Cities 108 (2021) 102941

Contents lists available at ScienceDirect

Cities

journal homepage: www.elsevier.com/locate/cities

A value of civic voices for smart city: A big data analysis of civic queries posed by Seoul citizens

Byungjun Kim[a], Minjoo Yoo[a], Keon Chul Park[b], Kyeo Re Lee[a], Jang Hyun Kim[a,*]

[a] Department of Interaction Science, Sungkyunkwan University, Republic of Korea
[b] Department of Digital Policy, Seoul Digital Foundation, Republic of Korea

ARTICLE INFO

Keywords:
Word2vec
Dynamic topic model
Civil complaints
Automatic classification
Smart city

ABSTRACT

Since urban problems are increasingly becoming complex and multifaceted due to the rapid urbanization, interest in the development of smart cities as an efficient solution to such problems is growing. The significance of smart cities lies in not having technology itself but using technology with a novel approach to solve urban problems, enhance the quality of life for urban residents, and optimize government performance. Such an approach includes scientifically processing civic query (including complaints, suggestions, and inquiry) data about a city and planning the city such that its policies reflect its residents' voices to ensure "throughput legitimacy." This study introduces a novel approach to analyze diverse informal civic query data for a city and plan the city in a way that its residents want. By analyzing 160,000 civic queries accumulated over 10 years from 2006 to 2017 from the "Oasis of 10 Million Imagination," which is a civic participation platform of Seoul, this study aims to contribute to the sustainable development of Seoul so that it can plan a citizen-centric and smart city that satisfies the demands of its residents. By applying Dynamic Topic Model, the authors intend to specify civic demands and forecast the demands of citizens.

1. Introduction of research questions and past literature

As urban problems are getting complicated rapidly, traditional ways of hearing citizens' complaints and suggestions are not useful. Therefore, the concept of smart city is spreading fast to accommodate urban problems efficiently. Smart city is essential to respond to the voices of citizens. A vast amount of the voices of citizens' data accumulated through various channels and methods should be analyzed efficiently, scientifically, and systematically. Through this process, it is possible to grasp civil complaints and to find optimal solutions to urban problems.

Smart cities and big data are topics that are actively studied in administrative studies, urban engineering, and information science. Ruijer, Grimmelikhuijsen, and Meijer (2017) proposed open data platform for democracy. They stressed importance of a context-sensitive open data design which could lead to form a basis for platforms supporting democratic processes. Abella, Ortiz-de-Urbina-Criado, and De-Pablos-Heredero (2017) also emphasized that smart cities are key elements that solve the complex problems of the city. And the study presented a model that included three stages of how big data from the smart cities' portal creates value for citizens and society. Cao, Giyyarpuram, Farahbakhsh, and Crespi (2020) cited data as an important element of smart cites, pointing out that there was a lack of systematic approach in existing smart cites studies. The paper also emphasized that sharing data from cities depends on trust from the perspective of the two different actors, data producers and consumers, and proposed a trustworthy data sharing platform. Aforementioned smart city data-related studies suggest a systematic model, in which the success or failure of smart cities depends on how data are used. However, they failed to present a practical methodology using public data from the perspective of government agencies or citizens. Our research provides actual analysis cases and algorithms necessary for building data-based smart cities.

Complaints and inquiries were data of interest in the fields of humanities, social sciences, and computer engineering. Kim (2010) assessed how the objectives of Seoul's platform for collecting complaints were achieved in terms of citizen participation. He stated that civic participation has been facilitated through online platforms such as "Oasis of Ten Million Imaginations (OTMI, hereafter)." He also introduced the details of OTMI and characterized it as a kind of decision support system. Hagen et al. (2016) analyzed the linguistic and semantic elements of the petition by examining online petition text data including civil complaints. The study analyzed petitions text based on dictionary and word frequency, and retrieved clusters of petitions by

* Corresponding author at: 9B307 International Hall, 25-2, Sungkyunkwan-Ro, Jongno-Gu, Seoul, Republic of Korea.
E-mail addresses: parkkc07@sdf.seoul.kr (K.C. Park), alohakim@skku.edu (J.H. Kim).

https://doi.org/10.1016/j.cities.2020.102941
Received 30 September 2019; Received in revised form 11 June 2020; Accepted 29 July 2020
Available online 01 October 2020
0264-2751/

unsupervised learning (topic modeling). On the other hand, our research can be used immediately to assign civil servants according to the demand for civil queries as it can be sophisticated in classification by supervised learning. Sinha, Guha, Varma, Mukherjee, and Mannarswamy (2019) applied LDA and CUrb (**C**ontext driven **Urb**an issue resolution) frameworks to analyze web data sources on urban civic issues of a city. They identified citizen views using semantic traits, location and time consumed for issue resolution with comparing topic classification techniques including Word2vec and Doc2vec. Vargas-Calderón and Camargo (2019) analyzed a set of tweets from Bogota's citizens by applying Word2vec model. They detected topics automatically related to politics, news, and religion and the results could be useful for the local government to make valuable decision. Sano, Yamaguchi, and Mine (2015) proposed a method for automatically categorizing complaints to reduce the workload for government side, but their study's data included only the complaints about city park, not city affairs in general. Hardaya, Dhini, and Surjandari (2017) studied classifications of Twitter data including complaints and proposals about Jakarta from 2013 to 2016 with the support-vector machine (SVM) algorithm. The accuracy rate of automatic classifications was 91.37% for six classes. However, the paper's data covers only four years with six categories. Son and Kim (2017) proposed an automatic classification system for processing civil complaints. Moreover, they suggested that the sorted complaints be assigned to proper department(s) or administrative unit(s), but they provided only methodological suggestions and failed to apply their arguments to real data. In this study, the authors found that past studies did not provide enough insights due to the limitation of their data and/or methodology. These existing studies show that it is necessary to explain how to use such method for urban administration processes.

From a methodological perspective, using a dynamic topic modeling approach can provide several benefits for analyzing a huge amount of the voices of citizens public data. Linton, Teo, Bommes, Chen, and Härdle (2017) constructed indicators of fraudulent schemes in cryptocurrencies using the data from relevant forums with dynamic topic modeling. They detected 50 topics according to event detection algorithm. Ha, Beijnon, Kim, Lee, and Kim (2017) used dynamic topic modeling to examine how user perception of smartwatch were described in online community. They suggested innovative ideas of enhancing smartwatch based on users' perceptions revealed online. Greene and Cross (2017) explained that dynamic topic modeling methodology consistently identifies semantic traits of the major topics of the political agenda. For that reason, they emphasized that the insight provided through analysis of a corpus of political speeches with dynamic topic modeling demonstrates the latent dynamics of political agenda and political system, which allows public to assess how the political system functions. Similarly, according to Wollard (2017), a topic model analysis of public comments to the San Francisco police commission showed which police-related topic categories including crime, accountability, and community issues were discussed between the public and decision-makers. The author explained that applying dynamic topic modeling in the study can evaluate what topics the public were interested in and were discussed over time.

The aim of this study was to analyze the text-based voices of citizens' big data to build a foundation for automatic monitoring of citizen complaints, and to promote citizen-centered policies. First, to training data (16,821 cases) that had been classified into 10 categories by civil affairs officers and machine learning methods, this research added unclassified testing data (143,496 cases). Second, keywords were extracted from an automatic classification of complaints, and dynamic topic model was applied. Finally, a methodology for efficiently classifying complaints for OTMI has been proposed.

Based on the literatures examined above, there are two research questions we suggest:

RQ1. What are the results of automatically classifying civic complaints on OTMI using machine learning algorithms?
RQ2. Can we utilize the classifications acquired from RQ1 analysis to track issues in topic and to label them through dynamic topic model?

In sum, past literatures are found to concentrate on developing quantitative methods, such as data algorithms. This study focuses on opinion mining including automatic classification of complaints and retrieving topics automatically. Method, Results, and Conclusion sections are following the current section.

2. Method

2.1. Data collection

2.1.1. Subject for analysis

The voices of citizens in Seoul were analyzed via the city's online platform, the “Oasis of Ten Million Imaginations.” The OTMI was designed to collect citizens' complaints and hand over the most liked ideas to city officials, who would respond to those citizens. For instance, if a citizen recommends a good solution to a city problem, other citizens and officials discuss it and may develop it as a policy initiative. Text data in the platform from 2006 to 2017—160,316 complaints—were used in this study.

2.1.2. Data columns

Civil inputs (suggestions, complaints, and inquiry) data consists of 13 columns as listed in Table 1 including title, user's name, classification, and so forth. A classification was chosen manually by a civil servant who has dealt with civil voices among 13 administrative fields (i.e., environment, traffic, planning, housing, safety, economy, health, female, culture, welfare, tax, construction, etc.). In this study, we applied the top 13 public administration fields defined in a guide for designing functions for informatization support system (MOIS & NIA, 2010). This is because the government's organization and definition of tasks were designed based on the guide, and thus this study applied the criteria by which civil voices are classified from it. A total of 16,821 complaints were categorized by the civil servant. In this study, the classification column of testing set (a total of 143,495 complaints) was a target variable for automatic classification (Table 1).

2.1.3. Processes

The processes of data analysis included four steps:1) data crawling, 2) data preprocessing, 3) data analysis, and 4) results analysis (Table 2).

The first step, data crawling, includes making sure personal information being deleted in raw data. The second step, data

Table 1
Data columns for civil complaints.

Index	Publicly viewable	Title	ID	Vote	Date created	Status
3	Yes	Natural city without ozone	Lost ways		October 11, 2006	
Complaint type	Detailed type	Classification	Department responsible	Contents		Department's opinion
Free offer	Simple query	Environment	Living environment department	“Seoul's environmental pollution is getting worse…”		

Table 2
The processes of data analysis.

1. Data crawling	2. Data preprocessing	3. Data analysis	4. Results analysis
Selecting data sets: "Oasis of 10 million imaginations" Data cleansing: Excluding personal information from raw data	Morphological analysis Removing stop words and setting up minimum units	Training: Modeling training dataset with Word2Vec. Testing: Putting testing data into model and classify them. Extracting keywords and conducting topic model	Analyzing voices of citizens Calculating accuracy score. Forecasting trends of civil demands by dynamic topic model Developing a new policy based on civil complaints

preprocessing, includes removing stop words and conducting morphological analysis. In this study, Mecab-ko[1] was used for morphological analysis. Only nouns were accepted. In this process, we removed needless words, punctuation marks, emojis, and typing errors. Mecab-ko has a system dictionary that covers only common words, so we added neologisms and compound nouns to the user dictionary for efficient analyses. We combined a title column with a 'Contents' column and defined stop words for excluding unnecessary words from nouns extracted by the morphological analyzer. For example, we added demonstrative pronouns (e.g., this, that, those) and redundant words like "Seoul" to the stop words corpus. After deleting stop words, we established a minimum number of words. If a document had less than 20 words, the complaint was excluded from the data sets. This decision was made because a document with less than 20 words tends to be junk texts. The third step, data analysis, is conducted for training dataset first for modeling by Word2Vec. Later, the model is used for classifying testing data set. The fourth step, results analysis, is to calculate accuracy score for testing dataset, where nouns were extracted with a morphological analysis. With this noun and machine learning model, "traffic" was used to label texts (Table 3).

2.2. Modeling

There were 10 classifications of complaints[2] (i.e., final target variables, shown in Table 4).

We divided data sets into two parts: training data and testing data. Categories for training data sets had already been determined by a civil servant who has specialty in handling civil complaints; the classification column of testing data sets was empty (Table 5).

2.3. Word2vec

Word2vec produces word embeddings; words in text documents can be transformed to countable vectors (e.g., Yao et al., 2017; Luo et al., 2018; Krishna et al., 2019). Even though there is a large corpus of text, Word2vec makes a vector space which includes hundreds or thousands of dimensions. In the vector space, a position is given to each word. This position can be expressed as coordinates (Fig. 1).

In the past, one-hot encoding was used in text mining. With one-hot encoding, a word is represented as 1 or zero. For example, "woman" could be expressed as [1,0,0], and "man" could be expressed as [0,1,0]. The distance of each word from point [0,0,0] is the same because one-hot encoding has only two options (1 or zero).

Moreover, Word2vec considers neighboring words back and forth. The size of a context window refers to the number of words near a given word; the algorithm checks relationships and contexts. If the size of a window is 20, 20 words before and after a given word are included as context words of that word. In this study, the skip-gram architecture

Table 3
An example of automatic classification.

Raw data (civil complaints text)
Title: A bathroom should be installed on the subway platform.
Contents: "When I use the subway, I often use the bathroom…"
↓
Extracting morphemes (nouns)
["Subway," "Platform," "Bathroom," "Installation," "Subway," "Usage," "Bathroom," "Usage," "Occurrence," "Present," "Platform," "Structure," "Bathroom," "A waiting room," "Time," "Standard," "Bathroom," "Parts," "Usage," "Recent," "Bathroom," "Remodeling," "Work," "Positive response," "A cause," "Standard," "Maintenance," "A cause," "Platform," "Citizens," "Comfort," "Usage," "Bathroom"]
↓
Automatic classification
The complaint text is classified as "traffic."

Table 4
Topics of complaints.

Classification	Main contents
Health	Public health, hygiene, etc.
Economy	Industry, payments for charges, prices, etc.
Traffic	Public transportation, pedestrians, etc.
Culture	Sightseeing, cultural experiences, cyber subculture, etc.
Welfare	Positive discrimination, labor welfare
Taxes	Default, tax collection, etc.
Safety	Firefighting, disaster prevention, declaration of emergency, etc.
Female	Infant care, multi-children families, etc.
Housing	Rentals, housing development, real estate policy, etc.
Environment	Waste disposal, energy, pollution, etc.

Table 5
Volume of data sets.

	Training data sets	Testing data sets	Sum
The number of documents	8238	90,574	98,812

model was used for predicting context window words based on target words.

Each word's vector was extracted, and the vector was divided by the number of words in each document. In other words, we calculated a mean feature vector of each complaint in the data sets. With the mean feature vectors in training data sets, we could predict the classification column in the testing data sets. We adjusted parameters and conducted repeated modeling for the best accuracy (Table 6).

Moreover, we employed the random forest method for fitting Word2vec's vectors and classifications. Random forest is an optimal method for classification by constructing a multitude of decision trees (Ho, 1995). The mean feature vectors are used as input variables in random forests. The classification column shows target variables in random forests. Random forests are representative of an ensemble learning method. The key point of the method is randomness, which is robust despite the noise of data.

[1] https://bitbucket.org/eunjeon/mecab-ko.

[2] Three classifications—planning, construction, etc.—were not included in the data analysis. Planning was too ambiguous to prescribe personality, and there were not enough secured cases in the rest of the classifications.

Fig. 1. Examples of Word2Vec.
Source: Modified from Mikolov, Sutskever, Chen, Corrado, and Dean (2013).

Table 6
The Word2Vec model's parameters.

Parameters	Value
Number of words' vector dimensions	5000
Minimum number of words	60
Size of windows	20
Downsample setting	0.0001

2.4. Dynamic topic model

Word2vec was employed in this study for automatic document classification. Word2vec was efficient for predicting types of complaints. However, the methodology used thousands of dimensions, so it was hard to interpret relationships among keywords for each topic. Even if multiple complaints belong to the same category, specific issues in the category vary widely, and issues change every year. Therefore, we employed Dynamic Topic Model (DTM) to capture changes in issues within the automatically categorized complaint category (e.g., Bhadury, Chen, Zhu, & Liu, 2016; Sleeman, Halem, Finin, & Cane, 2017). DTM is a methodology proposed by Blei and Lafferty (2006) that extends the concept of existing Latent Dirichlet Allocation (LDA) to find contiguous occurrences by adding series variables. We used Gensim package[3] based on Python for developing DTM. In addition, DTM in the Gensim package lacks the ability to visually show changes in issues over time, so we have utilized an open source plotter[4] that can show probability changes in keywords in topics in the form of plots. In order to track the issue change, the keywords representing the topics were extracted mainly based on their fluctuation of probability.

3. Results

3.1. Results of automatic classification

The results of automatic classification built on training data sets are as follows. The three major topics identified for both data sets were "environment," "traffic," and "safety." Accuracy of the results was confirmed by cross-validation. Cross-validation is a verification method, dividing training data sets into k-fold sets. In this study, k is 10. One of the 10-fold sets became testing data, and the rest were training data. The accuracy of automatic classification was 70.0% (Table 7).[5]

From 2006 to 2017, three major topics—traffic, the environment, and culture—ranked high in an annual proportion table. The proportion of culture decreased from 2008, and proportions of topics other than the three major ones generally increased (Table 8).

TF(Term Frequency)-IDF(Inverse Document Frequency) is an indicator to supplement the limitations of frequency analysis used in text analysis. As frequency analysis does not indicate relative importance of a specific word in multiple documents, we need to employ TF-IDF to represent relative importance of it. TF-IDF is a TF value multiplied by IDF value. This algorithm helps to identify relatively important words

Table 7
Results of automatic classification.

Training data sets		Testing data sets (predicted)	
Classification	Number of documents	Classification	Number of documents
Environment	2331	Environment	26,525
Traffic	1979	Traffic	22,938
Safety	1117	Culture	14,325
Culture	1028	Safety	6088
Health	855	Health	5006
Housing	785	Housing	4852
Welfare	640	Welfare	4482
Economy	571	Economy	4224
Female	474	Female	3313
Taxes	142	Taxes	908

by weighting words that appear frequently in a particular document rather than words that appear frequently in the entire document set. The following formula will help to clarify this concept. 'W' value means TF-IDF value. '(N/df)', meaning IDF, takes a log for preventing it from going too high score. 'i' means a certain word, and 'j' means a document in a specific order.

$$W_{i,j} = tf_{i,j} * \log\left(\frac{N}{df_i}\right)$$

$tf_{i,j}$ = *number of occurrences of i in j*

df_i = number of documents containing i

N = total number of document

The top 10 words according to this scoring system (by classification) are shown in Table 9. These top words were helpful for understanding the context of each topic.

Each category is represented by specific actors, institutions, and instruments, such as bicycle, bus, subway, taxi for traffic; park, Han river, garbage for environment; library, tourism for culture; road, construction, bridge for safety; restroom, cigarette, smoking, hospital for health; gift card, market, credit card for economy; old people for welfare; flat, building, lease for housing; children, daycare center, education for female; tax, pay, card for tax.

3.2. Dynamic topic model of three main categories

After the automatic classification process of testing data sets using Word2vec with random forests, each complaint was given a category name. Although 90,574 out of 98,812 complaints have been classified by machine learning algorithm, it still is difficult to represent the specific issues in those classifications. Therefore, we put those classified documents into the dynamic topic model (DTM) as input value.

The keywords in the graphs presented in the results were selected from the keywords with the high fluctuation of appearance probability in the topic. Here, fluctuation indicates the degree of changes in the importance of each keyword within a topic over time. The authors then chose a complaint case of each topic in each category that includes as many highly fluctuating keywords that had been selected as possible.

We analyzed the traffic, environment, and culture categories, which

[3] https://radimrehurek.com/gensim/.
[4] https://github.com/llefebure/un-general-debates.
[5] It rounds off numbers to the nearest hundredth.

Table 8
Proportions for topics of interest by year.

Year	Traffic	Environment	Culture	Safety	Health	Economics	Welfare	Housing	Female	Taxes
2006	24.44%	30.42%	28.35%	4.70%	2.24%	2.37%	2.07%	3.25%	1.80%	0.35%
2008	30.59%	29.67%	16.89%	5.11%	4.50%	3.82%	3.66%	2.82%	2.43%	0.50%
2009	26.91%	28.39%	15.93%	5.33%	6.36%	4.55%	4.29%	3.29%	4.25%	0.69%
2010	26.70%	25.68%	13.67%	5.45%	7.82%	5.93%	4.44%	3.33%	6.22%	0.77%
2011	24.90%	25.93%	13.44%	5.89%	8.74%	6.65%	5.10%	3.31%	5.24%	0.80%
2012	26.74%	25.10%	13.27%	5.66%	6.81%	6.42%	6.70%	4.20%	4.39%	0.71%
2013	31.22%	21.66%	11.13%	7.06%	7.12%	5.54%	5.31%	4.48%	5.43%	1.04%
2014	28.26%	21.21%	16.37%	6.12%	7.34%	5.40%	5.62%	3.95%	5.23%	0.50%
2015	28.67%	23.13%	12.18%	8.20%	5.51%	6.96%	4.96%	5.51%	3.95%	0.94%
2016	25.64%	21.62%	14.58%	9.36%	7.29%	6.09%	5.14%	5.80%	3.89%	0.58%
2017	28.32%	22.71%	11.06%	9.00%	7.82%	5.09%	5.68%	6.42%	3.02%	0.88%

Note. As 2007 data was incomplete, it was omitted from our analysis.

Table 9
TF-IDF top 10 words by classification.

	Traffic	TF-IDF	Environment	TF-IDF	Culture	TF-IDF	Safety	TF-IDF	Health	TF-IDF
1	Bicycle	1221.653	Han River	626.9990984	Background	306.7478	Road	147.7919	Restroom	227.9721
2	Bus	1085.828	Park	594.4348225	Library	286.6049	Install	134.6644	Cigarette	210.4678
3	Subway	912.997	Garbage	584.5929449	Citizen	272.8349	Construction	129.0809	Smoking	174.1035
4	Taxi	725.1894	Install	555.9587988	We	271.9514	Safety	119.4413	People	156.9151
5	Install	597.3353	Citizen	488.435885	Culture	269.8243	Accident	97.45439	Ban smoking	156.3705
6	Use	549.0282	Employ	475.2947481	Foreigner	248.8532	Bridge	92.1498	Hospital	141.1441
7	Road	546.8848	Thought	446.9024695	Designation	241.8161	Vehicle	89.23305	Health	140.6117
8	Time	544.3941	Person	424.4659845	Tourism	236.6658	Thought	87.48749	Installation	122.7626
9	Person	537.2111	Utilize	410.6962717	Think	233.4723	Sidewalk	87.18659	Thought	121.4571
10	Vehicle	530.6857	Use	381.5591435	Person	306.7478	Citizen	147.7919	Animal	117.5071

	Economy	TF-IDF	Welfare	TF-IDF	Housing	TF-IDF	Female	TF-IDF	Tax	TF-IDF
1	Market	189.13	Obstacle	198.3678	Flat	109.0949	Children	138.9706	Card	31.84233
2	Traditional	140.1138	Old people	138.7203	Housing	91.01293	School	132.1449	Pay	26.48376
3	Gift Card	109.7854	Volunteer	115.4821	Building	69.0073	Daycare center	119.1529	Tax	22.5154
4	Credit Card	102.533	Welfare	97.87913	Citizen	61.59454	Education	108.8713	Point	21.50334
5	Citizen	85.02789	Senior citizens	97.14574	Development	61.37265	Delivery	102.5882	Issue	18.38394
6	Public official	84.28577	Social	81.55724	Lease	58.42033	Student	95.43314	Use	17.09376
7	Thought	80.41214	Person	74.77106	Area	53.82145	Female	91.47398	Mileage	16.80298
8	Price	80.40577	Thought	68.40279	Install	50.0153	Support	91.40007	Credit	14.14177
9	Enterprise	79.14334	Activity	66.61759	Think	49.9888	Parents	85.40287	Bills	13.8045
10	Comment	78.0366	Job	64.26868	Business	49.75334	Home	85.2857	Fees	13.76684

accounted for the largest percentage of complaints that were automatically classified, as the subjects of our DTM. The DTM adds time series variables that traditional topic modeling could not deal with, allowing you to see trends in specific issues within a topic. By looking at the changes in the three internal issues that have had the highest demand for citizens over the past decade, we could predict future civil demand. Like LDA, DTM requires the researcher to set the number of topics in a hyperparameter. For example, if the number of topics is five in the environment category, DTM produces five topics as an output. In the process, we worked with the Seoul Digital Foundation and the Seoul Big Data Policy Officer to derive the optimal number of topics that best reveal the issue. Finally, we configure the number of topic's hyperparameters (k = 3) for each category.

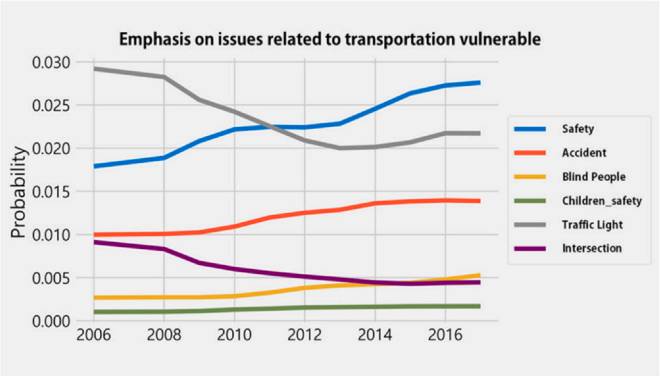

Fig. 2. Topic 1 in the traffic category.

3.2.1. Traffic

This topic is one of the civil complaint issues related to people who are alienated from transportation services. Over time, voices of 'safety' and 'accidents' are increasing. Social interest in traffic safety is appearing as a specific keyword such as 'visually impaired' and 'traffic lights'. These results imply that citizens' demands for traffic safety are moving towards the socially underprivileged such as blind people and children (Fig. 2).

Complaint example*: *2009 case

For the visually impaired, the sidewalk pedestrian block, 'Braille block,' is installed in the center of the pedestrian road, so it is convenient to walk on the sidewalk, and there are traffic light voice guides on both sides of the crosswalk to replace the traffic light. However, braille blocks stop at the crosswalk floor of the road… (continued)…

This topic is one of the civil complaint issues related to the distribution of exclusive city bus lanes and transfer service. In Seoul, a new exclusive bus lane was established by former mayor and later President Lee, Myung-bak. Dissatisfaction with the reserved lanes is gradually

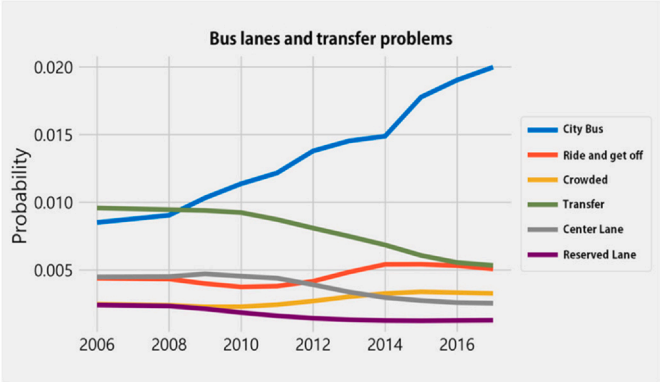

Fig. 3. Topic 2 in the traffic category.

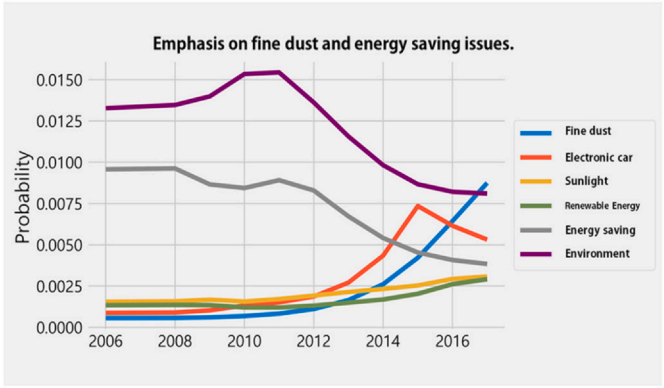

Fig. 5. Topic 1 in the environment category.

decreasing, but citizens' dissatisfaction with the complexity of the transit is expressed continuously. These results are due to the over-population of Seoul, which leads to the necessity of flexible operation of bus routes based on demand forecasting (Fig. 3).

Complaint example*: *2012 case

Pedestrian boarding of all city buses in Korea is done at the front door and pedestrian getting off at the back door. But I suggest getting on at the back door and getting off at the front door. Of course, there can be great confusion at the first attempt. Therefore, we will make a pilot route or a pilot bus to check the efficiency. And if it is found to be efficient, you can adopt this new policy… (continued)…

This topic tells the story of caring for the elderly and pregnant women in terms of the seating problem in the subway cabin. According to the graph, social interest in 'pregnant women' has been rising rapidly since 2014. This interest can be interpreted as a social issue of the birth rate in Korea. This contextual context has helped to enhance social support for 'pregnant women first' policy on the subway (Fig. 4).

Complaint example*: *2017 case

I would like to suggest a solution for problems with the designated seat for pregnant women in the subway cabin. Since there are always many people in the subway, most people sit in the caring (designated) seats for pregnant women. Even if a pregnant woman is given a designated seat in the subway cabin, most people are concentrating on smartphones. Even if you find a pregnant woman, you can't help her without knowing it… (continued) …

3.2.2. Environment

This topic indicates the story of 'fine dust' and 'energy saving' among environmental issues. Seoul is a representative city of the Republic of Korea, inhabited by about 10 million citizens, and fine dust is an important issue for civil health. As time goes by, the voice of 'fine dust' is increasing. Among various fine dust reduction techniques, citizens' attention is particularly paid to renewable energy (Fig. 5).

Complaint example*: *2017 case

Detailed explanations of eco-friendly renewable energy facilities such as solar, geothermal, and other renewable generators installed at Seoul City Hall helped us to learn a lot. In particular, I heard that the nation's largest geothermal facility installed in the city hall is more efficient than the original plan, so there is much energy left even after supplying air conditioning and heating for the Citizens' Office, restaurants and the Seoul Library… (continued)…

This topic demonstrates the story of 'garbage' and 'separate collection' among environmental issues. The garbage problem has always been an important issue, but the issue of garbage bin pickup seems to be spreading with the increase of quality sensitivity in citizen's life. This problem also needs to be supplemented by the management system to make waste discharge smoother at the administrative level (Fig. 6).

Complaint example*: *2013 case

Many things are mentioned because of various waste and recycling issues. But I would like to suggest food waste among them. First we need to clarify the distinction between food waste and garbage. We know lots of food waste, for example, fruit peels, vegetable peels, bones such as fish and meat, oils and burnt foods, leftovers, and seeds. There is a problem because non-food waste is mixed up… (continued)…

This topic addresses the need for alternative spaces due to the reduction of green space among environmental issues. In urban development, 'children' and 'environment' are important issues. In the reorganization of the city, citizens are stressing the need for parks. This is interpreted as the intention to solve the problem of children's play space and to improve the quality of life of citizens (Fig. 7).

Complaint example*: *2012 case

There are many sports facilities on hiking trails and parks to improve the physical strength and leisure of Seoul citizens, but most of these sports facilities are for adults.

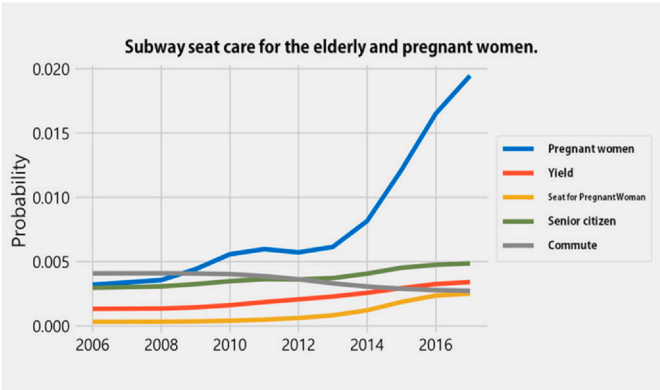

Fig. 4. Topics 3 in the traffic category.

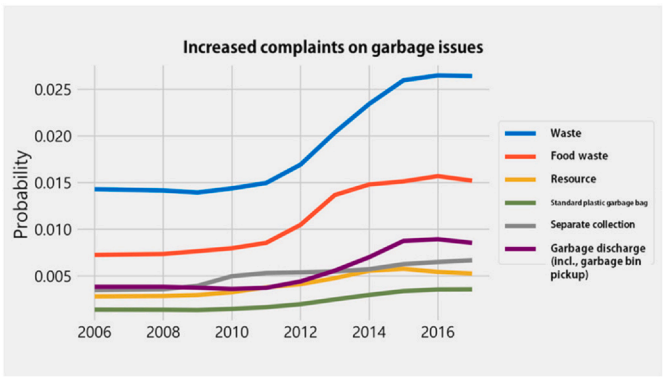

Fig. 6. Topic 2 in the environment category.

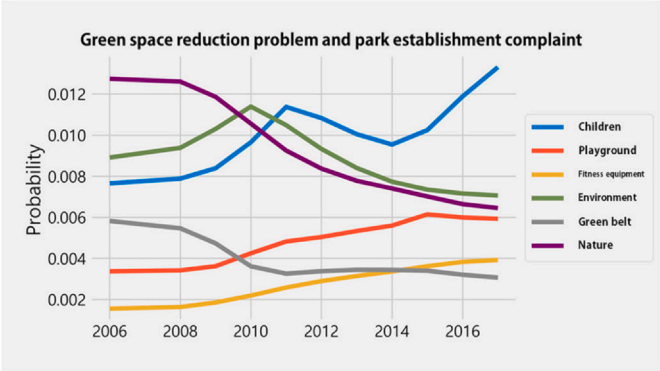

Fig. 7. Topic 3 in the environment category.

Therefore, it is not possible for children to play sports by using the sports facilities installed on hiking trails and parks… (continued)…

3.2.3. Culture

This topic expresses the agenda for foreign tourism in a cultural context. Seoul is a representative tourist destination of Korea. The paradigm of tourist information has changed due to smartphones and travelers' altering patterns from group to individual tours (Fig. 8).

Complaint example*: *2011 case

More and more foreigners are coming to Seoul. Therefore, I propose a package to organize the Seoul Experience Program for foreigners entering Gimpo International Airport in cooperation with travel agencies. Course 1: Seoul Tourism Course using Seoul City Bus… (continued)…

At the cultural level, this topic shows citizens' complaints about cultural properties and history. 'Cultural property' is an important means of showing the history of a country, and Gwanghwamun has an important implication in Seoul tourism. In this context, it is necessary to reinforce the image of Korean cultural assets by emphasizing the need to improve the Gwanghwamun area (Fig. 9).

Complaint example*: *2011 case

Gwanghwamun in Seoul has many cultural assets including Deoksugung Palace, Gyeongbokgung Palace, Changgyeonggung Palace, and Gyeonghuigung Palace.

Why don't you make a new tour of these old palaces? If you run a wagon and many foreigners can go to various palaces around Gwanghwamun, it should be a good tourist product… (continued)…

This topic talks about 'Korean Wave' and 'Branding' at the cultural level. The Korean wave has spread to various countries through K-pop, and tourists who are visiting Korea are wearing Korean traditional clothes called 'Hanbok' to increase their satisfaction. Therefore, citizens are proposing the need for more specific tourism products in order to inform the world about the positive image of Korea (Fig. 10).

Complaint example*: *2009 case

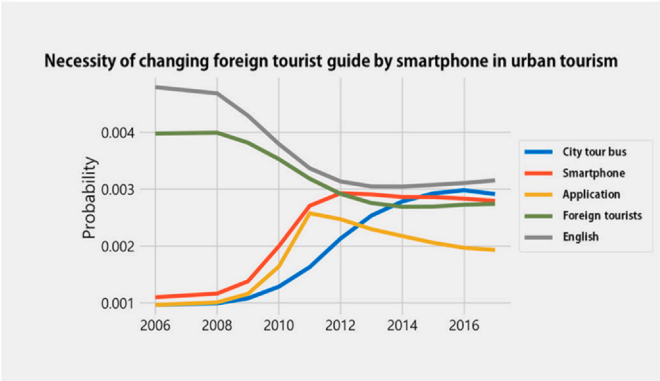

Fig. 8. Topic 1 in the culture category.

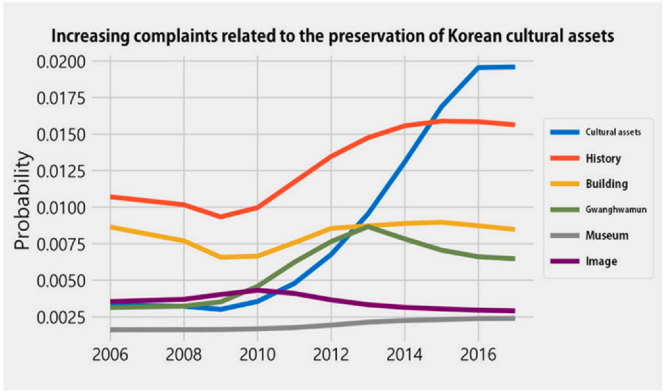

Fig. 9. Topic 2 in the culture category.

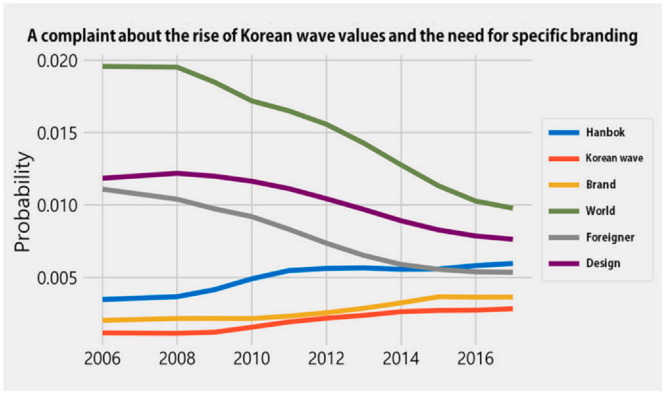

Fig. 10. Topic 3 in the culture category.

Hanbok is a culture in which our spirit permeates. How much would you like to see foreigners wearing Hanbok? I think it is a way not only to promote a tourist culture product but also to protect our culture and promote Seoul… (continued)…

4. Conclusion

4.1. Classification corresponding to social issues

This study is conducted to improve the quality of life for citizens. Citizens' voices were necessary big data to analyze. Each opinion was related to at least one social issue. In other words, civil demands have policy implications. For a smart city, we suggest that the government should manage big data from citizens' inputs. One of the key factors in establishing a smart city strategy for solving urban problems is the establishment of a citizen-centered innovation system; thus, citizens will move away from the majority-led approach promoted in the traditional policymaking processes. It is necessary to prioritize the areas that each citizen considers important to herself/himself and to promote policies to address issues from citizens' requests one by one. Table 10 shows the details of 10 classifications including the topics that were an outcome of preliminary study to the current research.

4.2. Composition of citizen inputs

A total of 24,548 people reported 98,812 complaints and other type of inquiries. Further, 7461 people posted more than two in 11 years, and 57 people filed more than 200. In addition, 57 heavy users submitted inputs as many as a total of 24,259. Hopefully, this research may contribute to the immediacy of and enhanced satisfaction from the conventional civil complaints processing procedures.

Table 10
Classifications of citizens' demands and social issues.

Classification	Civil demands	Related issues
Health	Public restrooms, healthcare services, Mental health, vaccinations, Public hygiene, smoking in public places	Improved hygiene in public restrooms and convenience Diffusion of social awareness about smoking
Economy	Payment fee, consumer price, Activation of the traditional market, Economic policy, creating jobs for youth, Citizen participation	Slow economic growth and income inequality Fee burdens A decrease in the productive population
Traffic	Improvements in the transportation system, pedestrian safety, Subway facilities, parking, Public transportation information, Bus station information	Inconveniences caused by the increased use of public transportation Increase of traffic in the city center, the lack of parking spaces, and infrastructure on the bike path
Culture	Improvements for the convenience of foreign tourists, Traditional culture experience, A library, Citizens' Square, Cultural symbolism	Surcharges for foreign tourists Cultural heritage safety management Lack of tourism information and programs
Welfare	Welfare for the aged, volunteer participation, basic livelihood security, Welfare policy, welfare for the disabled, Unemployment benefits	Increasing welfare costs because of aging Positive discrimination Lonely deaths among senior citizens
Taxes	Tax payment method and exemption, Earning points for payments, Insurance, delinquent collection	Collecting arrears on bad debts
Safety	Walking safety, natural disasters, Road maintenance work, firefighting, Emergency reporting	Walking and road safety Rising disaster risk Increased anxiety about natural damage and collapse of buildings
Females	Government support, multi-child benefits, Schools and childcare facilities, programs supporting women's health, school violence	Deepening social conflicts resulting from the increase in multicultural families and foreigners Nuclear family Gender disparity in economic activity participation
Housing	Rental housing, low-income housing, Parking problems, building safety, Redevelopment project, illegal banners	Shortage of houses Loss of community awareness and social conflicts
Environment	Waste disposal, green spaces in Seoul, Energy problem, eco-friendly transportation, Road facilities	Air pollution Expansion of eco-friendly energy use Garbage disposal and environmental pollution

5. Discussion

5.1. Feasible smart city solution

OTMI changed its name to Democracy Seoul[6] at the end of 2017 and started a new service. Democracy Seoul has been developed a digital democracy platform for collecting opinions from citizens and proposing policies. While OTMI simply focused on collecting and posting complaints, Democracy Seoul encourages more active participation from citizens. In OTMI, when citizens post complaints, the person in charge of handling the complaints read the complaints and delivered them to the relevant departments, and the department in charge responded to the complaints. OTMI's civil service handling policy had the trouble of having to read and confirm each of the complaints, and only played the role of simply connecting the civil service and the responsible department. On the other hands, Democracy Seoul offers a user experience for active participation of citizens. Like and comment function was strengthened. Civil petitions with more than 50 likes were answered by the relevant department, and if more than 100 likes were received, they were moved to a separate bulletin board to make it easier for more citizens to read the articles. Finally, the Mayor of Seoul will answer the complaints for more than 1000 likes.

As we see, the emergence of a more open civil service handling system will result in more citizens' participation in the civil service. Therefore, our proposed automatic civil service classification model is worth processing more civil service text quickly in the future. For smart city, this study is meaningful in that it proposed a prototype to efficiently handle civil complaints. For citizen, citizens' complaints can be quickly responded to, thus improving citizen satisfaction. For policy, new policy can be developed, and problematic policy can be revised through this study. For application of big data analysis, this study demonstrates a good example of how data science can be applied to enhance citizens' life.

5.2. Implications: automatic classification based on machine learning

Automatic classification to ensure accuracy in recording the voices of citizens based on machine learning/artificial intelligence (AI) would enhance the quality of responsive governance. It would be useful to categorize complaints frequently and automatically link them to related frequently asked questions (FAQ) contents for the users' convenience. We suggest that the government should provide data-driven insights for its citizens to save precious time and expense for reporting their complaints.

Automatic classification can dramatically shorten the time for classification work, previously done by civil servants. In addition, DTM can be used in automatically classified civil documents to identify the gravity of the issues over time. If a small number of people in charge have read, classified, and identified problems included in the complaints, they will now be able to effectively handle the complaints.

5.3. Suggestions for future research

Subsequent research suggests that noun-oriented word analysis should be extended to modifier and modified relationships, like adjective + noun and adverb + verb. Furthermore, if a specific word belonging to a cluster in the dynamic topic model is detected, a system that displays the FAQ contents immediately to the user would benefit her/him better. In addition, the 13 administrative fields that were applied as learning data criteria in this study were defined by government (MOIS & NIA, 2010). There are three areas where data do not exist or are extremely rare. Thus, future studies may contribute to data-based administrative innovation by characterizing civil voices through

[6] https://democracy.seoul.go.kr/.

unsupervised learning instead of the supervised learning approach applied in this study.

CRediT authorship contribution statement

Byungjun Kim: Conceptualization, Methodology, Software. **Minjoo Yoo**: Data curation, Validation. **Keon Chul Park**: Resources, Data curation. **Kyeo Re Lee**: Visualization, Data curation. **Jang Hyun Kim**: Writing - Review and Editing, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by the Seoul Digital Foundation grant funded by the Seoul Metropolitan Government (Seoul Smart City Policy Research).

References

Abella, A., Ortiz-de-Urbina-Criado, M., & De-Pablos-Heredero, C. (2017). A model for the analysis of data-driven innovation and value generation in smart cities' ecosystems. *Cities, 64*, 47–53.

Bhadury, A., Chen, J., Zhu, J., & Liu, S. (2016, April). Scaling up dynamic topic models. In Proceedings of the 25th international conference on world wide web (pp. 381-390). International World Wide Web Conferences Steering Committee.

Blei, D. M., & Lafferty, J. D. (2006, June). Dynamic topic models. In Proceedings of the 23rd international conference on machine learning (pp. 113-120). ACM.

Cao, Q. H., Giyyarpuram, M., Farahbakhsh, R., & Crespi, N. (2020). Policy-based usage control for a trustworthy data sharing platform in smart cities. *Future Generation Computer Systems, 107*, 998–1010.

Greene, D., & Cross, J. P. (2017). Exploring the Political Agenda of the European Parliament Using a Dynamic Topic Modeling Approach. *Political Analysis, 25*(1), 77–94.

Ha, T., Beijnon, B., Kim, S., Lee, S., & Kim, J. H. (2017). Examining user perceptions of smartwatch through dynamic topic modeling. *Telematics and Informatics, 34*(7), 1262–1273.

Hagen, L., Harrison, T. M., Uzuner, Ö., May, W., Fake, T., & Katragadda, S. (2016). E-petition popularity: Do linguistic and semantic factors matter? *Government Information Quarterly, 33*(4), 783–795.

Hardaya, I. S., Dhini, A., & Surjandari, I. (2017, October). Application of text mining for classification of community complaints and proposals. In 2017 3rd international conference on science in information technology (ICSITech) (pp. 144-149). IEEE.

Ho, T. K. (1995, August). Random decision forests. In Proceedings of 3rd international conference on document analysis and recognition (Vol. 1, pp. 278-282). IEEE.

Kim, C. D. (2010). Oasis of ten million's imagination in Seoul. *Korean Association of Governmental Studies, 2010*(12), 3–18.

Krishna, G. J., Ravi, V., Reddy, B. V., Zaheeruddin, M., Jaiswal, H., Teja, P. S. R., & Gavval, R. (2019, October). Sentiment classification of Indian banks' customer complaints. In TENCON 2019-2019 IEEE region 10 conference (TENCON) (pp. 429-434). IEEE.

Linton, M., Teo, E. G. S., Bommes, E., Chen, C. Y., & Härdle, W. K. (2017). *Dynamic topic modelling for cryptocurrency community forums. In Applied quantitative finance (pp. 355–372)*Berlin, Heidelberg: Springer.

Luo, J., Qiu, Z., Xie, G., Feng, J., Hu, J., & Zhang, X. (2018, October). Research on civic hotline complaint text classification model based on word2vec. In 2018 International conference on cyber-enabled distributed computing and knowledge discovery (CyberC) (pp. 180-1803). IEEE.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In Advances in neural information processing systems (pp. 3111-3119).

MOIS (Ministry of Public Administration and Security) & NIA (National Information Society Agency) (2010). A guide for designing functions for informatization support system. Retrieved Dec 10, 2018 from http://www.klid.or.kr/section/content/content.html?PID=task1a14.

Ruijer, E., Grimmelikhuijsen, S., & Meijer, A. (2017). Open data for democracy: Developing a theoretical framework for open data use. *Government Information Quarterly, 34*(1), 45–52.

Sano, Y., Yamaguchi, K., & Mine, T. (2015). Automatic classification of complaint reports about city park. *Information Engineering Express, 1*(4), 119–130.

Sinha, M., Guha, S., Varma, P., Mukherjee, T., & Mannarswamy, S. (2019, January). My city, my voice: Listening to the citizen views from web sources. In Proceedings of the ACM India joint international conference on data science and management of data (pp. 52-60). ACM.

Sleeman, J., Halem, M., Finin, T., & Cane, M. (2017, December). Discovering scientific influence using cross-domain dynamic topic modeling. In 2017 IEEE international conference on big data (big data) (pp. 1325-1332). IEEE.

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.

Vargas-Calderón, V., & Camargo, J. E. (2019). Characterization of citizens using word2vec and latent topic analysis in a large set of tweets. *Cities, 92*, 187–196.

Wollard, C. W., III (2017). *Responsiveness to citizen input: Topic model analysis of public comments to the San Francisco police commission*.

Yao, Y., Li, X., Liu, X., Liu, P., Liang, Z., Zhang, J., & Mai, K. (2017). Sensing spatial distribution of urban land use by integrating points-of-interest and Google Word2Vec model. *International Journal of Geographical Information Science, 31*(4), 825–848.