Fuzzy-Gibbs Latent Dirichlet Allocation Model for Feature Extraction on Indonesian Documents

by Putu Manik Prihatini

Submission date: 11-May-2023 03:20PM (UTC+0700)
Submission ID: 2090245456
File name: IIIA1c1a_Artikel_Fuzzy-Gibbs_Latent_Dirichlet_Allocation.pdf (922.6K)
Word count: 5919
Character count: 31028

Contemporary Engineering Sciences, Vol. 10, 2017, no. 9, 403 - 421
HIKARI Ltd, www.m-hikari.com
https://doi.org/10.12988/ces.2017.7325

Fuzzy-Gibbs Latent Dirichlet Allocation Model for Feature Extraction on Indonesian Documents

Putu Manik Prihatini

Doctoral Program of Engineering Science
Faculty of Engineering, Udayana University, Bali, Indonesia
Department of Electrical Engineering, Politeknik Negeri Bali, Indonesia

I Ketut Gede Darma Putra

Department of Information Technology, Faculty of Engineering
Udayana University, Indonesia

Ida Ayu Dwi Giriantari and Made Sudarma

Department of Electrical Engineering, Faculty of Engineering
Udayana University, Indonesia

Abstract

Latent Dirichlet Allocation is a topic-based feature extraction method that uses reasoning to find semantic relationship in corpus. Although Latent Dirichlet Allocation is very powerful in handling very large data sets, but it has a very high complexity along with increasing number of document to reach convergence. Latent Dirichlet Allocation generates probability for all topics in a document, which it contains uncertainty, so its relationship with number of iterations needs to be analyzed. In this paper, Latent Dirichlet Allocation modified by adding fuzzy logic in Gibbs sampling inference algorithm. Its purpose is to analyze the effect of fuzzy logic in handling uncertainty of the occurrence all topics in a document that affect number of iteration in reasoning. Fuzzy-Gibbs Latent Dirichlet Allocation algorithm is implemented on text data of Indonesian documents. Testing performed on three different sizes of data to determine the effect of the number of document to the number of iteration. The algorithm performance was also measured using Perplexity, Precision, Recall and F-Measure.

The test results show that Fuzzy-Gibbs Latent Dirichlet Allocation algorithm can reach convergence in a fewer iteration and has a better performance compared to Gibbs Sampling Latent Dirichlet Allocation algorithm.

Keywords: latent Dirichlet allocation, fuzzy logic, Gibbs sampling, Indonesian documents

1 Introduction

Digital document storage in very large quantities and increasing all the time requires an automatic method that allows users to find information. The effective method is clustering the documents by its category. To determine the category of a document, it needs to extract the unique features contained in the document. The results of feature extraction are processed to determine the category of document so that similar documents can be collected in a cluster. Therefore, the method of extracting features plays an important role in generating cluster of documents.

Feature extraction method that has been widely discussed in study is Term Frequency-Inverse Document Frequency (TF-IDF) [7, 16, 19, 23, 27, 30]. On TF IDF, the number of occurrences of each word in a document was used as unique features. For example, user was typing the query about "*darah* (blood)"; the result that appears is all documents that contain the word "*darah*" with the highest frequency. However, in the field linguistic, there are known the term of synonym and polysemy. Synonym refers to different words but have the same meaning, while polysemy refers to a word that has more than one meaning. As in the previous example, the word "*darah*" in the sentence "*Pria itu kehilangan banyak darah* (He lost a lot of blood)" which means "fluids in the human body". But, it has a different meaning in the sentence "*Pria itu naik darah* (He is angry)," which means "confirmed the sentence". The weakness of TF-IDF was not able to handle synonyms and polysemy. Moreover, TF-IDF was not able to find a connection structure of inter or intra-document [2]. Synonyms and polysemy has changed the feature extraction mechanism not only emphasizes the number of occurrences of words in the document, but more emphasis on the meaning of words contained in the document. This feature extraction mechanism appears in the method of topic models such as Latent Semantic Indexing (LSI), probabilistic Latent Semantic Indexing (pLSI) and Latent Dirichlet Allocation (LDA). LSI has been widely discussed in study [8, 11, 29]. pLSI is an enhancement of LSI which able to model every word as a representation of some topics that could overcome the problem of synonyms and polysemy [4, 14, 18, 31]. However, modeling which was done by pLSI only at document level. This leads to inability pLSI handle changes in the number of parameters in line with changes in the size of the corpus and the determination of the probability of documents outside the training document. LDA was developed to address the problem of pLSI, which works at the level of words, documents and corpus, so that it can capture the changes that occur to the words and documents [2].

LDA works with two mechanisms of reasoning and implementation. Reasoning performed to determine the distribution of words and documents on the topic, while the implementation uses reasoning results for the next retrieval requirements. There are several methods of reasoning used for the LDA in some studies, but widely used method is the Gibbs sampling [6, 15, 17]. Some studies have been done to improve the LDA performance by modifying LDA [1, 5, 10, 8]. The principle of reasoning in LDA is a repetition process of sampling the topic to reach convergence. Although LDA is very powerful in handling very large data sets, but this method has a very high complexity along with increasing number of document [13, 20, 22]. It is related with number of iteration performed in the reasoning to reach convergence. In addition, LDA generates proportion of occurrence for all topics in a document. This indicates that every document in corpus can be referring to all topics with different probabilities. In other words, each document has uncertainty about the topic. In the science of artificial intelligence, the uncertainty can be handled by either using fuzzy logic. Therefore, fuzzy logic can be added in the LDA method for improving the performance of LDA [3, 25]. However, the addition of these methods need to consider whether affecting the number of iteration or condition converging of LDA, because the increasing number of iterations will increase the computation time and computing resources of LDA.

In this paper, LDA modified by adding the concept of fuzzy logic in Gibbs Sampling inference algorithm. Its purpose is to analyze the effect of fuzzy logic in handling uncertainty of the occurrence all topics in a document that affect the number of iteration in the reasoning. Fuzzy-Gibbs Latent Dirichlet Allocation (FGLDA) algorithm is implemented on text data of Indonesian documents. At first, pre-processing performed on text data including tokenization, filtering, stemming and re-filtering which generates a large collection of words. Then, this algorithm uses number of occurrences of each word as an initial value in the reasoning. Testing performed on three different sizes of data. The algorithm performance was measured using the metric evaluation of Perplexity, Precision, Recall and F-Measure. The remainder of this paper is structured as follows. Section II describes research reviews that related to the proposed method. Section III describes the proposed work of FGLDA algorithm for Indonesian documents. Section IV shows simulation work and analysis of FGLDA algorithm for Indonesian documents. Section V presents conclusion with suggestions for future works.

2 Related Works

The study of LDA as topic-based feature extraction method has introduced by Blei [2]. LDA was developed to deal with the weaknesses of the LSI; LSI works only on two levels (documents and words), while the LDA works on three levels (corpus, documents, and words). LDA generates distribution of the entire topic on each document with different proportions. This is accordance with the reality that a document not only leads to one topic only, but may lead to many topics. The study

was used the variational inference reasoning methods to generate the probability of the entire topic in a document.

Chien and Chueh have used the LDA to develop hierarchical model segmentation, where the heterogeneous topic information in stream level and the word variations in document level are characterized [6]. The topic similarity between sentences is used to form a beta distribution reflecting the prior knowledge of document boundaries in a text stream. The method was used Markov chain to detect the stylistic segments within a document. Each segment is represented by a Markov state, and so the word variations within a document are compensated. The estimation problem was solved by applying the Gibbs sampling method. The whole model is trained by a variational Bayesian EM procedure.

Lau, Xia and Ye have improved the LDA to develop a novel weakly supervised cyber criminal network mining method to facilitate cybercrime forensics using Gibbs sampling algorithm [15]. The experimental results reveal that the proposed method significantly outperforms the Latent Dirichlet Allocation (LDA) and the Support Vector Machine (SVM) based method. It also achieves comparable performance as the state-of-the-art Partially Labeled Dirichlet Allocation (PLDA) Method.

Li, Zhou, Sun and Zhang have design the Weibo sentiment analysis based on LDA and Dependency Parsing [17]. A Gibbs sampling was used for inference and categorize emotion tendency automatically with the computer. In accordance with the lower ratio of recall for emotion expression extraction in Weibo, was used dependency parsing, divided into two categories with subject and object, and then, summarized six kinds of dependency models from evaluating objects and emotion words. The study proposed that a merge algorithm for evaluating objects can be accurately evaluated by participating in a public bakeoff and in the shared tasks among the best methods in the sub-task of emotion expression extraction.

Hitler and Newman have modified the Approximate Distributed LDA, or AD-LDA, to track an error bound on performance [13]. The method was proposed the parallel Gibbs sampler for LDA which enables to measure the fidelity aspect of tradeoff in terms of performance guarantees. The method upper bound the probability of making a sampling error at each step of the algorithm (compared to a sequential Gibbs sampler). The method shows empirically that the bound is sufficiently tight to give a meaningful and intuitive measure of approximation error in AD-LDA, allowing the user to track the tradeoff between accuracy and efficiency while executing in parallel.

Luo, Zhang, Ye, Wang and Cai have developed a representation model of text knowledge; model of power series representation (PSR); which has a low complex computation in text knowledge constructing process [20]. This method was proposed to leverage the contradiction between carrying rich knowledge and automatic construction. This method developed concept algebra of human concept learning to represent text knowledge as the form of power series, and a degree-2 power series hypothesis-based reasoning operations. The experiments and comparisons show that the method has better characteristics than LDA, vector space model and web ontology language.

Morchid, Bouallegue, Dufour, Linares, Matrouf, and Mori have proposed a method for categorization system based on a two-step process that expand the representation space by using a set of topic spaces and, then compact the representation space by removing poorly relevant dimensions [22]. This method was based on multi-view LDA-based representation spaces and factor-analysis models. The proposed categorization system reaches accuracy with a significant gain compared to the baseline method (best single topic space configuration).

Cau and Liu have proposed a novel type-2 fuzzy topic models (T2 FTM) to recognize human actions. This method use a type-2 fuzzy sets (T2 FS) to encode the higher-order uncertainty of each topic [3]. The primary membership function (MF) was used to measure the degree of uncertainty that a document or a visual word belongs to a specific action topic, and the secondary MF was used to evaluate the fuzziness of the primary MF itself. The method implements a two T2 FTM: 1) interval T2 FTM (IT2 FTM) with all secondary grades equal one; and 2) vertical-slice T2 FTM (VT2 FTM) with unequal secondary grades. Experiments on human action data sets demonstrate that T2 FTM performs better than other state-of the-art topic models for human action recognition.

Qi, Wu, Du and Su have developed the clustering method and topic model to extract latent driving states, which can elaborate and analyses the commonness and individuality of driving behavior characteristics with the longitudinal driving behavior data collected by the instrumented vehicle [25]. For describing the driving behavior comprehensively, the methods propose the driving state as the subordinate unit of the driving style. The method develop the ensemble clustering method (ECM) based on the kernel fuzzy C-means algorithm (KFCM) for obtaining required data dimensional reduction and designed the modified LDA model for driving state mining and analysis. Leveraging longitudinal driving behavior data, the proposed model can achieve the better understanding of the commonness and individuality of driving behaviors with objective and comprehensive analysis.

Based on some related work described above, there is no study that analyzes the effect of fuzzy logic in handling uncertainty of the occurrence all topics in a document to the number of iteration in the reasoning, which will be described in this paper.

3 Research Method

A corpus consists of a set of documents with many topics. In each document, there is a set of words; each word can refer to a particular topic. It means a document does not only refer to a particular topic, but can refer to more than one topic with different percentages. Topics and its percentage value will be the features for each document. Therefore, the topic model of LDA is very appropriate to perform this feature extraction. In this paper, the reasoning process of LDA uses Gibbs sampling algorithms. The addition of fuzzy logic is performed

on Gibbs Sampling reasoning process. This research consists of several phases, such as data acquisition, pre-processing of data, feature extraction of FGLDA, and evaluation, as in Figure 1.

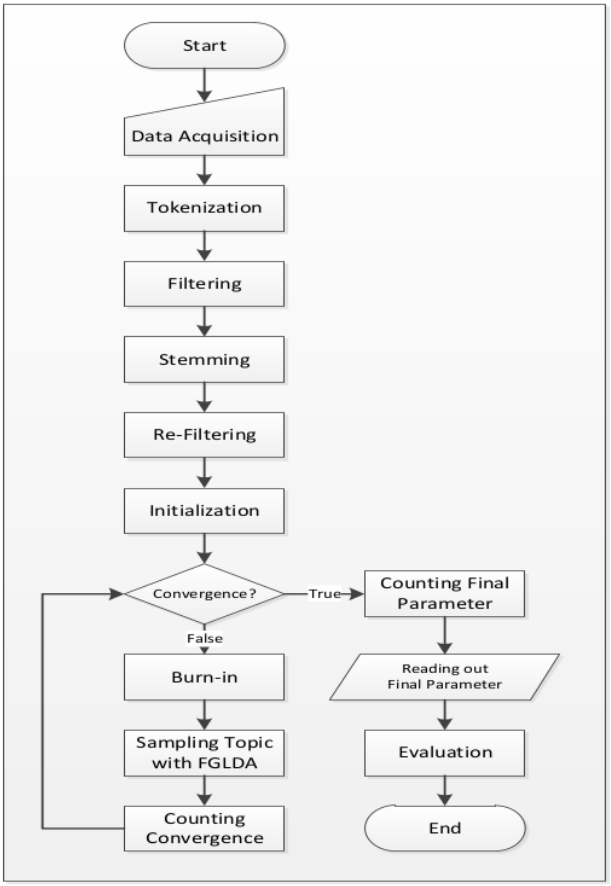


Figure 1. Research Design

3.1 Data Acquisition

In this paper, the data means "text of documents". Documents in this research are digital news which was acquired from Indonesian online news media. The number of documents collected was 1600 document, which is divided into three training groups of 100 documents, 500 documents and 1,000 documents.

The purpose of grouping the amount of data is to determine the effect of the number of documents to the number of iterations in the process of reasoning. Then, the digital news is stored in text files.

3.2 Pre-processing of Data

Some of words in text of news file have not meaningful for determining the value of a feature. Therefore, it is required to process the text before using it in the process of feature extraction. The process consists of several phases such as tokenization, filtering, stemming, and re-filtering.

Tokenization is used to decompose the text into the smallest unit which called "word" or "term". All the letters in the text is converted to lowercase. At this stage is also performed the process of removing punctuation, symbols or numbers that are not needed in the next process.

Filtering is a process to remove stop words such as prepositions, conjunctions, and words that are not meaningful. At the previous research, had been produced a list of 906 stop words that used in this research [24].

Stemming is process to find the root of each word, in other word, to find basic word. For Indonesian language, stemming has done by removing prefixes and suffixes. This process requires a dictionary of basic word and rules for prefixes-suffixes. At the previous research, had been produced a list of 30.342 basic words and five rules of prefixes and suffixes that used in this research [24].

Some of the basic words are generated by stemming included in the stop words list. Therefore, re-filtering process is required to remove those words from the pre-processing result. The results of re-filtering are the final result of the pre-processing stage which is a set of meaningful word for the process of feature extraction.

3.3 Feature Extraction of FGLDA

FGLDA algorithm consists of the initialization process, the burn-in process, the sampling process and the process of reading out final parameter, as in Figure 2 [12].

In the initialization process, FGLDA algorithm begins with the reading of a set of words from the result of re-filtering. Then, the algorithm calculated the number of occurrences each word on each document (term frequency - TF). The algorithm also makes fuzzy output curve based on the number of topics were set during the inference, as in Figure 3.

```
#inizialisation
create the value of fuzzy output curve
for all documents m do
    for all words n do
        count the value of term-frequency
        sampling topic index z0
        increment the number of document-topic (nmk)
        increment the number of topic-word (nkw)
        increment the sum of nmk (sumnmk)
        increment the sum of nkw    (sumnkw)
    end for
end for
#FGLDA sampling over burn-in period and sampling period
while not convergence
    for all documents m do
        for all words n do
            for all topics k do
                decrement the value of nmk
                decrement the value of nkw
                count the new value zi
                normalized zi with fuzzy logic (see Fig. 4)
                sampling new topic zi
                increment the value of nmk
                increment the value of nkw
            end for
        end for
    end for
    count the value of convergence
end while
#reading out final parameter
for all documents do
    for all words do
        count the value of parameter Φ
        count the value of parameter Θ
    end for
end for
```

Fig. 2. FGLDA Algorithm

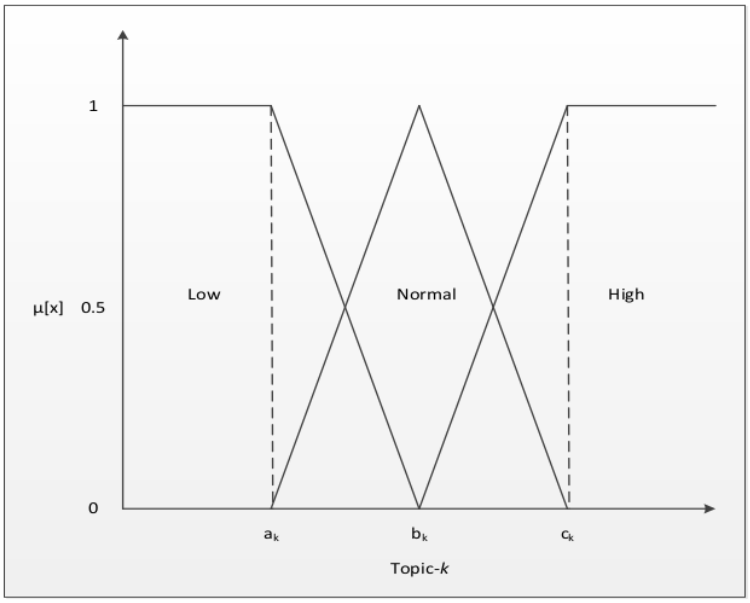


Fig. 3. Fuzzy Curve for Topic-*k*

In this paper, the number of topics is set to 10 topics. In the LDA algorithm, there is no initial value is assigned to each word in the document, thus making each word has uncertainty value. To overcome this problem, in this research, TF value becomes the initial value for each word in the document (z_o). Based on this initial value, the process of determining the topics of each word is done randomly by using a multinomial random number, as in (1) [9].

$$p(\vec{n}|\vec{p}) = \prod_{k=1}^{K} p_k^{n^{(k)}} = Mult(\vec{n}|\vec{p}, 1) \quad (1)$$

At the end of the initialization process, was done the process of calculating the number of each topic in the document (number of document topic-*nmk*) and the number of each word in the topic (number of topic word-*nkw*) that will used in the process of burn-in and sampling the topic. The total value of all *nmk* also calculated as *sumnmk* and the total value of all *nkw* as a *sumnkw*. The value of *sumnkm* and *sumnkw* used to subtract and add the value of *nmk* and *nkw* in any change of topic that occur on each word.

The process of burn-in and sampling the topic performed on several iterations to achieve the conditions of convergence. In this research, the convergence condition declared as the division between the values of the parameter α to β, as in (2). The value of α to β in LDA is a parameter on the level of corpus [2]. The value of α determine the mixing proportion of documents on the topics, while the

value of β determines the mixture components of words on the topics [12]. Therefore, the use of these values to achieve convergence is able to represent the model of topic mixture at the level of corpus, document and word, in accordance with the principles of LDA. The division operator used to reduce the threshold value to achieve convergence, so that the achieved results have a maximum evaluation value. At any iteration of inference, the process of calculating new value (z_i) performed for each word in each document, based on the value of *nmk*, *nkw*, α and β, as in (3) [12]. Furthermore, fuzzy logic process is carried out on a new value z_i. The value z_{new} of fuzzy logic process became a new probability value for each word in each document for its iteration, as in Figure 4. The value z_{new} used to determine the new topic of each word, which is done randomly by using a multinomial random number. This process is performed continuously until it reaches the convergence condition.

$$\|\sum_{n=1}^{N} z_i - \sum_{n=1}^{N} z_{i-1}\| \leq \frac{\alpha}{\beta} \tag{2}$$

$$p(z_i = k|\vec{z}_{\neg i}, \vec{w}) = \frac{n_{k,\neg i}^{(t)} + \beta}{\sum_{t=1}^{V} n_{k,\neg i}^{(t)} + W\,\beta} \frac{n_{m,\neg i}^{(k)} + \alpha}{\sum_{k=1}^{K} n_{m,\neg i}^{k} + K\,\alpha} \tag{3}$$

```
#fuzzification
count the degree of membership for z_i in fuzzy curve
#implication
for all fuzzy curve do
countα-predicate
    for all topics k do
        count the value of implication imp_i based on the value of fuzzy
        output curve and α-predicate
    end for
end for
#defuzzification
for all fuzzy output curve do
    for all topics k do
        count z based on the value of imp_i and α-predicate
    count the sum of z
    end for
end for
count z_new as division between z and sum of z
```

Fig. 4. Fuzzy Algorithm to Normalize z_i

The last process of the FGLDA algorithm is reading out final parameter. In this process, this algorithm calculates the number of documents for each topic (Φ) and the number of words for each topic (Θ), as in (4) and (5) [12].

These values will be used for subsequent retrieval process. In this paper, these values are used to measure the performance of FGLDA algorithm.

$$\boldsymbol{\varphi_{k,t}} = \frac{\boldsymbol{n_k^{(t)} + \beta}}{\boldsymbol{\sum_{t=1}^{V} n_k^{(t)} + \beta}} \tag{4}$$

$$\vartheta_{m,k} = \frac{n_m^{(k)} + \alpha}{\sum_{k=1}^{K} n_m^{(k)} + \alpha} \tag{5}$$

3.4 Evaluation of FGLDA

As mentioned previously, the purpose of this research is to analyze the effect of fuzzy logic in handling uncertainty of the occurrence all topics in a document that affect the number of iteration in the reasoning. Therefore, the last phase of this research is measuring the performance of the FGLDA algorithm. Based on its purpose, performance measurement conducted on the number of iterations produced at the stage of the reasoning for the three types of different amounts of data. This measurement will indicate two results: (1) Performance of FGLDA algorithm compare with LDA algorithm; (2) Effect of FGLDA algorithm performance towards number of documents.

Perplexity is a measurement of the model's ability to generalize the unseen data using the formula as in (6) and (7) [2, 12]. The smaller value of perplexity indicates the better performance of the algorithm. Variable N_m^t refers to the number of occurrences of the word t in document m.

$$\boldsymbol{P(\widetilde{W}|M) = \exp - \frac{\sum_{m=1}^{M} \log p(\vec{w}_m|M)}{\sum_{m=1}^{M} N_m}} \tag{6}$$

$$\boldsymbol{\log p(\vec{w}_m|M) = \sum_{t=1}^{V} N_m^t \log(\sum_{k=1}^{K} \varphi_{k,t} \cdot \vartheta_{m,k})} \tag{7}$$

Precision (P) is the fraction of retrieved documents that are relevant, as in (8); recall (R) is the fraction of relevant documents that are retrieved, as in (9); and F-Measure (F) is the weighted harmonic mean of precision and recall, as in (10) [21].

$$\boldsymbol{P(relevant|retrieved) = \frac{\#(relevant\ items\ retrieved)}{\#(retrieved\ items)}} \tag{8}$$

$$\boldsymbol{R(retrieved|relevant) = \frac{\#(relevant\ items\ retrieved)}{\#(relevant\ items)}} \tag{9}$$

$$\boldsymbol{F = \frac{2PR}{P+R}} \tag{10}$$

4 Experiment Result and Analysis

4.1 Experiment Result

In this research, experiments were performed on online news data with three groups of the number of documents, such as 100, 500, and 1000 documents. This group is used to determine the performance of FGLDA algorithm compared with LDA algorithm, and the effect of FGLDA algorithm performance towards the number of documents. The numbers of topic that are used in this research are 10 topics. The constant value for parameter β used in this research is 0.01, and the parameter α is $50/K$, where K is the number of topics [12, 26]. This performance of two algorithms is represented in graphs, where the x-axis stands for the number of documents and the y-axis represents the probability of each topic, as in Figure 5. The curve of convergence for the reasoning is achieved by the two algorithms as in Figure 6. The performance of two algorithms for the results of feature extraction, which was described in metric of Perplexity, Precision, Recall and F-Measure, as in Table 1, 2, and 3.

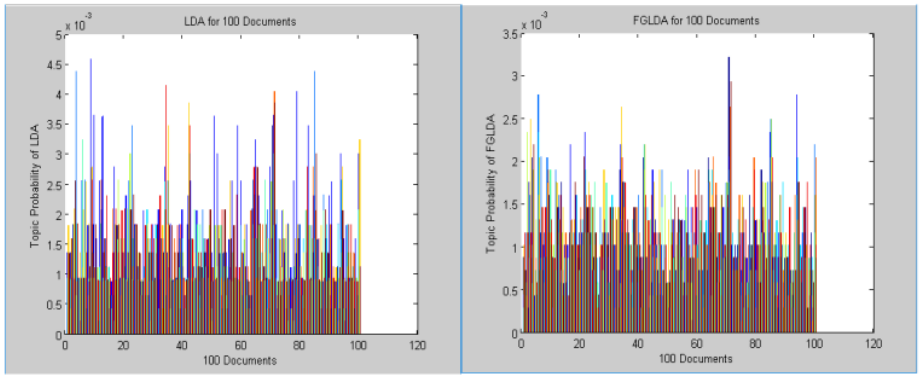
4.2 Analysis of Results

Results of experiments for FGLDA algorithm compared with LDA algorithm is represented in graph based on the number of documents processed.

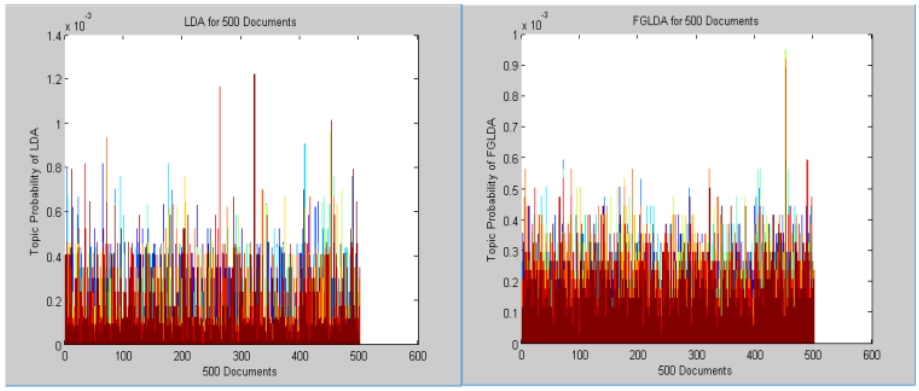
The analysis of graphs in Figure 5 which refers to 100 documents, 500 documents and 1,000 documents represents that:

1) LDA algorithm generates probability value that does not distributed to the whole document, in which the graph shows that there are many documents with probability values is tend to be equal, tends to be high, even tend to be very high.
2) FGLDA algorithm generates a probability value tend to be distributed to the whole document, in which the graph shows that there are only a few documents with probability values tend to be equal and very high, and there are several documents with a probability value tends to be low and high.

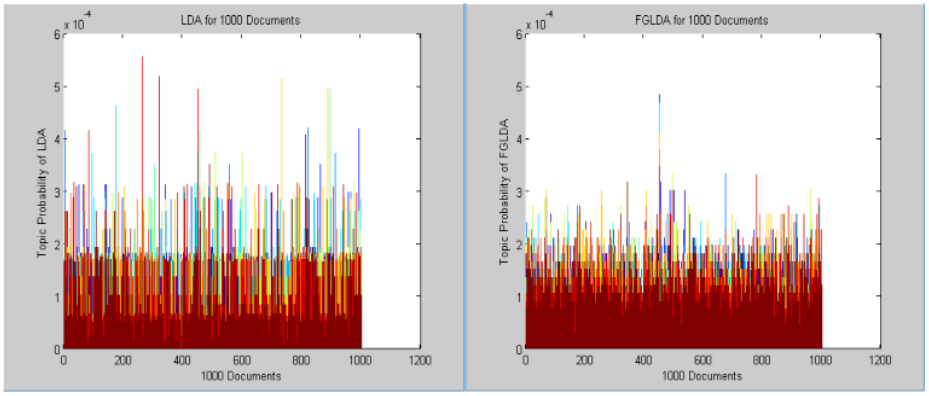
Thus, it can be concluded that FGLDA algorithm is capable of generating the probability distribution of all the topics in every document better than the LDA algorithm. This is in accordance with the principle of LDA that is resulted in the distribution of topics with the different percentage in every document.



(a)

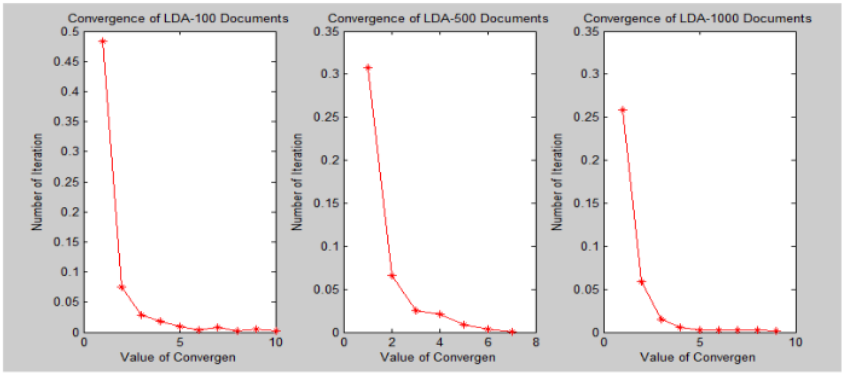


(b)

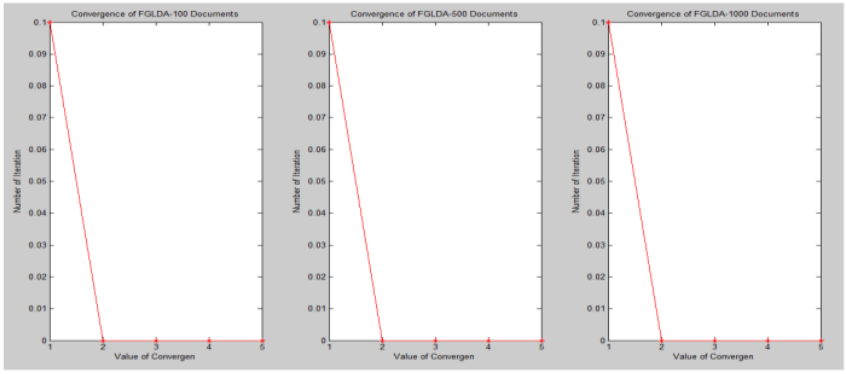


(c)

Figure 5. Topics Probability for (a) 100 Documents, (b) 500 Documents,



(a)



(b)

Figure 6. The curve of convergence for (a) LDA, (b) FGLDA

Table 1. Performance of Algorithm for 100 Documents

Method	Number of Iteration	Perplexity	Precision	Recall	F-Measure
LDA	10	0.1273	0.8690	0.7892	0.7770
FGLDA	2	0.1274	0.8975	0.8459	0.8387

Table 2. Performance of Algorithm for 500 Documents

Method	Number of Iteration	Perplexity	Precision	Recall	F-Measure
LDA	7	0.0589	0.8560	0.7756	0.7592
FGLDA	2	0.0590	0.8983	0.8490	0.8432

Table 3. Performance of Algorithm for 1000 Documents

Method	Number of Iteration	Perplexity	Precision	Recall	F-Measure
LDA	9	0.0376	0.8537	0.7697	0.7533
FGLDA	2	0.0376	0.8975	0.8486	0.8420

The analysis of Figure 6 which refers to 100 documents, 500 documents and 1,000 documents represents that FGLDA algorithm can reach convergence faster than LDA algorithm, which was indicated by the number of iteration.

The analysis of Table 1, 2, and 3 which refers to 100 documents, 500 documents and 1,000 documents, represents that:

1) LDA algorithm produces the number of iterations that is much greater than the FGLDA algorithm, indicates FGLDA algorithm have a better performance than LDA algorithm.

2) The number of documents does not affect the number of iterations is generated by FGLDA algorithm, while in LDA algorithm, the number of documents is affecting the number of iterations required in the process of reasoning.

3) Perplexity value between LDA and FGLDA algorithm is very small, indicating the two algorithms have the same performance in generalizing about unseen data.

4) The value of Precision, Recall and F-Measure for algorithms FGLDA is better than LDA algorithm, indicates FGLDA algorithm is more capable to retrieve the relevant documents.

Thus, it can be concluded that FGLDA algorithm has better performance than the LDA algorithm, based on the number of iterations, Perplexity, Precision, Recall and F-Measure.

5 Conclusions

FGLDA algorithm in this paper is a development of LDA algorithms, by adding the fuzzy logic in reasoning methods of Gibbs Sampling, to analyze the effect of fuzzy logic in handling uncertainty of the occurrence all topics in a document that affect the number of iteration in the reasoning. FGLDA algorithm is implemented on text data of Indonesian online news. The results of experiments on 100, 500, and 1000 documents, with a number of topics $K = 10$, the value of the parameter $\beta = 0.01$, and the value of the parameter $\alpha = 50/K$, indicate that FGLDA algorithm has better performance than the LDA algorithm, based on the number of iterations, Perplexity, Precision, Recall and F-Measure. FGLDA algorithm is capable of generating the probability distribution of all the topics in every document better than the LDA algorithm. FGLDA algorithm produces the number of iterations that is much smaller than the LDA algorithm. The number of documents does not affect the number of iterations is generated by FGLDA algo-

rithm. The value of Precision, Recall and F-Measure for algorithms FGLDA is better than LDA algorithm. Therefore, the FGLDA algorithm expected to be a better feature extraction method in modeling of topics with the capability of achieving faster convergence, thus saving of time and the use of computing resources for a very large number of documents.

In future work, this research will be developed to improve the application of fuzzy logic on LDA algorithm; so that it can achieve the performance value is much better than the results obtained in this paper.

References

[1] C. Archambeau, B. Lakshminarayanan and G. Bouchard, Latent IBP Compound Dirichlet Allocation, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **37** (2015), 321-333. https://doi.org/10.1109/tpami.2014.2313122

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

[3] X.-Q. Cao and Z.-Q. Liu, Type-2 Fuzzy Topic Models for Human Action Recognition, *IEEE Transactions on Fuzzy Systems*, **23** (2015), 1581-1593. https://doi.org/10.1109/tfuzz.2014.2370678

[4] H. Chen, L. Xie, C.-C. Leung, X. Lu, B. Ma and H. Li, Modeling Latent Topics and Temporal Distance for Story Segmentation of Broadcast News, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **25** (2017), 112-123. https://doi.org/10.1109/taslp.2016.2626965

[5] J.-T. Chien, Hierarchical Pitman–Yor–Dirichlet Language Model, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **23** (2015), 1259-1272. https://doi.org/10.1109/taslp.2015.2428632

[6] J.-T. Chien and C.-H. Chueh, Topic-Based Hierarchical Segmentation, *IEEE Transactions on Audio, Speech, and Language Processing*, **20** (2012), 55-66. https://doi.org/10.1109/tasl.2011.2143405

[7] G.N. Corrêa, R.M. Marcacini, E.R. Hruschka and S.O. Rezende, Interactive textual feature selection for consensus clustering, *Pattern Recognition Letters*, **52** (2015), 25-31. https://doi.org/10.1016/j.patrec.2014.09.008

[8] G. Cosma and M. Joy, An Approach to Source-Code Plagiarism Detection and Investigation Using Latent Semantic Analysis, *IEEE Transactions On Computers*, **61** (2012), 379-394. https://doi.org/10.1109/tc.2011.223

[9] W.M. Darling, A Theoretical and Practical Implementation Tutorial on Topic Modeling and Gibbs Sampling, School of Computer Science, University of Guelph, 2011.

[10] Y. Gao, Y. Xu and Y. Li, Pattern-based Topics for Document Modelling in Information Filtering, *IEEE Transactions on Knowledge and Data Engineering*, **27** (2015), 1629-1642. https://doi.org/10.1109/tkde.2014.2384497

[11] L. Gong, R. Yang, Q. Yan and X. Sun, Prioritization of Disease Susceptibility Genes Using LSM/SVD, *IEEE Transactions on Biomedical Engineering*, **60** (2013), 3410-3417. https://doi.org/10.1109/tbme.2013.2257767

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

[13] A. Ihler and D. Newman, Understanding Errors in Approximate Distributed Latent Dirichlet Allocation, *IEEE Transactions on Knowledge and Data Engineering*, **24** (2012), 952-960. https://doi.org/10.1109/tkde.2011.29

[14] S. Ji, W. Zhang and R. Li, A Probabilistic Latent Semantic Analysis Model for Coclustering the Mouse Brain Atlas, *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, **10** (2013), 1460-1468. https://doi.org/10.1109/tcbb.2013.135

[15] 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*, **9** (2014), 31-43. https://doi.org/10.1109/mci.2013.2291689

[16] Y. Li, A. Algarni, M. Albathan, Y. Shen and M.A. Bijaksana, Relevance Feature Discovery for Text Mining, *IEEE Transactions on Knowledge and Data Engineering*, **27** (2015), 1656-1669. https://doi.org/10.1109/tkde.2014.2373357

[17] Y. Li, X. Zhou, Y. Sun and H. Zhang, Design and Implementation of Weibo Sentiment Analysis Based on LDA and Dependency Parsing, *China Communications*, **13** (2016), 91-105. https://doi.org/10.1109/cc.2016.7781721

[18] C.-L. Liu, W.-H. Hsaio, C.-H. Lee and H.-C. Chi, An HMM-Based Algorithm for Content Ranking and Coherence-Feature Extraction, *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, **43** (2013), 440-450. https://doi.org/10.1109/tsmca.2012.2207104

[19] 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*, **27** (2015), 636-650. https://doi.org/10.1109/tkde.2014.2339850

[20] X. Luo, J. Zhang, F. Ye, P. Wang and C. Cai, Power Series Representation Model of Text Knowledge Based on Human Concept Learning, *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, **44** (2014), 86-102. https://doi.org/10.1109/tsmcc.2012.2231674

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

[22] M. Morchid, M. Bouallegue, R. Dufour, G. Linarès, D. Matrouf and R.D. Mori, Compact Multiview Representation of Documents Based on the Total Variability Space, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **23** (2015), 1295-1308. https://doi.org/10.1109/taslp.2015.2431854

[23] H. Noh, Y. Jo and S. Lee, Keyword selection and processing strategy for applying text mining to patent analysis, *Expert Systems with Applications*, **42** (2015), 4348-4360. https://doi.org/10.1016/j.eswa.2015.01.050

[24] P.M. Prihatini and I.K. Suryawan, Text Processing Application Development for Indonesian Documents Clustering, in The 1st International Joint Conference on Science and Technology (IJCST), Bali, Indonesia, 2016.

[25] G. Qi, J. Wu, Y. Du and M. Xu, Leveraging longitudinal driving behaviour data with data mining techniques for driving style analysis, *IET Intelligent Transport Systems*, **9** (2015), 792-801. https://doi.org/10.1049/iet-its.2014.0139

[26] M. Steyvers and T. Griffiths, Probabilistic Topic Models, Chapter in *Latent Semantic Analysis: A Road to Meaning*, D.M. T. Landauer, S. Dennis, and W. Kintsch, Editor, Laurence Erlbaum, 2006.

[27] M. Tutkan, M.C. Ganiz and S. Akyokuş, Helmholtz principle based supervised and unsupervised feature selection methods for text mining, *Information Processing & Management*, **52** (2016), 885-910. https://doi.org/10.1016/j.ipm.2016.03.007

[28] W. Wang, H. Xu and W. Wan, Implicit feature identification via hybrid association rule mining, *Expert Systems with Applications*, **40** (2013), 3518-3531. https://doi.org/10.1016/j.eswa.2012.12.060

[29] C.-H. Wu, H.-P. She and C.-S. Hsu, Code-Switching Event Detection by Using a Latent Language Space Model and the Delta-Bayesian Information Criterio, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **23** (2015), 1892-1903. https://doi.org/10.1109/taslp.2015.2456417

[30] Z. Zhao, X. He, D. Cai, L. Zhang, W. Ng and Y. Zhuang, Graph Regularized Feature Selection with Data Reconstruction, *IEEE Transactions on Knowledge and Data Engineering*, **28** (2016), 689-700. https://doi.org/10.1109/tkde.2015.2493537

[31] F. Zhuang, P. Luo, Z. Shen, Q. He, Y. Xiong, Z. Shi and H. Xiong, Mining Distinction and Commonality across Multiple Domains Using Generative Model for Text Classification, *IEEE Transactions on Knowledge and Data Engineering*, **24** (2012), 2025-2039. https://doi.org/10.1109/tkde.2011.143

Received: March 25, 2017; Published: April 21, 2017

Fuzzy-Gibbs Latent Dirichlet Allocation Model for Feature Extraction on Indonesian Documents

ORIGINALITY REPORT

17%	13%	15%	4%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Chien, Jen-Tzung, and Chuang-Hua Chueh. "Topic-Based Hierarchical Segmentation", IEEE Transactions on Audio Speech and Language Processing, 2012. Publication	1%
2	Xiangfeng Luo, Jun Zhang, Feiyue Ye, Peng Wang, Chuanliang Cai. "Power Series Representation Model of Text Knowledge Based on Human Concept Learning", IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2014 Publication	1%
3	www.arxiv-vanity.com Internet Source	1%
4	thesai.org Internet Source	1%
5	hal.archives-ouvertes.fr Internet Source	1%

6	Submitted to University of Illinois at Urbana-Champaign Student Paper	1 %
7	trid.trb.org Internet Source	1 %
8	"Computational Science and Technology", Springer Science and Business Media LLC, 2021 Publication	1 %
9	www.ieeeproject.in Internet Source	1 %
10	Submitted to Telkom University Student Paper	1 %
11	www.ics.uci.edu Internet Source	1 %
12	core.ac.uk Internet Source	1 %
13	doc.lagout.org Internet Source	1 %
14	I Made Bayu Permana Putra, Rukmi Sari Hartati, I Ketut Gede Darma Putra, Ni Kadek Ayu Wirdiani. "Optimized back propagation learning in neural networks with bacterial foraging optimization to predict forex gold index (XAUUSD)", Applied Mathematical Sciences, 2014	<1 %

Publication

15	gcris.ktun.edu.tr Internet Source	<1%
16	Submitted to National University of Ireland, Galway Student Paper	<1%
17	Ihler, A., and D. Newman. "Understanding Errors in Approximate Distributed Latent Dirichlet Allocation", IEEE Transactions on Knowledge and Data Engineering, 2012. Publication	<1%
18	eprints.qut.edu.au Internet Source	<1%
19	Rizka W. Sholikah, Agus Zainal Arifin, Diana Purwitasari, Chastine Fatichah. "Co-occurrence technique and dictionary based method for Indonesian thesaurus construction", 2017 5th International Conference on Information and Communication Technology (ICoIC7), 2017 Publication	<1%
20	Submitted to University of Surrey Student Paper	<1%
21	Mahmoud Oglah Al Hasan Baniata, Sohail Asghar. "Sentiment Analytics: Extraction of Challenging Influencing Factors from COVID-	<1%

19 Pandemics", Intelligent Automation & Soft Computing, 2021
Publication

22	technodocbox.com Internet Source	<1%
23	Ayse Cisel Aras, Ismail Gocer. "Driver Rating based on Interval Type-2 Fuzzy Logic System", IFAC-PapersOnLine, 2016 Publication	<1%
24	Ying Liu, Wei Song, Lizhen Liu, Hanshi Wang. "Document representation based on semantic smoothed topic model", 2016 17th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD), 2016 Publication	<1%
25	cps-vo.org Internet Source	<1%
26	docplayer.net Internet Source	<1%
27	icsgteis.unud.ac.id Internet Source	<1%
28	ojs.unud.ac.id Internet Source	<1%
29	www.deepdyve.com Internet Source	

		<1%
30	www.ijartet.com Internet Source	<1%
31	Laurence Hirsch, Masoud Saeedi, Robin Hirsch. "EVOLVING TEXT CLASSIFICATION RULES WITH GENETIC PROGRAMMING", Applied Artificial Intelligence, 2005 Publication	<1%
32	lxie.nwpu-aslp.org Internet Source	<1%
33	"Anticipating Future Innovation Pathways Through Large Data Analysis", Springer Science and Business Media LLC, 2016 Publication	<1%
34	Submitted to Assumption University Student Paper	<1%
35	Yunmei Liu, Min Chen. "The Knowledge Structure and Development Trend in Artificial Intelligence Based on Latent Feature Topic Model", IEEE Transactions on Engineering Management, 2023 Publication	<1%
36	jctjournal.com Internet Source	<1%
37	pnrsolution.org	

	Internet Source	<1%
38	www.acsij.org Internet Source	<1%
39	www.ijrar.com Internet Source	<1%
40	Zeng, Jia, William K. Cheung, and Jiming Liu. "Learning Topic Models by Belief Propagation", IEEE Transactions on Pattern Analysis and Machine Intelligence, 2012. Publication	<1%
41	mafiadoc.com Internet Source	<1%

Exclude quotes	On	Exclude matches	Off
Exclude bibliography	On		