Text Summarization using K-Means Algorithm

Ari Firdaus [a,1], Novi Yusliani *[b,2], Desty Rodiah[b,3]

[a, b, c] Computer Science Faculty, Universitas Sriwijaya Indonesia
[1] af.arifirdaus@gmail.com*; [2] noviyusliani@gmail.com; [3] destyrodiah@ilkom.unsri.ac.id
* corresponding author

ARTICLE INFO

Article history
Received 29 May 2021
Revised 5 Juni 2021
Accepted 14 Juni 2021

Keywords
Text Summarization
Indonesian News
K-Means
Clustering
Extraction Method

ABSTRACT

The quantity of news on the internet is massively increasing. It is challenging nowadays to extract the information in the news article efficiently. Text summarization can solve this issue. Text Summarization is a process to produce a short form of text containing important information from the original text. Extraction method is one of summarization approach. This approach extracted important sentences in the original text to be the summary. In this study, a summary of the text was generated using the K-Means algorithm. This algorithm clustered all sentences in text into 2 clusters. By using sentence weight, the initial centroid was selected based on the sentence with the largest weight and the sentence with the smallest weight. All sentences in cluster with the highest weight will be the candidate of the text summary. Number of sentences used to be in text summary are 30% from the total number of sentences in the original text. Sentences with high weight picked to form the final summary. From the questionnaire given to 50 respondents, the final summary generated by K-Means algorithm can describe the important information of the original text.

1. Introduction

News is very useful for informing the public about the events that are happened around them. However, the rapid development of the internet has an impact on the amount of news and the amount of information that makes it difficult to obtain information efficiently and effectively [1]. Therefore, a tool is needed to summarize information in new article so that it is easy for reader to get information. This tool is called automatic text summarization. Automatic text summarization is a tool used to generate a concise form of a text that contains important information in it that is required by the user automatically [2].

Based on the number of documents provided as input, the text summarization is divided into two, namely, Single Document Summarization and Multiple Document Summarization. Single Document Summarization uses one document as input, while Multiple Document Summarization uses more than one document [3].

In general, there are two methods of performing automatic text summarization, namely extractive and abstractive [4]. The extractive method summarizes the text by selecting existing sentences in the text, while the abstractive method summarizes the text by making new sentences [2]. According to [4] there are several techniques that can be used for extractive method approaches, including: TF-IDF, cluster based, graph theoretical approach, machine learning approach, and k-means clustering. The extractive method approach often yields better results than the abstractive approach. This is because problems in abstractive approaches such as semantic representation, inference and natural language generation are relatively more difficult than data-based approaches such as sentence extraction. Therefore, extractive methods assisted by unsupervised learning methods are preferred in grouping text summarizing [2].

K-Means method is the simplest and most widely used method in grouping methods. In text summarization, K-Means is used to cluster sentences based on their weight. K-Means has the ability to group large amounts of data quickly and efficiently [5]. However, K-Means has the disadvantage that it relies on initial clustering. If the selection of Centroids in the grouping is not

appropriate, the grouping results will be locally optimal, thus enabling the much-needed initial Centroid.

This paper will create a text summarization system with an extractive approach using the K-Means algorithm with the initial centroid determined based on the sentence with the largest weight and the sentence with the lowest weight.

2. Literature Study

This chapter describes the theoretical basis about text pre-processing, weighting, and K-Means algorithm.

1. Text Preprocessing

Text preprocessing is a process to prepare unstructured data so that it can be used for further processing [6]. Preprocessing plays a very important role in text summarization. There are four processes done in this stage.

a. Sentence Segmentation

Sentence segmentation is the process of split all paragraphs in text into a set of sentences. Sentence segmentation is done based on punctuation in the form of periods, exclamation marks, and question marks.

b. Case Folding

Case Folding is a process to change the letter into lower case or upper case and removing all characters other than letters [7].

c. Tokenization

Tokenization is the process of splitting sentences in a text document into terms. In the tokenization process, terms or tokens are separated by spaces, line breaks, and punctuation marks [7].

d. Stemming

Stemming is a tool to change word into its root. In this study, the stemming process done by a library that has been provided on [8] [9]. These libraries use Nazief and Adriani algorithm [10].

2. Weighting

Weighting is a process to give a weight to all sentences in the text. This weight has an important role in K-Means algorithm. This algorithm clustered all sentences in text based on the weight. Each sentence consists of a set of words that have a weight. All of the word weights in a sentence are added together so that the sentence has a weight [11]. In this research, the weighting process was done by Term Frequency – Inverse Document Frequency (TF-IDF). TF-IDF is a simple weighting and very good in describing the importance of a word [2].

- TF (Term-Frequency) is used to calculate the frequency of a term appear in the text. The formula of term-frequency can be seen in (1). To calculate the term-frequency, frequency of a term appear in the text is divided by the length of the text (the number of words in the text).

$$Tf_t = f(t,d) \,/\, f(d) \quad (1)$$

  Tf_t is the Term-Frequency of term *t*, *t* is token or term, *d* is text or document, $f(t,d)$ is the frequency of a term appear in the text, and $f(d)$ is the number of words in the text.

- IDF (Inverse Document Frequency) is used to describe the importance of a term in a text. When using TF, all terms are considered equally important. However, some of word like "and", "at", and "to" will appear frequently in the text but they are not very important in describing content of a text. Therefore, the weight of a term must be reduced when it's not important. On the contrary, the term that is really important must be increased. The formula to calculate IDF is shown in the equation (2).

$$idf_t = log_{10}(N / f(t,d)) \quad (2)$$

idf is the Inverse Document Frequency of term *t* and *N* is the number of sentences in the document.

- Equation (3) describes the calculation of tf-idf weight for each word in a sentence [2].

$$Tf - idf = tf \times idf \quad (3)$$

3. K-Means Algorithm

K-Means algorithm is an unsupervised learning algorithm for clustering large data sets. This algorithm clustered the data based on the assumption that the number of clusters are fixed. The first step of this algorithm is determining *k* initial centroids (center points) for each cluster. The next step is to cluster the data into a cluster based on the distance between the data points and the centroid. When all data in data sets have become members of a cluster, recalculate the centroid value of each group. After having a new centroid, re-select cluster members as in the previous step. Repeat this step until the centroid value does not change anymore [2].

Based on the explanation above, steps done in K-Means algorithm are:

1. Determining the initial centroid by taking the largest weight of a sentence and the smallest weight of a sentence
2. Clustering sentences based on the closest distance to the centroid
3. Determine the new centroid by calculating the average member value
4. Repeat steps 2 and 3 until the centroids value do not change.
5. The members in the cluster with the largest centroid value are sorted based on their weight and then some of the members are selected to be summary of the text.

In this study, the cluster was divided into two, cluster one is cluster where all the members in that cluster will be candidate sentences in summary of the text and cluster two is a cluster where all the members will not be used in a summary of the text. The initial centroid value for the two clusters were done by taking the sentence that had the highest weight as a centroid value for the cluster one and the sentence with the lowest weight as a centroid value for the cluster two [4].

To calculate the distance of a data point with the centroid, the Euclidean distance equation can be used. This equation is shown in equation (4).

$$J(p,q) = |p - q| \quad (4)$$

Where p and q are the sentence weight and centroid value that will be calculated the distance.

4. Related Research

Previous research using the K-Means algorithm in text summarization has been conducted by [2,4]. In a study conducted by [2], they tried to solve the problem of text summarization in English texts. In this study, the preprocessing used was case folding and tokenization, while for word weighting, TF-IDF was used. This study used English news article as the data. In determining the number of clusters, this study used a rule. If number of sentences (*N*) were less than 20, the number of clusters were N-4 and for more than 20, the number of clusters were N-20. After determining the number of clusters, to determine the initial centroid of each cluster, this study used sentence weight randomly. This study generated a summary of the text is 35% - 50% of the size the original text.

Another study conducted by [4], they tried to solve the problem in summarizing Single Document and Multiple Documents for Bangali texts. In this study, the preprocessing used is tokenization, stop word removal, noise removal, stemming, and sentence splitting. For word weighting this study used TF-IDF. If the sentence contained cue word or skeleton word, the weight of the sentence was increased by one. Number of clusters used in this research were two clusters. In determining the initial centroid for the two clusters, this research selected sentence with the highest weight and the lowest weight. The cluster with the initial centroid of the sentence with the highest weight was chosen as the text summary candidate, then 50% of the sentences based on the highest score from the cluster were selected as the text summary. This study generated a text summary is 30% of the size the original text.

Another relevant research was conducted by [12]. They used TextRank algorithm to make a summary of a text. In this study, the text used was generated by using web scraping Beauty Soap with the Python programming language. The preprocessing used is only tokenization and for sentence weighting using TextRank. This study obtained an f-score of 0.84 for a summary length of 10%, an f-score of 0.64 for a summary length of 15%, an f-score of 0.67 for a summary length of 20%, and an f-score of 0.70 for a summary length of 25%.

3. Methodology

Automatic text summarization is a tool used to generate a concise form of a text that contains important information in it that is required by the user automatically by implementing a certain algorithm or method [2]. One of approach for generating text summary is extractive-based summarization. The extractive method generated the summary by selecting the sentences in the original text. The summary generated by this method is part of the original text without modification.

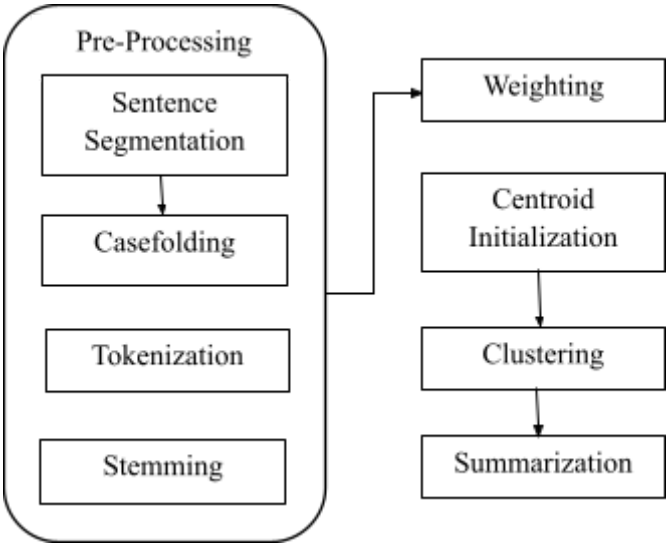

Fig. 1. Automatic Text Summarization Architecture

This study uses the clustering method which is one of the techniques in extractive-based text summarization. To get a summary of a text, the text was through the preprocessing phase first. Then, weighting all words in each sentence to generate the sentence weight. Weighting process done by TF-IDF method. After getting the sentence weight, next step is clustering using K-Means algorithm. Final phase is generated the summary. All of this process is shown in Figure 1.

Data used in this research was obtained from the Indonesian news online site, namely CNN Indonesia. The data used were 50 news articles. The average number of sentences were 17 sentences with the highest number of sentences 41 sentences and the smallest number of sentences 8 sentences. The tool used to get the data was a web scraper using cheerio.

4. Result and Discussion

The experiment was conducted by entering news links one by one into the software to see the results. The results taken are in the form of news content, summary results, document length, and summary length. Furthermore, the summary results and news content are compared and presented in the form of a questionnaire to see the feasibility of the summary results. Table 1 shows an example of the test. The results of the tests that have been conducted are presented in the form of tables which have been described before. Table 2 shows the test results in the form of a percentage of the number of summary sentences.

According to a survey conducted by the Pew Research Center, news readers age in range from 18 to 65 years. Therefore, the questionnaire was distributed to 50 random respondents with

different backgrounds such as students, employees, and housewives with an age range of 17 - 65 years who had filled out the questionnaire. Respondents make an assessment of the summary generated by the system by choosing a good or bad option. The results of the questionnaire can be seen in table 3. Based on table 2, the average length of the summary shown in table 4.

Table 1. Test Result Example

Document Link
https://www.cnnindonesia.com/teknologi/20210125091950-199-597903/lapan-duga-ledakan-buleleng-bali-efek-gelombang-meteor-besar
Document Summary
Lebih lanjut, Rhorom mengatakan sebagian besar meteor terbakar di atmosfer dan bisa jadi ada sebagian kecil yang tersisa dan jatuh ke permukaan bumi, darat atau laut. Fragmentasi meteor besar juga jamak terjadi ketika meteor tersebut mencapai ketinggian sekitar 100 kilometer di atas permukaan bumi. "Bila dibandingkan dengan kejadian di Bone, ada kemiripan sehingga diduga ledakan di Buleleng juga disebabkan adanya meteor besar yang jatuh," jelasnya. Rhorom juga menambahkan bahwa Meteor yang telah mencapai permukaan Bumi tidak berpotensi bahaya. Benda antariksa ini tidak mengandung unsur radioaktif yang membahayakan, mineral yang terkandung dalam meteor pun tidak berbahaya bagi lingkungan.

Table 2. Summary Length in Percentage

Document	Document length	Summary Length	Summary Length Percentage	Document	Document length	Summary Length	Summary Length Percentage
1	13	4	30.77 %	24	18	6	33.33 %
2	17	6	35.29 %	25	22	5	22.73 %
3	18	7	38.89 %	26	18	4	22.22 %
4	19	6	31.58 %	27	16	4	25.00 %
5	11	5	45.45 %	28	35	4	11.43 %
6	18	8	44.44 %	29	30	14	46.67 %
7	9	3	33.33 %	30	20	5	25.00 %
8	23	10	43.48 %	31	17	4	23.53 %
9	13	6	46.15 %	32	8	2	25.00 %
10	11	4	36.36 %	33	16	4	25.00 %
11	17	4	23.53 %	34	12	4	33.33 %
12	13	2	15.38 %	35	13	4	30.77 %
13	21	5	23.81 %	36	16	5	31.25 %
14	12	3	25.00 %	37	41	10	24.39 %
15	12	3	25.00 %	38	21	6	28.57 %
16	12	5	41.67 %	39	3	1	33.33 %
17	16	4	25.00 %	40	35	8	22.86 %
18	8	3	37.50 %	41	11	1	9.09 %
19	25	4	16.00 %	42	33	9	27.27 %
20	21	5	23.81 %	43	18	3	16.67 %
21	14	3	21.43 %	44	29	9	31.03 %
22	21	5	23.81 %	45	17	4	23.53 %
23	17	6	35.29 %	46	18	3	16.67 %
				47	16	3	18.75 %

Document	Document length	Summary Length	Summary Length Percentage
48	15	4	26.67 %
49	22	5	22.73 %
50	15	3	20.00 %

Table 3. Questionnaire Results

Document	Good	Bad
1	9	3
2	9	0
3	10	0
4	8	0
5	8	3
6	9	0
7	9	2
8	8	3
9	9	0
10	8	2
11	6	4
12	8	0
13	11	0
14	8	2
15	9	3
16	9	0
17	9	0

Document	Good	Bad
18	8	3
19	11	0
20	9	0
21	8	2
22	9	0
23	8	2
24	11	0
25	8	2
26	9	0
27	9	0
28	8	3
29	9	0
30	8	3
31	9	1
32	10	2
33	8	3
34	9	0
35	9	0

Document	Good	Bad
36	8	0
37	10	2
38	8	3
39	8	3
40	9	1
41	7	4
42	9	0
43	9	2
44	9	0
45	10	0
46	7	2
47	9	0
48	8	3
49	9	0
50	8	2

Table 4. The Average of Document Length and Summary Length

Document Length	Summary Length	Average Summary Length in Percentage
896	245	27.3%

The system succeeded in summarizing the article news with an average summary length of 27.3%. This result was obtained because the system chose the cluster with the highest score and then took half the members from that cluster. With an average summary length of 27.3%, users can read new article faster because they only need to read 27.3% of the entire article without losing any important information in the original article.

Table 5. Questionnaire Results

Result	Sum	Average
Good	435	87%
Bad	65	13%

Table 5 shows the percentage of summary goodness generated by the system based on questionnaire with 50 random respondents. Based on table 5, the summary generated by the system

gets 87% of good results and 13% of bad results. The results that are bad mostly come from new article that have a percentage of summary length smaller than the average percentage of summary length as in documents 41, 43, and 46. This summary does have a small percentage but the resulting summary is less accepted and made this bad summary. In addition, the summary results are bad due to:

1. There are article news that has a sentence with a period (not at the end of the sentence) which makes one sentence can be split into several sentences. As in document 48 with the sentence ""Yang paling penting di sini adalah pengawasan. Regulasi sebaik dan setegas apapun apabila tanpa pengawasan akan percuma," "kata Benny dalam keterangannya, Senin (25/1)." which is sentence in a quote is separated into two sentences by a period, after the word "pengawasan." The system extracted the first sentence in quote to be a summary that does not explain the contents of the new article.
2. The term that is not important becomes the term with the highest weight so that it affects the weight of the sentence. As in the 11th and 48th documents, the word "Yang" becomes the sentence with the highest score, so that the sentence with the term "Yang" has a high score even though the word "Yang" does not explain the contents of the new article.

Apart from that, the summary generated by the system gets pretty good results that are acceptable to users based on the questionnaire.

5. Conclusion

Based on the experimental result, text summarization using the K-Means Clustering method resulted in an average summary length of 27.3%. From the questionnaire distributed to 50 respondents, the summary text generated by the system is good and describing the content of the new article. Percentage of respondent that agreed the system is good in resulting text summarization is 87%. For the next research, adding stopword removal in pre-processing phase is needed to remove all terms that do not describing content of the text.

Acknowledgment

We thank to Informatics Engineering Department of Computer Science Faculty University of Sriwijaya for providing the environment and guiding in this research.

References

[1] J. Lewis, "News and the empowerment of citizens," *European Journal of Cultural Studies,* pp. 303-319, 2006.

[2] A. Agrawal and U. Gupta, "Extraction based approach for text summarization using k-means clustering," *International Journal of Scientific and Research Publications,* pp. 1-4, 2014.

[3] V. Pandya, "AUTOMATIC TEXT SUMMARIZATION OF LEGAL CASES: A HYBRID APPROACH," *arXiv preprint,* 2019.

[4] S. Akter, A. S. Asa, M. P. Uddin, M. D. Hossain, S. K. Roy and M. I. Afjal, "An Extractive Text Summarization Technique for Bengali Document(s) using K-means Clustering Algorithm," *IEEE,* 2017.

[5] B. K. Khotimah, F. Irhamni and T. Sundarwati, "A GENETIC ALGORITHM FOR OPTIMIZED INITIAL CENTERS K-MEANS CLUSTERING IN SMEs," *Journal of Theoretical and Applied Information Technology,* pp. 23-30, 2016.

[6] S. Mujilahwati, "Pre-Processing Text Mining pada Data Twitter," *Seminar Nasional Teknologi Informasi dan Komunikasi 2016,* pp. 49-56, 2016.

[7] M. R. Prathima and H. R. Divakar, "Automatic Extractive Text Summarization Using K-Means Clustering," *International Journal of Computer Sciences and Engineering ,* pp. 782-787, 2018.

[8] A. Librian, "sastrawijs," 18 Januari 2021. [Online]. Available: https://www.npmjs.com/package/sastrawijs.

[9] A. Librian, "damzaky/sastrawijs: Indonesian language stemmer. Javascript port of PHP Sastrawi project.," 18 Januari 2021. [Online]. Available: https://github.com/damzaky/sastrawijs.

[10] B. Nazief, M. Adriani, J. Asian, S. M. M. Tahaghoghi and H. E. Williams, "Stemming Indonesian," *Conferences in Research and Practice in Information Technology Series,* pp. 307-314, 2005.

[11] K. S. Jones, "A statistical interpretation of term specificity and its," *Journal of Documentation,* 1972.

[12] C. Mallick, A. K. Das, M. Dutta, A. K. Das and A. Sarkar, "Graph-Based Text Summarization Using Modified TextRank," *Soft Computing in Data Analytics Advances in Intelligent Systems and Computing,* pp. 137-146, 2019.