International Journal of Advanced Technology in Engineering and Science
Vol. No. 11, Issue No. 12, December 2023
www.ijates.com
ijates
ISSN 2348 - 7550

EFFICIENCY AND EFFECTIVENESS OF K-MEANS VERSUS FUZZY HIERARCHICAL METHODS FOR DOCUMENT CLUSTERING

Shashi Kant Jha[1], Dr. Deepika Sharma[2]

[1,2]Department of Computer Science, Mansarovar Global University, Sehore, M.P., India.

ABSTRACT

The 20 Newsgroup (NG20) dataset is a famous compilation of Usenet debates covering 20 different subjects, and this research compares clustering approaches that have been applied to this dataset. In this study, we compare an existing fuzzy hierarchical clustering approach with a suggested K-means clustering algorithm, focusing on how well the two perform in terms of accuracy, execution time, and cluster similarity. Execution time, accuracy, and average cluster similarity were the three main criteria used to assess the efficiency and usefulness of the K-means approach in clustering. We compared the K-means algorithm's execution time in milliseconds to the fuzzy hierarchical algorithm's execution time. The findings demonstrate that the K-means technique is computationally faster than the fuzzy hierarchical method. This is supported by the fact that the execution time of the K-means method is substantially reduced across different newsgroups.

Keywords: Document clustering, Accuracy, Time, Fuzzy, Similarity

I. INTRODUCTION

The explosion of digital material in today's information-rich society has made maintaining and extracting insights from massive databases very challenging. An essential tool in tackling these difficulties has been document clustering, a fundamental approach in data mining, machine learning, and information retrieval. This method does not need previous labeling or predetermined categories; instead, it groups documents into clusters or categories according to their content similarity. Finding underlying structures in unlabeled data is the main objective of document clustering, which aims to improve information organization, retrieval, and analysis.

The idea of similarity measurement is fundamental to document clustering. Clustering algorithms may group texts with similar themes, subjects, or content attributes by comparing their similarity or dissimilarity. Both the efficiency of information retrieval systems and the ease with which huge text corpora may be managed are improved by this method. An example of how document clustering may improve user experience and satisfaction is when users query a search engine. The results that are

shown are relevant and structured in a manner that matches the underlying content structure. The use of document vectors in a three-dimensional space is a basic strategy for document clustering. Common methods for converting text data into numerical vectors that represent the importance of words in the documents and the total corpus include Term Frequency-Inverse Document Frequency (TF-IDF) and Latent Semantic Analysis (LSA). Many clustering algorithms, such DBSCAN, Hierarchical Clustering, and K-Means, rely on the idea of grouping comparable vectors according to density or distance metrics; these vector representations make this possible.

Among the most popular techniques, K-Means clustering sorts the dataset into K separate clusters, with each document being assigned to the cluster that has the closest mean vector. With this technique, clustering is made easy, and it works wonders with big datasets. Nevertheless, K-Means may be limited in situations where the ideal number of clusters is uncertain, since it needs the cluster number to be stated in advance. This may be resolved by using techniques like the Elbow Method and the Silhouette Score to ascertain, from the dataset's properties, the optimal number of clusters. An further well-known method, Hierarchical Clustering, builds a structure of tiered clusters similar to a tree, with each tier representing a different degree of detail. You may explore data at different degrees of detail with this strategy, and you don't even have to predetermine the number of clusters. The versatility of Hierarchical Clustering lies in the fact that it may be either agglomerative (bottom-up) or divisive (top-down), depending on the situation. Divisive Hierarchical Clustering begins with all documents in a single cluster and recursively separates them, while Agglomerative Hierarchical Clustering starts with each document as its own cluster and iteratively merges the closest clusters.

Another significant approach that employs data point density to detect clusters is DBSCAN, which stands for Density-Based Spatial Clustering of Applications with Noise. It is not necessary to specify the number of clusters in advance when using DBSCAN as opposed to K-Means and Hierarchical Clustering. As an alternative, it describes clusters as dense point areas with sparser regions between them. When applied to datasets with noisy or unevenly formed clusters, this strategy excels where more conventional approaches fail. Important to document clustering is the assessment of clustering findings, which aids in determining the relevance and quality of the produced clusters. Internal assessment criteria like cohesion and separation and external measurements like Purity and Normalized Mutual Information (NMI) are among the metrics and approaches used to assess clustering success. Cohesion is a measure of how similar documents are within a cluster, whereas separation is a measure of how diverse clusters are from one another. You may learn a lot about how well the clustering process is working and where to focus your efforts for future improvements by looking at these indicators.

As a result of its adaptability and significance, document clustering has found uses in many different fields. When it comes to retrieving information, document clustering is a game-changer. It streamlines

search results, makes them more relevant, and improves the user experience. Clustering is a useful tool in content management for organizing massive document collections into topic groupings, which in turn allows for more effective administration and retrieval of material. Text mining, sentiment analysis, and social network analysis all rely heavily on document clustering to help them find trends and patterns in textual data. The benefits of document clustering aren't without their drawbacks, however. Clustering outcomes are very sensitive to the clustering method used, the parameters fine-tuned, and the documents represented. In addition, human judgment and domain knowledge may be necessary for the subjective interpretation of clusters and the significance of produced groups. Continuous research and innovation in clustering approaches, along with the development of increasingly complex algorithms and assessment methodologies, are necessary to address these issues.

K-Means for Document Clustering

In this age of information explosion, it is very necessary to efficiently handle and organize massive volumes of textual data for several applications, such as information retrieval, data mining, content recommendation, and knowledge discovery. An unsupervised learning approach known as document clustering is crucial in these fields because it allows for the easy access and analysis of information by grouping documents that are similar together. For document clustering, K-means clustering stands out among the many available clustering approaches thanks to its efficacy, simplicity, and efficiency. Assigning documents to the cluster with the closest centroid is the goal of K-means clustering, a partition-based technique. The goal is to split the dataset into a preset number of clusters. The approach seeks to reduce the within-cluster variance, sometimes called the sum of squared errors, using iterative operations. For all its apparent lack of complexity, K-means has shown to be a powerful tool in a wide range of applications, thanks to its user-friendly design and easy implementation. The ultimate clustering results may be affected by the number of clusters (K) and how sensitive it is to the original cluster centroids, which in turn affects its efficacy.

The K-means method starts by picking K initial centroids at random; these locations stand for the cluster centers. Based on a distance measure, usually Euclidean distance, each document in the collection is then allocated to the closest centroid. Assigning all documents causes the centroids to be recalculated as the mean of all clusters. Until convergence is achieved, whereby the cluster assignments remain constant or a maximum number of iterations is reached, the process of updating centroid and performing assignments continues repeatedly. Documents inside each cluster are more similar to each other than to those in other clusters; as a consequence, the dataset is partitioned into K clusters, with each cluster represented by its centroid. When it comes to large-scale datasets, one of K-means clustering's main strengths is how efficiently it handles computing. In most cases, the time complexity of the method is directly proportional to the number of documents and clusters, making it

capable of efficiently processing massive datasets. Furthermore, K-means is accessible to academics and practitioners in a wide range of domains because to its ease of implementation and comprehension.

Although K-means clustering offers many benefits, it does have certain restrictions that could affect how well it works and what kinds of problems it can solve. Predicting the exact number of clusters (K) in advance is a major hurdle that isn't always easy to accomplish in reality. Imperfect clustering outcomes could emerge from choosing the wrong value for K, which often requires subject expertise or empirical testing. Finally, the final clusters might be impacted by the sensitivity of K-means to the original location of centroids. It has been suggested to enhance the algorithm's resilience and convergence by running it numerous times with varied initializations and using approaches like K-means++. This should reduce the problem. One further thing that may go wrong with K-means is that it assumes clusters are equal in size and spherical, which isn't always how the data is really structured. When working with clusters that aren't perfectly round or have an unusual form, this assumption could provide inaccurate clustering results. One solution to this problem is the proliferation of K-means variants; for example, density-based approaches that don't assume spherical clusters and K-medoids, which swaps out centroids for real data points.

Topic modeling, text categorization, and information retrieval are just a few of the many document clustering-related applications of K-means. To make vast collections of text simpler to read and analyze, K-means groups documents with similar content to help organize and retrieve pertinent information. For instance, K-means can automatically sort papers and news items into appropriate categories by clustering them according to their themes, which is useful in academic literature and the news industry.

New developments in machine learning and natural language processing (NLP) have made K-means clustering even more useful for document analysis. More accurate and relevant clustering is now possible because to techniques like topic modeling and word embeddings, which have enhanced the representation of textual data. By capturing the semantic associations between words via word embeddings, K-means is able to group texts according to their true meaning instead of superficial characteristics. Hybrid systems that combine K-means with other machine learning algorithms and methodologies have also emerged, allowing users to take use of the best features of each method. Clustering performance may be enhanced by decreasing noise and concentrating on the most significant characteristics when dimensionality reduction methods like Principal Component Analysis (PCA) are used with K-means.

Fuzzy Hierarchical for Document Clustering

Documents may be assigned to several clusters with different levels of membership using fuzzy

clustering, as opposed to crisp clustering. When papers display traits of more than one category, rather than fitting cleanly into one, this method really shines. To account for the inherent complexity of textual data, Lotfi A. Zadeh proposed the notion of fuzziness in 1965 as a means to manage ambiguity and incomplete membership. Based on this idea, fuzzy hierarchical clustering incorporates it into a framework for hierarchical clustering, which uses a tree-like structure of layered clusters to arrange information.

There are a number of benefits to using fuzzy logic with hierarchical clustering. To begin with, it makes the clustering process more resilient by making allowances for the fact that document data is inherently imperfect and imprecise. When dealing with documents that have unclear or overlapping information, traditional hierarchical approaches could fail to produce correct or meaningful clusters. In contrast, fuzzy hierarchical algorithms provide more accurate and understandable clustering results by better representing the degree of similarity between texts and groupings.

Second, fuzzy hierarchical clustering is a great way to make clustering more efficient and scalable. Traditional approaches may need a substantial investment of time and computing resources in order to get acceptable results when applied to massive document collections. It is possible to decrease computational complexity and improve clustering efficiency by using fuzzy hierarchical algorithms, which can manage overlapping clusters and partial memberships. The capacity to swiftly and correctly handle and evaluate massive amounts of text data is of the utmost importance in real-world applications.

When it comes to assessing clustering outcomes and establishing cluster structures, fuzzy hierarchical approaches provide more leeway. We may get insights into both broad and particular trends in the data by exploring document linkages at multiple degrees of granularity, made possible by the hierarchical clustering method. When comprehending the interrelationships across several tiers of document classifications is crucial, as in topic modeling, information retrieval, and document summarization, this hierarchical structure may prove to be quite helpful. Different difficulties in document clustering have prompted the development of various fuzzy hierarchical clustering techniques. A hierarchical clustering method that combines fuzzy clustering principles with hierarchical clustering approaches is the Fuzzy C-Means (FCM) algorithm. Based on their resemblance to the centroids of the clusters, FCM allocates documents to several clusters with different levels of membership. A hierarchical representation of document connections is created by merging or breaking clusters depending on the fuzzy membership values.

The FAHC algorithm is another noteworthy method; it uses fuzzy logic to augment the agglomerative hierarchical clustering procedure. When FAHC merges clusters, it makes advantage of fuzzy membership values, which provide a more accurate and versatile way to describe document similarities. Deriving useful clusters and sub-clusters is possible from the resultant hierarchical

structure, which gives a thorough picture of document connections.

II. REVIEW OF LITERATURE

Lal, Chaman et al., (2021) It is simple for the user to group together papers with similar content using document clustering. Research in this fascinating field has yielded a plethora of new methods throughout the years. But studies focusing on English and other languages with a lot of resources tend to be the most common. With regard to Pakistani national anthems, this research gives an experimental estimate of clustering methods. Because Anthem is so short, it's hard to group its themes together. In this study, TF-IDF features, noise reduction, stemming, corpus tokenization, and stop-words were extracted before the song was clustered using the K-Means approach. The results show that a clustering method integrated with TF-IDF features may be used with a K-mean clustering approach.

Arivarasan, Aranga & Karthikeyan, Dr. (2019) The exponential growth of the internet is directly linked to the skyrocketing demand for printed materials. Gigabytes of processed text documents are the end product. Algorithms that are up to snuff boost performance by accurately indexing and retrieving text content. Additionally, data mining algorithms provide new ideas to the field. The result is a surge in academic interest in developing key models for text data mining. The proposed model uses a two-stage procedure that makes use of the K-Means method to group text texts. Prior to beginning the clustering method, there is a step called pre_processing. The tokenization approach is used for pre-processing in the procedure. The document feature vector is constructed autonomously by recognizing the various words, determining how often they appear, and assigning the TFIDF weights to each instance. In clustering, the feature vector is divided into smaller groups using various similarity metrics and the K-means method.

Lydia, Laxmi et al., (2018) Researchers in the area of data mining use a technique called text mining to correctly group the vast volumes of semi-structured data. The three primary features of maximum text documents are exploration, document structure, and rapid information retrieval.Document classification and text input data declaration is a laborious process. In order to make it easier to identify certain papers, this article primarily aims to provide a focused open source solution for organizing important document groupings into linked folders. Open research obligations provide problems in the form of algorithms that are now under consideration. We examine K-means partitioning for document clustering, centroid computation, and cluster similarity in this paper.

Wahyu, R.B & Vito, Arnold. (2018) The Internet has made it possible for everybody in our current digital era to easily access a plethora of information that was formerly only available in written form. Given the abundance of unstructured documents holding various sorts of information, a software that can automatically sort and classify digital documents is very necessary. In order to attain an accuracy

level of up to 85% according to the user's expectations, this desktop software employs the K-Means Algorithm to group documents based on their content similarity.

Kaur, Ramanpreet & Kaur, Amandeep. (2016) In this article, we found the outcomes of research that used many general methods for clustering and classifying documents. This work aims to improve clustering. The primary objective is to develop a system that expedites the retrieval of text documents from clusters. In this paper, we provide a new method for grouping and classification using MATLAB-based k-means with feedforward neural networks. We utilize k-mean to group text documents into clusters, and neural networks for classification. End result: Genetic algorithms, Guassian distributions, hybrid genetic algorithms, k-means clustering, fast k-means global, and semi-supervised models for labeled text like Partially Labeled Dirichlet Allocation and the Partially Labeled Dirichlet Process are just a few of the older methods that have been developed. Each of these approaches has its advantages and disadvantages, but one constant is the time it takes. This is why the study's main objective is to construct a model that is able to achieve document similarity using both supervised and unsupervised techniques. To resolve this tedious problem, we used neural networks for classification and k-means for grouping. We constructed a model using supervised and unsupervised approaches to attain document similarity.

Kim, Woosaeng & Kim, Sooyoung. (2014) Data availability is growing at an exponential pace due to improvements in computer power and the expansion of the internet. Since the document forms regulate these enormous data sets, it is vital to locate and evaluate them efficiently. The document clustering method allows for the automatic classification, searching, and processing of large datasets by grouping related texts based on their level of similarity. This article proposes a method for identifying the first seed points using principal component analysis to enhance the clustering performance of the K-means algorithm. In feature vector space, the documents are shown as vectors, and the procedure is used to cluster them. The experimental findings show that our method is superior to the standard K-means algorithm.

Li, Youguo & Wu, Haiyan. (2012) In order to develop a more effective K-Means clustering method, this research combines the traditional K-Means technique with the largest minimum distance algorithm. This revised strategy may make up for the shortcomings of the original K-Means algorithm when it comes to selecting the center of interest. The new K-Means technique fixes the two issues with the previous one: first, it doesn't rely too much on the choice of starting focus points, and second, it doesn't get trapped in the local minimum as easily.

III. PROPOSED METHODOLOGY

Dataset

An English newsgroup corpus will serve as the dataset for this study's investigation. This English

dataset is referred to as NG20, or 20 Newsgroup. There is a common dataset that uses this 20 Newsgroup, and it is widely utilized. Crawled from 20 separate newsgroup boards, the initial collection of the 20 Newsgroup dataset includes 19997 Usenet talks. All of the newsgroups have almost the same number of documents. The subjects covered range from politics and religion to computer science and sports.

Parameters

Ultimately, a clustering algorithm aims to generate high-quality clusters. Several factors may be used to assess the quality of the clusters that are created. The overall efficacy of the approach is shown by these criteria.

The overall runtime, frequent item dataset, and cluster formation count are the chosen metrics for evaluating the suggested strategy in this study.

A fuzzy clustering approach that uses a hierarchical aggregation algorithm is used to compare the suggested method to. Clustering medical records in a laboratory was the initial motivation for this approach. The principles of fuzzy set theory form the basis of this document clustering technique. In order to sort documents into specific clusters, it determines their membership function values prior to clustering.

After the documents are sent to the system, the total execution time displays how long it took for the complete process to calculate the values and generate the clusters. A shorter execution time indicates that the approach forms clusters more quickly and accurately.

An indication of how well the proposed and current systems can generate clusters with meaningful similarities between them is the system's accuracy. When comparing the system's overall performance, accuracy is a fantastic metric to employ.

IV. RESULTS AND DISCUSSION

In this part, you will find the values and graphical depiction of the evaluation parameters. The execution duration, accuracy, and average cluster similarity were the three measures used. When determining the system's overall performance, the three parameters used here are optimal. The system outperforms the current fuzzy hierarchical method in terms of these three parameters, and the accompanying graphics are also provided in this section.

Pictured in Fig. 1 are the millisecond-level comparisons between the current fuzzy clustering system and the suggested K-means clustering method. The two systems' execution times when fed the various newsgroups in the NG20 dataset are shown in Table II.

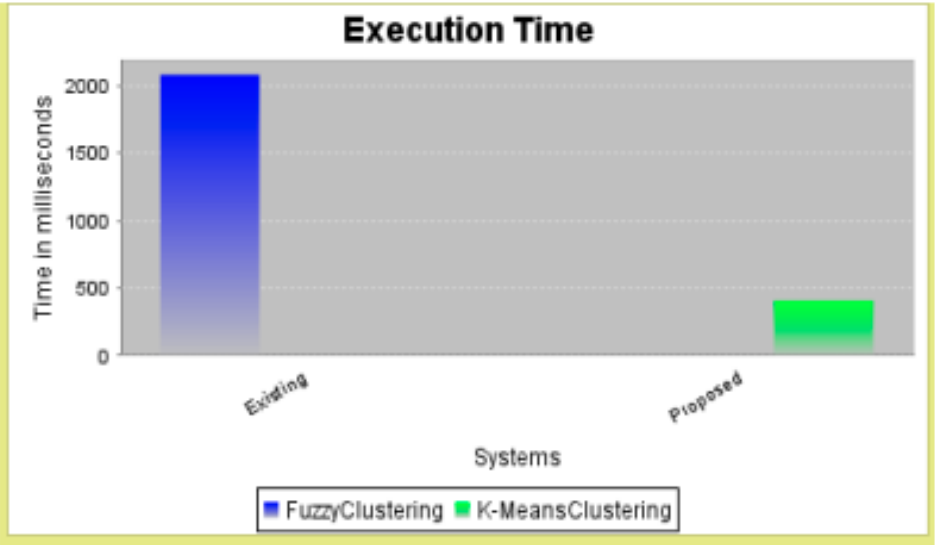

Figure 1 Execution time of NG20 dataset

Table II and Figure 1 show that the suggested technique outperforms the current algorithm in terms of execution. The execution time demonstrates how much faster and simpler the suggested approach is compared to the current ones in terms of processing capabilities.

Table 2: Execution Time for NG20 datasets

Sl. No	Topic	Execution time in milliseconds	
		Existing System	Proposed System
1	alt.atheism	3451	735
2	comp.graphics	3000	609
3	comp.os.mswindows.misc	1000	219
4	comp.sys.ibm.pc.hardware	986	202
5	comp.sys.mac.hardware	829	172
6	comp.windows.x	232	62
7	misc.forsale	767	157
8	rec.autos	951	203
9	rec.motorcycles	2515	516
10	rec.sport.baseball	1079	343
11	rec.sport.hockey	670	250
12	sci.crypt	813	516
13	sci.electronics	1078	234
14	sci.med	1204	250
15	sci.space	7500	1421
16	soc.religion.christian	1315	298
17	talk.politics.guns	2820	592
18	talk.politics.mideast	2640	546
19	talk.politics.misc	1078	219
20	talk.religion.misc	1328	279

In Fig. 2, we can see the system's visual depiction according to cluster similarity. The average similarity value of the documents in the final cluster that is formed is used to determine cluster similarity. Compared to the current fuzzy clustering approach, the documents inside each cluster are more closely related, as seen by the high similarity measure. The system's data processing accuracy in comparison to the fuzzy clustering approach is shown in Figure 3.

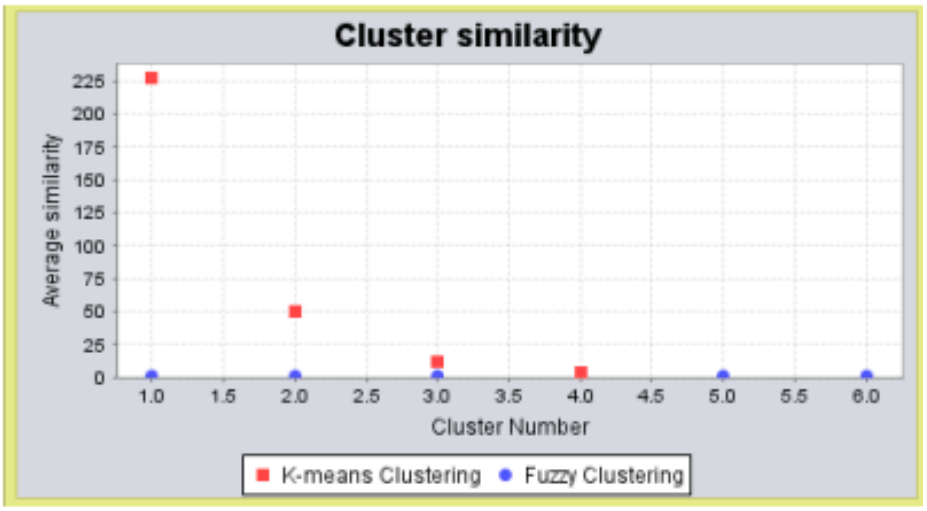

Figure 2: Results of Cluster similarity

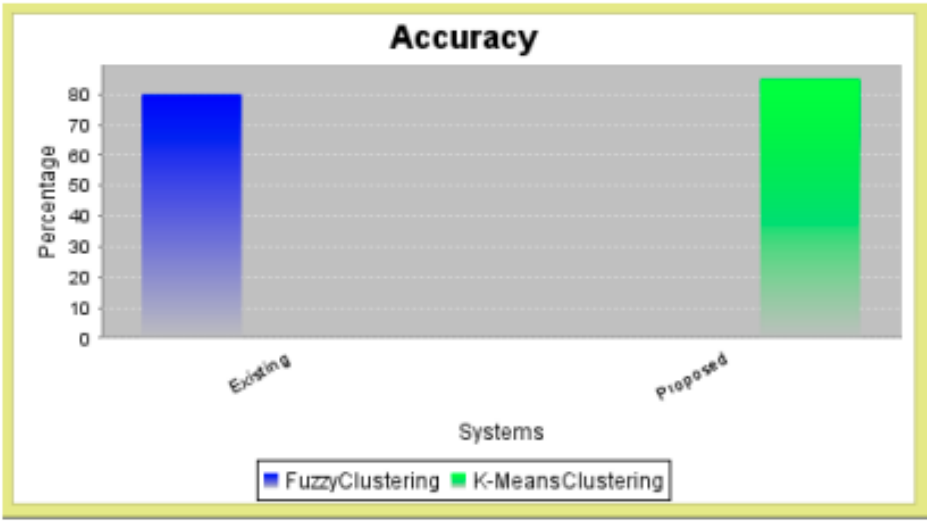

Figure 3: Results of Accuracy

V. CONCLUSION

This research compares and contrasts the efficiency and efficacy of two clustering approaches, K-means and fuzzy hierarchical, using the 20 Newsgroup (NG20) dataset. According to the findings, K-means clustering outperforms the fuzzy hierarchical technique in terms of computing speed,

drastically cutting execution time. Fuzzy hierarchical clustering offers a more comprehensive and understandable clustering structure, yet both approaches have strengths in terms of accuracy and cluster similarity. This is especially helpful when it's critical to comprehend the hierarchical connections and overlaps across clusters. The fuzzy hierarchical approach has a larger computational cost, but it is more resilient and improves cluster interpretability when dealing with papers that have overlapping properties.

REFERENCES: -

[1] C. Lal, A. Ahmed, R. Siyal, S. Kumar, S. Aftab, and A. Jamali, "Text Clustering using K-MEAN," International Journal of Advanced Trends in Computer Science and Engineering, vol. 10, no. 4, pp. 2892-2897, 2021, doi: 10.30534/ijatcse/2021/371042021.

[2] E. Oti, M. Olusola, F. Eze, and S. Enogwe, "Comprehensive Review of K-Means Clustering Algorithms," International Journal of Advances in Scientific Research and Engineering, vol. 07, no. 08, pp. 64-69, 2021, doi: 10.31695/IJASRE.2021.34050.

[3] F. M. Shamrat, Z. Tasnim, I. Mahmud, N. Jahan, and N. Nobel, "Application Of K-Means Clustering Algorithm To Determine The Density Of Demand Of Different Kinds Of Jobs," International Journal of Advanced Trends in Computer Science and Engineering, vol. 9, no. 02, pp. 2550-2557, 2020.

[4] N. Bakala, "K-Means Algorithm for Clustering Afaan Oromo Text Documents using Python Tools," International Journal of Recent Technology and Engineering (IJRTE), vol. 9, no. 1, pp. 1279-1282, 2020, doi: 10.35940/ijrte.A2284.059120.

[5] K. P. S and M.-S. Yang, "Unsupervised K-Means Clustering Algorithm," IEEE Access, vol. 12, no. 99, pp. 1-1, 2020, doi: 10.1109/ACCESS.2020.2988796.

[6] A. Arivarasan and K. Dr., "Data Mining K-Means Document Clustering using TFIDF and Word Frequency Count," International Journal of Recent Technology and Engineering (IJRTE), vol. 8, no. 2, pp. 2542-2548, 2019, doi: 10.35940/ijrte.B1718.078219.

[7] L. Lydia, P. Govindasamy, S. K. Lakshmanaprabu, and D. Ramya, "Document Clustering Based On Text Mining K-Means Algorithm Using Euclidean Distance Similarity," Journal of Advanced Research in Dynamical and Control Systems, vol. 10, no. 2, pp. 2-19, 2018.

[8] R. B. Wahyu and A. Vito, "Documents Clustering Using K-Means Algorithm," IT for Society, vol. 3, no. 02, pp. 2-19, 2018, doi: 10.33021/itfs.v3i02.589.

[9] T. Sardar and Z. Ansari, "An Analysis of MapReduce Efficiency in Document Clustering using Parallel K-Means Algorithm," Future Computing and Informatics Journal, vol. 3, no. 2, pp. 30.40, 2018, doi: 10.1016/j.fcij.2018.03.003.

[10] N. A. Mohd ariff, M. A. Abu Bakar, and M. I. Rahmad, "Comparative Study of Document Clustering Algorithms," International Journal of Engineering and Technology (UAE), vol. 7, no. 4, pp. 246-251, 2018, doi: 10.14419/ijet.v7i4.11.20816.

[11] R. Kaur and A. Kaur, "Text Document Clustering and Classification using K-Means Algorithm and Neural Networks," Indian Journal of Science and Technology, vol. 9, no. 40, pp. 2-18. 2016, doi: 10.17485/ijst/2016/v9i40/97722.

[12] W. Kim and S. Kim, "Document Clustering Technique by K-means Algorithm and PCA," Journal of the Korea Institute of Information and Communication Engineering, vol. 18, no. 3, pp. 625-630, 2014, doi: 10.6109/jkiice.2014.18.3.625.

[13] Y. Li and H. Wu, "A Clustering Method Based on K-Means Algorithm," Physics Procedia, vol. 25, no. 2, pp. 1104-1109, 2012, doi: 10.1016/j.phpro.2012.03.206.

[14] A. Šilić, M.-F. Moens, L. Žmak, and B. Dalbelo Bašić, "Comparing Document Classification Schemes Using K-Means Clustering," in Proceedings of the 5177th Lecture Notes in Computer Science, vol. 1, no. 2, pp. 615-624, 2008, doi: 10.1007/978-3-540-85563-7_78.

[15] Y.-K. Shin, "An Improved K-means Document Clustering using Concept Vectors," Journal of the Korean Data and Information Science Society, vol. 14, no. 4, pp. 1-30 2003.