Information Processing and Management 58 (2021) 102410

Contents lists available at ScienceDirect

Information Processing and Management

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

Automatic classification of citizen requests for transportation using deep learning: Case study from Boston city

Narang Kim, Soongoo Hong *

Department of Management Information Systems, Dong-A University, Busan, Republic of Korea

ARTICLE INFO

Keywords:
Citizen requests
Unstructured data
Convolutional neural network
Imbalanced data

ABSTRACT

Responding to requests from citizens is an essential administrative service that affects the daily life of people. The drastic increase in the volume of citizen requests in recent years has necessitated on-going studies on the automatic classification of citizen requests due to the time, effort, and misclassification errors involved in manual classification. Even though there have been prior studies that have analyzed citizen requests according to topic and frequency, they ignore the complicated and dynamic nature of such a dataset. Using a deep learning algorithm, this study proposes an automatic classification model for unstructured data by using transportation-related citizen requests from January 15th, 2016 until November 7th, 2018 of the City of Boston, USA, as an example. A combination of unsupervised and supervised learning was applied to the data. To address the issue of imbalance in data, this study also considered an equalization method. Five stepwise models were applied to increase the classification accuracy for the unstructured data. The final model uses achieved a classification accuracy of 90%. The model proposed in this study is expected to be generalized for classification of other citizen requests or unstructured text data on specific topics in the future. Moreover, this study has substantial academic importance given that it has proven diverse machine learning-related theories through their application to unstructured data.

1. Introduction

Governments have often reinvented the wheel on documentation and administration. In a 2017 survey of US state and local officials, it was reported that 53% of the officials faced difficulty in completing their work in 35 to 40 h, mainly due to the burden of excessive paperwork (Barrett & Greene, 2016). To address the inefficiency, governments around the world are increasing adopting the latest technologies, such as big data and artificial intelligence (AI), to automate the documentation process and ensure that the officials focus on the more essential tasks at hand (Agostino & Arnaboldi, 2016). For instance, Swedish Tax Agency adopted a chatbot that handles approximately 15,000 queries on tax returns an hour, making services more accessible to citizens as well as improving the efficiency of the government staff (AI Innovaion of Sweden, 2019).

Tending to citizens request is not only an essential part of governance, the requests themselves can be very useful and timely in the analysis of diverse political, social, and economic issues raised in the society (Lee & Choi, 2020). However, governments often fail to respond to them properly; they face many difficulties in properly handling the large number of requests received and in tending to public concerns (Jeong, Lee, & Hong, 2017). This is because citizen requests accepted by public institutions are more likely to be in the

* Corresponding author.
E-mail address: shong@dau.ac.kr (S. Hong).

https://doi.org/10.1016/j.ipm.2020.102410
Received 26 February 2020; Received in revised form 30 August 2020; Accepted 9 October 2020
Available online 9 November 2020

form of unstructured data; therefore, any mechanical processing, such as automatic classification, is bound to be relatively difficult. Usually, the staff in charge of these requests has to read the requests manually and categorize them. This is not only cost- and time-consuming, it also ties up precious human resources away from the more essential governance tasks at hand; thus, many governments are interested in developing automatic classification systems for such work.

Prior studies related to citizen requests have simply focused on identifying the frequency and topic of the requests (Cho, Choi, Na, Moon, & Kim, 2018). However, the unstructured nature of citizen request data has rarely been considered (Kim, Kim, Kim, & Lim, 2018a). A simple application of deep learning techniques to improve the classification accuracy is not successful in such a case as citizen requests are often concentrated around a specific topic, leading to an *imbalance in data (as explained later in text)*. An increase in the number of subject classifications further magnifies this issue (Kim, Yang, & Kim, 2007). Therefore, the present study tries to overcome the aforementioned issues by proposing an automatic classification model for unstructured data with acceptable accuracy. To do so, we examine transportation-related citizen requests received in Boston City, Massachusetts, USA from January 15th, 2016 until November 7th, 2018. The study also proposes the implications for categorizing unstructured data with a provision for diverse problems, including their solutions, that occur in the process of the case study.

The following section discusses related works and describes automatic classification of text based on machine learning. Next, we describe the process and results of the proposed automatic classification taking the examples of transportation-related citizen requests in Boston, Massachusetts. The last section summarizes the study and discusses its implications, research contributions, and limitations.

2. Literature review and model explanation

2.1. Machine learning and government affairs

In the past, many studies have applied machine learning methods to improve the efficiency of government affairs (Wirtz & Müller, 2019) or to predict an index of governance. Hagen (2018) proposed a framework to train and validate Latent Dirichlet Allocation (LDA) to analyze contents of e-petitions, and further showed a strong association with corresponding social events. Ryu, Hong, Lee, and Kim (2018) conducted keyword analysis and text mining on citizen request data in Busan City, South Korea, and examined the changes in these requests through an examination of the importance of keywords related to buses. Eshleman and Yang (2014) investigated the visual-spatial relationship between the citizen dissatisfaction recorded in the 311 Case Database and the sentimental aspect of Twitter posts, and reported the implications of relative happiness of cities and surrounding regions by modeling the sentimental characteristics of five metropolitan cities worldwide. Dalianis, Sjöbergh, and Sneiders (2011) analyzed two methods, manually created text patterns and machine learning-based methods, to detect important messages among a set of large number of emails sent by citizens to the Swedish Social Insurance Agency. Fu and Lee (2012) applied support vector machines (SVMs) to improve the classification accuracy for indistinguishable official documents in China. Ku and Leroy (2014) developed a decision support system where natural language processing techniques, similarity measures, and machine learning techniques were combined sequentially to support crime analysis and classify similar crimes to enhance the efficiency of law enforcement agencies. More recently, Lima and Delen (2020) tried to predict the corruption level of countries using combined machine learning algorithms including random forest, support vector machine, and artificial neural networks. Singh, Dwivedi, Kahlon, Pathania, and Sawhney (2020) suggested a quantitative forecasting technique for number of seats of Punjab Assembly election based on Twitter data in India.

2.2. Automatic classification of text based on machine learning

Text classification is defined as the establishment of a classification model that uses learning data and decides on the class of the given text. Machine learning technology is frequently used for text classification, including Naïve Bayes, SVMs, random forests, and neural networks (Montebruno, Bennett, Smith, & Lieshout, 2020; Zheng, Cai, Chen, & Rijke, 2020).

In a recent study on text classification through a conventional machine learning method, Bouazizi and Ohtsuki (2018) conducted a multi-category sentiment classification using Twitter data as the subject. Kim, Lee, Ryu, and Kim (2018b) investigated sentiment classification on Naïve Bayes using Wikipedia data as the subject. For marketing applications, Hartmann, Huppertz, Schamp, and Heitmann (2019) conducted a comparative analysis among the various classification schemes (SVM, Naïve Bayes, and random forest) using unstructured social media data.

Deep learning, which is a representative scheme of machine learning, has emerged as an effective solution for diverse problems in text mining, including classification and clustering of documents, document summary, web mining, and sentiment analysis. An increasing number of studies are using convolutional neural networks (CNNs) or recurrent neural networks (RNNs), both based on deep learning, for such purposes. CNN and RNN are supposed to be effective at extracting position-invariant features and modeling units in sequence, respectively. The state-of-the-art technology for many natural language processing (NLP) tasks often switches between CNNs and RNNs owing to the competition between these two deep neural networks (DNNs). Yin, Kann, Yu, and Schütze (2017) reported the first systematic comparison of CNNs and RNNs in a wide range of representative NLP tasks, aiming to provide basic guidance for DNN selection. Relevant classification research using deep learning includes Banerjee et al. (2018), who applied CNN and RNN to a radiology text report, and (Karakuş, Talo, Hallaç, & Aydin, 2018), who conducted sentiment analysis of movie reviews on a website using deep learning.

Different from the structure used in general, multi-layer perceptron, CNNs, which were first suggested by Le Cun, Bottou, Bengio, and Haffner (1998), consist of a convolutional layer and a pooling layer (Fig. 1).

In the convolutional layer, a multiplication operation is performed by multiplying a weighted region of an image expressed in the

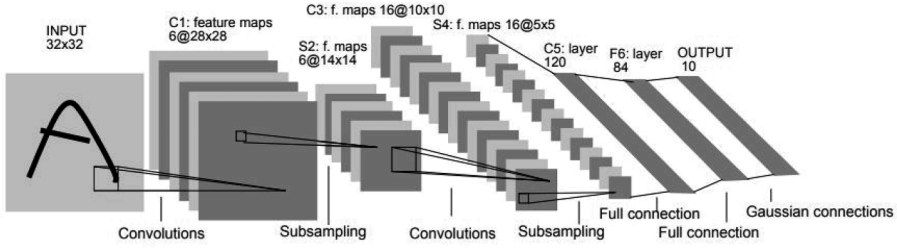

Fig. 1. LeNet-5 CNN design. (*Source:* Le Cun et al., 1998).

form of a weighted matrix with the corresponding element-wise while sequentially moving a filter having a smaller size than a corresponding image in a weighted matrix. In addition, in the pooling layer, a representative value such as a maximum value or an average value is sequentially extracted for each specific size region for the values obtained from the composite product. The convolutional layer and the pooling layer may be repeated several times in pairs. Repeated pairs of pooling layers followed by a convolutional layer can also be used. Through this operation, useful features are hierarchically extracted from the input image, and the extracted features are utilized to classify input data into a target class through one or more fully connected layers. The CNN model, which was originally devised for computer vision, has shown excellent performance in semantic parsing, search query retrieval, sentence modeling, and processing of other conventional natural languages. Kim (2014) proposed a CNN that uses vector expression of pre-trained words using the Word2Vec algorithm for classification at the sentence level. CNNs produce higher classification accuracy compared to logistic regressions (LRs) and SVMs and have been evaluated as machine learning tools that show excellent performance in diverse areas, including document classification and processing of natural language and text mining (Zhang & Wallace, 2017). The present study applies the CNN model by Kim (2014) for the automatic classification of transportation-related citizen requests in Boston (Fig. 2)

2.3. Imbalanced data

As the number of subject classification categories increase, collection of data becomes more difficult, which consequently leads to an imbalance in the data. This problem occurs when the amount of data in a specific category is extremely different from the data in another category. In most machine learning algorithms, learning proceeds under the assumption that the proportion of each category is similar. However, many machine learning algorithms encounter the problem of imbalanced data when dealing with real-world cases (Kim et al., 2007). This results in diminished performance of classifiers that use machine learning algorithms. Solutions to this problem can be largely divided into data-level approaches, algorithm-level approaches, and ensemble approaches. Data-level approaches include undersampling, oversampling, and a method of controlling the balance of the data by using both (Dubey, Zhou, Wang, Thompson, & Ye, 2014). Algorithm-level approaches involve using the error function of the existing machine learning algorithm. Several methods are used to handle oversampling and undersampling problems.

Adaptive synthetic is an imbalanced learning method using a weighted distribution to strengthen the decision boundary toward the minority class data. In kernel density estimation, methods have been applied to estimate the probability density distribution of minority class to sample additional minority data samples (He, Bai, Garcia, & Li, 2008). Synthetic minority oversampling technique (SMOTE) is an equalization method for imbalanced data. It does not require any professional knowledge in the application field, and there is no limitation on the classifier. SMOTE creates data in a category that has a small proportion. First, a sample of data with fewer classification members is selected, and the k-nearest neighbor of the sample is searched (Seo, 2014). The difference between the baseline sample and the k-nearest neighbor is obtained. A random number between zero and one is multiplied with the difference to be added to the original sample (Seo, 2017). The newly created sample is added to the training data. Instead of duplicating the sample similar to oversampling, SMOTE compensates the overfitting problem by creating new cases that are similar to the minority group. The edited nearest neighbor rule (ENN) was proposed by Wilson (1972); this procedure tests each sample using the k-NN rule with the rest of the data. Although the over-sampling method can balance the data distribution, it has some problems. For example, some majority class samples might overlap with the minority class samples, such that they cannot be well distinguished by the classifier. To address this problem, SMOTE can be combined with ENN, known as SMOTE + ENN. First, the training data are oversampled by using SMOTE. Second, the three nearest neighbors of each sample are found in the training data. Third, the misclassified samples are removed to produce cleaner data. In this manner, we can balance the data distribution, and the boundaries between classes become clearer (Lu, Huang, Zhao, & Zhang, 2019).

3. Automatic classification of citizen requests through deep learning

In this study, automatic classification using a CNN was conducted for the prompt processing of a large number of citizen requests. Citizen requests data has a particularly large number of categories compared to other types of unstructured data, and these categories are often mixed in practice. Furthermore, the data is often imbalanced, which makes quantification difficult. To address these shortcomings of citizen request data and to search for an optimal automatic classification scheme, this study applied the CNN model by Kim (2014) to transportation-related citizen requests in Boston between 2016 and 2018.

For the automatic classification of the citizen requests, machine learning should be conducted by first defining optimal categories

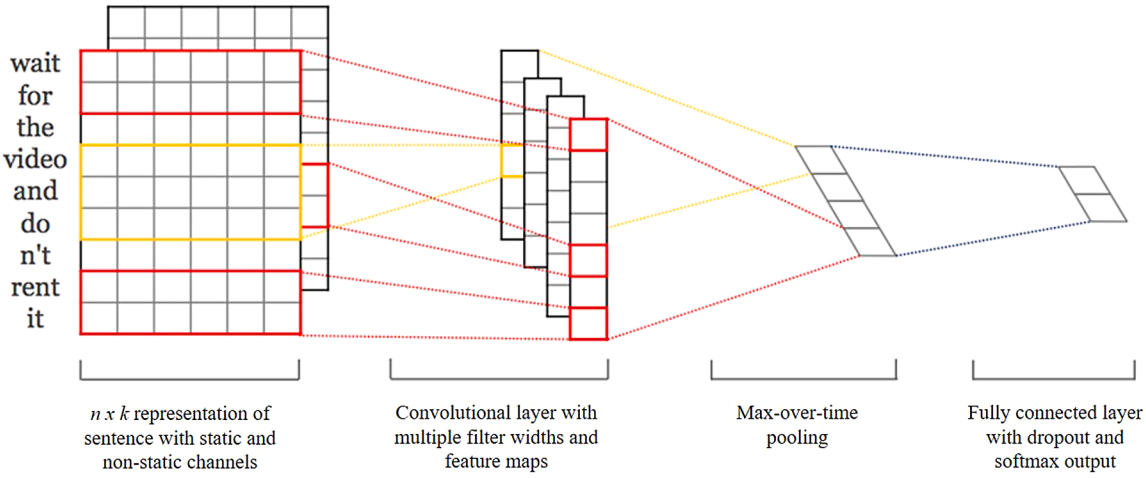

Fig. 2. Convolutional neural networks for sentence classification (Kim, 2014).

and creating labeled training data. Machine learning is divided into supervised learning and unsupervised learning depending on the availability of labeled data. Unsupervised learning is used for clustering unlabeled datasets. Supervised learning produces better performance than unsupervised learning because it conducts automatic classification by appropriately learning the data in the corresponding domains. In this study, unsupervised learning of clustering was first used for category optimization, and then supervised learning was used for the automatic classification. Prior to the measurements, the data were preprocessed using text cleaning, vectorization, and undersampling and oversampling. The flowchart representing the methodology used in the experiment is shown in Fig. 3. As a result of preprocessing and optimization, each step in the experiment uses, for the most part, the previous step's result in training and testing data. Through such a serial process, the accuracy of automatic classification using CNN was improved by merging the categories that were optimized through clustering, which resolves the overfitting and imbalanced data issues.

To test the deep learning and CNN both which consume high capacity of hardware, this study configured NVIDA GPU 1050 8 GB for GPU and 16 GB of Memory. All analyses were conducted using Python 3.6 with Tensorflow.

3.1. *Data collection and preprocessing*

The first step in the experiment and the proposed method involved data collection and preprocessing. This first step is important to ensure that the text data is clean and ready for further processing. To improve the accuracy and performance of the model, several methodologies were used to handle data preprocessing in the experiment. The original data weres obtained from Boston's public data website (https://data.boston.gov/dataset/vision-zero-entry) in Fig. 4 For text preprocessing, the duplicated data and data without value (N.A.) was removed, the text was cleaned, and checked for typos.

The result of the preprocessing is shown in Table 1. Our data were collected twice to understand the impact of data size on performance. First, the data were tested first and a more comprehensive dataset was analyzed, and the results were compared for

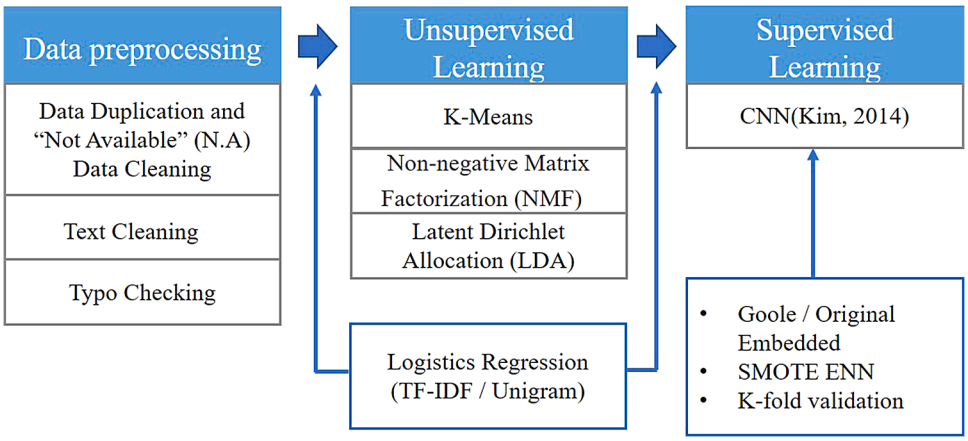

Fig. 3. Proposed classification sequence.

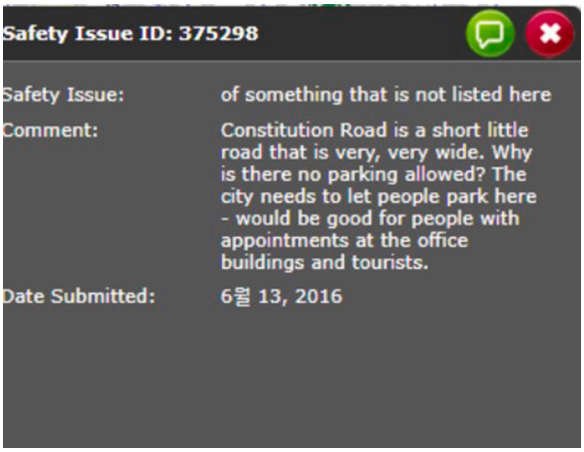

Fig. 4. Snapshot of a citizen's request on Boston's public website.

comparison.

The original data, which include transportation-related citizen requests in Boston city between 2016 and 2018, consists of approximately 10,000 data points and 21 categories. The descriptive statics are as listed in Table 2. An examination of the categories after the data preprocessing revealed that similar contents were classified into different categories, which indicates the necessity to merge the existing categories, as shown in Fig. 5.

Therefore, we merged the similar categories into one category through peer evaluation to avoid misclassifications, statistical errors, and delays in processing of citizen requests. Two peers were hired and asked to re-classify the original classifications, respectively. After the work, an author confirmed the outcomes. As the duplications in categories were obvious, no peer suggested incongruent classifications. After the peer evaluation, the original categories listed in Fig. 5 were re-categorized into 14 categories in Fig. 6. For example, the category of "*it's hard for people to see each other*" was merged into the category "*it's hard to see/low visibility*."

3.2. Logistic regression

The second step in the experiment was to run the LR algorithm to measure the accuracy of the categorization. The logistic F1-score on average (0.54) was low, as shown in Fig. 7. Therefore, to improve the accuracy, TF-IDF text preprocessing was used. Term Frequency-Inverse Document Frequency (TF-IDF) is a method used to calculate the weight of a word or phrase in a document in the field of natural language processing (Kim, Seo, Cho, & Kang, 2019). In addition, the Uni-gram model was utilized in the experiment to understand and obtain detailed information on the meaning of each word in the sentence. An n-gram is a type of probabilistic language model for predicting the next item in a sequence in the form of an (n1)-order Markov model. N-gram based techniques are predominant in modern NLP and its applications. Traditional n-grams are sequences of elements as they appear in texts (Siagian & Aritsugi, 2020). Specific words have a meaning when combined with other words (i.e., word groups) and studying them according to their word groups is considered to be more helpful in understanding their meaning.

After the manual classification, misclassification was still found to be an issue in the dataset. Fig. 8 shows the misclassification in the dataset using LR. This indicates that the categories were mixed-use, given that many people have classified these requests over several years and the initial categories failed to properly classify the issues that later became increasingly diverse following circumstantial changes.

3.3. Unsupervised learning

The third step involved unsupervised learning, which was used to cluster the data for another categorization optimization. This was done because accurate categorization is necessary to check for any errors in manual classification made in the second step. The n-gram data and TF-IDF preprocessing were used as the input for unsupervised learning. In the experiment, unsupervised learning consisted of three methods: k-means, non-negative matrix factorization (NMF), and latent Dirichlet allocation (LDA). We compared these methods in terms of their categorization optimization results.

Based on the training results, which are shown in Table 3, the LDA model was found to be more stable than the other methods,

Table 1
Data collection.

Year	Jan. 15, 2016–April 13, 2017	Jan. 15, 2016–Nov. 7, 2018
Before preprocessing	7166	9343
After preprocessing	5844	7184

Table 2
Description of the dataset.

Description	Values
Size of training and testing dataset	10,000 & 6000
Classes	21 categories
Maximum message length	133
Average message length	47.39
Minimum message length	3

bike facilities don't exist or need improvement
it's hard for people to see each other
it's hard to see / low visibility
it's too far / too many lanes to cross
of something that is not listed here
people are not given enough time to cross the street
people cross away from the crosswalks
people don't yield while going straight
people don't yield while turning
people double park their vehicles
people have to cross too many lanes / too far
people have to wait too long for the "Walk" signal
people run red lights / stop signs
people speed
sidewalks/ramps don't exist or need improvement
the roadway surface needs improvement
the roadway surface needs maintenance
the wait for the "Walk" signal is too long
there are no bike facilities or they need maintenance
there are no sidewalks or they need maintenance
there's not enough time to cross the street

Fig. 5. Original 21 categories by Boston City.

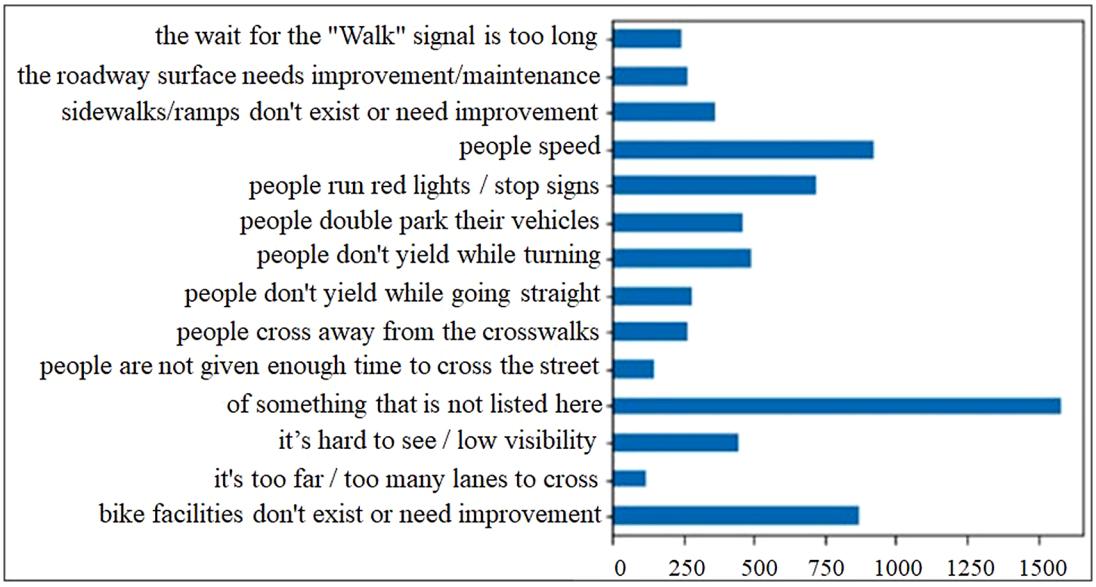

Fig. 6. Manual re-categorization optimization results.

	precision	recall	f1-score	Support
bike facilities don't exist or need improvement	0.65	0.75	0.70	651
it's too far / too many lanes to cross	0.50	0.06	0.10	87
it's hard to see / low visibility	0.58	0.54	0.56	337
of something that is not listed here	0.42	0.54	0.47	1186
people are not given enough time to cross the street	0.43	0.22	0.29	113
people cross away from the crosswalks	0.43	0.38	0.40	202
people don't yield while going straight	0037	0.27	0.31	210
people don't yield while turning	0.52	0.39	0.44	365
people double park their vehicles	0.74	0.64	0.69	343
people run red lights / stop signs	0.63	0.65	0.64	539
people speed	0.64	0.66	0.65	686
sidewalks/ramps don't exist or need improvement	0.57	0.53	0.55	275
the roadway surface needs improvement / maintenance	0.43	0.22	0.29	199
the wait for the "Walk" signal is too long	0.57	0.54	0.56	181
Avg /total	0.55	0.55	0.54	5374

Fig. 7. Result of logistic regression.

seen car bush child morning stop sign street even right way many near

true category : people don't yield while going straight
misclassified as : people run red lights / stop signs

every morning kiss walk school

true category : people run red lights / stop signs
misclassified as : of something that is not listed here

traffic turn right lane left need safety people bike driving car

true category : bike facilities don't exist or need improvement
misclassified as : of something that is not listed here

car south often travel lane take left give seen wrong turn oncoming traffic

true category : people run red lights / stop signs
misclassified as : of something that is not listed here

crossway car fast crossing

true category : people speed
misclassified as : it's too far / too many lanes to cross

Fig. 8. Misclassification in LR result.

including NMF that showed higher accuracy than LDA. The increase in the stability of the model or convergency, results in better training of the model (Wang, Yang, Wang, Xia, & Wang, 2020).

For the next step, we used the LDA category result to achieve convergence and a stable state, which produced 13 categories. LDA could provide more realistic results than K-means for topic assignment and categorization because K-means will partition N documents in K disjointed clusters or categories. On the other hand, LDA assigns a document to a mixture of categorizations and adds a Dirichlet prior in addition to the data generation process, which means that NMF qualitatively leads to worse mixtures. It fixes values for the

Table 3
Unsupervised learning results.

No.	Method	Accuracy	ARI	Variance
1	LDA	86%	0.12	7.76
2	K-Means	87%	0.07	58.84
3	NMF	90%	0.09	40.90

ARI: Adjusted Rand Index, compared with ground truth Variance: the variance of 10-fold validation results

probability vectors of the multinomials whereas LDA allows the topics and words themselves to vary.

3.4. Imbalanced data and deep learning (supervised learning)

The fourth step included various methods, such as deep learning (CNN), SMOTE, and ENN methods. After the data preprocessing, LR, and categorization optimization, we applied SMOTE-ENN. This technique is appropriate for undersampling and oversampling of data problems.

As mentioned before, owing to the nature of citizen requests, the data becomes imbalanced and concentrated on a specific category. The data used in this study also suffers from the same issue (see Fig. 9). Classification of datasets often has an unequal class distribution (oversampling and undersampling) among its examples (Rustogi & Prasad, 2019). This problem is known as imbalanced classification. SMOTE-ENN is one of the most well-known methods to balance the different number of examples among each class. However, the unequal class distribution is not regarded as a problem in itself anymore as performance degradation has also been associated with other factors related to data distribution (Sáez, Luengo, Stefanowski, & Herrera, 2015). One such problem includes having noisy and borderline data points in imbalanced datasets, which degrade a classifier's performance. To cope with such data points, SMOTE-ENN was used.

SMOTE was applied to each category that had an insufficient number of samples to adjust the balance of the sample data (Sun, Li, Fujita, Fu, & Ai, 2020). SMOTE-ENN equivalently balanced the data class distributions in the training data.

After this, CNN (Kim, 2014) was applied to the original and Google pre-trained embedded data, Word2Vec (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013), which was originally derived by the company from Google News data comprising hundreds of billions of words. However, Word2Vec is commonly used in sentiment analysis research on review texts in diverse domains aside from the news domain, including restaurants, movies, laptops, and cameras (Kim, 2014; Wang & Liu, 2015). Citizen requests usually consist of texts directly written by users themselves, which may include many typos, neologisms, and abbreviations. As a result, critical keywords can often be missing when analyzing such unstructured data. To solve this problem, the missing words were replaced with proper words through Word2Vec where words that have similar contexts were distributed in close distance in the word embedding space.

It is a common practice to use accuracy, precision, recall, and F1-score as metrics to evaluate the performance of methods in machine learning (Joshi, 2016).

According to precision and recall calculation results, the Google pre-trained model obtained higher accuracy than the original pre-trained model. This is because the Google model has pre-trained word vectors on several datasets, such as news and Wikipedia, in comparison to the original pre-trained dataset models, which were trained on only on a single dataset.

Figs. 10 and 11 show the differences between Google-pretrained and original-pretrained models using the same architecture and

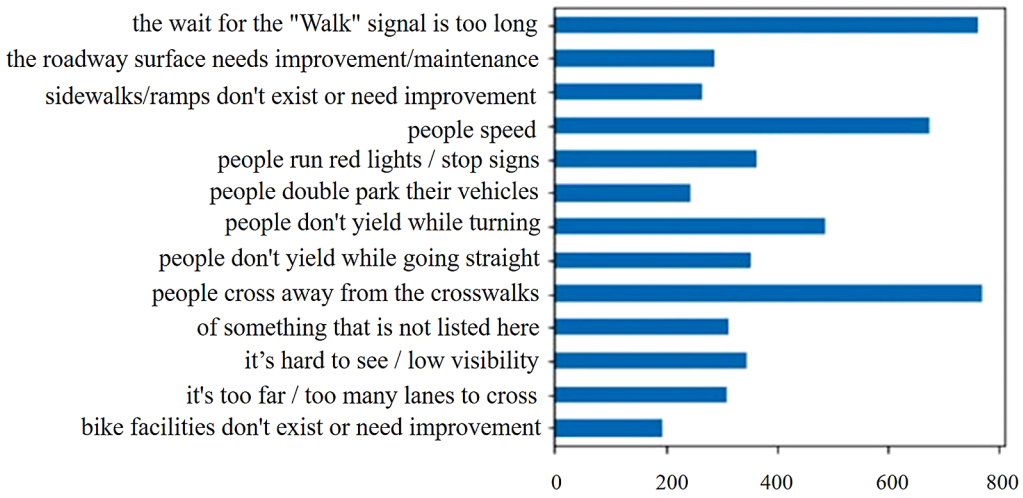

Fig. 9. Final result of categorization optimization result.

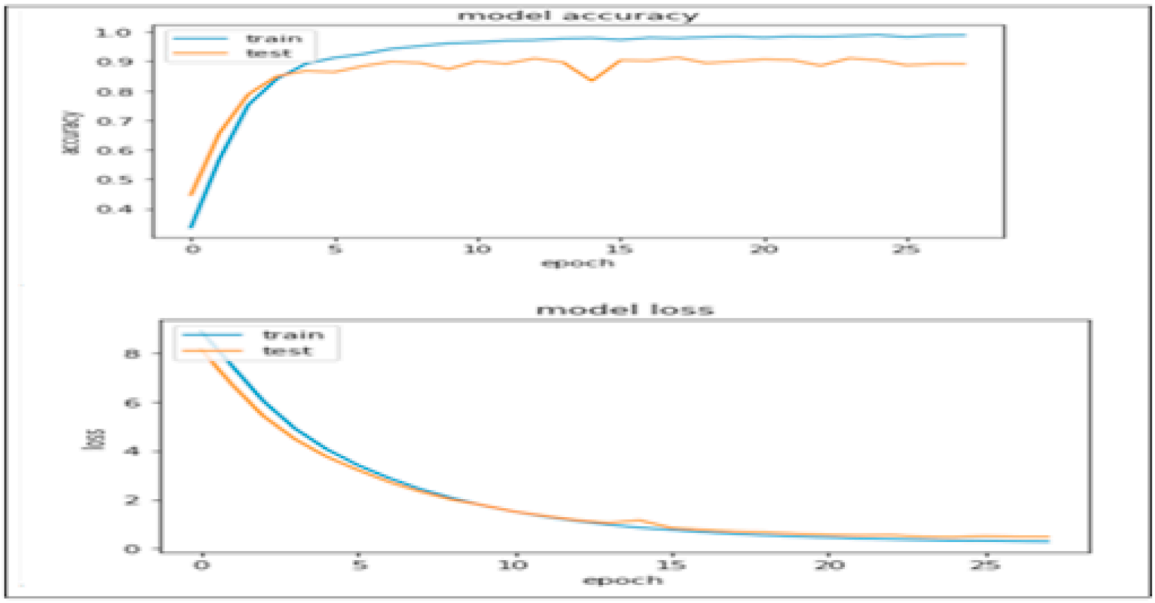

Fig. 10. Original-pretrained model CNN.

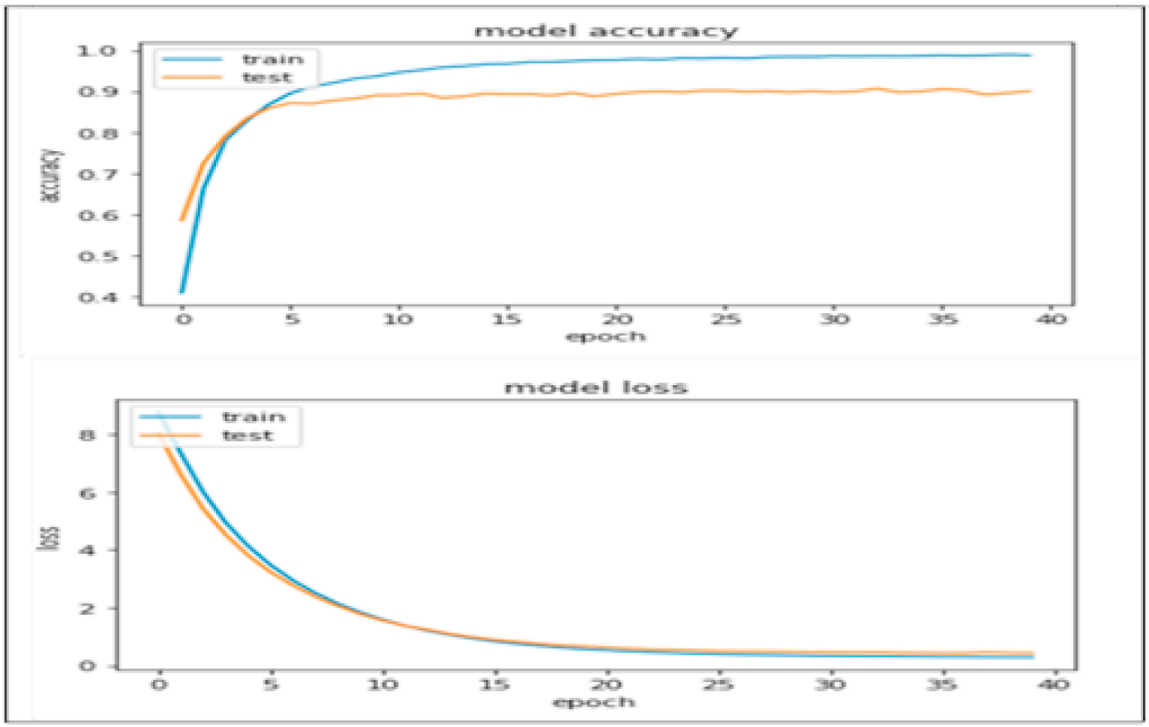

Fig. 11. Google-pretrained model CNN.

data, reaching 80% and 86–90% accuracy, respectively.

Next, K-fold cross-validation was conducted to prevent a biased concentration of the data when dividing it into training and testing sets for model evaluation. After partitioning the data ten times, the error rate for each fold was calculated and then the average was computed. Table 4 shows the result of the 10-fold validation. This table shows that the Google-pretrained system reached 89% accuracy on average, while the original system reached 86% accuracy.

Table 4
Final result of supervised learning.

Embedded model CNN	Accuracy mean (10-fold validation)
Google-Pretrained	89.38%(+/− 2.74%)
Original-Pretrained	86.02%(+/− 7.52%)

3.5. Results

In this study, a search process for the optimal automatic classification model among various contemporary models was described using the example of transportation-related citizen requests of Boston city. Each model was largely divided into two phases, namely, unsupervised learning (clustering) that searches for the optimal number of categories, and supervised learning (CNN) that automatically classifies the categories. To increase the accuracy of the unsupervised and supervised learning systems, five experimental models were applied in a consecutive order (see Table 5). For each model, the accuracy of category classification was checked via LR.

For model 1 (TF-IDF, CNN) transportation-related citizen requests data with 7166 observations were manually merged (researchers read the citizen request topic) into 14 categories from the original 21 categories. In the past, when a new request was logged, a person had to manually select and classify their request among the 21 categories. However, these categories were very similar to each other; therefore, the 21 categories were merged into 14 categories through peer-evaluation. The accuracy of the first model was 60%, lower than our expectation. It was assumed to be due to an overfitting problem, where the training dataset was excessively learned, owing to the large number of categories, and there being relatively insufficient amount of data in each category.

For model 2 (TF-IDF, CNN, more data), however, the result of the three years' data had a lower accuracy compared to when two years' data was saved. From the two results, it appears that our classification may not have a serious overfitting issue; however, the manual classification may result in a lower accuracy for the three years' data.

Therefore, for model 3 (TF-IDF, CNN, more data, Clustering), the categories were reestablished using the unsupervised learning of the clustering scheme. LDA, k-maens, and NMF were employed as the clustering schemes based on which the category and class were checked and optimized. In the category optimization result, 13 categories were determined by all three schemes. In other words, the classification accuracy was poor, and the 13 topics categorized through unsupervised learning became the new labels for optimizing the 14 topics. As a result of the second LR, all schemes showed 50% accuracy. The classification using CNN showed an accuracy of 54%, which is approximately 10% higher than model 2 but is still lower than model 1.

For model 4 (Unigram, CNN, more data, Clustering), Unigram was used to further enhance the outcome of clustering and LR obtained with the third model. Although each specific word has a meaning, the combination with other words (i.e., group of words) makes it easier to understand the meaning of a sentence. After using Unigram, the first LR accuracy increased to 55%, which is higher than that when using TF-IDF (54%) in previous models. The second LR result produced an accuracy of 86% (LDA), 87% (K-means), and 90% (NMF), which are higher than the 50% accuracy of model 3 and 70% of model 1. The classification using CNN showed 70% accuracy, which is approximately 16% higher than the prior model.

For model 5 (Unigram, CNN, more data, Clustering, SMOTE-ENN), a sampling scheme was employed to solve the problem of imbalanced data. This is a challenging problem involving classification, where the size of one class of target variables is overwhelmingly larger or smaller compared with other classes (Lee & Lee, 2014). In this case study, SMOTE-ENN was employed to address the problem of imbalanced data. Then, we compared originally pre-trained Kim's CNN model with the Google pre-trained one. K-Fold

Table 5
Result summary of stepwise models.

	Model 1	Model 2	Model 3	Model 4	Model 5	
Observation		7166	9343	9343	9343	9343
Pre-processing	Input	TF-IDF	TF-IDF	TF-IDF	Unigram	Unigram
	Accuracy	50%	54%	54%	55%	55%
Un-supervised Learning	LDA	Peer-screen	Peer-screen	50%	86%	86%
	K-means			54%	87%	87%
	NMF			50%	90%	90%
	Category	14	14	13	13	13
Supervised Learning	**Original**					
	(Accuracy)	60%	41%	48%	50%	86%
	(Recall)	58%	38%	55%	60%	85%
	(Precision)	61%	44%	47%	55%	86%
	(F-score)	60%	41%	51%	57%	85%
	Google					
	(Accuracy)	70%	40%	54%	70%	90%
	(Recall)	70%	40%	60%	70%	88%
	(Precision)	70%	40%	55%	70%	92%
	(F-score)	70%	40%	57%	70%	90%

Note: The number of layers in Deep Learning and CNN set 3 depth hidden layers and the dimensions of Word2Vec were 300 for all modesl.

validation was implemented 10 times to test the model, which yielded an accuracy of 86.02% and 89.38% for original and Google pre-trained models, respectively. This indicates that the Google-pretrained CNN model performed better. Google embedding, which is a model pre-trained with the corpus (3 billion execution words) and word vector model (3 million English word vectors) of Google news, is used in many application programs, and it has a higher accuracy than the original embedding.

4. Conclusions and managerial implications

Automatic classification of citizen requests is being actively investigated by researchers to enable governments to respond to their citizens more swiftly and to better utilize their human resources. Unlike previous automatic classification studies that were conducted using relatively high quality data such as academic materials and media articles (Fang et al., 2020; Gargiulo, Silvestri, Ciampi, & Pietro, 2019; Kim, 2019), the present study proposes the classification of a large number of citizen requests extracted from unstructured text comprising imbalanced categories. Specifically, the dataset used in the present study is different from those used in previous studies in the following aspects: (1) a larger number of categories are present, (2) misclassification errors and imbalance among categories are present, thereby making the task of quantification difficult, and (3) the learning set is not sufficiently large. To address the issues that arise due to manual classification, the categories were optimized. Furthermore, to resolve the data imbalance problems caused by the presence of several categories and specific requests, a new classification model was proposed for citizen requests using sampling methods. To this end, first, the categories that formed the basis of the classification of the citizen requests were confirmed and consequently optimized. Overfitting occurred when the CNN model was applied to the 7166 transportation-related citizen requests extracted from the dataset that the City of Boston published between 2016 and 2017. In the second phase, the algorithm was applied to approximately 9000 citizen requests, including those published up to 2018 to address the overfitting problem that arises when the amount of data used is insufficient. Consequently, the accuracy was lower than in the case of when only the data up to 2017 were used. Given that the existing categories were not accurately distinguished based on the type of the problem, the accuracy decreased as new citizen request data were added. Initially, the citizen requests were divided into various types without a determined standard. Furthermore, the low accuracy can also be attributed to the nature of the citizen requests, which changes depending on the circumstances. The existing categories in each topic should be rechecked, and optimization of the categories should be regularly conducted.

Second, the narrow topic of transportation that was selected in this study was further divided into smaller categories (means of transportation, road, facilities, and traffic regulations), and as a result, the task of classification became challenging. Among the unsupervised learning methods of clustering, LDA, K-Means, and NMF schemes were used. Among these, LDA continuously showed the best outcome and was selected to accurately define the categories. To optimize the categories, an appropriate clustering scheme should be selected to create better quality training data based on the coverage and characteristics of the topic.

Third, the data imbalance problem needed to be addressed. Citizen requests that reveal insights into public grievances with government policies are usually concentrated on a specific topic. The longer the data is accumulated, the stronger is the concentration on a specific topic. Consequently, the imbalanced data problem becomes more challenging. Among the variety of data sampling methods, SMOTE-ENN can be used to effectively balance the data.

Fourth, word embedding is a key problem in deep learning. However, citizen request data contain informal language that is of a quality lower than that of news or academic documents, and this in turn makes the task of processing difficult. In this study, desirable outcomes were obtained with the use of Google embedding. Recently, several studies have investigated the method of diverse embedding, such as GloVe by the Stanford University and FastText by Facebook's AI Research Laboratory. In this method, embedding that matches the characteristics of the data should be selected.

This study makes the following contributions. The approach of applying deep learning in the classification of citizen requests is expected to be used in the processing and classification of other citizen requests and low-quality text data on specific topics in the future. For instance, understanding citizens' needs via internet bulletin board, blog, and SNS is extremely crucial for proper municipal administration. The automatic classification methods proposed by our study can be extended for the identification of such needs and a fast response. Second, the reduced time and workloads will enable municipal employees to concentrate more on core tasks as observed in the case of the Swedish Tax Agency. Third, a more accurate classification will provide the government with objective criteria for policy decision making. Furthermore, this study has substantial academic importance given that diverse machine learning-related theories are proven by applying them to various unstructured data of citizen requests.

This study suffers from certain limitations owing to the use of several variables associated with automatic classification. Hence, a better model can be developed in future. Furthermore, a more optimized model for the automatic classification of citizen requests should be proposed using the clustering techniques of unsupervised learning, diverse sampling schemes, and embedding schemes that address the problem of imbalanced data and up-to-date deep learning algorithms, such as RNN-type long short-term memory (LSTM) and gated recurrent unit (GRU), including CNNs.

CRediT authorship contribution statement

Narang Kim: Conceptualization, Methodology, Software, Investigation. **Soongoo Hong:** Data curation, Writing - original draft, Visualization, Writing - review & editing.

Acknowledgement

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

Supplementary material

Supplementary material associated with this article can be found, in the online version, at 10.1016/j.ipm.2020.102410

References

Agostino, D., & Arnaboldi, M. (2016). A measurement framework for assessing the contribution of social media to public engagement: An empirical analysis on facebook. *Public Management Review, 18*(9), 1289–1307.

AI Innovaion of Sweden (2019). Artificial intelligence improves the swedish tax agency's customer service. https://www.ai.se/en/news/artificial-intelligence-improves-swedish-tax-agencys-customer-service. Accessed 28 Jan 2020.

Banerjee, I., Ling, Y., Chen, M. C., Hasan, S. A., Langlotz, C. P., Moradzadeh, N., … Lungren, M. (2018). Comparative effectiveness of convolutional neural network (CNN) and recurrent neural network (RNN) architectures for radiology text report classification. *Artificial intelligence in medicine, 97*, 79–88.

Barrett, K., & Greene, R. (2016). Is a 40-hour workweek enough in government? Governing. https://www.governing.com/columns/smart-mgmt/gov-time-usage-survey-government.html. July 21, 2016, Accessed 10 Jan 2020.

Bouazizi, M., & Ohtsuki, T. (2018). Multi-class sentiment analysis in twitter: What if classification is not the answer. *IEEE Access, 6*, 64486–64502.

Cho, T. I., Choi, B. G., Na, Y. W., Moon, Y. S., & Kim, S. H. (2018). A suggestion for spatiotemporal analysis model of complaints on officially assessed land price by big data mining. *Journal of Cadastre & Land InformatiX, 48*(2), 79–98.

Dalianis, H., Sjöbergh, J., & Sneiders, E. (2011). Comparing manual text patterns and machine learning for classification of e-mails for automatic answering by government agency. *International conference on intelligent text processing and computational linguistics* (pp. 234–243).

Dubey, R., Zhou, J., Wang, Y., Thompson, P. M., & Ye, J. (2014). Alzheimer's disease neuroimaging initiative. Analysis of sampling techniques for imbalanced data: An $n = 648$ ADNI study. *NeuroImage, 87*, 220–241.

Eshleman, R. M., & Yang, H. (2014). A spatio-temporal sentiment analysis of twitter data and 311 civil complaints. In 2014 IEEE 430 fourth international conference on big data and cloud computing, 477–484.

Fang, W., Luo, H., Xu, S., Love, P. E., Lu, Z., & Ye, C. (2020). Automated text classification of near-misses from safety reports: An improved deep learning approach. *Advanced Engineering Informatics, 44*, 101060.

Fu, J., & Lee, S. (2012). A multi-class SVM classification system based on learning methods from indistinguishable chinese official document. *Expert systems with applications, 39*(3), 3127–3134.

Gargiulo, F., Silvestri, S., Ciampi, M., & De Pietro, G. (2019). Deep neural network for hierarchical extreme multi-label text classification. *Applied Soft Computing, 79*, 125–138.

Hagen, L. (2018). Content analysis of e-petitions with topic modeling: How to train and evaluate LDA models? *Information Processing & Management, 54*(6), 1292–1307.

Hartmann, J., Huppertz, J., Schamp, C., & Heitmann, M. (2019). Comparing automated text classification methods. *International Journal of Research in Marketing, 36*(1), 20–38.

He, H., Bai, Y., Garcia, E. A., & Li, S. A. (2008). Adaptive synthetic sampling approach for imbalanced learning. *IEEE International joint conference on neural networks*.

Jeong, H. Y., Lee, T. H., & Hong, S. G. (2017). A corpus analysis of electronic petitions for improving the responsiveness of public services: Forcusing on busan petiton. *The Korean Journal of Local Government Studies, 21*(1), 423–436.

Joshi, R. (2016). Accuracy, precision, recall & f1 score: Interpretation of performance measures. *Retrieved April, 1*, 2016.

Karakuş, B., Talo, M., Hallaç, I. R., & Aydin, G. (2018). Evaluating deep learning models for sentiment classification. *Concurrency and Computation: Practice and Experience, 30*(21), e4783.

Kim, D., Seo, D., Cho, S., & Kang, P. (2019). Multi-co-training for document classification using various document representations: TF-IDF, LDA, and doc2vec. *Information Sciences, 477*, 15–29.

Kim, H., Kim, J., Kim, J., & Lim, P. (2018a). Towards perfect text classification with wikipedia-based semantic naïve Bayes learning. *Neuro Computing, 315*, 128–134.

Kim, H., Lee, T., Ryu, S., & Kim, N. (2018b). A study on text mining methods to analyze civil complaints: Structured association analysis. *Journal of the Korea Industrial Information Systems Research, 23*(3), 13–24.

Kim, M., Yang, H., & Kim, C. S. (2007). Improved focused sampling for class imbalance problem. *Journal of Information Processing Society Software and Data Engineering, 14*(4), 287–294.

Kim, P. J. (2019). An analytical study on automatic classification of domestic journal articles using random forest. *Journal of the Korean Society for Information Management, 36*(2), 57–77.

Kim, Y. (2014). Convolutional neural networks for sentence classification. *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1746–1751).

Ku, C., & Leroy, G. (2014). A decision support system: Automated crime report analysis and classification for E-government. *Government Information Quarterly, 31*(4), 534–544.

Le Cun, Y., Bottou, L., Bengio, Y., & Haffner. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE, 86*(11), 2278–2324.

Lee, H. J., & Lee, S. G. (2014). A comparison of ensemble methods combining resampling techniques for class imbalanced data. *The Korean Journal of Applied Statistics, 27*(3), 357–371.

Lee, J. H., & Choi, H. (2020). An analysis of public complaints to evaluate ecosystem services. *Land, 9*(3), 62.

Lima, M. S. M., & Delen, D. (2020). Predicting and explaining corruption across countries: A machine learning approach. *Government Information Quarterly, 37*(1), 101407.

Lu, T., Huang, Y., Zhao, W., & Zhang, J. (2019). The metering automation system based intrusion detection using random forest classifier with SMOTE+ENN. *2019 IEEE 7th international conference on computer science and network technology (ICCSNT)* (pp. 370–374). https://doi.org/10.1109/ICCSNT47585.2019.8962430.

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

Montebruno, P., Bennett, R. J., Smith, V. H., & Lieshout, C. (2020). Machine learning classification of entrepreneurs in british historical census data. *Information Processing & Management, 57*(3), 102210.

Rustogi, R., & Prasad, A. (2019). Swift imbalance data classification using SMOTE and extreme learning machine. *In 2019 international conference on computational intelligence in data science (ICCIDS)* (pp. 1–6). IEEE.February)

Ryu, S. E, Hong, S. G., Lee, T. H., & Kim, N. R. (2018). A pattern analysis of bus civil complaint in Busan city using the text network analysis. *Korean Computers and Accounting Review, 16*(2), 19–43.

Sáez, J. A., Luengo, J., Stefanowski, J., & Herrera, F. (2015). SMOTE-IPF: Addressing the noisy and borderline examples problem in imbalanced classification by a re-sampling method with filtering. *Information Sciences, 291*, 184–203.
Seo, J. H. (2017). A study on the performance evaluation of unbalanced intrusion detection dataset classification based on machine learning. *Journal of Korean Institute of Intelligent Systems, 27*(5), 466–474.
Seo, M. (2014). *Data processing analysis practice using r*. Seoul: Gil-Budd.
Siagian, A. H. A. M., & Aritsugi, M. (2020). Robustness of word and character N-gram combinations in detecting deceptive and truthful opinions. *Journal of Data and Information Quality (JDIQ), 12*(1), 1–24.
Singh, P., Dwivedi, Y. K., Kahlon, K. S., Pathania, A., & Sawhney, R. S. (2020). Can twitter analytics predict election outcome? An insight from 2017 Punjab assembly elections. *Government information quarterly* (p. 101444).
Sun, J., Li, H., Fujita, H., Fu, B., & Ai, W. (2020). Class-imbalanced dynamic financial distress prediction based on adaboost-SVM ensemble combined with SMOTE and time weighting. *Information Fusion, 54*, 128–144.
Wang, B., & Liu, M. (2015). Deep learning for aspect-based sentiment analysis. *Stanford university report*.
Wang, X., Yang, X., Wang, X., Xia, M., & Wang, J. (2020). Evaluating the competitiveness of enterprise's technology based on LDA topic model. *Technology Analysis Strategic Management, 32*(2), 208–222.
Wilson, D. L. (1972). Asymptotic properties of nearest neighbor rules using edited 782 data. *IEEE Transactions on Systems Man and Cybernetics, 2*(3), 408–421. https://doi.org/10.1109/TSMC.1972.4309137.
Wirtz, B. W., & Müller, W. M. (2019). An integrated artificial intelligence framework for public management. *Public Management Review, 21*(7), 1076–1100.
Yin, W., Kann, K., Yu, M., & Schütze, H. (2017). Comparative study of CNN and RNN for natural language processing. ArXiv preprint arXiv:1702.01923.
Zhang, Y., & Wallace, B. (2017). A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification. *In Proceedings of the Eighth International Joint Conference on Natural Language Processing, 1*, 253–263.November
Zheng, J., Cai, F., Chen, d. H., & Rijke, M. (2020). Pre-train, interact, fine-tune: A novel interaction representation for text classification. *Information processing & management* (p. 102215).