Optimization Number of Topic Latent Dirichlet Allocation

Bambang Subeno
Magister of Information System
Universitas Diponegoro
Semarang, Indonesian
bambang.subeno.if@gmail.com

Farikhin
Department of Mathematics
Universitas Diponegoro
Semarang, Indonesian
farikhin.math.undip@gmail.com

Retno Kusumaningrum
Department of Informatics
Universitas Diponegoro
Semarang, Indonesian
retno@live.undip.ac.id

***Abstract*— Latent Dirichlet Allocation (LDA) is a probability model for grouping hidden topics in documents by the number of predefined topics. If done incorrectly, determining the number of K topics will result in limited word correlation with topics. Too large or too small number of K topics causes inaccuracies in grouping topics in the formation of training models. This study aims to determine the optimal number of corpus topics in the LDA method using the maximum likelihood and Minimum Description Length (MDL) approach. The experimental process uses an Indonesian news articles with the numbers of documents are 25, 50, 90, and 600 with the numbers of words are 3898, 7760, 13005, and 4365. The results show that the maximum likelihood and MDL approach result in the same number of optimal topics. The optimal number of topics is influenced by alpha and beta parameters. In addition, the number of documents does not affect the computation times, but the number of words does. Computational times for each of those datasets are 2.9721, 6.49637, 13.2967, and 3.7152 seconds. The optimization model has resulted a number of LDA topics as a classification model. This experiment shows that the highest average accuracy is 61% with alpha 0.1 and beta 0.001.**

Keywords—optimization, number of topic, likelihood, minimum description length, latent dirichlet allocation, text clustering.

I. INTRODUCTION

The difference between news articles or textual articles disseminated through electronic media with other documents is the model of information flow. The news flow is a dynamic and continuously-updated stream; the more the news article in electronic media is, the larger the data collection as it always increases as well [1]. With enormous data variations, problems occur when needing to take on different news while having the same theme. So, to facilitate navigation, news articles must be grouped by the same topic. One way to get topic information contained within the corpus of a news article document is to use topic modeling. Latent Dirichlet Allocation (LDA) is a topic modeling technique that can group words into specific topics from various documents [2]. The number of topics contained in the corpus with various variations is necessary to optimize the number of topics contained within the corpus.

There are several estimation algorithms used in LDA including Expectation-Maximization algorithm [2]. Expectation-Propagation algorithm to obtain better accuracy [3], EM variations used require high computation, learning models to be biased, and inaccurate. Collapsed Gibbs Sampling [4]. All of these algorithms, the number of topics should be set beforehand.

Determining the number of K topics is very important in LDA. Incorrectly determining the number of K topics can result in limited word correlation with topic [5]. Too large or too small number of topics will affect the inference process and cause inaccuracies in grouping topics in the training model [6]. The use of Bayesian non parametric methods, such as Hirarchial Dirichlet Process (HDP) in determining the number of topics, experienced bottlenecks during high computation [7]. The use of stochastic variational inference and parallel sampling is not consistent in the determination of the number of topics in the LDA model [8].

In this study, we optimize the number of topic LDA using maximum likelihood and Minimum Description Length (MDL) towards the usage Indonesian news articles. Basically, LDA Collapsed Gibbs Sampling (CGS) runs based on the number of documents [9] [10], so that the number of documents greatly affects the computation time. In this study, the number of documents does not affect the computation time, while the number of words greatly affects the computing time. To obtain the optimal number of topic K based on likelihood, LDA CGS will run from the smallest number of K to the largest number of K. For each K, we will calculate log-likelihood value and perplexity with certain iteration. The iteration will stop itself if perplexity value convergences. The optimal number of topic will automatically be obtained based on maximum log-likelihood value of the K range. For MDL as opposed to likelihood, LDA CGS will run from maximum number of K to minimum number of K. The smallest MDL value of the K range represents the optimal number of topics.

II. METHODOLOGY

This section discusses the implementation of likelihood and MDL to find the optimal number of topic LDA. The process of optimizing the number of topic LDA is a one-time execution. The optimization process stages are documents input, pre-processing, Bag of Word (BoW), determining the maximum number of topic K, and optimizing number of topic. The process of optimizing the number of topic LDA can be seen in Figure 1.

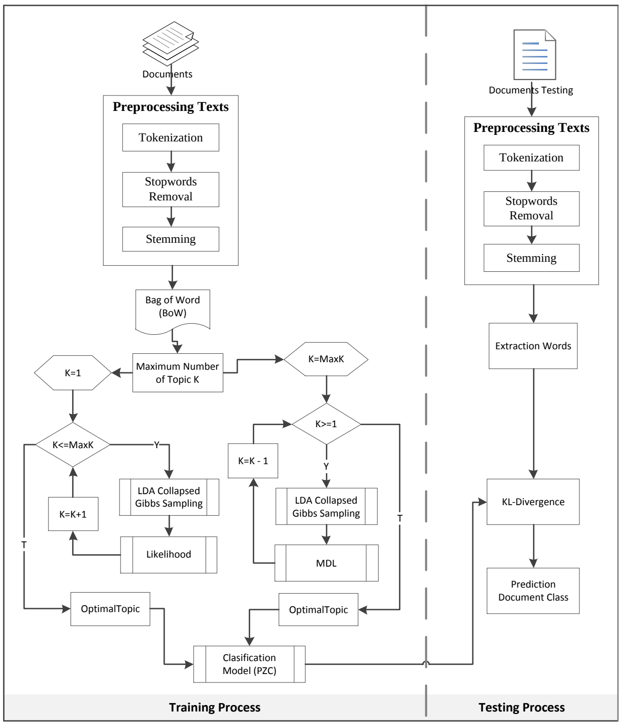

Fig 1. Process of Optimization Number of Topic LDA

A. Maximum Number of Topic

Bag of Word (BoW) pre-processing results still come in 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 [11] [12]. The formula for calculating the number of groups is as follows [11] [12]:

$$K = 1 + 3.322 \, log_{10}(N) \quad \approx 1 + log_2(N) \qquad (1)$$

where N is the number of data. For example, the resulted words are “makan”, “jeruk”, “mangga”, “beli”, “jeruk”, “apel”, “tarif”, “sopir”, “angkut”, “mahal”, “bbm”, “naik”, “bbm”, “solar”, “mahal”. Based on equation 1, the data can be grouped into 4 or 5 groups.

B. LDA Collapsed Gibbs Sampling

Latent Dirichlet Allocation is a topic modeling technique that describes the probability procedure of document [2]. Applying topic modeling to a document will be able to produce a set of low-dimensional polynomial distributions called topic. Each topic will be used to combine some information from documents that have the same word relationship. The resulted topic can be extracted into a semantic structure with comprehensive results, even in large data [13] [14]. LDA model is a probability model that can explain the correlation between words with hidden topics in the document, find topics, and summarize text documents [15]. The main idea of topic modeling assumes that each document can be represented as a distribution of several topics; each topic being the probability distribution of the words [16].

The development of LDA method used today is LDA as a generative model and LDA as inference model, that can be seen in Figure 2 [17].

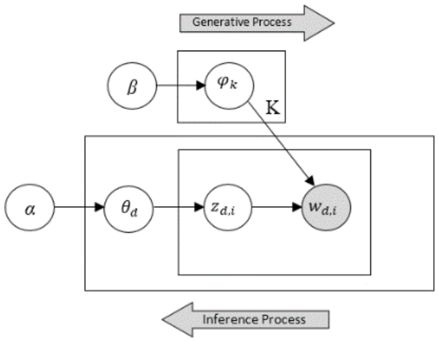

Fig 2. LDA Representation model

LDA as generative model is used to generate a document based on the probability value of word topic (φk) and proportion topic of document (θd). LDA as an inference model using Collapsed Gibbs Sampling (CGS) is the reverse of generative process as it aims to determine or find hidden value variables, i.e. probability word topic (φk) and proportion topic of documents (θd) from the predefined observation data [17]. In CGS processes, every word in the document will be determined at random at the beginning of the topic. Then, each word will be processed to determine a new topic based on the probability value of each topic. To calculate the probability value, the following formula is used [10]:

$$p(z_i = k|\vec{z}_{-i}, w) = \frac{n_{k,-i}^{(w)} + \beta}{n_{k,-i}^{(.)} + (V * \beta)} * \left(n_{d,-i}^{(k)} + \alpha\right) \qquad (2)$$

where V is number of vocabulary; $n_{k,-i}^{(w)}$ is the number of words w on topic k, except token i; $n_{d,-i}^{(k)}$ is the number of words in document d specified as topic k, except token i; and $n_{k,-i}^{(.)}$ is the total word on topic k, except the token i. To determine the probability words topic and proportion topic of document after going through the Gibbs Sampling process, the following formula is used [17]:

$$\varphi_{k,t} = PWZ = \frac{n_k^{(t)} + \beta_t}{\sum_{t=1}^{V}(n_k^{(t)} + \beta_t)} \qquad (3)$$

$$\theta_{d,k} = PZD = \frac{n_d^{(k)} + \alpha_k}{\sum_{k=1}^{K}(n_d^{(k)} + \alpha_k)} \qquad (4)$$

```
for ( d=1 to D ) do
  for ( i=1 to N_d) do
    v ← w_di , I_di ← N_di
    for ( j = 1 to I_di ) do
      k = z_dij
      N_wk ← N_wk − 1, NN_dk ← N_dk − 1
      for ( k = 1 to K ) do
        pk = (N_wk + β) x (N_dk + α) / (∑_v N_vk + Vβ)
      x ~ uniform (0, pk)
      k ← binarySearch(k: p_{k−1} < x < p_k)
      N_wk ← N_wk + 1, NN_dk ← N_dk + 1
      z_dij = k
```

Fig 3. Pseudo-code CGS Standard *[9]*

```
for ( d=1 to D ) do
  for ( i=1 to N_d) do
    v ← w_di , k = z_di
    N_wk ← N_wk − 1, N_dk ← N_dk − 1
    for ( k = 1 to K ) do
      pk = (N_wk + β) x (N_dk + α) / (∑_v N_vk + Vβ)
    x ~ uniform (0, pk)
    k ← binarySearch(k: p_{k−1} < x < p_k)
    N_wk ← N_wk + 1, NN_dk ← N_dk + 1
    z_dij = k
```

Fig 4. Pseudo-code Efficient CGS-Shortcut *[9]*

```
for ( i=1 to N_d) do
  v ← w_di , old_k ← z_di
  for ( k = 1 to K ) do
    if ( k = z_di) then
      N_wk ← N_wk − 1, N_dk ← N_dk − 1
    pk = (N_wk + β) x (N_dk + α) /
         (∑_v N_vk + Vβ)
    new_k ← index_k (max(p_k ))
    z_di = k
    if (old_k = new_k) then
      N_wk ← N_wk + 1, N_dk ← N_dk + 1
```

Fig 5. Collapsed Gibbs Sampling (CGS) – optimization

C. Likelihood

Maximum Likelihood is the estimated standard used to determine the point estimation of an unknown parameter of probability distribution with maximum probability. The estimation obtained by the likelihood maximum method is called likelihood maximum estimate [18]. There are several likelihood sample models developed for estimation on topic modeling such as Importance Sampling, Harmonic Mean, Mean Field Approximation, Left-to-Right Samplers, Left-to-Right Participant Samplers, Left-to-Right Sequential Samplers [19]. The log-likelihood function on topic LDA modeling is as follows [10]:

$$p(w_d|M) = \sum_{t=1}^{V} n_d^{(t)} \log(\sum_{k=1}^{K} \varphi_{k,t} \cdot \theta_{d,k}) \quad (5)$$

```
for ( v=1 to V ) do
  for ( d=1 to D ) do
    for ( k = 1 to K ) do
      // calculate_matrix
      C_v,d = C_v,d + ( φ_v,k x θ_d,k)
    Loglik = N_d,t x log(C_v,d)
    sumLoglik = sumLoglik + Loglik
```

Fig 6. Pseudo-code Likelihood Standard

```
B ← newBoW
foreach B ∈ { i=1 to N_d }
  d ← B_index_doc
  v ← w_di
  for ( k = 1 to K ) do
    // calculate_matrix
    C_v,d = C_v,d + (φ_v,k x θ_d,k)
  Loglik = N_d,t x log(C_v,d)
  sumLoglik = sumLoglik + Loglik
```

Fig 7. Pseudo-code Likelihood - optimization

D. Minimum Description Length

Minimum Description Length (MDL) is a method used to optimize parameter estimation of a statistical distribution and model selection in a modeling process. In this MDL principle, Bayesian theory is used to determine estimation by consideration of the likelihood data and existing knowledge of the prior probability [20]. Implementation of the MDL principle comes from the normalization of maximum likelihood to measure the model complexity of the data sets [21]. The formula for calculating the MDL is as follows [22]:

$$MDL = -\log(p(x|\theta)) + \frac{1}{2} L \log(NT), \quad (6)$$

$$L = \frac{1}{100}\left(1 + T + \frac{T(T+1)}{2} - 1\right)$$

where $\log(p(x|\theta))$ is *log-likelihood* value, T is the number of topic used, and N is number of words in the document .

E. Perplexity

Perplexity is another way to calculate the likelihood used to measure the performance of the LDA model. The smallest perplexity value is the best LDA model [10]. The formula for calculating the perplexity is as follows:

$$Perplexity = \exp\left\{- \frac{\sum_{d=1}^{D} \log p(w_d|M)}{\sum_{d=1}^{D} N_d}\right\} \quad (7)$$

where D is the number of document, $\log p(w_d|M)$ is log-likelihood according to the equation (5) and N is the number of words in the document.

III. IMPLEMENTATION ALGORITHM

In this study, we apply CGS-optimization algorithm and Likelihood-optimization using PHP programming language. The algorithms in Figures 4 and 6 of the document looping process are omitted because document index information appears in BoW results. Optimization process based on maximum likelihood and MDL once executed will automatically earn the optimal number of topic K, along with the value of perplexity, probability word topic, proportion topic for document, and probability topic of each class.

IV. EXPERIMENTS AND RESULT

Section IV consists of three sub sections, i.e. experiments setup, scenario of experiments, experiments result, and analysis.

A. Experiments Setup

In this study, we use Indonesian news articles from online portal of detik.com and Radar Semarang. The numbers of documents we use are 25, 50, 90, and 600 with the numbers of pre-processing words of each document are 3898, 7760, 13005, and 4365.

Implementation of experiments use PHP programming language, MySQL database, and hardware specifications as follows:

- Intel® Core™ i3 1.8GHz
- 4 GB of memory
- 500 GB of hard disk drive

B. Scenario of Experiments

Based on experiments setup, we perform four experimental scenarios using combinations of alpha 0.1, 0.001 and beta 0.1, 0.001. Scenario 1 aims to know comparison of execution time standard algorithm and optimization. Number of documents used 25, 90, 600 and alpha 0.1, beta 0.1. Scenario 2 aims to know the parameters that affect the time of optimization of the number of topics. Scenario 3 aims to know the parameters that affect the optimal number of topics by using Likelihood and MDL. Scenario 4 aims to know the application of resulted optimal number of topic with LDA CGS as the classifying model.

LDA CGS implementation result in the optimal number of topics as a classification model. We use 100 articles divided into 90%, or 90 document articles as training data and 10%, or 10 article documents as testing data. The article document is divided into 5 classes: each class for training data consisting of 18 news articles. In the testing process, we use Kullback-Leibler Divergence (KLD) to measure the distribution similarity between the proportion of document testing topics and the proportion of topics for each class produced in the training process. The prediction of the document testing class is taken from the smallest value of KLD. Detailed information of KLD can be found in [17]

C. Experiments Result and Analysis

The results of the experimental scenario 1 can be seen in Figure 8, and Figure 9. The results of the experimental scenario 2 can be seen in Table 1, Figure 10, and Figure 11. The results of the experimental scenario 3 can be seen in Table 2 and Figure 12. Furthermore, the result of experimental scenario 4 can be seen in Table 3 and Figure 13.

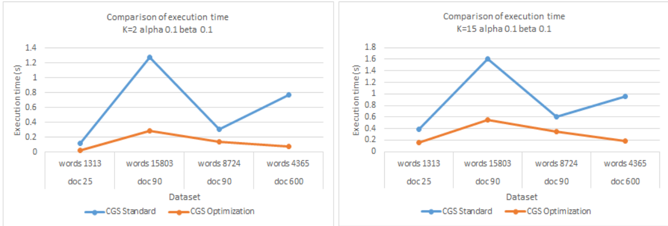

Fig 8. Comparison CGS Standard and Optimization

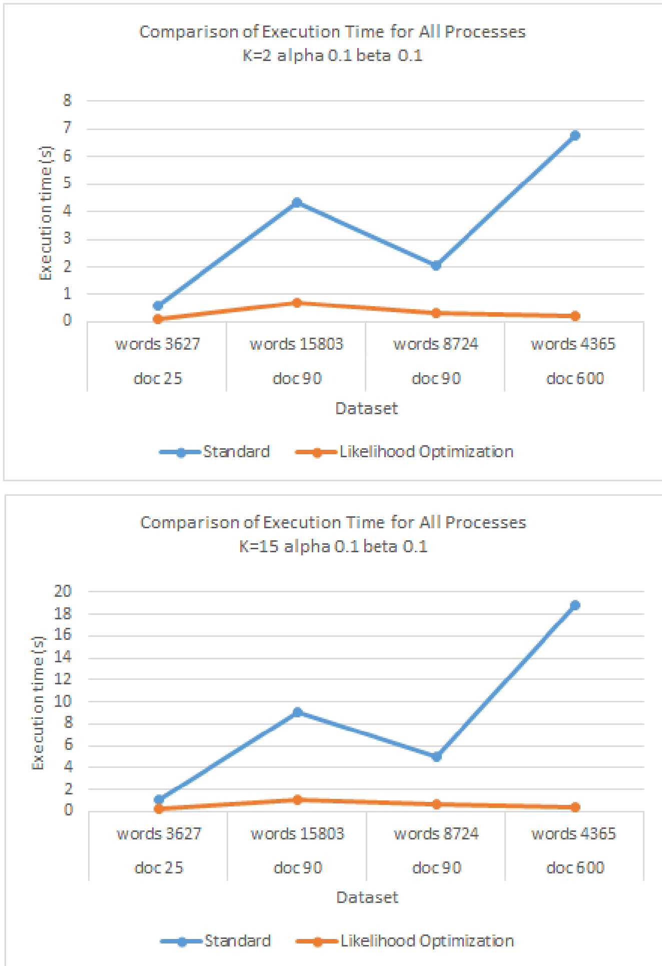

Fig 9. Comparison of Execution Time Standard and Optimization for All Processes

TABLE I. TIME OPTIMIZATION PROCESS RESULT

No	Doc	Words	Alpha	Beta	Computing Time(second)	
					Likelihood	MDL
1	25	3898	0.1	0.1	2.97216	2.97216
2	25	3898	0.1	0.001	2.96717	2.96717
3	25	3898	0.001	0.1	2.95516	2.95516
4	25	3898	0.001	0.001	2.97816	2.97816
5	50	7760	0.1	0.1	6.496371	6.496371
6	50	7760	0.1	0.001	6.467370	6.467370

7	50	7760	0.001	0.1	6.476370	6.477377
8	50	7760	0.001	0.001	6.457369	6.458369
9	90	13005	0.1	0.1	13.29676	13.29676
10	90	13005	0.1	0.001	13.31676	13.31676
11	90	13005	0.001	0.1	13.30975	13.30975
12	90	13005	0.001	0.001	13.30476	13.30476
13	600	4365	0.1	0.1	3.715208	3.725208
14	600	4365	0.1	0.001	3.715212	3.715212
15	600	4365	0.001	0.1	3.716212	3.716212
16	600	4365	0.001	0.001	3.715212	3.715212

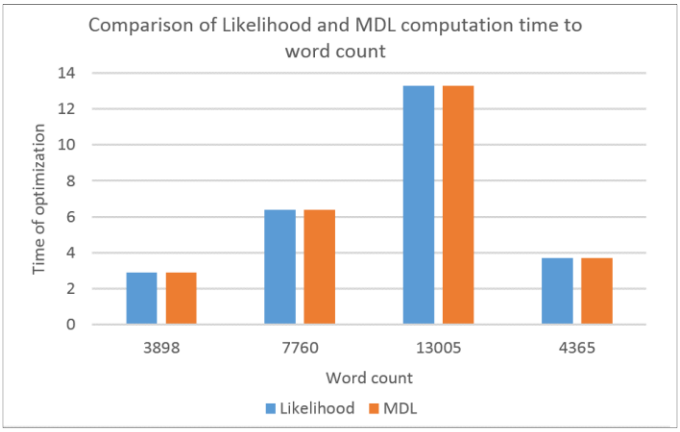

Fig 10. Comparison of Likelihood and MDL computation time to word count

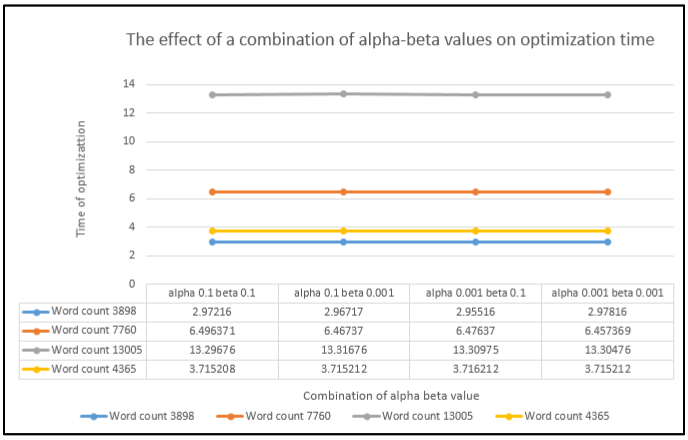

Fig 11. The effect of a combination of alpha-beta values on optimization time

Results in Table 1, Figure 10, and Figure 11 shows that the number of words used will affect the computational time: the greater number of words are, the bigger the computational time will increase. The number of documents and combinations of alpha, beta does not affect the computational time. The use of algorithms shown in Figures 5 and 7 greatly affects the optimization of the execution time. Looping document is removed because the Bag of Word (BoW) pre-processing results show a document index. This is shown in the experimental results of scenario first in Figure 8, and Figure 9 uses the optimization algorithm of execution time faster than standard algorithm.

TABLE II. OPTIMAL NUMBER OF TOPICS BASED ON LIKELIHOOD AND MDL

No	Doc	Words	Alpha	Beta	Optimal Number of Topic	
					Likelihood	MDL
1	25	3898	0.1	0.1	11	11
2	25	3898	0.1	0.001	12	12
3	25	3898	0.001	0.1	13	13
4	25	3898	0.001	0.001	13	13
5	50	7760	0.1	0.1	13	13
6	50	7760	0.1	0.001	14	14
7	50	7760	0.001	0.1	14	14
8	50	7760	0.001	0.001	14	14
9	90	13005	0.1	0.1	15	15
10	90	13005	0.1	0.001	15	15
11	90	13005	0.001	0.1	15	15
12	90	13005	0.001	0.001	15	15
13	600	4365	0.1	0.1	12	12
14	600	4365	0.1	0.001	12	12
15	600	4365	0.001	0.1	13	13
16	600	4365	0.001	0.001	13	13

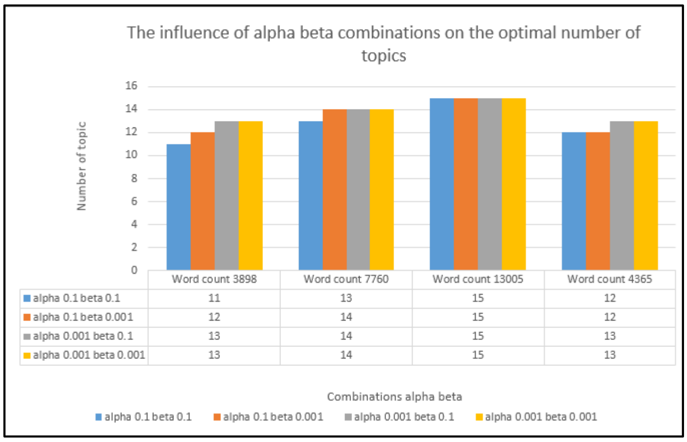

Fig 12. The influence of alpha, beta combinations on the optimal number of topics

Based on the experimental results in Table 2 and Figure 10, hyper-parameter alpha, beta can affect the optimal number of topics on likelihood and MDL. Although the use of alpha, beta values may affect the number of topics, the Likelihood and MDL processes will result in the same optimal number of topics.

TABLE III. AVERAGE ACCURACY CLASSSIFICATION OF EVERY FOLD

Fold	Accuracy of Document Classification			
	Alpha 0.1 Beta 0.1	Alpha 0.1 Beta 0.001	Alpha 0.001 Beta 0.1	Alpha 0.001 Beta 0.001
1	60%	70%	40%	50%
2	60%	50%	50%	40%
3	50%	50%	60%	50%
4	50%	80%	50%	50%
5	40%	60%	40%	50%
6	50%	70%	40%	50%
7	50%	70%	50%	70%
8	50%	50%	30%	60%
9	50%	60%	40%	50%
10	50%	50%	50%	50%
Average	**51%**	**61%**	**45%**	**52%**

Table 3 shows the result of LDA CGS implementation as a classification model using 10 fold. The highest accuracy of document classification is 0.80 or 80% with alpha 0.1 and beta 0.001.

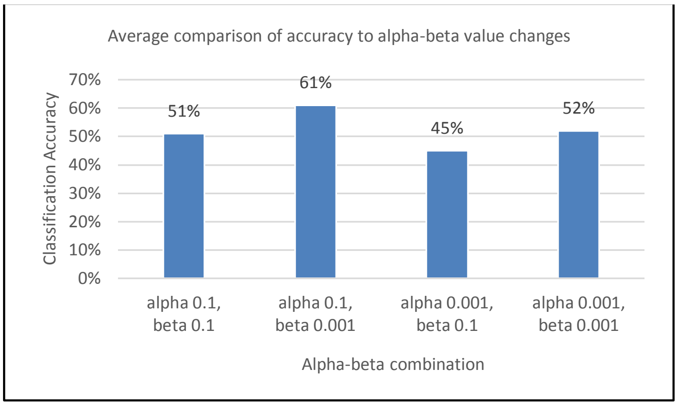

Fig 13. Average comparison of accuracy to alpha-beta value changes

Based on the experimental result in Table 3 and Figure 11, it is shown that the average highest classification accuracy of each fold is 61% with hyper-parameter alpha 0.1 and beta 0.001. The use of alpha and beta greatly affects the accuracy of document classification. The use of appropriate hyper-parameter alpha, beta will produce a high degree of accuracy as in fold 4 with 0.80 or 80% accuracy.

V. CONCLUSION

Optimization number of topic LDA, using Likelihood and MDL, yields the same optimal number of topic. The number of documents does not have a significant effect in the optimization process, but the number of words does. The more number the words used, the longer the computational time is. Combination of alpha, beta values will conduct an effect on the optimal number of topic, but does not give a significant effect on computational time.

Optimizing the number of topics LDA CGS can be applied as a classification model, but in order to get good accuracy, one should do several iterations and use appropriate alpha, beta values. The incorrect use of alpha, beta values will affect the optimal number of topics, and the classification accuracy is not good. In this study, the highest mean value earned for 10-fold is 0.61 or 61% with alpha 0.1 and beta 0.001. The best classification accuracy is shown on fold 4 with 0.80 or 80% accuracy value.

ACKNOWLEDGMENT

The authors would like to thank the research funding supported by Universitas Diponegoro under the grant of research for International Scientific Publication.

REFERENCES

[1] D. Bracewell, Y. Jiajun and R. Fuji, "Category Classification and Topic Discovery of Japanese and English News Articles," 2009.

[2] D. M. Blei, A. Y. Ng and M. I. Jordan, "Latent Dirichlet Allocation," *Journal of Machine Learning Research 3,* pp. 993-1022, 2003.

[3] T. Minka and J. Lafferty, "Expectation-propagation for the generative aspect model," *In UAI,* p. 352–359, 2002.

[4] T. L. Griffiths and M. Steyvers, "Finding scientific topics," *Proceeding of the National Academy of Sciences,* vol. 101, pp. 5228 - 5235, 2004.

[5] A. Kulesza, N. R. Rao and S. Singh, "Low-Rank Spectral Learning," *International Conference on Artificial Intelligence and Statistics (AISTATS),* vol. 33, 2014.

[6] J. Tang, Z. Meng, X. Nguyen and Q. Mei, "Understanding the Limiting Factors of Topic Modeling via Posterior Contraction Analysis," *International Conference on Machine Learning,* vol. 32, 2014.

[7] D. Cheng, X. He and Y. Liu, "Model Selection for Topic Models via Spectral Decomposition," *International Conference on Artificial Intelligence and Statistics,* vol. 38, 2015.

[8] S. Williamson, A. Dubey and E. P. Xing, "Parallel Markov Chain Monte Carlo for Parallel Markov Chain Monte Carlo," *Journal of Machine Learning Research,* vol. 28, pp. 98-106, 2013.

[9] T. S. Xiao Han, "Efficient Collapsed Gibbs Sampling For Latent Dirichlet Allocation," *Asian Conference on Machine Learning (ACML2010),* 2010.

[10] G. Heinrich, Parameter estimation for text analysis, 2.9 ed., Darmstadt, Germany: Fraunhofer IGD, 2009.

[11] H. Sturges, "The choice of a class interval," *Journal of the American Statistical Association,* pp. 65-66, 1926.

[12] D. W. Scott, "Sturges Rule," *Wires Computational Statistics,* 2009.

[13] S. Arora, R. Ge and A. Moitra, "Learning Topic Models - Going beyond SVD," *IEEE 53rd Annual Symposium on Foundations of Computer Science,* vol. 2, pp. 1-10, 2012.

[14] Z. Liu, High Performance Latent Dirichlet Allocation for Text Mining, London: Brunel University, 2013.

[15] D. M. Blei, "Probabilistic Topic Models," *Communication of The ACM,* vol. 55, no. 4, pp. 77-84, 2012.

[16] Z. Qina, Y. Cong and T. Wan, "Topic modeling of Chinese language beyond a bag-of-words," *Computer Speech and Language,* vol. 40, pp. 60-78, 2016.

[17] R. Kusumaningrum, R. Manurung and A. M. Arymurthy, "Integrated Visual Vocabulary in LDA base scene clasification for IKONOS images," *Journal of Appled Remote Sensing,* vol. 8, 2014.

[18] I. J. Myung, "Tutorial on maximum likelihood estimation," *Journal of Mathematical Psychology,* vol. 47, pp. 90 - 100, 2002.

[19] W. Buntine, "Estimating Likelihoods for Topic Models," *The 1st Asian Conference on Machine Learning,* 2009.

[20] J. I. Myung, D. J. Navarro and M. A. Pitt, "Model selection by normalized maximum likelihood," *Journal of Mathematical Psychology,* vol. 50, pp. 167 - 179, 2005.

[21] D. W. Heck, M. Moshagen and E. Erdfelder, "Model selection by minimum description length: Lower-bound sample sizes for the Fisher information approximation," *Journal of Mathematical Psychology,* vol. 60, pp. 29-34, 2014.

[22] W. Xiaoru, D. Junping, W. Shuzhe and L. Fu, "Adaptive Region Clustering in LDA Framework for Image Segmentation," *Proceedings of 2013 Chinese Intelligent Automation Conference,* pp. 591-602, 2013.