Aspect Based Sentiment Analysis in E-Commerce User Reviews Using Latent Dirichlet Allocation (LDA) and Sentiment Lexicon

Eko Wahyudi
Department of Informatics
DiponegoroUniversity
Semarang, Indonesia
eko@students.undip.ac.id

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

***Abstract*—User ratings on products sold by e-commerce greatly influence the number of purchases. Positive ratings will encourage other buyers to participate in buying the product. While negative ratings given by users will reduce the interest in purchasing products. Nonconformities between rating and user reviews sometimes provide a wrong assessment of a product. This happens because buyers also provide reviews on the quality of delivery services from e-commerce. Based on that issue, the utilization of the Latent Dirichlet Allocation (LDA) could be used on sentiment analysis of the user reviews. Sentiment analysis of the user reviews aims to facilitate e-commerce in informing the product quality as rating supporters that have been given by users. This research aims to determine the classification performance of sentiment analysis on e-commerce user reviews using the LDA algorithm with input data in the form of e-commerce user reviews. Then, compare the application of sentiment analysis of the user reviews with the use of general training data and per category training data. The result of this research showed that in the first iteration the best architecture was produced by the application of LDA with a combination of parameters of alpha 0.001, beta 0.001, and number of topics 15. The architecture had 67,5% accuracy level. From the best architecture then training data input is given based on each product review category. The result showed that the combination of the usage of general data and per category data indicate an increase in the average accuracy of 0,82% from the three-test data. Therefore, in order to produce the best performance of building a classification model of sentiment analysis of the user reviews, it should be performed by applying LDA with a combination of general data and per category data usage.**

Keywords*—*latent dirichlet allocation; sentiment analysis; e-commerce; user reviews; product quality;

I. Introduction

Internet plays an important role in the development of e-commerce in various countries. The increasing number of internet users in a country means the greater of e-commerce market in the country. In Indonesia, e-commerce is developing quite rapidly, it is marked by the number of online shops that have sprung up and competed with each other such as Tokopedia, Lazada, Elevania, OLX, Blibli, Bukalapak, and so on. In data released by the Indonesian Internet Service Providers Association (APJII), in 2014 internet penetration in Indonesia had reached 34.9% with an internet user number of 88.1 million out of a total population of 252.4 million. The penetration increases with the increase of internet users from year to year. From the data it was also mentioned that 27% of internet users in Indonesia had done online shopping[1]..

The internet makes buying and selling transactions easier and faster. E-commerce users are now using a lot of online store facilities in buying and selling, because it feels easy and users do not have to bother going to a physical store to buy the desired product. However, determining the quality of products when shoppingat online stores is still a problem for some people. Because not all online stores sell their own products. This will affect the quality assurance of the product being sold, because the goods sold are not always the same as what the buyer receives. This can be seen from the user reviews given on the products that have been purchased. Not a few products that have a low rating, because the quality of the product is not in accordance with the wishes of the user, although there are also many products that have a high rating because the product quality is good. The low rating will affect the decline in buyer interest in the future.

The problem that arises is the discrepancy of user reviews given the rating value. There are buyers who give low ratings but the reviews are pretty good and there are buyers who give high ratings but have less good reviews. The problem is caused by some buyers not only reviewing their products but also the quality of shipping services from e-commerce owners. The mismatch between ratings and reviews makes the process of determining product quality from e-commerce owners more difficult. To overcome this, it is necessary to analyze the existing user reviews, so that e-commerce owners can more easilydetermine product quality more objectively based on user reviews.

Sentiment analysis or opinion mining is a form of new technology that is widely used in research. Sentiment analysis is part of a system intended to explore user opinions provided through various media, both through print media, social media, blogs and websites to show user satisfaction and attitude towards a topic. Sentiment analysis itself can be defined as a form of classification of documents based on the overall expression of sentiment expressed by users or commenters[2].

Probabilistic topic modeling is now a method used in various studies in the field of sentiment analysis. Probabilistic

978-1-7281-4610-2/19/$31.00 ©2019 IEEE

topic modeling in general can extract features comprehensively, is stable in the form of latent semantic and can express the closeness of several data descriptions in large data capacities. [3]. Latent Dirichlet Allocation (LDA) is one of the methods of probabilistic topic modeling that is used to find out the topics that emerge from a document[4]. LDA was chosen because it has a good ability to handle large scale data and better structural stability because it provides parameters as random variables[5].

One type of the LDA method is LDA as an inference process. LDA as inference is characterized by the availability of data objects to be observed and the distribution of topics formed. The inference process on LDA can be implemented using the posterior approximation inference algorithm. Such inference algorithms, such as Mean Field Variational Method, Gibbs Sampling, and Collapsed Variational Inference [6].

This research will use the LDA Collapsed Gibbs Sampling algorithm as the inference process. The use of LDA as an inference process with the Collapsed Gibbs Sampling algorithm is done because in this study the data or documents to be observed have already been determined, i.e. the user review data.

II. Study Literature

A. *Sentiment Analysis*

Sentiment analysis also called opinion mining is a field of study that provides analysis of opinions, sentiments, evaluations, judgments, attitudes, and emotions towards entities such as products, services, organizations, individuals, problems, events, topics, and other attributes. In the industrial field the term sentiment analysis is more widely used, but in education both sentiment analysis and opinion mining have the same tendency to be used. Sentiment analysis and opinion mining mainly focus on opinions that express or express positive or negative sentiments[7].

Although linguistics and Natural Language Processing (NLP) have a long history, before 2000 research on opinions and sentiments was still small. Only then did the research topic become a very active research area. The first reason is because it has a very broad application, almost in every domain. The sentiment analysis industry has also grown due to the poliferation of commercial applications. Second, because it offers many challenging research problems, which have never been studied before. Then the third, because at this time the volume of data about opinions on social media is very large on the internet. Sentiment analysis research not only has an important impact on NLP, but also has a large impact on management science, political science, economics, and social sciences because they are all influenced by people's opinions [7].

B. *Online Review*

Online review sites continue to grow in popularity as more people seek advice from fellow users regarding services and products [8]. One type of online review that can be found is about product reviews. Product reviews can be found in various forms on the web, especially sites that offer specific types of products (such as MP3 players or films), news sites and magazines that have a variety of reviews (Rolling Stone), sites that combine reviews with e- commerce (Amazon), and sites that specifically collect professional reviews or user reviews in various areas (Rateitall.com) [9].

The amount of data about reviews on the internet, making users have difficulty in finding the desired information. This has led to increased research in the field of opinion mining and sentiment analysis, with the aim of providing a system that can automatically analyze user reviews and extract information that is most relevant to users[9].

C. *Latent Dirichlet Allocation (LDA)*

Probabilistic Toopic Modeling is a series of algorithms that aim to find the structure of hidden topics in a large number of document archives[10]. The main idea of probabilistic topic modeling is to assume that the document is a mixed distribution form of several topics, and each topic is a probability distribution of words. Thus, probabilistic topic modeling can be seen as a model of generating topics from a text [10].

Currently the famous probabilistic topic modeling model is the Latent Dirichlet Allocation (LDA) model which has a comprehensive assumption on text generation than the other models. [11]. LDA is part of the probabilistic topic model that inherits all the benefits of Probabilistic Latent Semantic Analysis (PLSA) and describes the semantic model with more faculties. LDA has better semantic conditions when implemented compared to other models and has stronger descriptive strength[11].

The LDA model is divided into two types of development, namely LDA as generative and LDA as inference. In the LDA generative process, it is assumed that the topic has been specified before the document is obtained. Then for each document in the collection extracts words with LDA to produce documents [10].

LDA as an inference process is the opposite form of LDA as a generative process [6]. LDA with the inference process works to learn the form of posterior distribution of the observed latent variables [12]. In the LDA inference process aims to determine the output in the form of latency variables word-topic probability (φ_k) and topic proportion for each document (θ_d) with input in the form of documents (D) in a document collection (corpus) where each document consists of a number of words (N), hyperparameters (α and β), and number of topics (K) [6].

The output of the LDA as an inference process to calculate the topic proportion value for each document ($\theta_{d,j}$) and the word-topic probability value ($\varphi_{j,i}$) can be mathematically defined as in the following (13).

$$\theta_{d,j} = p\,(z=j\,/d\,) = \frac{n_j^{(d)}+\alpha}{n_{\cdot}^{(d)}+T*\alpha} \tag{1}$$

$$\varphi_{j,i} = p\,(w=i\,/z=j\,) = \frac{n_j^{(w)}+\beta}{n_j^{(\cdot)}+W*\beta} \tag{2}$$

Information:
$\theta_{d,j}$= topic proportion

$\varphi_{j,i}$= word distribution

D. Preprocessing

At the preprocessing stage a mathematical representation of documents is made so that the classification of documents can be carried out by a machine, so Bag-of-Words (BoW) are formed[14]. Bag-of-Words that are implemented in this final project are information from i to nth token results, position of token words, and position of token documents. The steps consist of tokenization, stopword removal, and stemming.

1) Tokenization

Tokenization is the first step in preprocessing information retrieval. The task of tokenization is to separate the rows of words in the document into single words. The tokenization process also removes certain characters such as punctuation and changes all the letters in the word to lowercase letter. all author and affiliation lines.

2) Stopword Removal

Stopwords are words that often appear on documents that have no meaning in terms of information retrival [15]. Stopwords must be removed in preprocessing so as not to disturb the indexing process.

3) Stemming

Stemming is the process of transforming words in a text document into the basic form of words. Stemming is used in information retrieval which aims to increase the effectiveness and reduce the size of the data when indexing[15].

E. Evaluation

Performance evaluation of the formation of a classification model of user review data (product review) based on sentiment analysis is carried out using the cross validation method and confusion matrix.

1) K-Fold Cross Validation

Tokenization In k-fold cross validation data will be divided into k partitions of the same size $D_1, D_2, D_3, ..., D_k$. Training and testing are carried out k times. In the i-iteration, the *Di* partition will be test data, in addition to training data. In the first iteration, D_1 will become test data, $D_2, D_3, ..., D_k$ will become training data. Furthermore, the 2nd iteration, D_2 will become test data, $D_1, D_3, ..., D_k$ will become training data, and so on [16].

In this research, the method used is the 5 k-fold cross validation method. The way this method works is when testing as much as 20% of the total product review data is alternately used as test data 5 times against 80% of other product review data used as training data (training). Thus, each product review data will become training data and testing data in turn. The average accuracy is obtained from the average accuracy of the 5 tests.

2) Confussion Matrix

Confusion Matrix is a useful tool for analyzing how well the classifier recognizes tuples from different classes [16]. This confusion matrix method uses a matrix table. Where the confusion matrix model used in this study has been adjusted to the presence of 2 classes of product review quality namely positive class (Class 0), and negative class (Class 1) as shown in Table I.

Then enter the test data into the confusion matrix. After that, calculate the values that have been entered to determine the accuracy value. To calculate accuracy the following equation is used:

$$Accuracy = \frac{\sum_{i=j} C_{i,j}}{\sum_{i=0}^{1}\sum_{j=0}^{1} C_{i,j}} * 100\ \% \quad (3)$$

Information:
$C_{i,j}$= Test document

III. Methodology

This section will explain the steps taken in the research. The steps carried out in this study there are 6 processes, namely as follows:

1) Tokenization

The first process carried out is data tokenisation. This process will receive input in the form of user review data taken from e-commerce product review sites. The tokenization process is carried out on user review data in the form of sentences for removing punctuation marks, tags, and changing to lowercase letters.

2) Stopword Removal

The second process after tokenization is stopword removal. This process will process words from the user review data after tokenization, which will eliminate common words that are not used in the process of forming a classification model because they do not have characteristics as a determinant of the topic of a document (user review) such as words that are, in, but, though, but , and others. The output of the tokenization and stopword removal process is user review data that has met the criteria for the process of identifying data at a later stage.

3) Identification of Training Data and Test Data

The The third process is the identification of training data and test data. This process will divide a number of user review data into sections as training data and test data. Where, user review data from the results of the process of tokenization and stopword removal are carried out data sharing using 5 k-fold cross validation with a ratio of 80% for training data and the remaining 20% is used for testing data. The output of this process is training data for the training process and testing data for the testing process that has been divided according to the comparisons applied.

4) Training

The fourth process is training using the Latent Dirichlet Allocation (LDA) method as an inference process with Collapsed Gibbs Sampling. In this process, it will receive input in the form of a combination of parameters alpha, beta, T (number of topics), and N (number of iterations). Then, the next process will take a number of training data and initialize randomly to each word in the training data so that it includes a particular topic. Furthermore, each word with a particular topic that has been initialized before, will be repeated topic determination based on the results of the calculation of the probability of each topic in the training data. The output of this process is the classification model data in the form of the probability of a word including a particular topic, the probability of each topic, and the average probability of each topic to the document for the same (positive and negative) class.

5) Testing

The fifth process is testing. This process will use test data to assess the accuracy of the architecture of the results of the training process with training data. In testing, to determine the accuracy of the training process will be determined by the evaluation of each document in the test data that provides information in the form of predictions that the data includes a certain class based on calculations using the Kullback-Leibler Divergence (KLD). After all prediction information is available, then based on that information an accuracy calculation will be done using the amount of test data that provides class prediction information in accordance with the actual class that has been determined. The output of this process is the accuracy value of the architecture of each combination that exists to determine the best. The best combination is the combination with the greatest accuracy value and is used for the classification process of e-commerce user reviews in helping determine product quality.

6) Classification and Visualization

The final process is classification and displaying results. In this process, a visualization of the results of the formation of the LDA model for classification of e-commerce user reviews will help determine the quality of the product using input from the best architecture based on the results of the testing process for the formation of classification models with Collapsed Gibbs Sampling in the form of the probability value of a word including a particular topic, probability each topic, and the average probability of each topic in the document is based on the same data class (positive and negative) of the best architecture, as well as data on new user reviews that will be predicted. The output of this process is the classification of e-commerce user reviews to help determine the quality of the product against the new user review data based on its class according to visualization.

IV. Result and Discussion

A. Research Data

In this researchthe data used are user review data obtained from e-commerce sites specifically for gadget product reviews. User review data used in research is divided into two, namely: user review data for the formation of classification models and user review data which will be used as input in the classification process. User review data for the formation of this classification model was obtained from e-commerce website pages and gadget product review. Data is taken by criteria, namely: 50 product review data records for the camera category, 50 product review data records for the screen category, 50 product review data records for the ram category, and 50 product review data records for the battery category. Each of the 50 data records consists of 25 positive review records and 25 negative review records.

B. Research Scenario

The following are the scenarios used in this research:

1) Scenario 1

Scenario 1 illustrates research in sentiment analysis of e-commerce user reviews using the Latent Dirichlet Allocation method and general training data, ignoring the distribution of training data by category. In this scenario calculation is performed using a combination of parameters alpha values (of 0.1, 0.01 and 0.001), beta values (of 0.1, 0.01 and 0.001), the number of topics (2, 4, 8, 5, 10, and 15), and iteration of 1 for the process of forming the classification model with Collapsed Gibbs Sampling. So the total combination of calculation parameters for Collapsed Gibbs Sampling in forming the classification model is 54 paremeter combinations.

2) Scenario 2

Scenario 2 illustrates a sentiment analysis research of e-commerce user reviews using the Latent Dirichlet Allocation method and training data per category. In this scenario, the use of a combination of parameters is taken from the best combination of the results of the model formation carried out in the first scenario.

3) Scenario 3

Scenario 3 will compare the performance of sentiment analysis classification models of user reviews to help determine product quality using the Latent Dirichlet Allocation method using the best combination of parameters obtained in scenario 1 and scenario 2.

TABLE I. CONFUSSION MATRIX MODEL

Confussion Matrix (C)		Actual Class (*j*)	
		Class 0	Class 1
Predicted Class (*i*)	Class 0	$C_{0,0}$	$C_{0,1}$
	Class 1	$C_{1,0}$	$C_{1,1}$

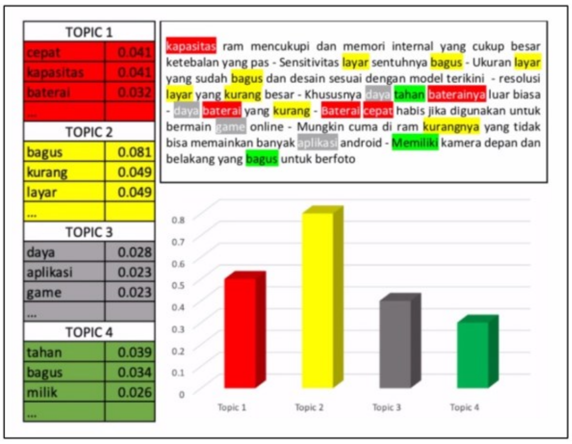
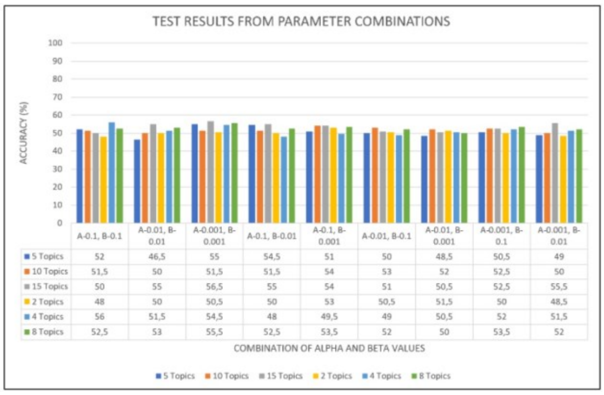


Fig. 1 Example Research Data Visualization



	A-0.1, B-0.1	A-0.01, B-0.01	A-0.001, B-0.001	A-0.1, B-0.01	A-0.1, B-0.001	A-0.01, B-0.1	A-0.01, B-0.001	A-0.001, B-0.1	A-0.001, B-0.01
5 Topics	52	46,5	55	54,5	51	50	48,5	50,5	49
10 Topics	51,5	50	51,5	51,5	54	53	52	52,5	50
15 Topics	50	55	56,5	55	54	51	50,5	52,5	55,5
2 Topics	48	50	50,5	50	53	50,5	51,5	50	48,5
4 Topics	56	51,5	54,5	48	49,5	49	50,5	52	51,5
8 Topics	52,5	53	55,5	52,5	53,5	52	50	53,5	52

Fig. 2. Graph of Average Testing Accuracy Results Scenario 1

C. *Test Results and Analysis*

The results of research on sentiment analysis on e-commerce user reviews that have been carried out are divided into three scenarios, the explanation of the results and analysis of each scenario are as follows:

1) *Scenario 1 Results*

The results of the tests conducted in scenario 1 use general training data on 54 combinations regardless of the division of training data by category.

Based on the average graph of the test accuracy results in Fig.2, it can be seen that the use of the 15th combination with parameters alpha value of 0.001, beta value of 0.001, number of topics 15, and iteration of 1 produces an average value of the best testing accuracy of 56.5 %. According to the best parameter combination, details of the accuracy value of each k-fold of the best parameter combination with the highest average accuracy value can be seen in Fig.3.

2) *Scenario 2 Results*

Scenario 2 uses the best combination produced by scenario 1 using training data divided by category. According to the best parameter combination, the following details the accuracy value of each k-fold of the best parameter combination with the highest average accuracy value that can be seen in Fig.4.

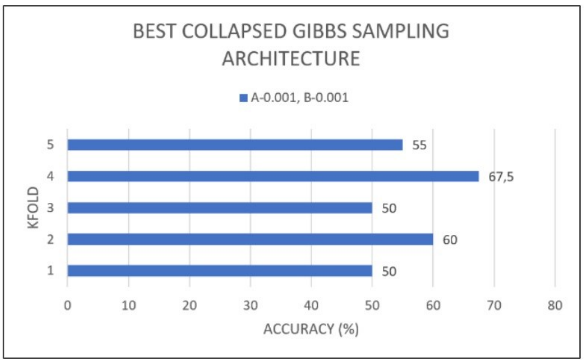


Fig. 3. Graph of Best Collapsed Gibbs Sampling Architectural

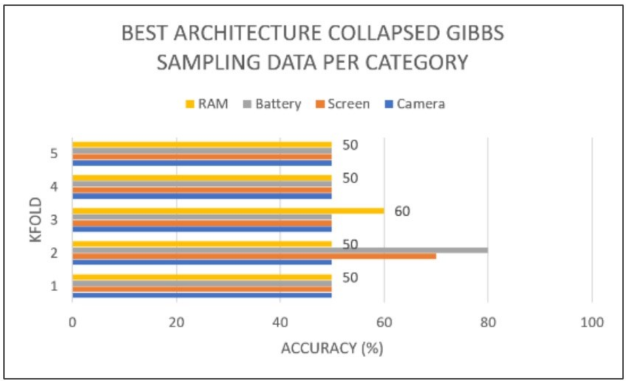


Fig.4. Graph of the Best Architecture Collapsed Gibbs Sampling Training By-Category

Based on Fig.4, it can be seen that the best combination of test results in the formation of the best classification model for camera category data has an accuracy value of 50% in all k-fold iterations. As for the screen category data, the formation of the best classification model lies in the 2nd k-fold with an accuracy value of 70%. As for the ram category data, the formation of the best classification model lies in the 3rd k-fold with an accuracy value of 60%. Then finally, in the battery category data, the formation of the best classification model lies in the 2nd k-fold with an accuracy value of 80%.

3) *Scenario 3 Results*

Scenario 3 explains the comparison of results from the use of test results from scenario 1 and the use of the best combination of classification models in scenario 2. The evaluation process carried out in scenario 3 is intended to compare whether the combination of general training data and training data per category affects the value of testing accuracy.

From the results of the evaluation, it is found that there is a difference in accuracy between the best parameter combination model and the use of general training data only and the best parameter combination model with a combination of general training data and training data per category. The comparison graph of the accuracy of testing the best architectural parameter combination for each scenario can be seen in Fig.5.

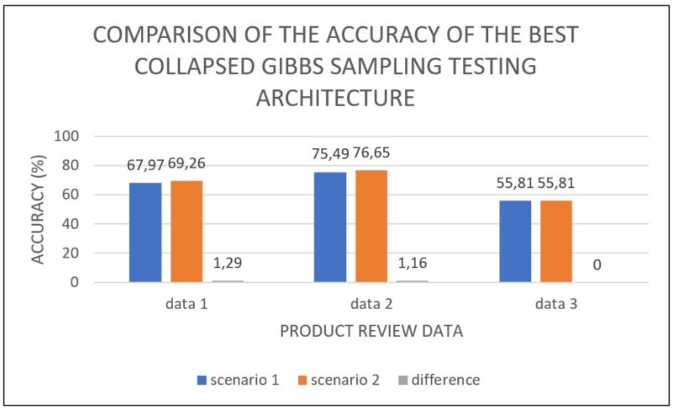


Fig.5. Graph of Comparison of Accuracy Results of Best Architecture Test Collapsed Gibbs Sampling

From the results of the comparison of the accuracy values contained in Fig.5 it is known that the use of combined training data using general training data and training data per category in the best architecture shows an increase in the accuracy of the classification model. For this reason, in the sentiment analysis study on e-commerce user reviews, the model chosen is the best architectural model produced from scenario 2 with a combination of alpha parameters 0.001, beta 0.001, number of topics 15, number of iterations 1, and combined training data between training data general and training data per category.

V. CONCLUSION

The conclusions from the reseach results of aspect based sentiment analysis in e-commerce user reviews are as follows:

The research produced the best architecture with a combination of alpha parameters 0.001, beta 0.001, number of topics 15, and number of iterations 1. The architecture, has a testing accuracy of 67.5%.

Research shows that the use of training data per category shows an increase in the value of accuracy. In this research, from the results of testing with 3 new user review data that has been labeled, the average difference in accuracy obtained increases by 0.82%.

Comparison of the performance of sentiment analysis classification in e-commerce user reviews using the LDA method shows that the architectural model with the combined use of general training data and training data per category produces the best testing accuracy.

REFERENCES

[1] APJII, Profil Pengguna Internet Indonesia 2014, Jakarta: Puskakom UI, 2014.

[2] B. Pang dan L. Lee, "Opinion Mining and Sentiment Analysis," *Foundations and Trends in Information Retrieval,* vol. 2, pp. 1-135, 2008.

[3] S. Arora, R. Ge dan A. Moitra, "Learning Topic Models - Going beyond SVD," *IEEE 53rd Annual Symposium of Computer Science (FOCS),* vol. 2, pp. 1-10, 2012.

[4] Y. Lu, Q. Mei dan C. Zhai, "Investigating Task Performance of Probabilistic Topic Models: an Empirical Study of PLSA and LDA," *Springer - Information Retrieval,* vol. 14, pp. 178-201, 2011.

[5] M. Girolami dan A. Kabán, "On an equivalence between PLSI and LDA," *SIGIR Forum (ACM Special Interest Group on Information Retrieval),* pp. 433-434, 2003.

[6] R. Kusumaningrum, W. Hong, R. Manurung dan A. Murni, "Integrated visual vocabulary in latent Dirichlet allocation – based scene classification for IKONOS image," *Journal of Applied Remote Sensing,* vol. 8, pp. 1-17, 2014.

[7] W. M. Darling, "A Theoretical and Practical Implementation Tutorial on Topic Modeling and Gibbs Sampling," *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies,* pp. 642-647, 2011.

[8] B. Liu, Sentiment Analysis and Opinion Mining, Synth. Lect. Hum. Lang. Technol: Morgan & Claypool Publishers, 2012, pp. 1-167.

[9] S. Brody dan N. Elhadad, "An Unsupervised Aspect-Sentiment Model for Online Reviews," *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL,* pp. 804-812, 2010.

[10] K. Dave, S. Lawrence dan D. M. Pennock, "Mining the peanut gallery: opinion extraction and semantic classification of product reviews," *WWW '03 Proceedings of the 12th international conference on World Wide Web,* pp. 519-528, 2003.

[11] D. M. Blei, "Probabilistic Topic Model," *Communication of The ACM,* vol. 55, no. 4, pp. 77-84, 2012.

[12] Z. Liu, High Performance Latent Dirichlet Allocation for Text Mining, Brunel University, 2013.

[13] T. L. Griffiths dan M. Steyvers, "Finding Scientific Topics," *Proceedings of the National Academy of Sciences of the United States of America,* vol. 101, no. Suppl 1, pp. 5228-5235, 2004.

[14] R. Feldman, B. Rosenfeld, R. Lazar, J. Livnat dan B. Segal, "Computerized Retrieval and Classiflcation: An application to Reasons," *Intelligent Data Analysis,* vol. 10, no. 2, pp. 183-186, 2006.

[15] R. B. Yates dan B. R. Neto, Modern Information Retrieval, Harlow: Addison Wesley, 1999.

[16] J. Han, M. Kamber dan J. Pei, Data Mining Concepts and Techniques, 3rd penyunt., San Francisco: Morgan Kaufmann Publishers, 2012.

[17] R. T.-W. Lo, B. He dan l. Ounis, "Automatically building a stopword list for an information retrieval system," *Journal on Digital Information Management: Special Issue on the 5th Dutch-Belgian Information Retrieval Workshop (DIR),* vol. 5, pp. 17-24, 2005.