IEEE

Conference Number #41709

# Certificate

Presented to:

## DJOKO BUDIYANTO SETYOHADI

as

Presenter

*The 2nd International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE-2017)*

**Yogyakarta, Indonesia | 1-3 November 2017**

**General Chair of ICITISEE 2017**

**Arief Setyanto, S.Si., M.T., Ph.D.**

www.icitisee.amikom.ac.id

Organized by:

UNIVERSITAS AMIKOM YOGYAKARTA

Sponsored by:

IEEE INDONESIA SECTION

IEEE Xplore

# Summarizing Indonesian Text Automatically By Using Sentence Scoring And Decision Tree

Periantu Marhendri Sabuna
Magister Teknik Informatika
Universitas Atma Jaya Yogyakarta
Yogyakarta, Indonesia
hendra1602@gmail.com

Djoko Budiyanto Setyohadi
Magister Teknik Informatika
Universitas Atma Jaya Yogyakarta
Yogyakarta, Indonesia
djoko.bdy@gmail.com

***Abstract*— Text summarization is a process of compressing a text from the source to be a shorter version, but the version still contains the main information there. By reading the summary, the readers might be easy and fast to understand the contents instead of reading all the text. Because of that, it needs a method to understand, clarify, and present the whole information needed clearly and succinctly in the summary. So, it allows the readers save the time and energy. This research combining sentence scoring and decision tree method for automatic text summarization in Indonesian language. It uses the decision tree algorithm to choose which of sentences will be selected in summarization system. To produce the rules for decision tree, it uses 50 news texts as the training data. The produced-model from the training stage will be implemented for sentence selection process to the summarization system. The result shows the highest f-measure score is 0, 80 and the average is 0, 58. Based on this, it concludes that the result of document summarization using *sentence scoring* and *decision tree* shows a better accuracy score for news text document.**

***Keywords*— *text summarization; sentence scoring; decision tree;***

## I. INTRODUCTION

The increasing number of information media nowadays also results in the increasing information sources as well. This condition is called Information overload, a condition where someone's efficiency to use the relevant and beneficial information for work is obstructed by a number of information sources [1]. Because of that number of information sources, documents on internet are increasing as well; and one of them is news documents. News document is a collection of information about a number of important and recent events periodically. To understand the news document contents through reading the text summary takes shorter time than reading the whole document contents so that text summary becomes very important. Because of that, it needs a method to understand, clarify, and present the whole information needed clearly and succinctly in the summary.

There are two techniques of summarization text; they are extraction and abstraction. Extraction technique is a technique to copy the most important and informative part of the text to make a summary and abstraction technique is a taking main idea technique from a text source and later it makes the summary by creating new sentences to represent the main idea itself in other words [2]. The earliest research of automatic text summarization is started with term frequency method by Luhn in 1958 [3]. After that, there are several methods appeared; there are Latent Semantic Analysis [4], Machine Learning [2], Genetic Algorithm [5], Graph [6], Sentence Scoring [7], and Naive-Bayes [8].

In this research, our work focused in Indonesian language, the official language of Indonesia. Indonesia is the fourth most populous nation in the world. With over 230 million speakers, making it one of the most widely spoken languages in the world [8]. Researches of automatic text summarization for Indonesian are still in a small number. One research from Budhi et al., they implemented Graph and Algorithm Exhaustive method [9]. Later in 2012, Aristoteles et al. did a research and implemented Algorithm Genetic method [10]. And the next research, it used Latent Dirichlet Allocation and Algorithm Genetic method by Silvia et.al.[11]. These researches are good, but there are found a problem of each whose average of accuracy level is still below 55%.

To summarize text automatically, sentence weighting is one of the important parts. Fereirra et al., did an evaluation for 15 features of sentence weighting to determine the more optimal one[7]. The research's result showed that every feature has different influential level toward summary system result. Besides, a large number of features also causes the longer time for computation, so that it needs some features to shorten the system time for counting the weight of every sentences [7]. If it only takes a shorter time to count the weight of the sentences, it implicates the total of summarization time in each document. Because of that, the research will use 8 features for a relevant sentence weighting with the characteristics from a news text.

Text processing is about vague data processing. There are many method can be developed for that. An example, Rough clustering is powerfull method for vague data processing[12]. However, supervised classification is easier than clustering approach since a partitioned clustering require some initial method for better result[13]. Decision tree is one of the popular classification algorithm for text mining since it performs a general to specific search of a feature space, and it use a tree structure representation. ID3 and C4.5, a frequent used decision tree algorithms, have been introduced by J.R Quinlan which produce reasonable decision trees. According to[14], C4.5 algorithm is better for the accuracy and is faster to compute than ID3 algorithm.

978-1-5386-0658-2/17/$31.00 ©2017 IEEE

Based on the early research analysis, this research will combining sentence scoring and decision tree method for automatic text summarization in Indonesian language. It will use C4.5 algorithm to choose which sentences that will be included in the summary system. To produce the rules of decision tree, it uses 50 news texts as the training data. The produced-model in the training phase will be implemented in sentence selection process which later will be selected to the summary system. The summary will get a test for the accuracy level by measuring the precision values, recall, and f-measure, later it will be compared to the previous researches.

## II. LITERATURE REVIEW

### A. Automatic Text Summarization

Text summarization is a process of compressing a text from the source to be a shorter version, but the version still contains the main information there[15]. Other definition, based on [5], is an automatic process to create a short version from a text or document by choosing the most important part and to result a relevant summary. There are two criteria in automatic text summarization, they are extraction and abstraction. Extraction technique is a technique to copy the most important and informative part of the text to make a summary and abstraction technique is a taking main idea technique from a text source and later it makes the summary by creating new sentences to represent the main idea itself in other words [2].

### B. Related Researches

The earliest researches for automatic text summarization were started with term frequency method making by Luhn in 1958 [16]. In the same year, Baxendale did a research and added ideas to use sentence position feature for a document as one of the determining factor [3]. The next research was in 1969 by Edmunson, he accumulated the weight of term frequency, sentence position, title, and key phrases. Generally, the sentence that starts with key phrases becomes a good indicator for a significant content from a text document [11].

The research of automatic text summarization algorithm is developing and it can be classified into some methods. The other research uses TF-IDF (Term Frequency-Inverse Document Frequency) method with assumption that the term in a proportional document is inverse to document number in the term-contained corpus [2]. In 2008, Fattah & Ren did and succeeded a research to make a text summarization using 10 text weighting features with genetic algorithm technique.

For Indonesian text, there are exhaustive graph and algorithm method by Budhi, Intan, Silvia, and Stevanus that implement virtual graph concept. One of the process stages is using Term Frequency-Inverse Document Frequency (TF-IDF) and exhaustive algorithm to make the graph [9]. The other research is by Aristoteles, Herdiyeni, Ridha, and Adisantoso; they made automatic text summarizer in Indonesian language with genetic algorithm. There are eleven weighting features used, where all the used-features were to do genetic algorithm model training and to gain the proper weight combination for each component.

Ferreirra et al., 2013 researched sentence scoring method using 15 weighting features[7]. In this method, there are 3 approaches: (i) word scoring – to determine the weight of each most important words; (ii) sentence scoring – to determine the word weight by verifying the word weight as sentence position in a document, the similarity to the title, etc.; and (iii) graph scoring – to determine sentence weight by analyzing the relation inter sentences.

### C. Decision Tree

Decision Tree is one of the classification methods that use a tree structure representation, where each of the tree nodes represents the tested-attributes. Each branches is the tested-results division and each leaf nodes represents certain class groups [17].

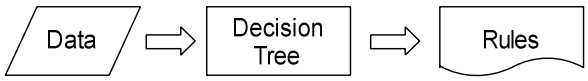


Fig 1. Decision Tree Concept

Figure 1 shows the process starts with classifying the random data to be decision rules. Generally, decision tree uses top-down searching strategy. The tree is built by dividing data recursively so that every part of the data comes from the same class [14]. C4.5 algorithm is one of the methods to make decision tree based on data training provided. C4.5 algorithm itself can solve the numeric data (continuous) and discrete.

## III. METHODOLOGY

The research itself uses four phases; the first phase is text documents collecting that will be used as the data training and testing. The next is training phase to produce a model or rule for using decision tree method. After it has a decision model, the next is testing phase to produce summarization system. And the last phase is evaluation, which to test the accuracy level between summarization system result and the manual summary.

### A. Document Collecting

The research also needs Indonesian text document intakes with file text-type document. So, it uses 50 national news text documents and the documents come from Harian Kompas online news which are the corpus of the research [10]. Each of documents has the manual summaries that are used to compare the result of summarization system.

### B. Training Phase

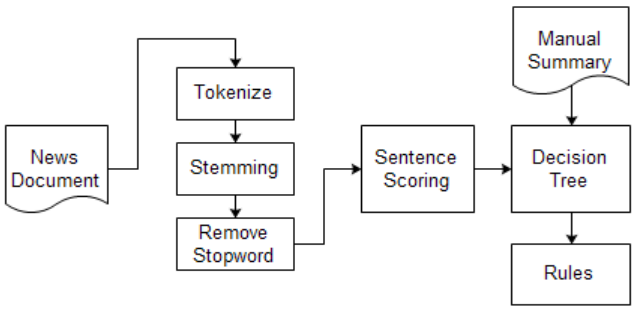


Fig 2. Training Phase

Figure 2 shows the data training phase is aimed to produce a decision model which consists of rules. In this phase, the first process is tokenize, where there is an input of news text. Tokenize is a process to delete punctuation in a text, and then to break the paragraph into sentences, and from the sentences, it can be easy to determine the weight of each sentence.

The next is stemming process to return the words to the basic form by deleting the affix, prefix, and suffix. After that, the process is to remove the stopword. Stopword is words which don't have meaning and is irrelevant. Usually, it is conjunction as and, that, at, from, and so on.

### C. Sentence Scoring

The next phase is sentence scoring, where every sentence will be given a score or weight based on 8 weighting features. There are TF/TDF, sentences with capital letter, sentences with verb, important phrases, number data, sentence length, sentence position, and the similarity between sentence and tittle. The explanation and the formula of each feature are provided above.

TF/IDF (F1)

The weighting is based on the number of term appearing or words in sentence (TF) and the number of term appearing or words in the whole sentences in the text (IDF).

$$TF(s,t) = \frac{term\ frequency(s,t)}{\max(term\ frequency(s,ti))}$$

$$Score\ f1(s,t) = TF(s,t) \times log\left(\frac{N}{sft}\right) \quad (1)$$

Uppercase (F2)

This method gives a higher weight for the words containing one or more capital letter, for example someone's name, city, country, and abbreviation. Below is the formula use:

$$CW(s) = \frac{number\ of\ uppercase\ words\ in\ s}{number\ of\ words\ in\ s}$$

$$Score\ f2(s) = \frac{CW(s)}{\max(CW(s))} \quad (2)$$

Proper Noun (F3)

The sentences containing more number of nouns get higher weight and have a tendency to be selected into the document summary. Nouns are like someone's name and place.

$$Score\ f3(s) = \frac{number\ of\ nouns\ in\ s}{number\ of\ words\ in\ s} \quad (3)$$

Cue Phrases (F4)

Generally, the sentence begins with phrases as "thus" and "investigation". It also emphasizes the phrases as "the best", "most important", "based on research", "especially", and other phrases that will be a good indicator for text document.

$$Score\ f4(s) = \frac{number\ of\ phrases\ in\ s}{number\ of\ phrases\ in\ doc} \quad (4)$$

Numerical Data (F5)

To summarize text, it needs to consider the numerical data in document; it is because usually the sentences with it give important information.

$$Score\ f5(s) = \frac{number\ data\ in\ s}{length(s)} \quad (5)$$

Sentence Length (F6)

A long sentence has higher weight. The length is counted based on the words total in a sentence times the number of average length in a document.

$$Score\ f6(s) = length(s) * average\ length(s) \quad (6)$$

Sentence Position (F7)

Sentence position is the position of a sentence in a paragraph. An assumption states that the first sentence in each paragraph is the most important sentence.

$$Score\ f7(s) = \frac{position(s)\ in\ paragraph}{number\ of\ sentences\ in\ doc} \quad (7)$$

Similarity to Tittle (F8)

Sentence that is similar to the document tittle is the same words appear in the sentence and the tittle.

$$Score\ f8(s) = \frac{keyword\ in\ s \cap keyword\ in\ title}{keyword\ in\ s \cup keyword\ in\ title} \quad (8)$$

### D. Decision Tree

Each sentence weight will be used as the training data of decision tree algorithm to produce decision model. Before doing data extraction in tree model, there are several process to shape the tree structure, there are:

a. Choose root based on the bigger gain ratio.
b. Choose the internal root or root branch based on the bigger gain ratio after deleting the chosen attribute as the root.
c. Repeat again till all the attributes are counted the gain ratio each of them.

Before looking for the gain score, firstly it needs to look for a chance of appearing a record in attribute (entropy). Mathematically, entropy score can be counted by this formula:

$$Entropy(S) = \sum_{i=1}^{n} - pi * log_2\ pi \quad (9)$$

Where S is the set of cases, n is the number of score in target attribute (class number), while pi is the number of sample in class i. From the formula above, we can look that if it is only two classes whose composition of samples is the same, thus the entropy score is zero. When we have got the entropy score, the next step is counting the information gain. Based on mathematic counting, information gain from A's attribute can be formulated as:

$$Gain(S,A) = Entr(S)\sum_{i=1}^{n}\frac{|S_i|}{|S|} * Entr(S_i) \quad (10)$$

Where S is the set of cases, A is attribute, n is the number of A's attribute partition, while i states a possible score of A's attribute and Si is the number of partition cases to i. The next is to count the gain ratio. Thing to know, there is a new term which is information splitting (SplitInformation) and it can be used as this formula:

$$SplitInformation(S,A) = -\sum_{t=1}^{c}\frac{S_i}{S}log_2\frac{S_i}{S} \quad (11)$$

Where S 1 to S c is c subset which is produced from S splitting. It is using A's attribute whose C score is many. Next, to count the gain ratio is by:

$$GainRatio(S,A) = \frac{Gain(S,A)}{SplitInformation(S,A)} \quad (12)$$

TABLE 1. DATA TRAINING EXAMPLE

| Sentence | f1 | f2 | f3 | f4 | f5 | f6 | f7 | f8 | Summary |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 0.1 | 0.2 | 0.0 | 0.0 | 0.1 | 0.3 | 0.5 | 0.1 | YES |
| 2 | 0.3 | 0.1 | 0.2 | 0.0 | 0.1 | 0.4 | 0.2 | 0.0 | YES |
| 3 | 0.4 | 0.2 | 0.1 | 0.2 | 0.2 | 0.2 | 0.3 | 0.3 | NO |
| 4 | 0.2 | 0.3 | 0.3 | 0.1 | 0.4 | 0.1 | 0.5 | 0.4 | YES |
| etc. | 0.1 | 0.4 | 0.0 | 0.0 | 0.1 | 0.4 | 0.1 | 0.0 | NO |

Table 1 explains that every sentence are the sample data, F1 to F8 are the attributes, and the summary columns are the attribute targets whose two attributes are Yes or No. Attribute Yes means the sentence appears in the summary, while attribute No means the sentence doesn't appear in the summary.

### E. Testing Phase

The next phase is data testing. In this phase, decision model is already produced by training process before for summarization system. Testing phase is described as:

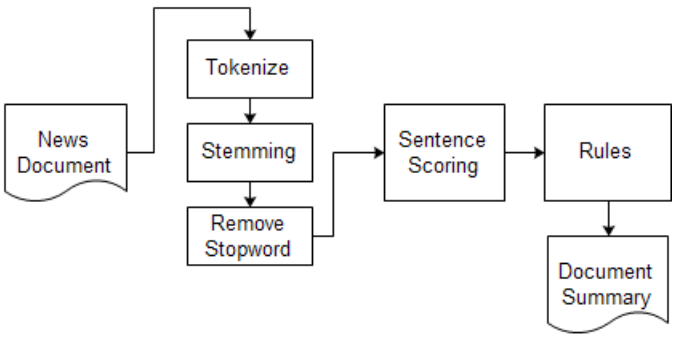


Fig 3. Testing Phase

### F. Evaluation

In this research, the evaluation is to measure the accuracy of the summarization system result and the manual summarization result. The testing of the summarization system uses precision method, recall, and f-measure. Precision method evaluates the accurate proportion to sentences in the summary, while recall is to evaluate the relevancy of sentence proportion of the summary [18].

$$Precision = \frac{S \cap T}{S}$$

$$Recall = \frac{S \cup T}{T}$$

$$F - measure = \frac{2 * Precision \times Recall}{Precision + Recall} \quad (13)$$

Where T is text which consists of manual summarization result sentences and S itself is the text of summarization system result.

## IV. RESULTS AND DISCUSSION

### A. Data Analysis

In the phase of making decision model, it needs a number of data as the training data. In this research, there are 50 text documents from the corpus data as the data training. The analyzed document text structure consists of news tittle, news contains, and the manual summarization result. The extraction result shows there 1237 sentences with 363 sentences are included in the summary and other 910 sentences are not. Using the data, the next step is weighting each sentence using 8 weighting features that has been explained in Methodology. The result is in the table.

TABLE 2. STATISTIC OF SENTENCE WEIGHTING RESULT

| Features | MIN | MAX | AVERAGE |
|---|---|---|---|
| TF/IDF | 0,00 | 1,85 | 0,28 |
| Uppercase | 0,00 | 0,82 | 0,17 |
| Proper Noun | 0,00 | 1,00 | 0,22 |
| Cue Phrases | 0,00 | 1,00 | 0,04 |
| Numeric Data | 0,00 | 1,00 | 0,03 |
| Sentence Length | 0,00 | 0,23 | 0,04 |
| Sentence Position | 0,10 | 1,00 | 0,69 |
| Similarity to tittle | 0,00 | 1,43 | 0,15 |

Table 2 shows the sentence weighting result statistics using 8 features, with minimal, maxima, and average score of each. Next, the result will be classified into some classes to make the decision tree.

TABLE 3. WEIGHT CLASSES

| Class | Criteria |
|---|---|
| Low | < 0,10 |
| Small | 0,11 - 0,30 |
| Medium | 0,31 - 0,50 |
| Big | 0,51 - 0,70 |
| High | > 0,71 |

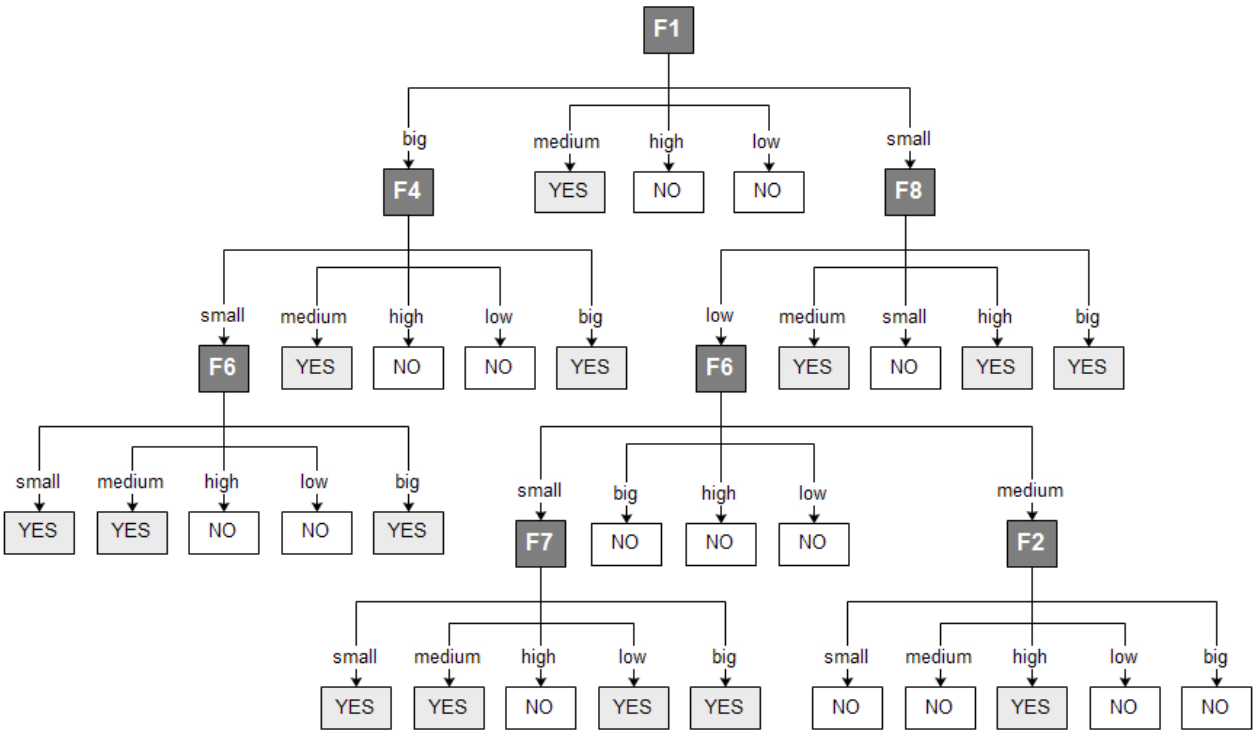


Fig 4. Decision Tree Result

Table 3 shows there are 5 weight classes; there are low, small, medium, big, and high. After identifying the class, training data is ready for the next process which is decision tree making.

### B. *Decision Tree Result*

To produce decision tree, it uses C4.5 algorithm to process the training data. Based on the test result, the decision tree can be looked as the figure above. Figure 4 shows the result of decision tree where feature F1 is the root. After decision tree has been shaped perfectly, the next step is rules making.

### C. *Evaluation*

The rules are used for the text summarization system to that 50 news documents. The summarization system result will be compared to the manual summary by counting the precision, recall, and f-measure. The result is in table 4 and figure 5.

TABLE 4. TEXT SUMMARIZATION TEST RESULTS

| Test | Attribute | MIN | MAX | AVERAGE |
|---|---|---|---|---|
| Source text | Number of Sentence | 11 | 94 | 25 |
| | Number of Words | 205 | 1527 | 463 |
| Manual summary | Number of Sentence | 3 | 31 | 7 |
| | Number of Words | 62 | 619 | 161 |
| System summary | Number of Sentence | 2 | 37 | 8 |
| | Number of Words | 61 | 699 | 168 |
| Evaluation | Precision | 0,38 | 0,88 | 0,54 |
| | Recall | 0,40 | 0,92 | 0,66 |
| | F-measure | 0,46 | 0,80 | 0,58 |

Table 4 shows that the average of the number of sentences and the number of words on the manual summary is not too different from the system summary. The average of the total sentences is 8 on the system summary and 7 on manual summary. While the average of the total words on the system summary is 168, manual summary gets 161 words.

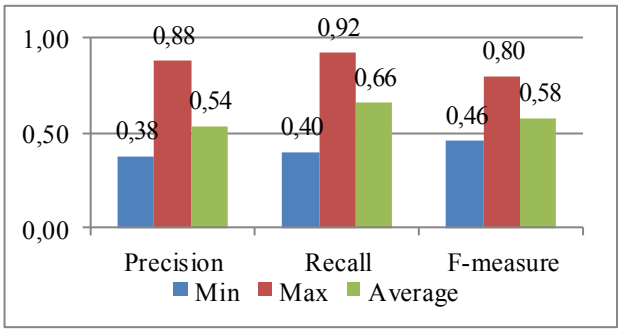


Fig 5. Text summarization test result graphic

Figure 5 shows the using of *sentence scoring* method and *decision tree* on Indonesian text summarization gets the results, where the precision average is 0,54, the recall is 0,66, and the f-measure is 0,58. The higher f-measure score is 0, 80 and the lowest is 0, 46.

According to the test results and discussion, the highest f-measure score is 0, 80 and the average is 0, 58. Based on this, it concludes that the proposed method shows a better accuracy score for text summarization on Indonesian language. Research from Aristoteles, Herdiyeni, Rida, and Adisantoso used genetic algorithm method and produced 0, 47 f-measure score. The other research from Silvia, Rukmana, Aprilia, Suhartono, Wongso, and Meilina, they used latent dirichlet allocation method and got 0, 55 f-measure score.

## V. Conclusions and Further Work

In this research, we have successfully combined the sentence scoring method and decision tree for the summarization of Indonesian text. Sentence scoring method used to generate weights in each sentence based on 8 text features are TF / IDF, uppercase, proper noun, cue phrases, numerical data, sentence length, sentence position, and similarity to tittle. Decision tree method is used to generate decision model or rule based on existing training data. From the rules that have been created, applied to select the sentences are important so as to generate a summary automatically.

For the next research, it is suggested to develop a score measuring formula for a better text feature. The hope is when many features have variety scores, new invention of better rule model will be found. The proposed method in this work can also be used as the basic to developing algorithms for multiple document summarizations for the Indonesian language. In addition, the document corpus and manual summary in the Indonesian language should also be made by professionals to achieve the standardization of testing and evaluation of text summarization algorithms.

## *References*


[1] David Bawden and Lyn Robinson, "The dark side of information: overload, anxiety and other paradoxes and pathologies," Journal of Information Science, 2008, 35 no 2, pp. 1-12,.

[2] Mahak Gambhir, Vishal Gupta,. Recent automatic text summarization techniques: a survey. Artificial Intelligence Review, 2017, 47.1: 1-66.

[3] Karel Jezek and Josef Steinberger, "Automatic Text Summarization ( The state of the art 2007 and new challenges )," Znalosti, 2008, pp. 1-12.

[4] Jen-Yuan Yeh, Hao-Ren Ke, Wei-Pang Yang, and I-Heng Meng, "Text summarization using a trainable summarizer and latent semantic analysis," Information Processing and Management, 2005, vol. 41, pp. 75-95,.

[5] Mohamed Abdel Fattah and Fuji Ren, "GA , MR , FFNN , PNN and GMM based models for automatic text summarization," Computer Speech and Language, 2009, vol. 23, pp. 126-144.

[6] Yogan Jaya Kumar and Naomie Salim, "Automatic Multi Document Summarization Approaches," Journal of Computer Science, vol. 8, pp. 133-140, 2012.

[7] Rafael Ferreira et al., "Assessing sentence scoring techniques for extractive text summarization," Expert Systems with Applications, 2013, 40, pp. 5755-5764,.

[8] Ahmad Najibullah, "Indonesian Text Summarization based on Naïve Bayes Method," Proceeding of the International Seminar and Conference 2015: The Golden Triangle (Indonesia-India-Tiongkok) , 2015, pp. 67-78.

[9] Gregorius S. Budhi, Rolly Intan, Silvia, and Stevanus , "Indonesian Automated Text Summarization," Proceeding 1st International Conference on Soft Computing, Intelligent System and Information Technology, 2007.

[10] Aristoteles, Yeni Herdiyeni, Ahmad Ridha, and Julio Adisantoso, "Text Feature Weighting for Summarization o f Documents in Bahasa Indonesia Using Genetic Algorithm," IJCSI International Journal of Computer Science Issues, 2012, vol. 9, no. 3, pp. 1-6.

[11] Silvia, Rukmana P., Aprilia V.R., Suhartono D., Wongso R., Meiliana, "Summarizing Text for Indonesian Language by Using Latent Dirichlet Allocation and Genetic Algorithm," Proceeding of International Conference on Electrical Engineering, Computer Science and Informatics, 2014, pp. 148-153.

[12] Djoko Budiyanto Setyohadi , Azuraliza Abu Bakar , Zulaiha Ali Othman , " Optimization Overlap Clustering Based On The Hybrid Rough Discernibility Concept and Rough K-Means," Intelligent Data Analysis, 2015, vol. 19, no. 4, pp. 795-823.

[13] Djoko Budiyanto Setyohadi , Azuraliza Abu Bakar ,Zulaiha Ali Othman, "An Improved Rough Clustering Using Discernibility Based Initial Seed Computation," Data Mining and Applications, 2010, vol. 6440, pp 161-168

[14] Badr Hssina, Abdelkarim Merbouha, Hanane Ezzikouri, and Mohammed Erritali, "A comparative study of decision tree ID3 and C4.5," (IJACSA) International Journal of Advanced Computer Science and Applications, Special Issue on Advances in Vehicular Ad Hoc Networking and Applications 2014, 2014, pp. 13-19.

[15] Nabil Alami, Mohammed Meknassi, and Noureddine Rais, "Automatic Texts Summarization : Current State Of The Art," Journal of Asian Scientific Research, 2015, vol. 5, no. 1, pp. 1-15.

[16] H.P Luhn, "The Automatic Creation of Literature Abstracts," IBM Journal, 1958, pp. 159-165.

[17] Chih-Chiang Wei and Jiing-Yun You, "C4.5 Classifier for Solving the Problem of Water Resources Engineering," Proceeding of the International Conference on Advanced Science, Engineering and Information Technology 2011, 2011, pp. 664-667.

[18] Rajesh Shardanand Prasad and Uday Kulkarni, "Implementation and Evaluation of Evolutionary Connectionist Approaches to Automated Text Summarization," Journal of Computer Science, 2010, vol. 6, no. 11, pp. 1366–1376.