모델링을 이용한 국내 산업경영공학 연구동향 분석. 〈한국경영공학회지〉, 21권 3호, 71-95.}

Kim, Y. H. (2018b). "Incheon Bupyeong-gu Civil Big Data Analysis. 2nd Half Best 7." *Maeil Ilbo*, February 11.

{김양훈 (2018b). "인천 부평구 민원 빅데이터 분석, 2018 하반기 Best 7 선정." 〈매일일보〉. 2월 11일.}

National Information Society Agency. (2015). *Strategy for Building Administrative Service Integration Delivery Platform*. Seoul: National Information Society Agency.

{한국정보화진흥원 (2016). 〈행정서비스 통합제공 플랫폼

구축 전략〉. 서울: 한국정보화지능원.}
Kwak, J. O. (2016). "Unkind Taxi Driver." *The Transportation News Korea*, May 31.
{곽재옥 (2016). "'불친절 택시기사' 정해져 있다." 〈교통신문〉. 5월 31일.}
Lee, J. M., Lee, J. A. & Jeong, J. H. (2017). "The Jeonse Price Forecasting Used by News Big Data - Focusing on Topic Modeling Analysis." *Korea Real Estate Academy Review, 69*, 43-57.
{이종민·이종아·정준호 (2017). 뉴스 빅데이터를 이용한 전세가격 예측. 〈부동산학보〉, 69권, 43-57.}
Lee, S. S. (2016). "A Study on the Application of Topic Modeling for the Book Report Text." *Journal of Korean Library and Information Science Society, 47*(4), 1-18.
{이수상 (2016). 독후감 텍스트의 토픽모델링 적용에 관한 탐색적 연구. 〈한국도서관·정보학회지〉, 47권 4호, 1-18.}
Liu, L., Tang, L., Dong, W., Yao, S. & Zhou, W. (2016). "An Overview of Topic Modeling and Its Current Applications in Bioinformatics." *Springerplus, 5*(1), 1608.
Mannila, H. (2000). "Theoretical Frameworks for Data Mining." *ACM SIGKDD Explorations Newsletter, 1*(2), 30-32.
Mergel, I., Rethemeyer, R. K. & Isett, K. (2016). "Big data in public affairs." *Public Administration Review, 76*(6), 928-937.
Mika, W., Seppo, L. & Mervi, R. (2018). "A Topic Modelling Analysis of Living Labs Research." *Technology Innovation Management Review, 8*(7), 40-51.
Ministry of the Interior and Safety. (2018). *Good Use of Big Data in Civil, Tourism and National Safety*. Sejong: Ministry of the Interior and Safety.
Na, Y. W., Park, H. J. & Jung, J. W. (2015). "Pattern analysis of environment complaint using the spatial big data." *Journal of the Korean Society of Civil Engineers, 63*(7), 29-35.
{나영우·박훈진·정진우 (2015). 공간 빅데이터를 활용한 환경민원 패턴분석. 〈대한토목학회지〉, 63권 7호, 29-35.}
Park, D. S., Moon, Y. S., Park, Y. H., Yoon, C. H., Jeong, Y. S. & Jang, H. S. (2014). *Big data computing technology*. Seoul: Hanbit Academy, Inc.
{박두순·문양세·박영호·윤찬현·정영식·장형석 (2014). 〈빅데이터 컴퓨팅 기술〉. 서울: 한빛아카데미.}
Park, H. J., Kim, H. N. & Hong, Y. J. (2017a). "A Topic Modeling Analysis on the Major Social Issues of the Students' Human Rights Ordinance in Korea." *Asian Journal of Education, 18*(4), 683-711.
{박현정·김한나·홍유정 (2017). 토픽모델링을 활용한 학생인권조례의 사회적 이슈 분석. 〈아시아교육연구〉, 18권 4호, 683-711.}
Park, J. S., Hong, S. G. & Kim, J. W. (2017b). "A study on science technology trend and prediction using topic modeling." *Journal of the Korea Industrial Information Systems Research, 22*(4), 19-28.
{박주섭·홍순구·김종원 (2017). 토픽모델링을 활용한 과학기술동향 및 예측에 관한 연구. 〈한국산업정보학회논문지〉, 22권 4호, 19-28.}
Park, S. H., Moon, H. S. & Kim, J. K. (2017c). "Online reviews analysis for prediction of product ratings based on topic modeling." *Journal of Information Technology Services, 16*(3), 113-125.
{박상현·문현실·김재경 (2017). 토픽 모델링에 기반한 온라인 상품 평점 예측을 위한 온라인 사용 후기 분석. 〈한국IT서비스학회지〉, 16권 3호, 113-125.}
Park, W. D. (2016). "Improvement Plan for the Civil Affairs Administration Service based on the Level of Resident Satisfaction." Master's Thesis. Myongji University.
{박원동 (2016). 〈주민만족도에 의한 민원행정서비스의 개선방안: 용인시 처인구를 중심으로〉. 명지대학교 일반대학원 석사학위논문.}
Ramirez, E. H., Brena, R., Magatti, D., Stella, F. (2012). "Topic model validation." *Neurocomputing, 76*(1), 125-133.
Seol, D. H., Ko, J. H., & Yoo, S. H. (2018). "Korean Sociological Association and sociological research: Changes in the areas of sociology in Korea 1964-2017." *Korean Journal of Sociology,*

52(1), 153-213.

{설동훈·고재훈·유승환 (2018). 한국사회학회와 사회학연구, 1964-2017년: 한국사회학회 발표 논문의 연구분야별 내용분석. 〈한국사회학〉, 52권 1호, 153-213.}

Shi, Z., Lee, G. M., Whinston, A. B. (2016). "Toward a Better Measure of Business Proximity: Topic Modeling for Industry Intelligence." *MIS Quarterly, 40*(4), 1035-1056.

Shin, H. C. (2009). "Administrative Service Improvement Program of Inhabitants Evaluation." Master's Thesis. Kyungpook National University.

{신희철 (2009). 〈주민평가에 의한 민원행정서비스의 개선방안: 대구광역시 수성구 민원행정을 중심으로〉. 경북대학교 행정대학원 석사학위논문.}

Son, N. R. & Kim, S. Y. (2017). "Complaints Statistics and Department of Automated Classifications System through Public Complaints Big Data Analysis." *The Journal of Korean Institute of Next Generation Computing, 13*(1), 22-35.

{손남례·김서영 (2017). 공공민원 빅데이터 분석을 통한 민원통계 및 담당부서 자동분류 시스템. 〈한국차세대컴퓨팅학회논문지〉, 13권 1호, 22-35.}.

Song, M. & Kim, S. Y. (2013). "Detecting the Knowledge Structure of Bioinformatics by Mining Full-text Collections." *Scientometrics, 96*(1), 183-201.

Stylios, G., Christodoulakis, D., Besharat, J., Vonitsanou, M. A., Kotrotsos, I., Koumpouri, A. & Stamou, S. (2010). "Public Opinion Mining for Governmental Decisions." Electronic *Journal of e-Government, 8*(2), 202-213.

Suh, J. H., Park, C. H. & Jeon, S. H. (2010). "Applying Text and Data Mining Techniques to Forecasting the Trend of Petitions Filed to E-People." *Expert Systems with Applications, 37*(10), 7255-7268.

van der Meer, T. G. (2016). "Automated Content Analysis and Crisis Communication Research." *Public Relations Review, 425*, 952-961.

Won, T. H. & Yoo, H. H. (2016). "Pattern Analysis for Civil Complaints of Local Governments Using a Text Mining." *Journal of the Korean Society of Surveying, Geodesy, Photogrammetry and Cartography, 34*(3), 319-327.

{원태홍·유환희 (2016). 텍스트마이닝에 의한 지자체 민원청구 패턴 분석. 〈한국측량학회지〉, 34권 3호, 319-327.}.

Yang, H. C. (2018). "Big Data Analysis on Gimpo City bus civil complaint, The Most Frequent Complaint is Nonstop Bus." *Kyeong Gi Ilbo*, October 11.

{양형찬 (2018). "김포시 버스민원 빅데이터 분석, 무정차 민원 가장 많아." 〈경기일보〉. 10월 11일.}

Yang, H. L., Chang, T. W. & Choi, Y. (2018). "Exploring the Research Trend of Smart Factory with Topic Modeling." *Sustainability, 10*(8), 2779.

Yoon, J. E. & Suh, C. J. (2018). "Research Trend Analysis on Smart Healthcare by Using Topic Modeling and Ego Network Analysis." *Journal of Digital Contents Society, 19*(5), 981-993.

{윤지은·서창진 (2018). 토픽모델링과 에고 네트워크 분석을 활용한 스마트 헬스케어 연구동향 분석. 〈디지털컨텐츠학회논문지〉, 19권 5호, 981-993.}

Yoon, M. Y. (2013). "Analysis of Major Data Promotion Strategies and Implications." *The Journal of Science and Technology Policy, 23*(3), 31-43.

{윤미영 (2013). 주요국의 빅데이터 추진전략 분석 및 시사점. 〈과학기술정책〉, 23권 3호, 31-43.}.

Yoon, S. Y, & Yoon, D. K. (2017). "A Trends Analysis on Disaster and Safety Management Using Topic Modeling." *Journal of Korean Society for Geospatial Information System, 25*(3), 75-85.

{윤소연·윤동근 (2017). 토픽모델링을 이용한 재난 및 안전관리 동향 분석. 〈대한공간정보학회지〉, 25권 3호, 75-85.}.

Yu, Y. L. (2017). "Analysis of Media Coverage on 2015 Revised Curriculum Policy using Big Data Analysis." Doctoral Thesis, Department of Education, Seoul National University.

{유예림 (2017). 〈빅데이터 분석 기법을 활용한 2015 개정 교육과정 정책에 대한 언론보도 분석〉. 서울대학교 일반대학원 박사학위논문.}