KINETIK, Vol. 3, No. 2, May 2018, Pp. 91-102
ISSN : 2503-2259
E-ISSN : 2503-2267

Paragraph Selection Methods Using Feature-Based on Segment-Based Clustering Process Using Paragraphs for Identifying Topics on Indication Detection of Plagiarism System

Denar Regata Akbi[*1], Arini Rahmawati Rosyadi[2]
[1,2]Universitas Muhammadiyah Malang
dnarregata@umm.ac.id [*1], arini.rosyadi@gmail.com[2]

Abstract

In segment-based clustering, the paragraphs selection as a dataset in the clustering process has a very important role. This is because the paragraph used as the dataset can affect the clustering result. This research uses paragraph selection using feature-based method which aims to optimize the clustering process conducted in the previous research. Based on the evaluation results using Silhouette Coefficient and Sum Square Errors evaluation methods to find the proper k value, it is found that with the utilization of Feature-based method, better results can be acquire compared to the evaluation result from the previous research.

Keywords: *Feature-Based, Paragraphs Selection, Segment-Based, Silhouette Coefficient, Sum Square Errors*

1. Introduction

Plagiarism is an act violating someone's copyright or work. It occurs in various fields. Plagiarism often occurs in scientific articles.

Several studies have been proposed to detect plagiarism. Brooke and Hirst in 2012 conducted a research on differentiating writing styles in one text document [1]. In addition, Brooke and Graeme (2012) conducted a study resulting of the effectiveness of n-Gram value showing a decrease to 30%. It is because the topic's data sets are not well regulated [2]. In 2013, Shrestha & Solorio proposed a method for detecting different types of plagiarism because the existing system was unable to recognize the type of plagiarism used. Hence, the application of variation of n-Gram method was proposed in order to detect the type of plagiarism [3]. Jiffriya, Jahan, Ragel, & Deegalla, in 2013, enacted the use of clustering to detect plagiarism because it can reduce detection time. According to the results, it led to four-time faster detection, but the clustering process only resulted in similarity values of the paired documents which are considered similar [4].

From some of these studies, no one has noticed the variations of the topics contained in a document. The variation of topics in a document may affect the results of plagiarism in the detection time and the accuracy of the results of detection. In 2015, Rosyadi and Arini used segment-based clustering aiming on identifying multiple topics in a set of documents [5]. However, this study had been found to be not optimal because the method of selecting paragraphs in each document is based only on the length of the paragraph, regardless of the core topic of the paragraph in the document. Thus, it may be possible for paragraphs, which have a core topic but having an unsuitable length with a given threshold value, will not be included in the process of plagiarism indication.

Therefore, this study proposes the use of paragraph selection method using feature-based on segment-based clustering process. It aims to improve the clustering results in the previous research leading to the assumption of improving the accuracy of the results obtained.

Feature-based methods are used to find important sentences in the text [6]. The study by Luhn (1999) saw the frequency of occurrence of words as an important thing in a document; Luhn assumed the words that often appear in documents should indicate something important in a paragraph or document. One of the features used in Luhn's research is the length of the sentence [7].

Meanwhile, this study utilizes feature-based method in the proximity of a paragraph to the title of the document. Moreover, it aims to select the appropriate paragraph for clustering.

Akbi, D., & Rosyadi, A. (2018). Paragraph Selection Methods Using Feature-Based On Segment-Based Clustering Process Using Paragraphs For Identifying Topics On Indications Detection of Plagiarism System. Kinetik, 3(2). doi:http://dx.doi.org/10.22219/kinetik.v3i2.593
Receive November 16, 2017; Revise December 15, 2017; Accepted January 03, 2018

2. Research Method

According to Rosyadi's study conducted in 2015, there are several major stages being undertaken [5]: (1) Running segment-based clustering process aiming at identifying multi topics from the set of documents [8]; (2) Identifying the topic using the weighting method tf-idf and tf-issf; and (3) Detecting plagiarism indications using Windowing Algorithm, n-Gram and hashing methods, presented in Figure 1.

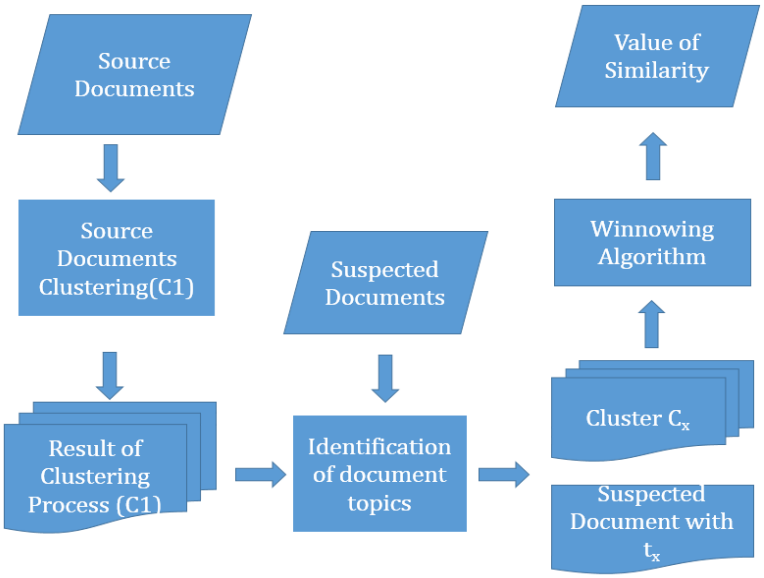

Figure 1. Flowchart of the Previous Research [5]

Figure 2 is a segment-based clustering process in a text document. This process begins by segmenting each document based on its own paragraph, generating segments on each document. Furthermore, the segments which have been generated are used in the clustering process using K-Means algorithm. The process yields clusters containing segments, each cluster called by a segment set. The process is called by paragraph clustering. The resulting segments are subsequently used as input to perform a segment-based clustering process using K-Means algorithm. This process is called by clustering paragraph process.

This research proposes improvements in the process stages of segment-based document clustering by adding a paragraph selection method using Feature Based which uses the proximity of the paragraph contents to the document. It is illustrated in Figure 3.

In this study, the dataset used is the same dataset as the previous research's dataset, utilizing 170 journal documents downloaded randomly with various topics as source documents or comparative documents.

Scenario testing is completed by comparing the system performance from previous research with research proposal. The process of testing the system carries out several tests including:

1. Testing the effect of the number of paragraphs of each source document on the use of Feature Based.
2. Testing clustering process evaluation using Silhouette Coefficient method [9] and Sum Squared Errors [10], performed to obtain the appropriate k value in the process:
   a) Clustering paragraph.
   b) Clustering clustered paragraph.

3. Result and Discussion

3.1 Feature-Based Method

In the implementation of feature-based methods differences are identified in the number of paragraphs in each source document. The differences are given in Table 1.

Based on Table 1, it is discovered that the difference in the number of paragraphs obtained from the use of feature based and without using feature based has an average of 18.64

paragraphs. It can be concluded that every document experience average decreasing number of paragraph by 42.26%.

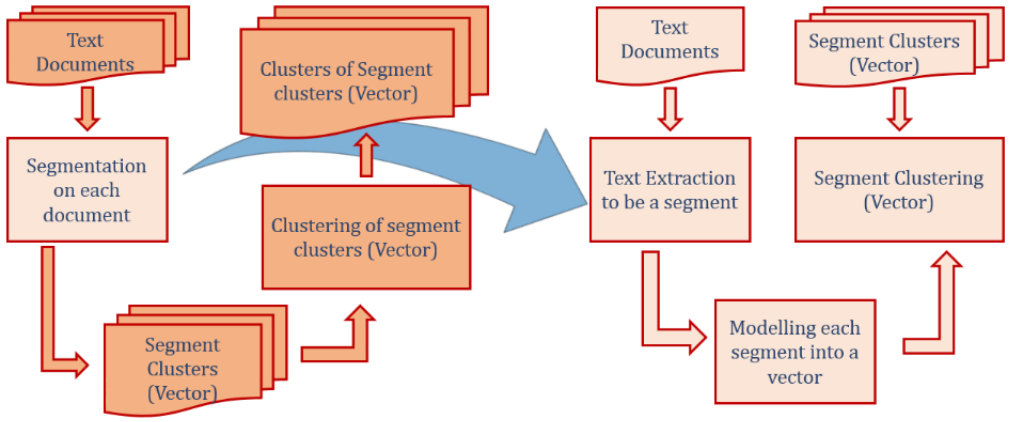

Figure 2. Document Clustering Process [5]

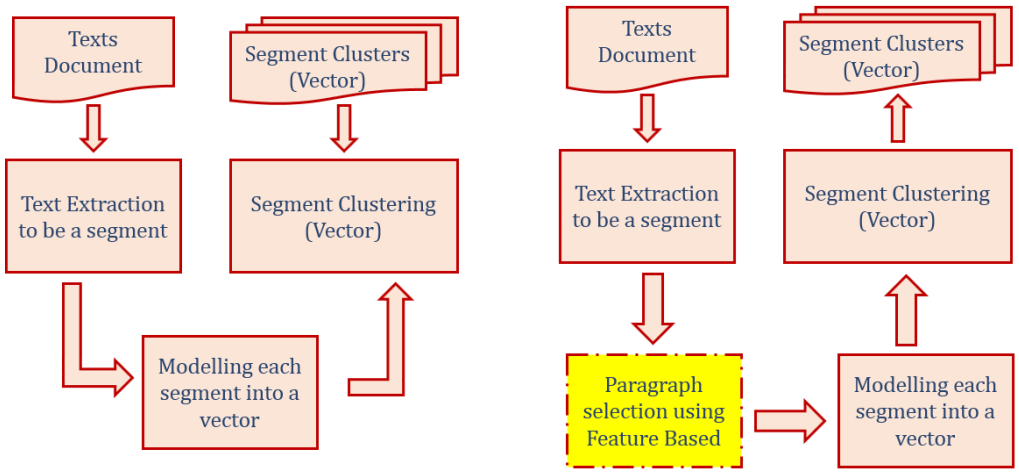

Figure 3. Proposed System: Paragraph Selection Method Using Feature based on Segment-Based Clustering Process

Table 1. The Effect of the Number of Paragraphs on Each Document on the Use of Feature-Based Method.

FIle Name	Without Feature-Based	Using Feature-Based	Difference number of paragraphs	Decreasing number of paragraphs (%)
1	43	23	20	46.51
2	8	3	5	62.5
3	58	40	18	31.03
4	21	17	4	19.04
5	42	22	20	47.61
6	18	10	8	44.44
…	…	…	…	…
167	54	32	22	40.74
168	21	12	9	42.85
169	29	16	13	44.82
170	24	16	8	33.33
	Average		18.64	42.26

3.2 Segmen-Based Clustering Process

3.2.1 Clustering Paragraph

Paragraph clustering process is the clustering process of each source document based on the paragraph. Clustering process on each document aims to identify the topic of the source

document. The previous research has assumed that any document may have a main topic as well as several sub-topics.

In paragraph clustering process, the clustering process is based on the number of paragraphs contained in a single source document. Thus, the grouping of source documents, in this process, is implemented based on the number of paragraphs. The groupings are presented in Table 2.

To evaluate segment-based clustering process, the evaluation method of Silhouette Coefficient and Sum Square Errors are employed to be able to determine the proper k value in clustering process. Each category of source document is evaluated against the value of Silhouette Coefficient and Sum Square Errors. The results of Silhouette Coefficient testing on short-length document can be seen in Table 3. The results of Silhouette Coefficient testing on medium-length document can be seen in Table 4 and long-Document can be seen in Table 5.

Table 2. The Document Category is based on the Number of Paragraphs on Each Document

Document category	Range of number of Paragraph	Number of documents	Variation of k value (previous research)	Variation of k value (proposed research)
Short-length document	0 < p ≤ 10	60	2, 3, 4	2, 3
Medium-length document	10 < p ≤ 25	70	3, 4, 5, 7, 8	3, 4, 5, 7, 8
Long-length document	P > 25	40	3, 4, 5, 7, 8, 10	3, 4, 5, 7, 8, 10

Table 3. Test Results of Short-Length Document Using Silhouette Coefficient

Short-length document	Silhouette Coefficient	
	k	
	2	3
2	0.390852135	0.39085214
6	0.481427355	0
7	0.046735954	-0.0807595
19	-0.001612083	0.41801883
26	0.633209367	0.4685907
35	0.024383622	0.34794663
…	…	…
106	0.403912423	0.55763751
108	0.41749931	0.42512052
109	0.05586646	0.7383077
146	0.446989844	0.22518114
Average	0.303931417	0.22518114

Based on the evaluation of each document category, then the average of each category is given for comparison with the evaluating result from the previous research.

Table 4. Comparison of Silhouette Coefficient Values between Prior and Proposed Research

Document Category	*Silhouette Coefficient*							*Research*
	Nilai k							
	2	3	4	5	7	8	10	
Short	0.265	0.159	-0.405					
Medium		0.126	-0.405	-0.507	-0.542	-0.497		Previous
Long		0.172	-0.362	-0.481	-0.622	-0.467	-0.560	
Short	0.304	0.225						
Medium		0.159	-0.482	-0.391	-0.549	-0.565		Proposed
Long		0.188	-0.364	-0.406	-0.504	-0.546	-0.488	

Table 5. Test Results of Short-Length Document Using Sum Square Errors

Short-length document	*Sum Square Errors*	
	k	
	2	3
2	0.01184275	0.01184275
6	0.066529975	0.07028802
7	0.0598	0.04422287
19	0.067018902	0.03406137
26	0.007186222	0.01073978
35	0.074491125	0.04197865
…	…	…
106	0.014876222	0.01082156
108	0.048648472	0.04828939
109	0.08816803	0.03455419
146	0.02950008	0.03553316
Average	0.042756424	0.03553316

The comparison is presented in Table 6.

Table 6. Test Results of Medium-Length Documents Using Sum Square Errors

Medium-length document	*Sum Square Errors*				
	k				
	3	4	5	7	8
1	0.2746580	0.359851	0.182467	0.180108	0.348634
4	0.0464370	0.063006	0.049113	0.061269	0.061271
5	0.2057690	0.173321	0.284618	0.295114	0.258354
8	0.1901160	0.249578	0.227327	0.318914	0.291835
9	0.076781	0.121907	0.102575	0.10876	0.078039
10	0.5529744	0.526287	0.51920	0.383899	0.432841
…	…	…	…	…	…
166	0.068249	0.123865	0.04425	0.105360	0.090193
168	0.0830165	0.075027	0.064868	0.05507	0.092750
169	0.1201828	0.084411	0.069776	0.112738	0.06632
170	0.0411803	0.119959	0.069730	0.095795	0.126474
Average	0.1654333	0.182003	0.180978	0.181009	0.172342

The results of Sum Square Errors testing on short-length document are presented in Table 7.

Table 7. Test Results of Long-Length Document Using Sum Square Errors

Long-Document	*Sum Square Errors*					
	k					
	3	4	5	7	8	10
3	0.665188317	0.261774	0.605816	0.451133	0.596655	0.40574
25	0.285211822	0.387600	0.440110	0.437871	0.348386	0.28184
27	0.258559112	0.365683	0.360917	0.332782	0.285851	0.31918
29	0.126975813	0.104532	0.222118	0.141240	0.222450	0.1724
30	0.084923534	0.24131	0.238510	0.29178	0.198640	0.10294
31	0.198656482	0.274430	0.279593	0.28140	0.301632	0.21486
…	…	…	…	…	…	…
162	0.127745114	0.180630	0.199754	0.174468	0.162809	0.13539
163	0.467067856	0.423199	0.473393	0.459720	0.430118	0.33602
164	0.055878453	0.154751	0.077321	0.150660	0.138073	0.14810
167	0.207109921	0.245965	0.171364	0.085035	0.212021	0.25737
Average	0.30509967	0.315170	0.325944	0.325155	0.321622	0.29238

Based on the evaluation of each document category, then the average of each category is given to be compared with the evaluation results from the previous research. The results of Sum Square Errors testing on medium-length document are presented in Table 8.

Table 8. Comparison of Sum Square Errors Values between Prior and Proposed Research

Document Category	Sum Square Errors k							Research
	2	3	4	5	7	8	10	
Short	0.050	0.042	0.055					
Medium		0.123	0.149	0.156	0.152	0.140		Previous
Long		0.314	0.334	0.358	0.335	0.332	0.315	
Short	0.043	0.036						
Medium		0.165	0.182	0.181	0.181	0.172		Proposed
Long		0.305	0.315	0.326	0.325	0.322	0.292	

The results of Sum Square Errors testing on Long-Document can be seen in Table 9.

Table 9. Clustered Paragraph Cluster Evaluating Using Silhouette Coefficient

Number of experiment	Silhouette Coefficient k			
	5	8	10	12
1	0.260	-1.000	-0.069	-0.200
2	0.264	-0.055	-1.000	-1.000
3	0.246	-1.000	-0.500	-0.600
4	0.172	-1.000	-1.000	-0.213
5	-0.001	-0.058	-0.600	-1.000
6	0.192	0.000	0.000	-1.000
7	-0.038	-1.000	-0.068	-0.429
8	-0.074	-0.049	-1.000	-0.059
9	0.321	-0.291	-1.000	0.200
10	0.443	0.156	0.148	0.166
Average	0.179	-0.430	-0.509	-0.414

The comparison is presented in Table 10.

Table 10. Clustered Paragraph Cluster Evaluating Using Sum Square Errors

Number of experiment	Sum Square Errors k			
	5	8	10	12
1	2.642	1.980	1.730	1.907
2	2.911	1.833	2.196	2.048
3	2.110	1.727	1.798	1.477
4	2.830	2.253	1.952	1.769
5	4.463	2.020	1.748	2.865
6	3.201	2.168	1.172	1.777
7	1.588	1.592	1.935	2.418
8	3.814	1.475	2.351	1.727
9	1.996	2.142	2.909	0.962
10	1.792	1.601	1.562	1.494
Average	2.735	1.879	1.935	1.844

Clustering process of cluster paragraph is a clustering process which involves the output of paragraph clustering process, i.e. clusters derived from each document to be re-clustered to obtain a new cluster of all source documents. It is intended to group the same sub-topic of all source documents.

Testing in this process is conducted ten times on each evaluation method. Table 11 and Table 12 present test results obtained. The evaluation results are comparable with testing from previous studies. The comparison results are presented in Table 13 and Table 14.

Table 11. Value Comparison of Silhouette Coefficient on Clustered Paragraph Cluster Process on Prior and Proposed Research

Research	*Silhouette Coefficient*			
	k			
	5	8	10	12
Previous	0.551	0.455	0.535	0.438
Proposed	0.179	-0.430	-0.509	-0.414

Table 12. Value Comparison of Sum Square Errors on Clustered Paragraph Cluster Process on Prior and Proposed Research

Research	*Sum Square Errors*			
	k			
	5	8	10	12
Previous	0.884	0.297	0.314	0.153
Proposed	2.735	1.879	1.935	1.844

Table 13. Test Results of Medium-Length Documents Using Silhouette Coefficient

Medium-length document	*Silhouette Coefficient*				
	k				
	3	4	5	7	8
1	0.021324763	-1	0.15826037	0.37792194	-1
4	0.344005399	0.00196448	-0.3333333	0.02668924	-0.6
5	0.229692838	-0.5186944	0.01634193	-1	-0.4433674
8	0.334985563	0.05286649	-0.3333333	-1	-0.298982
9	0.399646901	-1	-0.0041046	-1	0.41286848
10	0.025295214	-1	-1	-0.3333333	-0.5
…	…	…	…	…	…
166	0.313898468	-1	0.67828406	-1	-1
168	-1	-1	0.04857357	-0.5	-0.455205
169	0.050811567	-0.3333333	-0.3333333	-0.0290072	-0.5
170	0.726650351	-1	0.0664035	-1	-0.446746
Average	0.159430301	-0.4817275	-0.3906301	-0.5491882	-0.5647329

Table 14. Test Results of Long-Length Document Using Silhouette Coefficient

Long-length document	*Silhouette Coefficient*					
	k					
	3	4	5	7	8	10
3	-0.0030230	0.490605	-1	0.236921	-1	0.0171970
25	0.3474346	0.371276	-0.02474	-0.08422	-0.06023	-0.5
27	0.3255959	0.044699	0.028883	-1	-0.05395	-1
29	0.4030755	0.470939	0.004107	0.489032	-1	0.2273476
30	0.7284273	0.01636	-1	-1	-1	-0.333333
31	-1	-0.46120	-0.00952	-0.25927	-1	-0.5
…	…	…	…	…	…	…
162	0.3164308	-1	-0.55778	-1	-1	-0.008911
163	0.0196518	-0.47422	-1	-0.53956	-1	0.0416299
164	0.6519629	-1	0.04441	-1	-1	-1
167	0.2796938	-1	0.298497	0.654852	-0.04458	-1
Average	0.1882973	-0.36410	-0.40582	-0.50424	-0.54624	-0.488164

3.3 Discussion

3.3.1 Feature-Based Method

The results of Feature-Based usage testing given in Figure 4 show that when using Feature Based, the number of paragraphs in the document during the selection process is reduced compared to that of the previous study.

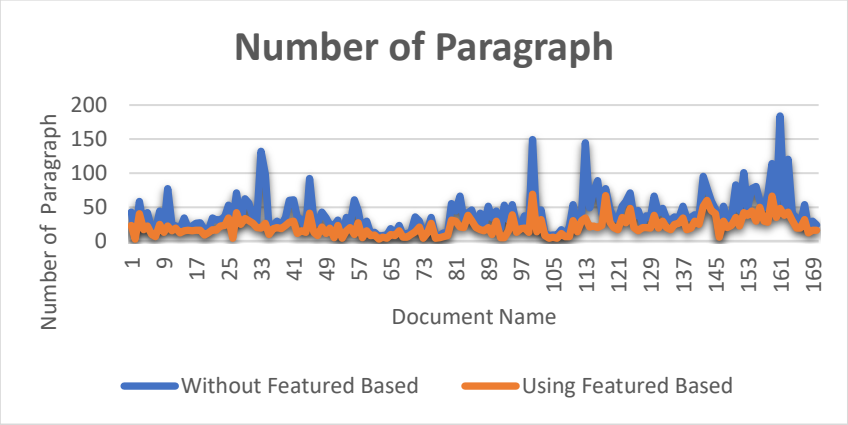

Figure 4. Comparison of Number of Paragraphs in Documents

3.3.2 Segment-based Clustering Process.

3.3.2.1 Cluster of Paragraph Process

The results of the test using the Silhouette Coefficient method in the previous research with the proposed research are shown in Figure 5 and Figure 6. This suggests that the proposed research obtained higher average value of Silhouette Coefficient if compared with the previous research, where the obtained values of Silhouette Coefficient in the proposed research on the k value of 2, 3, 5, 7 and 10 are better.

In Figure 5 and Figure 6, the peak value of Silhouette Coefficient is at the same k value. K value of short, medium, and long-length documents are 2, 3 and 3 respectively.

The results of the test using the Silhouette Coefficient method in the previous research with the proposed research are presented in Figure 7 and Figure 8. It shows that the proposed research scores lower than the previous studies. Based on the graphs from Sum Square Error resulting from both researches, the two did not get the elbow shape in accordance with the theory of Sum Square Error method. However, based on the previous research related to this evaluation method, it often occurs in the process of testing using the Sum Square Error method due to the determination of early centroid on clustering process using K-means algorithm.

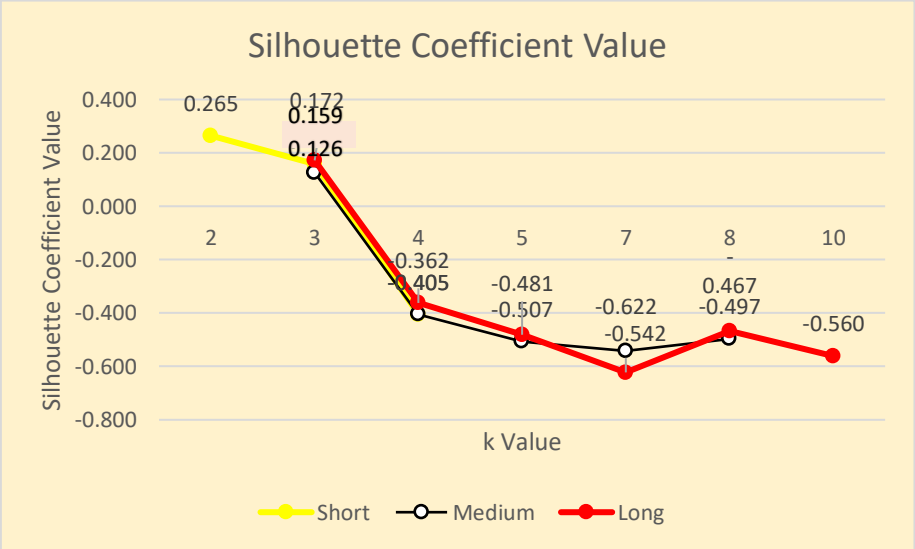

Figure 5. Values of Silhouette Coefficient on the Previous Research

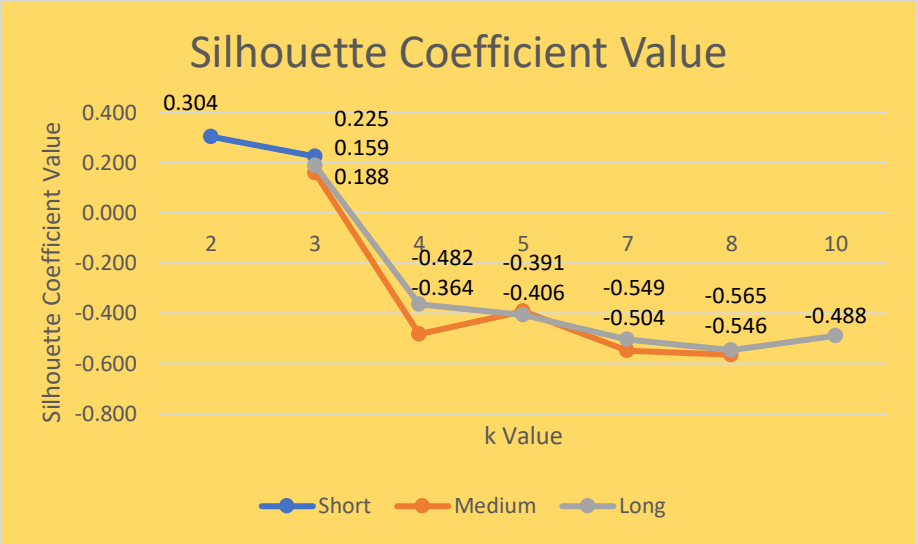

Figure 6. Values of Silhouette Coefficient on the Proposed Research

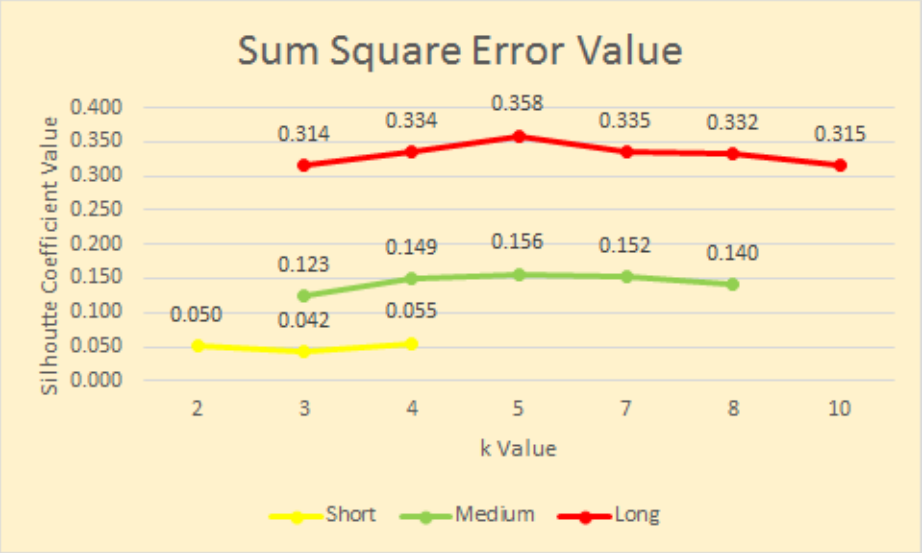

Figure 7. Values of Sum Square Errors on the Previous Research

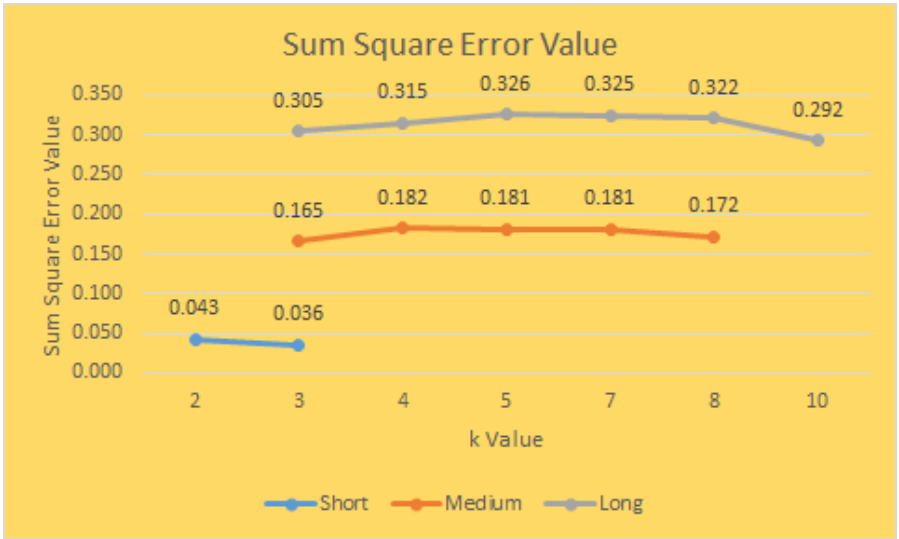

Figure 8. Values of Sum Square Errors on the Proposed Research

Therefore, testing the k value on the process of clustered paragraph cluster uses Silhouette Coefficient testing results. Based on the test, it has been assumed that the appropriate k values to use in clustering process of clustered paragraph for short, medium, and long-length documents are 2, 3 and 3 respectively.

3.3.2.2 Clustering Process of Clustered Paragraph

In testing clustered paragraph cluster, the results obtained are presented in Figure 9 and Figure 10.

Figure 9 shows that in the previous research, the value of Silhouette Coefficient of clustered paragraph cluster was higher than the proposed research. This may occur due to early centroid determination in the clustering process being not optimal; thus, it leads to the formation of less precise clusters.

Figure 10 shows that in the proposed research, the Sum Square Error score of clustered paragraph cluster is higher than that of the previous study. K value of 8 shows visible elbow point both in the previous research or the proposed one. At this point, there is a significant difference in the value of Sum Square Error between the k value of 5 and 8, compared to the difference of the k value of 8 and 10.

The test result uses Sum Square Error to determine the exact k value in clustering process of paragraph cluster. Based on Figure 10, it can be assumed that the exact k value is 8.

The process of paragraph selection using Feature Based gets better results in the selection of paragraphs, considered important based on the proximity of the paragraph with the title of the document.

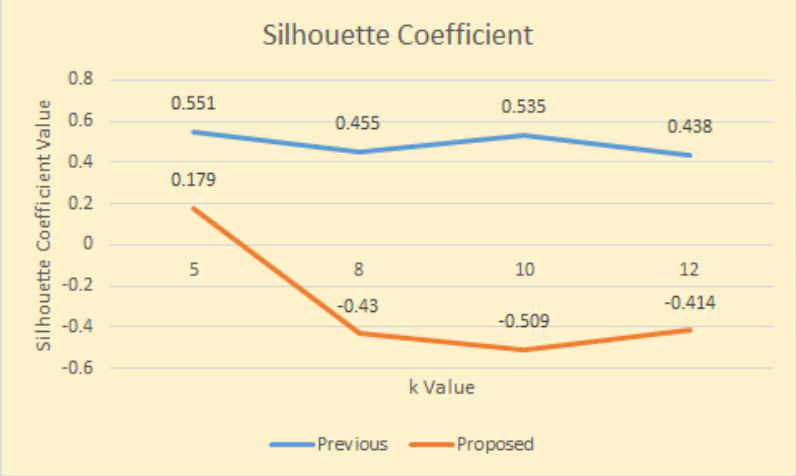

Figure 9. Values of Sum Square Error of Clustered Paragraph Cluster in the Previous Research

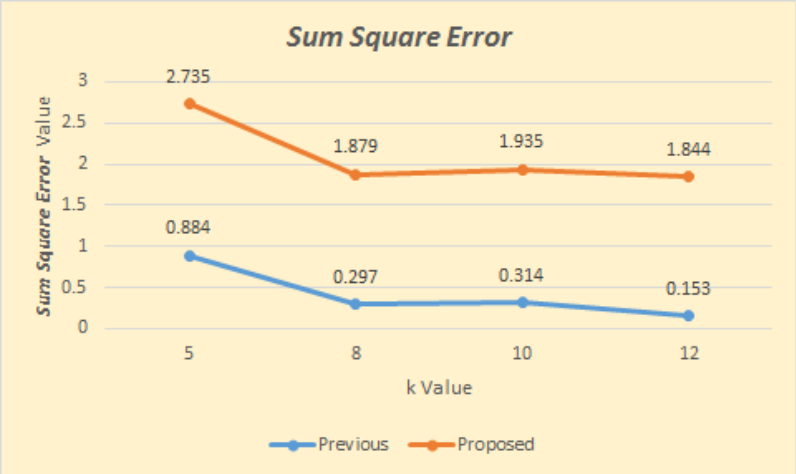

Figure 10. Values of Sum Square Error of Clustered Paragraph Cluster on the Proposed Research

4. Conclusion

The application of evaluation methods of Silhouette Coefficient and Sum Squared Errors indicates that the research proposal obtains better result from those of the previous research. It is shown by higher value of Silhouette Coefficient in paragraph clustering test scenario and the formation of elbow graph of the test on clustered paragraph cluster using Sum Square Errors evaluation method.

Meanwhile, elbow graph is not identified in the process of testing paragraph cluster using Sum Square Errors evaluation method. It can be caused by the early centroid determination on clustering using K-means.

References

[1] J. Brooke, and G. Hirst, *"Paragraph Clustering for Intrinsic Plagiarism Detection using a Stylistic Vector-Space Model with Extrinsic Features,"* Notebook for PAN at CLEF, 2012.

[2] J. Brooke, A. Hammond, and G. Hirst*, "Unsupervised Stylistic Segmentation of Poetry with Change Curves and Extrinsic Features,"* In CLfL@ NAACL-HLT, Pp. 26-35, June 2012.

[3] P. Shrestha, and T. Solorio, *"Using a Variety of n-Grams for the Detection of Different Kind of Plagiarism,"* CLEF, 2013.

[4] M. Jiffriya, M.A. Jahan, R.G. Ragel, and S. Deegalla, *"AntiPlag: Plagiarism Detection on Electronic Submissions of Text Based Assignments,"* Industrial and Information Systems (ICIIS) 8th IEEE International Conference, Pp. 376 – 380, Peradeniya: IEEE, 2013.

[5] A. Rosyadi, A.Z. Arifin, and & D. Purwitasari, *"Clusterization Based on Segment Using Paragraph to Identify Topic on Plagiarism Indication Detection,"* Inspiration Journal, Pp. 6(2), 2016.

[6] S. Ladda, N. Salim, and M.S. Binwahlan, *"Automatic Text Summarization Using Feature Based Fuzzy Extraction,"* Journal of Information Declaration, Pp.105-115, December 2008.

[7] H.P. Luhn, *"The Automatic Creation of Literature Abstracts: Advances in Automatic Text Summarization,"* Pp. 15, 1999.

[8] A. Tagarelli, & G. Karypis, *"A Segment-Based Approach to Clustering Multi-Topic Documents,"* Knowledge and Information System, Pp. 563-595, 2013.

[9] P.J. Rousseeuw, *"Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis,"* Computational and Applied Mathematics, 20: 53–65, 1987.

[10] N.P.E Merliana, & A.J. Santoso, "*Analysis of Best Cluster Number Determination on K-Means Clustering Method,"* Proceeding Sendi_U, 2015.