ISSN 2047-3338

Text Feature Weighting for Summarization of Documents Bahasa Indonesia by Using Binary Logistic Regression Algorithm

Aristoteles[1], Widiarti[2] and Eko Dwi Wibowo[3]
[1,3]Department of Computer Science, University of Lampung, Indonesia
[2]Department of Mathematics, University of Lampung, Indonesia
[1]aristo_tole@yahoo.com, [2]widiarti08@gmail.com, [3]edbowoo@gmail.com

***Abstract*— The research was conducted the text feature weighting on Indonesian text by using binary logistic regression algorithms. Features of the text using text features eleven [1]. Eleven text features used are sentence position, positive keyword negative keywords, similarity between sentences, sentences that resemble the title sentence, sentences containing names of entities, sentences that contain numeric data, length of sentence, the connection between sentences, the sum of the weight of the connection between sentences, and sentence semantics. The purpose of this research was to conduct the optimization of summarization text by using binary logistic regression algorithm and the influence of the eleven features text by using binary logistic regression algorithms. Binary logistic regression algorithm used in compression rate 30%. The results of this research show the accuracy of compaction on the 30% compression rate amount 91.1% and on "positive keyword (f2)" can represent the eleven text features to perform compaction of text.**

***Index Terms*— Binary Logistic Regression Algorithm, Compression Rate and Text Features**

I. INTRODUCTION

UNDERSTANDING the content of a long writing is not an easy thing. Thus, making a summary of writing, copying writing in another language or certain papers to know its point.

A summary of the content of the decision and restate sources [2] with no more than half of the original text [3] and has a proportionally comparison between sections are summarized in the summary [4]. The simple concept of the summary is taking an important part that describes the overall content of the document. In making a summary there are two manufacturing techniques that can be done, they are the abstractive and extractive.

Abstraction is a transformation of a sentence into shorter sentences and new sentences that are not in the original document or in different words, while extractive is producing a summary by selecting some of the sentences in the original document [5].

There are many techniques that have been used in automatic summarizing text such as word frequency statistical approaches [6], the position of the sentence [7], mathematical regression [8].

Past research involving ten text features to create a summary with a regression algorithm [8,9], but does not involve semantic features of the sentence. The results of research show that the accuracy of summarization by using text features ten is 42.84% [9]. Other studies by using semantic sentence, but only using genetic algorithms [1]. The results showed that the accuracy of summarization using eleven text features is 47.63% and four features, namely positive keywords (f2), the similarity between sentences (f4), which resembles the title phrase (f5), and sentence semantics (f11) has been able to represent eleven features [1]. Those results carried out a summary of the text using eleven features [1] by using binary logistic regression algorithm to determine the effect of these eleven features.

II. RELATED WORK

Text Feature Extraction: The processing of a sentence using the eleven features resulting values that will be used to summarize the documents. Eleven of these features is the sentence position, positive keyword, negative keyword, similarity between sentences, sentences that resemble the title phrases, sentences containing names of entities, sentences that contain numeric data, length of sentence, the connection between sentences, the sum of the weight of the connection between sentences, and sentences semantics.

Position sentence: numbering the sentences in sequence in a single paragraph. Suppose s is a sentence in the original text of the paragraph, X is the position of the sentence within a paragraph (e.g. there are five sentences in a paragraph, so the first sentence has five positions, the second sentence has four positions, etc.) N is the total number of sentences in a paragraph. So that the position of the sentence can be formulated as follows:

$$score_{f_1}(s) = \frac{X}{N} \quad (1)$$

Positive keyword: word in the original sentence that appears in the sentence of the summary. The calculation of positive keyword feature can be seen from the occurrence of the word in the text document and summary document. Therefore, it can be formulated as follows:

$$score_{f_2}(s) = \frac{1}{length(s)} \sum_{i=1}^{n} tf_i * P(s \in S | keyword_i) \quad (2)$$

Assume s is a sentence in the original text, S is a sentence in the summary, f_1 is positive text features keyword, n is the number of keywords in sentences, tf_1 is the number of keywords to-i, which appears in the sentence.

$$P(s \in S | keyword_i) = \frac{P(keyword_i | s \in S) P(s \in S)}{P(keyword_i)} \quad (3)$$

$$P(keyword_i | s \in S) = \left(\frac{sentence\ in\ the\ summary\ \&\ contain\ keyword_i}{(sentence\ in\ the\ summary)} \right) \quad (4)$$

$$P(s \in S) = \frac{(sentences\ in\ original\ text\ \&\ also\ in\ the\ summary)}{(sentences\ in\ the\ original\ text)} \quad (5)$$

Negative keyword: word in the original text phrase that does not appear in the sentence of the summary. Negative keyword is the opposite of positive keywords. Therefore, it can be formulated as follows:

$$score_{f_2}(s) = \frac{1}{length(s)} \sum_{i=1}^{n} tf_i * P(s \notin S | keyword_i) \quad (6)$$

Assume s is a sentence in the original text, S is a sentence in the summary, f_1 is negative keyword text features, n is the number of keywords in sentences, tf_1 is the number of keywords to-i, which appears in the sentence.

Similarity between sentences: the number of words in the same sentence with words in other sentences in the document and can be formulated as follows:

$$score_{f_4} = \left| \frac{keywords\ in\ s \cap keywords\ in\ between\ sentences}{keywords\ in\ s \cup keywords\ in\ between\ sentences} \right| \quad (7)$$

Assuming f_4 is feature similarity between sentences, and s is a sentence in the original text.

Sentence that resembles the title: the word in the same sentence with the word in the title of the document and is formulated as follows:

$$score_{f_5} = \left| \frac{keywords\ in\ s \cap keywords\ in\ title}{keywords\ in\ s \cup keywords\ in\ title} \right| \quad (8)$$

Assuming f_5 is a feature that resembles the title phrase, and s is a sentence in the original text.

Name of entity: a collection of words that have meaning or form the name of an organization, the name, region name, the name of the day, month year date, then the calculation of the sentence containing the name of the text features of entities can use the formula as follows:

$$score_{f_6} = \frac{name\ of\ the\ entity\ (s)}{length\ of\ sentence\ (s)} \quad (9)$$

Assuming s is a sentence in the original text, the text feature f_6 is a sentence containing name entities.

Sentences containing numerical data: a collection of words that have meaning or form the name of an institution, then the calculation of the sentence containing the name of the text features of entities can use the formula:

$$score_{f_7} = \frac{the\ numerical\ name\ (s)}{length\ of\ sentence(s)} \quad (10)$$

Where s is a sentence, f_7 is a sentence that contains a text feature numeric data.

Sentence length: calculated based on the number of words divided by the number of words in the phrase throughout the text, word count does not include common words. The calculation of sentence length text features using the formula:

$$score_{f_8} = \frac{Number\ of\ words\ in(s)}{unique\ words\ in\ the\ document(s)} \quad (11)$$

Assume s is a sentence, f_8 is sentence length text features.

Connections between sentences: the number of sentences that have the same word with the other sentences in the document and can be formulated as follows:

$$score_{f_9} = number\ of\ connections\ between\ sentences \quad (12)$$

Assume s is a sentence, f_9 is the connection between sentences of text features.

Connection weights between sentences: The function of this feature is to add up the weight of the connection between the sentence was calculated as follows:

$$score_{f_{10}} = weighting\ number\ of\ connections\ between\ sentences \quad (13)$$

Assume s is a sentence, f_{10} is the text feature connection weights between sentences.

Semantic sentence: a sentence that characterizes relations between sentences based on semantics. Assume D is a document, ($t = M$) is the number of words in D, and S($S = N$) is the number of sentences in D. With S_j is the sentence to-i in the document and t_i is the term$_i$ appears in the document. Then it can be made a matrix below:

$$A = \begin{array}{c|cccc} & S_1 & S_2 & \cdots & S_N \\ \hline t_1 & w_{1,1} & w_{1,2} & \cdots & w_{1,N} \\ t_2 & w_{2,1} & w_{2,2} & \cdots & wa_{2,N} \\ \vdots & \vdots & \vdots & \ddots & \cdots \\ t_M & w_{M,1} & w_{2,1} & \cdots & w_{M,N} \end{array}$$

Fig. 1: Matrix sentence semantics

tfi is the number of the appearance of the term to-i in the sentence. SFi Frequency of words is a sentence that contains a lot of the term to-i, while $ISFi$ = log $NSFi$ log is a measure of discriminate emergence term to-i in the document, N is the number of sentences in a document. Then $w_{i,j}$ is

$$w_{i,j} = tf_i \times ISF_i \quad (14)$$

Regression: a statistical method that can be used to investigate or build a model of the relationship between several variables [10].

The variables used include independent variables and the dependent variable. Regression model was used to determine the relationship between the independent variables with the dependent variable, and also to see whether each independent variable associated positive or negative and to predict the value of the dependent variable when the independent variable value has increased or decreased.

Mathematical regression: a good model for estimating the weight of a text feature [11], [12]. In this model of mathematical functions can relate output to input [8].

Logistic regression: relationship dependent variable (Y) with one or more independent variables (X) in which the dependent variable is categorical measurement scale and the independent variable (X) has a nominal measurement, level, scale, or ratio [13], [14], [15].

Logistic regression analysis with logic or circuit function is often called the binary logistic regression in which the dependent variable (Y) is worth dichotomy or binary (only two possibilities e.g. 1 = "pass" and 0 = "failed"). Line to describe the relationship between X and Y for binary logistic regression may not linear, but rather an S-shape.

Research methods: collection of text documents, training, and testing from the daily online Kompas obtained from the corpus [16] of 100 original documents and manuals for the research phase and 50 original documents and manuals for the testing phase.

Doing weight training phase modeling features with binary logistic regression algorithm. Response variable used consists of two categories, namely y = 1 which states "sentences into a summary" and y = 0, which states "the sentence was not included in the summary". The resulting model is integrated to each text feature for all documents in the testing phase.

Summary of the results of the system compared to manually summary. The goal is to find the accuracy of the system summary. Accuracy values between 0 and 1, which means if the value is 1 then the accuracy of the summary of the system 100%. Determination of accuracy using the F-measure, precision, and recall [17]:

$$F_{\beta=1} = \frac{(\beta^2+1)PR}{\beta^2 P+R} = \frac{2PR}{(P+R)} \quad (15)$$

$$P = \frac{|S \cap T|}{|S|}; R = \frac{|S \cap T|}{|T|} \quad (16)$$

With β is the weight of precision (P) and recall (R), β <1 the emphasis on precision and β > 1 an emphasis on recall. F-measure values between 0 and 1, and therefore the value of β = 1. Assume that S is a text summary of the results of the machine (the function scores of documents training) and T is the summary of the text manually.

III. Experiments and Result

Format documents: documents are used type plain text with a simple XML format. Plain text is a form of digital files that contain character data set without the addition of a specific format.

Type plain text is selected for this type of document is a public document that can be handled by almost any programming languages. This document uses two markers (tags). Tag <TITLE> and </TITLE> used to mark the title and tags <TEXT> and </TEXT> used to mark the contents of the document.

Implementation of the system: the interface text of summarization systems using Java and Perl as well as the use of summarization systems used data format XML text type.

The steps in the implementation of this system are reading the text, feature extraction and summarizing the text. Reading of the text is the way in cutting sentences, cutting words and word filtering.

Text extraction method is where summarizing is done based on a summary document ranking score of each sentence in the document. Cutting text on documents required to separate each one with the other sentences.

Word cutter is useful to cut the sentences into words with characteristic inter-word is a string of words separated by spaces or other punctuation (Spaces, tabs and newline).

Filtering is used to separate the words needed with words that are not meaningful (stop word) so as to reduce the complexity of calculation of text features. Example stop word is which, at, in, with, and others.

The extraction of feature makes the score of each feature of each sentence text which then will be used at the stage of summarization of text. Text summarizing is done based on the highest score of the sentence.

Binary logistic regression algorithm modeling: Binary logistic regression modeling process, the features of data extraction result for the entire sentence in the entire document put together and then regressed using binary logistic regression to obtain a binary logistic regression models.

Table 1: Examples of feature values

s	x1	x2	x3	x4	x5	x5	x7	x8	x9	x10	x11	y
1	1	706	294	46	5	172	0	101	0	0	77	1
2	0.5	583	417	23	0	143	0	24	0	0	79	1
3	1	115	885	69	0	133	22	157	0	0	89	0
4	1	103	897	41	5	0.15	0	0.14	0	0	28	0
5	0.5	45	955	41	0	0	0.12	87	0	0	97	0
6	1	85	915	23	0	71	0	49	0	0	0.07	0
7	667	105	895	41	5	0	0	0.08	0	0	54	0
8	333	84	916	18	5	0	0	35	0	0	62	0
9	1	76	924	55	0	0	0	122	0	0	29	0
10	1	52	948	46	0	63	0	56	0	0	52	0
11	667	53	947	41	0	0	0	56	0	0	0.09	0
12	333	619	381	18	5	0	0	42	0	0	85	1
13	1	607	393	14	0	167	0	21	0	0	112	1
14	0.5	708	292	14	5	125	0	28	0	0	74	1

x1 to x11 are the independent variables of the score of each features, and y is the dependent variable the value 0 and 1. Binary logistic regression modeling results can be seen in Table 2 and Table 3.

The test results of Binary Logistic Regression Model Algorithm: Testing method used F-measure is to compare the results of summarization systems is the sentence that has been generated manually summarization.

Each sentences in descending order to get the highest score. Sentence with the highest value is taken as a summary of the text. Many sentences are taken calculated based on compression is 30%. The test results of binary logistic regression modeling by using the F-measure result accuracy is 91.1%.

Coefficient model analysis: Based on the results of the binary logistic regression in Table 1, the results of analysis of the model parameters have a sig test under 0.05 which features the text "keyword positive" and "long sentences". It means that the text feature has significant influence on the results summary. Test the feasibility of the model results in Table 2, there is a sig 0.000 under sig 0.05. It means that the results of the model predictions are consistent with the observations. Test the feasibility of the model used to determine the difference between the values of the data generated by the system, while the test of the model parameters used to determine the real impact of the result summaries.

Tabel 2: Variable in the equation

	B	**S.E.**	**Wald**	**Sig.**
x1	-1197	.836	2050	.152
x2	23.976	1.716	195.136	.000
x4	3.533	3.344	1.116	.291
x5	-.107	13.307	.000	.994
x6	-.364	2.927	.016	.901
x7	-2127	4.549	.219	.640
x8	-40.877	8.466	23.315	.000
x9	.123	5.609	.000	.983
x10	.738	.503	2.155	.142
x11	1.795	7.591	.056	.813
Const	-9.391	.902	108.367	.000

Tabel 3: Hosmer and lemeshow test

Step	***Chi-square***	***Sig.***
1	73169	.000

Based on Table 4, the use of eleven features, eight features, six features, four features, five features, three features, two features, and one feature by including the second feature in it produces the same accuracy and greater than without involving a second feature in it. Meanwhile the using of eighth feature does not produce high accuracy if both features are excluded. Means the using of two features have an important influence on the testing phase and may represent the results of the accuracy of the eleven text features.

Comparison with [1] and [9]: the results of experiments that have been conducted, obtained a summary of the best accuracy of 91.1% F-measure calculation. Previous studies using the 11 feature sentences with genetic algorithm technique [1], which obtained 47.63% accuracy and using the 10 feature sentences with binary logistic regression technique [9] obtained accuracy is 42.84%.

Table 4: Accuracy results using F-measure

No	***Weight***											**Accuray (%)**
1	w1	w2	w3	w4	w5	w6	w7	w8	w9	w10	w11	91.1%
2	w2	w4	w5	w6	w7	w9	w10	w11				91.1%
3	w2	w4	w5	w6	w7	w11						91.1%
4	w2	w4	w5	w11								91.1%
5	w4	w5										40.1%
6	w1	w2	w4	w8	w10							91.1%
7	w2	w8	w10									91.1%
8	w2	w8										91.1%
9	w2											91.1%
10	w8											17.0%
11	w1	w4	w8	w10								13.4%
12	w1	w4	w10									19.1%
13	w4	w9	w10	w11								40.3%
14	w4	w11										39.6%
15	w10	w11										31.4%
16	w9	w10										43.5%
17	w4	w10										40.9%
18	w4	w9										40.7%
19	w9	w11										30.4%
20	w1	w5	w6	w7	w8							14.6%

IV. CONCLUSION

The research resulted 91.1% accuracy and greater than previous studies [1] by using 11 features and techniques of genetic algorithm, which is 47.63% and by using 10 features with binary logistic regression technique [9], which is 42.84%. The results of the analysis of model parameters test, only text feature "positive keywords (f2)" and the text feature "long sentences" which affect toward the summary. However, based on the results of observations show only text feature "positive keywords (f2" which affect the accuracy of the summary. It means, the text feature "positive keywords (f2)" could represent eleven text features to perform the summary text.

REFERENCES

[1]. Aristoteles, Herdiyeni Y, Ridha A, Julio A. 2012. Text Feature Weighting for Summarization of Documents in Bahasa Indonesia Using Genetic Algoritm. International Journal of Science Issues. Volume 9, Issue 3, 1-6. ISSN 1694-0814.

[2]. Mani,I., Maybury,M., 1999, *Advances in Automatic Text Summarization*, MIT Press, ISBN:0262133598, Cambridge, MA.

[3]. Radev D, Hovy E, McKeown K. 2002. *Introduction to the special issue on textsummarization. Computer linguist*.

[4]. Suyanto E. 2009. *Penggunaan Bahasa Indonesia Laras Ilmiah*. Ardana Media: Yogyakarta.

[5]. Jezek, K. and J. Steinberger. 2008. *Automatic text summarization (The state of the art 2007 and new challenges)*. Vaclav Snasel (Ed): Znalosti. 1-12.

[6]. Luhn, H. P. 1959. *The automatic creation of literature abstracts*. IBM Journal of Research and Development, 159-65.
[7]. Hovy E, Lin C. 1997. Automatic text summarization in summarist. ACL/EACL-97 Workshop on Intelligent Scalable Text Summarization. 18-24.
[8]. Fattah MA, Ren F. 2008. *Automatic text summarization*. Proceeding of Word Academic of Science, Engineering and Technology. ISSN 1307-6884.
[9]. Marlina.2012. Sistem Peringkasan Dokumen Berita Bahasa Indonesia Menggunakan Metode Regresi Logistik Biner [skripsi]. Bogor. Ilmu Komputer, Matematika dan Ilmu Pengetahuan Alam, Institut Pertanian Bogor.
[10]. Usman M. 2009. *Model Linear Terapan*. Sinar Baru Algensindo: Bandung.
[11]. Jann, B., 2005. *Making Regression Tables From Stored Estimates*. Stata Journal 5, 288–308.
[12]. Richard, W., 2006. *Review Of Regression Models Forcategorical Dependent Variables Using Stata*, Second Edition, by Long and Freese. The Stata Journal 6 (2), 273–278.
[13]. Agresti Alan (2002), *Catagorical Data Analysis* (2nd edition), John Wiley & Sons, Inc., New York. ISBN 0-471-36093-7
[14]. Dobson Annette J. (2002), *An introduction to generalized linear models* (2nd edition), Chapman & Hall, New York. ISBN 1-58488-165-8
[15]. Hosmer, D. W., dan Lemeshow, S., (2000), *Applied Logistic Regression* (2nd edition), John Wiley & Sons, Inc., New York. ISBN 0-471-35632-8
[16]. Ridha A. 2002. Pengindeksan otomatis dengan istilah tunggal untuk dokumen berbahasa indonesia [senior thesis]. Bogor. Ilmu Komputer, Matematika dan Ilmu Pengetahuan Alam, Institut Pertanian Bogor.
[17]. Baeza-Yates, Ricardo and Berthier Ribeiro-Neto. 1999. *Modern information retrieval*. ACM Press New York. Addison-Wesley.

Arisstoteles is B.Sc in Computer Science (University of Padjadjaran, Indonesia, 2004), M.Sc in Computer Science (Bogor Agricultural University, Indonesia, 2011). Since 2006 the author active as a lecturer in the Department of Computer Science, University of Lampung, Indonesia.

Widiarti is is B.Sc in Mathematics (University of Lampung, Indonesia, 2004), M.Sc in Computer Science (Bogor Agricultural University, Indonesia, 2011). Since 2005 the author active as a lecturer in the Department of Mathematics, University of Lampung, Indonesia.