Multi Document Summarization for the Indonesian Language Based on Latent Dirichlet allocation and Significance sentence

Agus Widjanarko
Department of Information System
Diponegoro University
Semarang, Indonesia
aguswidjanarko@student.undip.ac.id

Retno Kusumaningrum
Department of Informatics
Diponegoro University
Semarang, Indonesia
retno@live.undip.ac.id

Bayu Surarso
Department of Mathematics
Diponegoro University
Semarang, Indonesia
bayus@undip.ac.id

***Abstract*- Automatic Multi-document summarization in Bahasa Indonesia can help people get more comprehensive online news information. The clustering algorithm which is widely developed over a decade in the text data domains is Latent Dirichlet Allocation (LDA). The LDA method contributes quite well in the field of text classification and information retrieval. One of LDA's usages is a document summarization method, since LDA is able to get the framework in a document. The multi-document summarization in Indonesian language using unsupervised learning approach, especially LDA, is still limited. The LDA and Significance Sentence methods have the advantage of choosing representative sentences from source documents. The testing model was performed using a combination of alpha parameters 0.1 and 0.001 as well as beta 0.001 and 0.1, which is combined with compression rate at 10%, 30% and 50% in the sentence ranking process of each document. Testing results show that the best result was obtained under parameters combination as follows: alpha value is 0.01, beta value is 0.1, compression rate is 50% and cosine similarity value is 0.931.**

Keywords- extractive multi document summarization; Sentence LDA ;Significance Sentence; Topic Model

I. INTRODUCTION

Penetration of internet usage in Indonesia until 2016 has reached 132.7 million people or as much as 48.2% of the total population of Indonesia as much as 256.6 million people. The growth of Internet use is driving the development of digital information, from 132.7 million people 96.4% dominated access to online news information. With the growing Internet, users will result in greater overloading of information. To obtain valid information required an information processing mechanism so that users quickly get the information he needs.

Automatic text summarization systems help in this task by providing a quick summary of the information contained in the document. Multi-document summarization is a method for extracting relevant sentences from multiple documents into the summary.

Multi-document summarization to contain a summary by reducing the size of the document without losing the characteristics or central meaning of the document The primary purpose of multi-document summaries is to reduce the same data and identify contradictory information and maximum sentence cohesion [1] The extraction method emphasizes the relevant sentence between one sentence with another sentence between documents. Methods to measure the relevance of sentences are supervised and unsupervised learning methods [2][3].

Several supervised techniques method have been implemented to multi-document summarization is cue phrases and topic terms, for example terms with the Term Frequency, Inverse Document Frequency (TF/IDF) A summary method using word measurement, ie with TF / IDF approach on word weighting to get a representative sentence this methods has limitations on the selection of appropriate terms/words as the basis for weighting a word or sentence [4][5]. While using the LDA, a document is composed of a set of topics, the topic of the topic obtained the framework of thought of a document[4]. In addition to this, supervised methods are constrained by the availability of training sets that contain summaries of several documents [6].

In addition, the u nsupervised learning methods in the text data domains developed over a decade are the Latent Dirichlet Allocation (LDA). Clustering on documents today is an efficient method for organizing documents, one of which is the topic probability model or LDA. The LDA method contributes quite well in the field of text classification and information retrieval[7]. In LDA a document can be viewed as a distribution of latent topics, where each distribution of topics is through a word. In its development, the LDA method is widely used as an automatic multi-document summarization. Several unsupervised methods has been implemented to multi-document summarization in English document [8], [9].

While for Indonesian multi-document summarization. the various methods are used. They are Feature Scoring and rangking[10], Term Frequency dan Inverse Document Frequency (TF/IDF) and Singular Value Decomposition[11]. However, the use of Term Frequency in the field text summarization has recently been replaced by Latent Dirichlet Allocation [4]. LDA method in text classification can improve the quality of multi document summary, but in Indonesia language has not found the multi-document summary by using LDA method. Besides providing an overview of multi-document summarization techniques, we implemented LDA model for extraction-based summarization framework, which

utilizes several different techniques with the primary goal of finding out which of the applied methods contribute significantly to the quality of automatically generated summaries in Indonesia language.

II. METHODOLOGY

Section II explains 4 topics which give a foundation for implementing LDA for summarization (Fig 1). The first sub section explains preprocessing, The sub section explains LDA for sentence topic, where as the third sub section explains LDA with the significance sentence. The last sub section number of topic and step of multi document summarization for an Indonesia language news articles.

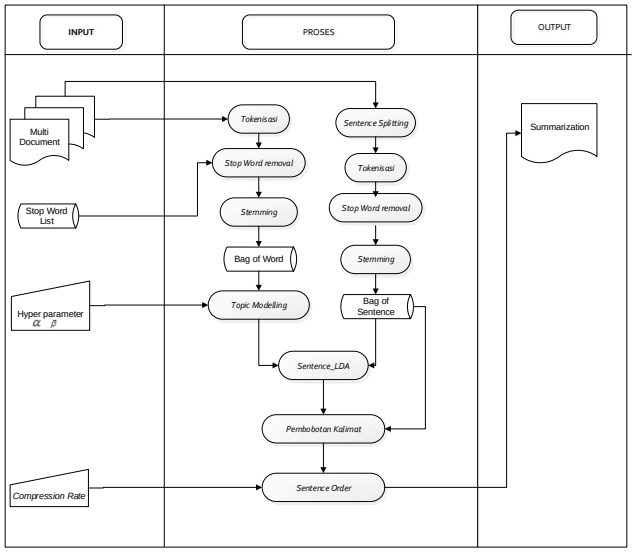

fig 1.Framework Automatic Multi-document Summarization

A. Preprocessing

In this research pre-processing is divided into 2 (two) ie pre-processing text to form Bag of Word and pre-processing text to form the Bag of Sentence. The preprocessing step of the text according to consists of noise removal, tokenization, stemming.

Sentence Splitting is the process of decoding sentences from a paragraph into a sentence token sentence, eliminating in addition to letters such as numbers and punctuation, as well as numbers and punctuation characters considered as delimiter or separator.

Tokenization is the process of cutting the order of terms from a sentence to a collection of tokens, eliminating characters other than letters such as numbers and punctuation, as well as numbers and punctuation characters are considered as delimiter or separator.

B. Basic summarization with LDA

The LDA (Latent Dirichlet Allocation) algorithm is a clustering algorithm in text data domains that have been developed over the past decade. Clustering in documents today is an efficient method for organizing documents, one of which is the topic probability model or LDA (Latent Dirichlet allocation). LDA Model to capture the latent topics and select the sentences to form the summarization. The paper assumes that a sentence only belongs to one topic and calculates the probability that the sentence belongs to the topic. Algorithms for scoring sentences based on LDA probability distributions. The basic idea is computing the probability of the sentence from probabilities of words and topics[4] [6].

$$P(S_r|T_j) = \frac{\sum_{W_i \epsilon S_r} P(W_i|T_j) * P(T_j|D_m)}{Length(S_r)} \quad (1)$$

Where Length (S_r) = length of the sentence

C. Significance Sentence

Significance Sentence is a document summary method using LDA approach. In this approach, significant or insignificant sentences are determined by 3 (three) parameters: Alpha, Beta, and Gamma [6]

a. Alfa (Diversity of sentence distribution)

Kullback Divergence Probability Sentence to topic T_k with Probability Sentence on Topic T_s

$$\alpha = KL(P(S_i|T_k)||P(S_i|T_s)) = \sum_t P(S_i|T_k) \log_2 \frac{P(S_i|T_k)}{P(S_i|T_t)} \quad (2)$$

b. Beta (Diversity of topic distribution)

Kullback-Leibler Divergence Probability of sentences on topic T_k on Document D_p and Topic T_k on Document D_M .

$$\beta = KL(P(T_k|D_p)||P(T_k\ D_M)) \sum_d P(P_k|D_p) \log_2 \frac{P(T_j|D_p)}{P(T_j|D_d)} \quad (3)$$

c. Similarity between sentences and titles

This equation uses the closeness of a sentence with the title using Jensen-Shannon distance.

$$\gamma = \frac{\sum_t P(W_i|T_k) \log_2 \frac{P(W_i|T_k)}{P(W_i|T_t)} + \sum_t P(W_k|T_i) \log_2 \frac{P(W_k|T_i)}{P(W_i|T_t)}}{2} \quad (4)$$

Normalization process of alpha, beta and gamma parameters from the calculation process of significance, sentence obtained the value of alpha, beta and gamma parameters in the sentence. All three are summed by first normalized

$$\Psi_r = \lambda_\alpha \alpha_N + \lambda_\beta \beta_N + \lambda_\Upsilon \Upsilon_N \quad (5)$$
$$\lambda_\alpha = 0.25, \lambda_\beta = 0.25, \lambda_\Upsilon = 0.5$$

D. Number of Topic

The number of topics K is very important for the Topic Modelling with LDA due to its great influence on the models. Choosing the topic number is very important. In the paper, we use Bag of Word (BoW) from preprocessing is still a random

data which can be made into group data. Lists containing grouped data by a specific interval class or by a particular category are called frequency distribution. The formula for number of groups where K = 1+3.322 $log10(N) \approx 1+log2(N)$, where N is the number of data [14][15].

E. Multi Document Summarization Algorithm

LDA-Significance Sentence multi document summarization algorithm is described in detail as follows

1) Set parameters for a corpus, the number of topics K and hyperparameter α, β.
2) Use Collaps Gibbs Sampling run the LDA [12][13] with number topic K follow distribution bag of word [14][15]. And hyperparameter α, β.
3) Get the probability distributions, Probability of Topic T_j given Document D_k $P(T_j|D_k)$ and Probability of Word W_i given Topic T_j $P(W_i|T_j)$.
4) Get Probability of Sentence S_r given Topic T_j $P(S_r|T_j)$ using LDA model by Equation 1.
5) Calculate α, β, γ for each sentence (Significance Sentences) using LDA model. (Equation 2-4).
6) Get weight each sentence $\Psi_r$with Equation 5 and then calculate for sentence position.
7) Form summarization according to sentence weight from each document with compression rate.

III. EXPERIMENT AND RESULT

This section consists of four sub sections, i.e. (a) experimental setup, (b) setting parameters (c) scenarios of experiments, and (d) experiments results and analysis.

A. Experimental Setup

This study uses a data set of news online with 5 news document from cnnindonesia.com. Each news document performs preprocessing, i.e. tokenization, stop word removal, stemming process, sentence detector and forming a bag of word –bag of the sentence. Next step is calculating LDA Collapsed Gibbs Sampling.

This experiment is implemented using PHP 7.1, WAMP 64 bit and the following hardware specifications: Intel® Core™ i3-6150 - 2.30 GHz, with 16 GB of memory and Solid State Disk 12b GB.

B. Setting Parameters

There are some parameters that needed to be determined in the experiment. Parameters could be set with alpha 0.1,0.01, 0.001 and beta 0.1, 0.01, 0.001[16].

C. Scenarios of Experiments

The approach of evaluation of the result of system summarization by using cosine similarity is the method of evaluation with the approach of an instrument of content-based, in this research compare the content of the summary document of human/gold summary with a result of system summary. From the experimental test using 9 test parameters alpha-beta with compression rate 10%, 30% and 50% summary from each document [7] [8].

D. Experiments Result and Analysis

The instrumental method refers to manual summaries made by humans to be compared with the summaries produced by a summarizing system. In this study using the Cosine Similarity used to measure the summarization performance. The accuracy of summarization results described in Table I, while the results for each compression rate and hyperparameter alpha 0.001 and beta 0.1.

TABLE I. SUMMARIZATION ON DATASET

Alpha	Beta	10 %		30%		50%	
		S	Sc	S	Sc	S	Sc
0,001	0,001	0,679	9	0,865	23	0,922	40
0,1	0,001	0,737	8	0,876	24	0,926	40
0,001	0,1	0,697	8	0,875	24	0,927	41
0,1	0,1	0,737	8	0,883	24	0,927	41
0,01	0,1	0,746	9	0,883	24	0,931	41
0,01	0,01	0,74	8	0,864	24	0,927	41
0,01	0,001	0,71	9	0,867	24	0,923	40
0,1	0,001	0,757	9	0,853	24	0,929	41
0,001	0,01	0,744	8	0,866	24	0,928	41

S : Similarity sentence

Sc : number of sentences summary results

From the test results obtained for the level of 10% obtained the highest similarity value in the combination of alpha test parameters 0.1 and beta 0.001 that is equal to 0.757 with the total number of sentences of 9. While the lowest similarity value of 0.679 in the combination of alpha parameter 0.001 and beta 0.001. In the test with 30% level of summation obtained the best results on a combination of alpha parameters 0.1 and beta 0.1 and alpha test parameters 0.01 and beta 0.1. With the value of similarity 0.883 with the number of sentences summary of 24 sentences. While the lowest similarity value is 0.853 obtained at the combination of alpha parameter 0.1 and beta 0.001.

At the 50% level of summary, there is a document similarity of 0.93 between the summary document and the test document or the human framework. This is due to the 50% level of summary, the more sentence 41 sentences. From these explanations the best similarity value is obtained at the 50% level with a combination of alpha 0.01 and beta 0.1. With the combination of alpha and beta parameters it can be concluded that the more number of words per topic will increase the weight of the sentence so that the summary obtained will approach the human summary. In addition to this the 50% level of summary with a combination of alpha parameter 0.01 and

beta 0.1 is the best level of this summary has been tested at the level of 90% concatenation obtained results similarity is 0.944 with the number of sentences summary 70 sentences and 70% level of summary obtained results similarities that is 0.940 with a sentence count of 54 sentences. So with the similarity value of 0.014 and 0.01 compared to the 50% level of summary, the sentence is presented fewer that is a number of 41 sentences.

IV. CONCLUSION

In this study, we proposed the implementation of Latent Dirichlet Allocation for automatic multi document summarization in Indonesian Language based on sentence extraction. Sentence extraction choosing sentences from the documents using some weighting mechanism. This paper using LDA model and Significance Sentence to calculate weight sentences. LDA model using Collapse Gibbs Sampling and number of topic based on the frequency distribution. Form summarization according to sentence weight ranking each document with compression rate. In addition to that in this study, it can be seen that the highest weighted sentence in each foreign document (news online) Indonesia Language in a deductive paragraph, where the main idea of the document lies in the first paragraph.

REFERENCES

[1] G. Yang, D. Wen, Kinshuk, N. S. Chen, and E. Sutinen, "A novel contextual topic model for multi-document summarization," *Expert Syst. Appl.*, vol. 42, no. 3, pp. 1340–1352, 2015.

[2] Aggarwal, C. C., and Zhai, C. X., 2013, *Mining text data*, Springer, New York. Al-Saleh, A. B., dan Menai, M. E. B., 2016, Automatic Arabic text summarization: a survey, *Artificial Intelligence Review*, *45*(2), 203–234.

[3] M. R. Amini and P. Gallinari, "Automatic text summarization using unsupervised and semi-supervised learning," *Princ. Data Min. Knowl. Discov.*, pp. 16–28, 2001.

[4] R. Arora and B. Ravindran, "Latent Dirichlet Allocation Based Multi-Document," *Proc. Second Work. Anal. Noisy Unstructured Text Data*, pp. 91–97, 2008.

[5] R. Arora, C. Science, B. Ravindran, and C. Science, "Latent Dirichlet Allocation and Singular Value Decomposition based Multi-Document Summarization," 2008.

[6] L. Na, L. Ying, T. Xiao-Jun, W. Hai-Wen, X. Peng, and L. Ming-Xia, "Multi-document summarization algorithm based on significance sentences," *Proc. 28th Chinese Control Decis. Conf. CCDC 2016*, no. 61402069, pp. 3847–3852, 2016.

[7] D. Blei, L. Carin, and D. Dunson, "Probabilistic topic models," *IEEE Signal Process. Mag.*, vol. 27, no. 6, pp. 55–65, 2010.

[8] H. Liu, H. Yu, and Z. Deng, "Multi-Document Summarization Based on Two-Level Sparse Representation Model," *Aaai*, pp. 196–202, 2015.

[9] C. Liu, C. Zhu, T. Zhao, and D. Zheng, "Extracting main content of a topic on online social network by multi-document summarization," *Proc. 2012 8th Int. Conf. Comput. Intell. Secur. CIS 2012*, no. 1, pp. 52–55, 2012.

[10] G. Yapinus, A. Erwin, M. Galinium, and W. Muliady, "Automatic Multi-Document Summarization for Indonesian Documents Using Hybrid Abstractive- Extractive Summarization Technique," *Inf. Technol. Electr. Eng. (ICITEE), 2014 6th Int. Conf.* , pp. 1–5, 2014.

[11] F. E. Gunawan, A. V. Juandi, and B. Soewito, "An automatic text summarization using text features and singular value decomposition for popular articles in Indonesia language," *2015 Int. Semin. Intell. Technol. Its Appl. ISITIA 2015 - Proceeding*, pp. 27–32, 2015.

[12] G. Heinrich, "Parameter Estimation for Text Analysis," *Web http//www. arbylon. net/publications/text-est. pdf*, pp. 1–31, 2008.

[13] H. Xiao and T. Stibor, "Efficient Collapsed Gibbs Sampling for Latent Dirichlet Allocation.," *Acml*, pp. 63–78, 2010.

[14] D. W. Scott, "Sturges' rule," *Wiley Interdiscip. Rev. Comput. Stat.*, vol. 1, no. 3, pp. 303–306, 2009.

[15] "The Choice of a Class Interval Author (s): Herbert A . Sturges Reviewed work (s): Source : Journal of the American Statistical Association , Vol . 21 , No . 153 (Mar ., 1926), pp . 65- Published by : American Statistical Association Stable URL : htt," vol. 21, no. 153, pp. 65–66, 2012.

[16] R. Kusumaningrum, M. I. A. Wiedjayanto, S. Adhy, and Suryono, "Classification of Indonesian news articles based on Latent Dirichlet Allocation," in *2016 International Conference on Data and Software Engineering (ICoDSE)*, 2016, pp. 1–5.