Sentiment Detection of Comment Titles in Booking.com Using Probabilistic Latent Semantic Analysis

Dewi Ayu Khusnul Khotimah
Information Technology Management
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
dewi.17092@mhs.its.ac.id

Riyanarto Sarno
Departement Of Informatics
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
riyanarto@if.its.ac.id

***Abstract* - In a competitive and dynamic environment, many hotels compete to provide the best quality for customers. Hotel quality affects the brand image of a hotel marked from whether or not customers are satisfied. Booking.com website is chosen as the ideal data source as being able to utilize User Generated Content (UGC). UGC, serves as a data source ensuring the authenticity of data for customer reviews in Ponorogo, Indonesia. This experiment uses English text classification, to determine customer satisfaction and dissatisfaction based on the text they write on the title of customer testimony. The classification method used is PLSA and python programming language is used for preprocessing data. The data sampling is performed by data crawling using WebHarvy. The test results show that PLSA slightly outperforms previous research methods, namely LSA.**

Keywords: ***Hotel Quality, Booking.com Website, PLSA, Text Classification.***

I. INTRODUCTION

Many hotels find it hard to compete for commodifying the quality of their products in a competitive and dynamic environment [1]. According to customers' expectations, the quality of hotel products can be identified based on the following criteria: capable to familiarize customer to hotel employees, hotel brand image, service implementation to customer loyalty, and special offer or value which customers can directly experience. It is crucial to understand the customer's satisfaction, so they would recommend the hotel to friends and relatives, of the hotel brand image, and their experiences during the hotel room, otherwise, they will not recommend it [2].

Customers' satisfaction will be the benchmark of brand image in a hotel. There are many positive and negative reviews they write in social media and websites which can be analyzed using appropriate techniques. Customers' reviews will be well-unstructured data, making it a challenge in processing large amounts of data [3]. In the previous paper [4], several researchers combined customers from all groups of countries, regardless the customers' nationality. The previous research lack in focusing on the customers' nationality and not checking of the determinants of customer satisfaction and dissatisfaction through customer testimony on hotel product and service attributes. The attributes of hotel products and services can be seen briefly from the customers' online reviews on the title of the testimonials on the booking.com website.

Previous researchers [4] analyzed hotel services based on travel destinations, hotel types, star ratings, and editor recommendations using the LSA (*Latent Se mantic Analysis*) method. However, this paper focuses on finding positive or negative word sensitivity on customer testimonials, especially in Ponorogo, Indonesian using PLSA (*Probability Latent Se mantic Analysis*) method. In this section, PLSA creates a document context from customer testimonials, which are distinguished by words with many meanings, and are grouped in the same words [5].

Customer testimonies on booking.com website are still in the form of text. Thus, a fragment of customer testimonial information also includes text mining and classification on positive or negative text categorization [6]. Customer testimonial data will be taken by means of crawling data. WebHarvy software is used in data crawling processing, then Python software is used to classify text, including Data Preprocessing and Topic Learning. In data Preprocessing, Tokenization, Stemming, Stopwords Removal, Punctuation removal and spelling correction processes will be executed. Topic learning will process Topic Modeling, Similarity Clustering and Cluster Evaluation. Both of these processes are used simultaneously, by the PLSA method to determine the positive or negative testimonials.

Another way to determine positive or negative sentiments is to find traceability in the process of matching words using the library on SentiWordNet. SentiWordNet will classify the existing word-for-word sentence in the testimonial sentence. Sentiment calculation on words is processed using SentiWordNet [7]. This method is easy to use, but the results are non-optimum in traceability. The non-optimum sentiments analysis is due to the classification process is more complicated word compares to the use of language [8]. So we use the PLSA method to improve the accuracy of undetectable words on SentiWordNet. The word is not detected, it will be difficult to detect the meaning and value of different sentiment (*ambiguity*)

ISBN: 978-1-5386-4571-0 (c) 2018 IEEE

[9]. Ambiguity, in the level of language, has many meanings that can be solved well using the PLSA method.

In the PLSA model, processing of positive and negative documents is categorized into a particular topic or context. However, PLSA does not process word or keyword similarity as is done by the Latent Semantic Analysis (LSA) model [10]. Topics and words that have similarities in documents may use method cosine similarity [11]. Cosine similarity is used as a search for similarities among documents and processes of positive and negative words. Positive and negative words will be calculated using the probability value of the document process on positive and negative topics [4].

The researchers write this paper with the discussion as follows: In the Literature Study, the researchers review some of the literature of several previous researchers, about the sensitivity of positive and negative words. The dataset and method we propose are listed in Methodology Section. In the Experimental Research Section, the researchers show the experimental and experimental results. Conclusions and future work are described in the Conclusion Section.

II. LITERATURE STUDY

Customer behavior identification regarding satisfaction and dissatisfaction has been done by Berezina [1]. Identification of customer behavior is supported by various disciplines such as techniques, management, marketing and hospitality [12]. The concept of customer satisfaction and dissatisfaction have been studied comprehensively by marketing and consumer behavior researchers. In his study, customer behavior is a very important topic among scientists. Customer behavior will reinforce a positive attitude towards the brand image of the hotel and customers will recommend word of mouth (*Word of Mouth Communication*) [1].

Previously, Harrison-walker [13] argued that all hotels should accept customer complaints for the benefits they make by developing websites, call centers, and live chats [1]. Testimonials on the website will be matched with the score of the hotel. The incompatibility of the testimony and the score of the hotel will have an impact sensitivity analysis on positive and negative words. The solution given from the previous paper [4] uses the LSA (*Latent Semantic Analysis*) method. The result is quite good, but the PLSA method is a new development and approach for automatic document feeding [14].

Automatic indexing is based on a semantic latent class model for analyzing factors of data to be calculated. PLSA is considered to be better than LSA method as it is capable to handle specific domain synonymy and polysemy words. PLSA categorizes documents into topic words based on the frequent level of wording in document topics. [14]. Text analysis on PLSA is able to find semantic topics implied in document through word representation [15]. The word representation can be categorized into positive and negative words according to the SentiWordNet library.

PLSA is suitable to be used in this paper as it provides a model for word grouping in the title of customer testimony. The title of customer testimony is taken from several topics every document by clustering method. The clustering method uses three clusters, k-means, k-means fast, and k-mediods. In the measurement of the similarity between words, the researchers used three methods: Cosine Similarity, Jaccard Similarity and Coefficient on Correlation [5].

III. METODHOLOGY

This paper used online textual review data as the source of data. Online textual review data was obtained through crawling data technique, then analyzed using PLSA. The following is a method proposed, data collecting details and data analyzing process data will be elaborated as the following:

A. Data Collection

Online customer review data were obtained by crawling method using WebHarvy software. WebHarvy collected data from the world's largest hotel booking site in the world, booking.com [16]. The reason underlying the consideration of the booking.com as the ideal data source was because it was able to utilize User Generated Content (UGC). UGC served as a data source that guaranteed the authenticity of data for customer reviews and customers who have booked a room at booking.com. Fig. 1 showed the screenshot of comment title customer online textual review webpage on Booking.com. The title of comment can be seen from letters that print in bold and pointed with an arrow.

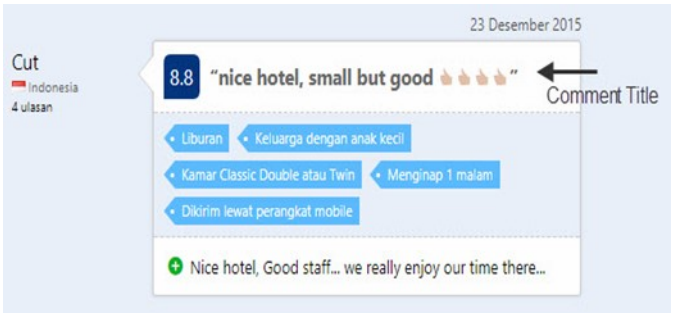

Fig. 1. Screenshot of comment title customer online textual review webpage on Booking.com

Fig. 2. The process of crawling with WebHarvy

B. Text Preprocessing

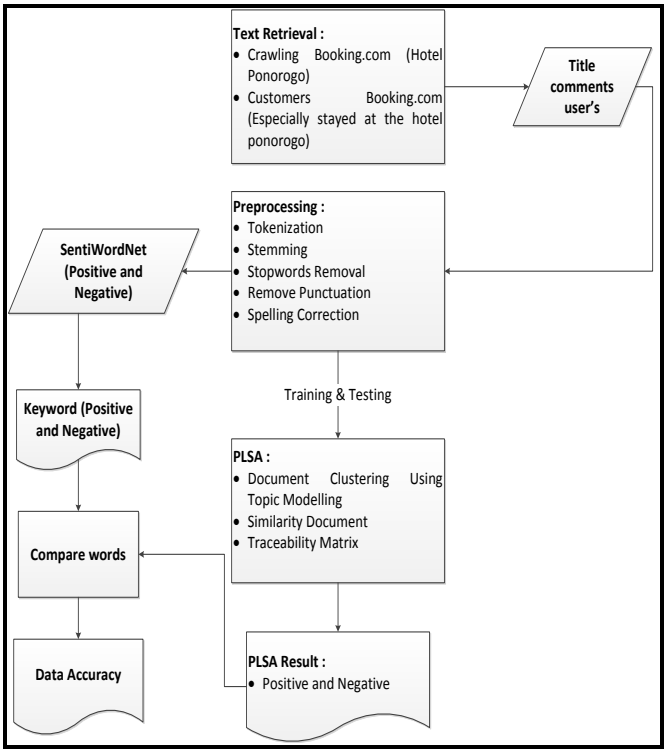

Fig. 3. Research overview diagram

Fig. 3 showed the diagram in this study. The system will retrieve user data which filled comments at booking.com website in Ponorogo city. The system used along with crawling data was WebHarvy. Crawling results was stored in excel file for processing. The crawling data contained the title of customer testimonials, customer names, customer scores in comment fields and hotel names. The titles of customer testimonies were identified in preprocessing. The programming language used was Python with the nltk (*Natural Language Toolkit*) library.

After the titles of the testimonies were taken, the classification phase of the text was represented in several steps as follows:

The first stage was tokenization. The tokenization process cut text into the smallest unit form in language processing (Word Cut). The second stage was Stemming. Stemming process defragmented a word into basic form by eliminating additives existed. The process of returning a word in this basic form used the Stemmer algorithm. The stemmer algorithm was used for English-speaking words [17]. The third stage was stopwords removal. Stopwords removal was used to remove stopwords in English in the title of customer testimonials. The fourth stage was punctuation removal. Punctuation removal was used to remove punctuation in customer testimonials. The last stage was spelling correction. Spelling correction, served to perfect the sentence in the title of customer testimony with writing errors, was described in Fig. 2.

After preprocessing was complete, the next process was using the word list from the preprocessing result, to calculate the appearance of each word in each document. The occurrence of words in each document, was calculated using PLSA.

C. Probabilistic Latent Semantic Analysis

The PLSA (Probabilistic Latent Semantic Analysis) method in this paper was used to calculate the probability of words and documents from customer testimonials. PLSA identified words into positive or negative by mapping these words in a variety of topics. Relationship between the result of preprocessing and how to determine positive or negative is described in Fig. 4.

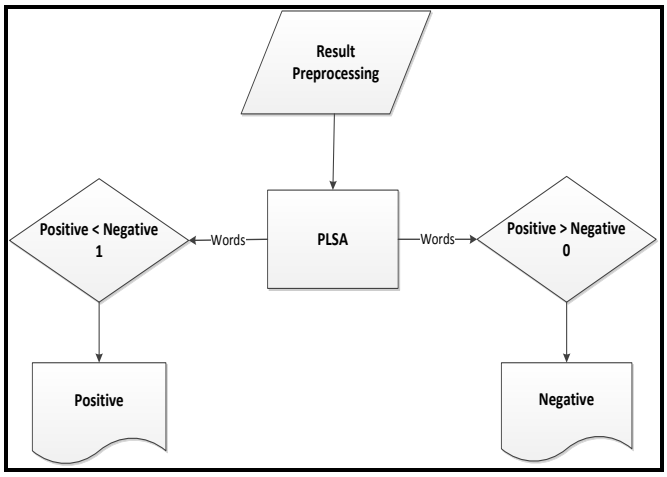

Fig. 4. Relationship among result preprocessing, and positive or negative.

PLSA was able to classify words into unknown topics. The unknown topic of every existing document had been grouped. The researchers used PLSA algorithm as follows: determining the number of topics in the title of the customer testimonials (z) then, attaching initialization parameters in probability: probability of topic P (z), probability document containing topic P (d | z) and random words in sentence contained in a topic P (w | z). The word measurement in the document was described in (1).

$$P(d_i, w_j) = \sum_{k=1}^{K} P(z_k) P(z_k) P(d_i \mid z_k) P(w_j \mid z_k) \tag{1}$$

The second step was calculating the probability of each parameter using Expectation Maximization in two steps. The first step was measuring E step. In E step the researchers measured the probability of the topic in the document, which can be seen in (2).

$$P(z_k \mid d_i, w_j) = \frac{P(w_j \mid z_k) P(z_k \mid d_i)}{\sum_{k=1}^{K} P(w_j \mid z_l) P(z_l \mid d_i)} \tag{2}$$

After measuring the E step, the second step was measuring the M step. M step was used to calculate the update values of document parameters and can be seen in (3) and (4).

The probability of a word in a topic and the probability of a word in a document was the result of a PLSA calculation.

$$P(w_j \mid z_k) = \frac{\sum_{i=1}^{N} {}^{n}(d_i \mid w_j) P(z_k \mid d_i, w_j)}{\sum_{m=1}^{M} \sum_{i=1}^{N} n(d_i \mid w_m) P(z_k \mid d_i, w_m)} \quad (3)$$

$$P(z_k \mid d_i) = \frac{\sum_{j=1}^{N} n(d_i \mid w_j) P(z_k \mid d_i, w_j)}{n(d_i)} \quad (4)$$

After measuring the PLSA, then the next step was measuring the similarity of documents. Similarity of documents was measured to know the categories in positive or negative words using Cosine Similarity. Cosine similarities was used to measure the similarity between two vector documents. The two documents were vector A and vector B. Vector A was the probability value of the positive document and vector B was the probability value of the negative document. The calculation of cosine similarities can be seen in (5).

$$\cos(\theta) = \frac{A.B}{\|A\| \mid \|B\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}} \quad (5)$$

IV. EXPERIMENTAL RESULT

This paper uses a data source from an online textual review with crawling techniques. Data were taken from booking.com website from respondents in Ponorogo city, Indonesia. Hotels located in Ponorogo city are Amaris hotel and Maesa hotel. The data are classified into the title of customer testimony using English language. The document in total is 25 with the preprocessing result that produces 42 terms. TABLE I reports the results of crawling and preprocessing data.

The next step is to calculate and report the results of the performance of PLSA in two stages. Phase 1 is the calculation in E step (*Expectation Step*) and the second step is M step (*Maximization step*). E step is used to obtain convergence and threshold values. Meanwhile, M step it is used to renew the existing values in document parameters. TABLE II reports the results of preprocessing data M step 2 and TABLE III reports the results of preprocessing data M step 3.

TABLE I. Crawling and Preprocessing Result

Document ID	Document	Preprocessing
Doc 1	“reasonable place but expensive compared to other comparable hotels in Java. Friendly helpful staff”	expensive, hotel, Friend, staff
Doc 2	"A fast breakfast can still be healthy!"	health, breakfast
Doc 3	“In overall, is good value for money”	good, value, money
Doc 4	"a suitable hotel for me"	suitable, hotel
Doc 5	“Probably the best hotel in Ponorogo”	best, hotel, ponorogo
Doc 6	“nice hotel, small but good 👍👍👍👍”	nice, hotel, small, good
Doc 7	"corridor air ventilation need to improve"	Improve, corridor, air, ventilation
Doc 8	“Price expensive and bad location”	price , expensive , bad, location
Doc 9	“Staff need some improvement”	improve, staff
Doc 10	“Not good”	bad
Doc 11	“The only choice in Ponorogo”	first, choice, ponorogo
Doc 12	“Value for money hotel.”	good, price, hotel
Doc 13	“I like the breakfast was good and the staffs were friendly.”	good, breakfast, friend, staff
Doc 14	“Staff were all very friendly.”	friend, staff
Doc 15	“the food is good for taste”	good, taste, food
Doc 16	"Unique hotel in ponorogo"	unique, hotel, ponorogo
Doc 17	“suitable for couples and solo traveller”	suitable, couple , solo, travel
Doc 18	“Very comfortable Transit stay on the way to major cities in east java”	comfort, transit ,major,city, east ,java
Doc 19	“A modern hotel”	modern, hotel
Doc 20	“A modern hotel and stylish spaces”	modern, hotel, stylish, space
Doc 21	“The english skills of the staff were good”	staff, good, skill, english
Doc 22	“Everything is good”	everything, good
Doc 23	“good location, comfort”	good, location, comfort
Doc 24	“location is good enough”	good ,location
Doc 25	"isolated hotels and villages in ponorogo"	isolate, hotel, ponorogo

TABLE II. Maximization Step (M step 2)

Term	Topic Positive	Topic Negative
expensive	0,028167085	0,028167085
hotel	0,028170972	0,028170972
friend	0,028170726	0,028174363
staff	0,028170727	0,028174367
health	0,028167085	0,028167085
breakfast	0,028167088	0,028167088
good	0,028174613	0,028174615

Term	Topic Positive	Topic Negative
value	0,028167085	0,028167085
money	0,028167085	0,028167085
suitable	0,028167086	0,028167086
best	0,028167085	0,028167085
ponorogo	0,028167333	0,028167322
nice	0,028167085	0,028167085
small	0,028167085	0,028167085
improve	0,028167085	0,028167085
corridor	0,028167085	0,028167085
air	0,028167085	0,028167085
ventilation	0,028167085	0,028167085
price	0,028167167	0,028167167
bad	0,028167085	0,028167085
location	0,028170804	0,028170804
first	0,028167167	0,028167157
choice	0,028167167	0,028167157
taste	0,028167167	0,028167167
food	0,028167167	0,028167167
unique	0,028167167	0,028167167
couple	0,028167086	0,028167086
solo	0,028167086	0,028167086
travel	0,028167086	0,028167086
comfort	0,028167167	0,028167167
transit	0,028167085	0,028167085
major	0,028167085	0,028167085
city	0,028167085	0,028167085
east	0,028167085	0,028167085
java	0,028167085	0,028167085
modern	0,028170724	0,028170724
stylish	0,028167086	0,028167086
space	0,028167086	0,028167086
skill	0,028167086	0,028167088
english	0,028167086	0,028167088
everything	0,028170722	0,028170722
isolate	0,028167167	0,028167167

TABLE III. Final Proofing (M step 3)

Document	Topic Positive	Topic Negative	Result Proofing
Doc FP 1	1,57483E-07	1,57538E-07	Topic Negative
Doc FP 2	0,00019842	0,00019843	Topic Negative
Doc FP 3	5,59079E-06	5,59032E-06	Topic Positive
Doc FP 4	0,00019844	0,00019846	Topic Negative
Doc FP 5	5,59012E-06	5,58964E-06	Topic Positive
Doc FP 6	1,57498E-07	1,57485E-07	Topic Positive
Doc FP 7	1,57421E-07	1,57434E-07	Topic Negative
Doc FP 8	1,57442E-07	1,57456E-07	Topic Negative
Doc FP 9	0,00019844	0,00019849	Topic Negative
Doc FP 10	0,00704428	0,00704489	Topic Negative
Doc FP 11	5,58938E-06	5,58891E-06	Topic Positive
Doc FP 12	5,59159E-06	5,59111E-06	Topic Positive
Doc FP 13	1,57558E-07	1,57504E-07	Topic Positive
Doc FP 14	*0,00019847	*0,00019854	*Topic Negative
Doc FP 15	5,59083E-06	5,59035E-06	Topic Positive
Doc FP 16	5,59014E-06	5,58966E-06	Topic Positive
Doc FP 17	1,57434E-07	1,57421E-07	Topic Positive
Doc FP 18	1,24906E-10	1,24896E-10	Topic Positive
Doc FP 19	*0,00019847	*0,00019849	*Topic Negative
Doc FP 20	1,57476E-07	1,57463E-07	Topic Positive
Doc FP 21	1,57517E-07	1,57483E-07	Topic Positive
Doc FP 22	*0,00019850	*0,00019851	*Topic Negative
Doc FP 23	5,59155E-06	5,59107E-06	Topic Positive
Doc FP 24	*0,00019850	*0,00019851	*Topic Negative
Doc FP 25	5,58966E-06	5,59014E-06	Topic Negative

*a mistake in identification

According to TABLE III, the researchers obtain the class actualization on true positive (TP) by 13, no mistake on false positive (FP), false negatives (FN) are 6, and True negatives are 6. The results are calculated by finding the data accuracy. The calculation of the accuracy value can be seen in (6).

$$\frac{(TP + TN)}{TP + FP + FN + TN} \tag{6}$$

The result of data accuracy is 76%. Overall, data accuracy is categorized as fair, because it is more than 0.5 (default 50%). In word classification, determining the topics and keywords taken is very crucial to the calculation of PLSA. Keywords are decided based on a combination of preprocessing data and English libraries on SentiWordNet. This combined method determines the classification of positive and negative words. Classification of positive and negative words will be better in the level of accuracy, if there is improvement in the classification of results. In this case, SentiwordNet still has a downside in word classification. The disadvantages of SentiWordNet is able to be managed by the PLSA method.

V. CONCLUSION

In this study, the positive and negative words sensitivity is predicted based on the title of testimony in booking.com website. Positive and negative words explain the level of customer satisfaction and dissatisfaction in the brand image of the hotel. The level of satisfaction and dissatisfaction are judged from the structure of the word meaning in a sentence. Customers will be satisfied or happy if in the testimony title is detected as positive. On the contrary, customers are considered as dissatisfied or disappointed if the sentence is detected as negative.

This study uses PLSA method that can work better than previous methods. In addition, the results indicate that customer satisfaction and dissatisfaction identification is 76% accurate. The results are able to analyze customer textual reviews online. Thus, this study capable to provide insight to business managers through word sensitivity of customer satisfaction and dissatisfaction of each hotel.

The researchers are further expected to be able to provide improvements by detecting more accurate dataset word sensitivity (above 76%). To improve the accuracy of the data, a better English language library than SentiWordNet is

required. Future studies may involve approaches with other methods for semantic approach. Another method used is expected to able to consider the meaning in each word better.

REFERENCES

[1] K. Berezina, A. Bilgihan, C. Cobanoglu, and F. Okumus, "Understanding Satisfied and Dissatisfied Hotel Customers: Text Mining of Online Hotel Reviews," J Hosp Mark Manag., 2016;25(1):1-24. doi:10.1080/19368623.2015.983631.

[2] M. Starkov, and J. Price. "Building a de-Commoditization strategy in hospitality," New York, NY: Hospitality eBusiness Strategies, Inc. Build a de-Commoditization Strateg Hosp New York, NY Hosp Ebus Strateg Inc. 2007. Journal of Consumer Satisfaction, Dissatisfaction and Complaining Behavior., 21, 66–77.

[3] A. Gandomi, and M. Haider. "Beyond the hype: Big data concepts, methods, and analytics," Int J Inf Manage., 2015;35(2):137-144. doi:10.1016/j.ijinfomgt.2014.10.007.

[4] X. Xu, X. Wang, Y. Li, and M. Haghighi. "Business intelligence in online customer textual reviews: Understanding consumer perceptions and influential factors," Int J Inf Manage., 2017;37(6):673-683. doi:10.1016/j.ijinfomgt.2017.06.004.

[5] F. Revindasari, R. Sarno, and A. Solichah. "Traceability Between Business Process and Software Component using Probabilistic Latent Semantic Analysis," 2016;(Icic):3-8. doi: 10.1109/IAC.2016.7905723.

[6] R-E. Fan, K-W. Chang, C-J. Hsieh, X-R. Wang, and C-J. Lin. "LIBLINEAR: A Library for Large Linear Classification," J Mach Learn Res., 2008;9:1871-1874. doi:10.1038/oby.2011.351.

[7] E.W. Pamungkas, and D.G.P. Putri. "An experimental study of lexicon-based sentiment analysis on Bahasa Indonesia," Proc - 2016 6th Int Annu Eng Semin Ina 2016. 2016:28-31. doi:10.1109/INAES.2016.7821901.

[8] I. Sunni, and D.H. Widyantoro. "Analysis sentiment and the extraction of the topic on the determinants of opinions regarding the sentiment of a public figure," J Sarge Institut Teknologi Bandung Bid Tek Elektro dan Inform. 2012;1(2):200-206.

[9] L.N. Pradany, and C. Fatichah. "Government policy analysis sentiment to content in indonesian language use twitter k-medoid svm and clustering," SCAN Journal of information technology and communication, 11(1), 59-66. Gov policy Anal Sentim to content Indones Lang use twitter k-medoid svm Clust SCAN J Inf Technol Commun 11(1), 59-66. 2016.

[10] T. Hofmann. "Unsupervised learning by probabilistic Latent Semantic Analysis. Machine Learning," 2001;42(1-2):177-196. doi:10.1023/A:1007617005950.

[11] L. Yuanchao, W. Xiaolong, X. Zhiming, and G. Yi. "A Survey of Document Clustering," A Surv Doc Clust. 2006.

[12] C.S.F Chow, and L.L Zhang. "Measuring Consumer Satisfaction and Dissatisfaction Intensities To Identify Satisfiers and Dissatisfiers," J Consum Satisf Dissatisfaction Complain Behav., 2008;21:66-79.

[13] L. Jean Harrison-Walker. "E-complaining: a content analysis of an Internet complaint forum," J Serv Mark., 2001;15(5):397-412. doi:10.1108/EUM0000000005657.

[14] J.F Pessiot, Y.M Kim, M.R Amini, and P. Gallinari. "Improving document clustering in a learned concept space," Inf Process Manag., 2010;46(2):180-192. doi:10.1016/j.ipm.2009.09.007.

[15] W. Ren, and K. Han. "Sentiment Detection of Web Users Using Probabilistic Latent Semantic Analysis," J Multimed., 2014;9(10):1194-1200. doi:10.4304/jmm.9.10.1194-1200.

[16] S. Gössling, and B. Lane. "Rural tourism and the development of Internet-based accommodation booking platforms: a study in the advantages, dangers and implications of innovation," J Sustain Tour., 2015;23(8-9):1386-1403. doi:10.1080/09669582.2014.909448.

[17] B.Y Pratama, and R. Sarno. "Personality classification based on Twitter text using Naive Bayes, KNN and SVM," Proc 2015 International Conference Data Software Engineering ICODSE 2015. 2016:170-174. doi:10.1109/ICODSE.2015.7436992.