

Latent Dirichlet Allocation (LDA) for Sentiment Analysis Toward Tourism Review in Indonesia

¹I R Putri, ²R Kusumaningrum

^{1,2}Department of Informatics, Faculty of Science and Mathematics, Diponegoro University, Semarang

E-mail: ¹indiati@student.undip.ac.id, ²retno@live.undip.ac.id,

Abstract. The tourism industry is one of foreign exchange sector, which has considerable potential development in Indonesia. Compared to other Southeast Asia countries such as Malaysia with 18 million tourists and Singapore 20 million tourists, Indonesia which is the largest Southeast Asia's country have failed to attract higher tourist numbers compared to its regional peers. Indonesia only managed to attract 8,8 million foreign tourists in 2013, with the value of foreign tourists each year which is likely to decrease. Apart from the infrastructure problems, marketing and managing also form of obstacles for tourism growth. An evaluation and self-analysis should be done by the stakeholder to respond toward this problem and capture opportunities that related to tourism satisfaction from tourists review. Recently, one of technology to answer this problem only relying on the subjective of statistical data which collected by voting or grading from user randomly. So the result is still not to be accountable. Thus, we proposed sentiment analysis with probabilistic topic model using Latent Dirichlet Allocation (LDA) method to be applied for reading general tendency from tourist review into certain topics that can be classified toward positive and negative sentiment.

1. Introduction

Nowadays, the tourism industry has experienced continued growth and deepening diversification to become one of the fastest growing foreign exchange sectors in the world that acknowledged by UN and other various international organizations such as World Bank and World Tourism Organization (WTO). These dynamics have turned tourism into a key driver for socio-economic progress in the world including Indonesia. In tourism industry, compared to other Southeast Asia countries such as Malaysia with 18 million tourists and Singapore 20 million tourists, Indonesia which is the Southeast Asia's largest country have failed to attract higher tourist numbers compared to its regional peers. Indonesia only managed to attract 8.8 million foreign tourists in 2013, with the value of foreign tourists each year which is likely to decrease such as in 2013 with 9.43% of foreign tourists every year but in 2014 only 7.19%. Meanwhile, if we compared to other countries, sceneries, and attractions that had been offered in Indonesia was not much different from other countries in South-East Asia. Apart from the infrastructure problems, marketing and managing also form of obstacles for tourism growth in Indonesia. Tourism management itself cannot be separated from the presence of data sources which manually managed by institutions that are directly related to the state of foreign tourists, i.e the Ministry of Manpower and Immigration Office, Central Bureau of Statistics (BPS), Bank Indonesia, Ministry of Culture and Tourism (Depbudpar), Department of Tourism regions (Disparda), and the Indonesian Hotel and Restaurant Association (IHRA). In general, the



data sources used still relies on statistical data in the form of secondary data that have been most affected by the data collection period (periodic) so that its implementation would need a huge number of resources. The difficulties encountered when the statistics then combined or compared with other numerical statistics data from other countries, as the result given an imprecise number and increase the risk of human error.

One of the technologies that have been developed to address these challenges is an online survey ever conducted by several websites like www.indo.com. But the accuracy of the online survey is still not accountable cause the user or the survey participants can not precisely be known. Other technologies such as grading and rating which are widely implemented in many cases of the online survey also did not give the results that can be assessed objectively because it is only given a fixed value based on numerical user given input. In this case, online opinion or online review can be considerate as one of most appropriate media that can be assessed more objective to represent user satisfaction or user impression through text. Online review is one of an unstructured text form that can be used in extracting data and gain more information related in user impression using sentiment analysis as one of *Information Extraction* process. Sentiment analysis is an *Information Extraction* and *Natural Language Processing* task that aims to obtain writer's feelings expressed in positive or negative review by analyzing large numbers of documents as a corpus.

The purpose of this research is to study, implement and analyze the most suitable techniques in the case of sentiment analysis for tourism review in Indonesia. In this research approach we proposed an unsupervised technique using probabilistic topic models to classify online review based on the sentiment behind those reviews. Currently, probabilistic topics models became a model which most developed and applied in a variety of research and well known to produced a pretty good effect on various applications especially in the field of classification of text and Information Retrieval [1]. As an advantage of probabilistic topics models, large data corpus can be interpreted as latent semantic in more comprehensive and stable manner [2]. Probabilistic topics models consist of Latent Semantic Analysis (LSA), Probabilistic Latent Semantic Indexing (PLSA) and Latent Dirichlet Allocation (LDA). This research use LDA method to performed classification on online review for sentiment analysis, since the LDA method designed to address problems in the LSA and the PLSA. LDA is also known to have a good ability and stability in handling large data scale because it gives the parameters as a random variable [3][4]. The rest of the paper as follows: in section 2 the overview and literature survey of some techniques which are commonly used in sentiment analysis. Then in section 3 present the proposed sentiment analysis approach method, techniques. Section 4 present the experimental process, result, and analysis. Finally, the main conclusions and future directions of research are presented in section 5.

2. Literature Survey

Sentiment Analysis is a large growing field which first brought by Liu in 2012. Sentiment Analysis also known as Opinion Mining is a field of study that provides analysis of opinion, sentiment, evaluation, attitudes, and emotions against the entities such as product, service, organization, individual, problem, events, topics, and other attributes [5]. The idea of Sentiment analysis is to determine the attitude of a writer through online text data toward certain topic or the overall tonality of a document. Sentiment analysis has been handled as natural language processing task in many ranges of field area including hotel, restaurant, tourism, devices and other consumer or marketing area. In the field of tourism, a corpus for sentiment analysis used online text data through travel sites such as TripAdvisor, Expedia, and Booking.com[8].

Various supervised or data-driven techniques such as Naïve Bayes, Maximum Entropy, SVM has been implemented for sentiment analysis in microblog data like Twitter. However some of those techniques tend to gives a rigid result, and some only give the best result in a feature based.

In this study, we design sentiment analysis using probabilistic topic model as one of topic modeling approach which assumes a document is a distribution over a topic and topic is represent as the distribution over word [1]. Probabilistic topic model is one of Topic Modelling approach which provides convenience to automatically organize, search and summarize large scale data. This model also proved to be well implemented on text fields and retrieval of information [1]. Implementation of

probabilistic topic models on document will generate a set of low-dimension polynomial distribution, and for each polynomial distribution then referred as topic. Topics will be used to capture information which related to each word in the document, so each topic can be generated and extracted into a semantic structure [2][6]. Recently one of probabilistic topic models that are very popular to use, and has more comprehensive assumptions on the generation of text compare to other methods is Latent Dirichlet Allocation (LDA)[6].

The development of LDA originally intended to address problems in PLSA (Probabilistic Latent Semantic Analysis). On PLSA, the input is only determined by one hyperparameter value symbolized as beta (β). While LDA perfected with the addition of hyperparameter alpha (α), then both values will generate a latent variable value which symbolized as theta (θ) and phi (ϕ). Therefore the number of parameters in LDA does not grow according to the size of train corpus.

LDA method is divided into two difference processes, i.e. LDA as generative process and LDA as inference process. LDA as inference process has been characterized by the availability of corpus data as observed variable to gain latent variable including the word distribution over topic (ϕ) and topic proportion for each document (θ). Otherwise, LDA as generative process implemented to create corpus data or a set of document from latent variable which already known. LDA as one of the probabilistic topic model that we implemented in this study applied as inference process by using a Gibbs sampling algorithm as approximate posterior inference algorithm.

3. Research Approach

3.1. Problem formulation

In this section, we present the problem formulation for the research approach. In a case of tourism in Indonesia, we propose a sentiment analysis that will be implemented to identify positive and negative sentiment given the user or visitors through the review on a travel site in order to gain the overall sentiment tendency toward tourism in Indonesia.

Sentiment analysis is composed of three major processes i.e. Subjectivity Classification, Orientation Detection, and Opinion Holder and Target Detection. The first process, Subjectivity Classification is applied when we want to determine whether a sentence is an opinion. For the second process, Orientation Detection is applied to determine whether the opinion is positive or negative. And the last process, Opinion Holder, and Target Detection are applied to determine who is opinionated and define the object of opinion [7]. Sentiment Analysis in this research implement orientation detection process, since we aim to identify positive and negative sentiment in tourism review.

The tourism review that we used as a corpus in this research is collected from TripAdvisor since these sites applied a systematic rule in publishing review so it will provide a better data as a corpus. Furthermore, the number of review's contributors that scattered in various countries in the world became a reasoning appeal for using TripAdvisor as a corpus data since we only used review in an English language.

According to the previous explanation, an appropriate technique to applied in this issue is Sentiment Analysis with orientation detection process in the document-level task. Thus, for each review given by user or visitors through travel site will be identified and classified as each one value of sentiment.

3.2. LDA for Sentiment Analysis

As mention before, this research implement LDA as inference process for sentiment analysis especially for document-level classification task. In the case of LDA for document-level classification task, there is two main processes with each difference data for those process, training data for training process and the testing data for testing process. Training and testing data is a result of the 10-fold cross validation process that implemented to divide the data in a balanced and objective manner. In advance of the main process, for both of training and testing data will be implement basic text processing i.e. tokenization, stopwords removal and stemming. Thus text processing aims to eliminate the words, symbols or characters were give insignificant meaning for information extraction.

Tokenization is a process to eliminating characters that likely insignificant for information extraction such as white space and punctuation marks. Afterwards, the second process is stopword removal, implemented to eliminating a word that likely insignificant for information extraction. In this research, there are 277 words in the English language that are considered as stop-words, which consist 55 slang words as an addition. The last text processing, stemming is a process to derive each word to their root form. This research implement porter's stemming for stemming process.

As the input of training process is Bag of Words (BoW) which consist a set of words token from all training document that has been processed through text processing. In training process, we implement LDA as inference process over the BoW to produce a classification model. In term of implementing LDA over the BoW, there is several parameter to be defined, there are number of iteration, number of topics, hyperparameter alpha (α) and beta (β). As the result of implementing LDA as inference process, there is topic distribution and topic proportion for each document and word distribution for each topic. Subsequently, we implement harmonic mean over topic proportion for each document to compute topic proportion for each class including positive and negative class since the aims of this classification process are to classify each document into two difference sentiment classes. Afterward, the classification model for LDA as inference process in document is value of topic proportion for each class, topic proportion for each document, word distribution for each topic and topic distribution, which are used in testing process.

The testing process used testing data, which consists an extracted word for each testing document. In testing process, we applied Kullback Leiber Divergence (KLD) to measures the similarity between two distribution, the topic proportion for each class from training process as defined as classification model and the topic proportion of testing data. Thus the smallest value of KLD indicated as the sentiment class of the document.

4. Experiment and Result

4.1. Experimental Setup

As mentioned in the section before, dataset we used as a corpus is collected from TripAdvisor tourist's reviews site which consists of 100 review in English language for two sentiment classes and each class consists 50 reviews. For dividing the dataset we applied the 10-folds cross-validation, thus one folds define as testing data and the rest folds as training data and it is performed repeatedly. Accordance with 10-folds cross-validation concept, for each fold will be assigned as testing data alternately so each data will be used as testing and training data. Hereafter, the accuracy value is measured for each fold and each average of accuracy over 10-folds and is computed to indicate performance for the classification model.

The experiment is implemented using Matlab programming in Windows 8.1 Pro – 64 bit, with the following hardware specification :

- Inter Core i7- 4710HQ CPU 3.50GHz
- 4 GB of memory (RAM)
- 1 TB of Hard Disk Drive

4.2. Scenario Experiment

In this study, we compare of several parameters in LDA as a combination experiment. The first parameter is number of iteration, N, with N = 9000, 10000, 30000, 50000, 70000, 90000, 100000, or 300000. The second, number of topics, T, with T = 5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 55, 60, 65, 70, 75 or 90. Another parameter is hyperparameter alpha (α), $\alpha = 0.1$ or $50/T$ (number of topics), which for all combination hyperparameter beta defined as 0.1.

Based on those parameter combinations, we conduct two scenario of experiments to obtain the best classification model for sentiment analysis. The aim of the first scenarios is to compare each combination and obtain the information of parameter affect for each accuracy in classification model. The second scanarion aims to identify the best parameter combination that given the highest accuracy value.

4.3. Experiment Result and Analysis

previous subsection, as the result of the first scenario showed in Figure 1. Figure 1(a) shows the comparison of accuracy value for each combination of number of iteration in two alpha. As seen in Figure 1 (a) we used eight parameter number of iteration as combination, while Figure 1 (b) shows the comparison of accuracy value for 16 parameter number of topics as combination in two alpha.

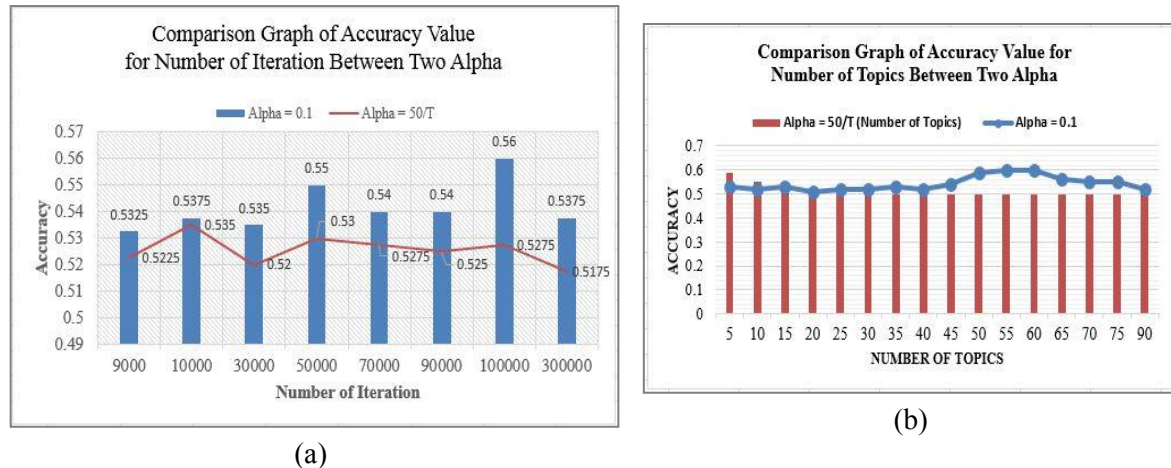


Figure 1. Comparison Graph of Accuracy Value for Number of Iteration (a) and Number of Topics (b) between two alpha.

Through comparison graph in Figure 1 (a) we can draw a conclusion that the number of iterations is also one of parameters that play an important role to obtain a good accuracy values by satisfying the stationary distribution in model classification using LDA. Although in science there are no definite methods in determining the appropriate number of iterations to reached a good accuracy and meet the given distribution, but in LDA as inference process, the word distribution of each topic and proportion topics will be strongly influenced by the number of iterations value. As conclusion the influence of number of iterations on the value of accuracy in case of sentiment analysis using LDA can be define that when the number of iteration is too small will increase the risk of Gibbs sampling to give a false representation of the posterior distribution, while the number of iterations is too high will certainly pose a greater computational cost. As well as the number of iteration, the least amount of number of topics that did not coincide with the data and the value of hyperparameter alpha will obtain an overly broad interpretation and tend to pose a high ambiguity. Otherwise, The number of topics that are too high and did not coincide with the data and the value of hyperparameter alpha can create a very distinctive interpretation so it will be difficult to be categorized or interpreted in a certain topic. The comparison of word distribution over topic to present the ambiguity interpretation can be seen in Figure 2.

Figure 2 shows the comparison of word distribution over topic in two difference number of topic. In small amount of number of topics shows the ambiguity interpretation when “monkeys” were interpret in several time for difference topic. As seen in Figure 2 for higher number of topics, the word presented in more distinctive manner.

Figure 3 shows, higher value of hyperparameter alpha determines the interpretation level of words over topic for more evenly distributed. Thus, the higher value of alpha will reach more stable posterior distribution when the number of topic is used in smaller value. The opposite applies, smaller value of alpha will interpret each words over topics diversely so it required higher value of number of topic to meet the stable posterior distribution.

Number of Topics = 55			
TOPIC_1	0,01545	TOPIC_2	0,02113
W-1066:steal	0,01266	W-1223:up	0,04011
W-1238:view	0,01266	W-212:close	0,03129
W-83:attract	0,00663	W-404:first	0,01807
W-101:bad	0,00663	W-134:bit	0,01366
W-109:bananasoverall	0,00663	W-447:getting	0,01366
W-129:better	0,00663	W-590:jumped	0,01366
W-150:bought	0,00663	W-929:rip	0,01366
W-154:breathtaking	0,00663	W-89:avoid	0,00926
W-186:centre	0,00663	W-100:backpack	0,00926
W-381:fee	0,00663	W-345:everything	0,00926

Number of Topics = 5			
TOPIC_1	0,18435	TOPIC_2	0,25842
W-711:monkeys	0,03894	W-711:monkeys	0,03724
W-750:not	0,02730	W-415:forest	0,02676
W-884:rabies	0,02471	W-710:monkey	0,02676
W-135:bite	0,01824	W-750:not	0,02581
W-137:bitten	0,01695	W-593:just	0,01629

Figure 2. The comparison of word distribution over topic

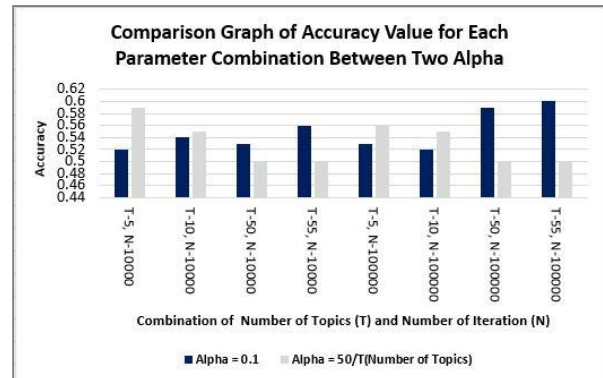


Figure 3. The Overall Comparison Graph of Accuracy Value for Each Parameter Combination Between Two Alpha

5. Conclusions

In this proposed study, we used Latent Dirichlet Allocation (LDA) as document classification method for sentiment analysis in TripAdvisor's review. The experiment was performed to evaluate the performance of the LDA-based document classification method in sentiment analysis by using 10-folds cross-validation techniques. As the result shows the best accuracy is about 60% as an average accuracy of all folds and the best accuracy is about 80% by fold 6 and 7.

Based on the experimental setup, a modification of each parameter induce different effects. A balance in each portion of the parameters is needed to obtain optimum levels of accuracy. The number of iterations or number of topics that are too small will certainly make the bias level of LDA going higher. Otherwise, the number of iterations that are too high will increase the cost computational, while the number of topics that are too high will lead LDA to difficult to interpreted the words in a certain topic.

In inversely related to the number of iterations and the number of topics, the value of hyperparameter alpha that are too small need to be balanced with the amount of the number of iterations and the number of topics to achieve the optimum accuracy. Where $\alpha \geq 0.1$ would be better used for the classification.

References

- [1] Blei D 2012 Probabilistic topic models *Communications of the ACM* 55 77
- [2] Arora S, Ge R and Moitra A 2012 Learning Topic Models -- Going beyond SVD 2012 *IEEE 53rd Annual Symposium on Foundations of Computer Science* 2 1-10
- [3] Girolami M and Kaban A 2003 On an equivalence between PLSI and LDA *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval - SIGIR '03* 433
- [4] Lu Y, Mei Q and Zhai C 2010 Investigating task performance of probabilistic topic models: an empirical study of PLSA and LDA *Information Retrieval* 14 178-203
- [5] Liu B 2012 Sentiment Analysis and Opinion Mining *Synthesis Lectures on Human Language Technologies* 5 1-167
- [6] Liu Z and Li M 2013 *High performance latent dirichlet allocation for text mining* (London: Brunel University)
- [7] Zagibalov T and Carroll J 2008 Automatic seed word selection for unsupervised sentiment classification of Chinese text *Proceeding of the 22nd International Conference on Computational Linguistics*
- [8] Liu S, Law R, Rong J, Li G and Hall J 2013 Analyzing changes in hotel customers expectations by trip mode *International Journal of Hospitality Management* 34 359-371