Journal of Autonomous Intelligence (2023) Volume 6 Issue 3
doi: 10.32629/jai.v6i3.996

REVIEW ARTICLE

Establishment of data mining-based public education administrative work automation system and student activity analysis

Seung-Ryeol Joo[1], Jong-Chan Kim[2], Sung-Jun Kim[3,*]

[1] *Department of Education, Konkuk University, Seoul 05029, South Korea*

[2] *Department of Education Policy Development, Innovalue Partners, Yongin 17055, South Korea*

[3] *Department of Big Data Content Convergence, Namseoul University Graduate School, Cheonan-si 31020, South Korea*

*** Corresponding author:** Sung-Jun Kim, mvstar@hanmail.net

ABSTRACT

There are several important factors in public education in Korea. Among them, it is very important to manage time to improve teachers' educational capabilities and students' grades. However, in Korea's public education, the existing school administrative work system has to deal with miscellaneous procedures, hindering teachers from guiding students. As a result, students also give low trust in public education. This study introduces procedures for an integrated public education data management system to automate administrative tasks of teachers and increase students' educational capabilities. By applying the Data Mining Problem Solving Methodology (ICAIS), we identified five stages in which data is processed. In addition, the activities required for students to go to college were processed with text mining techniques (from a simple word cloud to the construction of a neural network algorithm classification model), allowing students to check their grades themselves. Through this study, it reduces teachers' chores, concentrates student education, and provides students with the educational purpose of a self-directed method that determines their career path.

Keywords: data mining; text mining; self-directed learning; administrative work system automation

ARTICLE INFO

Received: 18 July 2023
Accepted: 14 August 2023
Available online: 25 September 2023

1. Introduction

The public education system in Korea has historically been known for its competitiveness and exam-oriented structure. The curriculum places a heavy emphasis on standardized testing, which has led to intense academic pressure on students to perform well in examinations. As a result, the system has become somewhat rigid, with a focus on rote memorization and academic achievement rather than fostering holistic learning and critical thinking. One significant challenge faced by the Korean public education system is the burden of administrative work on teachers. The current rigid system necessitates a considerable amount of administrative tasks, diverting teachers' time and energy away from focusing on innovative teaching methods and individualized student support. This can potentially impact the overall quality of education delivered in classrooms. Moreover, the college admission process in Korea heavily relies on students' high school records and standardized test scores, making it crucial for students to perform well academically throughout their high school years. The pressure to excel academically, combined with limited flexibility in course selection, can create challenges for

students in pursuing subjects that align with their interests and future career paths[1–5].

As a response to these challenges and the changing landscape of education due to the COVID-19 pandemic, there have been efforts to introduce IT technology into public schools. With the advent of e-learning and the availability of various data collection methods, there is an opportunity to streamline administrative tasks for teachers and utilize data-driven insights to improve educational practices. In contrast, the rise in private tutoring (known as "hagwon" in Korean) can be attributed to several factors. Many students and parents perceive private tutoring as a means to supplement the shortcomings of the public education system and provide additional support for achieving higher exam scores. Private tutoring institutes offer specialized and tailored instruction, exam-focused preparation, and personalized attention, which students may feel are lacking in the public education setting. Consequently, students often turn to private tutoring to bolster their academic abilities and improve their chances of getting into prestigious universities.

By considering these fundamental aspects of the Korean public education system and the factors driving students toward private tutoring, the article can present a more comprehensive and informative perspective on the challenges and potential solutions to improve the quality of education in Korea. We conducted two studies. Recently, with the introduction of IT technology in public education, various data can be easily collected in schools, and the data can be used to handle complex administrative tasks for teachers. In addition, by analyzing students' competency activities with text mining techniques, we intend to build a system that allows students to self-diagnose and confirm their current academic capabilities.

Research 1 aims to develop an automated school administration system using data mining techniques to streamline administrative tasks and improve efficiency.

Research 2 focuses on implementing a self-directed learning diagnosis system based on text mining to empower students in assessing and enhancing their academic competencies.

2. Related research

2.1. Data mining problem solving methodology

Data mining techniques were a key technology in problem-solving methodologies that emerged based on manufacturing in the mid-1990s. Based on statistics, data mining combined with computer science technology represents many technological advances in machine learning and has become a key tool used in the field. By building a database, you can organize several statistical techniques, such as visualization and hypothesis testing, into a single flow, starting with a structure in which data is collected through pipes. In particular, the ICAIS problem-solving methodology defines the flow of data from problem definition, data collection, analysis, application, and systemization based on manufacturing, and various statistical techniques and data mining techniques are mobilized. This type of data mining-based problem-solving methodology originated in manufacturing, but is widely used in various domains where data is generated. First, in the problem definition stage, teachers plan whether there is data corresponding to various complex problems among the various tasks that teachers encounter administratively, what kind of analysis can be performed, and how systematization will be carried out. In the data collection stage, the data generated from teachers and students in schools conducting public education are defined, and key data are set for the purpose of analysis[6–12]. And in the analysis stage, data analysis results that help self-directed learning, which can be monitored by teachers or students, are derived. Later, in the application and system phases, the design work is carried out so that the data that will occur in the future can be put into one system and expressed simply (**Figure 1**).

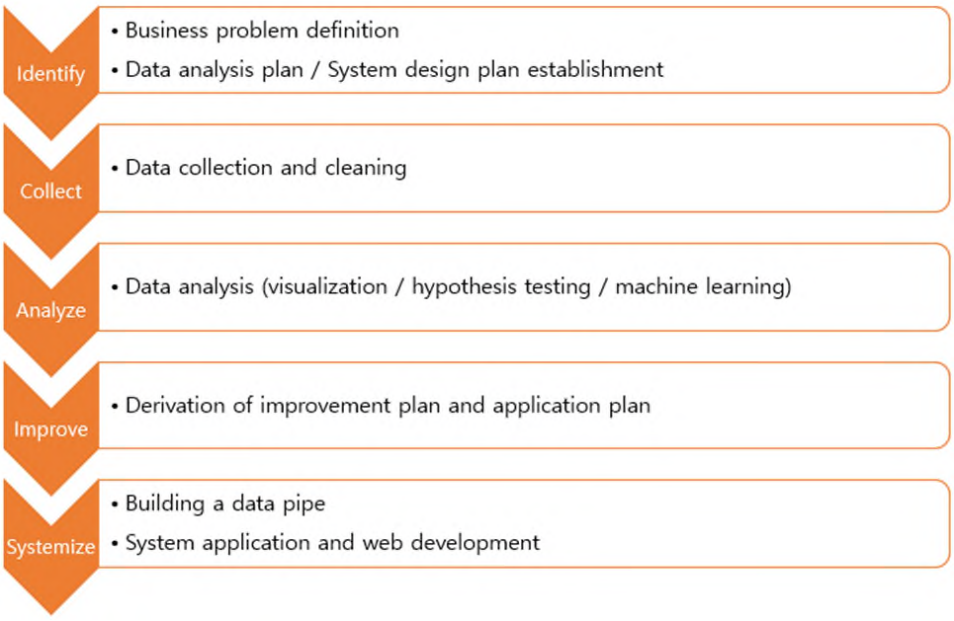

Figure 1. ICAIS problem solving methodology.

2.2. Text mining

Analyzing text is of great value along with looking at quantitative metrics. In particular, evaluation data for human resources, that is, data related to student activities, are often expressed in text. It is a technology that can easily evaluate and display not only the existing test scores of students, but also various activities related to career and competency. Text mining technology is a technology that applies statistics and data mining technology based on natural language processing. Representatively, there are word cloud, text classification, and association analysis. In the case of word cloud analysis, it is easy to find key words and expressions in a particular sentence or paragraph. Text classification is a technology that uses machine learning techniques to classify newly collected text according to labels specified by users in advance. Recently, neural network algorithms are used to utilize text classification technology. In the case of association analysis, analysis using unsupervised learning calculates the similarity between words in each sentence or paragraph to check how similar the sentences are to each other[13–20].

3. Research methodology

This research is divided into two perspectives. The first is to automate student attendance, which is one of the most worrisome tasks from the teacher's point of view, and configure it in a simple system format.

3.1. Data mining-based attendance processing

Data mining-based attendance processing systemization is divided into 5 steps. First, problem definition. Second, data collection. Third, data analysis. Fourth, derivation of application plan. Fifth, systemization. **Figure 1** shows how this procedure would be performed.

3.1.1. Problem definition

First of all, you can discover one of the many complex tasks that occur in schools that conduct public education: the problem of student attendance. It has a very high impact on student success and is one of the most time-consuming areas of teachers' work.

3.1.2. data collection

In this study, data were collected from about 500 students at D high school located in Gyeonggi-do, Korea. This data includes the student's personal information, grades, etc., and shows attendance by date.

Attendance information contains information such as tardiness, absence, and attendance, and the teacher must process it by uploading it to the payment system (**Table 1**).

Table 1. Collect D school student data.

Feature	Type	Describe
ID	ID	Student ID number
Birth date	Datetime	Student birth date
Gender	Object	Student gender
Admission year	Datetime	Admission year
Current grade	Numeric	Current grade
1st year class	Numeric	1st year class
2nd year class	Numeric	2nd year class
3rd grade class	Numeric	3rd grade class
Absence days	Numeric	Absence days count
Unexcused absence days	Numeric	Unexcused absence days count
Tardiness	Numeric	Tardiness days count

3.1.3. Data analysis

It shows the data of attendance-related data that teachers need to process through visualization. This can be analyzed later for the purpose of managing student grades. Currently, for the purpose of increasing work efficiency, teachers are responsible for checking the attendance status of students and checking their activities.

3.1.4. Application and systemization

Based on the analysis, we organize the process for attendance. And we build a system to handle it online easily. A picture of the system is shown below.

3.2. Text mining for student activities

It is very important for students to self-diagnose their academic and activity competencies. A lot of research has already been conducted on the task of analyzing and expressing the quantitative grades of existing students, and numerous services have been released. However, there is a lack of educational research or services that analyze students' activities and show them. Utilizing the text format used in Korean public education, the career-related activities entered by students are analyzed using text mining techniques. About 200 texts were mobilized here, and a word cloud analysis to find keywords for student activities and a text classification model to classify specific career activities were constructed. In this study, a classification model was created based on traditional machine learning techniques, and a model with high performance was derived compared to a naive Bayesian-based model. And it was combined with the previously configured administrative assistance processing system.

In particular, in traditional machine learning techniques, ensemble algorithms that combine multiple algorithms to create a more powerful classifier show extremely high performance. This algorithm, which is based on the decision tree algorithm of existing data mining, has several algorithms such as Voting, Bagging, and Boosting, which boasts a high-performance classification model.

In traditional machine learning techniques, ensemble algorithms play a crucial role in creating powerful classifiers that exhibit remarkably high performance. These algorithms are based on the decision tree algorithm, a popular method in data mining. Three well-known ensemble algorithms commonly used in text classification tasks are Voting, Bagging, and Boosting.

3.2.1. Voting

Voting is a straightforward ensemble technique where multiple individual classifiers are trained on the same dataset. Each classifier independently makes predictions, and the final classification decision is made based on a majority vote. This approach is particularly effective when different classifiers bring diverse perspectives to the problem, leading to improved overall accuracy.

3.2.2. Bagging (bootstrap aggregating)

Bagging involves training multiple instances of the same classifier on random subsets of the dataset, selected with replacement. Each classifier generates its predictions, and the final classification decision is determined through a weighted average or voting scheme. Bagging reduces the risk of overfitting and enhances the model's generalization ability.

3.2.3. Boosting

Boosting is another ensemble technique that iteratively trains multiple weak classifiers in sequence. At each iteration, the algorithm assigns higher weights to the misclassified instances from the previous iteration. This focuses subsequent classifiers on the previously misclassified samples, resulting in a strong ensemble model with improved accuracy.

However, to achieve optimal performance with ensemble algorithms, a process known as Hyper-parameter tuning is necessary. Hyper-parameters control the behavior of the algorithm and must be carefully selected to optimize model performance. Techniques like Grid Search or Random Search can be employed to systematically explore the hyper-parameter space and identify the best combination.

In the context of text classification, additional preprocessing steps are crucial to ensure smooth learning and effective model training. Padding is employed to standardize the length of input texts, ensuring all sequences have the same length, which is necessary for processing in neural networks. Text to sequence operations convert text data into numerical representations that machine learning models can process.

Python and R programming languages are widely used for data processing and machine learning tasks. In this study, the researchers opted for Python as their programming language of choice to process and learn from the data. Python offers a rich ecosystem of libraries and tools, making it well-suited for implementing text mining techniques and machine learning algorithms effectively.

4. Result and discussion

In order to apply the previously established procedure to the administrative system performed by teachers, we analyzed the experimental data collected from D high school and confirmed the expressed results.

4.1. Establishment of online attendance processing system

Based on the data collected earlier, after one student wrote the reason for attendance, the student's guardian was notified and a system was constructed to obtain consent. As shown in the figure below, student attendance must be conducted under the consent of the student's guardian and teacher, so if the student first registers the reason for attendance through the system, the parent is notified through a messenger, and if the parent agrees, the teacher and a system for administrators to pay. Payment requires the official seal of a teacher or administrator, and a function that can be uploaded in advance was designed, and it was developed in the form of an application (**Figure 2**).

In addition, when a teacher or administrator manages a class, the student attendance status is displayed on one screen so that the previously processed attendance status can be viewed. The figure below is a visualization of the attendance status of students confirmed by one teacher (administrator) among the 200

data of D high school (**Figure 3**).

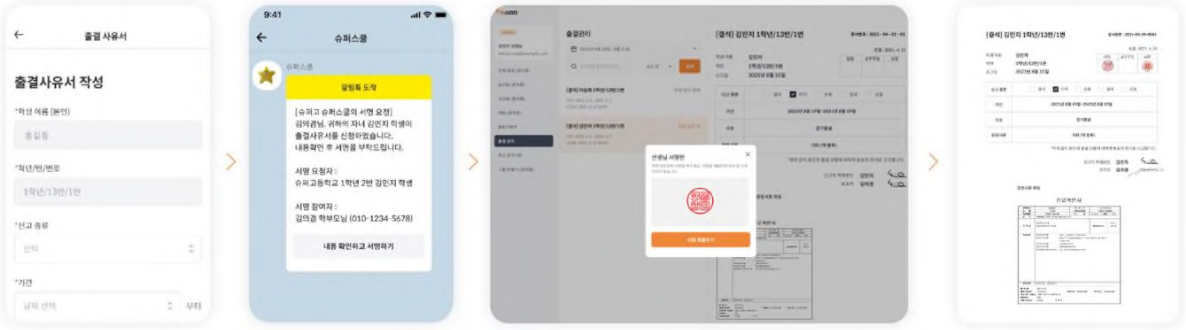

Figure 2. Implementation of attendance system.

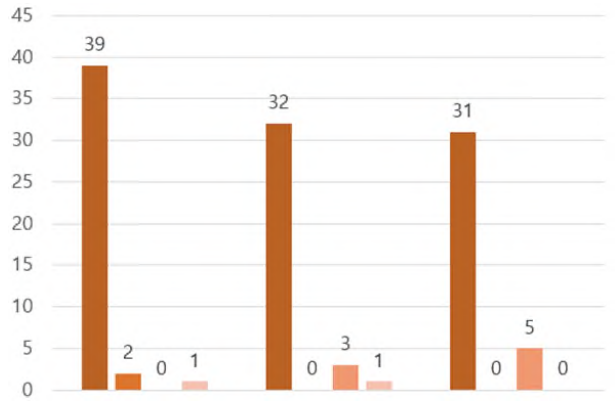

Figure 3. Visualize the attendance status of D school students in a bar graph.

4.2. Analysis of student activity data using text mining

Text mining analysis was performed using about 200 data related to students' activities. First, through the word cloud, the activity record written by one student was expressed as a graph (**Figure 4**).

Figure 4. Word cloud obtained from the report submitted by student A on student activities.

And using a machine learning model and a Bayesian model, a classification model was constructed to classify students' activities in relation to their college entrance exam competency. 70% of the total data was composed of learning data, and a multi-classification model was constructed by setting the basic competencies of the entrance examination to 'creative initiative', 'problem solving', 'leadership', 'major

related', and 'self-directed' items. It was configured using a programming language. First, a pipe was constructed in which text was processed and classified by using the Random Search technique of the machine learning technique (**Table 2**), and then Hyper-parameter tuning was performed using the Grid Search technique for the model with high performance.

The final selected model was selected as a random forest model, and the evaluation results were evaluated using accuracy as shown in the table below (**Table 3**). The verification work was also performed using 30% of the verification data.

Table 2. Performance table of text classification model on Random Search.

Model	Train acc	Test acc	Note
Decision tree model	0.613	0.602	-
Naive bayes model	0.612	0.655	-
Ensemble boosting model	0.916	0.516	Overfitting
Random forest model	0.885	0.813	Select
Support vector machine model	0.622	0.588	-

Table 3. Performance table of text classification model on Grid Search.

Model | parameter	Train acc	Test acc
Random forest model Estimator = 50 | Max depth = 5 | Max split = 10	0.876	0.772
Random forest model Estimator = 100 | Max depth = 10 | Max split = 20	0.886	0.684
Random forest model Estimator = 150 | Max depth = 15 | Max split = 30	0.899	0.887
Random forest model Estimator = 250 | Max depth = 25 | Max split = 50	0.842	0.832
Random forest model Estimator = 350 | Max depth = 35 | Max split = 100	1.00	0.521

Using the generated classification model, word cloud results, and other text mining results, a 'self-directed learning' system was constructed in which students evaluate themselves as shown below (**Figure 5**).

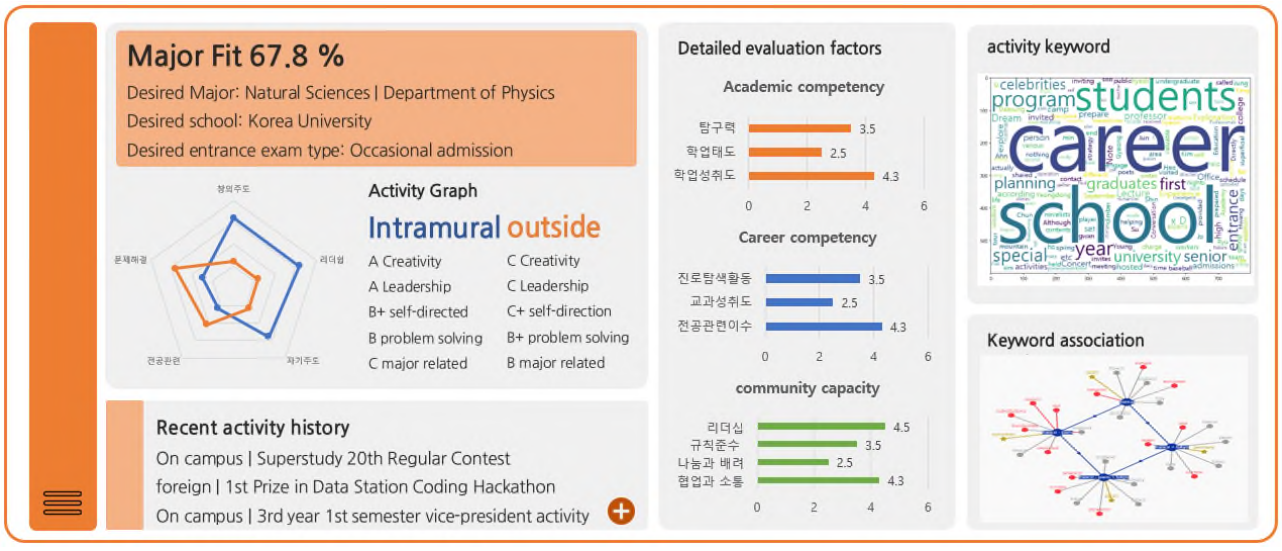

Figure 5. Student activity record system.

This is configured as an example of one student selected from over 200 data, and you can check your activity record according to each student's ID by linking with the database (**Figure 6**).

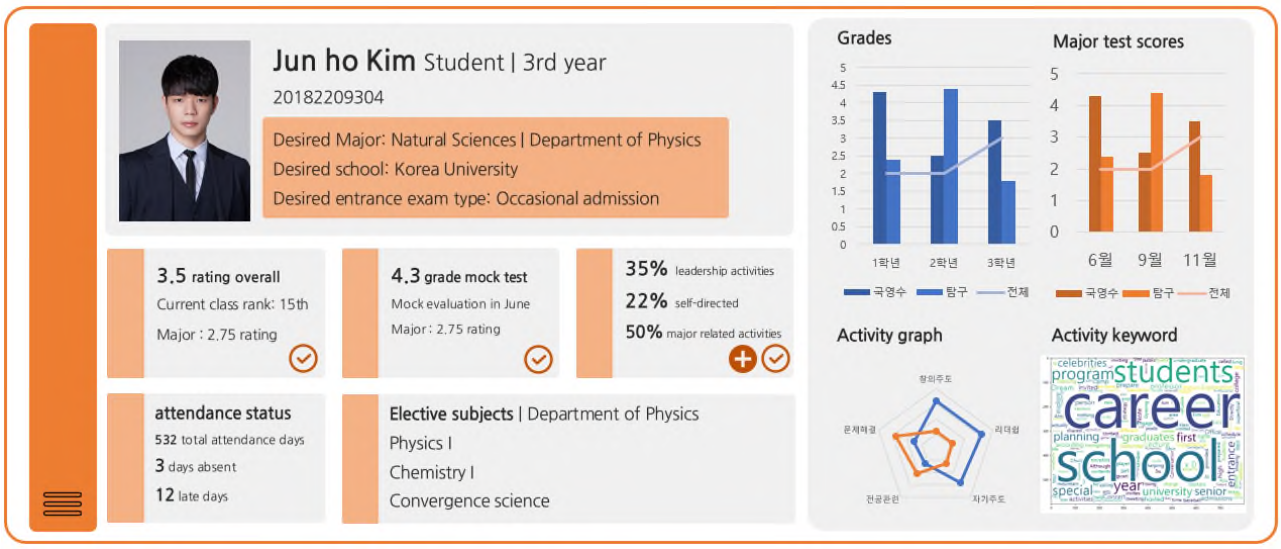

Figure 6. A system that can check quantitative grades and activities for each student.

5. Conclusion and future work

We constructed an automated system for teachers of educational institutions conducting public education and a student activity analysis system for students' self-directed learning using the previously collected data. As artificial intelligence and big data related technologies related to the 4th industry are increasingly emphasized in public education in Korea, we are trying to find a solution that can configure the previously configured system to suit each school. In addition, text mining techniques for student activity records will be analyzed using neural network-based algorithms as well as traditional machine learning algorithms. We will conduct research on how to add an auto-completion function to activity records. These improvements will have a positive impact on online public education systems in countries other than Korea in the future, and will have a great effect in areas where the private education market is large, along with improving class leadership for students through reducing the burden of administrative work on teachers, Public education could be further strengthened.

Conflict of interest

The authors declare no conflict of interest.

References

1. Roiger RJ. *Data Mining: A Tutorial-Based Primer*. Chapman and Hall/CRC; 2016.
2. Pyle D. *Data Preparation for Data Mining*. Morgan Kaufmann; 1999.
3. Jin Y. Development of word cloud generator software based on python. *Procedia Engineering* 2017; 174: 788–792. doi: 10.1016/j.proeng.2017.01.223
4. Hiemstra R. *Self-Directed Learning*. IACE Hall of Fame Repository; 1994.
5. Kowsari K, Jafari Meimandi K, Heidarysafa M, et al. Text classification algorithms: A survey. *Information* 2019; 10(4): 150. doi: 10.3390/info10040150
6. Minaee S, Kalchbrenner N, Cambria E, et al. Deep learning based text classification: A comprehensive review. *arXiv* 2020; arXiv:2004.03705. doi: 10.48550/arXiv.2004.03705
7. Mirończuk MM, Protasiewicz J. A recent overview of the state-of-the-art elements of text classification. *Expert Systems with Applications* 2018; 106: 36–54. doi: 10.1016/j.eswa.2018.03.058

8. Althobaiti MJ. BERT-based approach to Arabic hate speech and offensive language detection in Twitter: Exploiting emojis and sentiment analysis. *International Journal of Advanced Computer Science and Applications* 2022; 13(5): 972–980. doi: 10.14569/IJACSA.2022.01305109
9. Xu B, Guo X, Ye Y, Cheng J. An improved random forest classifier for text categorization. *Journal of Computers* 2012; 7(12): 2913–2920. doi: 10.4304/jcp.7.12.2913-2920
10. Virupakshappa R, Patil N. An enhanced segmentation technique and improved support vector machine classifier for facial image recognition. *International Journal of Intelligent Computing and Cybernetics* 2022; 15(2): 302–317. doi: 10.1108/IJICC-08-2021-0172
11. Rangayya, Virupakshappa, Patil N. Improved face recognition method using SVM-MRF with KTBD based KCM segmentation approach. *International Journal of System Assurance Engineering and Management* 2022; 1–12. doi: 10.1007/s13198-021-01483-3
12. Bühlmann P. Bagging, boosting and ensemble methods. In: Gentle JE, Härdle WK, Mori Y (editors). *Handbook of Computational Statistics*. Springer Berlin, Heidelberg; 2012. pp. 985–1022.
13. Vijayarani S, Ilamathi J, Nithya. Preprocessing techniques for text mining—An overview. *International Journal of Computer Science & Communication Networks* 2015; 5(1): 7–16.
14. Denny MJ, Spirling A. Text preprocessing for unsupervised learning: Why it matters, when it misleads, and what to do about it. *Political Analysis* 2018; 26(2): 168–189. doi: 10.1017/pan.2017.44
15. Fedus W, Goodfellow I, Dai AM. Maskgan: Better text generation via filling in the _. *arXiv* 2018; arXiv:1801.07736. doi: 10.48550/arXiv.1801.07736
16. Adeli H. Neural networks in civil engineering: 1989–2000. *Computer-Aided Civil and Infrastructure Engineering* 2002; 16(2): 126–142. doi: 10.1111/0885-9507.00219
17. Yu F, Xu X. A short-term load forecasting model of natural gas based on optimized genetic algorithm and improved BP neural network. *Applied Energy* 2014; 134: 102–113. doi: 10.1016/j.apenergy.2014.07.104
18. Abiodun OI, Jantan A, Omolara AE, et al. State-of-the-art in artificial neural network applications: A survey. *Heliyon* 2018; 4(11): e00938. doi: 10.1016/j.heliyon.2018.e00938
19. Sarkar D, Bali R, Ghosh T. *Hands-on Transfer Learning with Python: Implement Advanced Deep Learning and Neural Network Models Using TensorFlow and Keras.* Packt Publishing; 2018.
20. Bird S, Klein E, Loper E. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O'Reilly Media; 2009.