Detection of Reference Topics and Suggestions using Latent Dirichlet Allocation (LDA)

Setio Basuki
Faculty of Engineering
Informatics Department
Universitas Muhammadiyah Malang
Indonesia, Malang
Email: setio_basuki@umm.ac.id

Yufis Azhar
Faculty of Engineering
Informatics Department
Universitas Muhammadiyah Malang
Indonesia, Malang
Email: yufis@umm.ac.id

Agus Eko Minarno
Faculty of Engineering
Informatics Department
Universitas Muhammadiyah Malang
Indonesia, Malang
aguseko@umm.ac.id

Christian Sri Kusuma Aditya
Faculty of Engineering
Informatics Department
Universitas Muhammadiyah Malang
Indonesia, Malang
christianskaditya@umm.ac.id

Fauzi Dwi Setiawan Sumadi
Faculty of Engineering
Informatics Department
Universitas Muhammadiyah Malang
Indonesia, Malang
fauzisumadi@umm.ac.id

Ardiansah Ilham Ramadhan
Faculty of Engineering
Informatics Department
Universitas Muhammadiyah Malang
Indonesia, Malang
Email: ardiansahir@webmail.umm.ac.id

***Abstract*— *Pelatihan Aplikasi Teknologi Informasi* (PATI) is an activity of training required for new students in Universitas Muhammadiyah Malang (UMM) to provide knowledge and training on UMM or information technology concerned about general technology. At the end of the training, the students give the conclusions and suggestions to PATI. During this event, the training Committee gave less concern in term of the inference from students to provide a material evaluation. The primary factor originated from the commenting processes which should be performed one by one. Therefore, the comprehensive method should be implemented by modelling using Latent Dirichlet Allocation (LDA) in order to facilitate the Committee to undertake an analysis of the conclusions and suggestions. LDA is a "generative probabilistic model" of a collection of composites made up of parts. In terms of topic modeling, the composites are documents and the parts are words and/or phrases (n-grams). Conclusions and suggestions are taken as many as 1025 data from PATI 2016/2017. Based on such research, modelling of LDA identifies the 7 topics in the overall data. The process of analysis is done by external details each comment contains what topics. The evaluation is done by testing 250 data to determine the results of the conformity between the results of the analysis of the system as well as actual results obtained from respondents. The test results obtained accuracy of 83.6%.**

Keywords*— *Inference, Latent Dirichlet Allocation, PATI, Topic Modelling, UMM

I. INTRODUCTION

PATI is a training activity that must be followed by new students at the UMM [1]. Provided training and knowledge about technology and information owned by UMM or in general is an idea promoted by PATI. In the training accompanied by instructors regarding supporting materials for internal or external purposes. This activity was carried out in 8 laboratories owned by the campus. This activity guides students to practice immediately when attending the training for a week. At the end of the training, the students gave comments about the training that had been obtained. The comments are in the form of conclusions and suggestions, in which the data taken are the conclusions and suggestions of the students.

During this time, the training committee paid little attention to the conclusions and suggestions of students to be used as evaluation material because it was less effective to conclude by reading one by one student comments that were too many. While the comments can be searched for the main topics being discussed about something that we want to analyze. That way it can be used to conclude information that is hidden inside which can be used as evaluation material to determine strategies that must be taken in the future.

Therefore, a method is needed to provide a solution for topic modelling. Drawing from the name, topic modelling includes modelling textual data that aims to find hidden variables, namely a topic [2]. One model of topic modelling is the LDA method (Latent Dirichlet Allocation). The LDA method is a model that can be applied to topic modelling in a very large textual data collection. This model makes it easy to detect topics inside. Based on the available topics, the topic will be processed using the LDA method to produce topic modelling of student conclusions and suggestions. The data will be detected and produce a core topic from comments about the conclusions and suggestions.

Starting with the research [3] who used questionnaire data to evaluate the propensity of suggestions relating to various factors that contribute to the success of learning by using suggestions and comments as opinions. Researchers conducted opinion analysis and topic search with classification using the Naive Bayes Classifier (NBC). Based on research [4], researchers conducted a topic modelling at service centres owned by PT. Petrochemical Gresik. After the researcher gets the topic through the topic modelling process, then the results are adjusted to the company because the company has a topic category that has been provided. Next, the researcher analyzes the

978-1-7281-2133-8/19/$31.00 ©2019 IEEE

customer's voice data with the process of directing the data to the topic to find out what topics are contained. The results of the analysis are visualized in the form of a dashboard containing graphs. Not much different from previous researchers [5], also carried out the analysis of online user reviews of the amazon.com site. Researchers performed topic extraction to get what topics are in the customer review. Topic detection also can be performed using k-means that is a well-known and widely used partitional clustering method [6]. In this case, the researcher labelled the topic with subjective justification based on the terms that appeared on the results of the topic modelling, and because it uses LDA then document can have the possibility to enter into several topics.

This research focuses for building detection topics about conclusions and student suggestions on PATI using LDA. Thus, it will facilitate the training committee to find out the topics contained in student comments.

II. DATASET

The data used in this study are the data of conclusions and recommendations of the PATI 2016/2017 academic year at UMM. The entire data is in the form of HyperText Markup Language (HTML) files with a total of 4,485 data.

When preparing data, the conclusions and suggestions that are still raw are parsed. The data is then filtered in order to eliminate comments that do not contain meaning, comments that are too short, comments that do not use Indonesian, and comments that are not common in Indonesian, such as slang words, so that it can be implemented in research through several preprocessing stages including Case Folding, Tokenizing, Stopword Removal to get maximum results. The data is again checked to find the same data for each class during PATI implementation and after cleaning the same data, the data obtained is 1,025 data, and for the testing phase, it uses 250 test data.

III. METHOD

The system begins when the user inputs conclusions and suggestions, then preprocessing the data so that the features are processed selectively, and the data is in accordance with the needs of the main process. This preprocessing process goes through many stages, which among others eliminates HTML tags, case folding, stopword removal and tokenizing. Preprocessing result data is processed to do topic modelling using LDA. In this process, the preprocessing results are generated into the desired topic by determining how many topics we want to generate. In the topic is ordered to display words that have the probability of the topic. The system is shown in Fig. 1. In this case, the author uses the NlpTools Library which provides needs in Natural Language Processing, among others, text classifier, models, clustering and other types of PHP-based.

Furthermore, the process of similarity includes the calculation process of TF-IDF and Cosine Similarity. Topic results from the LDA modelling are used as queries where calculated the similarity of commentary data on the topic that has been obtained. In this process, the TF-IDF is weighted against the comment and query data in order to get the weight of each term on the entire document. The author uses the PHP-ML library to help calculate TF-IDF. Cosine Similarity plays a role in knowing each comment on any topic by approaching between queries and documents. The topic generated by the LDA becomes a query and is directed to the document containing the comment. After the process, the results of detection of the topic will be obtained.

A. Data Preprocessing

Most ways of topic modelling processing involve steps for data preprocessing and data cleaning. This will depend on the characteristics of the data to be analyzed. The first thing that needs to be done is importing data to retrieve content that is in the files. Then proceed with cleaning HTML tags that are still attached to the contents of the file. Then case folding is done to make the text in the document become a standard form in this case lower-case. The tokenizing stage is the stage of cutting the string based on each word that composes it. In addition, spaces are used to separate the words. Then it is needed to eliminate assumptions that lack meaning (common words). Stopword removal is the process of removing words that do not contribute much to the contents of the document. Words that include stopword are omitted because they have an unfavourable effect on searching for documents that the user wants.

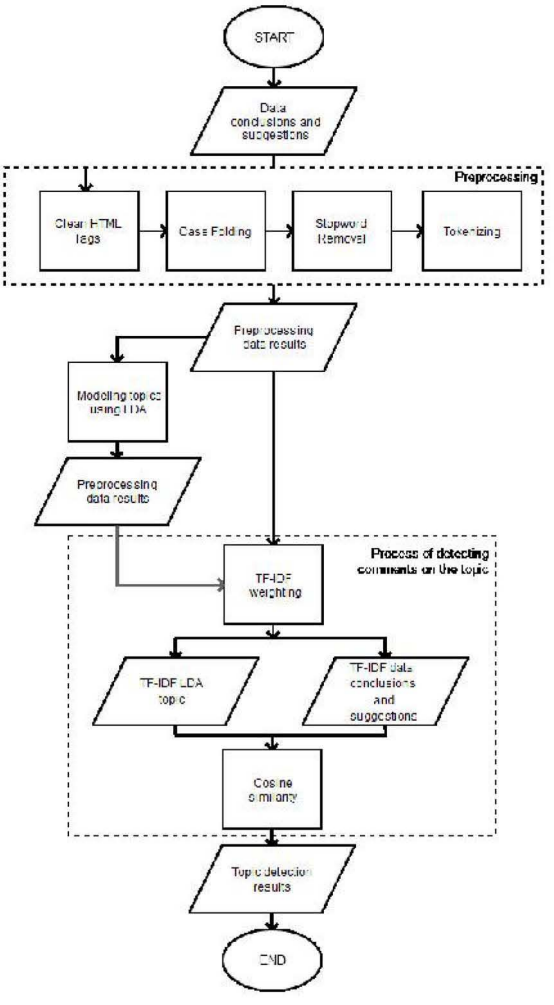

Fig. 1. Detection of reference topics and suggestions system

B. Topics Modeling Using LDA

Topic modelling will maximize the optimal results regarding the number of topics specified in 7 topics. On the other hand, INFOKOM DPP has a topic category that has been used as a reference, including computers, the internet, e-learning web, teacher appearance, material clarity, timeliness, and teaching interaction. Based on this foundation, topic modelling on PATI student comments includes conclusions and suggestions.

The topic modelling using the LDA algorithm aims to obtain any topic contained in the comment. The basic concept is that documents can represent as a mixed model that has various topics, where the topics are represented by the word. The basic intuition of LDA is a document containing various topics by defining the topic as a distribution on a fixed vocabulary. LDA represents documents with various topics that are made based on certain probabilities. The probability of the topic represents the clarity of a document. LDA is a generative probabilistic model from a set of the corpus which has the following process:

1. For each document w in the corpus D
   a. Choose N ~ Poisson(ξ)
   b. Choose θ ~ Dir(α)
2. For each word N in document wn
   a. Choose topic zn ~ Multinominal(θ)
   b. Choose word wn from p(wn | zn, β)

A Dirichlet k-dimensional random variable can take values in (k-1)-simplex (θ k-vector lies on (k-1)-simplex if θ i ⩾ 0, Σki =1, θ = 1) and has a probability formula like the following:

$$p(\theta \mid \alpha) = \frac{\Gamma(\sum_{i=1}^{k} \alpha_i)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \prod_{i=1}^{k} \theta_i^{\alpha_i - 1}$$

θ: topic distribution in the document

α: parameters for calculating how the topic is distributed in the document

k: number of topics

For parameters α dan β, merging the distribution of topics from the mixture θ, z, w, N has the following probability formula :

$$p(\theta, z, w|\alpha, \beta) = p(\theta|\alpha) \prod_{n=1}^{N} p(zn|\theta) p(wn|zn, \beta)$$

The modelling implementation in LDA uses PHP library, PHP-NLP-tools where the main step in modelling the topic using this LDA will be explained in Fig. 2.

From the flowchart in Fig. 2 is the implementation of the LDA Algorithm, the input is the document to be modelled, the number of topics wants to issue, and the number of terms wants to display for each topic. The next process is made sampling in order to obtain the sample of words identified in the document. Then each iteration and each number of words identified is accommodated in the sequence of words according to the iteration. Furthermore, topic modelling is based on full condition samples of words and documents. Sample full condition is a process for correcting a random distribution of values.

Re-sampling was carried out but directly distributed to the specified topic. The next process is an assignment of topics per word where the results of the sampling are probable and then entered into the number of topics that have been determined. Each topic will contain words with a probability of the topic.

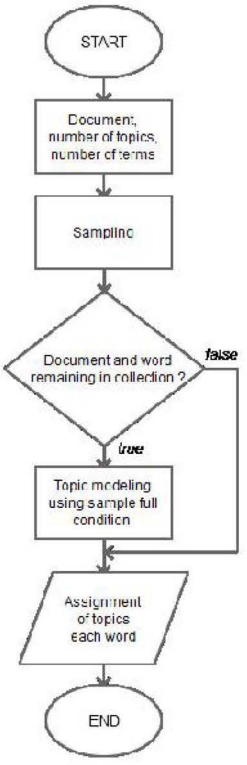

Fig. 2. LDA-based topic modelling process

C. TF-IDF Weighting

The TF-IDF weighting process begins with document input and query input. Queries are topic words obtained from the topic modelling results using LDA. The documents and queries are calculated using Term Frequency (TF) to get the number of terms that appear. Then do Inverse Document Frequency (IDF) to show the relationship of availability of a term in all documents and queries. Furthermore, the TF-IDF is weighted against the documents and queries in a process to determine how far the word (term) relationship is with the class. On TF-IDF there is a formula for calculating weighting as follows:

$$w_{ij} = \iota f \; x \; idf$$

$$w_{ij} = tf_{ij} \; x \; log \frac{N}{n}$$

W_{ij} = word weight t_j against documents d_i.

tf_{ij} = number of occurrences t_j in d_i.

N = number of all documents.

n = number of documents containing words t_j (there is at least one word, term t_j)

The results obtained from the TF-IDF query and the results of TF-IDF documents will later be continued in the Cosine Similarity process in order to find the closeness between comments on the topic. The TF-IDF weighting process uses the PHP library, PHP-ML.

D. Cosine Similarity

In this research, cosine similarity is used to find the closeness between documents (comments) on the topic to find out each document has proximity to any topic. The topic here is a query where the query contains topics from the analysis of topic modelling using LDA. Then the query is directed toward available documents. Each document will have value for each topic and then the value of each topic is sorted from highest to lowest to find out what topics have the highest value. The highest value topic shows the tendency of documents on the topic. To get this value there is a cosine similarity formula as follows:

$$Similarity = \cos(\theta) = \frac{A \cdot B}{|A||B|} = \frac{\sum_{i=1}^{n} Ai \, x \, Bi}{\sqrt{\sum_{i=1}^{n}(A_i)^2 \, x} \sqrt{\sum_{i=1}^{n}(B_i)^2}}$$

A = weight value *tf* x *IDF* from query (keyword)

B = weight value *tf* x *IDF* from document

$\sum A$ = sum of values *tf x IDF* from the query (keyword)

$\sum B$ = sum of values *tf* x *IDF* from document

The result of the approach to querying the document has been completed by getting the value of each document against the query. These processes are repeated against a number of queries. After each query has been calculated, the next step is to rank each query against each document. When a document has the highest query value, it can be concluded that the document tends to lead to the query.

IV. Results and Discussion

Tests were carried out using 250 data testing data. The data tested is data that has been labelled the results of previous detection analysis. The testing method used is accuracy because to determine how accurate a model is in classifying output. This research required respondents as many as 3 people because in order to get a variation of the 3 results of each respondent. Respondents labelled topics on documents subjectively based on looking for propensity in comments on each topic. After the respondent conducts labelling, the task of the researcher is to match the results of the system with the topic of the results of the respondents for each document. When a document found the results of the topic of the system match the results of the topic of the respondent, it can be said that the topics in the document are appropriate. Conversely, if it doesn't have a match then it can say that the topic is not suitable.

In the first test, out of a total of 250 data, the corresponding data was 206 data, while the data that did not match was 44 data. So, if measured using a percentage, you will get data accuracy as follows:

$$Accuracy = \frac{appropriate\ amount\ of\ data}{total\ of\ all\ data} x\ 100\%$$

$$Accuracy = \frac{206}{250} x\ 100\% = 82{,}4\%$$

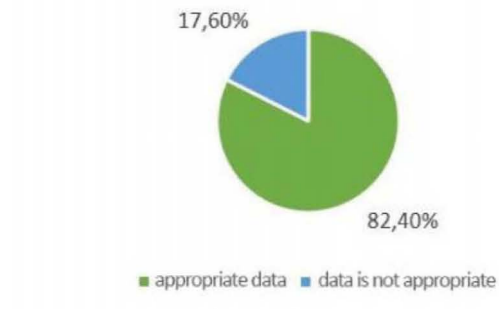

Fig. 3. First test result

The percentage in Fig. 3 is the result of the first test that gets accuracy = 206/250 or 82.4%. While the error rate is 44/250 or 17.6%

In the second test, out of a total of 250 data, the corresponding data is 212 data, while the data that is not suitable is 38 data.

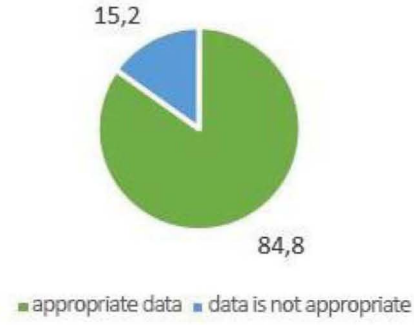

Fig. 4. Second test result

The percentage in Fig. 4 is the result of the second test that gets accuracy = 212/250 or 84.8%. While the error rate is 38/250 or 15.2%

It is possible in this study to find factors that influence the results of the above tests. Factors that most likely affect the level of accuracy are when calculating the similarity of documents (comments) to the topic. A document can have the possibility to enter into several topics. This will make the value of the similarity of a document a slight difference to the topics so that the document can be said to have meaning from several topics even though the value is not too strong. LDA looks at the topic as the number of clusters and probabilities as the proportion of cluster membership, thus LDA performs grouping softly, not like k-means where each entity can only be owned by one cluster. Another factor is the analysis parameters that have been determined in the boundary. Considering that the topic that will be issued has been determined it is likely to be a factor that causes the values to be obtained above.

V. Conclusion

Analysis of topic modelling using LDA is done with 3 parameters, namely document, number of topics and number of terms. The analysis uses 1025 data, the number of topics is 7 topics and 10 terms that want to be issued. After modelling, Topic 1 indicates the trend of meaning

regarding Training Tasks where the results are new topics of topics expected by the training. Topic 2 indicates the tendency of the meaning of the Teaching Appearance and Teacher Interaction where the results are one of the topics expected by the training. Topic 3 indicates the trend of meaning about Web E-learning where the results are one of the topics expected by the training. Topic 4 indicates the tendency of meaning regarding Material Clarity where the results are one of the topics expected by the training. Topic 5 indicates the trend of meaning regarding Training Outcomes where the results are new topics of topics expected by the training. Topic 6 indicates the tendency of meaning regarding Computer Facilities and the Internet / Network where the results are one of the topics expected by the training. Topic 7 indicates the tendency of meaning regarding Timeliness where the results are one of the topics expected by the training. Ultimately, the final step calculates the similarity between the comments on the topic where the results of conformity obtained the amount of data on the topic includes the topic 1 is 124 data, on topic 2 is 163 data, in topic 3 is 166, on topic 4 is 152 data, on topic 5 is 224 data, on topic 6 is 118 data, on topic 7 is 78 data. In order to measure the success of this study, the accuracy testing was carried out which resulted in an average value of 83.6%.

ACKNOWLEDGMENT

This work is partially supported by Laboratorium Informatika Universitas Muhammadiyah Malang. Authors wish to thank Universitas Muhammadiyah Malang for providing the funding.

REFERENCES

[1] P. P. UMM, "Pelatihan Aplikasi Teknologi Informasi (PATI) Universitas Muhammadiyah Malang," 2013.

[2] R. I. Kengken, "Pemodelan Topik Untuk Media Sosial Menggunakan Latent Dirichlet Allocation," Skripsi, pp. 1–9, 2014.

[3] A. Hamzah, "Sentiment Analysis untuk Memanfaatkan Saran Kuesioner Dalam Evaluasi Pembelajaran Dengan Menggunakan Naive Bayes Classifier (NBC)," Pros. Semin. Nas. Apl. Sains Teknol., no. November, pp. 211–216, 2014.

[4] A. Agustina, "Analisis dan Visualisasi Suara Pelanggan Pada Pusat Layanan Pelanggan Dengan Pemodelan Topik Menggunakan Latent Dirichlet Allocation (LDA) Studi Kasus: PT. PETROKIMIA GRESIK," Institut Teknologi Sepuluh November, 2017.

[5] N. Y. Wirawan, "Rancang Bangun Ekstraksi Topik Fitur Produk Dari Ulasan Pengguna Online Dengan Latent Dirichlet Allocation," Institut Teknologi Sepuluh November, 2017.

[6] Zhang, Dan, and Shengdong Li. "Topic detection based on K-means." 2011 International Conference on Electronics, Communications and Control (ICECC). IEEE, 2011.

[7] Zulhanif, "Pemodelan Topik Dengan Latent Dirichlet Allocation," Semin. Nas. Pendidik. Mat., pp. 1–8, 2016.

[8] D. Blei, A. Ng, and M. Jordan, "Latent Dirichlet Allocation (slide)," vol. 55, no. 4, 2012.

[9] A. Knispelis, "LDA Topic Models," Youtube. [Online]. Available: https://www.youtube.com/watch?v=3mHy4OSyRf0.

[10] E. F. Nurastuti, "Penerapan Algoritma Cosine Similarity Pada Sistem Pendektesian Kemiripan Jurnal Tugas Akhir (Studi Kasus : Stiki Malang)," SEKOLAH TINGGI INFORMATIKA DAN KOMPUTER INDONESIA MALANG, 2016.

[11] D. N. Ogie Nurdiana, Jumadi, "Perbandingan Metode Cosine Similarity Dengan Metode Jaccard Similarity Pada Aplikasi Pencarian Terjemah Al- Qur'an," JOIN, vol. I, no. 1, pp. 59–63, 2016.

[12] Ahli Hidayat, "Implementasi Metode Terms Frequency–Inverse Document Frequency (TF-IDF) dan Maximum Marginal Relevance untuk Monitoring Diskusi Online," pp. 1–13, 2016.

[13] Hikmah, Faizun Nuril, "Deteksi Topik Tentang Tokoh Publik Politik Menggunakan Latent Dirichlet Allocation (LDA)," Universitas Muhammadiyah Malang, 2017.

[14] Akbi, D. R., & Rosyadi, A. R.. Paragraph Selection Methods Using Feature-Based On Segment-Based Clustering Process Using Paragraphs For Identifying Topics On Indications Detection of Plagiarism System. *Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control*, *3*(2), 91-100, 2018.

[15] Basuki, S., Rizky, A., & Wicaksono, G. W. Case Based Reasioning (CBR) for Medical Question Answering System. *Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control*, *3*(2), 113-118, 2018.