Vol 3, Issue 10, 2024
Doi:- 10.38124/ijsrmt.v3i10.61
International Journal of Scientific Research and Modern Technology (IJSRMT) ijsrmt.com

Leveraging TF-IDF Matrix for Document Clustering with K-Means Algorithm

Shilpi Kulshrestha[1]
Department of CSE,
Jaipur National University, Jaipur, India

Dharmesh Santani[2]
Department of CSE,
Jaipur National University, Jaipur, India

Abstract:- Document clustering is an important task for information retrieval, it aims for grouping of similar kind of documents together for efficient organization and retrieval. This paper presents a new approach for document clustering by combination of the Term Frequency-Inverse Document Frequency (TF-IDF) matrix with the K-Means algorithm. The Proposed system overcomes the obstacles of the traditional methods integrating TF-IDF matrices to convey document semantics and K-Means clustering to get homogeneous document clusters. Key components of the system include text pre-processing techniques such as stop-word removal, stemming, and tokenization, which improve the quality of TF-IDF representations. Additionally, evaluation metrics like purity, F-measure, and silhouette score are applied to evaluate the system's clustering performance. Our proposed approach shows that it is feasible to process large volumes of documents and at the same time ensuring robustness by discarding outliers and noisy data in the data. The obtained results upon a benchmark dataset demonstrate the superiority of suggested approach in comparison to the baseline techniques and these results underline the effectiveness of the proposed method in terms of the efficiency of the document clustering and facilitating the streamlined document organization and retrieval in different domains.

Keywords:- *Document Clustering, TF- IDF Matrix, K-Means Algorithm, Evaluation Metrics, Text Pre-Processing, Robustness, Scalability.*

I. INTRODUCTION

The volume of text-based information especially that which lacks structure, due to the age of digital has tremendously increased. This era is characterized by various sources of textual data, which are coming from the worlds of social media, shopping, scientific research, industries, transactions among corporations, and others. On the one hand, such kind of datasets bring an additional opportunity in terms of extracting and breaking down the textual information, while on the other hand, they display a particular challenge for the mentioned tasks. Most part of the current non-targeted unstructured textual resources may go beyond the resolution limit of typical search techniques such as keyword-based searches and manual categorization. Unlike a human who carries out various processes of indexing, classification, and providing content from a huge body of texts, it is a computer which could automatically do such tasks [1].

An easy way to do computers understand natural language processing is to use magic square (TF-IDF matrix). It's computed through by the term frequency – which is the number of times a word appears in a document the phrase inverse document frequency (the percentage of all the documents that include the phrase scaled by logarithm). This term is an abbreviation utilized to capture the association between Features Term Frequency and Inverse Document Frequency (TF-IDF) [2]. By means of generated matrix, it is easy to compare documents and assign them to clusters after their relative values of phrases are described in every text. TF - IDF matrix in document clustering services has some impressive benefits which are demonstrated by its capacity to handle huge data sets, ability to identify the importance of individual terms in a document, and adapting to various text data formats. Nonetheless, TF-IDF matrix has its limitations in them not covering semantic relationships of the words as well as the possibility of not working on short or noisy texts.

One of the favourite clustering techniques that classify K sets for the alike data is the K-Means method. In order to operate, k centroids with point characteristics that represent each cluster are chosen at random. Afterwards, each data point is repeatedly distributed to the nearest centroid, and the centroid is calculated again using only the newly allocated data points. Until the centroids are stationary enough to hardly be noticed, the process will continue this time. The K-Means approach enables document clustering in a straightforward manner having many advantages for instance the simplicity, speed, and efficiency in handling big data. The K-Means technique does have some major deficiencies, which are the requirement to define the number of clusters in advance, the possibility to arrive at the local optimum, as well as the potential to perform badly with non-linearly separable data [3].

First, we turn the documents into TF-IDF vectors then we use the TF-IDF matrix in order to introduce documents clustering according to the K-Means approach. Then, the documents are assigned either of the K clusters according to the similarity of their vectors by the k-mean method. Then, we will evaluate the effectiveness of the clustering by means of metrics, which may be denoted as entropy, purity, and silhouette score [4]. Utilization of this method opens the way for a number of practical applications, including grouping news stories having the same theme into recommendation systems and data comparison that is legal to deal with plagiarism issues. It can hence be used for marketing, to determine the ways consumer's classified act for purchases.

A. *Problem Statement*

The sudden increase of digital documents in the big data age presents major hurdles to effectively classify and retrieve pertinent information. Conventional document clustering approaches that just use basic statistical techniques or keyword matching frequently fall short of capturing the underlying semantic structure of document collections. This restriction makes it more difficult for users to quickly locate and obtain pertinent documents, which reduces the efficacy of information retrieval systems. Thus, there is an urgent need for more sophisticated methods that can better capture the semantic links between documents and increase the accuracy of grouping.

B. *Objective*

➢ *The main Goals of this Work are Formulated as follows:*

- To conduct a comparative analysis to determine the incorporation of the TF-IDF matrix and the K-Means algorithm's extent of document clustering efficiency.
- To run the obtained results on the basis of alignment to the results of the benchmark experiments to the turned typical evaluation metrics.
- Considering the influence of the number of clusters and some feature selection tricks to learn how clustered the meta-representations are.
- To explore the practical implications of the research findings in real-world applications, such as information retrieval systems and document management platforms.

C. *Challenges*

➢ *Challenges Faced during this Research were as follow:*

- Data Quality: Ensuring the quality and consistency of the textual data, including handling noise, outliers, and data preprocessing errors.
- Parameter Tuning: Selecting appropriate parameters for the TF-IDF matrix construction and K-Means clustering, such as the number of clusters and distance metric, to optimize clustering performance.
- Interpretability: Interpreting and validating the clustering results, especially in the absence of ground truth labels, to ensure the meaningfulness of the clusters generated.
- Scalability: Dealing with the scalability of the clustering algorithm, particularly when clustering large datasets with millions of documents, to maintain computational efficiency.
- Evaluation Metrics: Choosing suitable evaluation metrics to assess the quality of clustering results and comparing them with baseline methods to ensure the effectiveness of the proposed approach.

II. RELATED WORKS

Techniques like text mining and document clustering are crucial for drawing knowledge out of massive text data sets. Based on shared features, document clustering puts related documents in one group. Numerous clustering algorithms, such as semantic-based approaches, Expectation Maximization techniques, and K-means variants, have been studied by Luo, Congnan et al [5]. The bag-of-words model, which is frequently used in traditional methods, may overlook synonymy between related documents. Bafna, Prafulla et al [6] have looked into other approaches, like clustering based on keywords, phrases, and concepts, to get around this restriction. V. Prabhas et al [7] proposed a clustering method using k mean for customer segmentation.

Balabantaray et al [8] developed a method for document clustering and used fuzzy logic, which permits varying degrees of truth in cluster membership. S. Chalechema et al [9] proposed a method a method for customer clustering for using k mean and RFM model. Kumbhar, Rutuja, et al [10] proposed a method which clusters documents according to normalized word frequencies by using the fuzzy c-means algorithm, text cleaning, stemming, and feature selection. Cluster purity, accuracy, and speed are three metrics used to assess the quality of clustering. S. Kulshrestha et al [11] proposed a method for customer classification and used k mean clustering algorithm.

Al-Obaydy, et al [12] proposed a method for document classification using term frequency-inverse document frequency and K-means clustering. J. Sarmah et al [13] proposed a method for performance analysis of deep CNN, YOLO, and LeNet for handwritten digit classification. M. Sohail et al [14] designed a system for hand written digit classification. M. Lal et al in their paper [15] and [16] proposed methods for hand written script recognition using deep learning methods [17]. These paper reviews text document clustering algorithms used in text mining to discover knowledge from textual data.

A. *Overview of Text Mining and Document Clustering*

Text mining is used for processing, evaluating, and drawing useful conclusions from a textual data by using different types of methods and approaches. The core task of Text mining is document clustering and this is important for no. of applications, including recommendation systems, text categorization, and information retrieval. Document clustering helps us with collecting information and decision-making processes by organizing and comprehending massive number of textual documents by clustering similar type of documents together.

B. TF-IDF (Term Frequency-Inverse Document Frequency) Matrix

TF-IDF (Term Frequency-Inverse Document Frequency) matrix is a useful component because it gives numerical representation to the documents, capturing their semantic content and term importance. The term frequency (TF) component helps us to counts how often a word appears in a document, whereas the inverse document frequency (IDF) component helps us by penalizing phrases that are common in the whole document. These two elements work together to create the TF-IDF matrix, which helps in document clustering by highlighting the phrases that are both unique and frequent in the whole document, indicating their significance in describing the content of specific documents.

C. K-Means Algorithm and its Application in Document Clustering

K-Means algorithm is a well-known clustering algorithm, which divides data points into K number of clusters and insert these data points into them according to similarity they are to one another according to their features. In this research, the K-Means algorithm utilizes the TF-IDF representations of documents for efficient grouping of them into the clusters. We optimize the cluster centroids to minimize the variance within-cluster and maximize the separation between-cluster this process is repeated in every iteration. Thus, K-Means help in efficient organization of documents of similar content into one cluster.

D. Previous Studies on Document Clustering Techniques

Document clustering algorithms have been extensively studied, covering both conventional and cutting-edge approaches. Conventional techniques comprise partitioning-based techniques such as K-Means and its variations, and hierarchical clustering, which builds a tree-like hierarchy of clusters. Some more sophisticated techniques have recently been studied, such as spectral clustering, which divides data using the eigenvectors of affinity matrices, and agglomerative clustering, which repeatedly combines data points based on their pairwise distances in the document text.

This research has focused on integrating feature selection techniques with domain-specific information to improve the interpretability and efficacy of document clustering. For increasing the robustness and quality of clustering model, ensemble clustering strategies can be integrated with various clustering algorithms. These types of algorithms have advanced features with high-dimensional data. Therefore, current research paper aims to create increasingly complex method which can measures, and investigates fresh uses of document clustering in fields including cybersecurity, social media analysis, and healthcare informatics.

III. METHODOLOGY

A. Data Collection and Preprocessing

The process starts with gathering the dataset of various documents to work with. These documents might come from academic journals, press releases, posts on social media, or any other type of domain-specific content. Following data collection, pre-processing techniques are used to clean and standardize the textual data. Managing special characters or numeric values, eliminating stop words, stemming or lemmatization, and tokenization are a few of these tasks. Following pre-processing, the data is prepared for further examination.

B. Construction of TF-IDF Matrix

To create the TF-IDF matrix, the pre-processed data is converted into a numerical representation that expresses the importance of phrases within each document in relation to the corpus as a whole. Every document has its term frequency (TF) calculated, which shows how frequently each term occurs in the text. In the next step, the inverse document frequency of each term is calculated, quantifying the rarity of the term across the entire corpus.

$$tf(t,d) = \frac{n_t}{\sum_k n_k}$$

$$idf(t,D) = \log \frac{|D|}{|\{\, d_i \in D \mid t \in d_i \,\}|}$$

In the last steps, the TF-IDF values for each term-document pair are calculated by multiplying the TF and IDF values. This process converts in a sparse matrix where rows correspond to documents, columns correspond to terms, and each entry represents the TF-IDF value of a term in a document.

$$tf\text{–}idf(t,d,D) = \mathrm{tf}(t,d) \cdot \mathrm{idf}(t,D)$$

C. Overview of K-Means Algorithm

K-Means algorithm first divides a dataset into k number of clusters, where k is given by the user. This technique uses a distance metric typically Euclidean distance measures. The working process is to repeatedly assign data points to the closest cluster centroid. The centroids of the clusters are updated in every step depending on the mean of the data points assigned to each cluster once all the data points have been assigned to clusters. This process continues until convergence, when the cluster assignments settle and centroids stop changing further. Large datasets can be clustered using K-Means algorithm because it is efficient, and scalable.

D. Integration of TF-IDF Matrix with K-Means Algorithm

K-Means algorithm uses the TF-IDF matrix as its input data where rows represents documents and columns represent terms together with the appropriate TF-IDF values. This algorithm process documents according to how similar their contents are. In the starting K centroids are created at random, and documents are sorted by their TF-IDF representations and allocated to the closest centroid in the first step. The next step calculates the mean of the TF-IDF vectors of the documents allocated to each cluster, and the centroids are updated repeatedly until there is change in the cluster centres. Clusters of documents with related content are produced as a result of this process, which continues until convergence.

E. Evaluation Metrics for Document Clustering

The quality of document clustering findings is rated using multiple assessment metrics. One frequently used metric is the Rand index, which evaluates how similar the actual and expected cluster assignments are to one another. The silhouette score and the Davies-Bouldin index are two metrics used to assess the cohesiveness and separation of clusters, respectively. Furthermore, external assessment criteria such as normalized mutual information (NMI) and purity may be used when ground truth cluster labels are assigned. These assessment metrics allow different clustering methods and parameter configurations to be compared and validated by providing quantifiable measurements of clustering performance.

IV. EXPERIMENTAL SETUP

A. Description of Dataset

An illustrative and differentiated collection of textual documents is needed for the experimental evaluation. For the suggested technique to be resilient across domains the documents should cover a range of themes, styles, and durations. In our study we used publicly accessible dataset. It needs pre-processing to ensure consistency in formatting across documents and to remove noise and irrelevant information.

B. Selection of Parameters

For maximizing the effectiveness of the k-means algorithm and the TF-IDF matrix, the following factors we adopted in our model.

- No. of clusters (K): In K-Means algorithm k is user defined and it provided at the time clustering that how many clusters we want.
- TF-IDF weighting scheme: This is used to assess the potential effects of various weighting schemes on the outcomes of clustering.
- Distance matrix: The results of clustering algorithm are greatly dependent on the distance matrix.
- Lemmatization vs. stemming: These techniques are used for the purpose of reducing words to their root forms to improved term matching.

C. Implementation Details

The TF-IDF matrix construction and method are integrated in the experimental implementation using the relevant libraries or programming frameworks such as, NLTK or scikit-learn. The pre-processed data is than used in creation of TF-IDF matrix and the documents are clustered together according to their representations from the K-Means method. Cluster centroids are updates iteratively. Convergence criteria and management of K-Means initialization techniques both take place in the implementation step.

➢ *Measures Included in Experimental Steps are:*

- Random seed: In this step random seed are selected to ensure reproducibility of results across multiple runs of the experiment.
- Cross-validation: The cross-validation techniques are used to check the robustness of the proposed model whether it is performing well or not.
- Performance monitoring: This performance evaluation parameter is used to monitor the execution time, memory usage, and convergence behaviour of the model.

V. RESULTS AND DISCUSSION

We use various evaluation parameters to measures the effectiveness of the document clustering algorithm. Clustering quality and computing efficiency performance measures are taken which gives the information of coherence and separation. The coherence gives closeness within the cluster and separation gives that how far the data points from another cluster. The clustering algorithm's computational efficiency is also calculated to evaluate scalability and performance, including both memory usage and execution time.

The results of proposed method are compared with the baseline approaches encompassing conventional document clustering methodologies like density based clustering, hierarchical based clustering and partitioned based clustering. For checking cluster quality feature representation methods are also compared like bag-of-words and word embedding.

The experimental results are analysed to learn more about how well the K-Means method and TF-IDF matrix works for document clustering. Patterns within the clusters, underlying semantic structure and trends are also analysed. Qualitative analysis also gets carried out to examine representative documents from each cluster and to evaluate how interpretable the clustering findings are.

The Effectiveness of TF-IDF and K-Means in the document clustering technique is the core issue for the discussion. The working efficiency of K-Means in dividing data into high-dimensional data along with the benefits and drawbacks of the TF-IDF format applicability are the subjects taken into account in this section. Along with the impact of parameter set on the clustering process discussing the number of clusters and distance metric are also taken into consideration. The experimental analysis observes the phenomenon on its run and make observations that can be used for the evaluation of postulating the report and proposing the different possible ways to examine the hypothesis.

What the analysis and discussion part are tasked to do is to lay out the results of a proposed approach to document clustering that integrates K-Means algorithm and TF-IDF matrix. This research is developing text mining methodology where easy understandable and using data can be used across multiple applications of which proposed technique is going to be evaluated to determine relevance and accuracy.

VI. IMPLICATIONS AND FUTURE

The conclusions from this research have a various application in several fields where document clustering is used crucially for information retrieval and management. To automatically clustering vast amounts of textual data, a scalable and effective approach that makes use of the TF-IDF matrix and K-Means algorithm is presented.

➢ *Practical Implications Include:*

- Better Information Retrieval: By giving users better ordered and pertinent search results, the efficiency of TF-IDF-based document clustering can improve information retrieval systems.
- Knowledge Discovery: By finding patterns, trends, and linkages across big text corpora, clustering related documents together aids in knowledge discovery.
- Automated document clustering facilitates the organization and summarization of documents, enhancing decision-making and document management procedures.
- Personalization and Recommendation Systems: By providing consumers with customized material, grouping documents according to content similarity can improve these features.
- Restrictions of the Research

➢ *Despite the Contributions of this Research, Several Limitations should be Acknowledged:*

- Data Representation: The Caliber of the textual material and the pre-processing methods used determine how well the TF-IDF matrix performs. The clustering findings might be impacted by noise or inaccuracies in the data.
- Algorithm Sensitivity: The K-Means algorithm's performance may vary depending on the parameters used, including the number of clusters and initialization technique. Results from clustering may be unsatisfactory if the parameters are set incorrectly.
- Even though K-Means is renowned for its scalability, there may still be computational difficulties when clustering massive datasets with millions of documents.
- Evaluation Metrics: Alternative metrics or qualitative evaluations can be required if the evaluation metrics selected to gauge the quality of the clustering process do not fully capture all facets of the clustering performance.

VII. SUGGESTIONS FOR FUTURE RESEARCH

➢ *Building upon the Findings of this Study, Several avenues for Future Research can be explored:*

- Enhanced Feature Representation: Investigation on alternative feature representation methods, such as word embedding or topic modelling, to capture more nuanced semantic relationships among documents.
- Advanced Clustering Techniques: Exploring of advanced clustering techniques, like spectral clustering, hierarchical clustering, and density-based clustering, to compare their performance with our method.
- Integration of Domain Knowledge: Incorporating domain-specific knowledge into the clustering process to improve accuracy and interpretability of clustering method.
- Dynamic Clustering: Could develop a dynamic clustering algorithms capable of adapting, i.e., to change in the dataset over time, thus enabling real-time clustering of streaming textual data.
- Interpretability and Visualization: By focusing on improving the interpretability and visualization of clustering results we can aid users in understanding and navigating large text corpora effectively.

VIII. CONCLUSION

➢ *Summary of Key Findings*

This study examined the use of the TF-IDF matrix to document clustering using the K-Means technique. To describe the textual data, the TF-IDF matrix was built for the study, and the K-Means technique was used to group the documents into clusters based on the similarity of their content.

➢ *Research's main Conclusions are:*

- The TF-IDF matrix which measures the significance of the documents, is used as a useful feature representation.
- Based on the TF-IDF based representations of documents is used in the K-Means technique to show scalability and efficiency in clustering of big dataset.
- Davies-Bouldin index and silhouette score are used to evaluate performance of suggested methods.

➢ *Contribution to the Field*

This research will make valuable contributions in the fields of text mining and document clustering in multiple aspects. After examining the TF-IDF matrix and k mean algorithm for document clustering, practical insights are provided on utilizing TF-IDF-based document clustering for document management and classification. When this proposed method is compared with baseline technique it enhances our understanding of current document clustering methods and their applications in the document clustering.

➢ *Final Remarks*

This study emphasizes how crucial it is to use sophisticated text mining methods for efficient document clustering, such as the TF-IDF matrix and K-Means algorithm. Integrating these approaches is a useful way to organize and draw conclusions from a large amount of data. Further research into more complex clustering algorithms and different feature representation strategies is required to enhance the performance of document clustering systems. All things considered, this study advances text mining and offers scholars and practitioners insightful information about using the TF-IDF matrix for document clustering with the K-Means algorithm.

REFERENCES

[1]. Chen, Tianqi, and Carlos Guestrin. "Xgboost: A scalable tree boosting system." *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. 2016.

[2]. Dai, Wenyuan, et al. "Co-clustering based classification for out-of-domain documents." *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*. 2007.

[3]. Saxena, Amit, et al. "A review of clustering techniques and developments." *Neurocomputing* 267 (2017): 664-681.

[4]. Xiong, Caiquan, et al. "An improved k-means text clustering algorithm by optimizing initial cluster centers." *2016 7th International Conference on Cloud Computing and Big Data (CCBD)*. IEEE, 2016.

[5]. Luo, Congnan, Yanjun Li, and Soon M. Chung. "Text document clustering based on neighbors." *Data & Knowledge Engineering* 68.11 (2009): 1271-1288.

[6]. Bafna, Prafulla, Dhanya Pramod, and Anagha Vaidya. "Document clustering: TF-IDF approach." *2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT)*. IEEE, 2016.

[7]. V. Prabhas, M. Lal Saini, C. Mohith, R. Kumar, and B. Tripathi, “Segmentation of E-Commerce Data Using K-Means Clustering Algorithm,” *2023 Glob. Conf. Inf. Technol. Commun. GCITC 2023*, 2023, doi: 10.1109/GCITC60406.2023.10426132.

[8]. Balabantaray, Rakesh Chandra, Chandrali Sarma, and Monica Jha. "Document clustering using k-means and k-medoids." *arXiv preprint arXiv:1502.07938* (2015).

[9]. S. Chalechema, M. L. Saini, I. Perla, and A. V. Shivanand, “Customer Segmentation Using K Means Algorithm and RFM Model,” *Proc. - 4th IEEE 2023 Int. Conf. Comput. Commun. Intell. Syst. ICCCIS 2023*, pp. 393–398, 2023, doi: 10.1109/ICCCIS60361.2023.10425556

[10]. Kumbhar, Rutuja, et al. "Text document clustering using K-means algorithm with dimension reduction techniques." *2020 5th International Conference on Communication and Electronics Systems (ICCES)*. IEEE, 2020.

[11]. S. Kulshrestha and M. L. Saini, “Study for the Prediction of E-Commerce Business Market Growth using Machine Learning Algorithm,” *2020 5th IEEE Int. Conf. Recent Adv. Innov. Eng. ICRAIE 2020 - Proceeding*, 2020, doi: 10.1109/ICRAIE51050.2020.9358275

[12]. Al-Obaydy, WN Ibrahem, et al. "Document classification using term frequency-inverse document frequency and K-means clustering." *Indonesian Journal of Electrical Engineering and Computer Science* 27.3 (2022): 1517-1524.

[13]. J. Sarmah, M. L. Saini, A. Kumar, and V. Chasta, “Performance Analysis of Deep CNN, YOLO, and LeNet for Handwritten Digit Classification,” *Lect. Notes Networks Syst.*, vol. 844, pp. 215–227, 2024, doi: 10.1007/978-981-99-8479-4_16.

[14]. M. Sohail, M. Lal Saini, V. P. Singh, S. Dhir, and V. Patel, “A Comparative Study of Machine Learning and Deep Learning Algorithm for Handwritten Digit Recognition,” *Proc. Int. Conf. Contemp. Comput. Informatics, IC3I 2023*, pp. 1283–1288, 2023, doi: 10.1109/IC3I59117.2023.10397956

[15]. M. Lal Saini, B. Tripathi, and M. S. Mirza, “Evaluating the Performance of Deep Learning Models in Handwritten Digit Recognition,” *Proc. - Int. Conf. Technol. Adv. Comput. Sci. ICTACS 2023*, pp. 116–121, 2023, doi: 10.1109/ICTACS59847.2023.10390027.

[16]. M. L. Saini, R. S. Telikicharla, Mahadev, and D. C. Sati, “Handwritten English Script Recognition System Using CNN and LSTM,” *Proc. InC4 2024 - 2024 IEEE Int. Conf. Contemp. Comput. Commun.*, 2024, doi: 10.1109/InC460750.2024.10649099

[17]. Shilpi K, Lokesh Lodha, “ Performance Evaluation of Word Embedding Algorithms” *International Journal of Innovative Science and Research Technology (IJISRT),* Volume 8, Issue 12, ISSN No: 2456-2165