Manuscript Submitted	27.09.2021
Accepted	28.12.2021
Published	31.12.2021

The impact of N-gram on the Malay text document clustering.

Rosmayati Mohemad[1], Nazratul Naziah Mohd Muhait[1], Noor Maizura Mohamad Noor[1] & Zulaiha Ali Othman[2]

[1]Faculty of Ocean Engineering Technology & Informatics, Universiti Malaysia Terengganu, 21030 Kuala Nerus, Terengganu, Malaysia.

[2]Faculty of Information Science & Technology, Universiti Kebangsaan Malaysia, 43600 UKM Bangi, Selangor, MALAYSIA.

rosmayati@umt.edu.my

Rosmayati Mohemad

[1]Faculty of Ocean Engineering Technology & Informatics, Universiti Malaysia Terengganu, 21030 Kuala Nerus, Terengganu, Malaysia.

rosmayati@umt.edu.my

Abstract

Document preprocessing is one of the crucial elements in text mining framework to provide a high-quality model for machine learning applications. The process including tokenizing, transform cases, stop word removal, and stemming. However, these sub-processes are not enough to optimize the clustering performance. Thus text preprocessing has to be improved by using N-gram features. N-gram is a sequence of words generate from a text document. Therefore, this study aims to evaluate the impact of using different N-gram models in text preprocessing. There are 1000 of the Malay documents were tested using N-gram on the K-means clustering algorithm. In addition, the document without N-gram is compared with the document that applies 2-gram,3-gram, and 4-gram. The result of text document clustering using 4-gram shows the highest accuracy with 92.48% compared to the text document clustering without using N-gram, which is 87.32%. The accuracy of the result indicates that applying N-gram in the Malay document clustering using K-means clustering algorithm could increase the cluster performance.

***Keywords*:** *N-Gram, document clustering, Malay documents, k-means.*

1. Introduction

Law enforcement personnel's ability to conduct investigations and prevent crime is dependent on their ability to acquire the data quickly and effectively. Currently, the police report contains various hidden information of the crime events, such as time, place, modus operandi, etc. These types of documents are commonly stored as simple text files in databases. However, these documents are difficult and take time to analyze all the documents that can be associated with the same modus operandi(MO). MO is the method of operation connected with a particular criminal event[1]. The crime analysis is very useful for law enforcement administration to take further precautions to formulate strategies to reduce the crime rate. Furthermore, the research also can be utilized both in investigative work and in court to provide information to criminal justice practitioners, and it can be revealed to the public. Thus, document clustering is one of the methods that might be used to deal with this challenge in the future.

A typical text document clustering framework contains text preprocessing phase, document representation, selection of clustering algorithm, and evaluation. The text preprocessing phase is the crucial phase that consists of a few tasks such as tokenization, stop-word removal, transform cases,

and stemming. Many improvements can be made in the text preprocessing phase, including the n-gram feature. However, the text preprocessing task should depend on the data types. For example, the Twitter data might have different tasks in text processing.

Text document clustering (TDC) is a text mining technique often used to group a collection of similar documents through the most relevant categories based on similarity or heterogeneity attributes. There are numerous of study that is related to text document clustering [2][3][4]. However, all these studies are not using n-gram model for their text representation. They just used the standard preprocessing for their document since their language is easy to handle. One of the studies that apply n-grams for the clustering approach is [5]. The author introduced text document clustering techniques using N-grams and 'improved sqrt-cosine similarity measure'.

Other than clustering, many classification approaches use the n-gram method to reduce the data representation. Similar work was also carried out by [6] using a Naïve Bayes algorithm to investigate N-gram's effect in document classification. However, the study results are very different from this study because the results obtained are decreased when using n-grams due to the system itself maybe it will be more accurate when classifying using the word rather than n-gram character. Sometimes the difference in results is also influenced by the algorithm itself. The same goes with another study by [7], the author has come out with the unsatisfied result when their experiment using bigram and trigram has no impact on their outcome. These studies aim to categorize the Turkish newspaper article based on the author by implementing Naïve Bayes, support vector machine, and random forest algorithm. The corpus consisted of 300 articles that have been collected from various newspapers. As a result, these studies retrieved a lower result for word-level and a high result for character-level. K-means clustering algorithm is one of the algorithms used in text mining. The K-Means technique is a non-hierarchical clustering method that is simple, fast, and flexible to data distribution [8][9][10]. This type of algorithm is commonly used to cluster unlabelled documents. Before choosing the algorithm, a prior study was done to identify whether this type of algorithm can conduct unstructured documents. Due to the efficient approach for text preprocessing, we discover new techniques for Malay text documents clustering by optimizing the accuracy performance using the n-gram method.

This study presented the current approach for clustering the Malay text documents by using N-grams document representation with cosine similarity measurement. An N-gram is a collection of a document's series of sequence characters. RapidMiner tools have been used in this study to validate the document clustering accuracy on the 1000 Malay housebreaking crime documents. The accuracy of the proposed clustering approach is evaluated for various n-gram values. The best result obtained using the N-grams is compared to the best result obtained using the baseline method.

The rest of the paper is organized as follows: Section 2 describes the proposed approach for this study. Section 3 outlines the experimental result and discussion, while section 4 details this study's conclusion and future work.

2. Proposed approach

One of the text mining approaches for unsupervised learning is clustering. Clustering is commonly used to cluster unlabelled documents into one cluster with the same characteristic or topic.

The documents are automatically sorted into categories when using these methods based on the number of clusters k. This document clustering enables the quick assessment of modus operandi classes. This method can overcome the problems and challenges faced by officers when analyzing the documents manually.

The text documents clustering is obtained by using the K-means clustering algorithm, one of the machine learning approaches for unsupervised learning. The set of unlabelled documents used for this study contains the five classes of MO: method, role, oddity, weapon, and location. To implement the algorithm, the following steps are required: randomly selecting the Malay text document, text preprocessing, N-gram document representation, and vector space model of the Malay document. Finally, k-means is applied with the 5 clusters by using cosine similarity measures on the different N-gram of the document representation. Finally is to evaluate the performance accuracy. The approach depicted in Figure 1 illustrates all of the phases in this study.

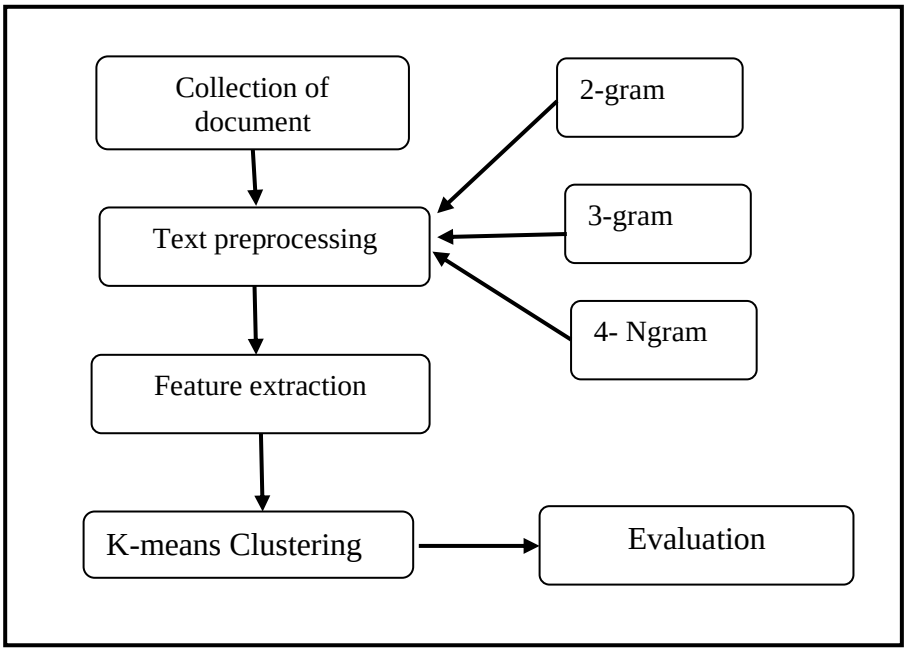

Figure 1 : Framework of the study.

The process begins by randomly extracting 1000 collections of documents and going through the filtering process in text processing. The housebreaking dataset used in this study is from 2010 to 2013. There are five classes need to be classify from data collection including cara (method), peranan (role), keganjilan (oddity), senjata (weapon), and tempat (location). The modus operandi classes of each crime report have been manually identified first by the crime investigator to evaluate the text document clustering performance.

The next task is the crucial task which is text preprocessing. This process needs to convert an unstructured data format into a structured data format before any other clustering algorithm can be performed. Preprocessing procedures have the most significant impact on the performance of machine learning algorithms. A key element of many machine learning approaches is a strategy that helps the system work better and efficiently. Some of the basic text processing include tokenization, stop word removal, and stemming, which are typically used in conjunction with the conventional "bag of words" techniques of document representation. This research uses a different text preprocessing approach by employing various N-grams (bi-gram, tri-gram,quad-gram). N-gram feature has a unique preprocessing method that improves transforming any character that is not a letter or multiple spaces into a single space that may include an error. Bi-gram refers to the 2-gram of terms, and tri-gram is for 3-grams of terms. However, Quad-gram of 4-grams of terms is rarely used by previous studies probably because the bi-gram or tri-gram gave the higher performance of their research. This text processing experiment uses RapidMiner tools to do the processing.

The first step is to import and load 1000 Malay raw documents into the repository. After that, drag all the relevant operators such as tokenization, transform cases, filter stopwords(dictionary), Stem(Dictionary), and generate n-Grams (term) and run the process. Tokenization is the process of breaking up the sentence or the text into pieces of work, while the transform cases operator is used to change all the letters in the document to the lower cases letter. Many words are not important in a document and do not give any meaning, called stopwords. It is needed to prepare the dictionary containing the list for all the Malay stopwords to use the filter stopwords operator in RapidMiner. The same goes with the stemming process, and the stem operator also needs to have the corpus to do the stemming process. Finally is to apply to generate the n-grams operator to generate the n-grams data representation. Figure 2 shows the example of text preprocessing in RapidMiner.

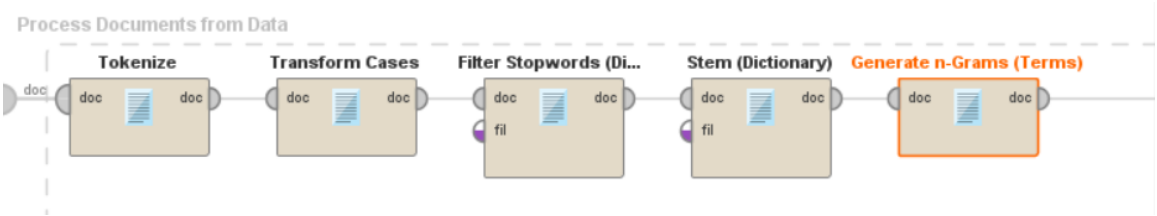

Figure 2: An example of preprocessing step in RapidMiner tool.

After that, feature extraction is performed by the proposed technique and the study's objectives. Feature extraction is a technique that involves finding and extracting unique features from textual documents. The N-grams document representation is a sequence of text 'n' words taken from a document and represented as a sequence of N-grams. Character, word, and byte can be recognized as 'textual units' based on their contexts of significance. Extracting character N-grams from a text is analogous to moving a "window" that is n characters wide over the document, one character at a time. The next process is to evaluate the effectiveness of N-Gram model, K-Means Clustering algorithm was chosen in this study that combined with the cosine similarity measure. The evaluation is assessed using precision, recall, accuracy, and f-measure. The table below shows the example of a feature or term extracted from the Malay text documents.

Table 1: Feature extraction example.

Feature	Result
Example words	bangunan belakang kedai makan
2-grams	bangunan_belakang, belakang_kedai, kedai_makan
3-grams	bangunan_belakang_kedai, belakang_kedai_makan
4-grams	banguna_belakang_kedai_makan

Table 1 shows an example text of the extraction using 2-grams, 3-grams, and 4-grams. The example word given in the table is the word obtained after undergoing a process in text preprocessing such as tokenization, stop word removal, and stemming. The text need to be stemmed to eliminates the unused work. Example word "bangunan belakang kedai" (*building behind shop*) is left after the filtering process. So, we can choose whether we want that word in a single word, 2-gram, etc. It means how many consecutive words we want to extract as keywords for related documents. If we choose the 3-gram(building_behind_shop), maybe using these keywords can easily clusteres the related English document, but it differs from Malay language.

3. Experimental Result & Discussion

The k-means clustering algorithm was used in this research to clustered the Malay unstructured documents. A real housebreaking crime data set is used to evaluate the proposed text processing method using Intel Core i7 processor, 8 GB of RAM, and Windows 10 machine. In addition, the proposed method was developed by using RapidMiner tools.

The investigation relies on 1000 Malay documents that contain five clusters of modus operandi. These unstructured documents will go through the text preprocessing process to clean documents. The following part of the research is generating features by using n-gram. After that, the n-gram document representation was applied to the k-means algorithm.

Our experiments' findings were examined by the accuracy, precision, recall, and f-measure of the proposed text processing approach using several N-gram windows to determine which types of n-gram had the highest accuracy. As a first step, the correctly clustered documents data were recorded for each experiment by other n-grams. Then the data were calculated to identify the accuracy, precision, recall, and f-measure of each output. Lastly, the various results of n-Grams are compared with the standard of text preprocessing techniques to acquire the best result and be considered a proposed method for this study. Table 3 depicts the list of relevant and irrelevant documents accomplished by k-means clustering.

Table 2: The relevant and irrelevant document obtain using k-means clustering algorithm.

Features	Cluster	Modus Operandi	Total of Documents Clustered	Relevant Documents Clustered	Irrelevant Documents Clustered
2-grams	Cluster 1	Role	162	85	77
	Cluster 2	Location	200	179	21
	Cluster 3	Method	339	194	145
	Cluster 4	Oddity	137	118	19
	Cluster 5	Weapon	162	131	31
		Total	**1000**	**707**	**293**
3-grams	Cluster 1	Role	197	98	99
	Cluster 2	Location	207	172	35
	Cluster 3	Weapon	106	98	8
	Cluster 4	Oddity	143	125	18
	cluster_5	Method	347	188	159
		Total	**1000**	**681**	**319**
4-grams	Cluster 1	Method	146	146	0
	Cluster 2	Oddity	219	165	54
	Cluster 3	Role	299	200	99
	Cluster 4	Weapon	171	136	35
	Cluster 5	Location	165	165	0
		Total	**1000**	**812**	**188**

As shown in table 2, three types of n-gram are implemented on 1000 data in the k-means clustering algorithm, including 2-grams, 3-grams, and 4-grams. The documents are classified into 5 clusters. A 4-grams model achieves the highest relevant document with 812 documents. A low performance was observed with a 3-grams model that predicted only 681 relevant documents clustered.

Table 3: Set for evaluation metric.

		Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
2-gram	TP	85	179	194	118	131
	FP	77	21	145	19	31
	FN	115	21	6	82	69
	TN	723	779	655	781	769
3-gram	TP	98	172	98	125	188
	FP	99	35	8	18	159
	FN	102	28	102	75	12
	TN	701	765	792	782	641
4-gram	TP	146	165	200	136	165
	FP	0	54	99	35	0
	FN	54	35	0	64	35
	TN	800	746	710	765	800

Table 3 portrays the metric evaluation assessment set for all n-grams based on the experiment results. True positive(TP), false positive(FP), false negative(FN), and true negative(TN) is calculated for each cluster to identify the performance value of a classified document. Based on the table, the true positive means how many documents clustered correctly in their class. In cluster 1, cluster 3 and cluster 4, the 4-gram gets the highest TP compared to 2-gram and 3-gram. Contradict for cluster 2, TP

is highest for 2-gram with 179 corrected document clustered. In cluster 5, the 3-gram model obtains the best performance for TP with 188 documents. These evaluation metrics only show the amount of correctly document clustered and incorrectly document clustered. The performance evaluation will present in table 4 below.

Table 4 : Overall experimental results.

n-gram	Cluster	Modus Operandi	Precision	Recall	Accuracy	F-measure
2-grams	Cluster 0	Peranan	0.5247	0.425	0.808	0.4696
	Cluster 1	Tempat	0.895	0.895	0.958	0.8688
	Cluster 2	Cara	0.5723	0.97	0.849	0.7199
	Cluster 3	Keganjilan	0.8613	0.59	0.899	0.7003
	Cluster 4	Senjata	0.8086	0.655	0.9	0.7238
		Average	**73.24%**	**70.70%**	**88.28%**	**69.65%**
3-grams	Cluster 0	Peranan	0.4975	0.4900	0.7990	0.4937
	Cluster 1	Tempat	0.8309	0.8600	0.9370	0.8452
	Cluster 2	Senjata	0.9245	0.4900	0.8900	0.6405
	Cluster 3	Keganjilan	0.8741	0.6250	0.9070	0.7289
	Cluster 4	Cara	0.5418	0.9400	0.8290	0.6874
		Average	**73.37%**	**68.1%**	**87.24%**	**67.914%S**
4-grams	Cluster 0	Cara	1.0	0.73	0.946	0.8439
	Cluster 1	Keganjilan	0.7534	0.8250	0.911	0.7876
	Cluster 2	Peranan	0.6689	1.0	0.9010	0.8016
	Cluster 3	Senjata	0.7953	0.68	0.9010	0.7332
	Cluster 4	Tempat	1.0	0.8250	0.9650	0.9041
		Average	**84.35%**	**81.2%**	**92.48%**	**81.40%**

Table 4 demonstrates the overall outcome of this study. The table shows the values for each cluster and n-gram. By carefully examining the data, it is found that 4-grams obtain the best result for precision, recall, accuracy, and f-measure with 83.35%, 81.2%, 92.48% and 81.40%. This method works well for the Malay documents and can cluster the higher relevant documents. The 2-gram and 3-gram did not show a significant difference among them. As we can see from the table, it is quite surprising that the 2-gram model is better than the 3-gram model. It can be concluded that an increase in the number of n-grams may not necessarily improve performance.

Table 5 : Average result for all docuemnts.

	Average values for all documents		
	2-gram	*3-gram*	*4-gram*
Precision	0.73	0.73	0.84
Recall	0.70	0.68	0.81
Accuracy	0.88	0.87	0.92
F-measure	0.69	0.67	0.81
Error rate	0.29	0.31	0.18

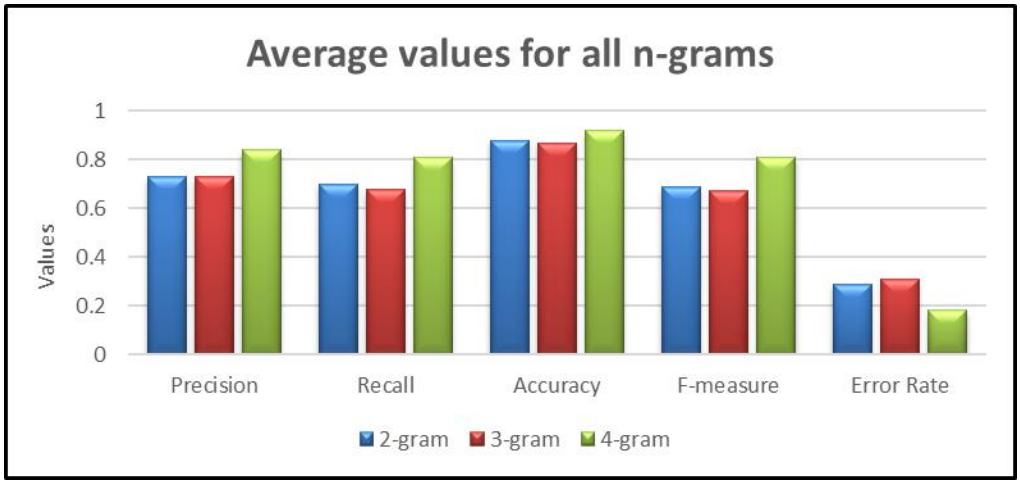

Figure 3 : The result for all n-gram.

This study focus on the text preprocessing techniques by implementing n-gram in the process. 1000 Malay text documents have been classified into 5 clusters by using a k-means clustering algorithm. As we can observe from figure 3, showed significant differences, while 2-grams and 3-grams had approximately the same value between them. The 4-grams obtain high importance for precision, recall, accuracy, and f-measure. At the same time, the error rate for 4-grams is the lowest among them. A total success achievement for this study is from 4-grams, and the values for all evaluations can be seen in table 5.

4. Conclusion and Future Recommendation

It was our goal in this study to propose an innovative technique to document categorization. Character N-gram-based format provides the best clustering performance while maintaining the smallest dimensionality possible. This result could be explained because the N-gram method is less prone to sparse-data issues and finds similarities between words and documents. Verbs patterns with N-grams that join two words can also be detected. To minimize dimensionality, we must disregard a large number of words, but frequent N-grams gather the frequency of many words and theoretically preserve more information by collecting frequent N-grams.

It is a complex task to find the efficient and effective technique to cluster the Malay documents. In this study, word-based clustering is not suitable due to the language dependence, as it does not perform well with the Malay document. Thus, this study aims to evaluate the impact of different N-gram models in text preprocessing using a real-world dataset from the Malaysian Police Department. Besides, we also discovered that feature selection based on document frequency might be used to document clustering using character N-gram representation when using character N-gram representation. Based on our findings, there is a limited number of studies that apply n-gram for documents clustering. After all, many classification approaches use this method to reduce text representation, but they mainly obtain unsatisfactory output performance. This is because the data or documents used are not suitable for applying the n-gram model, or else they might perform but the only used character or word level.

For this study, it has been demonstrated that the proposed approach uses 4-gram well on a real data set by achieving 92.48% accuracy higher than the standard procedure of text preprocessing, which is 87.32%. The use of the k-means clustering algorithm on 1000 of the Malay documents dramatically affects performance. It has been discovered that when the results of the test processes are analyzed, the technique presented within the study's scope is highly efficient and able to clustered the Malay document accurately. As a future work, the other clustering algorithm with 4-gram model can be implement to improve the Malay text document clustering.

Acknowledgement
This research is supported by Ministry of Education Malaysia, under Fundamental Research Grant Scheme (FRGS) with vote number 59541 (Reference Code: FRGS/1/2018/ICT04/UMT/02/3). The authors would also like to acknowledge Royal Police Department of Malaysia for their full support of this research.

References
[1] D. Birks, A. Coleman, and D. Jackson, "Unsupervised identification of crime problems from police free-text data," Crime Sci., vol. 9, no. 1, pp. 1–19, 2020, doi: 10.1186/s40163-020-00127-4.
[2] N. Abd Rahman, Z. Abu Bakar, and N. S. S. Zulkefli, "Malay document clustering using complete linkage clustering technique with Cosine Coefficient," ICOS 2015 - 2015 IEEE Conf. Open Syst., no. January 2016, pp. 103–107, 2016, doi: 10.1109/ICOS.2015.7377286.
[3] J. Agarwal, R. Nagpal, and R. Sehgal, "Crime Analysis using K-Means Clustering," Int. J. Comput. Appl., vol. 83, no. December, pp. 1–4, 2013, doi: 10.5120/14433-2579.
[4] M. Alruily, A. Ayesh, and A. Al-marghilani, "Using Self Organizing Map to Cluster Arabic Crime Documents," pp. 357–363, 2010.
[5] D. B. Bisandu, R. Prasad, and M. M. Liman, "Clustering news articles using efficient similarity measure and N-grams," Int. J. Knowl. Eng. Data Min., vol. 5, no. 4, p. 333, 2018, doi: 10.1504/ijkedm.2018.095525.
[6] F. Khoirunnisa, N. Yusliani, M. T, D. Rodiah, and M. T, "Effect of N-Gram on Document Classification on the Naïve Bayes Classifier Algorithm," vol. 01, no. 01, pp. 26–33, 2020.
[7] A. Deniz and H. E. Kiziloz, "Effects of various preprocessing techniques to Turkish text categorization using n-gram features," 2nd Int. Conf. Comput. Sci. Eng. UBMK 2017, no. May, pp. 655–660, 2017, doi: 10.1109/UBMK.2017.8093491.
[8] K. K. Purnamasari, "K-Means and K-Medoids for Indonesian Text Summarization," IOP Conf. Ser. Mater. Sci. Eng., vol. 662, no. 6, 2019, doi: 10.1088/1757-899X/662/6/062013.
[9] B. Aubaidan, M. Mohd, M. Albared, and F. Author, "Comparative study of k-means and k-means++ clustering algorithms on crime domain," J. Comput. Sci., vol. 10, no. 7, pp. 1197–1206, 2014, doi: 10.3844/jcssp.2014.1197.1206.
[10] R. T. Vulandari, W. L. Y. Saptomo, and D. W. Aditama, "Application of K-Means Clustering in Mapping of Central Java Crime Area," Indones. J. Appl. Stat., vol. 3, no. 1, p. 38, 2020, doi: 10.13057/ijas.v3i1.40984.