Education and Information Technologies (2020) 25:31–50
https://doi.org/10.1007/s10639-019-09901-7

Finding model through latent semantic approach to reveal the topic of discussion in discussion forum

Reina Setiawan [1,2] · Widodo Budiharto [1,2] · Iman Herwidiana Kartowisastro [1,3] · Harjanto Prabowo [1,4]

Received: 3 September 2018 / Accepted: 12 March 2019 / Published online: 27 March 2019

Abstract

There are lots of information and knowledge can be extracted from a discussion forum. Despite a discussion is opened by submitting a thread as the topic of discussion, however, the discussion may open out to different topics. This paper aims to present a model to find out a topic of discussion through latent semantic approach, named Topics Finding Model (TFM). The model proposes a complete step to reveal the topic of discussion from a thread in a discussion forum, consisting of the pre-processing text document, corpus classification and finding a topic. The model can be applied in various discussion forums and various languages with a few adjustments, such as stop-word removal list and stemming algorithm. The data were obtained from discussion forum in a learning management system. The data consist of 1050 posts divided into three different course subjects: information systems, management, and character building. The reason for using several course subjects is to observe consistency of the model. F-measure was used to measure the effectiveness of the model, and the results showed that the TFM was consistent and effective to reveal the topic of discussion, with a good precision. However, the recall can still be increased in a further study.

Keywords Topics finding model · Latent semantic · Discussion forum · Learning management system

✉ Reina Setiawan
reina@binus.edu

[1] Computer Science Department, BINUS Graduate Program - Doctor of Computer Science, Bina Nusantara University, Jl. K. H. Syahdan No. 9, Jakarta 11480, Indonesia

[2] Computer Science Department, School of Computer Science, Bina Nusantara University, Jl. K. H. Syahdan No. 9, Jakarta 11480, Indonesia

[3] Computer Engineering Department, Faculty of Engineering, Bina Nusantara University, Jl. K. H. Syahdan No. 9, Jakarta 11480, Indonesia

[4] Management Department, BINUS Business School – Undergraduate Program, Bina Nusantara University, Jl. K. H. Syahdan No. 9, Jakarta 11480, Indonesia

1 Introduction

In the present-day social environment, the discussion forum has become an increasingly popular tools to communicate and to share information among members. Through discussion forum, members update their knowledge about new things. At present, the discussion forum has rapidly become a part of Learning Management System (LMS) and part of Massive Open Online Course (MOOC) (Piña 2018; Kuran et al. 2017; Ruano et al. 2016). It means lots of knowledge and information arise in discussion forum. However, the posts are only known among the members. Recently, to improve content relevancies, some discussion forums have also included other external parties (e.g. from industries) to become part of the discussion forum members like lecturers. They can also raise questions to the students. The problem is how to retrieve the information or knowledge from discussion forums. Hence, this will in turn enrich the teaching learning process.

In the last five years, several researches about discussion forum have emerged, focusing on the association between discourse behavior and students' learning (Wang et al. 2015), effect of confusion in discussion forum (Yang et al. 2015), sentiment analysis of MOOC discussion forum (Wen et al. 2014) and unsupervised classification method to understand student posts (Ezen-can et al. 2015). On the other hand, a few works focus on the extraction information from the discussion. To enrich research in discussion forum field as research object, this study focuses on topic extraction of discussion forum posts using latent semantic. The topic become a label of post to retrieve the information and knowledge from discussion forum. This study will complement other people's works.

The study proposes a model for clustering posts based on the topic of discussion through latent semantic approach. The model is named Topics Finding Model (TFM), which is a new approach for the clustering posts. However, the characteristics of discussion forum have become challenging in a research environment. One of the challenges is by posting a discussion forum by a member without editing process, revealing that the post may consist of unstructured statements with some grammatical errors. Another characteristic of a discussion forum is about the topic of discussion. When a discussion is opened by a thread from a member; ideally, the thread ought to focus on one topic, however, the discussion may be opened out to other topics, which may diverge the members. The language used in a discussion forum is also another characteristic. Although there is a specific language to be used, several slangs may be used on several occasions. For example, if a discussion is in Indonesian language, members might use several English slangs during discussions. Thus, the latent semantic approach is used to handle the characteristics of these discussion forums.

Using LMS for the experiments, the study collects data from 1050 posts and divided them into three different course subjects: information systems, management and character building. The reason for using the course subjects in three different areas, computing, social and behavioral area, is to observe the consistency of the model. Actually, the language used in the discussion forum is the Indonesian language, and the effectiveness of the model is measured by an F-measure parameter. The result shows that the TFM is consistent and effective in revealing the topic of discussion.

The rest of the paper is organized as follows: Section 2 discusses the previous research related to this paper. Section 3 explains the proposed model and method. Section 4 discusses the evaluation and the result. Section 5 provides the conclusion.

2 Related works

2.1 Information retrieval (IR)

There are two scopes of research in IR. The first research is about how to index document. The second research is about document retrieval (Baeza-Yates and Ribeiro-Neto 2011; Sanderson and Croft 2012). This study focused on how to index in IR. The index is based on topic of discussion. In this research, to find out the topic, language modeling approach is used. In recent, many researches about modeling for information retrieval has arisen. There is a smoothing method for language modeling. This model used word probability estimation. Equation (1) is general form of smoothed model (Zhai and Lafferty 2017).

$$p(w|d) = \begin{cases} p_s(w|d), & if\ \ word\ w\ is\ seen \\ \propto_d p(w|\mathbb{C}), & otherwise \end{cases} \tag{1}$$

where the smoothed probability of a word seen in document is denoted by $p_s(w|\,d)$. The smoothed probability is a probability to adjust the maximum likelihood estimator of a language model. The collection language model is denoted by $p(w|\,\mathbb{C})$, meanwhile the coefficient controlling of probability mass assigned to unseen words is denoted by $\propto_d$.

Another approach used statistical to find a posteriori most likely documents given the query based on Bayes' law as Eq. (2). The d for which $p(d|\,q,\ U)$ is highest posteriori, q is the query, and U is the user's distill (Berger and Lafferty 1999).

$$p(d|q,U) = \frac{p(q|d,U)\,p(d|U)}{p(q|U)} \tag{2}$$

In this study the language modeling using latent semantic approach based on the probability of latent variable to find out the topic of discussion that can be used as label to index and to retrieve the document. The latent semantic approach is an approach to find out information from a text document, based on certain entities through latent variable. Meanwhile, the latent variable is an association between unobserved class variable with each observation based on co-occurrence data. The latent variable is adopted from a generative model from Probabilistic Latent Semantic Analysis (PLSA) (Hofmann 1999).

2.2 Corpus classification

Corpus classification is a process of classifying documents in a specific corpus based on certain approach. Several previous researches focus on clustering or classifying a corpus, using scatter/gather to cluster large corpus (Cutting et al. 2017), context semantic analysis (Benedetti et al. 2018), cluster word importance-based similarity

(Botev et al. 2017), cluster machine learning for text categorization (Sailaja et al. 2018) and cluster corpus classifier algorithm (Setiawan et al. 2019).

Ideally, a thread in discussion forum may discuss a topic, however the discussion can grow to other topics. The corpus classifier algorithm aims to identify the shortcoming that causes the variety topics discussion in a thread. In this approach, the similarity of documents is classified based on the similarity of words with 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 $d_1 = \{word_1, word_2, \cdots, word_m\}$ and $d_2 = \{word_1, word_2, \cdots, word_m\}$, respectively. Therefore, the i^{th} document containing m words in the corpus is expressed by Eq. (3):

$$d_i = \{word_1, word_2, \cdots, word_m\} \tag{3}$$

The similarity of documents is expressed in Eq. (4), and Fig. 1 shows the model of corpus classification approach (Setiawan et al. 2019). The algorithm needs two inputs, number of word with highest term-frequency denoted by m and number of similarity word denoted by n. The similarity between document A and document B can be expressed by Eq. (4) as follows:

$$sim(d_A, d_B) = \begin{cases} 1, if\ ((d_A \cap d_B)\ and\ (|d_A \cap d_B| \geq n)) \\ 0, otherwise \end{cases} \tag{4}$$

where:

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

The value of similarity is one, if the two conditions are fulfilled. The first, there is intersection between the two documents and the second is number of intersection element must be greater or equal than n. Otherwise, the value of similarity is zero.

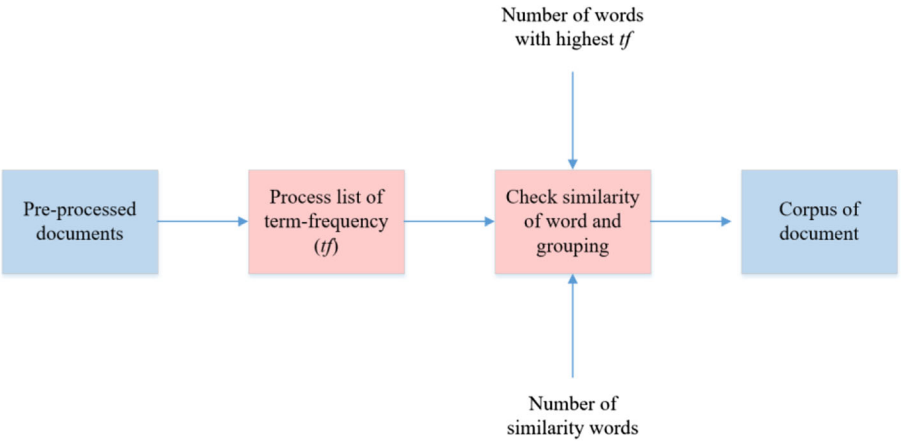


Fig. 1 The model of corpus classification approach (Setiawan et al. 2019)

The similar documents of classification result are defined as a corpus. The corpus is more specific and focused rather than a corpus based on a thread discussion. There is another model about corpus development in previous research. This model is based on Naïve Bayes, SVM and J48 with term weighting scheme ranking (Utomo and Bijaksana 2016). The model of corpus classification is based on similarity of words with highest term-frequency among documents. These models have different approach.

2.3 Probabilistic latent semantic analysis (PLSA)

Probabilistic Latent Semantic Analysis (PLSA) is a statistical approach to find out a categorized topic based on latent semantic through co-occurrence data analysis (Hong et al. 2008). PLSA is an aspect model introduced by Thomas Hofmann (Hofmann 1999), and it can be used for information retrieval. The aspect model associates co-occurrence with data in an unobserved class variable (topic), which an observation is occurrent of a word in a particular document (Hofmann 2001). Figure 2 shows the general structure of PLSA and describes its association among documents, topics and words. The probability $P(z|\,d)$ and $P(w|\,z)$ links topic layer to documents and words, respectively (Dan Oneata 1999).

The aspect model has two types: asymmetric parameterization and symmetric parametrization. The asymmetric parameterization is used when a number of topics is smaller than several documents and number of words (K << N, D).

A generative model for document and word co-occurrences with a joint probability is expressed by:

$$P(d_i, w_j) = P(d_i)P(w_j|d_i) \tag{5}$$

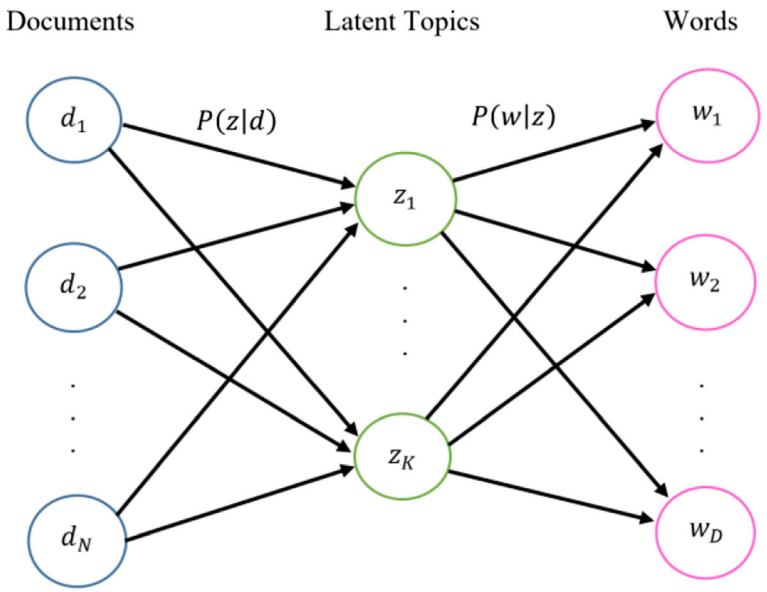


Fig. 2 The general structure of PLSA model (Dan Oneata 1999)

which

$$P(w_j|d_i) = \sum_{k=1}^{K} P(w_j|z_k)P(z_k|d_i) \quad (6)$$

The explanation of the symbols is:

$P(d_i)$	probability of a word occurrence in a particular document d_i
$P(w_j\|z_k)$	probability of class-conditional of a specific word conditioned on unobserved class variable z_k
$P(z_k\|d_i)$	a document specific probability distribution over the latent variable space

A generative model for document and word co-occurrences with a joint probability is expressed by:

$$P(d_i, w_j) = \sum_{k=1}^{K} P(z_k)P(d_i|z_k)P(w_j|z_k) \quad (7)$$

The explanation of the symbols is:

$P(z_k)$	probability of class-conditional in particular class variable z_k
$P(d_i\|z_k)$	probability of class-conditional of a particular document conditioned on unobserved class variable z_k
$P(w_j\|z_k)$	probability of class-conditional of a specific word conditioned on unobserved class variable z_k

The latent semantic approach based on PLSA approach, since one of discussion forum characteristics is ignored in the editing process. Thus, a statistical approach is relevant for these characteristics.

3 Proposed model

This paper proposes a latent semantic approach to find out the topic of discussion from a discussion forum. This approach is packaged in a model, named Topics Finding Model (TFM) as shown in Fig. 3. The TFM aims to find out topics of discussion in a corpus through three steps. A corpus is a set of posts of discussion, whereas a post in a discussion forum is a text document. The model consists of three steps: Pre-processing document, Corpus classification, and Finding topic. In pre-processing document, there are three activities: tokenization, stop-word removal, and stemming. The stemming process uses flexible affix classification approach, the stemming algorithm for Indonesian language (Setiawan et al. 2016). The corpus is obtained from discussion forum of Bina Nusantara University's learning management system and not publicly accessible. Since teaching and learning process in Bina Nusantara University uses Indonesian language, thus the discussion forum uses Indonesian language as well. The empty posts of discussion forum have been removed from the corpus. It is to ensure that the corpus meets the research object needs. The corpus is not validated by the authors, however, the corpus and stemming result are validated and verified by Language Center of Bina Nusantara University as an independent party.

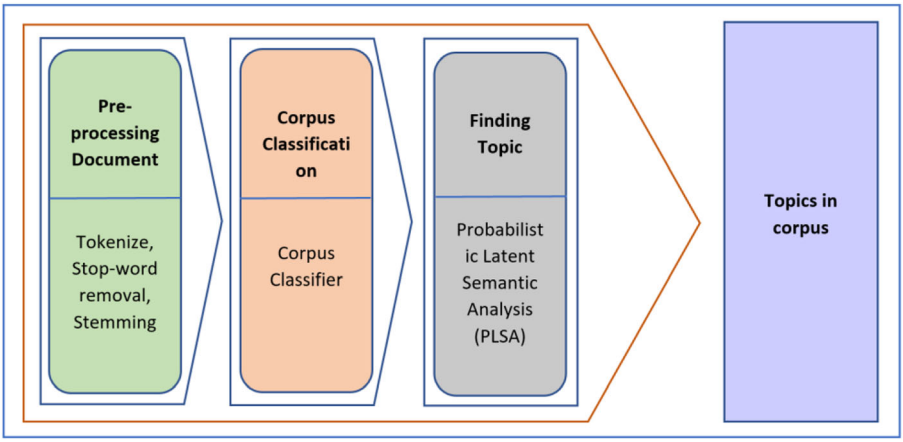


Fig. 3 The topics finding model (TFM)

In corpus classification step, the corpus classifier algorithm is used. Afterwards, in finding topic, Probability Latent Semantic Analysis (PLSA) approach is used.

The Topics Finding Model depicted in Fig. 3 above can be elaborated in the following equations:

- **Pre-processing Document step**:

In this step, the text document undergoes tokenization process, stop-word removal based on stop-word list and stemming. Equations (8) and (9) is mathematical model to show text document in tokenization process and stemming process, respectively. Meanwhile, Corpus contains of various stemmed text documents as illustrated in Eq. (10).

$$D_i = \{T_1, T_2, \cdots, T_p\} \tag{8}$$

where:

D denotes an original text document
i denotes number of original text documents
T denotes a token in original text document
p denotes number of tokens

$$d_j = \{w_1, w_2, \cdots, w_l\} \tag{9}$$

where:

d denotes a stemmed text document
j denotes number of stemmed text documents
w denotes a stemmed distinct word in text document
l denotes number of stemmed distinct words

$$C = \{d_1, d_2, \cdots, d_i\} \tag{10}$$

where C denotes a corpus that contains certain d

- **Corpus Classification step**:

In this step, the corpus is classified based on similar distinct words with highest term-frequency in several documents. There are two parameters in this step, i.e. m and n is number of words with highest term-frequency and number of similarity word, respectively. For the sake of convenience, Eq. (4) in Section 2, is rewritten as Eq. (11).

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

where:

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

The value of similarity is one, if the two conditions are fulfilled. The first condition, there is intersection between the two documents and the second is number of intersection element must be greater or equal than n. Otherwise, the value of similarity is zero.

- **Finding Topic step**:

There are eight steps to find the topic in corpus as follow explained:

1. Prepare a matrix to save term-frequency of the distinct word for each document. The term-frequency is a number of distinct words that occur in a document and denoted by *tf*. The matrix is created with size $J \times I$, where J is the number of distinct words in a corpus and I is the number of documents. Thus, tf_{11} means the number of first distinct word, which the first document occurs.
2. Prepare a matrix of the probability of a word of topic $P(word|\ topic)$. The size of the matrix is $J \times K$; J is the number of distinct words in a corpus and K is the number of topics. The values of the matrix are initialized with a random number and normalized using Eq. (12) as the probability. The normalization process aims to attain weight of word based on topic. The probability and the random number are symbolized by $P(w_j|\ z_k)$ and $w_j z_k$, respectively. The number of topics should be defined previously. Thus, $P(w_1|\ z_1)$ means probability a word w_1 become a topic z_1.

$$P(w_j|z_k) = \frac{w_j z_k}{\sum_{j=1}^{J} w_j z_k} \tag{12}$$

3. Prepare a matrix of the probability of topic of document $P(topic|\ doc)$. The size of the matrix is $K \times I$; K is the number of topics and I is the number of documents in the corpus. Similar with $P(word|\ topic)$, the values of the matrix are initialized with a random number and normalized using Eq. (13) as the probability. The normalization process aims to attain weight of topic based on document. The probability

and the random number are symbolized by $P(z_k|\, d_i)$ and $z_k d_i$, respectively. Thus, $P(z_1|\, d_1)$ is the probability of a topic z_1, which is part of the document d_1.

$$P(z_k|d_i) = \frac{z_k d_i}{\sum_{k=1}^{K} z_k d_i} \tag{13}$$

4. Prepare a matrix of the probability of a word of document $P(word|\, doc)$. The size of the matrix is $J \times I$; J is the number of distinct words in corpus and I is the number of documents. The values of the matrix are initialized with zeroes and the probability is defined with the Eq. (14). The probability is symbolized by $P(w_j|\, d_i)$. In equation, n denotes current iteration, therefore $n + 1$ means the next iteration. The number of topics determine the number of iterations.

$$P\left(w_j|d_i\right)_{n+1} = P\left(w_j|d_i\right)_n + P\left(w_j|z_k\right) \times P(z_k|d_i) \tag{14}$$

5. Prepare a matrix of the probability of the topic, given word and document $P(topic|\, doc,\ word)$. The size of the matrix is $K \times J \times I$; K is the number of topics, J is the number of distinct words in the corpus, and I is the numbers of documents in the corpus. The probability is symbolized by $P(z_k|\, d_i,\ w_j)$ and obtained in Eq. (15). This step is an estimation step which compute posterior probabilities for the latent variables.

$$P\left(z_k|d_i, w_j\right) = P\left(w_j|z_k\right) \times P(z_k|d_i)/P\left(w_j|d_i\right) \tag{15}$$

6. Update the probability of the topic of document $P(topic|\, doc)$ in Eq. (16) and followed by Eq. (13). This step is a maximization step to update $P(z_k|\, d_i)$.

$$P(z_k|d_i)_{n+1} = P(z_k|d_i)_n + \sum_{j=1}^{J} tf_{ji} \times P\left(z_k|w_j, d_i\right) \tag{16}$$

7. Update the probability of the word of topic $P(word|\, topic)$ in Eq. (17) and followed by Eq. (12). This step is a maximization step to update $P(w_j|\, z_k)$.

$$P\left(w_j|z_k\right)_{n+1} = P\left(w_j|z_k\right)_n + \sum_{i=1}^{I} tf_{ji} \times P\left(z_k|w_j, d_i\right) \tag{17}$$

8. The last step is a maximization step to update the probability of a word of document $P(word|\, doc)$ in Eq. (14). In the maximization step, the matrix of term-frequency impacts the update calculation of $P(topic|\, doc)$ and $P(word|\, topic)$, thus term-frequency of the distinct word influences the result of the topic.

The illustration of fourth step to eighth step are shown in Figs. 4, 5, 6 and 7. These illustrations explain step by step to find out topics and give a clear understanding. Figure 4 represents the calculation of probability of a word of document and part of maximization step as well. The $P(word|\, doc)$ is obtained from accumulation process of multiplication between $P(word|\, topic)$ and $P(topic|\, doc)$ as declared in Eq. (14). Figure 5 visualizes the estimation step. Eventually, Figs. 6 and 7 depict the

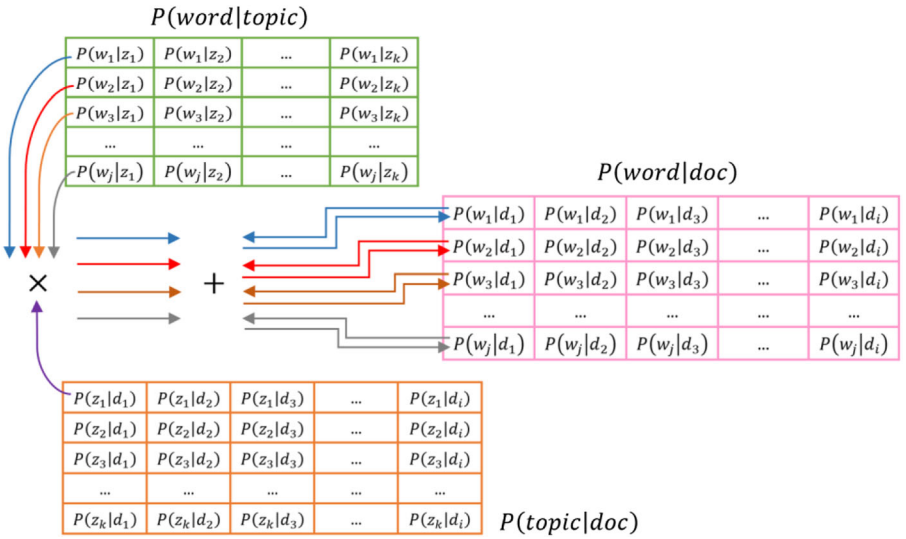


Fig. 4 The calculation of $P(word|\, doc)$ for word w_1 to w_j in document d_1 of topic z_1

maximization steps. In Figs. 6 and 7, the $P(topic|\, doc)$ and $P(word|\, topic)$ is attained from accumulation process of multiplication between sum of term-frequency and $P(topic|\, doc, word)$ as stated in Eqs. (16) and (17), respectively.

Figure 8 depicts the TFM in a flow chart form to explain the model for a better understanding. This flow chart represents step by step process in finding out the topic in corpus of discussion forum.

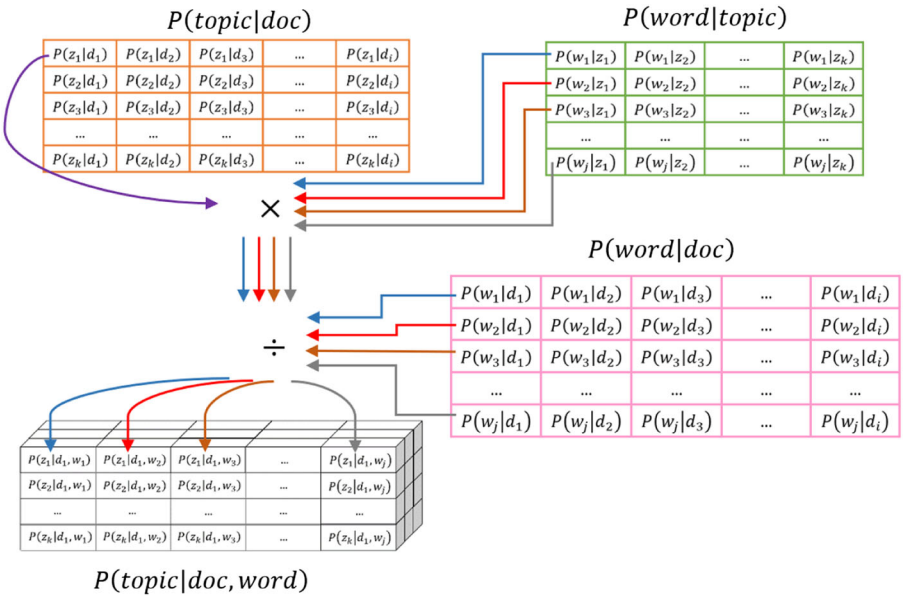


Fig. 5 The calculation of $P(topic|\, doc, word)$ for word w_1 to w_j in document d_1 of topic z_1

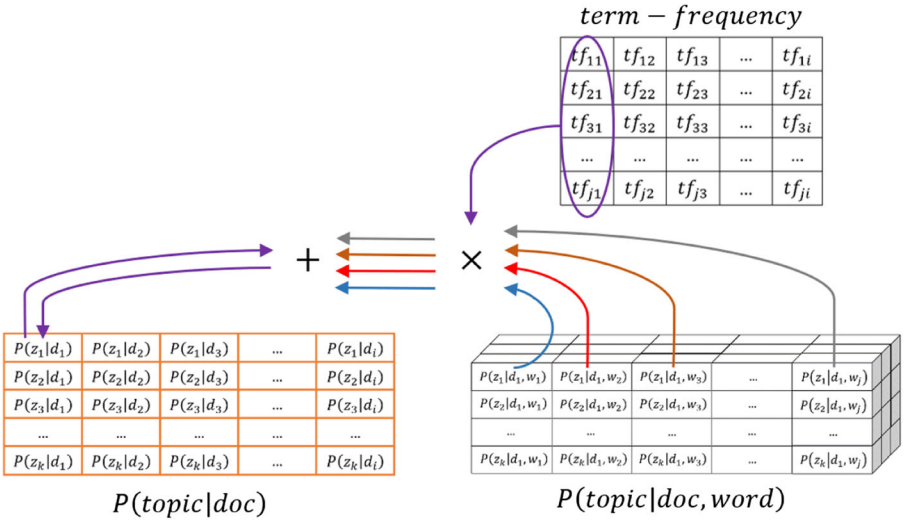


Fig. 6 The calculation of updated *P*(*topic*| *doc*) for topic z_1 of document d_1 and whole words

Based on the TFM process shown in Fig. 3, the flow chart in Fig. 8 describes flow of process from Pre-Processing document to Corpus Classification and Finding Topic. Steps of TFM is started by store discussion forum posts as text document as shown in first parallelogram in Fig. 8. Every post is allocated as a text document. The text documents are processed tokenization, stop-word removal and stemming. These processes are part of pre-processing document and impact to term-frequency of distinct word in every text document. The stop-word list and the stemming process are adjusted based on the language used in the discussion forum. The stemming process is required to produce a root word and it impacts to term-frequency of distinct word in a document. The term-frequency of distinct word is

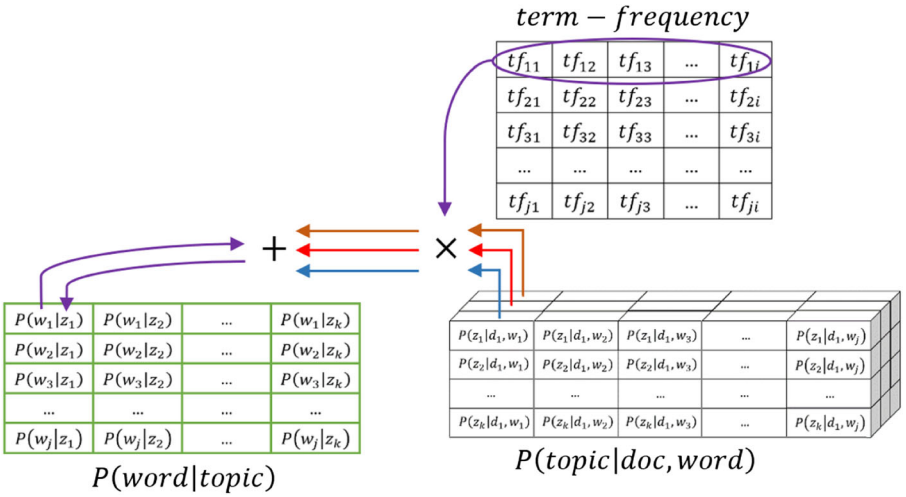


Fig. 7 The calculation of updated *P*(*word*| *topic*) for word w_1 of topic z_1 in whole documents

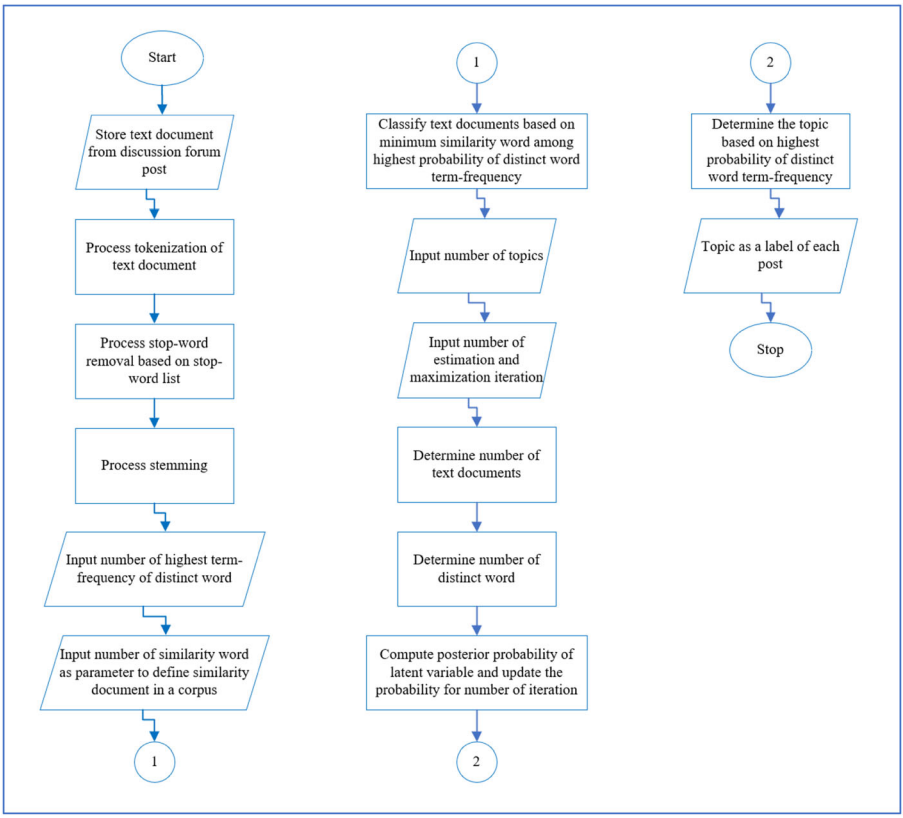


Fig. 8 The flow chart of TFM

needed in the second step, namely corpus classification. Corpus classification input consists of number of highest term-frequency of distinct word and number of similarity words as shown in first column of Fig. 8. The corpus classification groups documents based on similarity words. This output is then used in third process, i.e. finding topic. In second column of Fig. 8, beside the corpus that is created by corpus classification, there are two inputs of finding topic process; number of topics and number of iterations. The result of finding topic process is topic in corpus and it is used as a label of the post that is shown in the last parallelogram of Fig. 8. Generally, a discussion is opened through a thread and followed by replies or responses from among members. Ideally, a thread discusses a specific topic, however there is no a guarantee that it will be followed by a response. The discussion might be opened out to another topic in a thread. Therefore, a thread as a corpus is classified into the corpus classification step. This classification intent to group thread posts as a corpus into a specific corpus. The grouping is based on the number of similarities of distinct word with highest term-frequency. This similarity is determined among the number of highest term-frequency of distinct word.

Furthermore, the topic is found through a latent semantic approach. The topic is defined by the highest probability value of each document. The number of topics and

the number of iterations in estimation and maximization steps are two parameters needed in this process. The number of distinct words and the number of documents in a corpus is determined to compute the posterior probability of latent variable. The approach consists of eight steps (Setiawan et al. 2019). The process in detail explained in algorithm 1, namely Latent_Semantic Algorithm.

Algorithm 1 latent_semantic_algorithm

```
input k as number of topics
input x as number of iteration
determine i as number of documents
determine j as number of words
prepare matrix of term_frequency
foreach counter = 1 to length(i) do
    foreach counter = 1 to length(j) do
        store term_frequency of distinct word to matrix of term_frequency
    end
end
prepare matrix of prob_word_of_topic with random numbers
foreach counter = 1 to length(k) do
    normalize the values of prob_word_of_topic
end
prepare matrix of prob_topic_of_doc with random numbers
foreach counter = 1 to length(i) do
    normalize the values of prob_topic_of_doc
end
prepare matrix of prob_word_of_doc with zeros
foreach counter = 1 to length(i) do
    foreach counter = 1 to length(k) do
        store accumulation of multiplication between elements of
        prob_word_of_topic and prob_topic_of_doc to prob_word_of_doc
    end
end
foreach counter = 1 to length(x) do
    compute prob_topic_of_word_doc as estimation step
    update value of prob_topic_of_doc as maximization step
    update value of prob_word_of_topic as maximization step
    update value of prob_word_of_doc as maximization step
end
sort elements of prob_word_of_topic by descending mode
sort elements of prob_topic_of_doc by descending mode
determine the topic of each document based on highest probability of
prob_topic_of_doc element
```

Table 1 A profile data

Area of course subject	Course subject Id	Number of text documents	Number of threads discussion
Information System	1st-course subject	330 posts	5 threads
Management	2nd-course subject	370 posts	10 threads
Character Building	3rd-course subject	350 posts	13 threads

Table 2 The corpus profiles of 1st course subject

Thread	Number of text documents	Number of corpus
1st	82 posts	25
2nd	79 posts	13
3rd	73 posts	27
4th	47 posts	10
5th	49 posts	18

4 Evaluation and result

This study used 1050 text documents from a Learning Management System (LMS) of Bina Nusantara University to evaluate the model. The data are gathered from three different course subjects: information system, management and character building as the first, second and third-course subject, respectively. Online discussion characteristics can be grouped into: (1) highly confined discussion because the course is governed by math formula and physical law; (2) less confined discussion because math formula dan physical law are less exposed; (3) unrestricted discussion because of expressing personal experience and character. To accommodate all concerns cited above, those 3 courses were selected, i.e. Information Systems, Management, and Character Building represents group 1, group 2, and group 3, respectively. In period of gathering data, the number of taught courses were 62 courses which were grouped according to those characteristics. Another reason for choosing several subjects in three different areas: computing, social and behavioral area, is to observe consistency of the model. The number of documents of per course subject is 330 text documents, 370 text documents,

Table 3 The corpus profiles of 2nd course subject

Thread	Number of text documents	Number of corpus
1st	11 posts	4
2nd	32 posts	15
3rd	17 posts	7
4th	19 posts	8
5th	49 posts	12
6th	21 posts	6
7th	12 posts	7
8th	26 posts	11
9th	41 posts	12
10th	39 posts	18
11th	41 posts	12
12th	28 posts	9
13th	34 posts	15

Table 4 The corpus profiles of 3rd course subject

Thread	Number of text documents	Number of corpus
1st	13 posts	8
2nd	41 posts	14
3rd	28 posts	11
4th	37 posts	6
5th	43 posts	24
6th	36 posts	15
7th	25 posts	9
8th	23 posts	9
9th	35 posts	23
10th	69 posts	30

and 350 text documents. A text document represents one post in a thread discussion. Table 1 shows the profile of the data.

First, the data were processed with the pre-processing process: tokenization, stop-word removal and stemming. An example of tokenization statement in English is 'Knowledge can be obtained from learning and experience'. The tokenization consists of 8 tokens: 'Knowledge', 'can', 'be', 'obtained', 'from', 'learning', 'and' and 'experience'. Another example in the Indonesian language is '*Pengetahuan bisa didapat dari pembelajaran dan pengalaman*'. The result consists of 7 tokens: '*Pengetahuan*' ('Knowledge'), '*bisa*' ('can be'), '*didapat*' ('obtained'), '*dari*' ('from'), '*pembelajaran*' ('learning'), '*dan*' ('and') and '*pengalaman*' ('experience').

Moreover, the process is a stopped-word removal. Since mostly discussion is in the Indonesian language, the stop-word removal list used in the list is from Tala and completed with some common English words (Tala 2003). Using the previous example of an Indonesian statement, '*Pengetahuan bisa didapat dari pembelajaran dan*

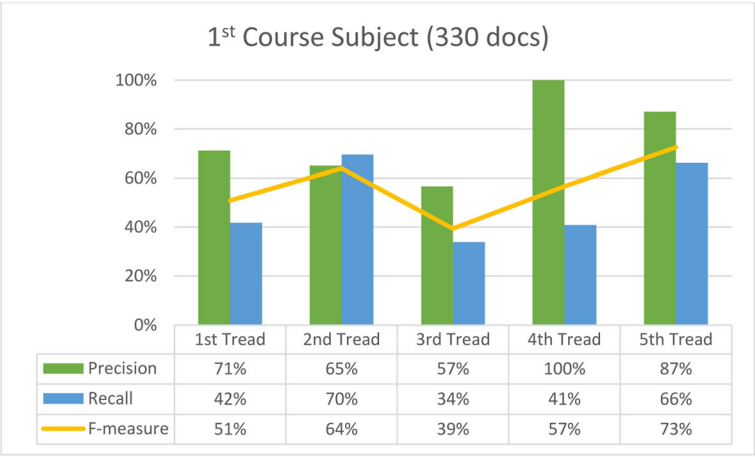
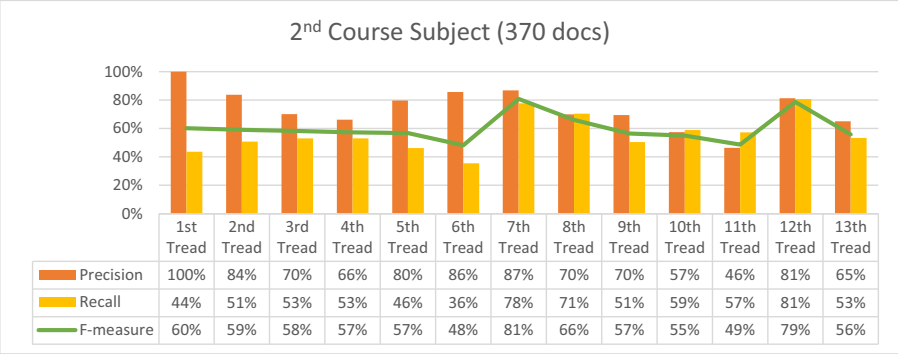


Fig. 9 The performance model of 1st course subject based on F-measure



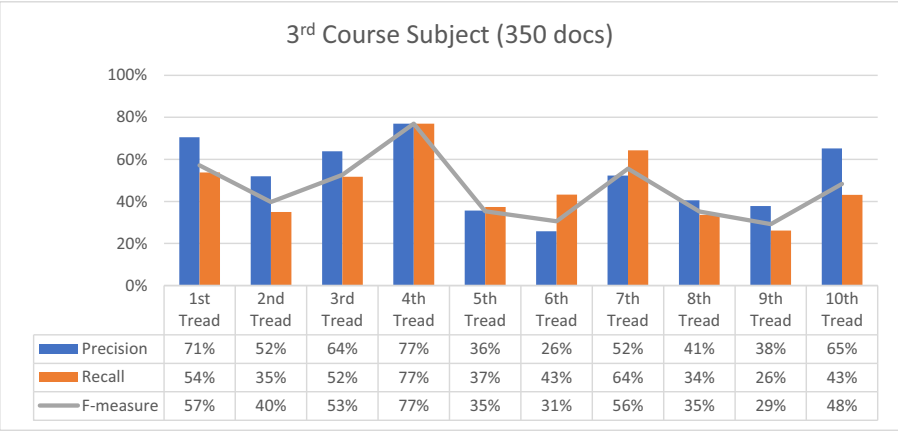
	1st Tread	2nd Tread	3rd Tread	4th Tread	5th Tread	6th Tread	7th Tread	8th Tread	9th Tread	10th Tread	11th Tread	12th Tread	13th Tread
Precision	100%	84%	70%	66%	80%	86%	87%	70%	70%	57%	46%	81%	65%
Recall	44%	51%	53%	53%	46%	36%	78%	71%	51%	59%	57%	81%	53%
F-measure	60%	59%	58%	57%	57%	48%	81%	66%	57%	55%	49%	79%	56%

Fig. 10 The performance model of 2nd course subject based on F-measure

pengalaman', the removed tokens are '*bisa*', '*didapat*', '*dari*' and '*dan*', therefore the remain tokens are '*Pengetahuan*', '*pembelajaran*' and '*pengalaman*'.

In this study, the stemming process is used for the flexible affix classification approach (Setiawan et al. 2016). This algorithm is used, since most of the discussion is in the Indonesian language and the algorithm is good to obtain high accuracy.

Second, the documents per thread were classified by a corpus classification approach (Setiawan et al. 2019). This process was needed to classify documents to be more specific corpus rather than a corpus based on a thread. Thus, the parameter of number of the word with highest term-frequency and number of similar words are 5 and 2, respectively. Tables 2, 3 and 4 show several the corpus from the result of the 1st, 2nd and 3rd-course subject classification. The examples in Table 2, the 1st thread consist of 82 posts classified into 25 corpora based on 2 similar words or more within 5 highest term-frequency words. This means that a thread of discussion ideally assumed as one corpus can be divided to into several corpora based on the certain similar words. It also happened in others thread of discussions.



	1st Tread	2nd Tread	3rd Tread	4th Tread	5th Tread	6th Tread	7th Tread	8th Tread	9th Tread	10th Tread
Precision	71%	52%	64%	77%	36%	26%	52%	41%	38%	65%
Recall	54%	35%	52%	77%	37%	43%	64%	34%	26%	43%
F-measure	57%	40%	53%	77%	35%	31%	56%	35%	29%	48%

Fig. 11 The performance model of 3rd course subject based on F-measure

Third, this is the last step in the model to find out a topic of discussion in a corpus using latent semantic approach. The documents per corpus were processed by PLSA eight steps as mentioned in Section 3. Figures 9, 10 and 11 show the performance of the model to find out the topic of discussions. The measurement used F-measure. Every post was read and defined the topic manually as a label in the text document. The precision and recall were measured based on the results of model's topic compared to a label per document. The F-measure was measured based on the precision average and recall average per thread.

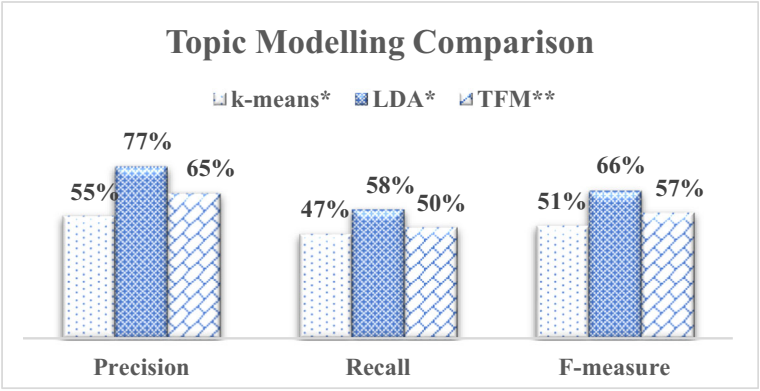
Figure 9 shows that the precision value is good. This reveals that the result of the topic in the model is correct; however, the trend of recall value is lower than the precision. This condition informs that there is a topic not found in the model. The chosen topic from the model is gathered only from the highest probability value of the latent variable. On that basis, the recall value is not good enough. An example is the following post below:

> dear pak menanggapi topik no 9 dan no 10 saya berpendapat : 9 porter's value chain model adalah model yang digunakan untuk membantu menganalisa aktifitas-aktifitas spesifik bisnis yang terjadi yang dapat menciptakan nilai dan keuntungan kompetitif bagi organisasi model ini terbagi menjadi 2(dua) aktifitas: a) aktifitas utama -inbound logistic (input) -operation (manufacturing and testing) -outbond logistic(storage and distribution) -marketing and sales -customer service b) aktifitas pendukung -organisasi infrastruktur (akuntansi keuangan manajemen) -manajemen sumber daya manusia -pengembangan produk dan teknologi (r&d) -pengadaan (procurement) 10 strategi untuk keunggulan kompetitif dari michael porter terdiri dari 5 strategi: a) strategi cost leadership : jika suatu organisasi memilik strategi ini maka mereka akan menghasilkan produk/jasa dengan biaya yang serendah mungkin b) strategi differentiation : pada strategi ini organisasi menawarkan bermacam-macam produk/jasa atau fitur produk melebihi dari yang ditawarkan oleh pesaingnya c) strategi innovation : dalam strategi ini organisasi mengedepankan penemuan produk/jasa baru serta fiturnya agar dapat selalu mengungguli pesaingnya d) strategi operational effectivenes : meningkat proses bisnis internal sehingga perusahaan dapat menjalankan aktifitas dengan lebih baik dibandingkan dengan pesaingnya e) strategi customer-orientation : strategi ini memfokuskan pada pelanggan dan bagaimana caranya untuk membuat pelanggan tersebut senang referensi : 1 http://kc99lounge blogspot com/2010/07/porters-value-chain html 2 lecture note terimakasih

The post is from Information System Concept course subject. Since mostly topics are in Indonesian language, for the sake of reader's convenience who are not Indonesian, the words in the brackets in the Topic column are written in English. The topics finding from the model consists of '*strategi*' ('strategy'), '*organisasi*' ('organization') and '*usaha*' ('business'). The topics from manually label contain of '*strategi*'

Table 5 The topics from the TFM order by probability value (English translated terminologies are added for the sake of reader's convenience)

Level	Topics
1st	'*strategi*' ('strategy'), '*organisasi*' ('organization'), '*usaha*' ('business')
2nd	'*produk*' ('product'), '*untung*' ('profit'), '*organisasi*' ('organization'), '*usaha*' ('business')
3rd	'*model*' ('model'), '*bisnis*' ('business'), '*produk*' ('product'), '*untung*' ('profit'), '*strategi*' ('strategy')
4th	'*strategi*' ('strategy'), '*model*' ('model'), '*usaha*' ('business'), '*produk*' ('product')
5th	'porter' ('porter'), '*organisasi*' ('organization'), '*bahan*' ('material'), '*aktifitas*' ('activity')



*(Rajasundari et al., 2017) and **result of this research.

Fig. 12 Performance of topic modelling comparison

('strategy'), '*usaha*' ('business') and 'porter' ('porter'). The topic 'organization' arise from the model, however not as topic from manually. It impacts to precision value. On the other hand, the topic 'porter' does not arise as the highest probability value from the model, though it is the topic for the post. It effects to recall value. Table 5 represent an example of topics in a corpus from the model based on highest probability value in descending mode. The 1st level is the highest probability value.

In consonance with Table 5, to improve the recall value, then it is necessary to explore topics that found out from the model in several next levels of the highest probability value.

Figures 10 and 11 provides similar results trend with Fig. 9, which include the precision, the recall and the F-measure. The overall results explain that TFM is consistent and effective to find out the topic of discussion, find out the result of the topic and find out whether the precision is good, however the recall can still be increased and observed in the future study.

Despite of not using the same dataset, an attempt was made to compare the research result of TFM approach to LDA and k-means. The result of k-means and LDA are obtained from the previous research (Rajasundari et al. 2017). The comparison among k-means, LDA, and TFM using precision, recall, and F-measure as parameters is shown in Fig. 12.

The TFM confirms that result of topic modelling algorithm gives a better performance than machine learning approach does. In this case, the topic modelling algorithm are LDA and TFM, while the machine learning approach is k-means. The Precision, Recall, and F-measure of TFM is 65%, 50%, and 57% which is greater than Precision, Recall, and F-measure of k-means, 55%, 47%, and 51%, respectively.

5 Conclusion

Through discussion forum, members enhance their knowledge about new things. Unfortunately, the knowledge or information only known among the members. This

paper has presented a model to extract the topic of discussion forum posts, namely Topics Finding Model (TFM). The topic from the model was used to label the post, then it is possible to retrieve knowledge from the discussion forum. The TFM consists of three steps: pre-processing text document, corpus classification and finding topic through latent semantic. To measure the effectiveness of the model, F-measure was used in this study. The result shows that TFM is consistent and effective in revealing the topic of discussion. The limitation of this study is the determined topic based on the highest probability value of the latent variable. Despite the precision value is good, the recall value can still be increased. This is an opportunity to explore another level of probability value of a topic to raise the recall value in the further study.

Since this approach has not been covered the slang and typographical error of posts, the normalization process may be added as part of pre-processing step. Thus, impact of the normalization process to the result can be observed. This process can be explored for further work.

References

Baeza-Yates, R., & Ribeiro-Neto, B. (2011). *Modern information retrieval: the concept and technology behind search (second edition)*. Addison-Wesley Professional Harlow.

Benedetti, F., Beneventano, D., Bergamaschi, S., & Simonini, G. (2018). Computing inter-document similarity with context semantic analysis. *Information Systems, 80*, 136–147. https://doi.org/10.1016/j.is.2018.02.009.

Berger, A., & Lafferty, J. (1999). Information retrieval as statistical translation. In *Proceedings of the 22nd annual international ACM SIGIR conference on research and development in information retrieval* (pp. 222–229). https://doi.org/10.1145/312624.312681.

Botev, V., Marinov, K., & Schäfer, F. (2017). Word importance-based similarity of documents metric (WISDM): Fast and scalable document similarity metric for analysis of scientific documents. In *Proceedings of the 6th international workshop on mining scientific publications* (Vol. 7, pp. 17-23). https://doi.org/10.1145/3127526.3127530.

Cutting, D. R., Karger, D. R., Pedersen, J. O., & Tukey, J. W. (2017). Scatter/gather: A cluster-based approach to browsing large document collections. In *ACM SIGIR Forum* (Vol. 51, pp. 148–159).

Dan Oneata. (1999). Probabilistic latent semantic analysis. In *Proceedings of the fifteenth conference on uncertainty* (pp. 1–7).

Ezen-can, A., Boyer, K. E., Kellogg, S., & Booth, S. (2015). Unsupervised modeling for understanding MOOC discussion forums: A learning analytics approach. In *Proceedings of the fifth international conference on learning analytics and knowledge* (pp. 146–150).

Hofmann, T. (1999). Probabilistic latent semantic analysis. In *Proceedings of the fifteenth conference on uncertainty in artificial intelligence* (pp. 289–296).

Hofmann, T. (2001). Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning, 42*(1–2), 177–196. https://doi.org/10.1023/A:1007617005950.

Hong, C., Chen, W., Zheng, W., Shan, J., Chen, Y., & Zhang, Y. (2008). Parallelization and characterization of probabilistic latent semantic analysis. In *Parallel processing, 2008. ICPP'08. 37th international conference on* (pp. 628–635). https://doi.org/10.1109/ICPP.2008.8.

Kuran, M. S., Pedersen, J. M., & Elsner, R. (2017). Learning management systems on blended learning courses: An experience-based observation. In *International conference on image processing and communications* (pp. 141–148). https://doi.org/10.1007/978-3-319-68720-9.

Piña, A. A. (2018). *An educational Leader's view of learning management systems. Leading and managing e-learning* (pp. 101–113). Springer.

Rajasundari, T., Subathra, P., & Kumar, P. (2017). *Performance analysis of topic modeling algorithms for news articles*. Journal of Advanced Research in Dynamical and Control Systems (11).

Ruano, I., Gamez, J., Dormido, S., & Gomez, J. (2016). A methodology to obtain learning effective laboratories with learning management system integration. *IEEE Transactions on Learning Technologies, 9*(4), 391–399. https://doi.org/10.1109/TLT.2016.2594771.

Sailaja, N. V., Sree, L. P., & Mangathayaru, N. (2018). New rough set-aided mechanism for text categorisation. *Journal of Information & Knowledge Management, 17*(2), 1850022. https://doi.org/10.1142/S0219649218500223.

Sanderson, M., & Croft, W. B. (2012). The history of information retrieval research. *Proceedings of the IEEE, 100*(Special Centennial Issue), 1444–1451. https://doi.org/10.1109/JPROC.2012.2189916.

Setiawan, R., Kurniawan, A., Budiharto, W., Kartowisastro, I. H., & Prabowo, H. (2016). Flexible affix classification for stemming Indonesian Language. In *2016 13th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology, ECTI-CON 2016* (pp. 1–6). https://doi.org/10.1109/ECTICon.2016.7561257.

Setiawan, R., Budiharto, W., Kartowisastro, I. H., & Prabowo, H. (2019). *Enhancing focus topic findings of discussion forum through corpus classifier algorithm*. Retrieved from http://socs.binus.ac.id/2019/01/15/enhancing-focus-topic-findings-through-corpus-classifier-algorithm/.

Tala, F. Z. (2003). A study of stemming effects on information retrieval in Bahasa Indonesia. *M.Sc. Thesis, Appendix D*, *pp*, 39–46.

Utomo, B. Y., & Bijaksana, M. A. (2016). Comprehensive comparison of term weighting method for classification in Indonesian corpus. 2016 4th *International conference on information and communication technology*, ICoICT 2016, 4(c), 1–5. https://doi.org/10.1109/ICoICT.2016.7571886.

Wang, X., Yang, D., Wen, M., Koedinger, K., & Rosé, C. P. (2015). *Investigating how Student's cognitive behavior in MOOC discussion forums affect learning gains*. International Educational Data Mining Society, 226–233.

Wen, M., Yang, D., & Rosé, C. P. (2014). Sentiment analysis in MOOC discussion forums: What does it tell us? In *Educational data mining 2014*.

Yang, D., Wen, M., Howley, I., Kraut, R., & Ros, C. (2015). Exploring the effect of confusion in discussion forums of massive open online courses. In *Proceedings of the second (2015) ACM conference on learning@ scale* (pp. 121–130).

Zhai, C., & Lafferty, J. (2017). A study of smoothing methods for language models applied to ad hoc information retrieval. *ACM SIGIR Forum, 51*(2), 268–276.