



ROAD SAFETY ISSUE EXPLORATION METHODOLOGY ON LIMITED SUBJECTS USING LDA TOPIC MODEL

Chang-Seok Oh

The Board of Audit and Inspection of Korea

112 Bukchonro, Jongno-gu Author address, Seoul, Korea

Phone: + 82-2-2011-3041 E-mail: csoh@bai.go.kr

Co-authors; Yongtaeck Lee, The Board of Audit and Inspection of Korea; Seung-Young Kho, Seoul National University

ABSTRACT

Study objective: The media represents the right of the people to know, and at the same time collects and forms public opinion about policy demands. For these reasons, media reports cannot be overlooked as a means of monitoring public opinion when it comes to policy formulation. Therefore, this study has developed a technique to search and evaluate policy issues based on quantification of media press releases on the topic of road safety.

Data sources: Data was consisted of media press releases searched in the news library of 'NAVER', Korea's leading internet portal. Crawling technique was employed to collect data.

Selection criteria: The search covered a time period of about 23 years, from 1990 to May of 2016. The analysis section was divided into three sections, the 3rd period referring to the last three years from May 2014 to May 2015, the 2nd period in which contains the inflection point for quantitative growth of civilian security related media reports from 2002 to 2013, and 1st period referring to the time prior to that.

Data screening and refinement: The initially collected data were subjected to certain screening and refining process. First, we created a limited keyword for the topic to be analyzed and collected press release data through crawling. In addition, we selected data that include both keywords and negative opinions based on one sentence of the press release.

Analysis method: Media data is comprised of unstructured information in text format. Therefore, in this study, topic analysis technique as a form of text mining was used to analyze unstructured data. Topic analysis is a method of clustering documents with similar themes by applying a text mining model to collected documents crawled based on search keywords. Through subject analysis, 'topic keywords' and 'topic contribution' are calculated, which can be used as quantitative indicators to help judge policy issues. The LDA model was used for the specific topic subject.

Results: In this study, we introduced the 'TF-IDF' index in order to examine the changes of keywords in the media reports. This index indicates how much and evenly the keyword was used in the collected press releases. The larger the value of keyword, the more frequently it has been mentioned in various media. Based on the TF-IDF index, the weighted words were classified as 'high speed', 'construction', 'safety', 'traffic' in the 1st period (1990-2002), 'high speed', 'safety', 'transportation', 'facilities', 'construction' in the 2nd period (2003-2013), and lastly, 'speed limit', 'speed bump', 'address', 'road signs' etc. in the 3rd period (2014-May 2016). Respectively, the safety of highway facilities has constantly



been raised as issues, and concerns on safety problems related to speed bumps and traffic signs are getting elevated more recently in the third period.

The policy issues related to the road traffic safety results, analyzed by the period via topic analysis on the press release, are as follows. The most important issues related to road traffic safety were related to the 'inconvenience to charging high pass fare (2001)' and 'inter-Korean railway and road diplomacy hardships (2002)' during the 1st period (1993-2002). In the 2nd period (2003-2013), concerns on the lack and unmeasured sizes of speed bumps, such as "no speeding brakes in front of elementary school, no braking brakes (2005)" and "speed bumps ignoring the standard measurements (2012)" have been raised. In the 3rd period (2014-May 2016), issues related to "dissatisfaction with road infrastructure in Seoul (2015)" and "risk of speeding bumps (2015)" were raised, which makes it noteworthy that the safety issue of the speed bump is keep mentioned as it did in the second period.

Keywords: Road safety policy issue analysis, Text mining, Topic model, LDA model, Issue Exploration

1. INTRODUCTION

How information related to transportation that is input in text format can be applied to transportation policies is becoming a focal point in various transportation modeling researches utilizing social media materials. Various atypical text materials such as social media are being utilized to analyze complexly interrelated traffic data such as predicting traffic volume, traffic demand surveys, etc. The following are some examples: Figuring out an individual's traveling behavior such as one's travelling purpose and route by collecting information on one's everyday life from the mobile phone (Farrahi et al., 2012), predicting a user's moving route by utilizing the GPS function of social networking services (Gao et al., 2012), evaluating the public transportation users' satisfaction level by using data from Twitter (Collins et al., 2013), deducing an individual's activity pattern by utilizing the user geographic information included in Twitter's check-in service (Hasan and Ukkusuri, 2014), predicting traffic volume during a future event by utilizing Twitter access information, hashtags, mentions, url links, etc. (Ni et al., 2014), estimating the TCS entry traffic volume and VDS speed by applying 'Google' and 'Naver' search traffic as input variables (Ryu et al., 2015), analyzing the correlation between the TCS traffic volume and 'Road Plus' (Webpage run by the Korea Expressway Corporation that provides traffic information) as well as highway traffic information applications (Ryu et al., 2016).

To this, this study utilizes text materials in the format of press releases to conduct a quantitative research on what kinds of issues the media has discussed until now in regards to road safety. The subway has consistently been brought up for various issues such as platform safety accidents for example the Guui Station incident, deterioration management of urban railway trains, abrupt stops of automated subways, lacking evacuee guidance facilities, etc.

2. LATENT DIRICHLET ALLOCATION ALGORITHM

The Latent Dirichlet Allocation (LDA) algorithm was proposed by Blei et al. (2003). Clustering algorithms that deal with structured data such as K-Means and artificial neural network's SOM follow a differentiation method that determines similarity according to coordinate distance and discovers centroid coordinates. In contrast, the LDA algorithm is what deals with text documents and figures out the documents' core topic and topic distribution within an arbitrary abstract space. Hasan and Ukkusuri (2014), based on the LDA algorithm, analyzed large-scale geographic location data through Twitter

access data to deduce Twitter users' individual activity patterns. Das et al. (2016) also applied the LDA Algorithm and analyzed the topics, keywords, and topic changes of journals presented in the TRB Annual Meeting from 2008 until 2014. Sun and Yin (2017) used the LDA model to deduce the changes in topic distribution over time on a total of 17,163 abstracts from 22 major traffic journals from 1990 until 2015.

The LDA algorithm has as its basic idea that documents have a potential topic and are expressed through random mixture (Blei et al., 2003). It is based on a bayesian mixture model that considers such potential topics as discrete data that are not interrelated (Das et al., 2016), and figure 1 shows this model shown in the format of a Bayesian Network Plate (Sun and Yin, 2017).

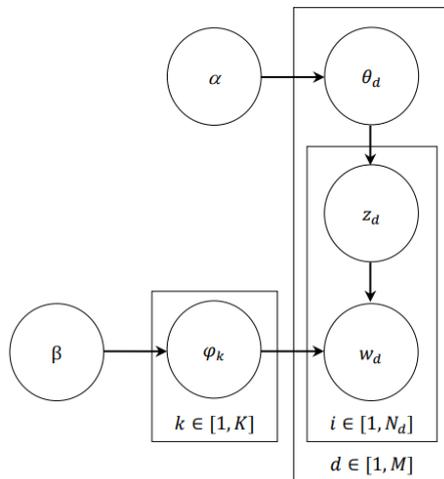


Figure 1. Bayesian Network Plate Representation of LDA (Source: Sun and Yin, 2017)

The principle of deducing a document subject in the LDA algorithm is as follows. When α is considered to be Dirichlet prior information regarding the topic distribution of each document, β Dirichlet prior information regarding word distribution of each topic, θ_d the topic distribution of document d , and the word distribution of topic k , LDA goes through a process of deducing θ_d and φ_k by observing the word ω from the documents when various documents have been observed already. When θ_d and φ_k are deduced from the words, the documents can be classified into topics. Here, $Z|\theta \sim \text{Multinomial}(\theta)$, $W|Z, \varphi \sim \text{Multinomial}(\beta)$ is assumed. The LDA model's joint probability distribution is $P(W, Z, \theta, \varphi; \alpha, \beta)$, and this distribution is a combination of a total of 4 sets of distributions: 2 sets of multinomial distributions and 2 sets of Dirichlet distributions. Meanwhile, α and β are values that were provided as prior information. Then, there is a marginalization process wherein all the effects of α and β are marginalized out. After this marginalization process, the $P(W, Z; \alpha, \beta) = \int_{\theta} \int_{\varphi} P(W, Z, \theta, \varphi; \alpha, \beta) d\varphi d\theta$ function can be deduced and Z can be found through W . This is when the Dirichlet distribution that is in a conjugated relationship with multinomial distributions is utilized to assume $\theta|\alpha \sim \text{Dirichlet}(\alpha)$, $\varphi|\beta \sim \text{Dirichlet}(\beta)$ in order to make integral calculation easier. The calculation of the possibility $P(W, Z, \alpha, \beta)$ that is derived from marginalization is modified to make Gibbs Sampling possible in order to deduce Z , whereafter θ and φ_k are deduced.



3. ANALYZED DATA AND ANALYSIS METHOD

3.1. Analyzed Data

3.1.1. Data Collection

The taxonomy for the topic analysis of media material regarding road safety is as shown in Table 1. This study considered all of the press release material shown on the ‘NAVER’ news portal when search keywords as shown in Table 1 were input for media material on the ‘NAVER’ portal. The time range for data collection was set to 23 years spanning from 1990 until May 2016. This time was divided into three intervals: ‘Interval 3’ as the most recent 3 years from 2014 until May 2016, ‘Interval 2’ as the time from 2002 until 2013 when there was a quantitative growth in the amount of press, and ‘Interval 1’ as any time before then.

The collection of media data for analysis goes through a process of ‘extraction → save → preliminary selection → final selection’. This research utilized the multistage sampling method for collecting data by using the keywords shown in Table 1 on daily newspapers provided on the ‘NAVER’ portal’s news service from January 1, 1990 until May 30, 2016. Multistage sampling is a method that samples after setting levels instead of directly sampling from the final sampling group when the cluster is divided into numerous upper and lower levels. This research set three levels and conducted sampling for each level. As such, when data is collected, the first stage of the upper level enables the researcher to collect data related to the analysis topic that is the focal point in the relative analysis, whereas the second and third stages of the lower levels enable the researcher to collect data related to the detailed content of the focal point in the relative analysis. Herein lies the advantage of this method because it satisfies both the sufficiency (upper level) and necessity (lower level) of related material.

Table 1. Taxonomy composition for topic analysis in this study

Hierarchy	Analysis Target	Search Keyword
class 1	Road traffic safety	Road safety facilities, traffic safety facilities, traffic safety regulations, road safety regulations, road traffic safety measures, traffic safety measures
class 1-1	Installation and management of road traffic safety facilities	Road traffic safety facility, road traffic safety facility installation, road traffic safety facility management, road facility, road safety facility
class 1-1-1	Speed bust	Speeding bikes for seniors, speeding bikes for disabled facilities, speeding bikes around schools
class 1-1-2	Road marking	Drawing display, road surface display, road management office order drawing display construction, drawing display, low cost paint, under specification paint
class 1-2	Management of traffic violation	Violation of traffic laws and ordinances, management of traffic violation, traffic violations
class 1-2-1	Management of traffic violation system	Administrative dismissal of violation of traffic laws and regulations, violation of traffic laws, penalties of driver and penalties, imposition of fines for vehicle owners, imposition of fines for unauthorized intervention
class 1-3	Manage the use of an aptitude test	Driver's license aptitude test, Driver's license aptitude test management, Driver's license aptitude test
class 1-3-1	Take advantage of the results of the aptitude test	Use of driver's license aptitude test judgment result, acquired physical disability driver's license occasional aptitude test
class 1-3-2	Dementia driver's license management	Manage dementia driver's license, manage driver's license for dementia patient



3.1.2. Data Processing

The data that is initially collected underwent a set selection and refinement process. First, data containing both keywords and negative meaning dictions in a single sentence in the press release material related to public safety was selected through an initial crawling process. Negative dictions are as shown in Table 2 (total 268 dictions).

The selected data underwent a set refinement process to select relevant material only and a further preconditioning process for topic analysis. First, the big data collecting technique, ‘crawling’, was used for data refinement because it collects large amounts of press release material. In this research, a separate data collecting and saving device for media data was developed and utilized for efficient data collection, and the developed crawler used ‘.NET’ to differentiate between saved data and conditional search through the developed data collection function and ‘Microsoft Access’. In summary, the refinement method using the self-developed crawler ① maintains just one url in a press release by eliminating overlapping multiple urls, ② deletes press releases that have no content or cannot be recognized, and ③ deletes press releases with unclear sources (reporting press company, date). Then, through the ‘TonK’ program, the preconditioning process ④ preconditions special symbols and Chinese letters in the content with ‘TonK’ and ⑤ conducts natural language processing through ‘TonK’.

3.1.3. Present Condition of Subject Material of Analysis

The present condition of media material related to subway safety that will be analyzed is shown in Table 2. A total of 72,790 press releases were initially collected through crawling, of which 29,567 were chosen as analysis subjects through a selection process. Regarding the topic area, 10,851 (36.7%) dealt with road traffic safety, 6,727 (22.8%) subway safety, and 11,989 (40.5%) airport safety.

Table 2. Number of selected press releases to be analyzed

[unit: case]			
Interval 1	Interval 2	Interval 3	Total
183	3,199	3,345	6,727

Note) Interval 1: Road traffic safety is from 1990 to 2002, Interval 2: from 2003 to 2013, Interval 3: 2014 May, 2016

3.2. Analysis Method

3.2.1. Method of Keyword Frequency Analysis

This research analyzes the changes in keyword frequency in press releases related to ‘road safety’. Through this analysis, one can observe which keywords were mentioned frequently in the media related to ‘road safety’ from 1990 until May 2016, and how keywords varied from year to year. This research implemented the terms shown in Table 3 for the keyword frequency analysis.

Table 3. Terms and Definition to Analyze Keyword Exposure Frequency

Terms	Meaning
TF	Term Frequency, which is a measure of how much the word appears throughout the media coverage
DF	Document Frequency, which is the frequency of the press coverage that contains the word, indicates how much the media is handling the word

IDF	Inversed Document Frequency, the reciprocal of the frequency of press coverage that contains the word
TF-IDF	TF multiplied by IDF, The higher the value of the keyword is, the more likely it is that the keyword is spreading widely in the press, which means that the keyword is treated frequently in many media

This research implemented the ‘TF-IDF’ index to observe the changes in keywords shown in press release. This index is a value revealing the overall even frequency of the respective keyword in collected press release material, and the higher the value the more frequent the keyword was mentioned in the media. The calculation principles of the TF-IDF index in the Topic Analysis Model are as the following.

The TF-IDF index is calculated by multiplying the TF value and IDF value squared, but generally in the topic analysis model, the word frequency in the document is divided by the total number of frequency of all words, using a TF value in a normalized format in order to find the TF-IDF index. The reason for this is to prevent a skewed TF value according to the text volume in the document. Moreover, when only big words with an overly large TF value are considered as important keywords, words that are not utilized can be included as well. In order to solve this problem, the IDF value is supplemented before being used (Lee and Kim, 2009). Generally in a topic analysis, the IDF value uses the value derived from dividing the number of documents included in the document set with the number of documents that use a specific word, in accordance with the TF value’s normalization theory.

The TF value is the following when the word w_i appears in the press release d_j .

$$tf_{i,j} = n_{i,j} / \sum_k n_{k,j} \quad (1)$$

Here, $n_{i,j}$ is the number of times the word w_i appeared in the press release d_j

$\sum_k n_{k,j}$ is the number of times all the words appeared in the press release d_j

Moreover, the IDF value is the following when the word w_i appears in the press release d_j .

$$idf_{i,j} = \log \frac{|D|}{|\{w_i | t_j \in d_j\}|} \quad (2)$$

Here, $|D|$ is the number of documents included in the document set

$|\{w_i | t_j \in d_j\}|$ is the number of documents in which the word t_j appears

Thus, the TF-IDF index can be defined as the following.

$$TFIDF_{i,j} = tf_{i,j} \times idf_{i,j} = \{n_{i,j} / \sum_k n_{k,j}\} \times \log \frac{|D|}{|\{w_i | t_j \in d_j\}|} \quad (3)$$

3.2.2. Analysis Method for Press Release Issues

In this research, the ‘TonK v1.0.12 (‘TonK’ henceforth)’ developed in KAIST (Korea Advanced Institute of Science and Technology) was used as an analysis tool in order to experimentally realize the LDA model. The topic analysis function in the ‘TonK’ program realizes topic analysis that applies the LDA model based on the theories shown in Blei et al. (2003) and Rosen-Zvi et al. (2010). Topic analysis based on the LDA model follows the method of 1 document following the Dirichlet distribution allocating topics to a K number of document sets with the Bayesian statistics method. When the LDA

model is estimated, a parameter value of α is derived for each document set among the K set of documents, and this study defined this as the ‘level of topic contribution’. The level of topic contribution is how much a topic that was similarly categorized in the LDA model takes up in the total K number of topic sets. Meanwhile, considering the size of the topic set that varies according to the period of analysis in this research, the ‘modified level of topic contribution’ was used to balance this. The theory on which the modified level of topic contribution was estimated varies on the number of topic sets (K) depending on the period of analysis, and can be considered as the modified level of topic contribution value for each time period after bringing into uniformity the number of topic sets to 10. Furthermore, the ‘standardized value of the level of topic contribution’ estimated by standardizing the arithmetic mean and distribution of the top 10 modified level of topic contributions was utilized as the final issue ranking of press releases. Meanwhile, the data crawler used to crawl media data underwent a separate coding process, and the collected material underwent natural language processing by linking it with the ‘TonK’ program. This research’s LDA topic analysis process adhered to the process used in Oh et al. (2016 a) and the detailed process flow chart is shown in appendix 1.

4. Analysis Results

4.1. Analysis Results of Keyword Frequency

The changes in the keywords in press releases related to road safety are as shown in Table 4. Based on the TF-IDF index, words that are being dealt with weightily are ‘express’, ‘construction’, ‘safety’, ‘transportation’, during interval 1 (1993~2002), ‘speed’, ‘safety’, ‘traffic’, ‘facility’, and ‘construction’ during interval 2 (2003~2013), and ‘speeding’, ‘speeding brake’, ‘address’. Safety problems related to expressway are steadily being mentioned as subjects, and it can be expected that safety issues related to speed bumps and traffic signs have been raised a lot recently in the third period.

Table 4. Road safety related press media keyword Trends (within TF-IDF index 10 rank)

rank	Interval 1 (1990~2002), of the total 183				Interval 2 (2003~2013), of the total 3,199				Interval 3 (2014~May 2016), of the total 3,345			
	key word	TF-IDF	DF	TF	key word	TF-IDF	DF	TF	key word	TF-IDF	DF	TF
1	high speed	11.165	1,037	5,524	high speed	32.12	2,388	13,094	persons	9.618	530	3,897
2	construction	9.004	1,076	3,818	accident	27.233	1,835	8,479	speeding	8.07	338	3,245
3	safety	7.472	451	1,827	traffic	22.657	3,277	13,180	bust	7.669	308	2,955
4	architecture	7.388	833	2,347	facility	21.317	1,675	6,825	address	6.016	608	2,734
5	traffic	7.367	1,188	4,518	construction	20.63	2,674	9,324	prevention	5.611	692	3,973
6	section	6.194	751	2,359	accident	17.268	1,751	5,646	guidance	5.505	440	1,851
7	orders	6.004	179	487	business	16.042	2,045	6,619	check	5.454	391	1,769
8	accident	5.708	446	1,620	construction	15.804	1,870	5,267	plate	5.32	536	2,193
9	Seoul	5.554	960	2,421	car	15.678	2,246	7,179	unmanned	4.983	228	1,714
10	management	5.36	1,045	2,669	section	15.525	2,018	5,711	group	4.908	360	1,533



4.2. Analysis Results of Press Release Issues

Appendix 2 shows a comparison of the issues the media has dealt with regarding road safety for each time interval. The issues that received the most focused reports from the media regarding road safety during Interval 1 (1993~2002) were those related to subway crimes such as ‘the frequent occurrence of sexual assaults in the Seoul subway (2001)’, ‘the subway is a hotbed for crime (2001)’, etc. During interval 2 (2003~2013), issues related to the safety and environment of the Busan subway were focused on such as ‘the discovery of asbestos in Seomyeon Station of the Busan subway (2008)’, ‘successive fires in the Busan subway (2012)’, ‘the Busan subway labor union demands manpower reinforcements for safety (2013)’, etc. During Interval 3 (2014~May 2016), issues related to the safety of the Seoul subway were in the spotlight such as ‘11 accidents in one year, the unsafe safety issues of the Seoul subway (2014)’, ‘Woman in her 80s dies from getting stuck in the screen doors in Seoul Station, line 1 (2016)’.

Appendix 2 periodically compares road safety related issues that the media has dealt with by time period. The issue that was dealt with the greatest emphasis in the first period (1990~2002) was related to ‘inconvenience in Hi-Pass fare recharge (2001)’ and ‘difficulties faced in South-North Railway ·road Dialogue (2001)’. Issues related to speed bumps not being installed and being non-compliant, such as ‘no speed bumps in front of elementary schools (2005)’, ‘speed bumps ignoring standards (2012)’, were primarily raised in the second period (2003~2013). In the third period (2014~May 2016), issues related to ‘complaints about the road infrastructure in Seoul (2015)’, ‘danger of speed bumps (2015)’ were raised, and it is noteworthy that the safety issue of speed bumps was raised, similarly to the second period.

The issues ranked second include the ‘opposition of Ilsan-Toegyewon outer-circle road construction by environmental organizations (2001)’ in the first period, and ‘Metro unmanned driving safety issue (2013)’ in the second period. In the third period, first and second significant issues both dealt with public inconvenience and safety issues related to speed bumps: ‘speed bumps as one pleases, increasing inconvenience (2016)’, ‘Gimpo, headaches caused by non-compliant speed bumps (2016)’, ‘frequent non-compliant speed bumps (2016)’.

The issues ranked third include ‘road safety investment source needs to be secured (2000)’, ‘significant changes to road signs (2002)’ in the first period, and in the second period, issues related to the effectiveness and safety of bicycle roads arose, such as ‘bicycle road problems (2006)’, ‘zero effectiveness of bicycle roads (2012)’, ‘sudden increase of bicycle road accidents, vulnerable safety facilities (2012)’. Similar to the first and second ranked issues, public safety issues related to speed bumps were dealt with in the third period, such as ‘no-standard speed bumps all require maintenance (2014)’, ‘speed bump, exceeding civil safety standards (2015)’.

The road safety related issue that was dealt with the greatest emphasis in the first period (1990~2002) was the ‘inconvenience in Hi-Pass fare recharge (2001)’ and ‘difficulties faced in South-North Railway·road Dialogue (2001)’. Issues related to speed bumps not being installed and being non-compliant, such as ‘speed bumps in front of elementary schools (2005)’, ‘speed bumps ignoring standards (2012)’, were primarily raised in the second period (2003~2013). In the third period



(2014~May 2016), issues related to speed bumps not being installed and being non-compliant, 'complaints about the road infrastructure in Seoul (2015)', 'danger of speed bumps (2015)' were raised.

Other issues discussed include 'government to perform large-scale crackdown on illegal roadside signboards', 'inconveniences for disabled people at the highway rest stops', 'unauthorized discharge of wastewater at highway rest stop', 'poor management of traffic safety facilities', 'mandatory road safety collision facilities', 'national compensation for wrong road signs', 'no signs for temporary opening roads', 'illegal driving at habitual road safety law violation reporting point', 'serious violation of traffic laws by city buses' in the first period. 'Drastic expansion of preliminary environmental review', 'poor road environmental impact assessment', 'abolition of ineffective road signs', 'poor road signage marking and installation', 'traffic control effect of unmanned airship', 'speeding risk of Daegu-Pohang highway' were discussed in the second period. In the third period, issues such as 'excessive speed bumps installed', 'fire safety frugidity of highway resting stops', 'Expressway Corporation violated 90% of resting facility arrangement regulations', 'low speed buses to disappear because of speed bumps', 'bus ignoring speed bump to compensate for injured passenger' were discussed.

5. Conclusion

This research utilized press releases in the media that can influence transportation policy making in order to experiment the data-driven policy issue searching method. The model used for the experiment was a probabilistic topic model based on the LDA algorithm, a crawler was composed to collect data, and an academic program was utilized for text mining analysis. The issue analysis was on subway safety. This research was unique in that a taxonomy comprising levels of specific keywords that will be focused on and analyzed during this time was created in order to limit the analysis topic, while negative meaning dictions were borrowed to collect media press releases dealing with subway safety issues. By implementing a TF-IDF index and analyzing a keyword frequency analysis, it was determined which keywords related to subway safety were most frequently used in the media during each time period, and various media issues were ranked according to the estimated level of topic contribution of the respective press release material through topic analysis. As a result of the analysis, the subjects of the most talked about the road safety were the subjects of the speeding brakes in the first period (1990~2002), and in the second period (2003~2013) In the third period (2014~May 2016), safety problems at the highway rest area were mentioned.

ACKNOWLEDGEMENTS

I hereby disclose that this paper is an edited/supplemented version of Oh (2016 b) so that it is suitable for this academic journal. Moreover, I would like to express my deep gratitude to professor Il-chul Moon at Korea Advanced Institute of Science and Technology(KAIST) for his theoretical and technical help in the 'TonK' program analysis process.



REFERENCE

- Blei, D. M., Ng, A. Y. and Jordan, M. I. (2003). Latent Dirichlet Allocation, *J. of Machine Learning Res.*, Vol. 3.
- Collins, C., Hasan, S. and Ukkusuri, S. V. (2013). A novel transit rider satisfaction metric: Rider Sentiments Measured from Online Social Media Data. *J. Public Transp.*, Vol.16, No.2.
- Das, S., Sun, X. and Dutta, A. et al. (2016). Text mining and topic modeling on compendium papers from transportation research board annual meetings, In: *The 95th TRB Annual Meetings for Presentation and Publication under Committee.*
- Farrahi, K. and Gatica-Perez, D. (2011). Discovering routines from large-scale human locations using probabilistic topic models, *ACM Trans. Intell. Syst. Technol.*, Vol.2, No.1.
- Gao, H., Tang, J. and Liu, H. (2012). gSCorr: modeling geo-social correlations for new check-ins on location-based social networks, In: *Proceedings of the 21st ACM International Conference on Information and Knowledge Management.*
- Hasan, S. and Ukkusuri, S. V. (2014). Urban activity pattern classification using topic models from online geo-location data, *Transp. Res. Part C*, Vol.44.
- Lee, S. and Kim, H.-J. (2009). Keyword extraction from news corpus using modified TF-IDF, *The J. of Soc. for e-Business Studies*, Vol.14, No.4.
- Moon, I.-C., Oh, A. H. and Carley, K. M. (2011). Analyzing social median escalating crisis situations”, *IEEE Intelligence and Security Informatics*, Beijing, China.
- Ni, M., He, Q. and Gao, J. (2014). Using social media to predict traffic flow under special event conditions”. In: *The 93rd Annual Meeting of Transportation Research Board.*
- Oh, C.-S., Lee, Y. and Ko, M. (2016 a). Establishment of ITS policy issues investigation method in the road section applied textmining, *J. Korea Inst. Intell. Trasp. Syst.*, Vol.15. No.6.
- Oh, C.-S. (2016 b). Exploring and analyzing national security-related audit problems using big data, *Audit and Inspection Research Institute.*
- Rosen-Zvi, M., Chemudugunra, C., Griffiths, T., Symth, P. and Steyvers, M. (2010), Learning author-topic models fro, text corpora, *ACM Transactions on Information Systems*, Vol. 28, No. 1.
- Ryu, I., Lee, J., Park, G. C., Choi, K. and Hwang, J-M. (2015), Analysis of highway traffic indices using internet search data, *J. Korean Soc. Transp.*, Vol.33, No.1.
- Ryu, I., Lee, J., Park, G. C., Choi, K. and Hwang, J-M. (2016), Exploring the temporal relationship between traffic information web/mobile application access and actual traffic volume on expressways analysis of highway traffic indices using internet search data, *J. Korean Soc. Transp.*, Vol.34, No.1.
- Sun, L. and Yin, Y. (2017). Discovering themes and trends in transportation research using topic modeling, *Transp. Res. Part C*, Vol.77.



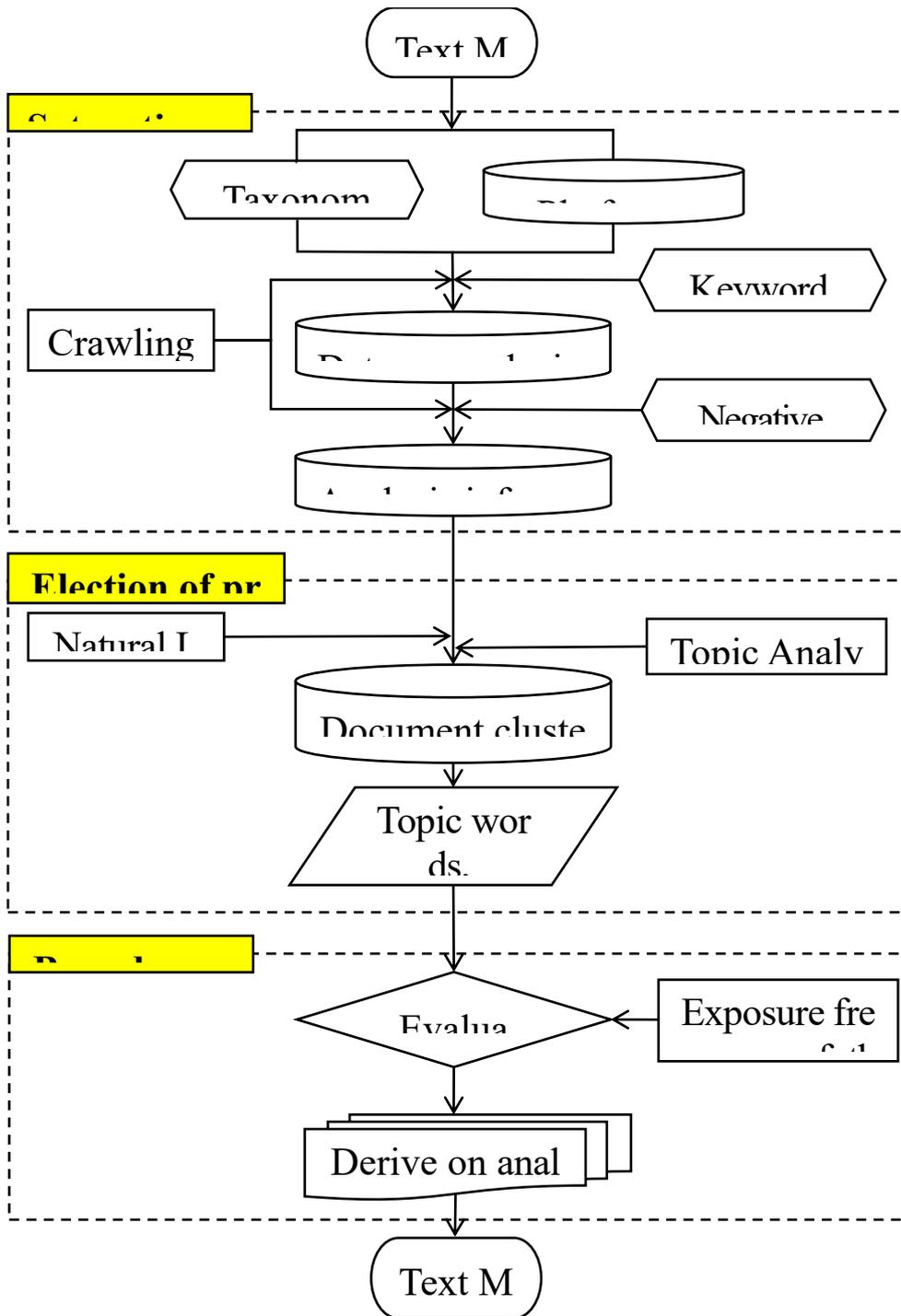
APPENDIX

Appendix 1. Negative meaning dictions adopted in this study (total 268 dictions)

process, snatch, aggravate, fake, reckless, conflict, topdog-underdog relationship, backward, dirty money, experienced, experience, experiencing, uneconomic, hierarchical selfishness, elastic band, high-handed, cat, purpose, broken, storeroom, public, public secret, scheme, over/understatement of property plant and equipment, over auditing, custom, provide an assortment, authoritative, rights and interests, The Its Know, league of their own, party of their own, no way, root cause, this is news to me, tail cutting, gulp, sharing the profit, bad, being bad, parachute, spoilt mood, abuse, taxpayer, waste, jittery, issue, omit, ostrichism, an ignorant person, blind money, dawdle, slow administration, efficiency, great moral discovery, attack, morally indifferent, moral laxity, bankruptcy, go back, disappear, beat a dead horse, to be revealed, plain water, tepid water, fool, scrape the surface, as one pleases, indulgence, aimless, other purpose, pretend not to hear, unreasonable, complacency, omnipotence, useless, silent, connivance, disorder, problem, no questions asked, future generation, insignificant, not implemented, insufficient, insensitive to public sentiment, civil complaint, bottomless pit, deprive, lax, abode of demons, indifferent attitude, unjust, weak, polarization between rich and poor, forbearance, absence, inappropriate, disqualified, inaccurate, not enough, poor, ministry selfishness, insensitivity to corruption, exaggerate, exaggeration, insensitivity, dissatisfaction, illegal, discomfort, uncomforTable truth, uneconomic, inefficient, corruption, secret, ineffective, inefficient, of the iceberg, about to lose, puff up, blind spot, low morale, groundless fact, truth be told, after the fair, fish, novice shaman, good will, washing the dishes, not yet ripe, sanctuary, silent outcry, alienation, to lock the sTable door after the horse is stolen, careless, all looks no substance, light punishment, tactic, hush up, urgent, be urgent, will be urgent, plant, reliability, prudent, cautious position, reality, serious, is it serious?, pocket money, useless, insensitive to danger, well-known, an open secret, pressure, underdog, what are you doing where, dumbfound, what now, oppressed, random, mess, nonsense, same old, Buddhist prayer, predicted, pre-notified, waste of budget, back and forth, misuse, malfunction, distortion, fixing the barn, worry, cry, long face, violation, violation of the constitution, nominal, ethics, conceal, doubt, egocentrism, how could this be, after all this time, double, double standard, consistency, a chip in porridge, recurrence, rice offered to Buddha, trawler, poor, collide, decrease, expose, proper timing, downfall, desperate, inspect, politics, exactness, each, men are blind in their own causes, fabricate, rough-and-ready, work without plans, masterless, smooth, overlap, until now, delay, criticize, slow progress, really?, collectivism, disciplinary punishment, manipulated, manipulated game, secret promise, failure, embezzlement, take, took, punishment, thorough, contract audit, request, reduce, silent outcry, partition, code, big problem, desk, negligent, shaken off, shook off, hollow, expose, favor, ripple, situation, splash out, expediency, defraudation, target, need, needed, unnecessary, sunflower, pant, false, tax paid by the sweat of one's brow, formality, unfair, unequal treatment, confusion, confirmed, environment, embezzle, ineffective, inefficient, harm



Appendix 2. Flow chart for audit issues investigation



Topic ranking	Period 1 (1993~2002)		Period 2 (2003~2013)		Period 3 (2014~June 2015)	
	Standardized value of level of contribution*	Major topics in media reports**	Standardized value of level of contribution**	Major topics in media reports**	Standardized value of level of contribution*	Major topics in media reports**
1	1.341	<ul style="list-style-type: none"> Setbacks in 'Nonstop Hi-pass card system' (2000) Inconveniences in recharging the Hi-pass (2001) Last minute pains in the talk for two Koreas' railways and roads (2002) 	2.644	<ul style="list-style-type: none"> No speed bumps in front of elementary schools (2005) Many speed bumps ignore regulations (2012) Speed bumps causing accidents because of their unnecessary height (2010) 	2.724	<ul style="list-style-type: none"> Citizens of Baegkok-myun, Jincheon, request speed bump installations (2015) Seoul citizens unhappy with road infrastructure facilities (2015) Speed bumps trapping people (2015)
2	1.296	<ul style="list-style-type: none"> Outer ring road from Ilsan to Toegyewon to pierce Bukhansan, aggravated opposition from environmental organizations (2001) Review of the legitimacy and validity of the outer ring piercing Bukhansan (2002) 	0.350	<ul style="list-style-type: none"> Automated operation of city railway must be stopped (2013) Self-driving cars needed for underprivileged people in traffic (2013) Hyundai Motors, currently developing an automated car (2005) 	0.260	<ul style="list-style-type: none"> Increased discomfort due to haphazard speed bumps (2016) Gimpo, headaches due to substandard speed bumps (2016) Many speed bumps ignore regulations (2016)
3	1.098	<ul style="list-style-type: none"> Investment sources for road traffic safety crucial (2000) Road signs to transform completely (2002) 	0.282	<ul style="list-style-type: none"> Road problems for cyclists experienced firsthand (2006) Bicycle roads have zero practicality (2012) Sudden increase in accidents on bicycle roads with poor safety control (2012) 	0.111	<ul style="list-style-type: none"> Non-standardized speed bumps need to be organized (2014) Speed bumps, surpasses the citizen safety level (2015)
4	0.777	<ul style="list-style-type: none"> Worries for traffic congestion due to tardy construction in the Anseong connecting road between the Two Koreas (2000) Startling accidents for 'Kickboard kids' speeding on roads at night (2000) Government, massive control on illegal road signs(2000) 	-0.086	<ul style="list-style-type: none"> Big increase in 'prior environment inspection' (2004) Controversy on the effects of roads on the environment between salespeople and 'Beommul'(2005) Prior environment inspections mandatory for national projects such as dams, canals, ports as well (2004) 	-0.283	<ul style="list-style-type: none"> ※ Could not find appropriate topic

* The standardized value of the issue contribution of the top 10 ranked issues using arithmetic average and variance
 **Summarized key issues in the order of the press release issue contribution(LDA model parameter estimates) within the program.

Commentary 1) Only issues in the top 7 issue ranking shown

Commentary 2) The press releases analysed were refined using the parts analysis, TF-IDF filtering and word length filtering functions in the 'TonK' program

Appendix 3. contd. Analysis of changes in subway safety issues reported in the media sorted by period(within issue raking 10)

Topic ranking	Period 1 (1993~2002)		Period 2 (2003~2013)		Period 3 (2014~June 2015)	
	Standardized value of level of contribution**	Major topics in media reports**	Standardized value of level of contribution**	Major topics in media reports**	Standardized value of level of contribution**	Major topics in media reports**
5	-0.490	<ul style="list-style-type: none"> • Inconvenient stairs in highway service stations, overlooking disability rights (2002) • Unauthorized discharge of waste water from highway service stations (2002) 	-0.337	<ul style="list-style-type: none"> • Unpractical road signs to be taken down (2006) • Road signs to exclude structure names and only include road names (2007) • Road signs hidden by roadside trees (2005) 	-0.312	<ul style="list-style-type: none"> • Speed bumps hindering those working in transportation (2014) • Excessive speed bumps in Manseong (2016)
6	-0.560	<ul style="list-style-type: none"> • Gimhae 'disgrace', named city with poor road traffic safety control (2002) • Concerns for 'series of explosions' when digging Yeochun complex roads (2000) 	-0.416	<ul style="list-style-type: none"> • Traffic control of automated aircraft (2013) • Selfish driving of automated aircraft to be controlled (2013) 	-0.357	<ul style="list-style-type: none"> • Insensitivity to fire safety management in highway service centers (2014) • Road constructions, violated more than 90% of regulations regarding resting facility assignment (2015)
7	-0.568	<ul style="list-style-type: none"> • Ministry of Construction and Transportation, designates collision facility construction crucial as road traffic safety control (2001) • Jeollabuk-do Jumbled up road signs (2001) • National reparation for incorrect road signs (2002) • Drivers experience discomfort due to lack of road signs about temporary roads (2000) • Controversy on illegal driving in area with a lot of reports on violating traffic laws (2001) 	-0.458	<ul style="list-style-type: none"> • Decreased effects of highway between Daegu and Pohang (2004) • Dangers of speeding on highway between Daegu and Pohang (2004) 	-0.365	<ul style="list-style-type: none"> • Low floor buses to disappear due to speed bumps (2014) • Bus that ignored speed limits to compensate for injured passenger (2016)

* The standardized value of the issue contribution of the top 10 ranked issues using arithmetic average and variance

**Summarized key issues in the order of the press release issue contribution(LDA model parameter estimates) within the program.

Commentary 1) Only issues in the top 7 issue ranking shown

Commentary 2) The press releases analysed were refined using the parts analysis, TF-IDF filtering and word length filtering functions in the 'TonK' program