KINETIK, Vol. 3, No. 2, May 2018, Pp. 91-100
ISSN : 2503-2259
E-ISSN : 2503-22677

Paragraph Selection Methods Using Feature-Based on Segment-Based Clustering Process Using Paragraphs for Identifying Topics on Indications 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 data-set in the clustering process has a very important role. This is because the paragraph used as the data-set can affect the clustering result. In this research used the method of paragraph selection using feature-based, which aims to optimize the clustering process done in previous research. Based on evaluating using Silhouette Coefficient and Sum Square Errors evaluation methods to find the proper k value, it was found that with the use of Feature Based method, there was better result compared to the evaluation result from previous research.

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

1. Introduction

Plagiarism is an violates act the copyright or work of someone, It's occurs in various fields. One case of plagiarism often occurs in scientific articles.

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

From some of these studies no one has noticed the variations of the topics contained in a document, whereas 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, Arini (2015) uses segment-based clustering with the aim of identifying multiple topics in a set of documents [5]. However, this study has been found to be less than optimal, because the method of selecting paragraphs in each document is based only on the length of the paragraph, regardless of the essence of the paragraph in the document, so that it may be possible for paragraphs that have an essence but have an unsuitable length with a given threshold value, will not be included in the indication process of plagiarism indication.

So in 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 previous research, so it is assumed to improve 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].

While in this study using feature-based method in the proximity of a paragraph to the title of the document. 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
Paper submitted on November 16, 2017; Revision on December 15, 2017; Received January 03, 2018

2. Research Method

In the Rosyadi 2015 study there are several major stages being undertaken [5]: (1) segment-based clustering process aimed at identifying multi topics from the set of documents [8]. (2) identifying the topic using the weighting method tf-idf and tf-issf. (3) detect plagiarism indications using the Winnowing Algorithm, n-Gram, and hashing methods, shown in Figure 1.

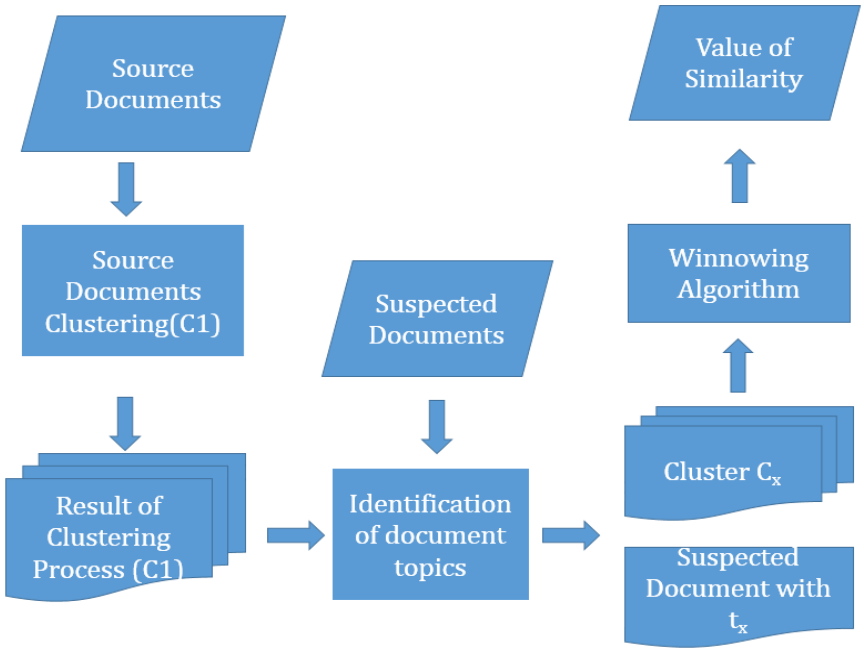

Figure 1. Previous Research Flow [5]

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 that have been generated are used in the clustering process using K-Means Algorithm. The process produces clusters containing segments that each cluster will call a segment-set. The process is called Cluster of Paragraph. The resulting segments are subsequently used as input to perform a segment-based clustering process using K-Means Algorithm. This process is called Cluster of cluster 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. This can be illustrated in Figure 3.

In this study the data-set used is the same data-set as the previous research data-set, which uses 170 journal documents with various topics as source documents or comparative documents. The documents are downloaded randomly.

Scenario testing is done by comparing the system performance from previous research with research proposal. In the process of testing the system carried out several tests including:

1. The effect of the number of paragraphs of each source document on the use of Feature Based.
2. Clustering process evaluate using the Silhouette Coefficient Method [9] and Sum Squared Errors [10], this test is performed to obtain the appropriate k value in the process:
   a) Cluster of Paragraph.
   b) Cluster of Cluster Paragraph.

3. Result and Discussion

3.1 Feature-Based Method

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

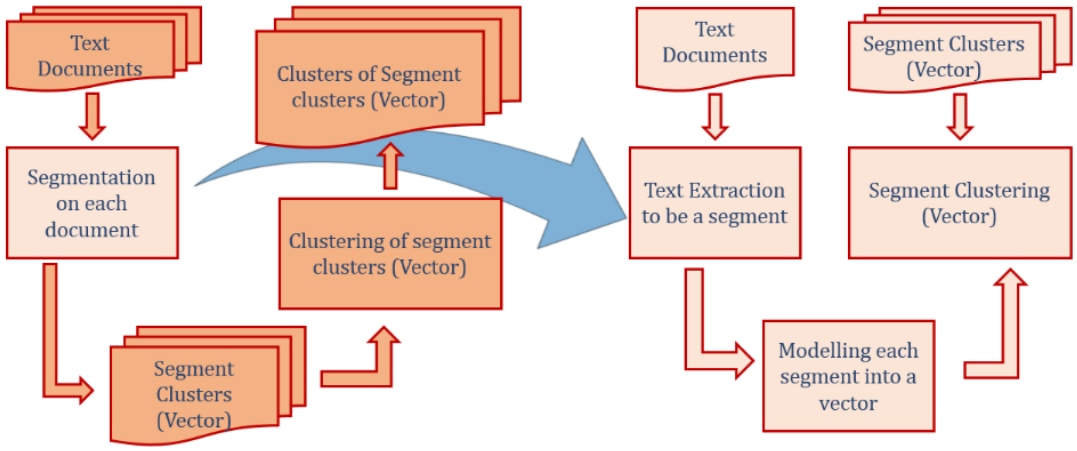

Figure 2. Document clustering process [5]

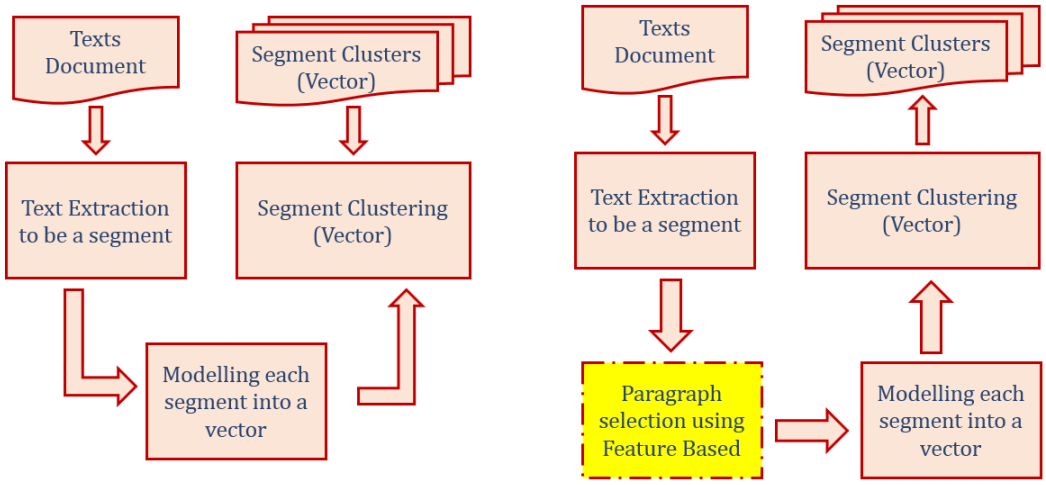

Figure 3. Proposed System: paragraph selection method using Feature Based on segment-based clustering process

Based on Table 1 it is known 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 or in percentage can be said every document has decreased the number of paragraphs on average by 42.26%.

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 Cluster of Paragraph.

The cluster of paragraph process is the clustering process of each source document based on the paragraph. Clustering process on each document aims to be able to know the topic of the source document. Previous research has assumed that in any document other than having a main topic can also have several sub-topics.

In the cluster of paragraph process, the clustering process is based on the number of paragraphs contained in a single source document, so in this process the grouping of source documents is done based on the number of paragraphs. The groupings are given in Table 2.

To evaluate segment-based clustering process are used the method of evaluation of Silhouette Coefficient and Sum Square Errors to be able to determine the proper value of k in clustering of cluster clustering process. each category of source document is evaluated against the value of Silhouette Coefficient and Sum Square Errors. The results of Silhoutte Coefficient testing on Short-Document can be seen in Table 3.

Table 2. The document category is based on the number of paragraphs on each document

Document Category	Range of amount paragraph	Amount of Documents	*Variation of k value* (Previous Research)	*Variation of k value* (Proposed Research)
Short-Document	$0 < p \leq 10$	60	2, 3, 4	2, 3
Medium-Document	$10 < p \leq 25$	70	3, 4, 5, 7, 8	3, 4, 5, 7, 8
Long-Document	$P > 25$	40	3, 4, 5, 7, 8, 10	3, 4, 5, 7, 8, 10

Table 3. Short-Document Test Results using Silhouette Coefficient

Short-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**

The results of Silhoutte Coefficient testing on Medium-Document can be seen in Table 4.

Table 4. Medium-Document Test Results using Silhouette Coefficient

Medium-Document	***Silhouette Coefficientt***				
	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**

The results of Silhoutte Coefficient testing on Long-Document can be seen in Table 5.

Table 5. Long-Document Test Results using Silhouette Coefficient

Long-Document	***Silhouette Coefficientt***					
	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**

Based on evaluating of each document category, then the average of each category is given for comparison with the evaluating result from the previous research. The comparison is given in Table 6.

Table 6. Comparison of Silhouette Coefficient values between prior and proposed research

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

The results of Sum Square Errors testing on Short-Document can be seen in Table 7.

Table 7. Short-Document Test Results using Sum Square Errors

Short-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 results of Sum Square Errors testing on Medium-Document can be seen in Table 8.

Table 8. Medium-Document Test Results using Sum Square Errors

Medium-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 Long-Document can be seen in Table 9.

Table 9. Long-Document Test Results 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 evaluating of each document category, then the average of each category is given for comparison with the evaluating result from the previous research. The comparison is given in Table 10.

Table 10. Comparison of Sum Square Errors values between prior and proposed research

Document Category	Sum Square Errors k=2	k=3	k=4	k=5	k=7	k=8	k=10	Research
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 Cluster of Cluster paragraph process is a clustering process that involves the output of a cluster of paragraph 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 done ten times on each evaluation method. So the test results obtained as in Table 11 and Table 12.

Table 11. Cluster of Cluster Paragraph evaluating using Silhouette Coefficient

Number of Experiment	Silhouette Coefficient k=5	k=8	k=10	k=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**

Table 12. Cluster of Cluster Paragraph evaluating using Sum Square Errors

Number of Experiment	Sum Square Errors k=5	k=8	k=10	k=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**

The evaluation results are comparable with testing from previous studies. The comparison results are given in Table 13 and Table 14.

Table 13. Comparison of Value of Silhouette Coefficient on Cluster of Cluster Paragraph 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 14. Comparison of Value of Sum Square Errors on Cluster of Cluster Paragraph 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

3.3 Discussion

3.3.1 Feature-Based Method

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

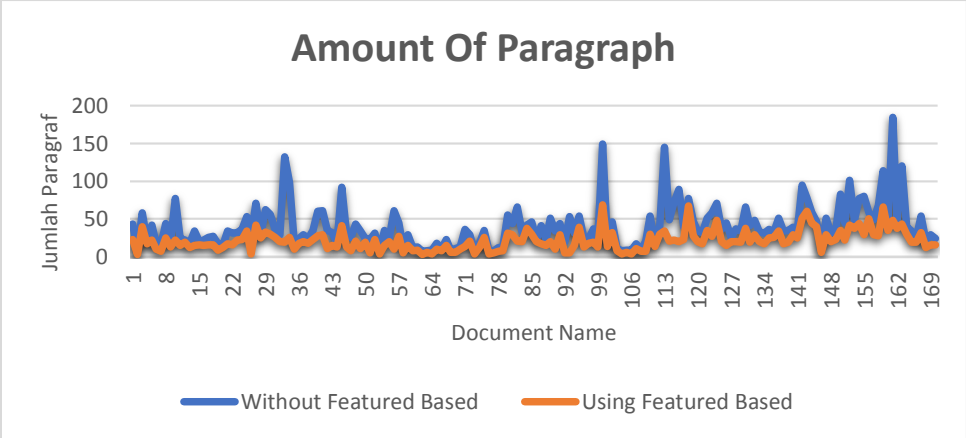

Figure 4. Comparison of Number of Paragraphs in Documents

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 previous research with the proposed research are shown in Figure 5 and Figure 6. This suggests that the proposed research obtained an average value of Silhoutte Coefficient higher if compared with the previous research, where the value of Silhoutte Coefficient obtained in the proposed research on the value of k equals 2, k equals 3, k equals 5, k equals 7, and k equal to 10 is better.

In Figure 5 and Figure 6 the peak value of Silhouette Coefficient is at the same k value, that is, on a Short-Document the value of k is at 2, the Medium-Document the value of k is at 3, the Long-Document of k is at 3.

The results of the test using the Silhouette Coefficient method in previous research with the proposed research are shown in Figure 7 and Figure 8. It's shows that the proposed research scored lower than previous studies. Based on the graphs from Sum Square Error resulting from both researches, both in previous research and research proposal, the two did not get the elbow shape in accordance with the theory of Sum Square Error method. However, based on research that has been done related to this method of evaluation, 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.

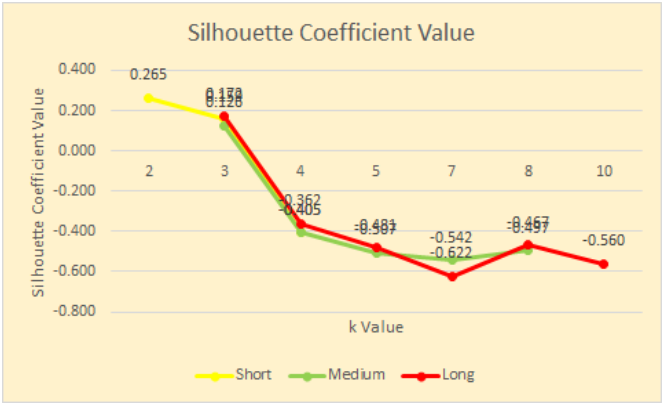

Figure 5. Value of Silhouette Coefficient Previous Research

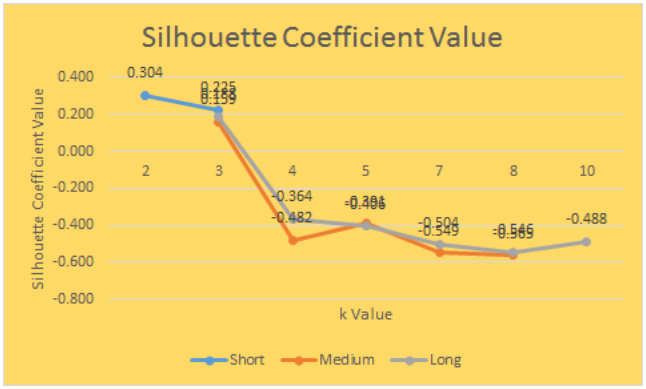

Figure 6. Value of Silhouette Coefficient Porpose Research

So in testing the value of k clustering process Cluster of Paragraph used Silhouette Coefficient testing results. Based on the test it has been assumed that the appropriate k value for use in Cluster of Paragraph clustering process is, the Short-Document using k equals 2, the Medium-Document using k equal to 3 and the Long-Document using k equals 3.

3.3.2.2 Cluster of Cluster Paragraph Process

In Cluster of Cluster Paragraph testing, the results obtained are shown in Figure 9 and Figure 10.

Figure 9 shows that in the previous research, the value of Silhouette Coefficient Cluster of Cluster Paragraph was higher than the proposed research. This may occur due to early centroid determination in the clustering process that is less than optimal, thus leading to the formation of less precise clusters.

Figure 10 shows that in the proposed research, the Sum Square Error Cluster of Cluster Paragraph score is higher than the previous study. At the value of k equals 8 visible elbow point or elbow either in previous research or proposal. At this point there is a significant difference in the value of Sum Square Error between the values of k equal to 5 and the value of k equals 8, compared to the difference of the k value equal to 8 and the value of k equals 10.

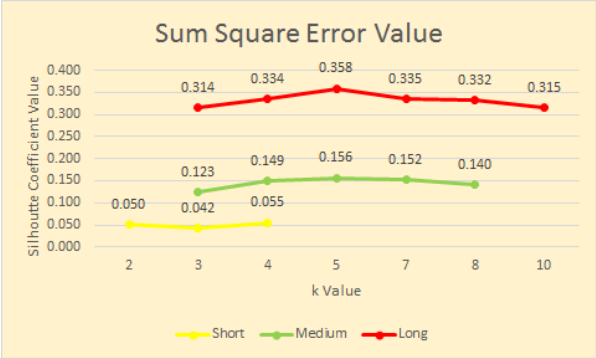

Figure 7. Value of Sum Square Errors Previous Research

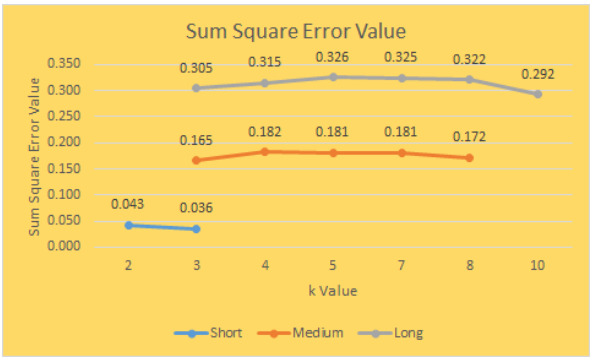

Figure 8. Value of Sum Square Errors Porposes Research

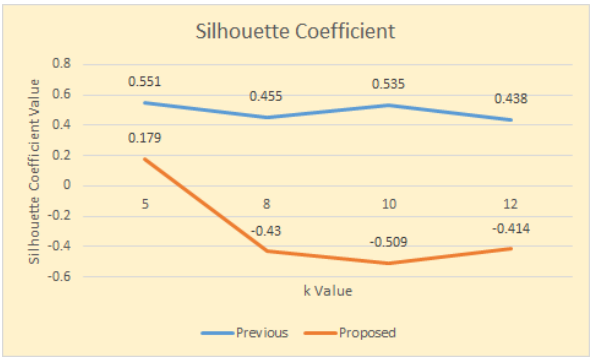

Figure 9. Value of Sum Square Error Cluster of Cluster Paragraph in previous research

So from this test, we use the test result using Sum Square Error to determine the exact k value in Cluster of Cluster Paragraph process. Based on Figure 10 it can be assumed that the exact k value is k equal to 8.

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

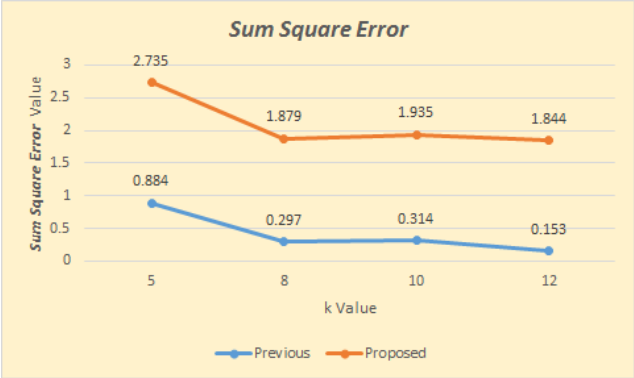

Figure 10. Value of Sum Square Error Cluster of Cluster Paragraph on the proposed research

4. Conclusion

The use of evaluation methods of Silhouette Coefficient and Sum Squared Errors, indicates that in the research proposal get better result from previous research. This is showed by the higher value of Silhouette Coefficient in Cluster of Paragraph test scenario and the formation of elbow graph on Cluster of Cluster Paragraph test using Sum Square Errors evaluation method.

While in the process of testing Cluster of Paragraph using Sum Square Errors evaluation method does not occur elbow graph. This can be due to the early centroid determination on clustering using K-Means

References

[1] Brooke, J., & Hirst, G. (2012). Paragraph Clustering for Intrinsic Plagiarism Detection using a Stylistic Vector-Space Model with Extrinsic Features Notebook for PAN at CLEF 2012.
[2] Brooke, J., Hammond, A., & Hirst, G. (2012, June). Unsupervised Stylistic Segmentation of Poetry with Change Curves and Extrinsic Features. In CLfL@ NAACL-HLT (pp. 26-35).
[3] Shrestha, P., & Solorio, T. (2013). Using a Variety of n-Grams for the Detection of Different Kind of Plagiarism. CLEF.
[4] Jiffriya, M., Jahan, M. A., Ragel, R. G., & Deegalla, S. (2013). AntiPlag: Plagiarism detection on electronic submissions of text based assignments. Industrial and Information Systems (ICIIS) 8th IEEE International Conference on (pp. 376 - 380). Peradeniya: IEEE.
[5] Rosyadi, A., Arifin, A.Z., & Purwitasari, D. (2016). Pengklasteran Berbasis Segmen Menggunakan Paragraf Untuk Identifikasi Topik Pada Deteksi Indikasi Plagiarisme. Jurnal Inspiration, 6(2).
[6] Ladda Suanmali, Naomie Salim, and Mohammed Salem Binwahlan, "Automatic TextSummarization Using Feature Based Fuzzy Extraction," Iurnal Teknologi Maklumat, pp.105-115, Desember 2008
[7] Luhn, H. P. (1999). The automatic creation of literature abstracts. Advances in automatic text summarization, 15.
[8] Tagarelli, A., & Karypis, G. (2013). A segment-based approach to clustering multi-topic documents. Knowledge and Information System, 563-595
[9] Peter J. Rousseeuw (1987). "Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis". Computational and Applied Mathematics. 20: 53–65.
[10] Merliana, N. P. E., & Santoso, A. J. (2015). Analisa Penentuan Jumlah Cluster Terbaik Pada Metode K-Means Clustering. Proceeding Sendi_U.