ISSN 1330-3651 (Print), ISSN 1848-6339 (Online) https://doi.org/10.17559/TV-20231214001207
Original scientific paper

Re-Clustering Documents to Enhance Search Accuracy with Imbalanced Abbreviation Data

Woon-Kyo LEE, Ja-Hee KIM*

Abstract: Abbreviation ambiguity poses significant challenges when searching academic literature. This study evaluated the accuracy of clustering algorithms on imbalanced datasets with varying ratios of target groups. A corpus consisting of 1052 papers focused on the study of abbreviations. The "MSA" dataset was clustered using TF-IDF, cosine similarity, and k-means. Clustering performance declined as the ratios in the target group deviated from balanced thresholds. A re-clustering method was introduced, involving the selective exclusion of non-target clusters. Re-clustering improved accuracy and *F*1 scores in most scenarios, demonstrating particular stability with higher cluster counts. The re-clustering performance of comparisons was stronger when compared to k-means and self-adaptive methods. The study highlights issues stemming from data imbalance and presents an effective strategy for enhancing abbreviation search efficiency.

Keywords: imbalanced data, K-means algorithm, Re-clustering, word sense disambiguation

1 INTRODUCTION

The digital age has seen a rapid increase in the volume of text media, especially academic papers. As noted by Professor Mike Thelwall [1], the number of papers registered in Scopus has been rapidly increasing. The use of abbreviations, particularly in academic and technical documents, has become a primary means of succinctly conveying information. However, ambiguities associated with abbreviations, such as 'MSA', which may stand for 'Multiple Sequence Alignments', 'Modern Standard Arabic', 'Microservice Architecture', or 'Microstrip Antenna', can lead to a significant expenditure of time and effort by researchers sifting through numerous documents to locate the desired information. This situation is especially complex in fields where abbreviations are frequently used, and accurately determining the precise meaning of each abbreviation becomes an even more challenging task. Therefore, there is a pressing need for new technical approaches that can effectively filter and select the documents researchers are seeking. In this context, fields such as systematic review studies require a complicated procedure where multiple experts read and judge literature according to the PRISMA guidelines [2], emphasizing the need to selectively include only relevant literature [3, 4]. In trend analysis and topic analysis studies, researchers also need to invest effort and time in manually categorizing the searched documents to find the ones they desire [5-7]. Especially, discerning documents that contain ambiguous abbreviations, such as "MSA", which can have multiple meanings, presents a major challenge. Automatically determining the precise meanings of these abbreviations is not a simple task [8, 9].

In scientific research, various abbreviations are commonly used. Traditionally, rule-based methods [10] and statistical methods [11-13] have been employed to resolve abbreviations ambiguity. However, these methods have limitations in accurately grasping the context and are constrained when dealing with large-scale data. Especially, while supervised learning offers high accuracy, preparing training data is time-consuming and expensive [14]. Accordingly, there has been an increasing emphasis on clustering methods among unsupervised learning techniques recently [15-18]. Nevertheless, research on clarifying abbreviations remains inadequate.

The challenge of imbalanced datasets emerges as a significant hurdle in machine learning, with extensive research efforts across various domains, including health, finance, engineering, and intelligent fault diagnosis, underscoring the complexity of this issue [35, 36]. Studies highlight a spectrum of solutions, from data pre-processing techniques to model construction methods and training optimization strategies, aimed at mitigating the bias towards majority classes [36]. Recent research has expanded the field by examining sampling techniques that modify data distribution to improve accuracy and more effectively tackle the issues present in imbalanced datasets [38, 39]. Despite these advancements, a notable research gap persists in unsupervised learning approaches, especially concerning abbreviation ambiguity within imbalanced datasets. A systematic mapping study points to a reliance on oversampling and classical machine learning models, while another review underscores the necessity for innovative data processing, model construction, and training optimization methods in the fault diagnosis domain [35]. However, both indicate that exploration into unsupervised learning techniques for addressing abbreviation-related imbalances is still in its early stages. This situation underscores the need for research to develop unsupervised learning strategies capable of effectively navigating the challenges posed by imbalanced datasets, thereby ensuring more accurate and efficient clustering outcomes in scenarios characterized by abbreviation ambiguity.

When researchers are looking for specific documents, there is an increasing interest in unsupervised learning methods that reduce the need for dictionaries or training data and eliminate the learning process [14]. Among these, clustering is evaluated as providing results similar to manual classification [14, 19]. The clustering process consists of steps such as data pre-processing, feature extraction, similarity measurement, execution of clustering algorithms, and evaluation, with various approaches existing at each stage [5, 9, 10]. Ongoing clustering research focuses on determining the appropriate number of clusters [20], resolving ambiguity in word meanings

[15-17], and handling imbalanced data [21]. Hence, the purpose of this study was to examine the effect of imbalanced dataset conditions on the accuracy of clustering results and to identify methods for enhancing these outcomes. Hence, the purpose of this study is to examine the effect of imbalanced dataset conditions on the accuracy of clustering results and to identify methods for enhancing these outcomes. In addressing the complexities introduced by imbalanced datasets, this research highlights the approach of modifying data distribution as a strategy to mitigate initial clustering challenges. Through the exploration of re-clustering methods, this study seeks to provide insights into achieving more balanced and representative clusters, offering a nuanced understanding of how to navigate the challenges associated with imbalanced data conditions.

Given the challenges and importance of abbreviation disambiguation, this study delved into the specific case of the abbreviation "MSA". The paper is organized as follows: Section 2 reviews related works. Section 3 details the research methodology. Section 4 discusses the results, and Section 5 concludes the paper and suggests directions for future research.

2 RELATED RESEARCH

2.1 Word Sense Disambiguation of Abbreviations

In the field of scientific research, the use of abbreviations is prevalent. This stems from both conventionally used abbreviations and those formed temporarily. It is essential to accurately grasp the meaning of these abbreviations, which leads to the issue of word sense ambiguity. In the field of natural language processing, the problem of word sense ambiguity has been a major research topic for a long time [9].

The challenge of word sense ambiguity has traditionally been approached using either rule-based methods with a set rule pattern, drawing from lexicons or ontologies, or statistical methods [10-13]. Firstly, rule-based methods have been in use since their emergence in the 1960s [11, 12]. This approach requires a predefined rule base to ascertain the meaning of abbreviations. If an abbreviation is not found within this rule base, a new rule has to be crafted and appended [13]. Secondly, statistical methods are classified into supervised and unsupervised learning depending on whether sense-tagged corpora are used for training [10]. With advancements in computational capabilities and artificial neural network methodologies, supervised learning methods are becoming increasingly popular in resolving word ambiguity [5, 6]. While supervised methods typically yield better results than unsupervised ones, generating training data and undergoing the training and evaluation process is time-consuming and costly. Recent attention has turned to unsupervised methods, which approach the problem without the need for lexicons or pre-training [14]. It is believed that unsupervised learning methods are apt for efficiently and accurately retrieving desired documents from a mixed set retrieved by an abbreviation search.

In the research on word sense disambiguation, unsupervised learning methods are widely used, among which clustering methods are predominantly employed [15-17]. While clustering methods have shown promise in abbreviation disambiguation, their broader application in document clustering also poses unique challenges.

2.2 Document Clustering

Clustering methods, while challenged by determining the number of clusters in real-world environments and handling high-dimensional data, are frequently employed to group documents with similar characteristics together [20]. This technique is applied across various research fields and is particularly useful for categorizing and searching specific documents without a distinct training phase [2]. In the field of systematic reviews, text mining techniques like clustering are assessed as potentially enhancing work efficiency and reducing workload [2-4].

The clustering of text documents follows a pipeline that includes text pre-processing, feature extraction, document similarity measurement, clustering algorithms application, and evaluation [28-30]. The first step, text pre-processing, is a process of cleaning and structuring the text data to enhance the model's performance. This procedure is carried out before implementing the model algorithm. The second step, feature extraction, involves identifying meaningful information from the text and effectively representing this data for model input. The third step, measuring similarity, evaluates the meaning and context encapsulated in the text data, enabling a comparison between different text data. The fourth step involves clustering algorithms, which group the input data into clusters based on similar characteristics. The fifth and final step is evaluation, which reviews the results of the clustering. If needed, a separate process may be required to create ground truth for assessment purposes.

In the clustering pipeline, various techniques are employed for each step. During the data pre-processing phase, tasks such as Tokenization, Stop Words removal, Capitalization, Noise Removal, and Stemming are executed. In the feature extraction phase, characteristics from sentences are extracted using methods like Bag of Words (BoW) and Term Frequency-Inverse Document Frequency (TF-IDF). For the similarity measurement phase, methods including Euclidean Distance, Cosine Similarity, Jaccard Coefficient, and Pearson Correlation are available. Notably, cosine similarity is frequently employed for measuring the similarity between two documents [28, 30]. In the application phase of clustering, algorithms like k-nearest neighbors, hierarchical clustering, and particularly the popular k-means, are utilized [18, 28, 30]. During the evaluation phase, model assessment primarily involves metrics extracted from the confusion matrix, such as accuracy, precision, recall, and the $F1$-score [28-30].

For clustering documents into groups, the k-means clustering algorithm is widely used. It outperforms many other algorithms and is employed in various research areas. It has demonstrated good performance in studies classifying articles in large-scale distributed environments [33]. Among its drawbacks is the need to preset the number of clusters and model parameters. However, methods such as the Elbow method and Silhouette Coefficient have been introduced to address these challenges [34]. The effectiveness of clustering algorithms, such as k-means, is

often influenced by the distribution of data, a factor particularly crucial in the context of imbalanced data sets.

2.3 Imbalanced Data Set

Imbalanced data refers to situations where the distribution of data across categories is skewed, resulting in certain categories having either very few or a disproportionately large number of data [22]. The problems posed by imbalanced data have been highlighted in various studies for quite some time [21-24]. Issues arising from imbalanced data include bias in the training data, degradation of model performance, and a mismatch between training and test data [21, 22]. To address imbalanced data, research has been conducted on sampling methods [25] and improvements in algorithms, but further investigation is still needed [21].

Research on imbalanced data spans various approaches, such as sampling methods and algorithmic improvements, which are applied in both supervised and unsupervised learning contexts [21, 24-26]. In supervised learning, research to address imbalanced data has focused on data preparation techniques like resampling, feature selection and extraction, improvements in learning processes like cost-sensitive learning, and enhancements in ensemble and classification algorithms [26]. Supervised learning methods require a labelling process for the creation of training data. In large data environments, the data labelling process can be time-consuming and costly [26]. In situations with a growing volume of data or where labelling operations are unfeasible, it is challenging to apply supervised learning. In the realm of imbalanced data research, unsupervised learning is particularly influenced when there is a skewed distribution of data [21].

Research conducted using unsupervised learning includes studies where, during data pre-processing, clustering is employed to find the center of the data and balance it through sampling [25], investigations are aimed at computing sub-datasets to address imbalanced data issues [23], and efforts are exerted to create subclusters, evaluate them, and subsequently merge to determine an appropriate number of clusters [24]. Despite these advances, there are still areas needing further exploration. These include indices for determining the right number of clusters, enhancements to clustering algorithms, clustering of minority groups, development of specialized classifiers, and detection of imbalanced data [21].

Various methods have been proposed to address imbalanced data, with adaptive competitive learning emerging as one of the effective approaches. The fundamental principle of competitive learning is the dynamic updating of cluster centers based on data characteristics [27]. This dynamic center-updating mechanism ensures that centers representing each cluster can be effectively identified even in imbalanced data. Additionally, the implementation of the self-adaptive algorithm allows for adaptation to changes in data distribution, thus better capturing the characteristics of imbalanced data. Applying this algorithm to clustering imbalanced data has been shown to yield beneficial results [24].

Research on how data imbalance affects clustering accuracy is limited. Although document clustering has been explored extensively, studying clustering for screening specific documents, like in the PRISMA Guide, remains underexplored, particularly in scenarios of imbalanced data. This paper presents an empirical test of clustering outcomes on sample data with varying data ratios and proposes an efficient clustering method for retrieving desired documents.

3 RESEARCH METHODS

When conducting research, researchers often face the challenge of sifting through a plethora of documents when they use abbreviations in their searches, aiming to pinpoint the truly relevant ones. The use of abbreviations in searches can result in a collection of documents that span a wide range of topics, mainly due to the inherent ambiguity of abbreviations. Clustering, an unsupervised learning method, has been put forward as a solution to mitigate this ambiguity. Yet, if the distribution of these document groups is skewed, isolating the desired documents becomes a non-trivial task. This study examined clustering outcomes according to data proportions and proposed a re-clustering method, as illustrated in Fig. 1, to enhance results in scenarios involving imbalanced data.

The research procedure unfolded in two principal phases: clustering and re-clustering. Each phase encompassed a series of steps such as sampling, clustering, and evaluation. The document type specified as pertinent for researchers within this study was "Microservice Architecture", with the abbreviation search keyword being "MSA". Documents procured via this abbreviation search were subjected to a labeling process to establish the full text of "MSA" as the ground truth value. Collections of documents sharing the same ground truth value were categorized as groups. Among these, the group designated as "Microservice Architecture" corresponds to the target group of interest. To examine the clustering results in contexts of imbalanced data the core aim of the study 19 samples were selected based on the data ratio of the target group and five based on the number of groups, cumulating in 95 distinct sampling scenarios. Initially, clustering was performed according to the sampling scenarios. Subsequently, specific clusters from the initial results were excluded from the sampling for re-clustering. Comprehensive details for each phase of the study will be provided in subsequent sections.

3.1 Collection and Labelling

The clustering process targeted documents related to "Microservice Architecture" identified through abbreviation searches. Microservice Architecture is frequently adopted as an architecture that facilitates agile and convenient scalability of applications in cloud environments. The abbreviation used for Microservice Architecture is "MSA". When searching "MSA" on Wikipedia, one can find various terms, such as "Multiple sequence alignment", "Maritime Safety Agency", "Microservice architecture", and "Measurement systems analysis". This study has chosen "MSA" as the search keyword.

For the period between 2010 and 2022, a search was conducted on the Web of Science (WOS) using the

keyword "MSA". Since the goal was to find papers related to Microservice Architecture, WOS categories that included the term "Computer", such as "Computer Science Information Systems", "Computer Science Artificial Intelligence", and "Computer Science Theory Methods", were selected. This search criterion yielded a total of 1521 papers. Data on the authors, titles, abstracts, and keywords of these papers were collected.

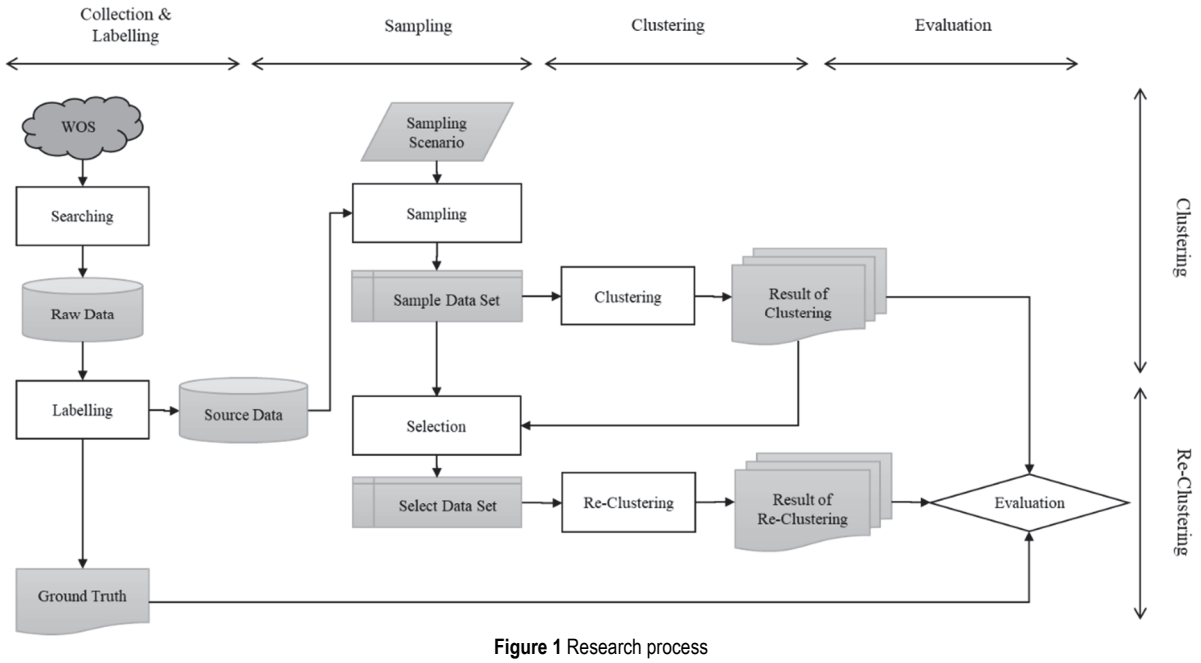


Figure 1 Research process

The full text "MSA" within the collected papers was examined, with initial emphasis placed on reviewing the title, abstract, and keywords. If the full text could not be discerned from these sections, the main content of the paper was examined. The distribution of the full texts for "MSA" from the 1521 papers is defined as groups, and this distribution is illustrated in Tab. 1. Papers were excluded from the analysis if "MSA" featured in author information or references, or if the abbreviation's meaning was unclear, corresponding to entries No. 16 and No. 17 in Tab. 1. The final primary dataset comprised 1076 papers. Any group with fewer than nine papers was categorized under "Others." The largest group, with 292 papers, pertained to "Multiple Sequence Alignments", and the "Microservice Architecture" group contained 68 papers.

Table 1 Numbers of papers by category

No.	Category	Number of papers
1	Multiple Sequence Alignments	292
2	Modern Standard Arabic	152
3	Microservice Architecture	68
4	Microstrip Antenna	56
5	Measurement System Analysis	22
6	Min-Sum Algorithm	19
7	Moth Swarm Algorithm	17
8	Multi-Head Self-Attention	15
9	Method of Successive Averages	15
10	multimodal sentiment analysis	14
11	multi-scale attention	14
12	Maritime Situational Awareness	13
13	Multiple system atrophy	10
14	Multi-Scale Auto convolution	9
15	Others	360
16	MSA (Author or Reference)	279
17	Not found	166

The full text of "MSA" was typically indicated when it was indicated in the abstract using parentheses, as in "MSA (Microservices Architecture)". The title, keywords, abstract, and main content were examined in sequence to determine the appropriate full text for "MSA" in the absence of such parenthetical full-text indications. If the context where "MSA" was used could not be identified, it was categorized as "not found". The identified full text of "MSA" was added to the collected data and used as the ground truth value.

3.2 Sampling

A sample dataset was constructed from the gathered data for analyzing clustering results, considering both the number of groups and the distribution of the target group. The sample dataset was constructed by selecting samples based on the number of groups and the proportion of data in the target group. Fig. 2 illustrates the structure of the sampling scenario, which is determined by the number of groups and the data ratio of the target group.

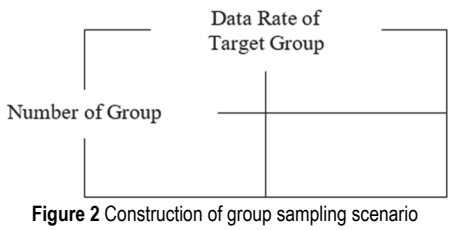


Figure 2 Construction of group sampling scenario

The number of groups was determined by keeping the target group, "Microservices Architecture", fixed and adding groups based on their count. Scenario 1 incorporated "Multiple Sequence Alignments", Scenario 2 included both "Multiple Sequence Alignments" and "Modern Standard Arabic", Scenario 3 expanded to encompass "Multiple Sequence Alignments", "Modern Standard Arabic", and "Microstrip Antenna", while

Scenario 4 further integrated "Measurement System Analysis". For Scenario 5, the target group "Microservices Architecture" remained unchanged, and data was randomly drawn from all other groups excluding the target group. Tab. 2 displays the proportions of each group per scenario, introducing the Group Distribution Rate (GDR) to quantify each group's relative size within the overall dataset for a given scenario. The GDR, adjusting for the varying number of groups across different scenarios, guides the proportional sampling of data across each group according to their respective sampling scenarios, ensuring an accurate reflection of document distribution.

Table 2 Data rate of group by scenario

Scenario (Number of Groups)	Group of Ground truth	Count	GDR
Scenario 1 (2)	Microservice Architecture	68	19%
	Multiple Sequence Alignments	292	81%
Scenario 2 (3)	Microservice Architecture	68	13%
	Multiple Sequence Alignments	292	57%
	Modern Standard Arabic	152	30%
Scenario 3 (4)	Microservice Architecture	68	12%
	Multiple Sequence Alignments	292	51%
	Modern Standard Arabic	152	27%
	Microstrip Antenna	56	10%
Scenario 4 (5)	Microservice Architecture	68	12%
	Multiple Sequence Alignments	292	49%
	Modern Standard Arabic	152	26%
	Microstrip Antenna	56	9%
	Measurement System Analysis	22	4%

The target group data proportion was set based on the "Microservices Architecture" group, increasing in increments of 5% up to 95%. Data extraction from the target group was conducted by random sampling according to the designated proportion. For groups other than the target group, sampling was performed uniformly at random. In Tab. 3, the number of data samples for each sampling scenario is shown, with a focus on the Scenario Inclusion Rate (SIR). The SIR quantifies the representation of the target group's samples within each scenario relative to the overall sample count, providing insight into the dataset's diversity and balance. By indicating the ratio of target group samples to total samples collected, this metric enables evaluation of each scenario's ability to capture the dataset's features and handle any imbalances.

In this study, a targeted sampling method was employed to create balanced sample datasets for various scenarios, ensuring fair representation of each group within the samples. This approach involved setting a fixed number of data points for the target group, followed by proportionally selecting data from the remaining groups to achieve a 1:1 ratio. Specifically, in Scenario 2, where the target group's data ratio is set at 5%, 16 data points were randomly chosen from the target group to match this ratio. To ensure a balanced distribution, 304 data points were randomly selected from the other two groups, not including the target group, resulting in a total sample dataset of 320 data points. This strategy ensured that the dataset for Scenario 2 accurately reflected the intended 5% data ratio for the target group, generating balanced datasets as intended for comprehensive analysis of clustering performance across different data distribution scenarios.

Table 3 Initial clustering distribution of target group documents in sample data by scenario

	Scenario Inclusion Rate (SIR) (Count of documents in target group / Total Count of documents)																		
Scenario (Group's Number)	5%	10%	15%	20%	25%	30%	35%	40%	45%	50%	55%	60%	65%	70%	75%	80%	85%	90%	95%
Scenario 1 (2)	15/ 307	32/ 324	52/ 344	68/ 340	68/ 272	68/ 226	68/ 194	68/ 170	68/ 151	68/ 136	68/ 124	68/ 113	68/ 105	68/ 97	68/ 91	68/ 85	68/ 80	68/ 76	68/ 72
Scenario 2 (3)	16/ 320	34/ 338	55/ 359	68/ 340	68/ 276	68/ 226	68/ 194	68/ 170	68/ 150	68/ 136	68/ 124	68/ 114	68/ 104	68/ 98	68/ 92	68/ 84	68/ 80	68/ 76	68/ 72
Scenario3 (4)	9/ 177	19/ 187	30/ 198	42/ 210	56/ 224	68/ 227	68/ 194	68/ 170	68/ 149	68/ 137	68/ 125	68/ 113	68/ 104	68/ 98	68/ 92	68/ 86	68/ 80	68/ 77	68/ 71
Scenario 4 (5)	5/ 93	10/ 98	16/ 104	22/ 110	29/ 117	38/ 126	48/ 126	59/ 147	68/ 152	68/ 136	68/ 124	68/ 112	68/ 104	68/ 96	68/ 92	68/8 4	68/ 80	68/ 76	68/ 72
Scenario 5 (all)	53/ 1061	68/ 680	68/ 453	68/ 340	68/ 272	68/ 227	68/ 194	68/ 170	68/ 151	68/ 136	68/ 154	68/ 113	68/ 104	68/ 97	68/ 91	68/ 85	68/ 80	68/ 76	68/ 72

3.3 Clustering and Evaluation

The clustering of documents in the sample data set involved a four-step procedure: Cleansing, Vectorization, Similarity Assessment, and Clustering Execution. Fig. 3 outlines the methodology used to isolate the desired target group from the mixture of groups.

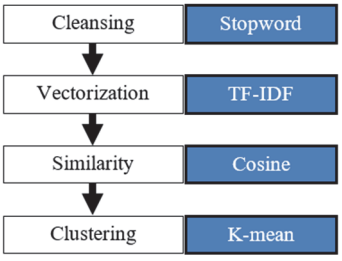


Figure 3 Process for clustering in the sample dataset

For the composition of a corpus pertaining to each paper within the sample data set, the title, abstract, and keywords were combined. A cleansing process was then applied to this combined corpus to filter out any non-pertinent words, thereby highlighting the significant terms. In the subsequent vectorization phase, the TF-IDF (Term Frequency-Inverse Document Frequency) method was implemented to evaluate the importance of each term within the corpus, considering the term's frequency against the backdrop of its inverse document frequency. With the TF-IDF values ascertained for every term, the calculation of document similarity was possible. The cosine similarity metric was employed for this purpose, leveraging the cosine angle as a measure of similarity between document vectors. Finally, the clustering step incorporated the k-means algorithm, a widely recognized method for clustering analysis.

To extract features from documents, TF-IDF is a method that computes a word's importance in a specific document by multiplying its frequency with its inverse document frequency [31]. The TF-IDF method is a commonly employed approach to represent documents as vectors due to its simplicity and widespread use. While alternative techniques have been explored, the TF-IDF remains a prevalent method in document classification and clustering studies to this day [32, 33]. The computation for TF-IDF is as follows [28].

$$TF(t) = \frac{\text{Number of times term tapprears in a document}}{\text{Total number of termsin the documents}} \quad (1)$$

$$IDF(t) = \log_d \frac{\text{Total number of documents}}{\text{Number of documents}} \quad (2)$$

$$TF - IDF = TF(t) \times IDF(t) \quad (3)$$

The cosine similarity measure is used to calculate the similarity between two documents [28]. The cosine similarity expression, which uses the cosine angle between two vectors to determine the correlation between them, is as follows [28].

$$cosinesimilarity(d_1, d_2) = \frac{d_1 \cdot d_2}{\|d_1\|_2 \ \|d_2\|_2}$$

The operation indicates dotproduct, (4)

$\|d_2\|_2$ is $L2$ norm

The clustering process was conducted using the scikit-learn library within Python programming. For the vectorization process, the 'TfidfVectorizer' function provided by the 'scikit-learn' library was utilized. The k-means clustering was also performed using the "K-Means" module in 'scikit-learn'. To run k-means, the user must define the number of clusters, also known as the k-value.While the k value can be determined by assessing the SSE (Sum of squared errors) over a range of k values or using the Silhouette Coefficient, this research set the k-value based on the predetermined number of clusters in each sample scenario. For instance, in Scenario 1 with two clusters, the k value was set to 2, while in Scenario 2 with three clusters, the k value was set to 3.

In this study, the evaluation of document clustering effectiveness was enhanced by employing a comprehensive set of metrics, including Purity, Entropy, Accuracy, $F1$-Score, and computation times (CPU time and memory consumption). Purity measures cluster quality by assessing the dominance of a single category within each cluster [37].

$$Purity = \sum_{r=1}^{n} \frac{1}{n} max_i \left(n_r^i\right) \quad (5)$$

Entropy assesses the distribution diversity within clusters by measuring the information content of their compositions [37].

$$Entropy = \sum_{r=1}^{n} \frac{n_r}{n} \left(-\frac{1}{\log q} \sum_{i=1}^{q} \frac{n_r^i}{n_r} \log \frac{n_r^i}{n_r} \right) \quad (6)$$

Accuracy and $F1$-score were computed using confusion matrices. While accuracy provides a general measure of performance, the $F1$-score is particularly informative in cases of imbalanced data distributions, as it considers both precision and recall [28, 33]. The metrics used are as follows.

$$accurancy = \frac{TP + TN}{TP + FP + TN + FN} \quad (7)$$

$$precision = \frac{TP}{TP + FP} \quad (8)$$

$$recall = \frac{TP}{TP + FN} \quad (9)$$

$$F1 - \text{score} = 2 \times \frac{precision \times recall}{precision + recall} \quad (10)$$

where: TP is true postitive, TN is true negative, FP is false positive, FN is false negative.

The results of clustering showed variations based on changes in the sample scenarios. Notably, the number of groups and the data ratio significantly influenced the accuracy and $F1$-score of the clustering. To analyze this, balanced data ratios were calculated by considering the number of instances in each group. The calculation formula is provided as Eq. (9). For instance, in Sample Scenario 1, which has two clusters, a balanced data ratio of 50% was derived. Similarly, in Sample Scenario 4 with five clusters, a balanced data ratio of 20% was computed.

$$\text{balanced data ration} = \frac{1}{\text{Count in cluster}} \quad (11)$$

This research employed the TF-IDF algorithm for feature extraction and cosine similarity to measure inter-document relatedness, while the K-means algorithm was applied for cluster analysis. The clustering outcomes were evaluated using a confusion matrix to derive accuracy and $F1$-score metrics, particularly focusing on the 'Microservice Architecture' category to segregate relevant clusters from unrelated ones.

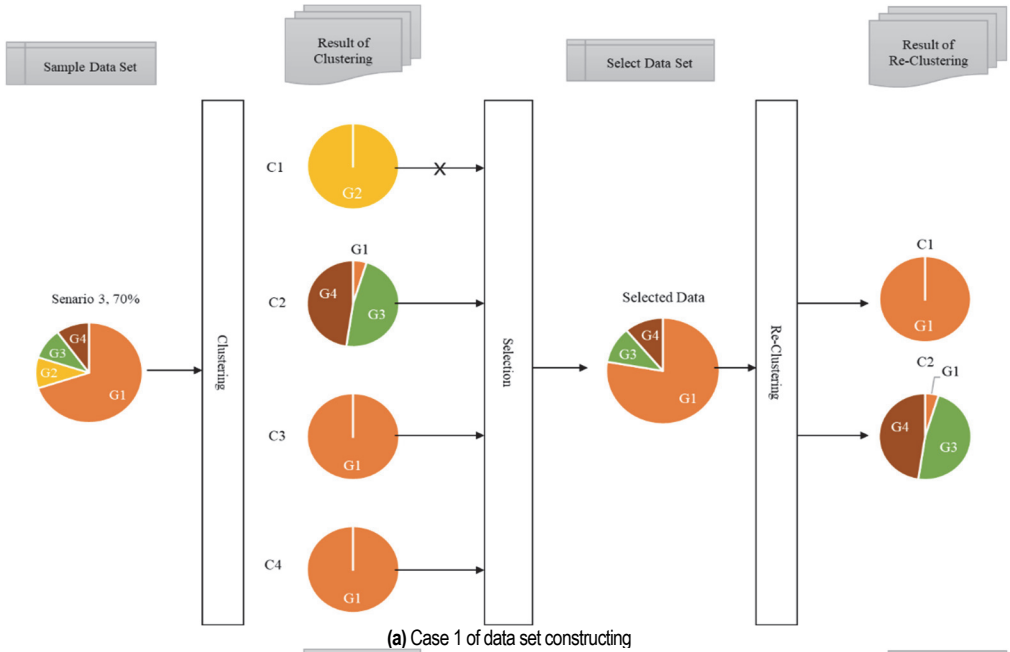
To assess the performance of the algorithm, both CPU time and memory usage were measured using Python's profiling tools. CPU time was captured using the '%% time' magic command, which provided an accurate measurement of the algorithm's execution time excluding any wait times, ensuring that the measurement reflected pure processing time. Memory usage was evaluated using the '%% memit' magic command, which measured the increase in memory consumption during the algorithm's execution. This approach allowed for a comprehensive analysis of the algorithm's efficiency in terms of processing speed and resource consumption. The tests were conducted in a cloud computing environment, which offered a scalable and

controlled setting for evaluating the algorithm's performance under various computational loads.

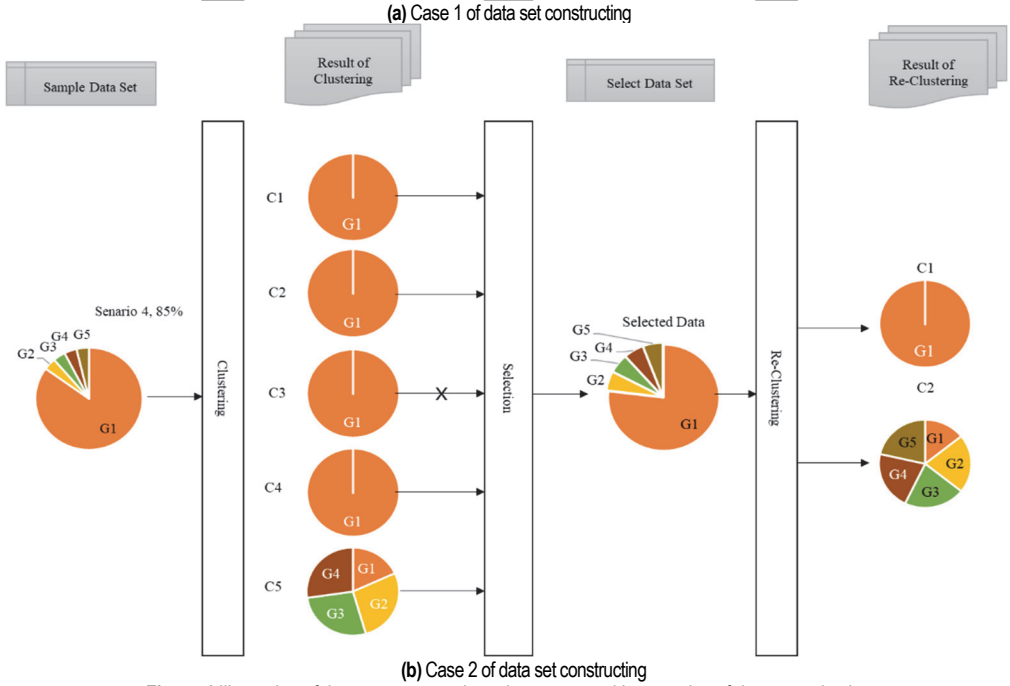
3.4 Re-clustering

After clustering, if the target group was not identifiable, refinement of the clustering results was undertaken. To isolate the target group, the volume of target data was reduced by excluding certain clusters based on the initial clustering outcomes. The dataset was adjusted by discarding one of the clusters from the clustering results, under the assumption that it would be discernible which clusters were extraneous to the research aims. Typically, clusters unrelated to the target group were omitted. However, in scenarios where the data ratio for the target group was significantly high, clusters containing the target group were removed. This strategy is predicated on the premise that a high data ratio of the target group may induce the uniformity effect [21], resulting in the target group being divided into sub-clusters.

In the process of refining the clustering parameters, an elbow analysis was conducted using Distortion and Silhouette Score to determine the optimal number of clusters (k). However, it is important to note that the sample dataset used for clustering was manipulated to reflect specific clustering values, rather than following the k values suggested by the elbow analysis. This deviation was intentional, as the sample data set's clustering values were applied directly to accommodate the unique distribution and characteristics of the manipulated sample datasets. This approach allowed for a more tailored analysis of clustering performance, particularly in scenarios where standard parameter tuning methods might not accurately capture the nuances of the dataset's structure.



(a) Case 1 of data set constructing



(b) Case 2 of data set constructing

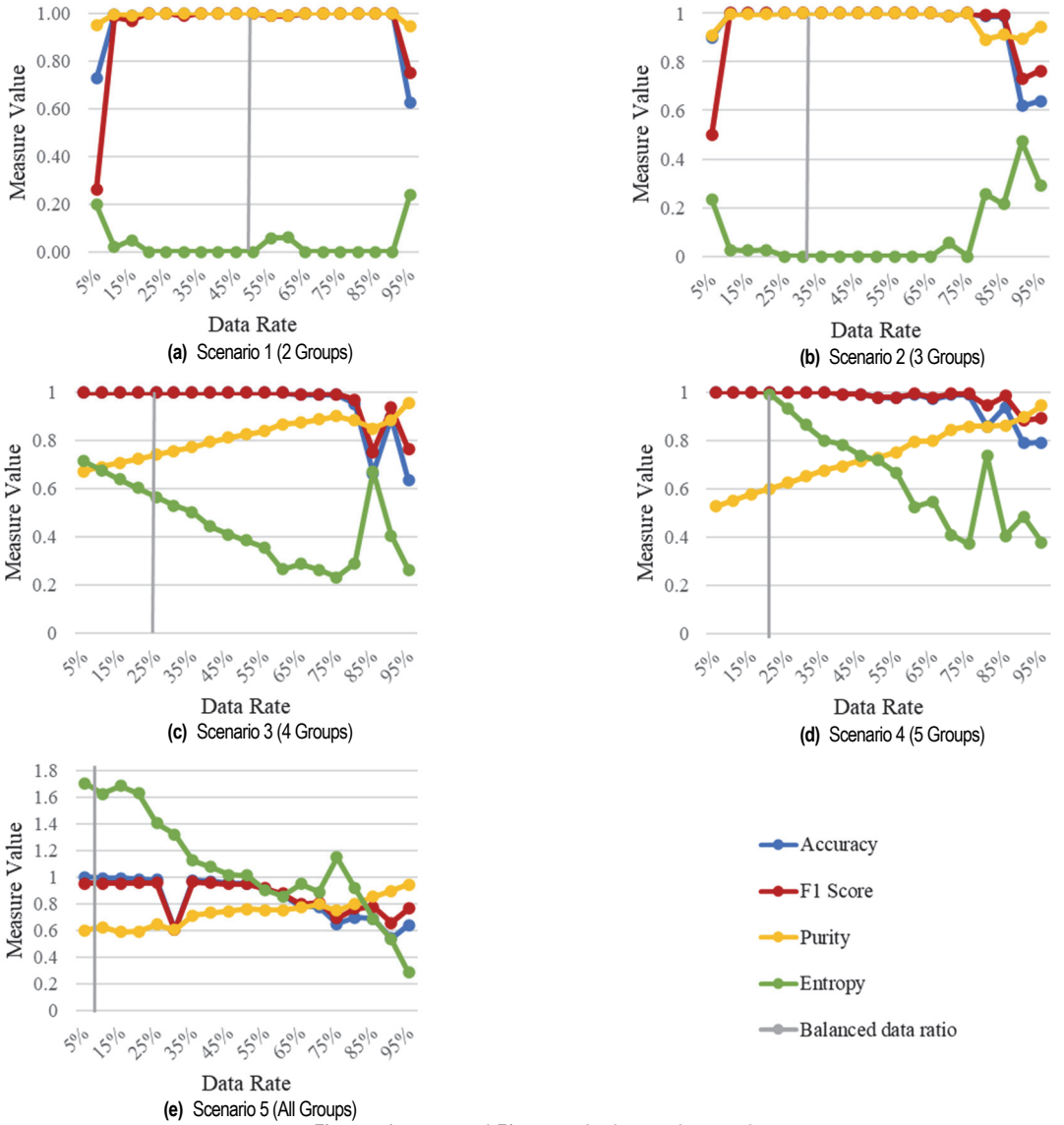
Figure 4 Illustration of the two-stage re-clustering process with examples of data set selection

Fig. 4 demonstrates the procedure of curating a dataset derived from clustering results, particularly when the target group is identified as G1. In Case 1, cluster C1, which emerged from the clustering outcomes, lacks the target group. Typically, in most scenarios, clusters that do not encompass the target group, such as C1, are removed when assembling a dataset for re-clustering. Meanwhile, in Case 2, cluster C3 from the clustering outcomes contains the target group. In scenarios where the data ratio of the target group is significantly high, clusters like C3 that incorporate the target group are omitted. This approach predominantly applies to high target group data ratio scenarios. Subsequent to these exclusions, re-clustering is performed to more accurately identify the target group. The re-clustering process followed the same four steps as the clustering process: Cleansing, Vectorization, Similarity, and Clustering. The results of re-clustering were measured using a confusion matrix to calculate Accuracy and *F*1-score values. The evaluation was conducted based on the target groups, similar to the evaluation of the clustering method.

4 STUDY RESULTS

4.1 Imbalance Rate Data

Examination of clustering results based on balanced data ratios revealed that the accuracy and *F*1-score values remained stable as the data ratio of the target group approached the balanced data ratio. It was observed that a significant deviation from the balanced data ratio resulted in a degradation of both accuracy and *F*1-score values. These analytical findings are shown in Fig. 5. Fig. 5 presents graphs depicting accuracy, *F*1-score, purity, and entropy values per sample scenario, with respect to data ratios. The *x*-axis of each graph represents the data ratio of the target group, while the y-axis represents accuracy and *F*1-score values. The grey baseline in each graph represents the scenario-specific balanced data ratio. Each graph is divided into scenarios based on the number of included groups in the sample data. Scenario 1 includes two groups, while Scenario 2, 3, and 4 encompass three, four, and five groups respectively. Scenario 5 includes sampling all groups.



(a) Scenario 1 (2 Groups)
(b) Scenario 2 (3 Groups)
(c) Scenario 3 (4 Groups)
(d) Scenario 4 (5 Groups)
(e) Scenario 5 (All Groups)

Figure 5 Accuracy and *F*1-score value by sample scenario

In particular, within Sample Scenario 5, a decline in performance was observed as the data ratio of the target group exceeded 20%. This phenomenon was apparent in scenarios with a higher number of groups, and it became

evident that as the data ratio of specific groups increased, the *F*1-score demonstrated a more sensitive response compared to accuracy. The clustering results for imbalanced data can be summarized into three main observations. Firstly, regardless of the number of groups, imbalanced data ratios adversely impact the accuracy of clustering. Particularly, as the number of groups increases and the data ratio of the target group becomes higher, there is a tendency for accuracy to decline. This suggests that clustering performance is suboptimal when specific group data ratios are high. Secondly, clustering performance was compared based on balanced data ratios. It was found that significant deviations in the data ratio of the target group from the balanced data ratio led to decreases in both accuracy and *F*1-score. Lastly, the results from Sample Scenario 5 highlight the performance degradation of clustering when the data ratio of the target group is high. Notably, the *F*1-score demonstrates a more sensitive response than accuracy in such cases.

Further analysis incorporating purity and entropy metrics provided additional insights into the clustering performance across different scenarios. It was observed that scenarios 1 and 2 yielded identical results, indicating a stable clustering outcome when the data distribution was relatively balanced. However, scenarios 3, 4, and 5 presented contrasting results, with purity values indicating instability as the proportion of non-target groups increased. This instability in purity suggests that a higher presence of data from non-target groups contributes to less homogeneous clusters. Conversely, entropy showed opposite trends to purity, with lower values indicating more favorable clustering outcomes. This inverse relationship between purity and entropy underscores the complexity of achieving optimal clustering in imbalanced datasets. The findings suggest that while purity becomes unstable with an increase in non-target group data, aiming for lower entropy can lead to more defined and well-separated clusters, enhancing the overall clustering quality.

These observations underscore the nuanced impact of data distribution on clustering metrics beyond traditional measures such as accuracy and *F*1-score. The interplay between purity and entropy highlights the importance of considering multiple dimensions of clustering performance, especially in scenarios characterized by significant data imbalances.

The experimental results revealed that both the number of groups and the data ratio impact the performance of clustering. In the case of imbalanced data, clustering outcomes might not be accurate, and to address this, attempting re-clustering could enhance performance. The following section delves into a detailed analysis of the re-clustering results.

4.2 Re-clustering Results for Clusters not Identifying the Target Group

The performance deteriorated as data ratios deviated from the balanced data ratio in the clustering results. To enhance the identification performance of the target group, a re-clustering process was conducted by reconstructing the target dataset by excluding certain clusters from the clustering results of scenarios with degraded performance. The re-executed sample scenarios and data counts are presented in Tab. 4. In scenarios 1 and 2, where the clustering results failed to identify the target group, it was unable to identify the target group at data ratios of 5% and 10%, as well as at ratios exceeding 95% and 80%, respectively. In scenario 3, identification failure occurred at data ratios of 65% and above, and in scenario 4, it was at data ratios of 50% and above. For the purpose of re-clustering, the selected dataset for reconstruction excluded clusters in which the target group was not present in the clustering results.

Table 4 Secondary clustering distribution of target group documents in sample data by scenario

	Rate of Data (Count of documents in target group / Total Count of documents)																		
Scenario	5%	10%	15%	20%	25%	30%	35%	40%	45%	50%	55%	60%	65%	70%	75%	80%	85%	90%	95%
Scenario 1	15/161	32/177	-	-	-	-	-	-	-	-	-	-	-						49/53
Scenario 2	16/181	34/187	-	-	-	-	-	-	-	-	-	-	-			57/73	57/69	40/48	36/40
Scenario3	-	-	-	-	-	-	-	-	-	-	-	-	68/92	68/88	68/84	68/80	39/51	19/28	49/52
Scenario 4	-	-	-	-	-	-	-	-	-	68/119	68/110	68/101	68/95	68/89	68/86	68/80	40/52	45/53	50/54
Scenario 5	-	-	-	68/292	68/224	68/206	68/158	68/139	68/123	68/115	68/104	68/96	68/92	68/88	48/71	52/69	49/61	55/63	56/60

Fig. 6 illustrates a graph displaying accuracy and *F*1-Score values before and after re-clustering. The *x*-axis represents the measured values before re-clustering, while the *y*-axis represents the measured values after re-clustering. The points on the graph are denoted by blue for accuracy and red for *F*1-Score. The upper blue region above the diagonal line indicates an improvement in re-clustering results, while the lower red region below the diagonal line signifies a decline in re-clustering results. Labels for the points in the lower region indicate the scenario number and the data ratio of the target group. The overall outcome of the before-and-after results indicates an enhancement in the identification of the target group through re-clustering.

In summary, the re-clustering results reveal that firstly, accuracy and *F*1-score values improved in most sample scenarios. However, in some scenarios, the opposite outcome was observed. Particularly, in cases of extreme data ratios (either very high or very low), there was no improvement in performance even after re-clustering. Secondly, the success rate of re-clustering was greatly influenced by the selection process of the target data and the balanced data ratio. In most cases, as data imbalance was alleviated, performance improved. However, in

scenarios with extremely high or low data ratios, the extent of performance enhancement was limited. Thirdly, the degree of performance improvement varied across scenarios. Specific scenarios, such as Scenarios 2 and 3, showed significant changes, while in cases like Scenario 5, there was relatively little to no performance improvement.

To enhance the performance through re-clustering, it is evident that careful consideration of data distribution and the number of groups is necessary. While re-clustering led to improved overall performance results, instances where the data of the target group was too abundant or too scarce exhibited a uniformity effect, resulting in diminished performance.

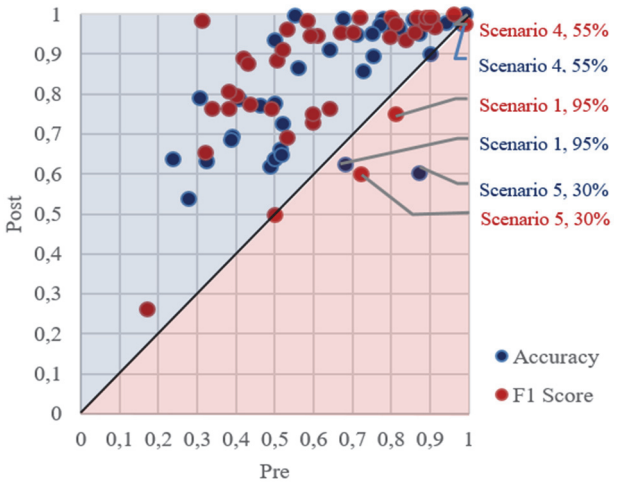


Figure 6 Comparative analysis of pre and post re-clustering accuracy and *F*1-score values

4.3 Comparative Analysis of Clustering Results

To evaluate its effectiveness in resolving abbreviation ambiguity and addressing imbalanced data issues, the proposed re-clustering method was compared with other clustering methods. The proposed re-clustering method involved removing clusters from the initial clustering results. Clustering was performed using k-means clustering, resulting in a cluster count of 2. The Self-adaptive competitive learning method, effective for imbalanced data, divided and then merged clusters. The cluster count was also set to 2.

A detailed comparison of clustering methods can be observed in Fig. 7. The x-axis of the graph represents the data ratio of the target group, while the y-axis displays performance metrics for each clustering method. The metrics are divided into accuracy and *F*1-score for each scenario. The scenarios were constructed by increasing the number of clusters. This was an attempt to analyze the data by subdividing it. As the number of clusters increases, the complexity of the data also increases. The comparison of performance evaluation metrics across different clustering methods is shown in Fig. 8 using a box plot. The length of the box plots increases as the number of scenarios increases, indicating the incorporation of more groups in the data distribution. The elongation of the data distribution demonstrates increased variability, which is caused by the imbalanced data state and supported by the presence of outliers. The relationship between purity and entropy values becomes inversely proportional as the scenario number increases, which is of particular significance. There is a noticeable decline in purity accompanied by an increase in entropy, indicating that both the uniformity and separateness of clusters decrease as the dataset becomes more imbalanced with the addition of more classes. This trend highlights the inherent difficulty of clustering in data conditions that are becoming more imbalanced. However, the re-clustering method stands out for its relatively stable performance, evidenced by its box plots showing minor variations despite the rising scenario numbers. The effectiveness of re-clustering in managing imbalanced datasets is emphasized by its adaptability, which offers a promising approach to improving clustering outcomes in complex data distributions.

When examining the results of re-clustering, it becomes evident that maintaining good performance on imbalanced data was possible when the number of clusters was low, specifically in Scenarios 1 and 2. It showed strong performance in areas with low data ratios. In contrast, with a higher number of clusters, such as in Scenarios 4 and 5, performance declined in regions with high data ratios. This likely results from the heightened clustering complexity, increasing sensitivity to imbalanced data.

In K-means clustering, performance fluctuations are not substantial in scenarios with a low number of clusters, but there is a gradual decline in performance as the number of clusters increases.

Overall, the self-adaptive competitive learning method displayed lower accuracy at lower data ratios. However, this method showed improved performance at higher data ratios in scenarios with more clusters. Notably, in Scenario 5, characterized by a high number of clusters and significant data imbalance, it demonstrated superior accuracy. This outcome, contrasting with other clustering methods, suggests that self-adaptive competitive learning may be particularly effective in specific scenarios of high data imbalance. While this does not universally indicate enhanced data segmentation, it highlights the method's unique capability to maintain performance under challenging conditions of data imbalance and high cluster counts.

The F-1 score consistently exhibited higher and more consistent values in comparison to accuracy in the dataset utilized for this study, signifying that the F-1 score is a more dependable performance metric in this particular context. As the number of clusters increased, the complexity of the data also increased, resulting in varying accuracy results across different clustering methods. Overall, the re-clustering consistently demonstrated stable performance across various scenarios and changes in data ratio. However, in scenarios with a higher number of clusters, such as Scenario 5, re-clustering displayed a decrease in accuracy at high data ratios, possibly due to heightened sensitivity towards data imbalance as data complexity increased. Interestingly, the self-adaptive competitive learning method exhibited strong performance in Scenario 5, particularly in cases with high data ratios, suggesting its potential as an effective approach in imbalanced data conditions. On the other hand, in imbalanced data situations, K-means clustering generally had lower accuracy. This could be attributed to the uniformity effect [21], which occurs when a predominant category in the dataset causes excessive subdivision of data

into sub-clusters, consequently negatively impacting the overall clustering accuracy. Among the three methods, K-means clustering demonstrated the greatest impact from the uniformity effect, suggesting limitations in its application in the presence of data imbalance.

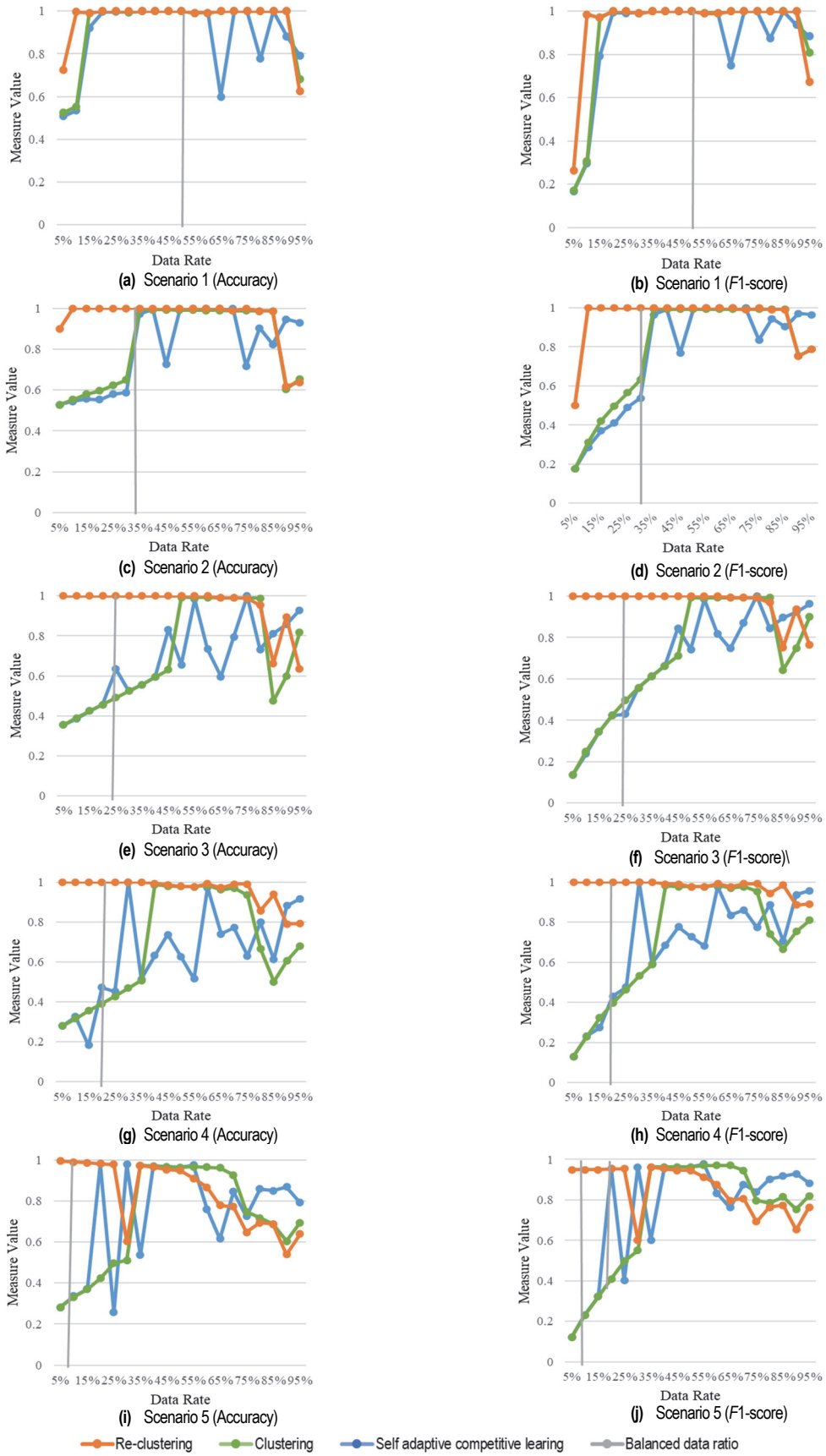


Figure 7 Comparison of Accuracy and *F*1-Score Across Five Scenarios by Clustering Method

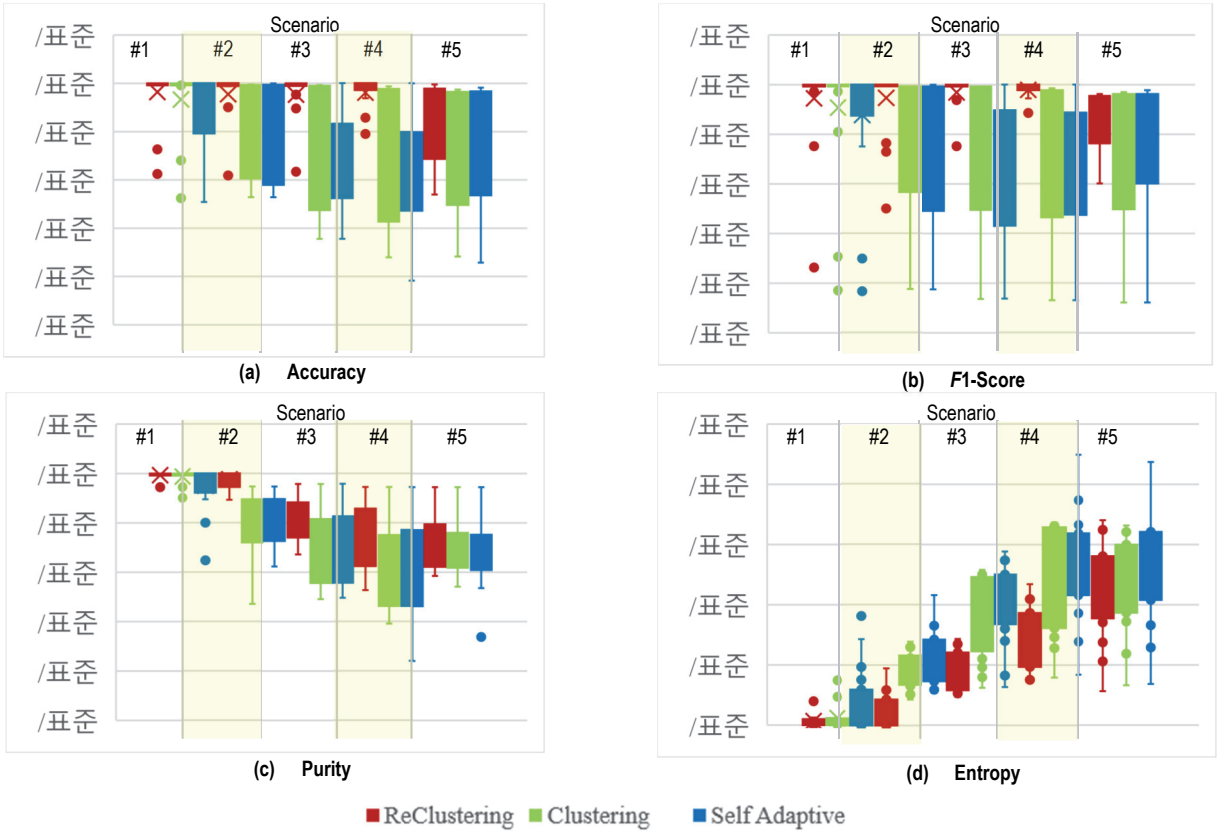


Figure 8 Box plot for comparison of accuracy, *F*1-score , purity, entropy by clustering method

SUMMARY

Method	Count	Sum	Average	Variance
ReClustering	10	1048.71	104.871	194.6489
Clustering	10	972.88	97.288	0.739084
Self Adaptive	10	971.56	97.156	0.469404

ANOVA

Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	390.1351	2	195.0676	2.987901	0.067263	3.354131
Within Groups	1762.717	27	65.28581			
Total	2152.852	29				

Figure 9 ANOVA analysis of memory increse by method

SUMMARY

Method	Count	Sum	Average	Variance
ReClustering	10	260.24	26.024	0.603382
Clustering	10	208.3	20.83	0.291222
Self Adaptive	10	654	65.4	2.711111

ANOVA

Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	11879.77	2	5939.887	4942.059	2.38E-35	3.354131
Within Groups	32.45144	27	1.201905			
Total	11912.22	29				

Figure 10 ANOVA analysis of CPU time by method

The performance evaluation of three distinct algorithms was conducted through the measurement of memory increase and CPU time, with the results subjected to ANOVA tests as depicted in Fig. 9 and Fig. 10. The analysis of memory consumption revealed no significant differences among the three algorithms, with a *p*-value exceeding 0.05. This outcome suggests that the memory requirements of these algorithms are comparable, likely due to the application of a single dataset across all clustering processes. Conversely, the CPU time analysis demonstrated significant differences among the algorithms, indicating variability in computational efficiency. The self-adaptive competitive learning approach required the most CPU time, followed by the re-clustering and then the standard clustering method. This ranking suggests that the complexity and computational demands of the self-adaptive competitive learning algorithm are higher compared to the other techniques. The increased CPU time for this method may reflect its intricate processing mechanisms, which, while potentially offering enhanced clustering accuracy or adaptability, also result in longer execution times.

Table 5 Result of ANOVA (Single Factor) by Method

Benchmarks	Scenario	*SS*	*df*	*MS*	*F*	*p*-value
Accuracy	#1	0.05	2	0.02	1.105	0.339
	#2	0.24	2	0.12	3.907	0.026
	#3	0.90	2	0.45	11.839	0.000
	#4	1.18	2	0.59	13.283	0.000
	#5	0.17	2	0.08	1.645	0.203
*F*1-Score	#1	0.04	2	0.02	0.453	0.638
	#2	0.35	2	0.17	2.981	0.059
	#3	0.94	2	0.47	9.072	0.000
	#4	1.03	2	0.52	10.110	0.000
	#5	0.16	2	0.08	1.448	0.244
Purity	#1	0.02	2	0.01	3.574	0.035
	#2	0.38	2	0.19	22.973	0.000
	#3	0.21	2	0.10	6.770	0.002
	#4	0.16	2	0.08	3.038	0.057
	#5	0.03	2	0.01	0.974	0.384
Entropy	#1	0.24	2	0.12	4.039	0.023
	#2	2.37	2	1.18	41.234	0.000
	#3	3.73	2	1.86	21.691	0.000
	#4	3.82	2	1.91	13.365	0.000
	#5	0.31	2	0.15	0.982	0.381

To compare the performance of three clustering algorithms, an ANOVA test was carried out in different

scenarios. The analysis showed notable differences in performance among the algorithms, particularly in Scenarios 2, 3, and 4. On the other hand, Scenarios 1 and 5 had comparable results in certain metrics, suggesting that the re-clustering method performed similarly well in these situations. However, it did not show statistically significant progress compared to other methods. In general, the results indicate that although all the algorithms tested have their strengths, the re-clustering method has the potential to enhance clustering outcomes in particular situations. This observation suggests that re-clustering may be a beneficial method for tackling specific challenges that were identified in the dataset configurations examined in this research. Tab. 5 displays the analysis's detailed outcomes, offering a thorough perspective on how the performance comparisons vary in various scenarios.

5 CONCLUSIONS

This research presented a re-clustering method designed for effectively clustering documents of target groups under imbalanced data scenarios, particularly focusing on abbreviation searches. The re-clustering method consisted of two steps. The first step involved a sampling process that constructed a new dataset by excluding specific categories from the initial clustering results. The second step involved a subsequent re-clustering of this newly formed dataset. The results indicated that the re-clustering method improved clustering results in imbalanced data environments. There were improvements in both accuracy and *F*1-Score values, with only a few specific cases as exceptions. The results of re-clustering were compared to those of other methods, revealing its superior performance. The re-clustering method demonstrated better results compared to k-means clustering and self-adaptive competitive learning methods. An analysis of variance (ANOVA) was used to identify differences in clustering results under different scenarios. This highlights the efficacy of the approach in confronting the obstacles presented by imbalanced data.

This approach has provided a versatile testing framework for assessing various clustering methods under imbalanced conditions. To explore the effects of clustering within imbalanced data environments, this study strategically modified data distributions and the number of groups in the datasets. The clustering results are significantly dependent on changes in the distribution of data based on the balanced data ratio. This observation highlights the importance of carefully determining the optimal number of clusters and accurately assessing the distribution of target groups to effectively cluster imbalanced datasets. The constructed sample datasets are a valuable resource for future research, as they provide a basis for determining the optimal number of clusters and analyzing data distribution states.

Future research should focus on broadening the variety of datasets and investigating the automation of method selection to improve the efficacy of identifying imbalanced conditions. Moreover, the exploration of incorporating unsupervised learning strategies to enhance document clustering methods shows potential. By studying the challenges related to parameter tuning and the assessment of data distribution techniques, including data visualization, this study will enhance its established foundation. Efforts should also be made to expand the research across a wide range of academic fields using various keywords to effectively generalize the findings. These efforts aim to address the issues of abbreviation ambiguity and imbalanced data, consequently improving the effectiveness and precision of document clustering and contributing to the broader field of unsupervised learning in data science.

Acknowledgements

This research was partially supported by the Seoul National University of Science&Technology in 2022 (Research Grant Number: 2022-1111). We extend our heartfelt thanks to the Seoul National Universityof Science & Technology for their generous support and contribution to our study.

6 REFERENCES

[1] Thelwall, M. & Sud, P. (2022). Scopus 1900-2020: Growth in articles, abstracts, countries, fields, and journals. *Quantitative Science Studies*, *3*(1), 37-50. https://doi.org/10.1162/qss_a_00177

[2] O'Mara-Eves, A., Thomas, J., McNaught, J., Miwa, M., & Ananiadou, S. (2015). Using text mining for study identification in systematic reviews: a systematic review of current approaches. *Systematic reviews*, *4*(1), 1-22. https://doi.org/10.1186/2046-4053-4-5

[3] Burns, J. K., Etherington, C., Cheng-Boivin, O., & Boet, S. (2021). Using an artificial intelligence tool can be as accurate as human assessors in level one screening for a systematic review. *Health Information & Libraries Journal*. https://doi.org/10.1111/hir.12413

[4] Gartlehner, G., Wagner, G., Lux, L., Affengruber, L., Dobrescu, A., Kaminski-Hartenthaler, A., & Viswanathan, M. (2019). Assessing the accuracy of machine-assisted abstract screening with DistillerAI: a user study. *Systematic reviews, 8*(1), 1-10. https://doi.org/10.1186/s13643-019-1221-3

[5] Kim, M. & Kwon, H. C., (2021). Word Sense Disambiguation Using Prior Probability Estimation Based on the Korean WordNet. *Electronics*, *10*(23), 2938. https://doi.org/10.3390/electronics10232938

[6] Popov, A. (2017). Word sense disambiguation with recurrent neural networks. *Proceedings of the Student Research Workshop associated with RANLP, 2017,* 25-34. https://doi.org/10.26615/issn.1314-9156.2017_004

[7] Na, S. T., Kim, J. H., Jung, M. H., & Ahn, J. E. (2016), Trend Analysis using Topic Modeling for Simulation Studies, *Journal of the Korea Society for Simulation (JKSS), 25*(3), 107-116. https://doi.org/10.9709/JKSS.2016.25.3.107

[8] Charbonnier, J. & Wartena, C. (2018). Using word embeddings for unsupervised acronym disambiguation. *Proceedings of the 27th International Conference on Computational Linguistics, Santa Fe, New Mexico, USA, August 20-26, 2018,* 2610-2619.

[9] Bevilacqua, M., Pasini, T., Raganato, A., & Navigli, R. (2021). Recent trends in word sense disambiguation: A survey. *International Joint Conference on Artificial Intelligence,* 4330-4338. https://doi.org/10.24963/ijcai.2021/593

[10] Navigli, R. (2009). Word sense disambiguation: A survey.*ACM computing surveys (CSUR), 41*(2), 1-69. https://doi.org/10.1145/1459352.1459355

[11] Karystianis, G., Thayer, K., Wolfe, M., & Tsafnat, G. (2017). Evaluation of a rule-based method for epidemiological document classification towards the automation of systematic reviews. *Journal of biomedical informatics, 70,* 27-34. https://doi.org/10.1016/j.jbi.2017.04.004

[12] Aubaid, A. M. & Mishra, A. (2020). A rule-based approach to embedding techniques for text document classification. *Applied Sciences, 10*(11), 4009. https://doi.org/10.3390/app10114009

[13] Park, Y. & Byrd, R. J. (2001). Hybrid text mining for finding abbreviations and their definitions. *Proceedings of the 2001 conference on empirical methods in natural language processing.*

[14] Ciosici, M., Sommer, T., & Assent, I. (2019). Unsupervised Abbreviation Disambiguation Contextual disambiguation using word embeddings. *arXiv preprint arXiv:1904.00929*.

[15] Navigli, R. & Lapata, M. (2009). An experimental study of graph connectivity for unsupervised word sense disambiguation. *IEEE transactions on pattern analysis and machine intelligence, 32*(4), 678-692. https://doi.org/10.1109/TPAMI.2009.36

[16] Xu, H., Wu, Y., Elhadad, N., Stetson, P. D., & Friedman, C. (2012). A new clustering method for detecting rare senses of abbreviations in clinical notes. *Journal of biomedical informatics, 45*(6), 1075-1083. https://doi.org/10.1016/j.jbi.2012.06.003

[17] Pelevina, M., Arefyev, N., Biemann, C., & Panchenko, A. (2017). Making sense of word embeddings. *arXiv preprint arXiv:1708.03390.* https://doi.org/10.18653/v1/W16-1620

[18] Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern recognition letters, 31*(8), 651-666. https://doi.org/10.1016/j.patrec.2009.09.011

[19] Zhong, Q., Zeng, G., Zhu, D., Zhang, Y., Lin, W., Chen, B., & Tang, J. (2021). Leveraging domain agnostic and specific knowledge for acronym disambiguation. *arXiv preprint arXiv:2107.00316.*

[20] Ezugwu, A. E., Ikotun, A. M., Oyelade, O. O., Abualigah, L., Agushaka, J. O., Eke, C. I., & Akinyelu, A. A. (2022). A comprehensive survey of clustering algorithms: State-of-the-art machine learning applications, taxonomy, challenges, and future research prospects. *Engineering Applications of Artificial Intelligence*, *110,* 104743. https://doi.org/10.1016/j.engappai.2022.104743

[21] Krawczyk, B. (2016). Learning from imbalanced data: open challenges and future directions. *Progress in Artificial Intelligence, 5*(4), 221-232. https://doi.org/10.1007/s13748-016-0094-0

[22] Sun, Y., Wong, A. K., & Kamel, M. S. (2009). Classification of imbalanced data: A review. *International journal of pattern recognition and artificial intelligence, 23*(04), 687-719. https://doi.org/10.1142/S0218001409007326

[23] You, C., Li, C., Robinson, D. P., & Vidal, R. (2018). Scalable exemplar-based subspace clustering on class-imbalanced data. *Proceedings of the European Conference on Computer Vision (ECCV)*, 67-83. https://doi.org/10.1007/978-3-030-01240-3_5

[24] Lu, Y., Cheung, Y. M., & Tang, Y. Y. (2019). Self-adaptive multiprototype-based competitive learning approach: a k-means-type algorithm for imbalanced data clustering. *IEEE transactions on cybernetics, 51*(3), 1598-1612. https://doi.org/10.1109/TCYB.2019.2916196

[25] Lin, W. C., Tsai, C. F., Hu, Y. H., & Jhang, J. S. (2017). Clustering-based under sampling in class-imbalanced data. *Information Sciences, 409,* 17-26. https://doi.org/10.1016/j.ins.2017.05.008

[26] Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H., & Bing, G. (2017). Learning from class-imbalanced data: Review of methods and applications. *Expert systems with applications, 73,* 220-239. https://doi.org/10.1016/j.eswa.2016.12.035

[27] Rumelhart, D. E. & Zipser, D. (1985). Feature discovery by competitive learning. *Cognitive science, 9*(1), 75-112. https://doi.org/10.1016/S0364-0213(85)80010-0

[28] Naik, M. P., Prajapati, H. B., & Dabhi, V. K. (2015). A survey on semantic document clustering. *2015 IEEE international conference on electrical, computer and communication technologies (ICECCT)*, 1-10. https://doi.org/10.1109/ICECCT.2015.7226036

[29] Kowsari, K., Jafari Meimandi, K., Heidarysafa, M., Mendu, S., Barnes, L., & Brown, D. (2019). Text classification algorithms: A survey. *Information, 10*(4), 150. https://doi.org/10.3390/info10040150

[30] Shah, N. & Mahajan, S. (2012). Document clustering: a detailed review. *International Journal of Applied Information Systems, 4*(5), 30-38. https://doi.org/10.1016/0306-4573(88)90021-0

[31] Salton, G, & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information processing & management, 24*(5), 513-523. https://doi.org/10.1016/0306-4573(88)90021-0

[32] Jalal, A. A. & Ali, B. H. (2021). Text documents clustering using data mining techniques. *International Journal of Electrical & Computer Engineering, 11*(1), 664-670. https://doi.org/10.11591/IJECE.V11I1.PP664-670

[33] Bafna, P., Pramod, D., & Vaidya, A. (2016). Document clustering: TF-IDF approach. *2016 international Conference on electrical, electronics, and optimization techniques (ICEEOT)*, 61-66. https://doi.org/10.1109/ICEEOT.2016.7754750

[34] Kodinariya, T. M. & Makwana, P. R. (2013). Review on determining number of Cluster in K-Means Clustering. *International Journal, 1*(6), 90-95.

[35] Ren, Z., Lin, T., Feng, K., Zhu, Y., Liu, Z., & Yan, K. (2023). A systematic review on imbalanced learning methods in intelligent fault diagnosis. *IEEE Transactions on Instrumentation and Measurement, 72,* 1-35. https://doi.org/10.1109/TIM.2023.3246470

[36]Werner de Vargas, V., Schneider Aranda, J. A., dos Santos Costa, R., da Silva Pereira, P. R., & Victória Barbosa, J. L. (2023). Imbalanced data preprocessing techniques for machine learning: a systematic mapping study. *Knowledge and Information Systems, 65*(1), 31-57. https://doi.org/10.1007/s10115-022-01772-8

[37]Kou, G., Peng, Y., & Wang, G. (2014). Evaluation of clustering algorithms for financial risk analysis using MCDM methods. *Information sciences, 275,* 1-12. https://doi.org/10.1016/j.ins.2014.02.137

[38]Shi, Y. (2021). A Novel Oversampling Method for Imbalanced Datasets Based on Density Peaks Clustering. *Tehnički vjesnik, 28*(6), 1813-1819. https://doi.org/10.17559/TV-20210608123522

[39]Liu, Q., Yun, F., Dong, M., Djoric, D., & Zivlak, N. (2024). Health Prognosis for Equipment Based on ACO-K-Means and MCS-SVM under Small Sample Noise Unbalanced Data. *Tehnički vjesnik, 31*(1), 24-31. https://doi.org/10.17559/TV-20230505000608

Contact information:

Woon-Kyo LEE, Student
Seoul National University of Science & Technology Graduate school of Public Policy and Information Technology,
232 Gongneung-ro, Nowon-gu, Seoul, Korea
E-mail: johntato@seoultech.ac.kr

Ja-Hee KIM, Professor
(Corresponding author)
Seoul National University of Science & Technology Graduate school of Public Policy and Information Technology,
232 Gongneung-ro, Nowon-gu, Seoul, Korea
E-mail: jahee@seoultech.ac.kr