Research Article

Optimizing Text Categorization for Indonesian Text Using Clustering Label Technique

Syopiansyah Jaya Putra[1*], Teddy Mantoro[2], Muhamad Nur Gunawan[3], Ismail Khalil[4]

[1,3]Department of Information System, Faculty of Science and Technology, UIN SyarifHidayatullah, Jakarta, Indonesia
[2]Faculty of Engineering and Technology, Sampoerna University, Jakarta, Indonesia
[4]Institute of Telecooperation, Johannes Kepler University Linz, Austria
[1*]syopian@uinjkt.ac.id

Article History: Received: 10 November 2020; Revised: 12 January 2021; Accepted: 27January 2021; Published online: 05April 2021

Abstract: Text Categorization plays an important role for clustering the rapidly growing, yet unstructured, Indonesian text in digital format. Furthermore, it is deemed even more important since access to digital format text has become more necessary and widespread. There are many clustering algorithms used for text categorization. Unfortunately, clustering algorithms for text categorization cannot easily cluster the texts due to imperfect process of stemming and stopword of Indonesian language. This paper presents an intelligent system that categorizes Indonesian text documents into meaningful cluster labels. Label Induction Grouping Algorithm (LINGO) and Bisecting K- means are applied to process it through five phases, namely the pre-processing, frequent phrase extraction, cluster label induction, content discovery and final cluster formation. The experimental result showed that the system could categorize Indonesian text and reach to 93%. Furthermore, clustering quality evaluation indicates that text categorization using LINGO has high Precision and Recall with a value of 0.85 and 1, respectively, compare to Bisecting K-means which has a value of 0.78 and 0.99. Therefore, the result shows that LINGO is suitable for categorizing Indonesian text. The main contribution of this study is to optimize the clustering results by applying and maximizing text processing using Indonesian stemmer and stopword.

Keywords: Text categorization, text mining, clustering, Information retrieval, Lingoalgorithm

1. Introduction

Text categorization using clustering is a grouping of text documents which has capability handling categorization of high volume data. There are three types of data contained on a computer – structured, semi-structured, and unstructured. Currently, there are a lot of data stored in unstructured models such as full-text documents provided on the website, email, and others [1]. Therefore, text categorization has a role in putting a text document into the appropriate groups, so they can help in the process of finding information from large data sources [2,3].

Several techniques can be used for text categorization, while one of them is used clustering technique. Clustering is the grouping of the data set into the meaningful smaller groups, or also called clusters. One example of an application that uses clustering is Google News app that takes news from several sites, and then the news is grouped into specific topics such as business, technology, entertainment, sports, science, and others. Furthermore, text clustering plays a significant role in navigation and browsing process, and also can manage a large amount of stored electronic data [4]. Therefore, text categorization using clustering technique is automatic text grouping which has the capability to handle a significant amount of data and using the principle of maximizing the similarity between documents in the same group and minimize the similarity between groups.

The problem for huge unstructured data would be having the difficulty of getting information about the characteristics of the desired document from the data source. This is due to too much spread data and there are no proper tools to solve the problem. Search engines usually provide numerous search results when using common words as keywords, so users have difficulty in finding the desired information [5]. Furthermore, there are several challenges in text clustering, such as determining the similarity between the text and determines how a text is suitable to fit into a cluster. Moreover, a proper text consists of a set of words from a particular language, while every language has a morpheme, word, and different grammar. Therefore, text categorization using clustering techniques is also one way to categorize unstructured text for easy management and access.

Study on text categorization applies various techniques, namely classification, clustering, semi-supervised, and machine learning. For categorization, applied classification technique uses these algorithms: K-Means [6], and Support Vector Machine [7]. For clustering technique, it uses these algorithms: Frequent-term Based [8], LINGO [9], K-Means [10], Bisecting K-means [11] , and Term Weighting [12]. For the semi-supervised learning uses these algorithms: Semi-Supervised Agglomerative Hierarchical Clustering and Semi-Supervised Fuzzy c-Means [13]. For machine learning uses Feature Selection algorithm [14].

Text categorization using LINGO clustering algorithm has been applied for English [15], Polish [9], and Marathi [5]. Text categorization in English using LINGO showed that evaluation value for Precision (89.5-91) and Recall (90-91), therefore, based on algorithm comparison indicates that LINGO is better than K-Means [15]. Likewise, text categorization for the Polish evaluated by some users' shows that the useful clusters are equal to 70-80%, and 80-95% of snippets inside those clusters matching their topic [9]. Similarly, for Marathi, it shows good evaluation results which indicate Precision and Recall with values of 86.58 and 96.33, respectively [5].

Research for Indonesian text categorization has also been applied using clustering technique, namely the following algorithms: Fuzzy C-Means [16], K-Means [10], [17], and Single Pass Clustering [18]. However, it is only Single Pass Clustering algorithm using evaluation techniques of Precision and Recall, which has the values of 0.79 and 0.88, respectively. Based on this, the Indonesian text categorization using Single Pass Clustering algorithm shows a lower value than the use of LINGO in other languages mentioned above.

Another method is Bisecting K-means clustering technique also applied to process the data set into four stages [11]: (1) Select a cluster to split, (2) Find 2 sub-clusters using the basic K-means algorithm, (3) Repeat step 2, the bisecting step, for a fixed number of times and take the split that produces the clustering with the highest overall similarity, (4) Repeat step 1, 2, and 3 until the desired number of clusters is reached. Finally, to determine experimental performance result uses Precision, Recall, and F-measures formulas. Therefore, research on text categorization of Indonesian text documents using LINGO and Bisecting K-Means algorithms need to be examined.

This paper presents an intelligent system that categorizes Indonesian text using two clustering techniques. In the proposed clustering technique, clustering algorithm methods are applied to process the data set into several stages: (1) pre-processing, (2) extract frequent phrase, (3) cluster label induction, (4) content discovery, and (5) final cluster formation. Pre-processing of the dataset performs tokenization, stemming and stopword mark. In this phase, we optimize the stemming and stopword to increase the clustering result [28-29]. The next phase, it extracts the phrases to be candidates label. In cluster label induction phase, the term-document matrix (TDM) is constructed based on term frequency- inverse document frequency (TF-IDF) weighting scheme. Also, Singular Value Decomposition (SVD) technique is applied to identify labels of each cluster. Cluster content discovery phase uses Vector Space model to assign input document. Finally, text clusters are sorted based on the score. To determine experimental performance result uses Precision, Recall, and F-measure formulas.

The experimental result of this study demonstrated that the system could categorize Indonesian text contained in the data set to reach 93% and 80%. Also, it showed that LINGO which has a value of Precision, Recall, and F-measure are 0.85, 1, and 0.92, better than Bisecting K-means reach which only reach a value to 0.78, 0.99, and 0.87, respectively. The evaluation results are almost the same to the evaluation of Precision and Recall using LINGO forEnglish. Therefore, LINGO is suitable for categorizing Indonesian text. Furthermore, this study is applied and maximized Indonesian stemmer and stopword so that system can optimize the clustering label results, especially performance metrics Precision and Recall.

This study proposes optimizing text categorization for Indonesian text. For text processing uses Indonesian stemmer to produce the right word root, then for stopword marking process using Indonesian stopword. In this step, both LINGO and Bisecting K-means algorithms are used.

2. Methodology

For this research experiment, a dataset of Indonesian text documents containing Indonesian Translation Quran (ITQ): Surah Al-Kahf verses one trough 30 collected from http://tanzil.net/#trans/id.indonesian with the last update on June 4, 2010 was used. This dataset has 398 of words. Stopwords used contains 759 common words that are often appeared in Bahasa Indonesia [22]. Each dataset was translated and attributes selected by Carrot controller and categorized using LINGO and Bisecting K-means algorithms.

Clustering algorithm method processes the dataset in five phases, namely Pre-Processing, Frequent Phrase Extraction, Cluster Label Induction, Cluster Label Discovery, and Final Cluster Formation [9]. Figure 1 presents text categorization process using Clustering algorithm method.

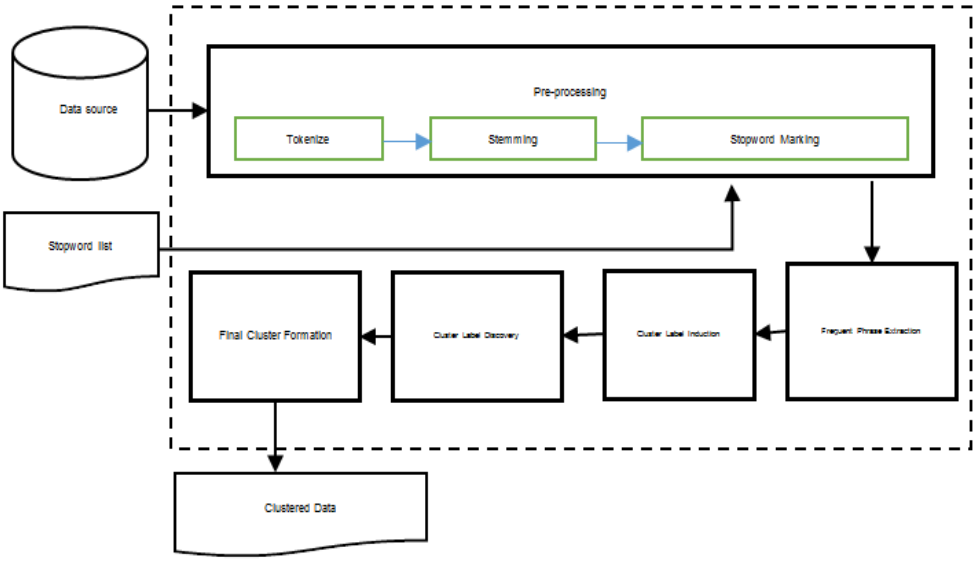

Figure 1. Text Categorization Process

The detail of text categorization process using clustering algorithm method will be described in the following section.

2.1 Pre-processing

In the pre-processing phase, the data selected is cleansing through some process that starts with tokenization on a dataset to produce token data. The tokens are normalized into a letter format by converting it to lowercase. At the tokenization process, input texts tokenize using tokenize script, based on space characters, and lowercase, for the example, the verse: *Merekakekal di dalamnyauntukselama – lamanya* (Where they will abide forever), will tokenize to: *mereka* | *kekal* | *di* | *dalamnya* | *untuk* | *selama – lamanya*.

The next process is apply stemming to make root form from words in preprocess text [23]. After tokenization process, word from the token above, *mereka* | *kekal* | *di* | *dalamnya* | *untuk* | *selama – lamanya*, will steam to *mereka* | *kekal* | *di* | *dalam* | *untuk* | *lama*. The stemming script, will steam the word using Indonesian Stemmer which already setup based on prefixes, suffix, and preposition in the Indonesian language.

The last method is to do stopword marking based on stopwords list indexed. At the stopword marking process, the script will mark the stopword with -1 (minus one), 0 (zero) for the Verse title, and 1 (one) for term or word. So, for the words above, the script will give mark *mereka* |1| *kekal* |1| *di* |-1| *dalam* | 1 |*untuk* |1| *lama*|1|. Pseudo-code Algorithm for Preprocessing:

1: D ← input documents (Indonesian texts)
STEP 1: Preprocessing
2: for all d ←D do (D: Input documents)
3: perform tokenization of d ;(d: document)
4: normalize into letter format;
5: if language of d recognized then
6: apply Indonesian stemming and mark stop words using stopword list (in d);
7: end if
8: end for

2.2 Frequent Phrase Extraction

In this phase, the phrases are extracted into candidates label if considered to meet several requirements [9]. The following requirements are: (1) if a phrase or a single term in the input documents appeared at least the same as term frequency threshold, (2) does not cross sentence boundaries, (3) being a complete phrase and not begin nor end with a stopword.

Pseudo-code algorithm for Frequent Phrase Extraction:
STEP 2: Frequent Phrase Extraction

8: concatenate all documents;
9: Pc← discover complete phrases; (Pc: Discover complete phrase)
10: Pf← p: {p ←Pc ∧ frequency (p) > Term Frequency Threshold}; (Pf: Discover number frequent phrase)

2.3Cluster Label Induction

Cluster label induction performs the following four steps; term-document matrix building, abstract concept discovery, phrase matching, and pruning label. Firstly, term-document matrix building is constructed based on term frequency-inverse document frequency (TF-IDF) [24]–[26] which is generated in the first phase. Secondly, the term-document matrix is calculated using the method of Singular Value Decomposition to find its orthogonal basis, which supposedly represents the abstract concepts appearing in the input documents [9]. Thirdly, the phrase matching process uses standard cosine distance to calculate how well a phrase or a single term represents an abstract concept, resulting in a value that is also used as the score of a label.

Cosine between document vector aj and the query vector q is calculated by the formula (1):

$$cos\theta_j = \frac{a_j^T q}{\|a_j\|\|q\|} = \frac{\sum_{i=1}^{t} a_{ij}\, q_i}{\sqrt{\sum_{i=1}^{t} a_{ij}^2}\,\sqrt{\sum_{i=1}^{t} q_i^2}} \quad (1)$$

Where aij is the degree of relationship between term i and document j, aj is the jth document vector, t is the number of terms, and ‖a‖ denotes the length of vector a, and T is the sequence of elements (t1, t2, t3,…, tn).

Fourth, during pruning label process, all pairs of candidate labels are calculated to get the similarities by using classic Vector Space Model, then select one label with the highest score for each group of similar tags.

Pseudo-code Algorithm for Cluster Label Induction:

STEP 3: Cluster Label Induction
11: A ← term-document matrix of terms not marked as stop-words and with a frequency higher than the Term Frequency Threshold ;(A: Term document matrix)
12:Σ, U, V ← SVD (A); {Product of SVD decomposition of A}
Σ : is a t × d diagonal matrix having the singular values of A ordered decreasingly along its diagonal.
U: is a t × t orthogonal matrix whose column vectors are called the left singular vectors of A.
V: is a d × d orthogonal matrix whose column vectors are called the right singular vectors of A.
13: k ← 0; {Start with zero clusters}
14: n ← rank (A);
15: repeats
16: k ←k+l;
17: q ←($\sum_{i=1}^{k} \Sigma ii$)/ ($\sum_{i=1}^{n} \Sigma ii$); (q: percentage quality threshold – Candidates Label Threshold)
18: until q < Candidate Label Threshold;
19: P ← phrase matrix for Pf;
20: for all columns of U_k^TP do
21: find the largest component mi in the column;
22: add the corresponding phrase to the Cluster Label Candidates set;
23: labelScore ← mi ;(mi = Vector cosines of the angles between the ith abstract concept vector and phrase and term vectors can be calculated m)
24: end for
25: calculate cosine similarities between all pairs of candidate labels;
26: identify groups of labels that exceed the Label Similarity Threshold;
27: for all groups of similar labels do
28: select one label with the highest score;
29: end for

2.4 Cluster Label Discovery

Cluster label discovery phase uses classic Vector Space Model to assign the input text to the cluster labels induced in the previous stage, and then matches the input snippets against a series of queries, each of which is a single cluster label [9]. Snippet assignment threshold values fall within the 0.0–1.0 range and empirically verified that thresholds within the 0.15–0.30 range produce the best results. For a certain query label, if the

similarity between a snippet and the label exceeds the Snippet Assignment Threshold, it allocates the text to the corresponding cluster. For those snippets that don't match any cluster labels are assigned to "Others" [20].

Pseudo-code algorithm for Cluster Content Discovery:

STEP 4: Cluster Content Discovery
30: for all L ← Cluster Label Candidates do
31: create cluster C described with L; (L: Label, C: Cluster)
32: add to C all documents whose similarity
to C exceeds the Snippet Assignment Threshold;
33: end for
34: put all unassigned documents in the "Others" group

2.5 Final Cluster Formation

In this final cluster formation phase, clusters are sorted based on their score. The score is calculated using the following formula (2):

$$C_{score} = label_{score} \times \|C\| \quad (2)$$

Where C is cluster, label score is label score, and $\|C\|$ is the total number of documents assigned to cluster C[9].

Pseudo-code Algorithm for Final Cluster Formation:
STEP5:FinalClusterFormation do
36:ClusterScore←LabelScorex$\|C\|$;
37: end for
C:Cluster
||C||: The number of documents assigned to cluster *C*
Clusterscore: Cluster Score
Labelscore:Label Score

2.6 Evaluation

In this phase, it measures the experimental testing and evaluates the performance using precision and recall, and F-measure [27]. The following formulas are Precision (3), Recall (4), and F-measure (5):

$$Precision = \frac{tp}{(tp + fp)} \quad (3)$$

$$Recall = \frac{tp}{(tp + fn)} \quad (4)$$

$$F - measure = \frac{2\ Recall\ x\ Precision}{(Recall + Precision)} \quad (5)$$

Where Precision is the fraction of the retrieved documents which are relevant, Recall is the fraction of the relevant documents which have been retrieved, tp (true positives) is a number of relevant elements which have been retrieved, fp (false positive) is a number of irrelevant element which have been retrieved, and fn (false negative) is a number or relevant elements which have not been retrieved. The F-Measure values are within the interval [0, 1] and larger values indicate higher detection quality. On the basis of these measures, overall precision and recall values as well as an overall F-measure value were computed as the average mean of the precision, recall and F-measure values for all documents.

3. Results and Discussion

This section explains each process of the experimental data and also evaluates the performance of the experimental testing. The dataset used in this experiment is the Indonesian text given on the Indonesian Translation of Qur'an (ITQ). Text categorization is performed on each of the datasets. First, it does parsing of the dataset and then chooses selected attributes. The selected attributes from the dataset are the name of surah and verse content.

3.1 Pre-processing

In tokenization phase, the data are cleansed from the characters and terms that may affect the quality of the group's description. To clean the text is done by making a cut word/token by a space character and punctuation found as seen in the first row in Table 1. Table 1 shows the results of sample documents that contain text cleansing "*Segalapujibagi Allah yang telah...*" ("All Praise be to Allah who has…"). Line Field Index in Table 1 shows that the type of field based on the attribute specified, namely a value of -1 for stopword, the value of 0 for the name of the letter, and the value of 1 to the contents of the paragraph.

Table 1. Sample data tokenization and stopword marking

Token	*Al Kahfi*(Al Kahf)	*Segala* (All)	*puji* (praise)	*bagi* (be to)	*Allah* (Allah)	*yang* (who)	*telah* (has)
Token Type	term	stopword	term	term	term	term	term
Field Index	0	-1	1	1	1	1	1

In the phase of normalization, case letter on the token image is normalized to lowercase, and in the steaming to get the word essence, so to token "*siapakah*" (who), it is turned into "*siapa*" (who), as well as for other tokens. However, irregularities were found in this phase such as token "*seorang*" (a man) that was supposed to be an "*orang*" (man).

3.2 Frequent Phrase Extraction

In this phase, the resulting candidate's phrases will be the cluster label. Table 2 shows the phrases were extracted successfully from the token data that have been made in the previous stage. Phrases column in this table is the cluster labels candidate to be used. Tokens will be a candidate if they meet the requirements determined, that does not begin and end with a stopword has a value of term frequency (TF), that exceed the value of the minimum term-predetermined frequency, namely 1. So, the only phrase which has a value of TF more than 1 will be the candidate, such as phrases: "*Tahun*" (Year) and "*Niscaya*" (Undoubtedly) as the cluster labels which have a value of TF: 3 and 4, respectively.

Table 2. Sample of Frequent Phrase Extraction Result

Phrases	Term Frequency (TF)	Inverse Document Frequency (IDF)	
		Number Document	Total Occurrence
Tahun (Year)	3	10	1
		24	2
*Niscaya (*Undoubtedly*)*	4	15	1
		19	2
		28	1

3.3 Cluster Label Induction

Based on the label candidates from the previous phase, this phase, 16 cluster labels index were generated with scores for each cluster label, such as index 476 with the score of 23.79, index 111 with the score of 16.97, and so on. The number of cluster label index represents the position of a phrase within the dataset. Therefore, if it is checked on the cluster label data, cluster label index 476 is a presentation of the phrase "*Tahun* (Year)," index 111 is a presentation of the phrase "*Niscaya* (Undoubtedly)," and so on. The cluster label score shows the value of the label similarity with an abstract concept. Table 3 presents sample result of cluster label induction with cluster label index and score.

3.4 Cluster Content Discovery

In this phase, the document associated with the cluster label is found using classic Vector Space Model. Table 4 presents a sample of the documents related to the cluster label index "476" where the actual content is the "*Tahun"*(Year) and the clustered index label "111" that has the content "*Niscaya*" (Undoubtedly). The documents associated with each group label are found by using classic Vector Space Model. For cluster label "*tahun"* (year), two related documents were found (Table 4).Documents are chosen because it has the phrase "*tahun"* (year) in it. For documents area not assigned to any cluster, it will be put in a cluster labeled "Others."

Table 3. Sample Result of Cluster Label Induction

Cluster Label Index	476	111	216	22	240	356	520	186	513	342

Cluster Label Score	23.79	16.97	16.45	14.71	14.19	13.27	13.09	11.67	11.67	11.48

Table 4. Sample of Discovered Content of some Clusters

Cluster Label Index	**Cluster Label Name**	**Assigned Documents**
476	*Tahun* (Year)	*"Maka Kami tutuptelingamerekabeberapatahundalamguaitu"* So, We struck them several years in a cave
		Dan merekatinggaldalamguamerekatigaratustahun dan ditambahsembilantahun (lagi). And they stayed in their cave three hundred years and added nine years (more)
111	*Niscaya (*Undoubtedly*)*	*"Sesungguhnyajikamerekadapatmengetahuitempatmu, niscayamerekaakanmelemparkamudenganbatu, ataumemaksamukembalikepada agama mereka, dan jikademikianniscayakamutidakakanberuntungselamalamanya".* Indeed, if they can figure out where you are, they will throw you with a stone, or force you to return to their religion, and if so*undoubtedly* you will not win forever.

Table 5. Comparison Precision, Recall, and F-measure of Algorithm Method

Algorithm Method	Number of Documents	Number of Categories/ Cluster Labels	Number of Documents on All Categories (%)	Number of Documents Cannot be Categorized (%)	Evaluation		
					Precision	Recall	F-measure
LINGO	30	16	93%	7%	0.85	1	0.92
Bisecting K-means	30	7	80%	20%	0.78	0.99	0.87

3.5 Final Cluster Formation

This is the cluster that has been generated by another vote in this phase. This phase presents the results of scoring from the resulting cluster. Cluster score shows only clusters that are considered good by LINGO and Bisecting K-means. In this phase, the resulting score from each cluster. For cluster "*Tahun* (Year)" has a 23.79 score, cluster "*Niscaya* (Undoubtedly)" have a score 16.97, and so on. Value scores do not indicate the level of quality of a cluster; the higher score shows just how good these clusters for LINGO, so it does not have any influence on the outcome of performance metrics.

3.6 Evaluation

In this experiment, the result of study shown in Table5, the prepared dataset consists of 30 documents, Cluster label induction using LINGO method was generate 16 categories or cluster labels. The result contains 93% of documents, but there are 7% of the documents that cannot be categorized. Cluster label using Bisecting K-means method was generate 7 categories or cluster labels, contains 80 % of documents, and 20 % of the documents that cannot be categorized.

Based on the measurement of the accuracy of the documents placed in each category (Precision) shows an average value of 0.85, while the measurement of relevant documents called for each category (Recall) shows an average value of 1. Compared with Bisecting K-Means with a value of Precision0.78 and Recalls0.99, thus LINGO clustering algorithm has a better quality than Bisecting K-means. F Measure Using LINGO 0.92 and F Measure Using Bisecting K-Means 0.87.

Previous research used the K-Means algorithm for the categorization of Indonesian text [17], performed using the F-Measure (0.67) and purity (0.61) evaluation. Therefore, our study is better than using the K-means algorithm [17], as our research optimizes the stemming and stopword processes that are compatible with the Indonesian language. Similarly, our experimental results are also better than the Fuzzy C-Means algorithm [16] which uses tf-idf weighting for cluster label determination, whereas our research uses Singular Value Decomposition and Vector Space Model to create cluster labels and cluster documents.

Based on the experiment result, there is a limitation to suit for text categorization of large data set since LINGO algorithm takes up much memory and a high number of matrix transformations. Similarly, the discovery of stemming error from the pre-processing phase which greatly affects the subsequent phases, it influences the

outcome of precision in performance metrics. Thus, a more optimal method using stemming algorithm is crucial to producing clusters with high accuracy. Another limitation is the lexical use of Indonesian texts that can influence performance metrics, including the lexical to clustering the documents which have to be considered [21].

4. Conclusion

This paper presents an achievement of the intelligent text categorization for Indonesian text into meaningful cluster labels by adopting LINGO and Bisecting K-means clustering algorithms. There are 93%and 80% of Indonesian text documents obtained from the dataset can be categorized as the cluster labels. Furthermore, clustering quality evaluation indicates that the text categorization has values of Precision, Recall, and F-measure of0.85, 1, and 0.92, respectively, and 0.78, 0.99, and 0.87 for Bisecting K-means.LINGO clustering algorithm has a better result when compared to Bisecting K-means method. Therefore, LINGO clustering algorithm is suitable for categorizing the Indonesian text documents.

The experiment result proves that users can obtain Indonesian text categorization using clustering technique more exactness and completeness. The main contribution of this study is to optimize the clustering results by applying and maximizing text processing using Indonesian stemmer and stopword.

Future research direction is required applying improved Indonesia stemming algorithm and stopword, as well as Indonesian lexical databases to get a better cluster quality. Furthermore, text categorization using large text documents should be implemented to maximize the performance and clustering results.

References

1. K. Sumathy and M. Chidambaram. 2013. Text Mining: Concepts, Applications, Tools and Issues–An Overview. *International Journal of Computer Applications*. 80(4). 29–32.
2. E. S. Han, G. Karypis, and V. Kumar. 2001. Text Categorization Using Weight Adjusted k-Nearest Neighbor Classification. *Data Mining and Knowledge Discovery*. 3918.5 3–65.
3. F. Sebastiani. 2005. Text Categorization, Text Mining and its Applications to Intelligence, *CRM and Knowledge Management*. 109–123.
4. M. J. Basha, K. P. Kaliyamurthie, T. Nadu, and A. Info. 2017. An Improved Similarity Matching based Clustering Framework for Short and Sentence Level Text. *International Journal of Electrical and Computer Engineering*. 7(1). 551–558.https://doi.org/10.11591/ijece.v7i1.pp551-558
5. S. R. Vispute and P. M. A. Potey. 2013. Automatic Text Categorization of Marathi Documents Using Clustering Technique. *Advanced Computing Technologies (ICACT), 2013 15th International Conference on. IEEE*. 1–5. https://doi.org/10.1109/ICACT.2013.6710543
6. X. Zhou, Y. Hu, and L. Guo. 2014. Text categorization based on clustering feature selection. *2nd International Conference on Information Technology and Quantitative Management, ITQM*.
7. Pilászy. 2005. Text categorization and support vector machines. *The Proceedings of the 6th International Symposium of Hungarian Researchers on Computational Intelligence.* vol. 1.
8. Beil, M. Ester, and X. Xu. 2002. Frequent term-based text clustering. *KDD '02 Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining.* 436.https://doi.org/10.1145/775047.775110
9. S. Osinski, J. Stefanowski, and D. Weiss. 2004. Lingo: Search Results Clustering Algorithm Based on Singular Value Decomposition. *Intelligent Information Processing and Web Mining*. 358–368.
10. T. Khotimah. 2014. *Pengelompokan Surat Dalam Al Qur'an MenggunakanAlgoritma K-Means* (Grouping Surah in the Qur'an Using K-Means Algorithm). *Jurnal SIMETRIS*, 5(1).
11. M. Steinbach, G. Karypis, and V. Kumar. 2000. A Comparison of Document Clustering Techniques. *KDD workshop on text mining.* 400(X). 1–2.
12. M. Lan, C. L. Tan, J. Su, and Y. Lu. 2009. Supervised and Traditional Term Weighting Methods for Automatic Text Categorization, *Pattern Analysis and Machine Intelligence, IEEE Transactions.* 31(4). 721–735.
13. Benkhalifa and A. Bensaid.1999. Text Categorization using the Semi-Supervised Fuzzy c-Means Algorithm, *in 18th International Conference of the North American. IEEE.* https://doi.org/10.1109/NAFIPS.1999.781756
14. Y. Peng, Z. Xuefeng, Z. Jianyong, and X. Yumhong. 2009. Lazy learner text categorization algorithm based on embedded feature selection, *Journal of Systems Engineering and Electronics.* 20(3). 651–659.
15. J. Zhang and S. Chen. 2013. A study on clustering algorithm of Web search results based on rough set, *in Software Engineering and Service Science (ICSESS).*

16. S. Y. Charezita and Suyanto. 2012. *Clustering Dokumen Bahasa Indonesia DenganMenggunakan Fuzzy C-Means* (*Indonesian Document Clustering Using Fuzzy C-Means*), *Universitas Telkom*.
17. Y. Dwi, P. Negara, and M. Syarief. 2015. *ClusterisasiDokumen Web (Berita) Bahasa Indonesia MenggunakanAlgoritma K-Means* (Clustering Web Documents (News) Indonesian Language Using K-Means Algorithm). 4(3). 159–166.
18. A. Z. Arifin and A. N. Novan. 2002. KlasifikasiDokumenBeritaKejadianBerbahasa Indonesia denganAlgoritma Single Pass Clustering (Classification of Indonesian News Events Documents with Single Pass Clustering Algorithm), *Proceeding of Seminar on Intelligent Technology and Its Applications (SITIA), 2002*.
19. G. Krishnasamy, A. J. Kulkarni, and R. Paramesran. 2014. A hybrid approach for data clustering based on modified cohort intelligence and K-means, *Expert Systems with Applications*. 41(13). 6009–6016.https://doi.org/10.1016/j.eswa.2014.03.021
20. Osiński and D. Weiss. 2005. A concept-driven algorithm for clustering search results, *IEEE Intelligent Systems.* 20(3). 48–54. https://doi.org/10.1109/MIS.2005.38
21. Demir, E. Sezer, and H. Sever. 2014. Modifications for the Cluster Content Discovery and the Cluster Label Induction Phases of the Lingo Algorithm, *International Journal of Computer Theory and Engineering.* 6(2). 87–90. https://doi.org/10.7763/IJCTE.2014.V6.842
22. F. Z. Tala, 2003. *A Study of Stemming Effects on Information Retrieval in Bahasa Indonesia*, M.Sc. Thesis. Appendix D. 39–46.
23. P. Bhole and A. J. Agrawal. 2014. Extractive Based Single Document Text Summarization Using Clustering Approach, *IAES International Journal of Artificial Intelligence.* 3(2). 73–78.
24. P. V. Amoli and O. S. Sh. 2015. Scientific Documents Clustering Based on Text Summarization, *International Journal of Electrical and Computer Engineering (IJECE)*. 5(4). 782 - 787.
25. S. J. Putra, K. Hulliyah, N. Hakiem, R. P. Iswara, and A. F. Firmansyah, 2016. Generating Weighted Vector for Concepts in Indonesian Translation of Quran. *Proceedings of the 18th International Conference on Information Integration and Web-based Applications and Services. ACM, 2016*.https://doi.org/10.1145/3011141.3011218
26. S. J. Putra, R. H. Gusmita, K. Hulliyah, and H. T. Sukmana. 2016. A semantic-based Question Answering System for Indonesian Translation of Quran. *Proceedings of the 18th International Conference on Information Integration and Web-based Applications and Services*.2016. 506–509.https://doi.org/10.1145/3011141.3011219
27. E. Rendón, I. Abundez, A. Arizmendi, and E. M. Quiroz. 2011. Internal versus External cluster validation indexes, *International Journal of Computers and Communications*. 5(1). 27- 34.
28. Tunali, Volkan, and TurgayTugayBilgin. 2012. Examining the impact of stemming on clustering Turkish texts. *Innovations in Intelligent Systems and Applications (INISTA), 2012 International Symposium on. IEEE*. https://doi.org/10.1109/INISTA.2012.6246966
29. Uysal, AlperKursat, and Serkan Gunal. 2014. The impact of preprocessing on text classification. *Information Processing & Management*. 50(1). 104 -112. https://doi.org/10.1016/j.ipm.2013.08.0s06