Performance analysis in text clustering using k-means and k-medoids algorithms for Malay crime documents

Rosmayati Mohemad[1], Nazratul Naziah Mohd Muhait[1], Noor Maizura Mohamad Noor[1], Zulaiha Ali Othman[2]
[1]Faculty of Ocean Engineering Technology and Informatics, Universiti Malaysia Terengganu, Terengganu, Malaysia
[2]Center for Artificial Intelligence Technology, Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia, Selangor, Malaysia

Article Info

Article history:

Received Jul 16, 2021
Revised Apr 19, 2022
Accepted May 12, 2022

Keywords:

Clustering
K-means
K-medoids
Malay language

ABSTRACT

Few studies on text clustering for the Malay language have been conducted due to some limitations that need to be addressed. The purpose of this article is to compare the two clustering algorithms of k-means and k-medoids using Euclidean distance similarity to determine which method is the best for clustering documents. Both algorithms are applied to 1,000 documents pertaining to housebreaking crimes involving a variety of different modus operandi. Comparability results indicate that the k-means algorithm performed the best at clustering the relevant documents, with a 78% accuracy rate. K-means clustering also achieves the best performance for cluster evaluation when comparing the average within-cluster distance to the k-medoids algorithm. However, k-medoids perform exceptionally well on the Davis Bouldin index (DBI). Furthermore, the accuracy of k-means is dependent on the number of initial clusters, where the appropriate cluster number can be determined using the elbow method.

Corresponding Author:

Rosmayati Mohemad
Faculty of Ocean Engineering Technology and Informatics, Universiti Malaysia Terengganu
21030 Kuala Nerus, Terengganu, Malaysia
Email: rosmayati@umt.edu.my

1. INTRODUCTION

The increase of unstructured texts in various forms on the internet has contributed to information overload, which is becoming a growing problem for many organizations to transform into insightful and actionable information. Kalra and Aggarwal [1] reported that 80-90% of potential growth data is expected to be available in the form of unstructured text that may probably contain hidden patterns and trends. To obtain valuable information, this raw data from social media posts, articles, documents, reports, images, audios, or videos must first be filtered, analyzed, and processed using machine learning (ML) techniques. By leveraging ML, meaningful insights contained within unstructured data could significantly empower decision-making processes in every field, including criminal investigation.

Criminal investigation entails investigative tasks, which necessitate collecting evidence and gathering information by a criminal investigator to make an accurate decision [2]. A police report is one of the valuable information sources for triggering a criminal investigation. In general, it is a first-hand information report containing narratives about an incident in which a crime or offence is suspected [3]. The hidden but useful information that lies within police reports necessitates a high level of skill on the part of criminal investigators to manually and regularly analyze the reports to find crime patterns and trend correlations, which is a challenging task due to the large volume of reports in an unstructured format [4]. The issues arise when too many reports are analyzed at the same time, and not enough criminal investigators are skilled at data analysis. This approach requires human intelligence, but it can be ineffective and prone to errors. Understanding crime patterns is critical for criminal investigators to comprehend crime. The emergence of text analytics tools that leverage ML can parse through large amounts of unstructured data from criminal reports to reveal criminal patterns of offenders, thereby assisting criminal investigators in making effective and better decisions on crime prevention [5]. The tools can aggregate, query and analyze criminal records to unveil crucial crime patterns automatically.

Recent advancements in computer and internet of things (IoT) technologies have resulted in the rapid development of text analytics, prompting many researchers and academicians to focus on data mining and text mining techniques, particularly in the criminal domain. The goal is to integrate human knowledge with machine intelligence through analytical processes that provide pertinent information on crime patterns and trend correlations to support human insight and decision-making [4], [6]. Data mining applications of crime analysis highly rely on structured data sets that employ data modelling techniques such naïve Bayes [7], neural networks [8], long short-term memory (LTSM) [9], decision trees (DT) [10] and k-means clustering [6], [11]. Meanwhile, text mining applications apply natural language processing techniques to deal with unstructured text before recognizing crime patterns in various languages, such as English, Arabic, and Swahili [4], [12], [13]. However, very little research has been conducted on the processing of Malay text for analyzing crime. Text mining is the process of identifying and extracting meaningful information that is widely utilized to solve real-world problems in the criminal domain for a variety of purposes, including preventing crime [13], detecting criminal activity [14], [15], matching crimes [4], and detecting crime hotspots [16], [17]. Clustering is one of the potential techniques in text mining for recognizing crime patterns.

Text document clustering (TDC) is an effective text mining technique that has been used with the aim of grouping a collection of comparable documents into the most relevant categories based on homogeneity or heterogeneity attributes. TDC can be accomplished in two ways; hierarchical-based and partition-based [18]–[20]. The hierarchical approach clusters similar documents that are either divisively or agglomerative. A divisive method is a top-down approach that starts by grouping a set of documents into a single cluster, which is then split into smaller and heterogeneous clusters. The agglomerative method, on the other hand, is a bottom-up approach that considers each document as a single cluster and then, homogenous clusters are grouped together to form a new cluster [21]. In contrast to the hierarchical, the partitioning approach decomposes a set of documents into groups of disjoint clusters based on their homogeneity using iterative processes.

The k-means and k-medoids algorithms are widely used partition-based clustering algorithms for dealing with unstructured text efficiently and have been used in recent studies such as in [22]–[24]. Both algorithms divide a set of documents into predefined clusters, with the average dissimilarity between documents within the cluster being minimal. The goal of the k-means algorithm is to minimize the total squared differences between each document and its cluster centroid, which is calculated as the mean value. In comparison, the k-medoids algorithm employs medoids as cluster centers, attempting to minimize the total number of dissimilarities between documents and their cluster centers. K-means and k-medoids algorithms, however, have received less attention for clustering Malay text documents, particularly in the crime domain [25]. These algorithms are chosen based on the type of data and the purpose of the research. Given that the purpose of the study is to identify crime reports with a common modus operandi pattern, k-means clustering and k-medoids are the optimal partitioning algorithms for clustering unstructured text. Due to the efficiency with which these clustering techniques organize and analyze data, they present promising opportunities in the criminal domain. Furthermore, both algorithms are capable of predicting crime by analyzing criminal patterns in textual data.

Due to their simplicity and ability to deal with unstructured text, the k-means and k-medoids algorithms remain as two of the most commonly used partition-based clustering algorithms. Therefore, prior studies on k-means and k-medoids algorithms have been classified according to their input data type, language, similarity measure, document representation, and the method of evaluation. Table 1 summarizes the related works on k-means, k-means fast, k-means++, and k-medoids algorithms that have been conducted in a variety of fields, including education, medical, news, manufacturing, finance, natural disasters, and crime.

The classical k-means algorithm has been applied to cluster structured data [26], and unstructured textual data as demonstrated in [27]–[30] where the initial clusters are randomly chosen. Meanwhile, several studies have compared the k-means and k-medoids algorithms for clustering structured data [31]–[33] and unstructured textual data [34], [35]. Studies on text document clustering using the classical k-means algorithm have shown a declining trend in recent years, particularly for processing English-based textual documents. Choosing the initial cluster numbers at random has a significant impact on the clustering time and performance [36]. The number of clusters must be proportional to the amount of data. Due to this, the k-means algorithm has been enhanced to speed up its execution time and improve clustering performance.

k-means fast and k-means++ are enhanced versions of k-means that have been explored and compared for clustering crime articles [37], abstracts [38], and news articles [39]. The k-means++ was introduced to aid in the identification of a suitable initial centroid point. Meanwhile, k-means fast is an accelerated version of k-means. Furthermore, Lakshmi and Baskar [40] proposed a dissimilarity-based initial centroids selection algorithm to improve k-means clustering performance, while [41] applied latent semantic indexing and pillar algorithm.

Table 1. Summary of related studies on k-means and k-medoids algorithms for text document clustering

Ref.	Domain	Algorithm				Data Type		Language	Similarity Measure				Document Representation		Evaluation		Best Result
		k.m	k.md	k.mf	k.m++	strc	unstrc		cos	cor	euc	jac	vsm	tf-idf	si	db	
[34]	education	/	/				/	Indonesian	/					/			k.md
[27]	education	/					/	Indonesian	/				/				k.m
[38]	education	/	/	/			/	English	/	/		/	/			/	k.m+k.md+cos
[35]	education	/	/				/	English			/						k.md
[28]	random	/	/				/	English			/			/			k.m
[37]	crime	/			/		/	English	/			/		/			k.m++ +cos
[26]	crime	/				/		Indonesian							/		k.m
[31]	medical	/	/			/		English			/						k.md
[42]	news	/					/	English					/				k.m +chi sq
[32]	finance	/	/			/		English			/					/	k.m
[40]	education	/					/	English	/					/			k.m +cos
[39]	news				/		/	English	/				/				k.m++
[43]	manufacturing	/					/	English	/					/			k.m +LSA
[41]	news	/					/	English	/					/			k.m +LSI
[29]	business	/					/	English	/	/	/			/			k.m +cos
[30]	news	/					/	Indonesian			/			/			k.m
[33]	natural disaster	/	/			/		Indonesian								/	k.m

Indicators:
*k.m=k-means *k.md=k-medoids *k.mf=k-means fast *k.m++=k-means++ *strc=structured *unstrc=unstructured *cos=cosine *cor=correlation coefficient *euc=euclidean distance *jac=jaccard *vsm=vector space model *si=silhouette index *db=Davies-Bouldin index *LSI= latent semantic indexing *LSA=latent semantic analysis

As shown in Table 1, k-means, k-means fast, k-means++, and k-medoids have demonstrated considerable potential for clustering unstructured textual data for a variety of purposes, including text summarization [28], [34], trend and pattern detection [27], [35], [38], [43] crime prediction [37], and opinion mining [29]. The unstructured text dataset used is collected from multiple resources, including both local and online repositories. The dataset stored in a local repository is referred to as closed domain data, whereas the online repository dataset is regarded as open domain data. Thesis reports and abstracts [27], [34], [38], local news [30], and customer reviews [29] are among the textual datasets of local repositories used in the studies. Meanwhile, several open data sources, such as purchase transactions [35], Reuters-8 and WebKB [40], BBC news [39], [42], Usenet articles [41], Bernama news [37], and websites [43] available in online repositories. However, less research has been done to explore the potential of k-means and k-medoids algorithms for clustering crime documents. Most of the research on the performance of k-means algorithms for clustering English documents makes use of publicly available data. On the other hand, several studies in the fields of crime, medicine, and finance make use of structured data [26], [31], [32].

In terms of language, numerous studies have been conducted on the k-means clustering method, which is used to cluster English documents. There has been an increase in the number of studies on TDC using the k-means [27], [30] and k-medoids [34] algorithms for processing textual data in Indonesian over the last few years. As far as we know, no research has been conducted on clustering Malay text documents using partitioning clustering algorithms such as k-means and k-medoids. Only a few clustering techniques have been studied recently for grouping Malay documents. These techniques include complete linkage clustering [44], latent semantic indexing [45], and fuzzy c-means clustering [46].

In text document clustering, k-means and k-medoids algorithms work by determining the correlation between documents via similarity measurement metrics. The metrics are the main components used by both algorithms to group similar documents together, while dissimilar documents are placed in different clusters. Similarity measures significantly influence k-means and k-medoids algorithms' performance. Various similarity measures have been proposed for document clustering, but most of the research has focused on cosine similarity [27], [29], [34], [37]–[41], [43] rather than Euclidean distance [28]–[32], [35]. Meanwhile, only a few studies use the correlation coefficient [29], [38] and Jaccard index [37], [38]. Some studies [40], [41] achieve the best results by utilizing cosine similarity tests. However, determining which similarity measure produces the best results across datasets is challenging due to the fact that each dataset clusters differently when considering the correlation between words appearing in the text.

Document representation is the crucial step before applying text document clustering algorithms. The quality of document representation affects the performance of text clustering algorithms. Each document is transformed into a bag of words. The bag-of-words method represents a document based on the frequency of its words, which is determined by the term frequency-inverse document frequency (TF-IDF) relationship and the vector space model (VSM). The representation model can be chosen between n-gram, unigram, or bigram. TF-IDF is a technique for counting terms across all documents [47], whereas VSM is a collection of vectors that may contain split term indexes. Due to its effectiveness in terms of frequency, most prior studies have used TF-IDF [28]–[30], [34], [37], [40], [41], [43]. It is more effective to implement in a lengthy document than in a short document, as long documents contain a greater number of terms. However, when it comes to predicting a group of documents, a weakness may occasionally occur. In addition, the variation in document sizes is another challenge that affects representation.

Cluster validity analysis is used to evaluate the performance of the text clustering algorithm. The silhouette index (SI) and the Davies-Bouldin index (DBI) are both frequently used evaluation metrics that take cohesion and separation factors into account [20]. SI validates the clustering performance by comparing the similarity of documents within a cluster to the dissimilarity of documents between clusters. DBI, on the other hand, calculates the average similarity between clusters. However, not all of the studies in Table 1 employ these metrics to assess their performance. Only one study used SI [26], while the other two used DBI [32], [38].

Overall, the results show that k-means clustering is the most commonly used method and produces the best results when combined with other methods. As discussed in the literature, however, the type and quantity of data used should be appropriate for the method chosen. Therefore, in this study, we explored and compared the performance of k-means and k-medoids algorithms using Euclidean distance for clustering Malay housebreaking crime reports. The performance of clustering algorithms depends on the distance similarity measurement, which can be quantified using a variety of measurement metrics such as distance-based similarity (Euclidean distance, Manhattan distance, Chebyshev distance, Minkowski distance, Jaccard distance), cosine similarity, and distribution-based similarity [20], [48]. The aim of this paper is to cluster crime reports with similar modus operandi patterns. In this study, five predefined modus operandi are used as references: *cara* (method), *peranan* (role), *keganjilan* (oddity), *senjata* (weapon), and *tempat* (location). The experimental result shows that k-means performed better than k-medoids in grouping crime reports with similar modus operandi patterns, with a 61.33% efficiency rate.

2. METHOD

This section presents a brief overview of the research method by providing a high-level summary of the main steps involved in the methods implemented. Figure 1 depicts the main framework of our research method. As shown, the framework consists of five phases: document collection, text preprocessing, document term weighting, text clustering, and analysis and evaluation of clustering. The Malay text preprocessor, a Java-based tool, is used to perform the text preprocessing steps. The tool was developed by employing Malay morphological rules, which is beyond this paper's discussion scope. The subsequent procedures of document term weighting and text clustering are carried out with the assistance of RapidMiner, a data analysis tool. RapidMiner is a versatile data mining tool as it can generate all the predictive modelling required to make a real impact in the field of data sciences.

The first step of the framework is to prepare the raw Malay text documents, which must be imported into RapidMiner. The RapidMiner tool supports a diverse range of document formats, including txt, xml, and csv. Due to the fact that the raw documents for this study are contained in an excel file, this file type must be successfully installed in RapidMiner. The relevant operators, such as tokenizing, transforming cases, filtering stop words, and stemming, are then used to perform text preprocessing. Following that, the clustering operator is applied to a dataset to build new clusters.

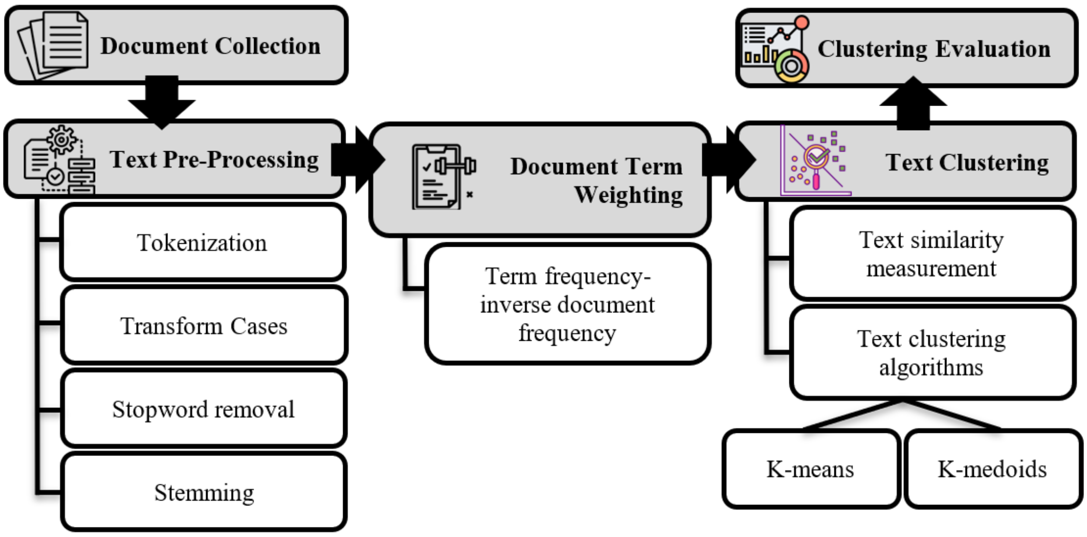

Figure 1. Research method framework

2.1. Document collection

The dataset for this study is a collection of housebreaking crime reports from 2010 to 2013. The dataset for this study is a collection of housebreaking crime reports from 2010 to 2013. High-quality research datasets are carefully chosen because previous research has demonstrated that low-quality datasets have an adverse effect on the results of machine learning, particularly on the accuracy of machine learning [47], [49]. It is a closed domain dataset obtained from the Malay Royal Police Department. There are 100,383 Malay crime reports in the corpus. For the experiment, 1,000 crime reports were randomly selected and filtered using five different modus operandi, including *cara* (method), *peranan* (role), *keganjilan* (oddity), *senjata* (weapon) and *tempat* (location). Each crime report has been manually identified its modus operandi by the crime investigator. The number of documents for each modus operandi category is distributed evenly and summarized in Table 2. The size of the documents varied between 1 and 145 words, with the total size of words is 11,524. Meanwhile, Table 3 shows an excerpt of the narrative from the reports. 72 documents are blank and are thus classified as a noise dataset. These documents are subsequently replaced with random non-empty crime reports.

Table 2. Summary of housebreaking crime dataset

Modus Operandi	Number of Document	Range Length (Words)	Total Words
Method (*cara*)	200	3-116	3,428
Role (*peranan*)	200	1-145	2,244
Oddity (*keganjilan*)	200	5-44	237
Weapon (*senjata*)	200	1-26	1,151
Location (*tempat*)	200	10-68	1,864

Table 3. Excerpt of narratives from the housebreaking crime reports

Modus Operandi	Narratives of crime report
Method	*Saspek telah mengumpil dinding kiosk yang boleh dibuka dan telah masuk ke dalam kiosk sebelum mengambil barangan milik kiosk berkenaan serta wang tunai hasil jualan*
Role	*4 lelaki cina tidak dikenali ketinggian 5 kaki 7 inci berbadan sederhaan dan seorang berbadan gemukbermuur awal 30-an hingga 40-an.*
Oddity	*Sebelum kejadian adik pengadu telah mempastikan pintu serta semua tingkap dikunci dengan baik.*
Weapon	*Penjenayah di percayai menggunakan besi pengumpil untuk mengumpil papan dinding.*
Location	*Rumah kediaman apartment taman perumahan*

2.2. Text preprocessing

Preprocessing task is an important part of text mining. It aims to clean and convert text to a machine-readable format, thus enhancing the efficiency of clustering algorithms. In this stage, text preprocessing of documents is performed to transform Malay documents into meaningful terms by applying tokenization, transforming cases, stopword removal, and stemming. These steps are integrated into a Java-based tool called the Malay text preprocessor.

Tokenization is the process of splitting a text into a single significant word or term, referred to as a token. This study employs a white space tokenization approach in which a collection of sentences is chunked into words whenever a white space is encountered. Then, the words are transformed into lowercase in order to prevent confusion caused by similar terms. Despite the fact that it contributes to output consistency, lowercasing addresses the sparsity issue by reducing the text dimension.

The following step is to eliminate the stopwords. Stopwords are a group of less significant words found in text documents. The most frequently occurring words in the text are usually derived from pronouns, prepositions, conjunctions, numbers, punctuation marks, or symbols. These terms lack meaning in documents and are not significant for clustering tasks. *di* (at), *ke* (to), *dengan* (with), *kemudian* (next), *walau bagaimanapun* (however), *tiada* (nothing), and *tidak* (no) are examples of Malay stopwords. There are 323 stopwords defined and used in this study.

The process of stemming is the discovery of the root word. It is concerned with the Malay language's morphology. The purpose of stemming is to break the word down into its constituent words. This is due to the fact that the Malay word pattern has a variation of affixation and derivation rules. For instance, the English terms 'connect', 'connected', and 'connection' are all derived from the root word 'connect'. In contrast, in Malay, the words *menyambung* (connect), *disambung* (is connected), and *sambungan* (connection) all originate from the word *sambung*. In this study, the Malay stemmer considers several different variations of morphological rules with the aid of a Malay dictionary, including single prefixes (*di, ke, se, me, ber, ter, men, mem, pen, pem, per, meng, peng*), single suffixes (*i, an, kan, lah, kah, nya*), double prefix (*mem+per*) and prefix-suffix pairs, i.e; *kehidupan* (life), *menghadapi* (facing), *sedikitnya* (little). However, a detailed discussion of the rules is outside the scope of this paper.

2.3. Document term weighting

Text documents are represented in high dimensionality space with a high degree of complexity. In order for the clustering algorithm to work, the textual documents had to be transformed into a numerical representation. The transformation is based on bag-of-words. The document term matrix was implemented in this study using the term frequency-inverse document frequency (TF-IDF) weighting scheme. This is the most frequently used technique for document representation in the vector space model and conversion of textual information to numerical format. The TF-IDF determines the value of each word in the document based on its frequency of occurrence. Each word or term has a unique value that indicates its frequency of occurrence in documents. The value ranges from 0 to 1. When a word has a value of 0, it indicates that it is unimportant. The critical and pertinent words have a value of 1. A word with a high TF-IDF value indicates a strong connection to that document.

The term frequency (TF) indicates the frequency of a word's occurrence in a document relative to the total number of words in the document. For calculating TF is used (1).

$$TF(w,d) = \frac{\text{word } w \text{ frequency in document } d}{\text{total number of words in document } d} \tag{1}$$

Meanwhile, in (2) depicts the inverse document frequency (IDF) for each word contained in the document to quantify the significance of a term within a collection of documents.

$$IDF(w,C) = \log\left(\frac{\text{total number of documents in the collection C}}{\text{number of documents containing word } w}\right) \tag{2}$$

To identify the TD-IDF score of a word in a document, in (3) is calculated. TD-IDF is used to weight words that are uncommon in a collection of documents, as well as words that frequently appear in the document.

$$TF - IDF(w,d,C) = TF(w,d) * IDF(w,C) \tag{3}$$

2.4. Text clustering

Two partitional clustering algorithms were used in this study: k-means and k-medoid. Both algorithms segment documents into a group of clusters entirely determined by users, with the initial set of *k* mean serving as the cluster centroids based on similarity. This study used the Euclidean distance to quantify similarity. The distance computes the root of square differences between a pair of documents. It is calculated by (4).

$$Distance\ d = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots} \tag{4}$$

2.4.1. K-means algorithm

The simplest unsupervised method is the k-means algorithm, which clusters documents by calculating the mean of the documents in the cluster. This calculation is repeated until no changes are detected in any of the clusters. The number of clusters must be specified, but not more than the number of documents. The distance between each data point and the centroid is calculated using a predefined distance measure. The shortest distance between the document and the centroid is the optimal distance. The cluster's centroid value is determined by the group's documents. The k-means algorithm is:

```
Input:
k: number of clusters (k=1,2…n)
D: data set of d data points
Output: set of k clusters (documents within cluster)
Begin
   Step 1 Choose the value of k clusters to assign
   Step 2 Randomly choose k data points as initial centroids $c_k$ for k cluster
   Step 3 Repeat
   Step 4 Assign each d data point to the nearest centroid based on the distance measure
   Step 5 Calculate the mean of the centroid for each cluster
   Step 6 Until convergence is reached or no changes in each cluster
End
```

2.4.2. K-medoids algorithm

K-medoids is a clustering algorithm similar to k-means, but it is less sensitive to outliers than k-means clustering. The fundamental concept of k-medoids is to identify a document in a cluster using a randomly generated *k* cluster. Each remaining document is clustered with the most closely related medoid. Instead of using the mean value of the documents in each cluster, the k-medoids algorithm makes use of representative documents as a reference data point. The algorithm for the k-medoids clustering is depicted:

```
Input:
k: number of clusters (k=1,2…n)
D: data set of d data points
Output: set of k clusters (documents within cluster)
Begin
   Step 1 Choose the value of k from d data points as the medoids
   Step 2 Repeat
   Step 3 Assign each remaining d data point to the nearest medoid (cluster) based on the
   distance measure
   Step 4    For each $m_k$ medoid point
   Step 5       For each d data point (non-medoid point)
   Step 6       Calculate the sum of dissimilarity distance from d data point to its
   nearest $m_k$ medoid
   Step 7       Swap d with $m_k$ if the computed sum of dissimilarity distance is minimal to
   form a new medoid
   Step 8 Until convergence is reached or no changes in each cluster
End
```

2.5. Clustering evaluation

Clustering evaluation is critical for quantifying the quality of generated clusters. Two approaches are used to evaluate the cluster quality; cluster accuracy and cluster validity analysis. Cluster validity is determined using the Davies Bouldin index. Meanwhile, cluster accuracy is assessed using accuracy, precision, recall, and the F-measure.

2.5.1. Cluster accuracy

In this experiment, precision, recall, and F-measure are used to evaluate the accuracy of clusters generated using the k-means and k-medoids algorithms. These parameters are essential to evaluating how well both algorithms cluster the relevant documents. The precision class contains the correct values, while the recall class contains the actual predicted values. The F-measure is the average of the precision and calibration values. The detailed equations discussed in this section are derived from [50].

a. Precision

Precision is the ratio of true positive documents, that is, correctly predicted documents clustered, to the total number of predicted documents clustered. It indicates the number of instances where the actual document is correctly grouped into the cluster. The higher the precision value, the more accurate the model. It is calculated:

$$Precision = \frac{\mathrm{TP}}{\mathrm{TP+FP}} \quad (5)$$

where TP denotes true positive, and FP is a false positive.

b. Recall

Recall is the percentage of predicted documents that are correctly grouped into a cluster. The value is determined by the number of relevant predicted documents clustered and the total number of actual positive documents. The recall values are calculated:

$$Recall = \frac{\text{TP}}{\text{TP+FN}} \tag{6}$$

where TP denotes, true positive, and FN represents false negative.

c. F-measure

The F-measure is derived from the precision and recall values. It is difficult to compare a high recall value with a low precision value, or vice versa. As a result, the f-measure would assist in balancing the measurement between precision and recall value. The equation (7) denotes the calculation of the F-measure.

$$f - measure = \left(\frac{2*recall*precision}{recall+precision}\right) \tag{7}$$

d. Accuracy

Accuracy is a straightforward metric for determining the proportion of clustered documents to the total number of actual and predicted documents. The equation (8) is used to calculate the accuracy measure:

$$Accuracy = \frac{\text{TP+TN}}{\text{TP+FP+FN+TN}} \tag{8}$$

where TP denotes true positive, FP signifies false positive, TN is true negative, and FN represents false negative.

2.5.2. Cluster validity

Cluster validity analysis is used to evaluate the k-means and k-medoids algorithms. It aims to determine the significance of the disclosed cluster structure created. The Davies-Bouldin index (DBI) is used in this experiment to determine the quality of the clustering performed on the basis of the dataset's quantities and features.

3. RESULTS AND DISCUSSION

3.1. Text representation

TF-IDF transforms text into numerical values. Table 4 contains an excerpt of term frequency values generated in the experiment. The total occurrences of each word in a document are computed. A term that appears multiple times in a document is considered significant. After that, the TF-IDF is calculated for each occurrence term. The weighting of terms determines the importance of documents. Table 5 illustrates an example of a document term matrix after applying the TF-IDF function.

The word cloud facilitates analysis by separating the text from the dataset into tokens of words. The size of the font in a word cloud is determined by the frequency with which those words appear in the text. If a word frequently appears in the text, the font size in the word cloud will be larger. The smaller font size in the word cloud indicates that the word has a low vectorization weight. Figure 2 depicts the most frequently occurring words in the dataset used in this study. It demonstrates that the most frequently occurring word in crime reports is *pintu* (door), as the majority of housebreaking crimes involve the effect of a pryed door. Another word that frequently appears in the reports is *umpil* (lever), *tingkap* (window), and grill.

Table 4. Partial view of the documents term matrix

# of document	*abdullah*	*abu*	*acu*	*ada*	*adik*	*agama*
Document 1	0	0	0	0	0	.3
Document 2	0	0	0	0	7	0
Document 3	0	0	0	0	4	0
Document 4	0	0	0	1	0	0
Document 5	0	0	2	0	0	0
Document 6	0	5	0	0	0	0
Document 7	4	0	0	0	0	0
Document 8	3	0	0	0	0	0

Table 5. Term frequency-inverse document frequency values

	abdullah	*abu*	*acu*	*ada*	*adik*	*agama*
Document 1	0	0	0	0	0	0.325
Document 2	0	0	0	0	0.683	0
Document 3	0	0	0	0	0.362	0
Document 4	0	0	0	1	0	0
Document 5	0	0	0.176	0	0	0
Document 6	0	0.497	0	0	0	0
Document 7	0.424	0	0	0	0	0
Document 8	0.344	0	0	0	0	0

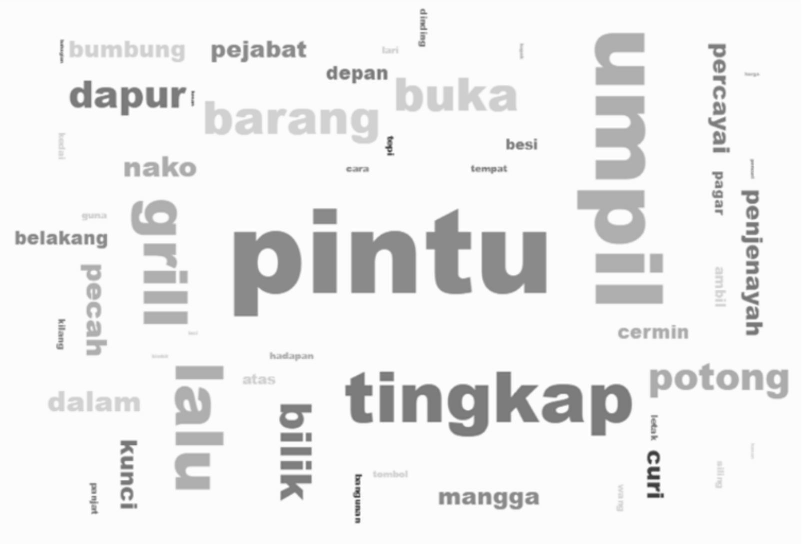

Figure 2. Word cloud for the highest frequencies of the word

3.2. K-means and k-medoids clustering results

The value of *k* cluster is set to five for the clustering experiments because the collection of documents falls into five pre-classified modus operandi, namely method (*cara*), oddity (*keganjilan*), role (*peranan*), weapon (*senjata*), and location (*tempat*). The prior classification has been done by the domain expert of crime investigation. The 1,000 documents are classified into five clusters using k-means and k-medoids clustering algorithms, with each document being assigned its TD-IDF value. The result of the cluster analysis for both k-means and k-medoids algorithms on the dataset is shown in Table 6. As indicated in the table, cluster_4 contains the most relevant documents clustered using the K-means algorithm, with 163 documents. In the meantime, the clusters with the fewest relevant documents obtained are cluster_2 and cluster_5, each with only 47 documents. Meanwhile, for the k-medoids algorithm, cluster_3 produces the most relevant documents, with 156 documents, while cluster_5 produces the least relevant documents with only 23 documents.

3.3. Clustering evaluation

3.3.1. Results for cluster accuracy

For the evaluation of cluster accuracy produced by k-means and k-medoids algorithm, precision, recall, F-measure and accuracy matrix are applied. The cluster results obtained are compared to the actual classification of documents as done by the domain expert. Table 7 summarizes the evaluation results of precision, recall, F-measure and accuracy for both algorithms.

The highest precision value for the k-means algorithm is 96% for cluster_3 (weapon). Meanwhile, cluster_1 (oddity) produces the lowest precision value at a rate of 26%. Nonetheless, cluster_4 (method) has the highest recall value at 82%, while cluster_2 (role) and cluster_5 (location) has the lowest recall values at only 24%. For the k-medoids algorithm, cluster_5 (role) achieves a maximum precision rate of 100%, and the minimum score of precision is from cluster_1 (method) at 24%. In the meantime, the highest recall rate is scored by cluster_3 (oddity) with 78%, and cluster_5 (role) produces the lowest rate of recall with only 12%. As we can see in the table, for both algorithms, the scores of precisions and recall for every cluster are unbalanced. For instance, cluster_5 (role), as generated by the k-medoids algorithm, has the highest percentage of precision (100%) but the lowest recall rate (12%). This is because a precision score of 100% for cluster_5 indicates that all of the 23 documents contained within the cluster have been correctly grouped as actual labelled documents. The precision rate, on the other hand, does not take into account the total number of documents that do not cluster correctly. The lowest recall rate obtained from cluster_5 is due to the fact that only 23 documents were assigned to the correct cluster out of 200 actual clustered documents.

For the k-means algorithm, cluster_4 (method) produces the highest f-measure score with 65%, while cluster_5 (location) produces the lowest score, around 24%. In terms of the k-medoids algorithm, the highest f-measure rate comes from cluster_3 (oddity) with 53%, while the lowest rate comes from cluster_5 (role) with 21%. A high f-score indicates that the cluster contains a small number of irrelevant documents. Meanwhile, the k-means algorithm obtains the highest accuracy in cluster_3 (weapon) with 86%. In the k-medoids algorithm, cluster_4 (location) achieves 87% accuracy.

Table 6. Clusters produced from the k-means and k-medoids algorithms

Clustering Algorithm	Cluster	Modus Operandi	Total # of Documents Clustered	# of Relevant Documents Clustered	# of Irrelevant Documents Clustered
K-means	cluster_1	Oddity	480	123	357
	cluster_2	Role	72	47	25
	cluster_3	Weapon	69	66	3
	cluster_4	Method	299	163	136
	cluster_5	Location	80	47	33
		Total	1000	446	554
K-medoids	cluster_1	Method	230	55	175
	cluster_2	Weapon	286	100	186
	cluster_3	Oddity	385	156	229
	cluster_4	Location	76	72	4
	cluster_5	Role	23	23	0
		Total	1000	406	594

Table 7. Precision, recall and f-measures of cluster accuracy

Algorithms	Cluster #	Modus Operandi	Precision	Recall	f-Measure	Accuracy
K-means	cluster_1	Oddity	26%	62%	36%	57%
	cluster_2	Role	65%	24%	35%	82%
	cluster_3	Weapon	96%	33%	49%	86%
	cluster_4	Method	55%	82%	65%	83%
	cluster_5	Location	59%	24%	34%	81%
	Average		60%	45%	44%	78%
K-medoids	cluster_1	Method	24%	28%	26%	68%
	cluster_2	Weapon	35%	50%	41%	71%
	cluster_3	Oddity	41%	78%	53%	73%
	cluster_4	Location	95%	36%	52%	87%
	cluster_5	Role	100%	12%	21%	82%
	Average		59%	41%	39%	76%

On average, k-means achieves a precision of 60%, slightly higher than k-medoids, which obtains a precision of 59%. Simultaneously, k-means has a higher recall rate of 45% compared to k-medoids, which has a recall rate of only 41%. The same is true for the average of the f-measure and the accuracy. The k-means algorithm achieves the highest scores for both measurements, 44% and 78%, respectively. This shows that k-means algorithms are superior to k-medoids at grouping documents into the correct clusters. In addition, each cluster generated by the k-means algorithm has higher sensitivity than k-medoids to the total number of correctly clustered documents.

3.3.2. Cluster validity

The performance vectors for the k-means and k-medoids algorithms are shown in Table 8. The DBI is used to evaluate the performance of clusters. DBI is a technique for validating the internal cluster. The minimum DBI value indicates the highest intra-cluster similarity between the documents in the cluster, whereas the maximum DBI value indicates the lowest intra-cluster similarity between the documents in the cluster. The average distance from the centroid is used to determine the location of observation within a cluster. The centroid value indicates the location of the element’s center in the cluster to which it has been assigned. The centroid distance with the smallest value has a high probability of being a group in that cluster. According to the Table 8, the k-means algorithm outperforms the k-medoids algorithm in terms of average centroid within cluster distance by -0.903 value. However, it is surprising to discover that k-medoids outperform k-means by obtaining -1.814 of the DBI.

Table 8. Performance vector of k-means and k-medoids algorithm

Algorithm	Cluster	Performance Vector	Average centroid	Davies Bouldin
K-means	cluster_1	-0.978	-0.903	-4.844
	cluster_2	-0.685		
	cluster_3	-0.715		
	cluster_4	-0.896		
	cluster_5	-0.836		
K-medoids	cluster_1	-1.826	-1.834	-1.814
	cluster_2	-1.956		
	cluster_3	-1.851		
	cluster_4	-1.859		
	cluster_5	0		

4. CONCLUSION

This study aims to perform a comparative analysis between the k-means and k-medoids algorithms to locally cluster Malay text documents. Several experiments were conducted using the RapidMiner tool, while text preprocessing was performed using the Malay text preprocessor, a Java-based application that incorporates Fatimah's Malay morphological rules. According to the findings, the accuracy of clusters is significantly moderate in terms of precision, recall, and f-measure for the k-means algorithm. Meanwhile, for cluster validity, it demonstrates that k-means outperforms the k-medoids algorithm in terms of average centroid within the cluster. In conclusion, the cluster quality obtained from both k-means and k-medoid clustering experiments is significantly moderate. Overall, the k-means and k-medoids algorithms work well for Malay unstructured data but have a few flaws. The data points for the k-means are chosen at random so that the documents produced can be grouped differently. Because of the low value of the data point, the grouping result cannot be optimal. Furthermore, the cluster's performance is influenced by a variety of factors. K-medoids are less effective on large datasets and take less time to cluster the dataset than k-means. This study produces satisfactory results in text document clustering, but there are some challenges in analyzing Malay documents, such as filtering out noise in the documents, such as typo errors, and incomplete sentence structure. Furthermore, because Malay is rich in morphological rules and word patterns, the accuracy of stopword removal and stemming could have a significant impact on clustering results. In the future, textual documents will be morphologically stemmed using a linguistic approach before text clustering is carried out. Further experiments are expected to be conducted in the future using variants of the k-means algorithm, including k-means++, k-means fast, and k-means kernel.

ACKNOWLEDGEMENTS

This research is supported by the Ministry of Higher Education Malaysia, under the Fundamental Research Grant Scheme (Reference Code: FRGS/1/2018/ICT04/UMT/02/3). The authors would also like to thank the Royal Police Department Malaysia for their unwavering support for this study.

REFERENCES

[1] V. Kalra and R. Aggarwal, "Importance of text data preprocessing and implementation in RapidMiner," in *Proc. First Int. Conference on Information Technology and Knowledge Management*, 2018, vol. 14, pp. 71–75, doi: 10.15439/2017km46.

[2] G. Rod and P. Darryl, "Introduction to criminal investigation: processes, practices and thinking-open textbook," First Edit. BCcampus, pp. 122–136, 2017.

[3] H. Yu and N. Monas, "Recreating the scene: an investigation of police report writing," *Journal of Technical Writing and Communication*, vol. 50, no. 1, pp. 35–55, Jan. 2020, doi: 10.1177/0047281618812441.

[4] N. Qazi and B. L. W. Wong, "An interactive human centered data science approach towards crime pattern analysis," *Information Processing and Management*, vol. 56, no. 6, 2019, doi: 10.1016/j.ipm.2019.102066.

[5] Y.-S. Li and M.-L. Qi, "An approach for understanding offender modus operandi to detect serial robbery crimes," *Journal of Computational Science*, vol. 36, Sep. 2019, doi: 10.1016/j.jocs.2019.101024.

[6] J. Agarwal, R. Nagpal, and R. Sehgal, "Crime analysis using k-means clustering," *International Journal of Computer Applications*, vol. 83, no. 4, pp. 1–4, Dec. 2013, doi: 10.5120/14433-2579.

[7] M. Jangra and S. Kalsi, "Naïve Bayes approach for the crime prediction in data mining," *International Journal of Computer Applications*, vol. 178, no. 14, pp. 33–37, May 2019, doi: 10.5120/ijca2019918907.

[8] Q. Wang, G. Jin, X. Zhao, Y. Feng, and J. Huang, "CSAN: A neural network benchmark model for crime forecasting in spatio-temporal scale," *Knowledge-Based Systems*, vol. 189, Feb. 2020, doi: 10.1016/j.knosys.2019.105120.

[9] M. Feng *et al.*, "Big data analytics and mining for effective visualization and trends forecasting of crime data," *IEEE Access*, vol. 7, pp. 106111–106123, 2019, doi: 10.1109/ACCESS.2019.2930410.

[10] P. Yerpude and V. Gudur, "Predictive modelling of crime dataset using data mining," *International Journal of Data Mining and Knowledge Management Process*, vol. 7, no. 4, pp. 43–58, 2017, doi: 10.5121/ijdkp.2017.7404.

[11] A. Joshi, A. S. Sabitha, and T. Choudhury, "Crime analysis using k-means clustering," in *2017 3rd International Conference on Computational Intelligence and Networks (CINE)*, Oct. 2017, pp. 33–39, doi: 10.1109/CINE.2017.23.

[12] M. Alruily, A. Ayesh, and A. Al-Marghilani, "Using self organizing map to cluster Arabic crime documents," in *Proceedings of*

the International Multiconference on Computer Science and Information Technology, Oct. 2010, vol. 5, pp. 357–363, doi: 10.1109/IMCSIT.2010.5679616.

[13] G. Matto and J. Mwangoka, "Detecting crime patterns from swahili newspapers using text mining," *International Journal of Knowledge Engineering and Data Mining*, vol. 4, no. 1, 2017, doi: 10.1504/ijkedm.2017.10006068.

[14] R. N. Zaeem, M. Manoharan, Y. Yang, and K. S. Barber, "Modeling and analysis of identity threat behaviors through text mining of identity theft stories," *Computers and Security*, vol. 65, pp. 50–63, Mar. 2017, doi: 10.1016/j.cose.2016.11.002.

[15] T. Siddiqui, A. Y. A. Amer, and N. A. Khan, "Criminal activity detection in social network by text mining: comprehensive analysis," in *2019 4th International Conference on Information Systems and Computer Networks (ISCON)*, Nov. 2019, pp. 224–229, doi: 10.1109/ISCON47742.2019.9036157.

[16] S. Mukherjee and K. Sarkar, "Analyzing large news corpus using text mining techniques for recognizing high crime prone areas," in *2020 IEEE Calcutta Conference (CALCON)*, Feb. 2020, pp. 444–450, doi: 10.1109/CALCON49167.2020.9106554.

[17] H. Kaur, T. Choudhury, T. P. Singh, and M. Shamoon, "Crime analysis using text mining," in *2019 International Conference on contemporary Computing and Informatics (IC3I)*, Dec. 2019, pp. 283–288, doi: 10.1109/IC3I46837.2019.9055606.

[18] V. Renganathan, "Text mining in biomedical domain with emphasis on document clustering," *Healthcare Informatics Research*, vol. 23, no. 3, Jul. 2017, doi: 10.4258/hir.2017.23.3.141.

[19] A. K. Abasi, A. T. Khader, M. A. Al-Betar, S. Naim, Z. A. A. Alyasseri, and S. N. Makhadmeh, "A novel hybrid multi-verse optimizer with K-means for text documents clustering," *Neural Computing and Applications*, vol. 32, no. 23, pp. 17703–17729, Dec. 2020, doi: 10.1007/s00521-020-04945-0.

[20] C. Zong, R. Xia, and J. Zhang, "Text clustering," in *Text Data Mining*, Singapore: Springer Singapore, 2021, pp. 125–144.

[21] L. C. Leong and R. Alfred, "Optimizing terms reduction process for bilingual clustering of Malay-English Corpora," in *Lecture Notes in Electrical Engineering*, vol. 520, 2019, pp. 279–287.

[22] C. Oktarina, K. A. Notodiputro, and I. Indahwati, "Comparison of k-means clustering method and k-medoids on twitter data," *Indonesian Journal of Statistics and Its Applications*, vol. 4, no. 1, pp. 189–202, Feb. 2020, doi: 10.29244/ijsa.v4i1.599.

[23] S. A. Curiskis, B. Drake, T. R. Osborn, and P. J. Kennedy, "An evaluation of document clustering and topic modelling in two online social networks: Twitter and Reddit," *Information Processing and Management*, vol. 57, no. 2, 102034, Mar. 2020, doi: 10.1016/j.ipm.2019.04.002.

[24] M. Afzali and S. Kumar, "Text document clustering: issues and challenges," in *International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon)*, Feb. 2019, pp. 263–268, doi: 10.1109/COMITCon.2019.8862247.

[25] J. C. Ling Lee, P. Lee Teh, S. Lun Lau, and I. Pak, "Compilation of malay criminological terms from online news," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 15, no. 1, Jul. 2019, doi: 10.11591/ijeecs.v15.i1.pp355-364.

[26] R. T. Vulandari, W. L. Y. Saptomo, and D. W. Aditama, "Application of k-means clustering in mapping of Central Java crime area," *Indonesian Journal of Applied Statistics*, vol. 3, no. 1, Jul. 2020, doi: 10.13057/ijas.v3i1.40984.

[27] M. F. Riyadhi, "Text mining application for automation of determining thesis topic trends using the k-means clustering method (case study: computer systems study program) (In Indonesian)," *Jurnal Sistem Komputer*, vol. 2, no. 1, pp. 1–6, 2019.

[28] R. C. Balabantaray, C. Sarma, and M. Jha, "Document clustering using k-medoids," *International Journal of Knowledge Based Computer System*, vol. 1, no. 1, pp. 7–13, 2013.

[29] J. Žižka, K. Burda, and F. Dařena, "Clustering a very large number of textual unstructured customers' reviews in English," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2012, pp. 38–47.

[30] N. G. Yudiarta, M. Sudarma, and W. G. Ariastina, "Application of the clustering text mining method for grouping news on unstructured textual data (in Indonesian)," *Majalah Ilmiah Teknologi Elektro*, vol. 17, no. 3, Dec. 2018, doi: 10.24843/MITE.2018.v17i03.P06.

[31] S. A. Abbas, A. Aslam, A. U. Rehman, W. A. Abbasi, S. Arif, and S. Z. H. Kazmi, "K-means and k-medoids: cluster analysis on birth data collected in City Muzaffarabad, Kashmir," *IEEE Access*, vol. 8, pp. 151847–151855, 2020, doi: 10.1109/ACCESS.2020.3014021.

[32] M. Aryuni, E. D. Madyatmadja, and E. Miranda, "Customer segmentation in XYZ bank using k-means and k-medoids clustering," in *2018 International Conference on Information Management and Technology (ICIMTech)*, Sep. 2018, pp. 412–416, doi: 10.1109/ICIMTech.2018.8528086.

[33] M. Herviany, S. P. Delima, T. Nurhidayah, and Kasini, "Comparison of k-means and k-medoids algorithms for grouping landslide prone areas in West Java Province (in Indonesian)," *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, vol. 1, no. 1, pp. 34–40, 2021.

[34] K. K. Purnamasari, "K-means and k-medoids for indonesian text summarization," *IOP Conference Series: Materials Science and Engineering*, vol. 662, no. 6, Nov. 2019, doi: 10.1088/1757-899X/662/6/062013.

[35] P. Arora, Deepali, and S. Varshney, "Analysis of k-means and k-medoids algorithm for big data," *Procedia Computer Science*, vol. 78, pp. 507–512, 2016, doi: 10.1016/j.procs.2016.02.095.

[36] S. Mousavi, F. Z. Boroujeni, and S. Aryanmehr, "Improving customer clustering by optimal selection of cluster centroids in K-means and K-medoids algorithms," *J. of Theoretical and Applied Information Technology*, vol. 8, no. 10, pp. 3807–3814, 2020.

[37] B. Aubaidan, M. Mohd and M. Albared, "Comparative study of k-means and k-means++ clustering algorithms on crime domain," *Journal of Computer Science*, vol. 10, no. 7, pp. 1197–1206, Jul. 2014, doi: 10.3844/jcssp.2014.1197.1206.

[38] S. Al-Anazi, H. AlMahmoud, and I. Al-Turaiki, "Finding similar documents using different clustering techniques," *Procedia Computer Science*, vol. 82, pp. 28–34, 2016, doi: 10.1016/j.procs.2016.04.005.

[39] W. Arif and N. A. Mahoto, "Document clustering-a feasible demonstration with k-means algorithm," in *2nd Int. Conference on Computing, Mathematics and Engineering Technologies (iCoMET)*, 2019, pp. 1–6, doi: 10.1109/ICOMET.2019.8673480.

[40] R. Lakshmi and S. Baskar, "DIC-DOC- K-means: dissimilarity-based initial centroid selection for document clustering using K-means for improving the effectiveness of text document clustering," *Journal of Information Science*, vol. 45, no. 6, pp. 818–832, Dec. 2019, doi: 10.1177/0165551518816302.

[41] S. Adinugroho, Y. A. Sari, M. A. Fauzi, and P. P. Adikara, "Optimizing K-means text document clustering using latent semantic indexing and pillar algorithm," in *2017 5th International Symposium on Computational and Business Intelligence (ISCBI)*, Aug. 2017, pp. 81–85, doi: 10.1109/ISCBI.2017.8053549.

[42] A. I. Kadhim and A. K. Jassim, "Combined chi-square with k-means for document clustering," *IOP Conference Series: Materials Science and Engineering*, vol. 1076, no. 1, Feb. 2021, doi: 10.1088/1757-899X/1076/1/012044.

[43] R. Sabbagh and F. Ameri, "A framework based on k-means clustering and topic modeling for analyzing unstructured manufacturing capability data," *Journal of Computing and Information Science in Engineering*, vol. 20, no. 1, Feb. 2020, doi: 10.1115/1.4044506.

[44] N. A. Rahman, Z. A. Bakar, and N. S. S. Zulkefli, "Malay document clustering using complete linkage clustering technique with

cosine coefficient," in *2015 IEEE Conference on Open Systems (ICOS)*, 2015, pp. 103–107, doi: 10.1109/ICOS.2015.7377286.

[45] N. A. Samat, M. Azrifah, A. Murad, M. T. Abdullah, R. Atan, and I. Technology, "Malay documents clustering algorithm based," *Information Retrieval*, 2009.

[46] M. S. Salleh, S. A. Asmai, H. Basiron, and S. Ahmad, "Named entity recognition using fuzzy c-means clustering method for Malay textual data analysis," *J. Telecommunication, Electronic and Computer Engineering*, vol. 10, no. 2–7, pp. 121–126, 2018.

[47] T. Botsis, G. Hartvigsen, F. Chen, and C. Weng, "Secondary use of EHR: data quality issues and informatics opportunities.," *Summit on translational bioinformatics*, vol. 2010, pp. 1–5, Mar. 2010.

[48] A. S. Ramkumar and R. Nethravathy, "Text document clustering using k-means algorithm," *International Research Journal of Engineering and Technology*, vol. 6, 2008, [Online]. Available: www.irjet.net.

[49] W. Dai and D. Berleant, "Benchmarking deep learning classifiers: beyond accuracy," *arXiv preprint arXiv:2103.03102*, 2021.

[50] T. Zeugmann *et al.*, "Precision and recall," in *Encyclopedia of Machine Learning*, C. Sammut and G. I. Webb, Eds. Boston, MA: Springer US, 2011.

BIOGRAPHIES OF AUTHORS

Rosmayati Mohemad is currently an associate professor at the Faculty of Ocean Engineering Technology and Informatics, Universiti Malaysia Terengganu. She received her Ph.D. degree in Computer Science from Universiti Kebangsaan Malaysia in 2013. Her research interest is in information system and knowledge engineering, specializing in decision support systems, ontology modeling, and text mining. This also includes research contributions in the design, development, and evaluation of intelligent decision support systems for analyzing and improving decision-making processes in various domains such as tender management, forensics, crime investigation, and education. To date, she has produced more than forty peer-reviewed journal articles. She also has authored and co-authored 3 books and 4 chapters in book. Other than that, she has presented and published more than 25 conference proceedings. She is an editor of books, conference proceedings, and reviewer of international journals and established conferences. She is a member of the Malaysia Board of Technologists. She can be contacted at email: rosmayati@umt.edu.my.

Nazratul Naziah Mohd Muhait received her bachelor's degree in Information Management System from the Universiti of Technology Mara in 2014. Then she obtained her Master of Business Administration from Universiti Sultan Zainal Abidin in 2018. Currently, she is a graduate research assistant at the Faculty of Ocean Engineering Technology and Informatics, Universiti Malaysia Terengganu. Her Ph.D. research is in the area of text analytics. She can be contacted at email: nazz1799@gmail.com.

Noor Maizura Mohamad Noor is a professor at the Faculty of Ocean Engineering Technology and Informatics, Universiti Malaysia Terengganu. She obtained her Diploma and Bachelor of Computer Science from Universiti Pertanian Malaysia in 1991 and 1994 respectively. She received her Master of Science (Computer Science) from Universiti Putra Malaysia in 1997. She later acquired her doctoral degree in Computer Science from The University of Manchester in 2005. Her recent research work focuses on improving organizational decision-making practices using technologies. This includes research interest in the design, development, and evaluation of decision support systems for analyzing and improving decision processes. Her research interest also focuses on the areas of computer science, intelligent decision support systems, clinical decision support systems, and information system as well as action research in education. She has presented and published over two hundred of research papers on the decision support system at various international and local refereed journals, conferences, seminars, and symposiums. She can be contacted at email: maizura@umt.edu.my.

Zulaiha Ali Othman is currently an associate professor at Center for Artificial Intelligence Technology, Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia. She received her first degree in Computer Science from Universiti Kebangsaan Malaysia (UKM) in 1990, her master's degree in software technology from University of Sheffield, United Kingdom in 1997, and Ph.D. degree in Computing (Agent Oriented Methodology) from Sheffield Hallam University, in 2003. Her research interests include artificial intelligent, agent technology and data mining. She can be contacted at email: zao@ukm.edu.my.