Automatic cluster labeling using ontology and Latent Dirichlet Allocation (LDA)

Rifki Adhitama
Department of Information System
Diponegoro University
Semarang, Indonesia
adhitama@student.undip.ac.id

Rahmat Gernowo
Department of Physics
Diponegoro University
Semarang, Indonesia
gernowo@yahoo.com

Retno Kusumaningrum
Department of Informatics
Diponegoro University
Semarang, Indonesia
retno@live.undip.ac.id

***Abstract*— Latent Dirichlet Allocation (LDA) is a topic modeling method that provides the flexibility to organize, understand, search, and summarize electronic archives that have proven well implemented in text and information retrieval. The weakness of the LDA method is the inability to label the topics that have been formed. This research combines LDA with ontology scheme to overcome the weakness of labeling topic on LDA. This study uses datasets of 50 news documents taken from the online news portal. The ontology scheme used in this study is based on the dictionary of the field contained in *Kamus Besar Bahasa Indonesia* (KBBI). The experiment aims to find the best word count representation for each topic in order to produce the relevant label name for the topic. Cohen's kappa coefficient is used to measure the reliability of the label based on the agreement of two linguistic experts, while the mean relevance rate is used to measure the average of the relevant value of linguistic experts on a label with particular words representation that has more than 41% of the kappa value. The results of this study indicate the highest kappa value is in the five words representation of each topic with 100% value, while the highest mean relevance rate is in the 5 words and 30 words representation of each topic with 80% value. The average of kappa value is 61%, and the average value of mean Relevance rate is 71%.**

Keywords—text clustering; cluster labeling; latent dirichlet allocation; ontology

I. INTRODUCTION

According to We Are Social's compendium of world digital stats, Indonesia has 88.1 millions of active internet users, increasing 15% from January 2015 to January 2016. [1]. Online-based media articles are the most popular medium to share the current state of what is happening and are the most accessible media of the community to get information [2]. Classification of news articles is one of the data examples in test domain that are widely studied.

Several methods have been implemented to classify and cluster documents, especially in Indonesian news articles like Naïve Bayes Method [3] and Latent Dirichlet Allocation (LDA) [4]. LDA shows better performance in classification compared to the naïve Bayes [4]. Although LDA provides satisfying performance in classifying documents, LDA still has limitations of its ability, which is to label the formed cluster when grouping words in documents into a certain cluster.

On the other hands, there are several resources that can be used to label the cluster, such as dictionaries and linguistics experts, which in the Indonesian language the most commonly used reference is Kamus Besar Bahasa Indonesia (KBBI). KBBI is a dictionary containing a collection of words and their definition. The linguistics expert is people who understand the meaning and relation of words that contain in KBBI. However, the use of dictionary and linguistics experts directly to a create generic label of clustered documents is less effective. For example, if there is a cluster containing plane, train, and motorcycle, then the perfect label of this cluster would be "vehicle", but it also can be label as "transportation", "motor powered vehicle", and so on. The problem of labeling was often solved by human experts when cluster titles are given manually [5] [6].

This study aims to automatically create a generic label in clustered documents for easier interpretations. To make the automatic labeling feasible, we combine ontology scheme with semantic similarity measure to map the terms in LDA formed cluster into their generic titles. There are several methods of semantic similarity measures existed, such as path based, IC based, feature based and hybrid method [7]. According to research [8], hybrid similarity gives better performance compared to the other model, Because of that, this research we use Hybrid measure (Zhou similarity) to calculate the similarity between words in clustered documents to create generic label of the cluster.

The rest of this paper is organized as follows. Section 2 describes the methodology of the research and proposed model. Section 3 describes the experimental setup, scenarios and results. Subsequently, the conclusion is drawn in Section 4.

II. METHODOLOGY

In this study, we implement LDA to form the cluster and Zhou similarity to measure between terms in a cluster and to

generate the label of the cluster. The ontology used in this research is based on "kata bidang" on Kamus Besar Bahasa Indonesia (KBBI). The general process of proposed model is described in figure 1.

The first step of this research is preprocessing, which aims to process raw data to be ready for the next step. Preprocessing in this research includes tokenization, stop word removal, and n-gram splitting. The next step is to create clusters consisting of a collection of words using LDA. The last step is using Zhou similarity and ontology to generate the generic label of the clusters.

To get the relevance value of the label, we use Cohen's Kappa coefficients with two linguistic experts and calculate the relevant value of each label. It's called as mean relevance rate.

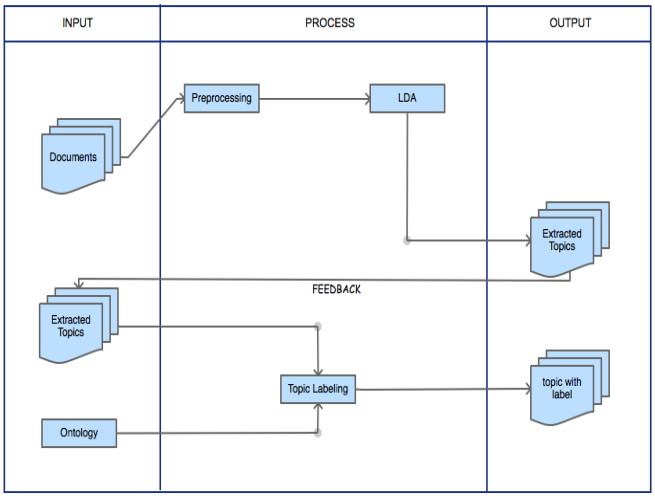

Fig. 1. Research Framework

A. Pre-processing

The aim of pre-processing is to make raw data ready for the next step. In this research, preprocessing stage is divided into three steps described as below.

- Tokenization

  Tokenization is a process of breaking a document into single term by converting the words in a document into lowercase and eliminating white space and punctuation marks (e.g. full stop, comma, semicolon, slash, underscore, question mark, etc.). Subsequently, we perform stop word removal for the next step.

- Stop word removal

  Stop word removal is a process of removing the words which have less significant meaning. We use 759 words that are considered as stop words in Indonesian language, such as "walau", "yang", "sebut", "ada", "adalah", "agar", etc. The next step is n-gram splitting.

- N-Gram Splitting

  n-gram splitting is a process to split the term to a contiguous sequence of n term. Because there are words in the field dictionary of KBBI which contain up to 4 phrases (e.g. amal makruf nahil munkar), the n-gram process will split the term from unigram up to four-gram in order to match the words in the document to the words in the dictionary.

B. LDA-based Classification method

LDA is a probabilistic topic model of which each document is represented as a random mixture over a set of latent topics, and each of the topics is represented as a distribution over vocabulary [9].

LDA uses a Dirichlet distribution to generate a set of words with a specific topic. The topic is a distribution probability of words, and each document is a random mixture of several topics according to a certain proportion [10]. This study uses Gibbs sampling as inference algorithm, which the formula is described below [11]:

$$p(z_i = k | z^{(-i)}, w) = \frac{n_{k,-i}^{(w)} + \beta}{n_{k,-i}^{(*)} + V * \beta} * \left(n_{d,-i}^{(k)} + \alpha\right) \quad (1)$$

$$\theta_j{}^{(m)} = \frac{n_j{}^{(m)} + \alpha}{n_{\cdot}{}^{(m)} + k\alpha} \quad (2)$$

$$\varphi_j{}^{(w)} = \frac{n_j{}^{(w)} + \beta}{n_j{}^{(\cdot)} + w\beta} \quad (3)$$

where:
m: documents
z: topics
w: words
α: hyperparameter topic – documrent distribution
β: hyperparameter words – topic distribution

In this study, LDA is used to form the topics of the documents. Topics formed from LDA process are still unlabeled, or just a mixture of word probability of each topic. Next step is labeling the topics.

C. Ontology

Ontology is a formal description of a concept explicitly in a domain, the property of each concept and its limits. A concept in the ontology can have objects (instances). Technically, ontology is represented in the form of class, property, facet, and instance [12].

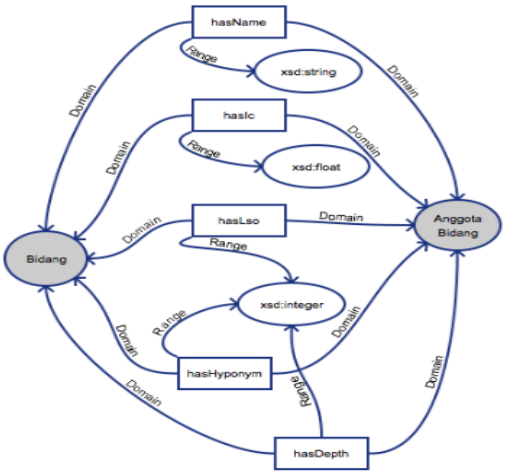

Fig. 2. Ontology Graph

In this research, the ontology scheme is built based on domain dictionary field contained in KBBI where each word will be divided into two classes namely the field and its members. Each of them has attributes attached to each individual of class each, such as name, IC, Lowest sub-ordinate (lso), hyponym, and depth. Ontology graph in this study is presented in Fig. 2.

D. Topic Labeling

Semantic similarity measures can be used to overcome the ambiguity of a sentence or term [13], text segmentation [14], and to examine consistency and coherency of ontology [7]. This study uses ontology scheme to build word-to-word relationship and Zhou similarity to measure the similarities between words in each topic to generate the topic label.

Semantic similarity measures are divided into four categories, which are path based, information content based, feature based, and hybrid measure. This study uses Zhou similarity that belongs to hybrid measure to calculate the similarity between words on each topic. The formula of Zhou similarity is described below [8]:

$$sim_{zhou}(c_i, c_j) = 1 - k\left(\frac{\log(len(c_i, c_j) + 1)}{\log(2 * (deep_{max} - 1))}\right) - (1 - k) * ((IC(c_i) + IC(c_j) - 2 * IC(lso(c_i, c_j)))/2) \quad (4)$$

where:
len(c_i, c_j): the shortest path from c_i to c_j
lso(c_i, c_j): the lowest common *subsumer* of c_i and c_j
$deep_{max}$: the maximum depth of the taxonomy
IC(c): information content of c
k: weight factor

The weight factor (k) can be manually adjusted. Based on previous research [8], the weight factor used in this study is 0.5.

III. Experiment and Results

This section consists of three sub sections, i.e. (i) experimental setup, (ii) scenarios of experiments, and (iii) results and analysis.

A. Experinmental setup

The dataset used in this study consists of 50 Indonesian news articles taken from various online news portals. The experiment is divided into two scenarios: the first scenario divides the dataset into 15 topics using LDA with different words representation, while the second scenario calculates the relevance value of each words representation that has a kappa value greater than 0.4. This threshold aims to keep the reliability value not too low.

This research implementation uses Code Igniter framework, which runs on hardware specification as follows:

- Intel Core i5 2.5GHz
- 10 GB of Memory
- 500 GB of hard disk drive

B. Scenarios of Experiments

As explained before, we compare the number of words representation for each topic as the parameter. The topic is divided into 15 topics, and the words representation of each topic is divided into 7, which are 5 words per topic, 10 words per topic, 15 words per topic, 20 words per topic, 25 words per topic, 30 words per topic, and 35 words per topic.

Based on the scenario explained before, the first scenario aims to identify which words representation has kappa values above the threshold. The second scenario aims to identify which words representation has the higher mean relevance rate.

C. Experiments Result and Analysis

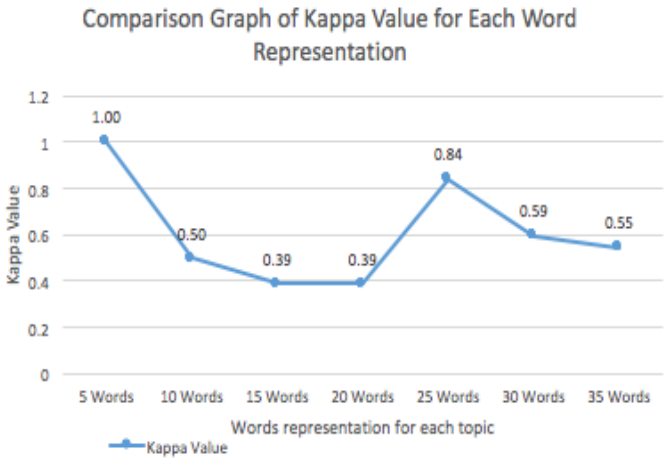

Fig. 3. Comparison Graph of Kappa Value for each words representation

Fig. 2 shows the experiment result for the first scenario that is the comparison of kappa value for each words representation. According to Fig. 2, the highest kappa is 1 or 100%, which is obtained at 5 words representation for each topic. The second rank obtained at 25 words representation for each topic with the kappa value 0.8 or 80%, and the lowest kappa value is 0.39 or 39%, which is obtained at 15 and 20 words representation of each topic.

Based on Fig. 2, there is two Kappa value below the threshold, which are at 15 words representation and 20 words representation. High kappa values indicate the labels of the topics have a high consistency of relevant agreement between experts. The relevant agreement between experts can be influenced by a word that has a unique meaning, where the word has only one meaning and only belongs to one field.

Furthermore, 5 words representation, 10 words representation, 25 words representation, 30 words representation, and 35 words representation are processed to the second scenario.

TABLE I. RELEVANCE VALUE AND RELEVANCE RATE FOR EACH WORDS REPRESENTATION

Words representation	Kappa Value	Relevance Value		Relevance Rate		Mean relevance Rate
		Expert 1	*Expert 2*	*Expert 1*	*Expert 2*	
5 Words	1.00	12	12	0.80	0.80	0.80
10 Words	0.50	10	6	0.67	0.40	0.53
25 Words	0.84	11	10	0.73	0.67	0.70
30 Words	0.59	13	11	0.87	0.73	0.80
35 Words	0.55	12	9	0.80	0.60	0.70
Average Value	0.61	11.6	9.6	0.77	0.64	0.71

a.

b.

Based on Table 1, 5 words representation has 1 or 100% kappa value, 12 relevant labels out of 15 labels according to expert 1 judgment, and 12 relevant labels to expert 2 judgment with the mean relevance rate for 5 words representation is 0.80 or 80%. 10 words representation has 0.5 or 50% kappa value, 10 relevant labels out of 15 labels according to expert 1 judgments, and 6 relevant labels to expert 2 judgment with 0.53 or 53% of mean relevance rate value. At 25 words representation, expert 1 gives 11 relevant value, and 10 from expert 2 with the mean relevance rate is 0.70 or 70%. 30 words representation has 13 relevant judgments valued from expert 1 and 11 relevant judgments valued from expert 2 with 0.80 or 80% of mean relevance rate value. 35 words representation has 12 relevant labels according to expert 1 judgment and 9 relevant labels according to expert 2 judgment with 0.70 or 70% of mean relevance rate value. 5 and 30 words representation have the highest value of mean relevance rate, thus indicate the experts provide a highly relevant value for the label with 5 and 30 words representation.

At 5 words representation, the value of relevant label is quite high due to the existence of label “umum”. When a mixture of words does not fit into any specific topic, the label of the topic will be set as “umum”. This label united the opinions between two experts when both have different interpretations, so their opinions can be put together under an “umum” label. Different conditions occur in 30 words representation. Since the topic contains many words, there are some words having more than one meaning, or belong to more than one field, so the experts have many choices to associate words with labels. This is why the relevant value on 30 words representation is quite high.

IV. CONCLUSION

In this study, the results of LDA, to classify topics, and ontology scheme, to label the topics formed by LDA, give the average kappa value of 0.61 or 61%, with the highest is at 5 words representation for each topic with 1.00 or 100% kappa value, and the lowest are at 15 and 20 words representation with 0.39 or 39% of kappa value. High kappa value indicates the experts have high value of consistent relevant agreement (expert 1 and expert 2 give the same agreement on the same topic). The high value of rate agreement can be influenced by two things. The first is general topic called “umum” that can unite judgments of the experts. The second is word meaning. If a topic contains many words with unique meaning, the experts can easily agree to give the relevant value of the label, as long as the label matches to its words representation.

With the average kappa value is 0.61 or 61%, it can be said that the results obtained from this study is relatively good, while the measurements of mean relevance rate give the best result of 0.8 or 80% value at 30 and 5 words representation with the average value of mean relevance rate is 0.71 or 71%. Finally, it can be concluded that the ontology scheme used in this study is quite relevant to label the datasets.

ACKNOWLEDGMENT

The authors would like to acknowledge the research funding supported by Universitas Diponegoro under the grant of research for international scientific publication – Year 2017 (number 276-36/UN7.5.1/PG/2017).

REFERENCES

[1] J. Balea, "Tech in Asia," Tech inAsia, 28 1 2016. [Online]. Available: https://www.techinasia.com/indonesia-web-mobile-statistics-we-are-social. [Accessed 15 2 2017].

[2] C. Za'in, M. Pratama, E. Lughofer and S. Anavatti, "Evolving Type-2 Web News Mining," *Applied Soft Computing,* 2016.

[3] I. Saputra, Y. Sibaroni and A. P. Kurniati, "Analysis and Implementation of Classification Indonesian News with Naive Bayes Method," 2008.

[4] R. Kusumaningrum, S. Adhy, M. I. A. Wiedjayanto and Suryono, "Classification of Indonesian News Articles based on Latent Dirichlet Allocation," in *International Conference on Data and Software Engineering*.

[5] K.-K. Lai and S.-J. Wu, "Using the patent co-citation approach to establish a new patent classification system," *Information Processing and Management: an International Journal,* vol. 41, no. 2, pp. 313-330, 2005.

[6] P. Glenisson, W. Glanzel, F. Janssens and B. De Moor, "Combining full text and bibliometric information in mapping scientific disciplines," *Information Processing and Management: an International Journal - Special issue: Infometrics,* vol. 41, no. 6, pp. 1548-1572, 2005.

[7] L. Meng, R. Huang and J. Gu, "A Review of Semantic Similarity Measures in WordNet," *International Journal of Hybrid Information Technology,* vol. 6, no. 1, pp. 1-12, 2013.

[8] Z. Zhou, Y. Wang and J. GU, "New Model of Semantic Similarity Measuring in WordNet," in *3rd International Conference on Intelligent System and Knowledge Engineering*, 2008.

[9] D. M. Blei, "Probabilistic Topic Models," *Surveying a suite of algorithm that offer a solution to managing large*

document archives, vol. 55, no. 4, pp. 77-84, April 2012.

[10] Z. Liu, *High Performance Latent Dirichlet Allocation for Text Mining,* London: Brunel University, 2013.

[11] G. Heinrich, "Parameter Estimation for Text Analysis," Fraunhofer IGD, Darmstadt, Germany, 2009.

[12] N. F. Noy and D. L. McGuinness, "Ontology Development 101: A Guide to Creating Your First Ontology," Stanford, 2001.

[13] S. Patwardan, S. Banerjee and T. Pedersen, "Using Measures of Semantic Relatedness for Word Sense Disambiguation".

[14] H. Kozima , "Computing Lexical Cohesion as a Tool for Text Analysis," 1993.