Indonesian Text Feature Extraction using Gibbs Sampling and Mean Variational Inference Latent Dirichlet Allocation

PM Prihatini, IKGD Putra, IAD Giriantari, M Sudarma
Doctoral Programmed of Engineering Science, Faculty of Engineering
Udayana University
Bali, Indonesia

***Abstract*— Latent Dirichlet Allocation has been developed as topic-based method which uses reasoning to determine the topics of a document. There are many methods of reasoning used for Latent Dirichlet Allocation, including the Gibbs Sampling and Mean Variational Inference, the most widely used in research. However, there have not been many studies that discuss the implementation of these methods on the Indonesian text, so analysis is needed to compare its performance in generating feature extraction. Therefore, in this paper, will be implemented the method of reasoning Gibbs Sampling and Mean Variational Inference for Latent Dirichlet Allocation on Indonesian text. The objective is determining the performance of both algorithms on Indonesian text so it can provide a reference about the better reasoning method for Latent Dirichlet Allocation on Indonesian text. The research was implemented on digital Indonesia news text data with 100 documents. The tests are conducted on feature data as the result of extraction process using three type of evaluation metric. The test results show that Gibbs Sampling has a better performance than Mean Variational Inference for Latent Dirichlet Allocation on Indonesian text.**

Keywords—feature extraction; latent dirichlet allocation; gibbs sampling; mean variational inference; Indonesian text

I. INTRODUCTION

Data on the internet continues to grow reaching unlimited amounts. Information retrieval system is the solution that can help the user to search information needed in the enormous data. Information retrieval system consist of two main stage, that are offline and online. A set of very large documents processed first through the offline stage to determine the unique features within each document. The unique features are extracted through feature extraction process. Then, the information searched by the user that have represented by query, will be processed online by matching the query with the features of the document. Based on this matching process, the relevant information is displayed, starting from the highest relevant value to the lowest.

Feature extraction has an important role in information retrieval system to acquire the unique features of a document. The unique features are going to be the candidate answers to the information required by users. To generate the unique features, it is needed an automated process so it can improve the performance of information retrieval systems. The commonly method used for feature extraction process is Term Frequency-Invers Document Frequency or TF-IDF [1-6]. The extraction results from this method directs a document into a single topic, such as inflation news that is included in the economic topic. But in reality, one document discusses many topics, such as inflation news included in economic, social, and political topics. Therefore, Latent Dirichlet Allocation or LDA has been developed as topic-based method which uses reasoning to determine the topics of a document.

LDA extracts document features from the word level, continues to the document level, and finally reaches the corpus level [7]. Feature extraction with LDA is done through the reasoning stage and the results can be implemented for the next mining process. Reasoning becomes important in LDA process because it determines topic distribution in a document. Therefore, selection process of reasoning methods needs a special attention. There are several methods that use for the reasoning of LDA. Blei uses the method of reasoning Variational Inference for LDA, which proved capable of finding the topic distribution in the corpus [7-12]. But, Gibbs Sampling are the most widely method of reasoning used in research of LDA [13-17]. Both methods have proven to work well on English text. However, there have not been many studies that discuss the implementation of these methods on the Indonesian text, so analysis is needed to compare its performance in generating feature extraction.

Therefore, in this paper, will be implemented the method of reasoning Gibbs Sampling and Mean Variational Inference for Latent Dirichlet Allocation on Indonesian text. The objective is determining the performance of both algorithms on Indonesian text so it can provide a reference about the better reasoning method for Latent Dirichlet Allocation. The research was implemented on digital Indonesia news text data with 100 documents. All documents through pre-processing stage. The main process is a document features extracted using Gibbs Sampling and Mean Variational Inference for Latent Dirichlet Allocation. At the end, the tests are conducted on feature data as the result of extraction process using three type of evaluation metric to measure the performance of both algorithms. The next section of this paper will discuss the research method, experimental result, conclusion and upcoming work.

II. RESEARCH METHOD

A. Dataset

The research uses a digital news dataset from online news media in Indonesia. Digital news was taken from the website manually and stored in a text file. The number of files collected was 100 documents, which are extracted through Gibbs Sampling and Mean Variational Inference LDA, as shown by research method in Fig 1.

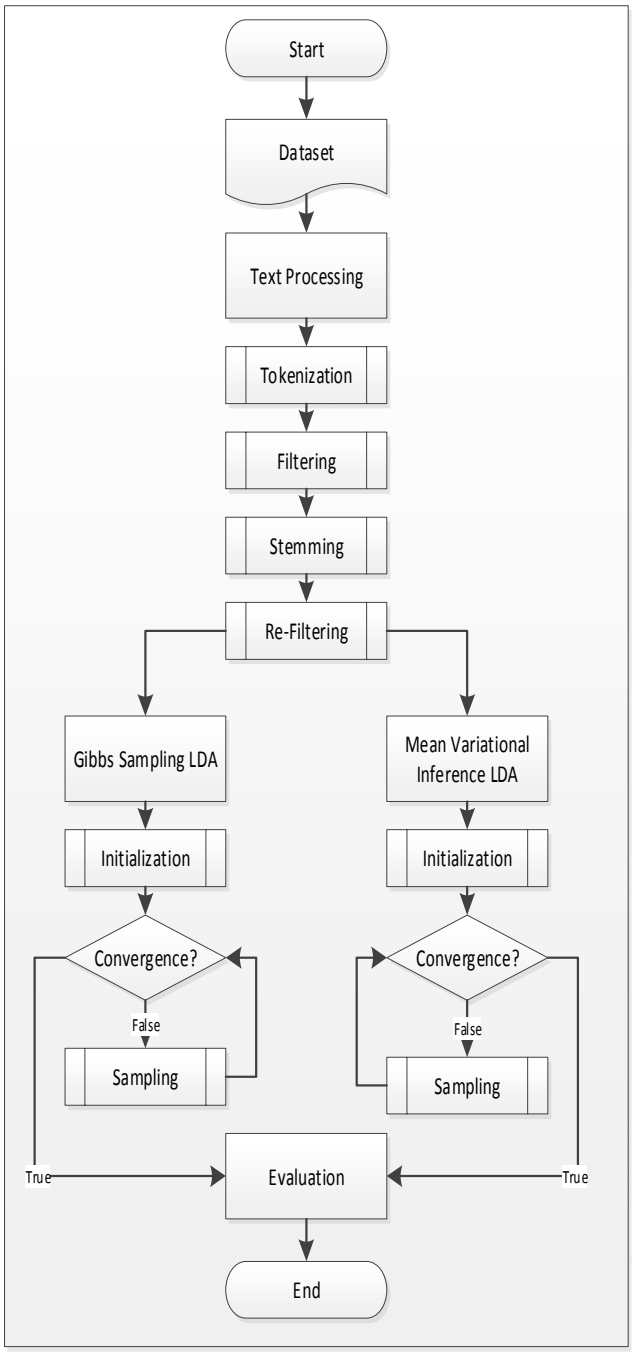

Fig. 1. Algorithm of Gibbs Sampling LDA

B. Text Pre-processing

The text file consists of a collection of sentences that would be split into the smallest unit called a term or word through tokenization process. To avoid errors toward identifying the term, then all the terms converted to lowercase. The result is a text file that contains a list of terms. The list of terms as a result of tokenization checked to find out if there is a term that is included in the stop word. If there is, then the term will be deleted from the list. The aim of this filtering process is to prevent the emergence of terms that are not meaningful as a unique feature of the document. The research used 906 stop words [18]. Stemming is the next processing stage, which checks whether each term is a root word in the dictionary. Otherwise, stemming will remove prefix and suffix of each term in accordance with the rules of Indonesian grammar. The research used 30342 root words. There are five rules used to remove affixes [18]. The last process is rechecking the results stemming whether the resulting root words included in the stop words. This stage is the re-filtering process to ensure that the resulting term list is a unique feature of each document [19].

C. Feature Extraction of Gibbs Sampling LDA

The reasoning process of Gibbs Sampling LDA algorithm begins with initialization. Then the initialization results are used to do a new topic sampling, as shown in Fig. 2 [15]. Variable *nkm* represents the number of topic *k* for each term of document *m*. Variable *nm* represents the count of topic for document *m*. Variable *ntk* represents the number of term *t* for each topic *k*. Variable *nk* represents the count of term for topic *k*. A multinomial random value is used as the initial value at the initialization stage. Then the value of all variables is incremented for each term and for each document in the corpus.

```
#initialization stage
nkm=0, nm=0
ntk=0, nk=0
for m=1 to M do
      for n=1 to Nm do
            zmn=k = Mult (1/K)
            nkm=nkm + 1, nm=nm + 1
            ntk=ntk + 1, nk=nk + 1
      end for
end for
# sampling stage
while not converged do
      for m=1 to M do
            for n=1 to Nm do
               nkm=nkm – 1, nm=nm – 1
               ntk=ntk – 1, nk=nk – 1
               k = p(z,w)
               nkm=nkm + 1, nm=nm + 1
               ntk=ntk + 1, nk=nk + 1
            end for
      end for
end while
```

Fig. 2. Algorithm of Gibbs Sampling LDA

At the sampling stage, the value of all variables is decremented for each term and for each document in the corpus. Then, a new topic is sampling for each term (1). Variables *a* and *b* are corpus parameter [7, 15]. After sampling a new topic, the value of all variables is incremented again. The process of sampling new topics is done repeatedly until the final state is converged. In this research, it is achieved by ten times iteration.

$$p(z,w) = \frac{ntk_{-t} + b}{\sum_{t=1}^{V} ntk_{-t} + Wb} \frac{nkm_{-t} + a}{\sum_{k=1}^{K} nkm_{-t} + Ka} \quad (1)$$

Gibbs Sampling LDA algorithm also generates matrix of document-topic *dt* and term-topic *tt* (2)(3) [15].

$$dt = \frac{ntk + b}{\sum_{t=1}^{V} ntk + b} \quad (2)$$

$$tt = \frac{nkm + a}{\sum_{k=1}^{K} nkm + a} \quad (3)$$

D. Feature Extraction of Mean Variational Inference LDA

As with Gibbs Sampling, Mean Variational Inference also begins with the initialization process prior to sampling the topic, as shown in Fig. 3 [7].

Variable *t*, *a* and *b* refer to Dirichlet parameter. Variable *c* refers to multinomial parameter. In the initialization process, both variables were set to value depending on LDA dimension *k* and the number of terms for all documents *N*. This is done for each topic *i* and term *n*. Sampling process on the Mean Variational Inference algorithm is also done repeatedly until it reaches the convergent condition. For each term and topic, the value of *t* updated depending on the value of *b* and *c,* using the function of exponential and digamma *dg*. Then, the value of *t* normalized to sum to 1, and the value of *c* updated using the new value of *t*. The convergence condition for this algorithm uses ten times iteration.

```
for n=1 to N do
    for i=1 to I do
        t =1/k
    end for
end for
for i=1 to N do
    c = a + N/k
end for
while not converged do
    for n =1 to N
        for i = 1 to k
            t = b exp (dg(c))
        end for
        normalize t to sum to 1
    end for
    c = a + Σt
end while
```

Fig. 3. Algorithm of Mean Variational Inference LDA

E. Evaluation of Algorithm

Measurements were performed on the results of feature extraction obtained through the algorithm using three type evaluation metrics.

Precision (*P*) represents how many relevant the retrieved documents (4) [20]. Variable *tp* is true positive, indicates relevant items retrieved. Variable *fp* is false positive, indicates retrieved items. Variable *fn* is false negative, indicates relevant items.

$$P = \frac{tp}{(tp + fp)} \quad (4)$$

Recall (*R*) represents how many relevant documents can be retrieved (5) [20].

$$R = \frac{tp}{(tp + fn)} \quad (5)$$

F-Measure (*F*) represents the harmonic mean value of *P* and *R* (6)[20].

$$F = \frac{2PR}{P + R} \quad (6)$$

III. Result and Analysis

This research uses 100 news documents taken from online news media in Indonesia. This algorithm requires the number of topics to determine the topic of each unique feature in the document. Unique features are the terms that characterize the document that determine the topic of document. In this research, the number of topic *K* is 10 topics. The constant value for *b* is 0.01. The constant value for *a* is 5 [15, 17].

The result of experiment for convergence condition is represented in graph; *x*-axis represents number of iteration; *y*-axis represents difference value of convergence, as shown in Fig. 4. The values of evaluation metric are shown in Table I. It is also shown in graph; the *x*-axis represents number of topic; *y*-axis represents value of evalution metric; the blue dotted line represents the value for Gibbs Sampling LDA, while the red solid line represents the value for Mean Variational Inference LDA, as shown in Fig. 5.

The experimental result in Fig. 4 shows that Gibbs Sampling algorithm has decreased in value from the first iteration to tenth iteration to achieve the zero value, while the value of Mean Variational Inference LDA algorithm fluctuated. This means that Gibbs Sampling LDA algorithm tends to be more stable in achieving convergent conditions than Mean Variational Inference LDA algorithm.

Metric measurements of the experimental result in Table I show that Gibbs Sampling LDA algorithm has a better value than Mean Variational Inference LDA algorithm. This is strengthened by the graph shown in Fig. 5, where F-Measure of Gibbs Sampling LDA algorithm has no significant value changes for all topics (more stable) than Mean Variational

Inference LDA algorithm. This shows that Gibbs Sampling LDA algorithm more relevant in extracting unique features in documents.

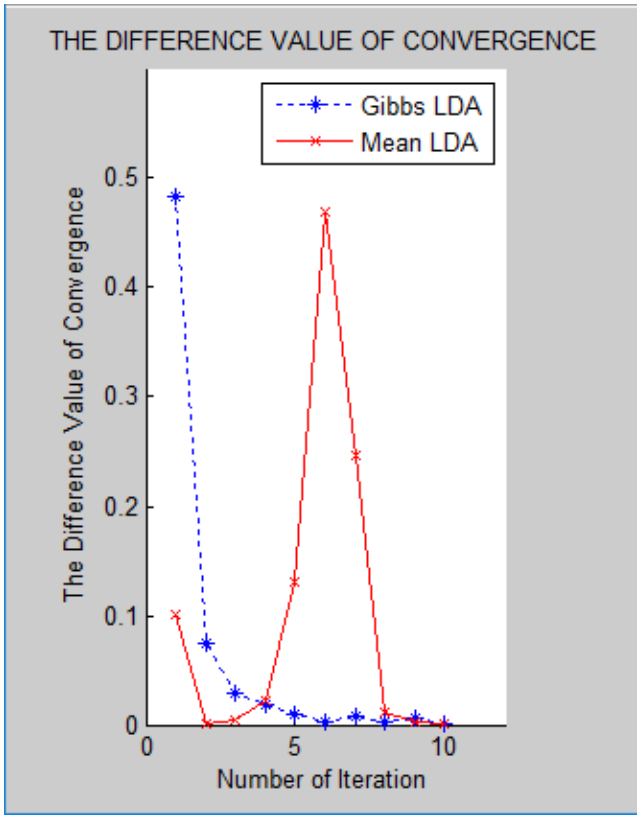

Fig. 4. The Difference Value of Convergence for Gibbs Sampling and Mean Variational Inference LDA

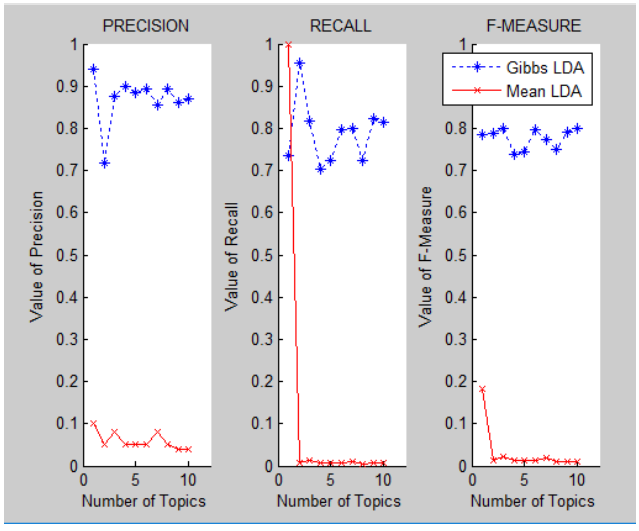

Fig. 5. The Performance of: (a) Precision, (b) Recall, (c) F-Measure

TABLE I. PERFORMANCE OF ALGORITHMS

Method	Precision	Recall	F-Measure
Gibbs Sampling LDA	0.8690	0.7892	0.7770
Mean Variational Inference LDA	0.5001	0.6890	0.5850

IV. CONCLUSION AND FUTURE WORK

Gibbs Sampling and Mean Variational Inference are two famous methods for Latent Dirichlet Allocation that is compared in this paper to extract unique features of an Indonesian text. Its purpose is to determine the performance of both algorithms on Indonesian text. The research was implemented on digital Indonesia news text data with 100 documents, K=10, b=0.01, and a=5. The results of experiment show that Gibbs Sampling has better performance than Mean Variational Inference for LDA. Gibbs Sampling LDA algorithm is more relevant in extracting unique features in a document to find hidden topics in the document. Therefore, Gibbs Sampling LDA algorithm can be implemented to extract Indonesian text.

In the upcoming work, this research will improve the ability of Gibbs Sampling LDA algorithm more better, so it can achieve the most optimal measurement values to extract unique features in documents.

REFERENCES

[1] Z. Zhao, X. He, L. Zhang, W. Ng, and Y. Zhuang, "Graph regularized feature selection with data reconstruction," IEEE Transactions on Knowledge and Data Engineering, vol. 28, pp. 689-700, 2016.

[2] M. Tutkan, M.C. Ganiz, and S. Akyokuş, "Helmholtz principle based supervised and unsupervised feature selection methods for text mining," Information Processing & Management, vol. 52, pp. 885-910, 2016.

[3] K. Liu, L. Xu, and J. Zhao, "Co-extracting opinion targets and opinion words from online reviews based on the word alignment model," IEEE Transactions on Knowledge and Data Engineering, vol. 27, pp. 636-650, 2015.

[4] Z. Hai, K. Chang, J.-J. Kim, and C.C. Yang, "Identifying features in opinion mining via intrinsic and extrinsic domain relevance," IEEE Transactions on Knowledge and Data Engineering, vol. 26, pp. 623-634, 2014.

[5] M. Ceci, C. Loglisci, and L. Macchia, "Ranking sentences for keyphrase extraction: a relational data mining approach," Procedia Computer Science, vol. 38, pp. 52-9, 2014.

[6] E. Haddi, X. Liu, and Y. Shi, "The role of text pre-processing in sentiment analysis," Procedia Computer Science, vol. 17, pp. 26-32, 2013.

[7] D.M. Blei, A.Y. Ng, and M.I. Jordan, "Latent dirichlet allocation," Journal of Machine Learning Research, vol. 3, pp. 993-1022, 2003.

[8] J. Paisley, C. Wang, D.M. Blei, and M.I. Jordan, "Nested hierarchical dirichlet processes," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 37, pp. 256-270, 2015.

[9] S. Mandt and D. Blei, "Smoothed gradients for stochastic variational inference," Neural Information Processing Systems, 2014.

[10] M.D. Hoffman, D.M. Blei, C. Wang, and J. Paisley, "Stochastic variational inference," Journal of Machine Learning Research, vol. 14, pp. 1303-47, 2013.

[11] D.M. Blei, "Probabilistic topic models," Communications of the ACM, vol. 55, pp. 77, 2012.

[12] D.M. Blei, "Introduction to probabilistic topic models," Communications of the ACM, 2011.

[13] R.Y.K. Lau, Y. Xia, and Y. Ye, "A probabilistic generative model for mining cybercriminal networks from online social media," IEEE Computational intelligence magazine, pp. 31-43, 2014.

[14] W.M. Darling, A theoretical and practical implementation tutorial on topic modeling and Gibbs Sampling, School of Computer Science, University of Guelph, 2011.

[15] G. Heinrich, Parameter estimation for text analysis, University of Leipzig, Germany, 2008.

[16] M. Dowman, V. Savova, T.L. Griffiths, K.P. Körding, J.B. Tenenbaum, and M. Purver, "A probabilistic model of meetings that combines words and discourse features," IEEE Transactions on Audio, Speech and Language Processing, vol. 16, pp. 1238-1248, 2008.

[17] M. Steyvers and T. Griffiths, Probabilistic topic models, Laurence Erlbaum, 2006.

[18] P.M. Prihatini and I.K. Suryawan, "Text processing application development for Indonesian documents clustering," The 1st International Joint Conference on Science and Technology (IJCST), Bali, Indonesia, 2016.

[19] P.M. Prihatini, I.K.G.D. Putra, I.A.D. Giriantari, and M. Sudarma, "Fuzzy-gibbs latent dirichlet allocation model for feature extraction on Indonesian documents," Contemporary Engineering Sciences, vol. 10, pp. 403-421, 2017.

[20] C.D. Manning, P. Raghavan, and H. Schutze, An Introduction to Information Retrieval, England: Cambridge University Press, 2008.