Indonesian News Extractive Text Summarization Using Latent Semantic Analysis

1st Rizka Ainur Rofiq
IT Department
Telkom University
Bandung, Indonesian
rizkarofiq@gmail.com

2nd Suyanto
IT Department
Telkom University
Bandung, Indonesian
suyanto@telkomuniversity.ac.id

2021 International Conference on Computer Science and Engineering (IC2SE) | 978-1-6654-0046-6/21/$31.00 ©2021 IEEE | DOI: 10.1109/IC2SE52832.2021.9792010

***Abstract*—News text is a text that contains important information that is happening to be disseminated to the public. In the news, the more information, the more text is displayed. Of course, it takes a lot of time to read the entire text of the news. Automatic text summarization is needed to help readers understand the content of the news text quickly. In this study, the application of the latent semantic analysis method with the GongLiu, Steinberger Jezek, and Cross techniques will be applied to automatic text summarization. The test data will be tested by using local news about politics. By comparing the rate the three methods previously mentioned, Gongliu is considered the best amongst the three methods since it has the highest Rogue value and the fastest processing time.**

Keywords: text summarization, semantic analysis, comparing rate

I. INTRODUCTION

Text summarization is a technique of processing a text document with a computer program to produce a summary that retains the essential essence of a text document [1]. In this globalization era, the use of text summarization is increasingly widespread. By using text summarization it is easier for someone to get information by just reading a text summary [6][7][21].

Text summarization in general has two text summarizing approaches which are classified into extractive and abstractive. Extractive summarization is a system's process of copying the main features of the original document and merging it into a shorter version. Abstractive summarization is generated from extracting the original document and then generated by adding new sentences that are different from the previous original document [6][9][22]. The extractive approach has the characteristics of simple text summarization. This simplicity makes the extractive approach more efficient than the abstractive approach [1][2][4].

Latent Semantic Analysis (LSA) is a method of sequence-based machine learning. The choice of LSA method for extractive summarization is because LSA can be used for term similarity machines based on hidden topics and can be used for clustering [5]. LSA is also well known for its success in summarizing using an extractive approach [2][5]. However, this method has a weakness in that its measurement is only based on the relationship between all sentences, which makes the larger the sentence the required supporting features that have greater intelligence [2][5][7]. Text summarization research was conducted by Xiong and Luo by comparing the Latent Semantic Analysis method and the Maximal Marginal Relevance method which results in the Latent Semantic Analysis method being better at summarizing text [14]. Another study conducted by Luthfiarta Zeniarja and Salam applied latent semantic analysis techniques to the document clustering process to improve accuracy with good results[15].

LSA has several ways to solve text summarization, one of which is the TF-IDF method [5][7]. TF-IDF assigns each word weighting value in the summarized document [2][10]. TF-IDF is often used in factor weighting or weighting in information retrieval [5][10][18].

According to Makbule Gulcin Ozsoy, Ilyas Cicekli, and Ferda Nur Alpaslan in the Gong & Liu summary algorithm, the first concept is selected, and then the sentence most related to this concept is selected as part of the resulting summary. Then the second concept is selected, and the same steps are executed. Repetition selects the concept and sentences most related to that concept continue until a predetermined number of sentences is extracted as part of the summary [11]. In the research of Makbule Gulcin Ozsoy, Ilyas Cicekli, and Ferda Nur Alpaslan, it was also explained that the SteinbergerJezek method works by choosing sentences that relate to all important concepts and at the same time choosing more than one sentence from an important topic [11].

Cross is an enhanced result of the SteinbergerJezek method. The cross method works by processing the summary results not only based on the similarity of words and sentences in a document but also the length of the sentence determines the success of increasing the precision value of the LSA F measurement in previous studies, that's what has been done [11]. Then, the next step taken by Jamhari, Noersasongko, and Subagyo is latent cross semantic analysis, which is a method that applies the extraction of hidden semantic meanings or called semantics in a text sentence [13].

Each summary should be scored to see how good the summary results are. Initially, summary evaluation is done manually concerning the human evaluation of the summary results. However, manual evaluation has some drawbacks, such as assessments that are not objective because everyone has different standards in assessing a summary. In addition, manual evaluation is associated with a lot of time and effort. To address this weakness, an investigation is conducted for the automatic implementation of the summary evaluation. And one of the results of this study is ROUGE, which is from ChinYew Lin[12][20]. Elements of ROUGE elements include:

A. Rouge N

ROUGE N is a unigram, bigram, trigram, and higher-order n-gram overlap between the candidate summary and the specified reference summary. Ngram in RougeN is the number of word lengths in an abstract candidate that will be compared to the reference to the reference to the reference [12].

B. Rouge L

Rouge L works by measuring a group of words using the longest match with LCS. The advantage of using LCS is that it does not require sequential matches, but sequential matches that reflect the order of words at the sentence level. Because it automatically contains general programs in the longest order, you do not need a predetermined program duration [12].

C. Rouge S.

With skip bigram metrics, these metrics are a collection of multiple pairs of words in a sentence. You can search for sequential words in reference text that appear in the output of the model but are separated by one or more words. ROUGE S allows us to add some leeway to our n-gram match[12].

Based on the research described above, this study presents a summary of Indonesian news documents using the Gongliu, Steinberger Jazek, and Cross method, which aims to create a system that can provide important information in summary form with quick results. Then the results of the text summarization will be compared with the three methods with the comparison parameters obtained from the Rouge score and the time during extractive automatic text summarization processing takes place.

II. RESEARCH METHOD

The Indonesian news text summary system developed in this study is designed to learn how to summarize Indonesian news texts from a series of sentences on an online news site.

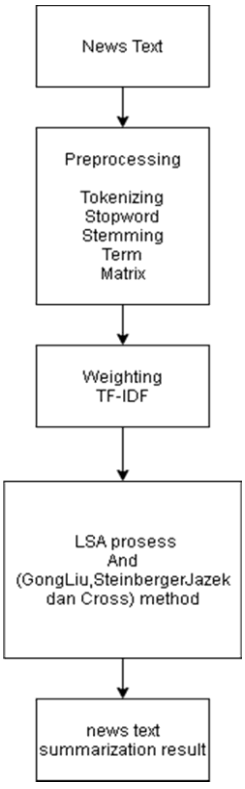

Fig. 1. The Indonesian news text summary system developed

The data that is used for testing is an article or news text in the Indonesian language for preprocessing. The purpose is to change the text of Indonesian news or articles from online news websites kompas, tribunnews, and liputan6, etc. Then it becomes data that will be processed in the next stage as many as 5 documents. The length of the summary is certainly shorter than the original text because it only takes the essence of the text and does not add text.

Preprocessing is the process of the early stages of processing summary text. The goal is to produce index terms so that they can be processed to the next stage in the TF-IDF processing stage and the reduced LSA method using SVD aims to clean the noise that exists during processing[18][19]. There are several steps involved in preprocessing. The first stage of tokenization is the process of cutting the input strings of each constituent word or separating every word in the document that runs through all sentences resulting from the parsing process[23]. At this stage, numbers, punctuation marks, and non-alphabetical characters are ignored because they are considered separators or word separators that do not affect the word processor. The case-sensitive process is also performed during the lexing phase by converting letters into lowercase letters. . It's a step toward converting filtered data into a basic word by eliminating the suffix that's on each word. In this process I used the python sastrawi library, removing the addition from the word and making a basic word to reduce the variation of the word with a word that also has the same root[25]. This is done to turn a word into a basic word. Furthermore, entering the stopword stage is the stage where the removal of some common or unimportant words that are better known as a list of unimportant words from the results of tokenization that have been done before[24]. From this process will be generated words that will then be used as terms. The process of removing unnecessary words is done by eliminating unimportant words obtained from keyword list data.

The next stage after preprocessing is the TF-IDF calculation which aims to give the weight value of each word in a text. The weighting is carried out on each sentence using an index term and weight will be generated from a word which will result in the number of times the word appears in the text. The following is the formula for calculating[16].

$$TFIDF\ (d, \mathrm{w}) = tf(d, \mathrm{w})\ x \log N\ /\ d\mathrm{fw} \qquad (1)$$

Tf (d,w) = frequency of occurrence of term w in the text
d, n = total number of text
dfw = number of documents containing term

For calculations, the LSA method is used to analyze the relationship of phrases in a set of texts. The goal is to show the similarity of the meaning of several words in the text[17]. The stages in the LSA algorithm are as follows:

- Matrix Input Creation
  The input matrix is the result of the TF-IDF calculation so that it can form a matrix.
- Singular Value Decomposition
- The function of SVD as a matrix modifier is simpler by becoming several matrix components. SVD is used to reduce noise to improve accuracy.

$$Amn = Umm\ \sum VnnT \qquad (2)$$

A = input matrix
U = matrix with dimension m xk
V = matrix with dimension k x n
$\sum$ = diagonal matrix with dimension k xk
m = number of rows of matrix
n = number of matrix columns

To determine the eigenvalues and eigenvectors of the N matrix:

$$N = AT\ A \qquad (3)$$

Information:
N = Square matrix dimension m or n
A = TF-IDF preprocessing input matrix
AT = Matrix transpose of input TF-IDF

III. Result And Discussion

TABLE I COMPARISON METHOD

Comparison Aspect		Method			
		Gongliu	***Steinberge Jezek1***	***Steinberge Jezek2***	***Cross***
Rouge 1	Precision	0.75	0.594771242	0.628571429	0.681818182
	Recall	0.829268293	0.739837398	0.715447154	0.731707317
	Fmeasure	0.787644788	0.739837398	0.669201521	0.666666667
Rouge 2	Precision	0.568627451	0.568627451	0.607142857	0.643939394
	Recall	0.804878049	0.707317073	0.691056911	0.691056911
	Fmeasure	0.764478764	0.630434783	0.646387833	0.666666667
Processing Time(ms)		0.308000002	2.1858	0.4587	0.5601
Processing Time(s)		0.000308	0.0021858	0.0004587	0.0005601

After trying automated text with the gongliu method, SteinbergerJezek and Cross used the Rouge comparison aspect. In each Rouge are displayed Precision, Recall, and FMeasure values. Precision values are calculated in almost the same way, but rather than dividing them by the number of n-grams the reference is better to divide them by the number of n-grams of the model. For value recall obtained from the results of the summary evaluation conducted by calculating the acquisition value and precision. The recall is the ratio of the portion of the number of sentences in the reference summary and the system summary, with the number of sentences in the reference summary. Precision is the comparison between the number of sentences in the reference summary and the system summary, with the number of sentences in the system summary, and F measure 0from the calculation of the combination of the calculation of the value of recall with the value of Precision. The three values aim to see the accuracy rate value in the automatic system summary and reference summary, the higher the value obtained, the better the level of accuracy of the text.

Then obtained from the results of comparisons of the four methods. In the Gongliu method obtained the highest Rouge calculation results Rouge1 precision value 0.75, recall 0.8292682926829268, and F Measure 0.7876447876447877 and Rouge 2 Precision 0.7279411764705882, recall 0.8048780487804879, and F Measure 0.7644787644787646478646. For the lowest Rouge calculation results obtained by the SteinbergerJezek method with Rouge 1 Precision value 0.5947712418300654, recall 0.7398373983739838 and F Measure 0.7398373983739838 and Rouge 2 Precision 0.5686274509803921, recall 0.7073170731707317 and F Measure 0.76444787644787646.

For time comparison parameters on the four methods, the time comparison method is done using the timeit library in python. The performance by simply calling the timeit library into the program and then automatically the program will process the time calculation done and automatic text processing in each method will be tested. From the results obtained, the Steinberger jezek method has the acquisition time of 0.00218579999999497224 seconds in processing. Thus Steinberger jezek method became the fastest process of summarizing text among the other three methods. And the cross method with a time gain of 0.00056010000000012846 seconds makes it the longest method of processing compared to the other three methods.

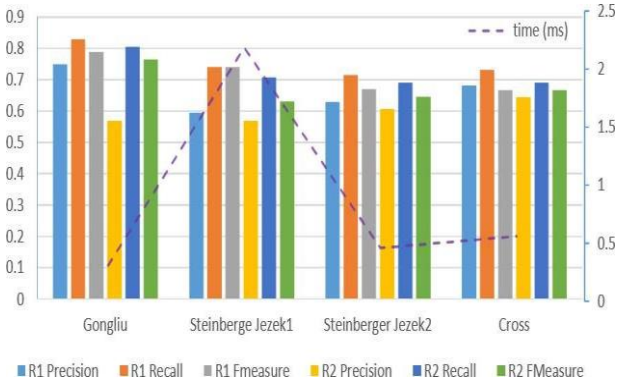

Fig. 1 Result Comparison Method

The chart shows that of the three methods tested, the Gongliu method is considered the best method among the other three methods because it has the highest Rouge value and the fastest process time. As shown in the chart above, the longer processing time does not always provide a better Rogue value. The longer processing time does not always provide a better Rogue value.

IV. Conclusion

More and more automatic text developments in the modern era. Research on automatic texting with different methods in contrast to the existence of this extractive method was developed into different methods to get optimal results in the detection of text. This study explained several methods of electronic texting using Latent Semantic Analysis. Some approach methods consist of gongliu, Steinberger Jazek and Cross methods.

The three methods described in the study were evaluated using Rouge and Rouge2 values and the time resulting from the process of each method in extractive texting. The results showed that the gong liu method has good Rouge1 and Rouge 2 values from the other three methods. For time comparison parameters the Steinberger Jazek method becomes the fastest in automatic text processing.

The evaluation of the comparison of each method using Rouge and the time done in this study is expected to help for the development of automatic texting in Indonesian more optimal in its commemoration.

Acknowledgment

We thank our friends and colleagues at Telkom University Computer Faculty for their support. In compiling this study, researchers did not forget to thank and praise and gratitude to all those who had helped researchers in completing this thesis for the graduation requirements of Strata One at the Faculty of Informatics, Telkom University. Researchers receive a lot of help from various parties both morally, support and materially. Therefore, the researcher would like to thank:

1) God for giving all gifts, graces and reinforcements so that researchers can complete the writing of this study well.
2) Mr. Dr. Suyanto, S.T., M.Sc as a thesis guidance lecturer who patiently guided researchers in completing this study.
3) My parents were patient and always encouraged me in doing this research.
4) Mr and Mrs of lecturers and staff of the Faculty of Informatics, Telkom University for all patience in educating and helping and willingly bothered by researchers during activities in lectures.
5) For all the writer's friends as a research hangout who has provided a platform for researchers to play, discuss, and eliminate fatigue

REFERENCES

[1] Ganiger, S. (2018). Algorithms for Single Extractive Document. 2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS), (Iciccs), 1284–1287.
[2] Geetha, J. K., & Deepamala, N. (2015). Kannada text summarization using Latent Semantic Analysis. 2015 International Conference on Advances in Computing, Communications and Informatics, ICACCI 2015, 1508–1512.
[3] Jain, A., Bhatia, D., & T hakur, M. K. (2018). Extractive Text Summarization Using Word Vect or Embedding. Proceedings - 2017 International Conference on Machine Learning and Data Science, MLDS 2017, 2018–January, 51–55.
[4] Varalakshmi K, P. N., & Kallimani, J. S. (2018). Survey on Extractive Text Summarization Methods with Multi-Document Datasets. 2018 International Conference on Advances in Computing, Communications and Informatics, ICACCI 2018, 2113–2119.
[5] Ardytha Luthfiarta, Junta Zeniarja, Abu Salam (2014). Integrasi Peringkas Dokumen Otomatis Dengan Algoritma Latent Semantic Analysis (Lsa) Pada Peringkas Dokumen Otomatis Untuk Proses. 13(1), 61–68.
[6] Pragantha, J., Informatika, T., Informasi, F. T., & Tarumanagara, U. (2017). Automatic Summarization Pada. 1(1), 71–78.
[7] Chowdhury, S. R., Sarkar, K., & Dam, S. (2018). An Approach to Generic Bengali Text Summarization Using Latent Semantic Analysis. Proceedings - 2017 International Conference on Information Technology, ICIT 2017, (1), 11–16.
[8] Mustaqhfiri, M., Abidin, Z., & Kusumawati, R. (2016). Peringkasan Teks Otomatis Berita Berbahasa Indonesia Menggunakan Metode Maximum Marginal Relevance. Matics.
[9] Merchant, K., & Pande, Y. (2018). NLP Based Latent Semantic Analysis for Legal Text Summarization. 2018 International Conference on Advances in Computing, Communications and Informatics, ICACCI 2018, 1803–1807.
[10] Saputra, Jerry. Fachrurrozi, M. Y. (2017). Peringkasan Teks Berita Berbahasa Indonesia Menggunakan Metode Latent Semantic Analysis (LSA) dan Teknik Steinberger & Jezek. Computer Science and ICT, 3(1), 215–219.
[11] MakbuleGulcinOzsoy, I1yasCicekli, FerdaNurAlpaslan. 'Text Summarization of T urkish Texts using Latent Semantic Analysis" Proceedings of the 23,d International Conference on Computational Linguistics (Co ling 20 I 0), Beijing, August 20 I 0, pp. 869-876
[12] Lin, Chin-Yew. 2004. " ROUGE: A Package for Automatic Evaluation of Summaries". Available on : 1 januari 2009
[13] Jamhari, M., Noersasongko, E., & Subagyo, H. (2014). PengaruhPeringkas Dokumen Otomatis Dengan Penggabungan Metode Fitur dan Latent Semantic Analysis (LSA) Pada Proses Clustering Dokumen Teks Berbahasa Indonesia. 1, 2355-5920.
[14] Xiong, S., & Luo, Y. (2015). A new approach for multidocument summarization based on latent semantic analysis. In Proceedings - 2014 7th International Symposium on Computational Intelligence and Design, ISCID 2014 (Vol. 1, pp. 177 180).
[15] Luthfiarta, A., Zeniarja, J., & Salam, A. (2013). Algoritma Latent Semantic Analysis (LSA) Pada Peringkas Dokumen Otomatis Untuk Proses Clustering Dokumen. Seminar Nasional Teknologi Informasi & Komunikasi Terapan 2013 (SEMANTIK 2013), 2013(November), 13 18.
[16] Okfalisa & Harahap, A. H., 2016. Implementasi Metode Terms Frequency-Inverse Document Frequency (TF-IDF) dan Maximum Marginal Relevance untuk Monitoring Diskusi Online. Jurnal Sains,Teknologi dan Industri, Volume 13, pp.151-159.
[17] Gunawan, F. E., Juandi, A. V., & Soewito, B. (2015). An automatic teks summarization using teks features and singular value decomposition for popular articles in Indonesia language. 2015 International Seminar on Intelligent Technology and Its Applications (ISITIA) (pp. 27-32). Surabaya: IEEE. doi:10,1109/ISITIA.2015.7219948.
[18] Evan, F. H., Pranowo, & Purnomo, S. Y. (2014). Pembangunan Perangkat Lunak Peringkas Dokumen dari Banyak Sumber Menggunakan Sentence Scoring dengan Metode TF-IDF. Seminar Nasional Aplikasi Teknologi Informasi (SNATI), 17-22.
[19] Martin, C. D., & Porter, M. A. (2012). The extraordinary SVD. American Mathematical Monthly. https://doi.org/10.4169/amer.math.monthly.119.10.838
[20] A. Romadhony, Z. R. Fariska, N. Yusliani, and L. Abednego, "Text Summarization untuk Dokumen Berita Berbahasa Indonesia," J. Telkom Univ., pp. 408–414, 2017.
[21] H. P. Luhn, "The Automatic Creation of Literature Abstracts," IBM J. Res. Dev., vol. 2, no. 2, pp. 159–165, 2010.
[22] E. Hovy and C.-Y. Lin, "Automated text summarization and the SUMMARIST system," p. 197, 1996.
[23] R. Feldman and J. Sanger, The Text Mining Handbook. 2006.
[24] M. W. Berry, Survey of Text Mining Clustering, Classification, and Retrieval Scanned by Velocity, vol. 32, no. 10. 2004.
[25] N. S. W. Gotami, Indriati, R. K. Dewi. (2018, September). Peringkasan Teks Otomatis Secara Ekstraktif Pada Artikel Berita Kesehatan Berbahasa Indonesia Dengan Menggunakan Metode Latent Semantic Analysis. Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer. 2(9), pp. 2821-2828. Available: http://jptiik.ub.ac.id/index.php/j-ptiik/article/view/2430/90