International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-8 Issue-2, July 2019

# Enhancing Focus Topic Findings of Discussion Forum through Corpus Classifier Algorithm

**Reina Setiawan, Widodo Budiharto, Iman Herwidiana Kartowisastro, Harjanto Prabowo**

*Abstract*: ***In learning management system, a discussion forum, in which the students and lecturers are involved actively as part of the learning method, enriches the context of communication, thereby enhancing the students' learning and performance. The aim of this paper was to determine the appropriate topics for a discussion forum for learning management systems through enhanced probabilistic latent semantic analysis (PLSA) with the corpus classifier algorithm. In preparing the paper, the methods used were PLSA and the classifying process, which classifies the documents to become a corpus based on the similarity word approach. The similarity word is influenced by the term-frequency of the word in the document. The novel concept in this paper is the corpus classifier algorithm. The experiment was conducted using three approaches to discover the topic, and it used 4,868 distinct words from 234 documents. The documents were contained in three threads subject. The post of the discussion forum is the text document. The performance of the result was measured by the f-measure, which was calculated for each thread subject. The corpus classifier algorithm was used in the second approach, and third approach increased the average f-measure values for the second and third thread subjects by approximately 24 and 17%, respectively.***

***Keywords: Corpus Classification, Discussion forum, PLSA, Similarity word, Topic findings***

## I. INTRODUCTION

The use of information technology (IT) is growing rapidly and impacting the daily lives of most people. IT is defined as using the computer and other gadgets to manage information [1]. Many fields are affected by IT, including the educational field, where, in daily activities, teachers apply IT to improve the teaching and learning processes [2].

**Revised Manuscript Received on 30 July 2019.**
***** Correspondence Author

**Reina Setiawan***, Computer Science Department, BINUS Graduate Program – Doctor of Computer Science, Computer Science Department, School of Computer Science, Bina Nusantara University, Jakarta, Indonesia.

**Widodo Budiharto**, Computer Science Department, BINUS Graduate Program – Doctor of Computer Science, Computer Science Department, School of Computer Science, Bina Nusantara University, Jakarta, Indonesia.

**Iman Herwidiana Kartowisastro**, Computer Science Department, BINUS Graduate Program – Doctor of Computer Science, Computer Engineering Department, Faculty of Engineering, Bina Nusantara University, Jakarta, Indonesia.

**Harjanto Prabowo**, Computer Science Department, BINUS Graduate Program – Doctor of Computer Science, Management Department, BINUS Business School - Undergraduate, Bina Nusantara University, Jakarta, Indonesia.



From an educational perspective, IT has created a new method of teaching and learning, which is known as the Learning Management System (LMS). In the LMS, a discussion forum is a medium that is used by lecturers and students to learn and to discuss the topics of teaching and learning material with each other [3], [4], [5]. The discussion forum, as part of teaching and learning methods, is to enrich the context of communication and enhance students' learning and their performances in the course; in the discussion forum, both students and lecturers are involved actively [6], [7]. In the discussion forum, a discussion is initiated by introducing a thread subject to be discussed. Ideally, the content of a thread subject discusses only one topic. However, more than one topic can be posted in a thread subject, but this has an impact on the topic of the reply. Hence, it may become into several topics. The topic is defined as an information focus, and it is used to state the focus of the content of the discussion [8], [9].

Prior research about topic modeling includes the software framework [10], unsupervised topic modeling [11], cross-language that uses interlingual topic modeling [12], automatic evaluation [13], latent topic modeling [14], [15], leveraging unstructured information [16], and the dynamic discovery of a topic [17]. There has been no research related to topic modeling of a discussion forum in a learning management system. Since a post in a discussion forum has not been edited, the use of Probabilistic Latent Semantic Analysis (PLSA) as unsupervised learning is appropriate for this situation, especially discussion forums in the Indonesian language. PLSA is used to identify a latent variable, which is called the "topic" [18]. In the Indonesian language, an incomplete sentence is a part of the type of an Indonesian sentence [19], [20]. Thus, it is possible that such a sentence could be used as a part of a post. In recent years, there has been increased research based on PLSA. PLSA has been implemented in several fields, such as land cover classification analysis [21], human motion analysis [22], and word disambiguation analysis [23].

In this paper, we focused on the implementation of topic modeling from PLSA to determine the appropriate topic of discussion for the posts in a forum. The characteristics of the discussion posted on a forum influence the focus of the topic findings from PLSA. Therefore, we proposed an algorithm to enhance the focus of the topic findings of each post. The algorithm considers the term-frequency of distinct words in a post as a text document, and it classifies the document based on the similarity of a number of distinct words $(n)$ from a number of certain words with highest term-frequency $(m)$ throughout all of documents in the corpus.

This classification is required to classify the documents so that they become a group of corpora before being processed by PLSA to determine the appropriate topic.

In this paper, we used the experiments that processed 4,868 distinct words from 234 documents within three subject threads of a course. Three approaches were used in conducting the experiments. The 234 documents were from a discussion forum of the learning management system of Bina Nusantara University. For the most part, the documents were written in the Indonesian language, but some words were English, and some words were abbreviations, such as IT for Information Technology or BP for Business Process. The results of the accuracy topics of each document from these three approaches were compared using the f-measure. The precision, recall, and f-measure were measured for each document, and the f-measure was averaged for each thread subject. The results proved that the appropriateness of the topics was enhanced by the corpus classifier algorithm.

The paper consists of the following sections. The section on materials and methods describes the methods used in the paper, such as the eight steps of PLSA; the corpus classifier algorithm as the new algorithm proposed, including the mathematical model; and the experimental design. The results and discussion section explain and discuss the results of the experiment. Our conclusions are presented in the last section of the paper.

## II. MATERIALS AND METHODS

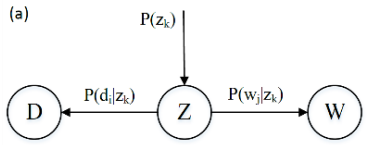
### A. Probabilistic Latent Semantic Analysis

Probabilistic Latent Semantic Analysis is a statistical approach to identify and distinguish the context of words using latent semantic variable (i.e. topic), and it is known as an aspect model [24]. Thomas Hofmann introduced PLSA, and it can be used to retrieve information [25]. PLSA contains two types pf parameterization, i.e., symmetric and asymmetric parameterization. Fig. 1 shows graphical models of symmetric and asymmetric parameterization [18]. The symmetric parameterization model is expressed by:

$$P(d_i, w_j) = \sum_{k=1}^{K} P(z_k) P(d_i \mid z_k) P(w_j \mid z_k) \quad (1)$$

where:

- $P(z_k)$ is the probability of class-conditional in particular class variable, $z_k$.
- $P(d_i \mid z_k)$ is the probability of class-conditional of a particular document conditioned on unobserved class variable, $z_k$ and a particular document is identified by $d_i$.
- $P(w_j \mid z_k)$ is the probability of class-conditional of a specific word conditioned on unobserved class variable, $z_k$ and a specific word is symbolized by $w_j$.




(b)

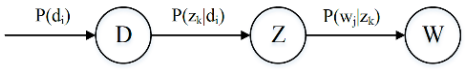



Fig. 1. (a) Symmetric Parameterization; (b) Asymmetric Parameterization

The asymmetric parameterization model is expressed by:

$$P(d_i, w_j) = P(d_i) P(w_j \mid d_i) \quad (2)$$

and

$$P(w_j \mid d_i) = \sum_{k=1}^{K} P(w_j \mid z_k) P(z_k \mid d_i) \quad (3)$$

where:

- $P(d_i)$ is the probability of a word's occurring in a particular document $d_i$.
- $P(w_j \mid z_k)$ is the probability of class-conditional of a specific word conditioned on unobserved class variable, $z_k$ and a specific word is identified as $w_j$.
- $P(z_k \mid d_i)$ is a document's specific probability distribution over the latent variable space.

Since the number of topics is smaller than the number of words and the number of documents, we used asymmetric parameterization in this paper.

The eight steps in the asymmetric parameterization of PLSA are presented in this paper. The first step is to prepare a matrix to save the term-frequency $(tf)$ of distinct words for each document. The size of the matrix is $J \times I$, where $J$ is the number of words in the corpus, and $I$ is the number of documents in the corpus. Therefore, the size of the matrix depends on the corpus, and, in this paper, there is a variety of corpuses, depending on the three approaches mentioned in Introduction.

The second step is to prepare a matrix to save probability of word of topic, and it is initialized with random numbers and is symbolized by $w_j z_k$. This probability is denoted by $P(w_j \mid z_k)$ and the values are obtained using (4).

$$P(w_j \mid z_k) = w_j z_k \div \sum_{j=1}^{J} w_j z_k \quad (4)$$

The size of the matrix is $J \times K$, where $J$ is the number of distinct words in the corpus, and $K$ is the number of topics. The variable $j$ is an index of words, and the variable $k$ is an index of topics, and they range from one up to the value of $J$ and $K$, respectively.

The third step is to prepare a matrix to save the probability of topic of the document, and it also is initialized with random numbers that are symbolized by $z_k d_i$. This probability is denoted by $P(z_k \mid d_i)$, and the values are obtained using (5).

$$P(z_k \mid d_i) = z_k d_i \div \sum_{k=1}^{K} z_k d_i \quad (5)$$

The size of the matrix is $K \times I$, where $K$ is the number of topics and $I$ is the number of documents in the corpus. The variables $k$ and $i$ are an index of topics and an index of documents that range from one up to the values of $K$ and $I$, respectively. The fourth step is to prepare a matrix to save probability of word of document, and it is initialized to zero. This probability is denoted as $P(w_j \mid d_i)$, and the values are obtained using (6).

$$P(w_j \mid d_i)_{n+1} = P(w_j \mid d_i)_n + P(w_j \mid z_k) \times P(z_k \mid d_i) \quad (6)$$

The size of the matrix is $J \times I$, where $J$ and $I$ are the number of distinct words and the number of documents in the corpus, respectively. The variable $n$ denotes the current iteration, so $n+1$ means the next iteration. The number of iterations depends on the number of topics.

The fifth step is the estimation step (E-Step), and three-dimensional matrices are prepared to save the probability of topic of word and document, and they are initialized to zero. This probability is denoted as $P(z_k \mid d_i, w_j)$, and the values are obtained using (7).

$$P(z_k \mid d_i, w_j) = P(w_j \mid z_k) \times P(z_k \mid d_i) \div P(w_j \mid d_i) \quad (7)$$

The size of matrix is $K \times J \times I$, where $K$ is the number of topics, $J$ is the number of distinct words in the corpus, and $I$ is the number of documents in the corpus.

The sixth step is the maximization step (M-Step) to update the probability topic of the document using (8), followed by (5). In (8), $n$ is the current value of probability topic of document, and $n+1$ is the next value of the probability topic. The term $tf_{ji}$ is the term-frequency of a specific word.

$$P(z_k \mid d_i)_{n+1} = P(z_k \mid d_i)_n + \sum_{j=1}^{J} tf_{ji} \times P(z_k \mid w_j, d_i) \quad (8)$$

The seventh step is the maximization step (M-Step), and it is used to update the probability word of the topic using (9), followed by (4). In (9), $n$ is the current value of probability word of the topic, and $n+1$ is the next value of the probability word.

$$P(w_j \mid z_k)_{n+1} = P(w_j \mid z_k)_n + \sum_{i=1}^{I} tf_{ji} \times P(z_k \mid w_j, d_i) \quad (9)$$

The eighth step is the maximization step (M-Step) in which the probability word of document is updated using (6).

### B. Corpus Classifier Algorithm

In this paper, we propose a new approach in which a layer is used before PLSA to determine the topic. This layer is needed to group the documents that have similar words based on the term-frequency of the words. The proposed algorithm, which is called the corpus classifier algorithm, is part of the layer, and the novelty of this paper is that it is used to improve the result of the appropriateness of the topic from PLSA. Fig. 2 shows a model of this approach. The corpus classifier algorithm classifies documents based on how many similar words they contain and classifies certain words with the highest term-frequency $(tf)$ into one group. The number of similar words $n$ and the number of words in the highest term-frequency $m$ are determined initially by the user. In this paper, $n$ and $m$ were determined to be 2 and 5, respectively. First, the algorithm creates the list $tf$ of distinct words, maps the words for each document in a corpus, and then sorts $tf$ in a descending manner. Second, the algorithm creates groups of documents based on the number of similar words that were provided as input in the beginning. This similarity occurs in certain words with the highest term-frequency. A document in a corpus is illustrated as a set of words, and it contains $m$ words, e.g., the first document and the second document are denoted as $d_1 = \{word_1, word_2, ..., word_m\}$ and $d_2 = \{word_1, word_2, ..., word_m\}$, respectively.

Therefore, the document can be expressed by (10):

$$d_i = \{word_1, word_2, ..., word_m\} \quad (10)$$

The similarity of the documents can be expressed by (11):

$$sim(d_A, d_B) = \begin{cases} 1, if\ (d_A \cap d_B) and (|\ d_A \cap d_B\ | \geq n) \\ 0, otherwise \end{cases} \quad (11)$$

where:

- $sim(d_A, d_B)$ is the similarity between two of the documents
- $n$ is the number of similar words within the highest term-frequency $m$ words

The value of the similarity is one if there is an intersection between the two sets and if the number of elements in the intersection is greater than or equal to $n$, meaning that the similarity is 'yes'. However, the meaning of the zero value is not similar. The process to check the similarity is in a module, entitled "check_similarity."The pseudo code of the corpus_classifier_algorithm is shown by Algorithm 1, and the pseudo code of the check_similarity module, as part of the corpus classifier algorithm, is shown by Algorithm 2. The maximum iterations required to check the similarity is expressed by (12).

$$\max_{iter} = \sum_{i=1, m>i}^{m-1} (m-i) \quad (12)$$

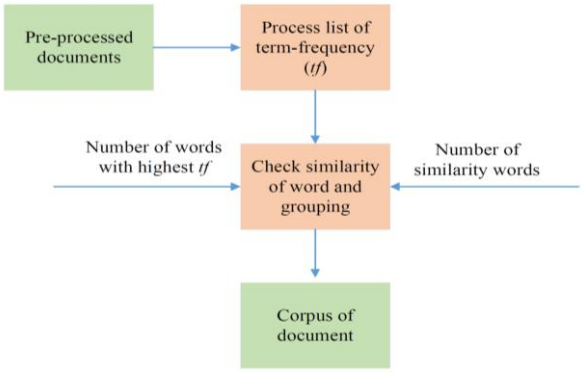



**Fig. 2. Model of the Corpus Classifier Approach**

**Algorithm 1** corpus_classifier_algorithm

```
input m as number of words with highest tf
input n as number of similar words
foreach counter = 1 to Length(a_thread_subject) do
    create list tf of distinct word and mapping word
    sort tf by descending mode
    if first document then
        save as a group
    else
        do check_similarity
        if similarity is yes then
            save the group_number
        else
            save as a group
end
```

**Algorithm 2** check_similarity module

```
foreach group_number = 1 to Length(group_number) do
    flag ← 0
    foreach counter = 1 to Length(m) do
        if word in list of m words then
            increment flag
    end
    if flag ≥ n then
        return similarity is yes
end
return similarity is no
```

### C. Experimental Design

Fig. 3 shows the design of the experiment described in this paper. The first step is to conduct pre-processing of the text document. One text document is one post. The second step is to determine the topic with PLSA directly. The third step is to classify the text document with the corpus classifier algorithm so that it becomes a group classification and then continue to determine the topic with PLSA. The fourth step is the process of document classification. The process is to combine the classification of groups into one group so that it contains only one text document and then continue to determine the topic with PLSA. The process is done based on a thread subject.

In order to measure the accuracy of the topic, the next process is to calculate the precision, recall, and f-measure for each document and to calculate the average f-measure of the three approaches. Precision, recall, and the f-measure are used to measure the relevance of the topic of the document [26]. In this paper, all of the programs use the Python programming language.

## III. RESULT AND DISCUSSION

In this paper, we described our experiments with 4,868 distinct words from 234 documents. The documents are posts of a discussion forum concerning the learning management system at Bina Nusantara University. The posts are text documents that, for the most part, are written in the Indonesian languages.

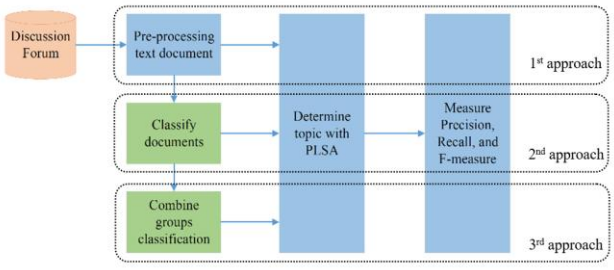



**Fig. 3. Experimental Design**

The documents consisted of three threads subjects, and there were 1,721 distinct words from 82 documents, 1,729 distinct words from 79 documents, and 1,418 distinct words from 73 documents in the first thread subject, second thread subject, and third thread subject, respectively. All of the documents were from one course, entitled "Information System Concept". The posts of the discussion forum were extracted from the database, and the processing included tokenization, stop-word removal, and stemming. Since most of the discussion in the forum was in the Indonesian language, the stop-word list used in this paper is in that language. The stop-word list was selected from Tala with the addition of several common words [27]. There were 927 general words or symbols in the stop-word list. The stemming also used an algorithm for the Indonesian language. It used a flexible affix classification algorithm to remove affixes and determined the root word [28]. Pre-processing of the information that was retrieved was required to improve the accuracy of the results, [26], [29].

Three approaches were used in the experiment. In the first approach, after pre-processing the documents, the experiment was continued to determine the topic with PLSA. There were eight steps in the asymmetric parameterization of PLSA. In this approach, the corpus was based on the thread subject, so there were three corpora. In the second approach, after pre-processing the documents, the experiment was continued to classify the documents using the corpus classifier algorithm. After the documents were classified into several groups, the process was continued to determine the topic with PLSA. In this approach, there were three corpora in the first approach, there were 25, 13, and 27 corpora from the first, second, and third thread subjects, respectively. These corpora were based on the thread subjects and grouped by the corpus classifier algorithm. The third approach was similar to the second approach, but the corpus was determined in a different manner. In this approach, the documents were grouped into only one document comprising one corpus. Thus, in this approach, there were 11 corpora, 8 corpora, and 17 corpora from the first, second, and third thread subjects, respectively.

The results of these experiments were measured by their precision, recall, and f-measure. The 234 documents were read by the people involved in the experiment, and they defined the topics of each document. One document can have more than one topic. Furthermore, the topics of each document, which were determined through PLSA of the three approaches, were compared to the topics identified in the manual process. Thus, the precision, recall, and f-measure of each document were calculated. The average value of the f-measure of each thread subject was used as a parameter to measure the performance of the three approaches. The average values of the f-measures obtained from the three approaches were compared. All of the algorithms in this paper were coded by the Python programming language.

In the first approach of the experimental design, three corpora were used in the experiment, and they were defined based on the thread subject. The result of the topic findings and the average values of the f-measures from the experiment are provided in Table 1. The topic results are '*usaha*', '*informasi*', and '*komputer*' meaning '*company*', 'information', and 'computer', respectively. The topic finding from all of the documents in the third corpus was

Table 1. Topic Result of the First Approach

| Thread subject | Corpus | Number of documents | Topic findings | F-measure |
|---|---|---|---|---|
| 1 | 1 | 82 | *usaha*, *informasi* | 61% |
| 2 | 2 | 79 | software, *komputer* | 40% |
| 3 | 3 | 73 | *informasi* | 22% |

'information'. This topic finding was too general, so the average f-measure was too low, i.e., only 22%.

The findings of the topic of the second approach are represented by the first thread subject and the second thread subject, as shown in Tables 2 and 3. Since some topics are in Indonesian, to facilitate the understanding of the topics, the words in the brackets in the Topic column are written in English. Tables 2 and 3 show that, in the second approach, the topic generated from the PLSA was more specific and there was less variance than in the first approach.

Table 2. Results of the topic findings of the first thread subject of the second approach

| Corpus | Number of documents | Topic findings |
|---|---|---|
| 1 | 5 | *strategi*, *organisasi*, *usaha* (strategic, organization, company) |
| 2 | 29 | *informasi*, system (information, system) |
| 3 | 6 | *organisasi*, IT (organizational, Information Technology) |
| 4 | 14 | *usaha*, *produk* (company, product) |
| 5 | 1 | IT, *usaha* (Information Technology, company) |
| 6 | 1 | BP, functional, cross (Business Process, functional, cross) |
| 7 | 1 | *informasi*, *dukung*, computer (information, support, computer) |
| 8 | 2 | pressures, information, *sebut* (pressures, information, mention) |
| 9 | 1 | *bisnis*, *definisi*, davenport (business, definition, davenport) |
| 10 | 1 | *aktivitas*, *bisnis*, *tuju* (activity, business, goal) |
| 11 | 1 | information, advantage, IT (information, advantage, Information Technology) |
| 12 | 1 | advantage, *organisasi*, *mesin* (advantage, organization, mechine) |
| 13 | 1 | *aktivitas*, business, *terima* (activity, business, receive) |
| 14 | 2 | *usaha*, *ancam* (company, threat) |
| 15 | 1 | *hasil*, strategic, *mudah* (result, strategic, easy) |
| 16 | 3 | *kuat*, *produk*, *beli* (strong, product, buy) |
| 17 | 1 | pressure, *yg*, *langgan* (pressure, which, customer) |
| 18 | 2 | process, business (process, business) |
| 19 | 2 | customer, product, to (customer, product, to) |
| 20 | 2 | activities, value, buyer (activities, value, buyer) |
| 21 | 1 | ``, market (``, market) |
| 22 | 1 | business, organization, the (business, organization, the) |
| 23 | 1 | *langgan*, *tekan* (customer, pressure) |
| 24 | 1 | *simpan*, data (save, data) |
| 25 | 1 | *usaha*, competitive (company, competitive) |

Table 3. Results of the topic findings of the second thread subject of the second approach

| Corpus | Number of documents | Topic findings |
|---|---|---|
| 1 | 6 | service, web (service, web) |
| 2 | 23 | software, server, computing (software, server, computing) |
| 3 | 19 | input, computer (input, computer) |
| 4 | 15 | software (software) |
| 5 | 4 | computing, *layan*, cloud (computing, service, cloud) |
| 6 | 1 | best, *terimakasih*, regards (best, thank you, regards) |
| 7 | 2 | service (service) |
| 8 | 1 | computer, *yg*, computers (computer, which, computers) |
| 9 | 1 | komputer (computer) |
| 10 | 4 | software (software) |
| 11 | 1 | server, computer (server, computer) |
| 12 | 1 | server (server) |
| 13 | 1 | *organisasi*, service, web (organization, service, web) |

The average f-measure of the second approach is shown in Table 4. The average f-measure of the second approach was better than that of the third approach. The explanation is that the combination of several corpora did not provide better results, especially if there was no classification used, as was done in the first approach. The average f-measures of the second and third thread subjects increased significantly compared to the first approach, in which the decreased f-measure of the first thread subject was not significant. The f-measures for the second and third thread subjects were approximately 24 and 17%, respectively.

Most of the posts in the discussion forum were in the Indonesian language, but several words and sentences were in English. Since the stop-word list in this paper is in Indonesian, some common English words, such as 'the', 'is', 'are', 'you', and 'regards' were not removed, and they arose as topics. This condition occurred because these common words were mentioned often in the document and affected the value of term-frequency. Thus, there is a possibility that they could occur as topics. In Table 2, there is a special character, i.e., "", which could arise as a topic. This condition occurred because the stop-word list did not yet cover all special characters of symbols.

In the third approach, there was a combination of several corpora that were contained in only one document. In the first, second, and third thread subjects, there were 15, 6, and 12 corpora, respectively, and they were combined as one corpus. Therefore, the numbers of corpora of the first, second, and third thread subjects were 11, 8, and 17. The average f-measure of the third approach is shown in Table 5.

Table 4. Average f-measure of the second approach

| Thread subject | Number of corpora | F-measure |
|---|---|---|
| 1 | 25 | 51% |
| 2 | 13 | 64% |
| 3 | 27 | 39% |

Table 5. Average f-measure of the third approach

| Thread subject | Number of corpora | F-measure |
|---|---|---|
| 1 | 11 | 54% |
| 2 | 8 | 61% |
| 3 | 17 | 37% |

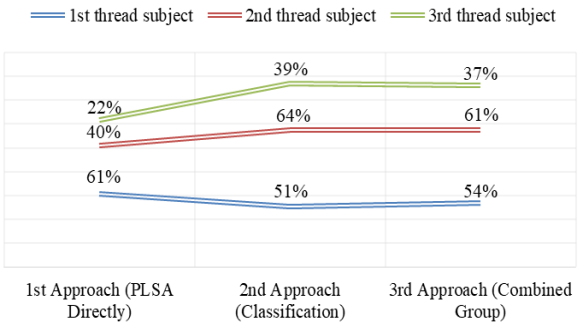



**Fig. 4. Comparison of the average f-measure values**

Fig. 4 shows a comparison of the performances of the three approaches. The performance was measured based on the average f-measure. Fig. 4 shows that the second and third approaches have increased f-measure values, whereas there was an insignificant decrease in the f-measure of the first thread subject. The corpus classifier algorithm showed enhancement in terms of its ability to find topics based on the average value of the f-measure.

## IV. CONCLUSION

The Corpus Classifier Algorithm was proven to increase the performance of topic modelling from PLSA. It increased the f-measure by approximately 24 and 17% for second and third thread subjects, respectively. Even though there was a decrease of about 10% for the first thread subject, the percentage was lower than the percentage increases of the other threads subject. The variety of posts in the discussion forum can be classified using the corpus classifier algorithm. Another aspect that must be considered is the stop-word list. Since the posts in the discussion forum can be written in several languages, the stop-word list also must contain the common words of these languages. Based on the analysis, these common words influence the topic result because they influence the value of term-frequency. The term-frequency has significant impact in process of PLSA and in the process of preparing the classification document. Another consideration is to include special characters or symbols on the stop-word list.

## ACKNOWLEDGMENT

This is a prior research of another research that has been published in Education and Information Technologies on March 27, 2019 with title "Finding Model through Latent Semantic Approach to Reveal the Topic of Discussion in Discussion Forum". Because of some reasons, this research has been delayed in publication.

## REFERENCES

1. A. K. Chowdhury and V. Shanmugan, "Information Technology: Impacts on Environment and Sustainable," *Pertanika J. Sci. Technol.*, vol. 23, no. 1, pp. 127–139, 2015.
2. M. N. Khambari, W. S. Luan, A. Fauzi, and M. Ayub, "Promoting Teachers' Technology Professional Development through Laptops," *Pertanika J. Soc. Sci. Humanit.*, vol. 20, no. 1, pp. 137–145, 2012.
3. J. Schoonenboom, "Using An Adapted, Task-Level Technology Acceptance Model to Explain Why Instructors in Higher Education Intend to Use Some Learning Management System Tools more than Others," *Comput. Educ.*, vol. 71, pp. 247–256, 2014.
4. M. S. Kuran, J. M. Pedersen, and R. Elsner, "Learning Management Systems on Blended Learning Courses: An Experience-Based Observation," in *International Conference on Image Processing and Communications*, 2017, pp. 141--148.
5. A. A. Piña, *An Educational Leader's View of Learning Management Systems*. Springer, 2018.
6. M. S. Balaji and D. Chakrabarti, "Student Interactions in Online Discussion Forum: Empirical Research from 'Media Richness Theory' Perspective," *J. Interact. Online Learn.*, vol. 9, no. 1, 2010.
7. C. K. Cheng, D. E. Paré, L. M. Collimore, and S. Joordens, "Assessing the effectiveness of a voluntary online discussion forum on improving students' course performance," *Comput. Educ.*, vol. 56, no. 1, pp. 253–261, 2011.
8. J. K. Gundel and T. Fretheim, "Topic and Focus," *Handb. Pragmat.*, vol. 175, pp. 1–19, 2004.
9. T. Saracevic, "Relevance: A review of and a framework for the thinking on the notion in information science," *J. Am. Soc. Inf. Sci.*, vol. 26, no. 6, pp. 321–344, 1975.
10. R. Rehurek and P. Sojka, "Software Framework for Topic Modelling with Large Corpora," in *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, 2009.
11. M. Purver, P. K. Konrad, J. B. Tenenbaum, and T. L. Griffiths, "Unsupervised Topic Modelling for Multi-Party Spoken Discourse," in *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL*, 2006, no. July, pp. 17–24.
12. W. De Smet, "Cross-Language Linking of News Stories on the Web using Interlingual Topic Modelling," in *Proceedings of the 2nd ACM workshop on Social web search and mining*, 2009, pp. 57–64.
13. D. Newman, J. H. Lau, K. Grieser, and T. Baldwin, "Automatic Evaluation of Topic Coherence," in *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 2010, no. June, pp. 100–108.
14. B. Chen, "Latent Topic Modeling of Word Co-occurrence Information for Spoken Document Retrieval," in *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009*, 2009, no. 2, pp. 3961–3964.
15. C. Zhai, "Probabilistic Topic Models for Text Data Retrieval and Analysis," in *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2017, pp. 1399–1401.
16. J. W. Uys, N. D. Preez, and E. W. Uys, "Leveraging Unstructured Information Using Topic Modelling," in *PICMET 2008 Proceedings*, 2008, no. c, pp. 27–31.
17. N. Li, W. Luo, K. Yang, F. Zhuang, Q. He, and Z. Shi, "Self-organizing Weighted Incremental Probabilistic Latent Semantic Analysis," *Int. J. Mach. Learn. Cybern.*, vol. 0, no. 0, pp. 1--12, 2017.
18. T. Hofmann, "Unsupervised learning by probabilistic Latent Semantic Analysis," *Mach. Learn.*, vol. 42, no. 1–2, pp. 177–196, 2001.
19. H. Alwi, S. Dardjowidjojo, H. Lapoliwa, and A. M. Moeliono, *Tata Bahasa Baku Bahasa Indonesia*. Jakarta: Balai Pustaka, 2003.
20. Kushartanti, U. Yuwono, and M. R. M. T. Lauder, *Pesona Bahasa Langkah Awal Memahami Linguistik*. Gramedia Pustaka Utama, Jakarta, 2007.
21. J. Shi, X. Tian, Z. Jiang, D. Zhao, and M. Lu, "Sparsity-constrained probabilistic latent semantic analysis for land cover classification," in *Geoscience and Remote Sensing Symposium (IGARSS)*, 2016, no. 61071137, pp. 5453–5456.

22. J. Wang, P. Liu, M. F. H. She, A. Kouzani, and S. Nahavandi, "Neurocomputing Supervised learning probabilistic Latent Semantic Analysis for human motion analysis," *Neurocomputing*, vol. 100, pp. 134–143, 2013.
23. G. S. Tomar, M. Singh, S. Rai, A. Kumar, R. Sanyal, and S. Sanyal, "Probabilistic Latent Semantic Analysis for Unsupervised Word Sense Disambiguation," *IJCSI Int. J. Comput. Sci. Issues*, vol. 10, no. 5, pp. 127–133, 2013.
24. C. Hong, W. Chen, W. Zheng, J. Shan, Y. Chen, and Y. Zhang, "Parallelization and Characterization of Probabilistic Latent Semantic Analysis," in *Parallel Processing, 2008. ICPP'08. 37th International Conference on*, 2008, pp. 628–635.
25. T. Hofmann, "Probabilistic Latent Semantic Analysis," in *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*, 1999, pp. 289--296.
26. R. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval: the concept and technology behind search*, Second. Addison-Wesley Professional Harlow, 2011.
27. F. Z. Tala, "A Study of Stemming Effects on Information Retrieval in Bahasa Indonesia," *M.Sc. Thesis, Append. D*, vol. pp, pp. 39–46, 2003.
28. R. Setiawan, A. Kurniawan, W. Budiharto, I. H. Kartowisastro, and H. Prabowo, "Flexible affix classification for stemming Indonesian Language," in *2016 13th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology, ECTI-CON 2016*, 2016, pp. 1--6.
29. C. D. Manning, P. Raghavan, and H. Schutze, *An Introduction to Information Retrieval*. Online edition (c) 2009 Cambridge UP, 2009.

## AUTHORS PROFILE

**Reina Setiawan** is a faculty member of Computer Science Department of Bina Nusantara University. She obtained her bachelor degree, major in Information Management from STMIK Bina Nusantara, Jakarta-Indonesia in 1996, her master degree, major in Management from Pelita Harapan University, Banten-Indonesia in 2005, and her Ph.D from BINUS Graduate Program - Doctor of Computer Science, Bina Nusantara University, Jakarta-Indonesia. Her research interest is in information retrieval, data mining, and text processing. words.

**Widodo Budiharto** is a professor of Artificial Intelligence at School of Computer Science, Bina Nusantara University, Jakarta-Indonesia. He obtained his bachelor degree, major in physics from Indonesia University, Jakarta-Indonesia. He continued his study in information technology major at STT Benarif, Jakarta-Indonesia and obtained his Master in Information Technology. He obtained his Ph.D. in Electrical Engineering from Institute of Technology Sepuluh Nopember, Surabaya-Indonesia. He took his Ph.D. Sandwich Program in Robotics at Kumamoto University, Japan and Postdoc in Robotics and Artificial Intelligence at Hosei University, Japan. He took offering as a visiting professor at Erasmus Mundus French Indonesian Consortium (FICEM)-French, Hosei University-Japan, and Erasmus Mundus Scholar EU Universite de Bourgogne-French in 2017, in 2016, in 2007, respectively. His research interest is in intelligence systems, data science, robot vision, and computational intelligence.

**Iman Herwidiana Kartowisastro** is Quality Assurance Director and Provost, BINUS Higher Education. He obtained his B.Sc. in Electronics & Telecommunication from Trisakti University, Jakarta-Indonesia in 1986. He then continued his study at City, University of London-UK and obtained his M.Sc. in Information Engineering in 1987. Still at the same university, he then obtained his Ph.D. in Robotics Control. In 2000 he got his postgraduate diploma in Business Administration from De Montfort University-UK. His research interest is in robotics control, vision and intelligence.

**Harjanto Prabowo** is a professor of Information System Management and Rector of Bina Nusantara University. He earned his doctoral degree in business management with cum laude from University of Padjadjaran, Bandung-Indonesia, Magister Management Information System from Bina Nusantara University and Engineer degree, major in electrical Engineering (best graduate) from Diponegoro University, Semarang-Indonesia. His research focus is in the strategic knowledge management and innovation areas. Currently, he is chairman of information system compartment of APTISI (Association of Indonesia Private Universities), member of KADIN (Indonesian chamber of commerce and industry), an active member of APTIKOM (Higher education association of information technology and computer science), ISEI (Indonesian Economist Association), International data warehousing association, American Society for Quality (ASQ), and American Marketing Association (AMA). He has published numerous research paper in both international and Indonesian journals and has been a speaker in various conference and seminars (for higher education and universities), especially conferences relating to implementation of quality management systems, information system, good governance, quality management and global competitiveness.