



Article

A Semantics-Based Clustering Approach for Online Laboratories Using K-Means and HAC Algorithms

Saad Hikmat Haji ¹, Karwan Jacksi ^{2,*}  and Razwan Mohmed Salah ³ 

¹ Department of Information Technology, Technical College of Informatics-Akre, Duhok Polytechnic University, Akre 42003, Iraq

² Department of Computer Science, University of Zakho, Zakho 42002, Iraq

³ Department of Computer Science, University of Duhok, Duhok 42001, Iraq

* Correspondence: karwan.jacksi@uoz.edu.krd

Abstract: Due to the availability of a vast amount of unstructured data in various forms (e.g., the web, social networks, etc.), the clustering of text documents has become increasingly important. Traditional clustering algorithms have not been able to solve this problem because the semantic relationships between words could not accurately represent the meaning of the documents. Thus, semantic document clustering has been extensively utilized to enhance the quality of text clustering. This method is called unsupervised learning and it involves grouping documents based on their meaning, not on common keywords. This paper introduces a new method that groups documents from online laboratory repositories based on the semantic similarity approach. In this work, the dataset is collected first by crawling the short real-time descriptions of the online laboratories' repositories from the Web. A vector space is created using frequency-inverse document frequency (TF-IDF) and clustering is done using the K-Means and Hierarchical Agglomerative Clustering (HAC) algorithms with different linkages. Three scenarios are considered: without preprocessing (WoPP); preprocessing with steaming (PPwS); and preprocessing without steaming (PPwOS). Several metrics have been used for evaluating experiments: Silhouette average, purity, V-measure, F1-measure, accuracy score, homogeneity score, completeness and NMI score (consisting of five datasets: online labs, 20 NewsGroups, Txt_sentoken, NLTK_Brown and NLTK_Reuters). Finally, by creating an interactive webpage, the results of the proposed work are contrasted and visualized.

Keywords: document clustering; semantic similarity; online laboratories; crawling; TF-IDF; K-means; HAC

MSC: 68U15



Citation: Haji, S.H.; Jacksi, K.; Salah, R.M. A Semantics-Based Clustering Approach for Online Laboratories Using K-Means and HAC Algorithms. *Mathematics* **2023**, *11*, 548. <https://doi.org/10.3390/math11030548>

Academic Editors: Abeer Alsadoon and Luis Coelho

Received: 7 December 2022

Revised: 12 January 2023

Accepted: 13 January 2023

Published: 19 January 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

In the domains of information retrieval and text mining, analyzing and utilizing enormous numbers of text documents are crucial challenges. Clustering data into meaningful categories is an essential task that entails subdividing a collection of data objects into smaller groups. This method is used in data mining, information retrieval and knowledge discovery to identify hidden patterns and objects inside diverse types of data. Text clustering is the process of grouping a set of unlabeled texts so that texts within the same cluster are more similar to those within other groups [1–4].

Document clustering (or text clustering) is an effective approach to organizing text documents into meaningful groups for navigating and mining valuable information [5–8]. It groups documents into relevant clusters that can be used to peruse a collection of documents or to organize search engine results in response to a user's query. Traditionally, the characteristics used for clustering consisted of single, unique, or compound words from the document collection and did not consider semantic relationships. This could lead to synonym and polysemous problems. A bag of original words cannot effectively reflect the content of a

document or produce meaningful clusters. Therefore, using semantic clustering can improve document clustering approaches that incorporate the meaning of words [9,10]. The main process of document clustering is shown in Figure 1 [11].

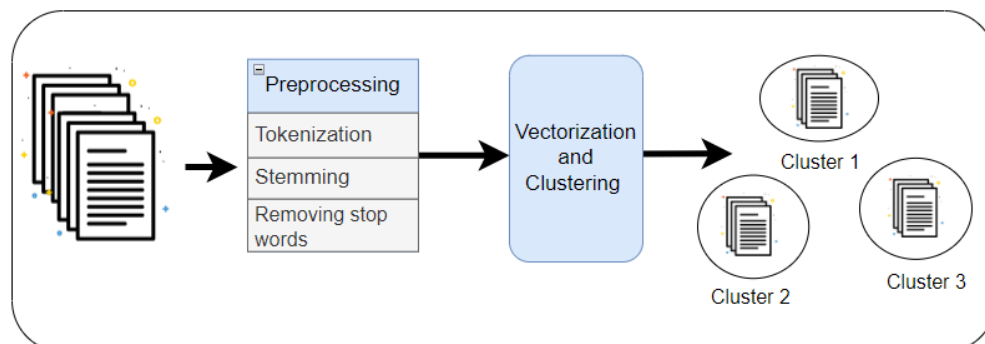


Figure 1. Document clustering phases.

Semantic clustering is a technique for categorizing data that are semantically related to one another. It refers to the point at which a dataset is divided into distinct clusters, such that two items within the same cluster are semantically equivalent. In comparison, two items from different clusters are dissimilar. By discovering semantic dissimilarities, semantic document clustering offers a substantial advantage for removing irrelevant documents [2,12,13].

Several ways of clustering documents have been presented. Before applying a clustering algorithm, the Term Frequency Inverse Document Frequency (TF-IDF) is a standard method for defining a corpus. In addition, Word Embedding techniques (i.e., Glove and Word2vec) are used to represent words as n-dimensional vectors grouped by a clustering algorithm such as K-means, hierarchical agglomerative clustering (HAC), DBSCAN, etc. [10,14–17].

This research proposes an approach to semantically clustering online laboratories in real time. Online laboratories are remotely conducted experiments that are intended to support, not replace, hands-on laboratories. They are intended to enhance students' skills, provide them with experience and aid them in becoming acquainted with real-life phenomena [18]. Online experiments are found in various domains including electronics, mechatronics, informatics, etc. Online laboratories have played an essential role in science and engineering teaching during unusual times such as the COVID-19 pandemic. Online laboratories can supplement rather than replace traditional laboratories by providing students with specific engineering experience and by allowing them to investigate systems and their real-world behaviors [19,20].

In this work, the initial step is to crawl the descriptions of online laboratories using web crawling technology and use preprocessing methods to tokenize, stem and remove stop words. Next, the TF-IDF is applied to the preprocessed data to transform it into an integer form so that clustering algorithms can use it. Finally, two algorithms, K-Means and HAC, were utilized for clustering and a comparison of the outcomes with different datasets was performed using several internal and external evaluation measures.

This study investigates the use of ML clustering algorithms on small datasets (which consist of online laboratories' descriptions) and applies two different ML clustering algorithms (K-Means and HAC clustering algorithms). In the clustering use case, we aim to find relevant groups within the online laboratory dataset. Our main contributions are the handling of small datasets in ML clustering and real-time datasets using the crawling system to have an up-to-date online laboratory dataset.

This study is divided into eight sections. The first section consists of a general introduction and summary of the research. The second section consists of the literature review. The methodology of the proposed system is described in Section 3. Section 4 describes the representation of clustering. The experimental results and implementation are presented in Section 5. Section 6 presents the results and a discussion thereof. In Section 7, the

proposed work is compared to previous works and the conclusion of the research is shown in Section 8.

2. Related Work

Various comparative studies on clustering approaches have been conducted, but no single approach has been determined to be superior to others. Issues including precision, training time and scalability are key factors in finding the optimal method for semantically comparable document clustering.

Salih and Jacksi, 2020 [21] applied K-Means and Wards algorithms to document clustering. Each algorithm has been implemented in three different ways: without preprocessing; preprocessing with stemming; and preprocessing without stemming. For data representation, the TF-IDF method was used to vectorize the data. Silhouette and other metrics were used to determine how similar the five unique datasets were. As a result, the Wards method is ineffective for huge datasets. Their research demonstrated the challenge of finding a single strategy that works well for clustering all types of datasets.

Jacksi et al. 2020 [5] developed a new technique for document clustering on the basis of semantic similarity. It has been used for K-Means and HAC clustering algorithms. Their technique generated the vector space that was generated by TF-IDF, then compared the results of the algorithms using multiple datasets and internal and external evaluation metrics. Based on their conclusions, the K-Means algorithm has excellent performance but is slower than the HAC algorithm. In addition, in 2021, Jalal and Ali [22] proposed document clustering that could cluster research paper text documents into useful groups. TF-IDF was utilized for data vectorization. The documents are classified into principal groups using the highest cosine similarity score. The findings revealed that over 96% of papers with similar scopes could be clustered.

Mehta et al. 2021 [23] proposed a clustering strategy that combines the effectiveness of statistical characteristics using the TF-IDF method and semantic features using lexical chains to integrate semantic features with WordNet relations, including identity, synonymy, hypernymy and meronymy. K-Means clustering techniques were also used in their investigation. Experimental results show that the suggested method has outperformed numerous clustering approaches based solely on semantic variables and statistical data. On the other hand, Mohammed et al. [15] proposed the use of Glove word embedding and DBSCAN clustering for semantic document clustering. Following preprocessing, they employ the Glove word embedding algorithm with the data's PPwS and PPWoS, then the DBSCAN clustering technique. Experimentally, the proposed system outperforms a system using the TF-IDF and K-Means clustering methods when the dataset is large and meaningful. However, if the dataset is small, the TF-IDF and K-Means algorithms perform better than the suggested method. Moreover, Ma and Zhang, 2015 [24] preprocessed the 20 newsgroups dataset with the word2vec and the K-Means clustering algorithms. A high-dimensional word vector has been generated via the word2vec generator for selecting features. A linear calculation was done to reduce the massive quantity of word vectors to word vectors with semantic similarity. Then, they applied the K-Means clustering approach to classify sentences with identical or closely related semantic meanings. The authors used the glove model rather than the word2vec model to convert the dataset to a word vector and used principal component analysis (PCA) to present and visualize the datasets into a data frame to find a smaller number of unmatched words, as opposed to a linear calculation to reduce a word's dimensionality vector and visualize and extract features from word vectors.

Furthermore, Adebisi et al. 2020 [25] presented a model for document grouping based on semantic similarity and utilized publishing documents that shared particular keywords. For this case study, published materials from 2010 to 2018 from randomly selected Nigerian universities were retrieved and used. This strategy was described as a clustering problem. The DBpedia and WordNet ontologies were employed to enhance the precision of the clustering results and capture domain terms that are semantically connected in the publishing dataset. Text documents were modelled using LSI and TF-IDF to generate

feature vectors. Feature vectors were clustered using the K-Means clustering approach. The silhouette analysis technique was used to examine the clustering results, which revealed an average intra-cluster similarity of 0.80 across all data points. The proposed method solves the difficulties of sparse data and high dimensionality that are associated with conventional document clustering methods.

Based on the results of the literature review, the performance and accuracy of each clustering algorithm vary because of the reliance on the dataset type and methods used for preprocessing and vectorization. Each clustering algorithm cannot give the best results for all types of datasets. Some clustering algorithms can provide good results for large datasets while others give good results for small ones. As a result, we proposed a semantic approach for clustering the small datasets, which is the description of online laboratories (real-time datasets) by using the most commonly used methods of preprocessing, vectorization and clustering algorithms for small datasets.

3. The Methodology of Proposed work

The proposed system consists of several significant steps that play a vital role in the proposed semantic approach, as illustrated in Figure 2.

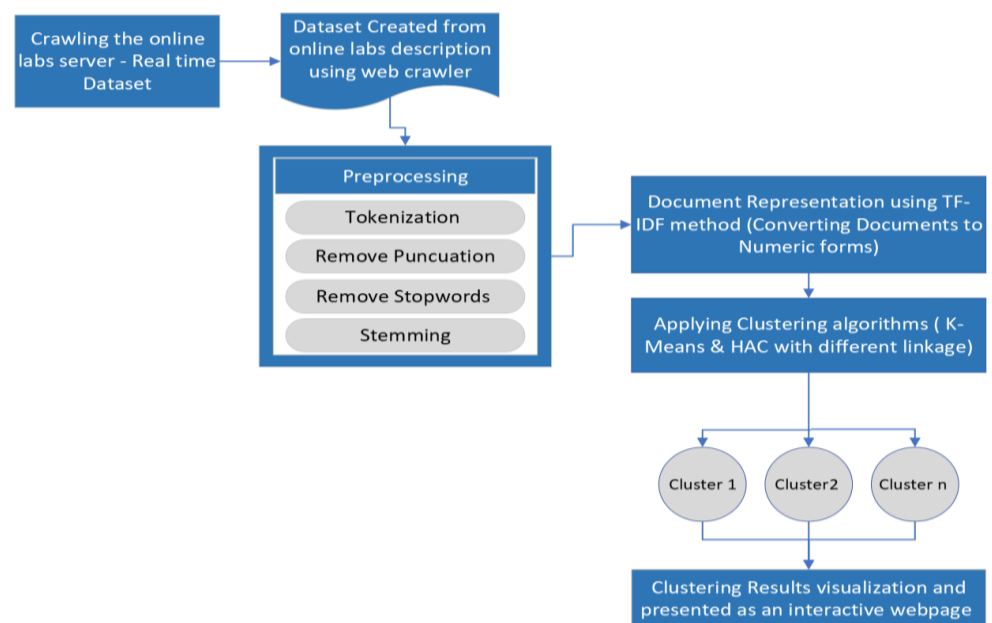


Figure 2. Methodological steps of the proposed system.

3.1. Data Gathering and Crawling

The first step is crawling the description of online laboratories using web crawling technology: separating the description, genres and titles by using the crawler or web crawling (data crawling). This is employed in data mining, gathering information from the Internet to have recent data datasets for use in online labs. The reason for applying clustering algorithms to online laboratories is that many universities and institutes around the world have been widely using online laboratories in their education system and they have a vital role in science education during unusual times (such as the COVID-19 pandemic). By clustering these labs, the online laboratories' interests reach the specific type of labs more easily and more accurately using less time.

3.2. Document Preparation

Preprocessing is necessary for the documents to become more convenient. This is a data mining method that turns raw data into a format that is easier to understand, more

useful and more efficient. After data preprocessing, the data will be easier to see and understand. Document preprocessing decreases the size of the datasets.

The proposed system includes the following preprocessing stages. Tokenizing the documents is the first stage. Tokenization is the process of separating the raw text into tiny parts. It transforms raw text into tokens or individual words. These tokens aid in comprehending the context and constructing the NLP model. By evaluating the sequence of words, tokenization facilitates the interpretation of the text's meaning. Stemming is conducted on the tokens to recover each word's root in the second stage, which refers to the reduction of a word to its origin or root stem. For example, the word "develop" can either appear as "developed" or "developing." These three terms become "develop" when stemmed.

Stop word removal (such as the, a, an, un, as, he, she, it, they, you, their, etc.) is one of the most prevalent preprocessing procedures in NLP applications. The goal of this stage is to exclude words that often appear in all corpus documents.

3.3. Representing Document

This step aims to represent document input as a fixed-length vector describing the document's content in order to simplify documents and make them easier to manipulate. TF-IDF is a statistical metric that analyzes the relevance of a word in a document relative to a corpus of documents. Using TF-IDF, the documents are converted to a numeric format following preprocessing. TF identifies the frequency with which a term appears in a document, whereas IDF identifies the importance of a phrase. By multiplying TF and IDF, the numerical weight of the words is calculated. This is often utilized in Natural Language Processing and Information Retrieval [26,27].

$$TF(t) = \frac{\text{Number of times term } t \text{ occurs in a document}}{\text{total number of terms in the document}} \quad (1)$$

$$IDF(t) = \text{Log}\left(\frac{\text{Total number of documents}}{\text{Number of documents with term } t}\right) \quad (2)$$

$$TF - IDF(t, \text{Document}) = TF \times IDF \quad (3)$$

3.4. Using the K-Means and HAC Clustering Algorithms

Clustering is an issue of unsupervised learning. Its primary goal is to collect similar data inside a cluster so that data within the same cluster are more similar to data in other clusters. It is a collection of objects classified according to their similarities and differences [5,28]. Therefore, the TF-IDF matrix for both clustering algorithms, namely K-Means and HAC, has been selected for application.

3.4.1. K-Means Algorithm

K-Means Clustering is a technique for unsupervised learning that clusters unlabeled datasets into discrete groups. It is an iterative strategy that divides the unlabeled dataset into K clusters so that each dataset belongs to only one group with similar characteristics [21,29]. The steps of K-Means are as described below:

- Determine the number of clusters by selecting value K.
- Select K points or centroids randomly.
- Assign each data point to its nearest centroid, so constructing the K clusters specified in advance.
- Determine the variance and add a new centroid to each cluster.
- Repeat the third step, this time reassigning each data point to the cluster's new nearest centroid.
- Proceed to step 4 if reassignment occurs; otherwise, proceed to FINISH.
- The model is complete.

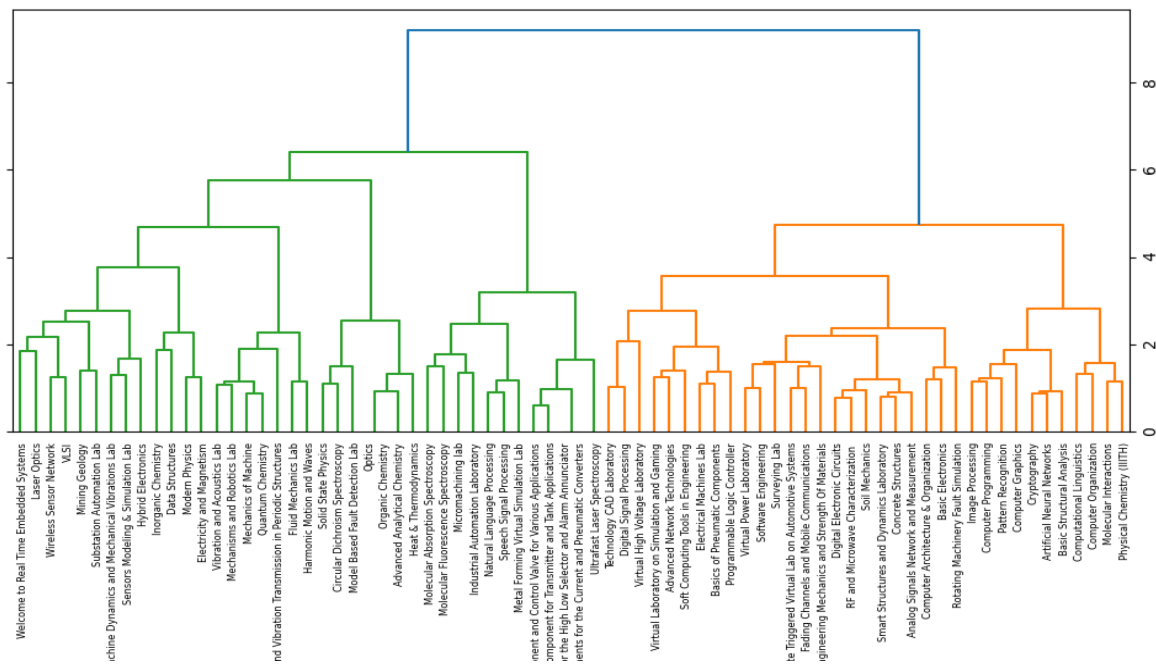


Figure 4. Optimal number of clusters for HAC algorithm using dendrogram.

The clustering results allow users to filter the online laboratory types based on semantic clustering using a web application which is installed on the Semantic Web Lab website on the servers of the University of Zakho. The results of the proposed application are accessible via this link: swlab.uoz.edu.krd/scolabs. A snapshot of the system is shown in Figure 5.

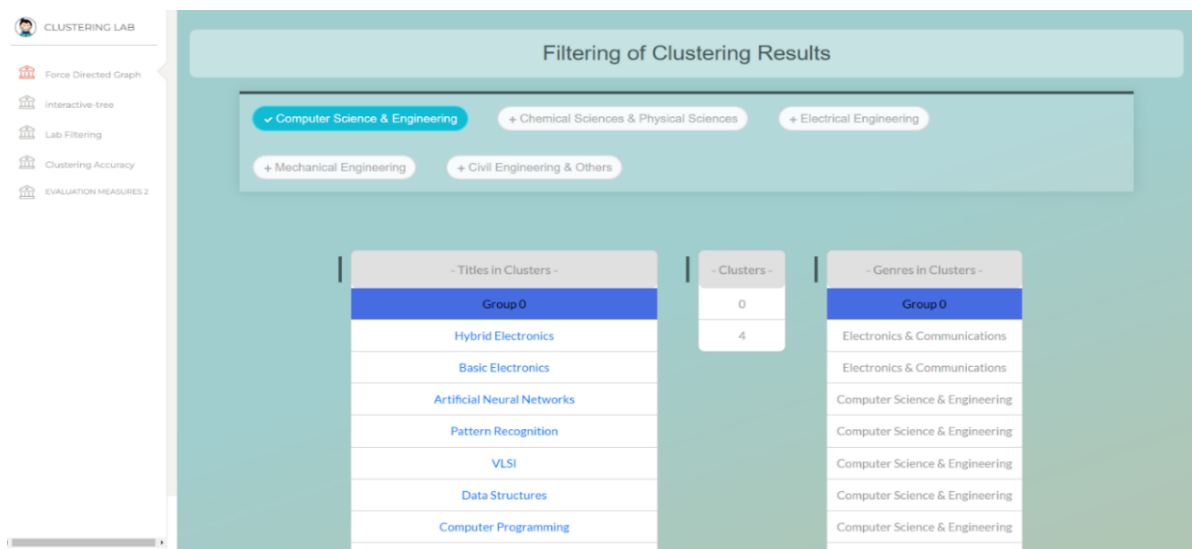


Figure 5. Filtering the clustering results.

5. Experimental Results and Implementation

Two different algorithms have been selected for five different datasets. These two algorithms are K-means and HAC. Results were made for the K-Means algorithm so they could be compared with the HAC algorithm. The results that are used are based on three different scenarios: WoPP, PPwS and PPwoS. The main dataset chosen in this proposed system is the description of the online laboratory repository. This is done by crawling the

short real-time description of the online laboratory repository from the online laboratory servers using web crawling technology (www.vlab.co.in). Four other datasets are used: NLTK_Brown, 20 newsgroups, Txt_sentoken and NLTK_Reuters. In addition, the proposed system is evaluated using internal and external evaluation metrics.

5.1. Datasets

For the proposed system, one primary dataset from an online repository of laboratory data and four regularly used datasets in document clustering research are utilized. They are comparable in terms of document number, class number and number of words. The largest of these data sets consisted of 19,715 documents, while the smallest consisted of 95 items.

- NLTK_Reuters: Reuters-21578 is the most frequent dataset used to evaluate document categorization and document clustering. It has 19,715 unique documents.
- NLTK_Brown: The corpus comprises one million words of American English writing published in 1961 and comprises 500 unique documents.
- Txt-Sentoken: Including negative and positive folders of review movies, this medium-sized dataset has 2000 unique documents.
- 20 Newsgroups: This is a highly frequent and valid dataset containing 18,846 unique documents and is used to test numerous data mining methods, text application and machine learning methods, etc.
- Online Labs: This is a small dataset which contains the descriptions of 95 online laboratories obtained by crawling the short real-time description of the online laboratories' repository from the online laboratories' servers. Each lab's titles, genres and descriptions are included.

5.2. Optimal Cluster Number

The elbow approach and the silhouette coefficient are two of the most commonly used methods to determine the optimal number of clusters for the K-Means algorithm [31]. The elbow method, depicted in Figure 6, is probably the most well-known technique, in which the sum of squares at each number of clusters (Equation (4)) is calculated and graphed. The user looks for a shift in slope from steep to shallow (an elbow) to determine the best number of clusters. Based on the most prevalent approaches for choosing the best number of clusters, 13 is the optimal number.

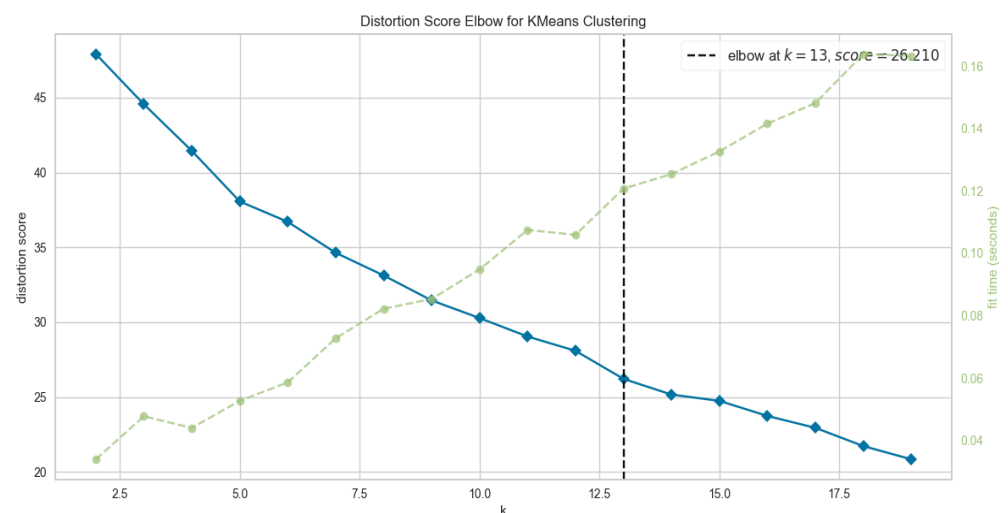


Figure 6. Optimal number of clusters using elbow method.

Additionally, the silhouette coefficient is a method for calculating the appropriate number of clusters and understanding and confirming cluster consistency. The silhouette technique computes silhouette coefficients for each point, representing the degree to which

a point resembles its cluster compared to other clusters. A high silhouette value indicates that an object is well-adapted to its cluster but poorly suited to clusters in its surroundings. In addition, according to the silhouette coefficient approach illustrated in Figure 7, the optimal number of clusters is 13.

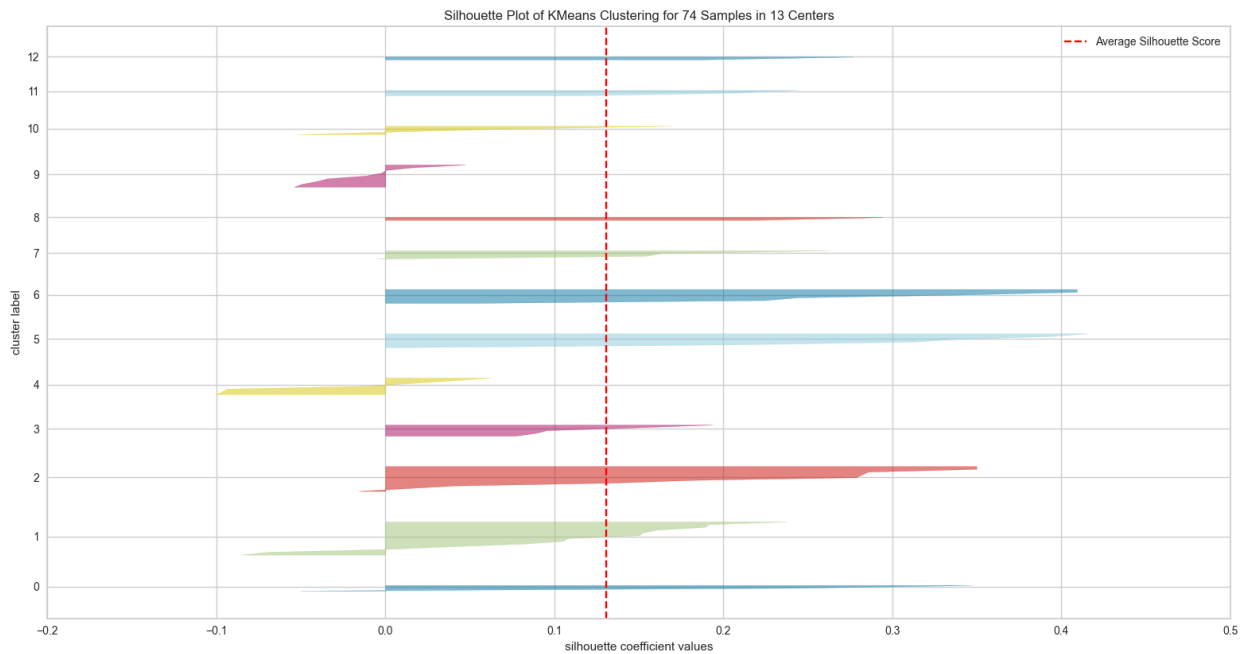


Figure 7. The Silhouette score when the number of cluster K = 13.

5.3. Evaluation Measures

Cluster analysis involves evaluating clustering results. Clustering is evaluated by comparing the result to the standard data structure. The most challenging part of the clustering process is evaluating the results. Our research uses seven evaluation measures, including internal evaluation metrics such as Silhouette Ratio:

$$\bar{S} = \frac{1}{n} \sum_{i=1}^n \left(\frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \right) \tag{4}$$

$a(i)$ represents the average distance between sample i and other samples in the cluster, while $b(i)$ represents the minimum distance between sample i and the other clusters.

Other external evaluation metrics are also used such as purity, homogeneity, completeness, V-measure, F1-measure and accuracy.

6. Results and Discussion

For each of the datasets used, K-Means and HAC algorithms are proposed. The K-Means method produced results that were compared to the HAC algorithm. Table 1 shows the similarity ratio using the internal evaluation metric silhouette score for all three scenarios: WoPP, PPWoS and PPwS.

Table 1. Similarity Ratio for the optimal cluster number = 13 with different scenarios.

Scenarios	Datasets	Partitioning Clustering	HAC Algorithm			
		K-Means	Ward	Single	Complete	Average
WoPP	Online Lab	0.130429856	0.1281765	−0.1262827	0.1327986	0.1600882
	20 Newsgroups	0.14729813	0.0646622	−0.1027132	0.0934307	0.0732552
	txt_sentoken	0.025685631	−0.0067918	−0.0107076	−0.001784	−0.0072473
	NLTK_Brown	0.029879957	0.0241643	−0.0196633	0.0243814	0.0353152
	NLTK_Reuters	0.385767347	0.3211324	0.0703577	0.2430639	0.320531
PPWoS	Online Lab	0.14830793	0.1472155	−0.0367298	0.134025	0.1433745
	20 Newsgroups	0.163033891	0.0910541	−0.1187512	0.0851621	0.0791069
	txt_sentoken	0.028361561	−0.0095231	−0.0115982	−0.009855	−0.0096819
	NLTK_Brown	0.02926897	0.0194938	−0.006007	0.0313465	0.0311442
	NLTK_Reuters	0.335450842	0.2775758	0.0457569	0.225533	0.2775008
PPwS	Online Lab	0.142849649	0.1494705	−0.1247673	0.1216341	0.1274566
	20 Newsgroups	0.095702334	0.0355147	−0.1545218	−0.00648	0.0144914
	txt_sentoken	0.012576785	−0.0126844	0.0057517	−0.010246	−0.0079978
	NLTK_Brown	0.032447364	0.0152929	0.0046581	0.0218663	0.0286168
	NLTK_Reuters	0.360457389	0.3131754	0.0532977	0.239106	0.3018736

6.1. Internal Evaluation for the Proposed System

Table 1 shows the outcome of the proposed system after applying the Silhouette score, an internal evaluation metric, to all datasets generated by the HAC (various linkages) and K-Means algorithms, with the optimal number of clusters determined by the elbow method and the silhouette coefficient.

For the first scenario, implementation without preprocessing, Table 1 illustrates the similarity ratio for the optimal cluster number = 13, which is obtained using the elbow method and silhouette coefficient. The highest similarity ratio for this scenario using the internal evaluation metric is obtained for the dataset NLTK_Reuters for the two clustering algorithms. In contrast, the worst similarity ratio was achieved for both algorithms with the dataset Txt_sentoken. Finally, to conclude the self-evaluation for our proposed real-time dataset online laboratories, the algorithm HAC (average linkage) outperforms the K-Means clustering algorithm as shown in Figure 8.

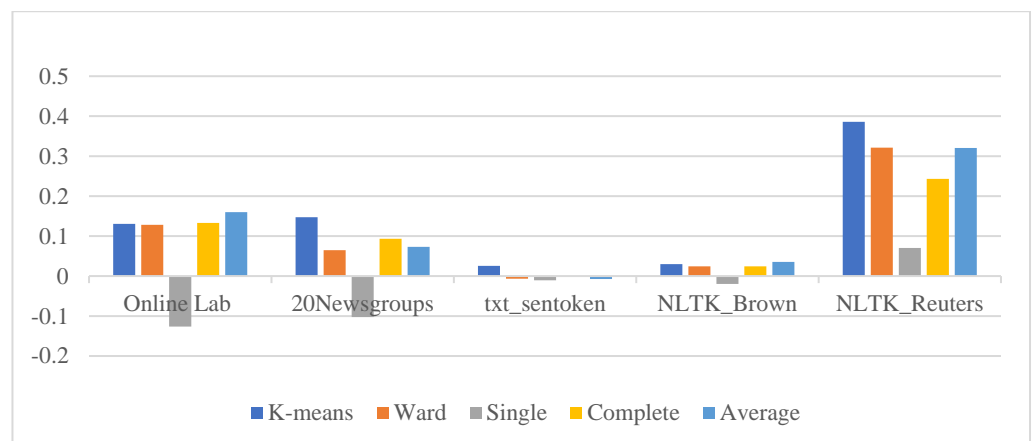


Figure 8. Similarity ratio without preprocessing (WoPP).

For the second scenario, PPWoS, Table 1 illustrates the similarity ratio for the optimal cluster number = 13. The highest similarity ratio for the second scenario is obtained for the dataset NLTK_Reuters for the two clustering algorithms. In comparison, the worst similarity ratio was achieved for both algorithms with the dataset Txt_sentoken. As a self-

evaluation for our proposed real-time dataset online laboratories, the K-Means algorithm outperforms the HAC algorithm as shown in Figure 9.

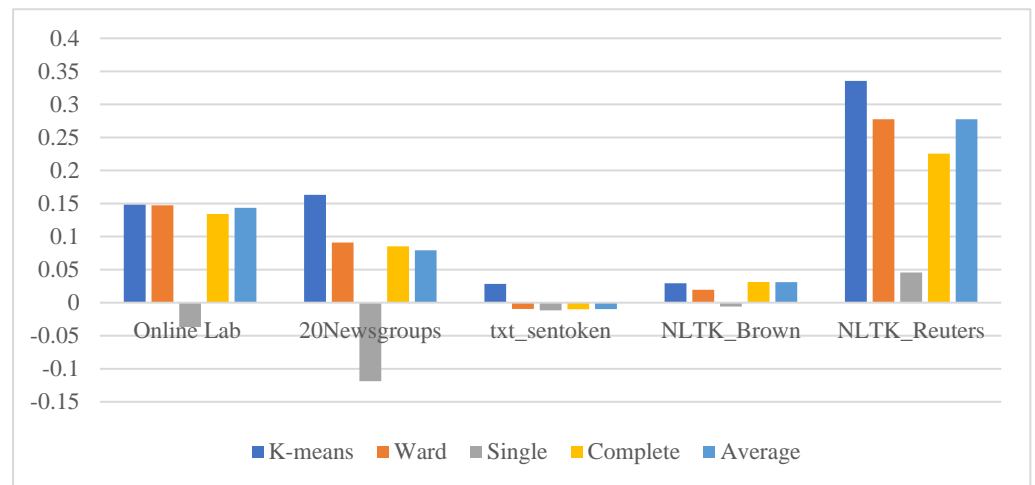


Figure 9. Similarity ratio preprocessing without stemming (PPWoS).

For the third scenario, PPwS, Table 1 shows the similarity ratio for the optimal cluster number = 13. Like the two previous scenarios, the highest similarity ratio is achieved for the dataset NLTK_Reuters and the worst similarity ratio is achieved for the dataset Txt_sentoken.

As a self-evaluation for our proposed real-time dataset online laboratories using the third scenario, the HAC (Ward method) outperforms the K-Means algorithm, as seen in Figure 10.

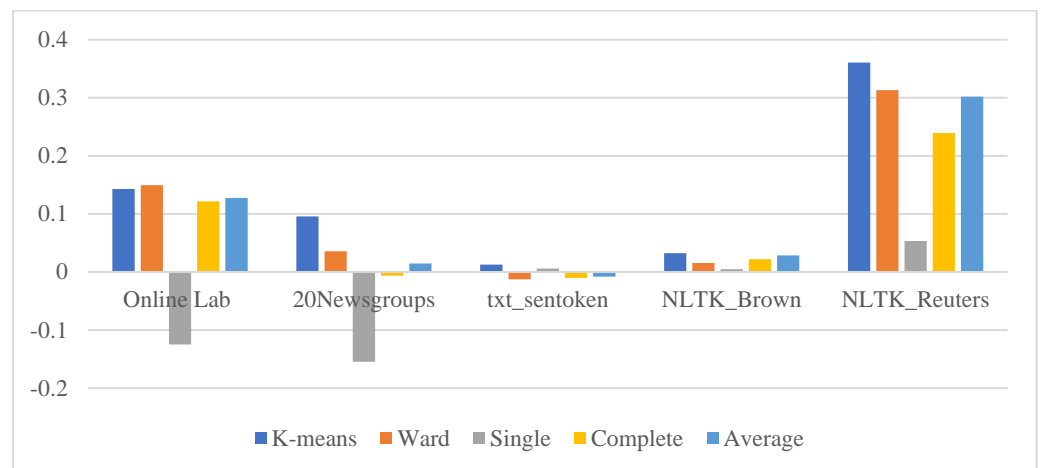


Figure 10. Similarity ratio preprocessing with stemming (PPWS).

6.2. External Evaluation for the Proposed System

Multiple criteria are employed for external evaluation utilizing the five datasets for both K-Means and HAC (Ward linkage) algorithms for various scenarios and optimal cluster number = 13. The outcomes of external measurements are depicted in Table 2.

- Evaluation metrics for the K-Means algorithm of optimal cluster number K = 13 as shown in Table 2:

The implementation WoPP for all five different datasets using external evaluation measures shows that the online dataset labs, our proposed dataset, recorded the highest similarity ratio for V-measure, homogeneity and NMI score. While the dataset Txt_sentoken

gives the best similarity ratio for the metric’s purity and accuracy, the dataset NLTK-Brown recorded the highest similarity ratio for F-measure and completeness.

In addition, for the scenario of PPWoS, the dataset NLTK_Brown gives the best results for V-measure, F-measure, accuracy, completeness and NMI score. In contrast, the dataset Txt_sentoken provides the best outcome for the purity metric and the online lab’s dataset recorded the best result for the homogeneity metric.

For the scenario of PPwS, the proposed dataset, which is an online lab, also gives the highest results using the evaluation metrics V-measure, F-measure, accuracy, homogeneity, completeness and NMI score, while the dataset Txt_sentoken provides the highest result for the purity evaluation metric.

- Evaluation metrics for the HAC (Ward method) algorithm of optimal cluster number K = 13 as shown in Table 2:

The implementation without preprocessing (WoPP) for all five different datasets using external evaluation measures shows that the proposed dataset, an online lab, gives the best results using the evaluation metrics V-measure, homogeneity and NMI score. While the dataset NLTK_Brown provides the highest results using F-measure, accuracy and completeness metrics, the dataset Txt_sentoken gives the highest result using the external purity metric.

For the PPWoS, the results show that the online dataset labs provide the best results using the metrics V-measure, homogeneity and NMI score. In comparison, the Txt_sentoken dataset provides the highest result using the metrics purity, F-measure and accuracy and the NLTK_Brown dataset gives the highest result using the completeness metric.

In addition, for the PPwS, the results show that for our proposed dataset, which consists of online labs, the highest results were achieved using the metrics purity, V-measure, homogeneity, completeness and NMI score. In contrast, the Txt_sentoken gives the highest outcomes using F-measure and the accuracy metrics.

Table 2. External Evaluation metrics for both algorithms with different scenarios for the optimal cluster number 13.

Algorithms	Scenarios	Datasets	Purity	V-Measure	F-Measure	Accuracy	Homogeneity	Completeness	NMI-Score
K-Means	WoPP	Online Lab	0.462162162	0.345549506	0.014574899	0.024324324	0.397271606	0.305743723	0.345549506
		20 Newsgroups	0.102408999	0.043879554	0.036345756	0.047808554	0.039965172	0.048643981	0.043879554
		txt_sentoken	0.6075	0.019775376	0.016650513	0.0635	0.045449135	0.012636911	0.019775376
		NLTK_Brown	0.36	0.326133575	0.057811277	0.06	0.320056571	0.332445818	0.326133575
		NLTK_Reuters	0.542404301	0.134296242	0.001147675	0.025210863	0.148022042	0.122899964	0.134296242
	PPWoS	Online Lab	0.445946	0.33343	0.062572	0.094595	0.383353	0.295011	0.33343
		20 Newsgroups	0.104266	0.042329	0.043036	0.058633	0.038738	0.046653	0.042329
		txt_sentoken	0.611	0.021128	0.015683	0.057	0.048037	0.013542	0.021128
		NLTK_Brown	0.342	0.339315	0.06665	0.096	0.336702	0.34197	0.339315
		NLTK_Reuters	0.534202	0.124766	0.003014	0.046992	0.138458	0.113538	0.124766
	PPwS	Online Lab	0.554054	0.438731	0.081511	0.108108	0.501941	0.389661	0.438731
		20 Newsgroups	0.109201	0.052181	0.048011	0.065266	0.047916	0.05728	0.052181
txt_sentoken		0.6135	0.027836	0.018374	0.0685	0.064871	0.01772	0.027836	
NLTK_Brown		0.368	0.372691	0.016878	0.024	0.365426	0.380252	0.372691	
NLTK_Reuters		0.535082	0.124158	0.002725	0.062656	0.137668	0.113062	0.124158	
HAC (Ward)	WoPP	Online Lab	0.513513514	0.387974527	0.060164835	0.094594595	0.448631654	0.341766126	0.387974527
		20 Newsgroups	0.105752	0.044192	0.035488	0.051682	0.039858	0.049583	0.044192
		txt_sentoken	0.593	0.022797	0.033348	0.151	0.046642	0.015085	0.022797
		NLTK_Brown	0.358	0.328791	0.098994	0.184	0.310418	0.349476	0.328791
		NLTK_Reuters	0.536287	0.125513	0.002942	0.067708	0.136927	0.115856	0.125513
	PPWoS	Online Lab	0.4864865	0.3583059	0.0508277	0.0810811	0.4107279	0.3177507	0.3583059
		20 Newsgroups	0.0984294	0.0358565	0.043526	0.0579964	0.032702	0.0396845	0.0358565
		txt_sentoken	0.591	0.0182871	0.0672182	0.3495	0.0349447	0.0123839	0.0182871
		NLTK_Brown	0.366	0.3403492	0.0095858	0.012	0.3214758	0.3615769	0.3403492
		NLTK_Reuters	0.5352674	0.1268945	0.0049828	0.0910186	0.1391429	0.116628	0.1268945

Table 2. Cont.

Algorithms	Scenarios	Datasets	Purity	V-Measure	F-Measure	Accuracy	Homogeneity	Completeness	NMI-Score
		Online Lab	0.5945946	0.4435351	0.0390575	0.0675676	0.5098827	0.3924661	0.4435351
		20 Newsgroups	0.1034702	0.0418827	0.0374483	0.0497718	0.0383529	0.0461281	0.0418827
	PPwS	txt_sentoken	0.568	0.0207478	0.0403083	0.216	0.0398396	0.0140262	0.0207478
		NLTK_Brown	0.354	0.3190606	0.0273966	0.038	0.3056518	0.3336998	0.3190606
		NLTK_Reuters	0.5359626	0.1304612	0.0017955	0.0375846	0.1441178	0.1191688	0.1304612

7. Comparison with Existing Studies

The proposed study is compared to the previous studies that are presented in the literature review in Section 2. Table 3 illustrates a detailed comparison of study objectives, similarity measures, evaluation measures, clustering algorithms and datasets used.

Table 3. Comparison of the proposed system with the literature.

Reference	Date	Semantic Approach	Dataset	Similarity Measures	Clustering Algorithms	Other Measures
Jalal and Ali [22]	2021	Text clustering based on semantic similarity	Research papers in BEEI journal	Precision and Recall	Cosine similarity	TF-IDF
Mehta et al. [23]	2021	Clustering large text datasets	Newsgroup, Reuters, Classic3	Silhouette coefficient, Purity, AMI,	K-Means	TF-IDF, WordNet
Salih and Jacksi, [21]	2020	Movies Clustering based on semantic similarity	IMDB and Wikipedia	Silhouette score, purity, V-Measure, F1-Measure, accuracy, homogeneity, completeness, NMI-Score	K-Means and HAC (Ward method)	TF-IDF
Jacksi et al. [5]	2020	Document clustering based on semantic approach	IMDB and Wikipedia	Purity, accuracy, F1-measure, NMI, Silhouette score	K-Means and HAC	TF-IDF
Adebiyi et al. [25]	2020	Research document Clustering	Publications from Nigerian universities	Silhouette analysis	K-Means	TF-IDF and WordNet
Mohammed et al. [15]	2020	Document Clustering based on semantic similarity	IMDB and Wikipedia	Silhouette average, purity, accuracy, F1, completeness, homogeneity and NMI score	DBSCAN, K-Means	Glove
Ma and Zhang, 2015 [24]	2015	Cluster large text dataset	20 Newsgroups	F1-micro score	K-Means	Word2Vec
Proposed system	2022	Online laboratorial document clustering based on semantic approach	Realtime educational online labs	Silhouette score, purity, V-Measure, F1-Measure, accuracy, homogeneity, completeness, NMI-Score	K-Means and HAC with different linkages (Single, Complete, Average, Ward)	TF-IDF

8. Conclusions

This study implemented a document clustering system based on semantic similarity using K-Means and HAC clustering algorithms. It applied them to the online laboratory repository by crawling the repository’s short real-time description from the online laboratory servers. The outcomes of clustering algorithms are presented in a format that facilitates comparison. Four more datasets were applied and analyzed to evaluate the proposed system using internal and external assessment measures. It is concluded that a general method capable of optimally grouping all sorts of datasets is unachievable. We have therefore sought to use two algorithms that can operate effectively with five distinct sorts of datasets. Each approach has been utilized in three scenarios: WoPP, PPWoS and PPwS. This study employed the Silhouette metric and seven other external evaluation criteria to determine the resemblance between the five distinct datasets. Applying the K-Means algorithm, the highest similarity ratio was obtained using the Silhouette score with the NLTK-Reuters dataset. In contrast, the HAC approach produced the lowest results with

the Txt-Sentoken dataset. Our research concluded that the HAC algorithm outperforms the K-Means algorithms for small datasets as a self-evaluation for our proposed online educational dataset. The limitation of the research was lay in finding good datasets from online educational laboratories.

The following are possible future directions for expansion of the proposed work on real-time educational online laboratories:

- Enlarge the system by including more clustering algorithms and approaches for small datasets such as Optics, Affinity propagation and K-Medoids.
- Applying these algorithms on other Virtual Learning Environments.
- Using word embedding methods such as Glove or Word2vec for word representation instead of the TF-IDF method.

Author Contributions: Writing—original draft, S.H.H.; Writing—review & editing, R.M.S.; Supervision, K.J. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Huang, A. Similarity Measures for Text Document Clustering. In Proceedings of the New Zealand Computer Science Research Student Conference (NZCSRSC), Christchurch, New Zealand, 14–18 April 2008.
2. Fatimi, S.; El, C.; Alaoui, L. A Framework for Semantic Text Clustering. *IJACSA* **2020**, *11*, 451–459. [[CrossRef](#)]
3. Djenouri, Y.; Belhadi, A.; Djenouri, D.; Lin, J.C.-W. Cluster-based information retrieval using pattern mining. *Appl. Intell.* **2021**, *51*, 1888–1903. [[CrossRef](#)]
4. Haji, S.H.; Abdulazeez, A.M.; Zeebaree, D.Q.; Ahmed, F.Y.H.; Zebari, D.A. The Impact of Different Data Mining Classification Techniques in Different Datasets. In Proceedings of the 2021 IEEE Symposium on Industrial Electronics & Applications (ISIEA), Virtual Event, 10–11 July 2021; IEEE: Langkawi Island, Malaysia, 2021; pp. 1–6.
5. ADC: Advanced document clustering using contextualized representations. *Expert Syst. Appl.* **2019**, *137*, 157–166. [[CrossRef](#)]
6. Shan, C.; Du, Y. A Web Service Clustering Method Based on Semantic Similarity and Multidimensional Scaling Analysis. *Sci. Program.* **2021**, *2021*, 1–12. [[CrossRef](#)]
7. Lwin, W. Impressive Approach for Documents Clustering Using Semantics Relations in Feature Extraction. In Proceedings of the 2019 the 9th International Workshop on Computer Science and Engineering, WCSE, Changsha, China, 18–20 October 2019.
8. Absalom, E.; Ezugwu, A.M.I. A Comprehensive Survey of Clustering Algorithms: State-Of-The-Art Machine Learning Applications, Taxonomy, Challenges, And Future Research Prospects. *Sci. Direct* **2022**, *110*, 165–193.
9. Al-Azzawy, D.S.; Al-Rufaye, F.M.L. Arabic words clustering by using K-means algorithm. In Proceedings of the 2017 Annual Conference on New Trends in Information & Communications Technology Applications (NTICT), Baghdad, Iraq, 7–9 March 2017; IEEE: Baghdad, Iraq, 2017; pp. 263–267.
10. Bafna, P.; Pramod, D.; Vaidya, A. Document Clustering: TF-IDF Approach. In Proceedings of the 2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT), Chennai, India, 3–5 March 2016; IEEE: New York, NY, USA, 2016; pp. 61–66.
11. Shaban, K. A Semantic Approach for Document Clustering. *JSW* **2009**, *4*, 391–404. [[CrossRef](#)]
12. Nair, S.R.; Gokul, G.; Vadakkan, A.A.; Pillai, A.G.; Thushara, M. Clustering of Research Documents—A Survey on Semantic Analysis and Keyword Extraction. In Proceedings of the 2021 6th International Conference for Convergence in Technology (I2CT), Maharashtra, India, 2–4 April 2021; IEEE: Maharashtra, India, 2021; pp. 1–6.
13. Alian, M.; Awajan, A. Arabic Semantic Similarity Approaches—Review. In Proceedings of the 2018 International Arab Conference on Information Technology (ACIT), Werdanye, Lebanon, 28–30 November; IEEE: Werdanye, Lebanon, 2018; pp. 1–6.
14. Ibrahim, R.K.; Zeebaree, S.R.M.; Jacksi, K.; Sadeeq, M.A.M.; Shukur, H.M.; Alkhayyat, A. Clustering Document based Semantic Similarity System using TFIDF and K-Mean. In Proceedings of the 2021 International Conference on Advanced Computer Applications (ACA), Maysan, Iraq, 25–26 July 2021; IEEE: Maysan, Iraq, 2021; pp. 28–33.
15. Mohammed, S.M.; Jacksi, K.; Zeebaree, S.R.M. Glove Word Embedding and DBSCAN algorithms for Semantic Document Clustering. In Proceedings of the 2020 International Conference on Advanced Science and Engineering (ICOASE), Duhok, Iraq, 23–24 December 2020.
16. Zhou, Y. Application of K-Means Clustering Algorithm in Energy Data Analysis. *Wirel. Commun. Mob. Comput.* **2022**, *2022*, 1–8. [[CrossRef](#)]
17. Jacksi, K.; Ibrahim, R.K.; Zeebaree, S.R.M.; Zebari, R.R.; Sadeeq, M.A.M. Clustering Documents based on Semantic Similarity using HAC and K-Mean Algorithms. In Proceedings of the 2020 International Conference on Advanced Science and Engineering (ICOASE), Duhok, Iraq, 23–24 December 2020; IEEE: Duhok, Iraq, 2020; pp. 205–210.

18. Salah, R.M.; Alves, G.R.; Abdulazeez, D.H.; Guerreiro, P.; Gustavsson, I. Why VISIR? Proliferative activities and collaborative work of VISIR system. In Proceedings of the 7th International Conference on Education and New Learning Technologies (EDULEARN15), Barcelona, Spain, 7–8 July 2015.
19. Radhamani, R.; Kumar, D.; Nizar, N.; Achuthan, K.; Nair, B.; Diwakar, S. What virtual laboratory usage tells us about laboratory skill education pre- and post-COVID-19: Focus on usage, behavior, intention and adoption. *Educ. Inf. Technol.* **2021**, *26*, 7477–7495. [[CrossRef](#)] [[PubMed](#)]
20. Qona'ah, N.; Devi, A.R.; Dana, I.M.G.M. Laboratory Clustering using K-Means, K-Medoids, and Model-Based Clustering. *IJAS* **2020**, *3*, 64. [[CrossRef](#)]
21. Salih, N.M.; Jacksi, K. Semantic Document Clustering using K-means algorithm and Ward's Method. In Proceedings of the 2020 International Conference on Advanced Science and Engineering (ICOASE), Duhok, Iraq, 23–24 December 2020.
22. Jalal, A.A.; Ali, B.H. Text documents clustering using data mining techniques. *IJECE* **2021**, *11*, 664. [[CrossRef](#)]
23. Mehta, V.; Bawa, S.; Singh, J. Semantic clustering: Combining statistical and semantic features for clustering of large text datasets. *Expert Syst. Appl.* **2021**, *174*, 114710. [[CrossRef](#)]
24. Ma, L.; Zhang, Y. Using Word2Vec to process big text data. In Proceedings of the 2015 IEEE International Conference on Big Data (Big Data), Washington, DC, USA, 29 October–1 November 2015; IEEE: Santa Clara, CA, USA, 2015; pp. 2895–2897.
25. Adebisi, M.O.; Adigun, E.B.; Ogundokun, R.O.; Adeniyi, A.E.; Ayegba, P.; Oladipupo, O.O. Semantics-based clustering approach for similar research area detection. *TELKOMNIKA* **2020**, *18*, 1874. [[CrossRef](#)]
26. Stanchev, L. Semantic Document Clustering Using a Similarity Graph. In Proceedings of the 2016 IEEE Tenth International Conference on Semantic Computing (ICSC), Laguna Hills, CA, USA, 4–6 February 2016; IEEE: Laguna Hills, CA, USA, 2016; pp. 1–8.
27. Vinoth, D.; Prabhavathy, P. A Short Text Clustering Approaches in Social Media. *ECS Trans.* **2022**, *107*, 1375–1386. [[CrossRef](#)]
28. Zandieh, P.; Shakibapoor, E. Clustering Data Text Based on Semantic. *Int. J. Comput.* **2017**, *26*, 8.
29. Huang, S.; Kang, Z.; Xu, Z.; Liu, Q. Robust deep k-means: An effective and simple method for data clustering. *Pattern Recognit.* **2021**, *117*, 107996. [[CrossRef](#)]
30. Liu, L.; Mosavat-Jahromi, H.; Cai, L.; Kidston, D. Hierarchical Agglomerative Clustering and LSTM-based Load Prediction for Dynamic Spectrum Allocation. In Proceedings of the 2021 IEEE 18th Annual Consumer Communications & Networking Conference (CCNC), Las Vegas, NV, USA, 9–12 January 2021; IEEE: Las Vegas, NV, USA, 2021; pp. 1–6.
31. Das, T.; Paitnaik, S.; Mishra, S.P. Identification of the Optimal Number of Clusters in Textual Data. In *Advances in Distributed Computing and Machine Learning*; Sahoo, J.P., Tripathy, A.K., Mohanty, M., Li, K.-C., Nayak, A.K., Eds.; Lecture Notes in Networks and Systems; Springer: Singapore, 2022; Volume 302, pp. 215–225. ISBN 9789811648069.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.