Sentiment Analysis for Popular e-traveling Sites in Indonesia using Naive Bayes

Tata Sutabri*, Syopiansyah Jaya Putra**, Muhammad Ridwan Effendi*,
Muhamad Nur Gunawan**, Darmawan Napitupulu***
*Faculty of Computer University MH.Thamrin Jakarta
Jl. Pondok Gede Raya No.38-39 Jakarta Timur
** Department of Information System, Syarif Hidayatullah State Islamic University Jakarta
Jl. Juanda 95 Ciputat, Tangerang Selatan, 15412, Indonesia
*** Lembaga Ilmu Pengetahuan Indonesia
Jl. Jend. Gatot Subroto 10, Jakarta, 12710, Indonesia

***Abstract*- Users of online ordering and/or purchase services on the marketplace often face difficulties in determining which objects or services are selected closest to the criteria of potential users. The rating or rating feature used by many marketplaces is sometimes not objective and does not match the content of reviews given by the reviewers. This results in a decrease in the level of user confidence in the ratings and ratings provided by the service. Therefore, the prospect will seek to obtain a thorough analysis, by reading and analyzing any reviews related to a particular product or service. The burden for users is the number of reviews that are not small and the use of different language styles. This Research proposes a method that can provide a rating value that is more in line with the content of the review with respect to the sentiments present in the review. The method developed utilizes a corpus built on the topic model of reviews on the site of the hotel service provider as well as articles relating to the hotel. The sentiment analysis was obtained by using the Naïve Bayesian classification method and the use of probabilistic value of the corpus. The results of the test show the success rate of methods in analyzing sentiment is 89%. The result of sentiment analysis is used as reference of calculation of rating value.**

Keywords— Sentiment Analysis; Corpus; Naïve Bayesian; Topic Model; Hotel Review.

I. INTRODUCTION

The growth of online media encourages the emergence of unlimited textual information, resulting in the need for presentation, without prejudice to the value of the information. Textual information is categorized into two, namely facts and opinions. Fact is an objective expression of an entity, an event or a trait, whereas opinion is a subjective expression depicting people's sentiments, opinions or feelings about an entity, event, or trait. [1]. The amount of information available in the form of user reviews for different items ranging from mobile phone products, holiday travel, hotel services to movie reviews [2-4]. It is now a valuable source of knowledge to help other users, find the desired information, and make more informed decisions for the various things that are needed [3].

Tourism is one object that has a great opportunity to be developed and promoted through the website. The most of tourist sites that exist today, made it easier for the tourists to supply accommodation during the tour. Hotel is one of the tourism products that are very important to consider in terms of facilities, services or travel distance. Currently there are many tourist websites, such as trivago.com, booking.com, traveloka.com, tripadvisor.co.id, wisatakita.com, misteraladin.com, pegi-pegi.com, and others, which provide facilities for users the internet writes review of personal opinions and experiences online.

In this discussion, the tourist website that became the object of research is traveloka.com site; the reason is because traveloka is the first online site in Indonesia, since February 2012, for booking airline tickets and hotels online. In addition, the number of users of hotel services that use traveloka services. There are 18,227 hotels in Indonesia that are promoted through traveloka.com website, so this tourist website most in demand and used by domestic tourists. The problems arise that visitor to the site should read the overall reviews, so it takes a long time. In addition, it was found that the rating or score raised in the review sometimes did not match with the reviews written by the users of the hotel services.

When a user searches for a hotel in a particular tourist city, it usually looks for reviews of hotels in the city, to make a decision to book a place in one of the hotels. This review is doubtful for new users, or potential customers, because it's hard enough to read and understand all reviews in a short span of time.

In this study, looking at the limitations and accepting the challenge, it is necessary to analyze the sentiment to determine the positive / negative and the rating of the reviews of the hotel, which is done by applying the topic models (TM) approach using generative techniques in compiling or modeling the topic contained in the document [5-7]. The model topics are built to fit the traveloka travel satisfaction measurement category as cleanliness, comfort, food, location, and service. The corpus built on expert knowledge is used to analyze sentiments in review documents or online reviews of hotels using classification methods [8].

The research related to the sentiment of analysis has been done by many researcher [1, 5-23], classified the text as an alternative to organizing digital documents, so that it can simplify and accelerate the search for information needed. The method used is the Naïve Bayes method [22, 23], in which a text document is represented as a set of words (bag of words); each word in the document is assumed to be independent of each other.

The research of Ghulam Asprofi Buntoro, Teguh Bharata Adji and Adhistya Erna Purnamasari in 2014, conducted research on the level of community sentiment toward a social media problem, especially twitter, using the combination of Lexicon based and Double Propagation method which resulted in 7 (seven) sentiment analysis parameters, as very positive, positive, somewhat positive, neutral, somewhat negative, negative and very negative with 23.43% accuracy [11].

Moay El-Din, Hoda M.O Mokhtar, and Osama Ismael's research in 2015, conducted an analytical or Opinion Mining sentiment that was used for automatic detection of the subject of information such as emotion and feelings of emotional opinions. Online review on a paper can be a source of reference, where information can save time in reading paper [24].

The technique described in the research is Sentiment Analysis of Online Paper (SAOOP). Presents a comparison based on the accuracy of the performance measurement and the value of the meaning of the sentence. This SAOOP technique uses a lexical English dictionary and uses a Bag of Word (BOW) model that aims to get accuracy. The method used is Topic Domain Independence, Negation, Creation Lexicon World Knowledge Requirement, Spam and Fake Review.

In the next section will be described method Multinomial Naïve Bayesian Classifier which is the basis of the proposed method development is by utilizing the corpus that has been formed. Then proceed with the discussion of research results and concluded with conclusions.

II. RESEARCH METHOD

2.1 Analysis of sentiment by classification method

The sentiment analysis in this study is the process of classifying textual documents about online reviews or hotel reviews, which are divided into two classes, namely positive and negative sentiment class. The classification method used in this study is the Naive Bayes Classifier [22, 23]. The process begins with preprocessing [12] consisting of tokenizing, stop words filtering and stemming [25], which then carried out the Naïve Bayes classification process.

2.1.1 Sentiment Analysis with Multinomial Naïve Bayesian Classifier

The classification method used in the sentiment analysis research for this traveloka.com review is the Naïve Bayesian Classifier. Several variations of the Naïve Bayesian Classifier have been developed to calculate the probabilistic sizes for each word and scoring on each class. One is the Multinomial Naïve Bayesian model developed by Manning et al (2008) [26].

This method estimates the conditional probability of a term or token that has a given class given by the class as the relative frequency of the word t in the document belonging to the class c.

$$P(t|c) = \frac{T_{ct}}{\sum_{t' \in V} T_{ct'}} \quad (1)$$

Thus, the Multinomial Naïve Bayes model takes into account the number of occurrences of the word t in the training document of class c, including multiple events. The document training process with Multinomial Naïve Bayesian can be shown with Algorithm 1 below.

Algorithm_1. Training documents with Multinomial NBC

INPUT : Document *D*, Class *C*
OUTPUT : Vocabulary *V*, Prior Knowledge*prior*, Likelihood *condprob*

1. Extract vocabulary *V* from document *D*
2. Calculate the number of *N* documents *D*
3. For every $c \in C$
   Calculate N_c as number of *D* documents that have class c
   a. Calculate *prior [c]* = N_c / N
   b. Combine all text in document *D* that has class *c* into $text_c$
   c. For every $t \in V$
      Calculate T_{ct} as the number of tokens appearing from $text_c$ which has class *c*
   d. For every $t \in V$
      Calculate Likelihood *condprob [t] [c]* = $\frac{T_{ct}}{\sum_{t' \in V} T_{ct'}}$

Then to perform testing based on the results of training data can be used Algoritma 2 below.

Algorithm_2. Testing document with Multinomial NBC

INPUT : Class *C*, Vocabulary *V*, Prior Knowledge*prior*, Likelihood *condprob*, Test document *d*
OUTPUT : $\underset{c \in C}{arg\,max} \quad score[c]$

1. Extract token *W* from test document *d* based on Vocabulary *v*
2. For each c ∈ C
   Calculate *score [c] = log prior [c]*
   For every $t \in W$
   Calculate *score [c] + = log condprob [t][c]*
   Count $\underset{c \in C}{arg\,max} \quad score[c]$

2.1.2 Analysis of Sentiment with Multinomial Naïve Bayesian Classifier and Corpus.

To improve the performance of Naïve Bayesian Classifier (NBC) we use the corpus data that has been developed in the previous stage. The use of a K corpus aims to give more weight to value of the likelihood, for each term listed in the corpus. The corpus used is the corpus that deals with the topic of the hospitality field, namely cleanliness, comfort, food, location, and service.

The weight of the corpus *wk* is obtained from the probabilistic value of the occurrence of term t' on the existing topic. To normalize the weights, we used the proportionality of the term number for each class c, the positive class p+ = 0.65 and the negative p- = 0.35 in the train data. Thus, Likelihood *condprob* can be calculated by the following formula.

$$condprob[t][c] = \frac{T_{ct}}{\sum_{t' \in V} T_{ct'}} \times \left(1 + \left(\sum_{t' \in K} wk_{t'} \times p_c\right)\right) \quad (2)$$

While to get score on each class *score [c]* can be used formula as follows.

$$score[c] = \sum_{t' \in V} log\left(condprob[t][c] \times \left(1 + \left(\sum_{t' \in K} wk_{t'} \times p_c\right)\right)\right) \quad (3)$$

2.2. Rating calculation based on score

In this study one of the problems that must be solved in addition to classification performance in determining the sentiment of the review, which is to improve the rating provided by commentators / reviewers / customers traveloka.com. Many ratings are found that do not match the review content. To calculate the rating to match the review content, a score of [c] was obtained by combining the Naïve Bayesian Classifier and multinomial models of the corpus.

A positive rating is obtained by multiplying the positive score [c +] by the number 5, and then summing it with number 5 as the initial positive rating. For a while the negative rating without added to the number 5. Here is the formula used to find the rating.

$$rating = \begin{cases} 5 + (score[c_+] \times 5), & c = positif \\ (score[c_-] \times 5), & c = negatif \end{cases} \quad (4)$$

From the formula can be exemplified if the score obtained from the review is 0.75 with POSITIVE sentiment then the rating = 5 + (0.75 x 5) = 8.75.

III. RESULTS AND DISCUSSION

The learning process of the Naïve Bayesian classification method consists of Prior Knowledge and Likelihood. To store the value of prior knowledge can also be done by storing the value of occurrence of Nc term on each class of review (positive and negative). This is done to save storage media. Prior knowledge is the result of the occurrence of term by the number of terms throughout the data set. Thus the result is a floating number which requires more space than the integer number.

In addition to prior knowledge, Likelihood or probabilistic values are generated each term for each class (positive/negative). Examples of learning outcomes in train data with the Naïve Bayesian Classifier multinomial method can be seen in table 1.

Similar to the process of learning with the Naïve Bayesian classification method, the proposed method also produces Prior Knowledge and Likelihood. When compared with the results of the learning process with the original Naïve Bayesian method, the proposed method produces a likelihood value that has a longer relative distance between the positive and negative terms. For example, the term "holiday", in the original method has a positive likelihood value of -5.46 and negative -7.17. As for the proposed method has a positive likelihood value of -5.46 and negative -36.08. This high range is due to the use of a "holiday" corporation with a weight of 0.356. Examples of learning outcomes in the training data with the proposed method can be seen in table 2.

The proposed classification method results in better performance than the original Multinomial Naïve Bayesian method. The test results show an increased accuracy of 0.3 (converted into percentage to 3%). The original method has an accuracy of 0.86, while the proposed method has an accuracy of 0.89. Completely classification performance calculations for both the Multinomial Naïve Bayesian Classifier method and the proposed method can be seen in table_3.

TABLE 1.
EXAMPLES OF LEARNING OUTCOMES WITH NBC MULTINOMIAL METHODS

No.	Term	Nc Pos	Nc Neg	Likelihood Positive	Likelihood Negative
1	alat	37	9	-6.70895	-6.81991
2	dapur	25	4	-7.08844	-7.51305
3	ganti	40	18	-6.63297	-6.17805
4	kompor	8	1	-8.14931	-8.42935
5	tempat	217	20	-4.96204	-6.07797
6	dalam	11	1	-7.86163	-8.42935
7	lorong	4	0	-8.7371	-9.12249
8	lewat	3	0	-8.96024	-9.12249
9	tingkat	95	7	-5.78219	-7.04305
10	kebersihannya	1	0	-9.65339	-9.12249
etc.	…	…	…	…	…

TABLE 2
LEARNING RESULTS WITH PROPOSED METHODS

No.	Term	Nc Pos	Nc Neg	Likelihood Positive	Likelihood Negative
1	alat	37	9	-6.70895	-6.81991
2	dapur	25	4	-7.08844	-7.51305
3	ganti	40	18	-0.74203	-6.17805
4	kompor	8	1	-8.14931	-8.42935
5	tempat	217	20	-4.96204	-6.07797
6	dalam	11	1	-7.86163	-8.42935
7	lorong	4	0	-8.7371	-8.75832
8	lewat	3	0	-8.96024	-9.12249
9	tingkat	95	7	-5.78219	-10.5798
10	kebersihan	1	0	-9.65339	-9.12249
etc.	…	…	…	…	…

TABLE 3
CLASSIFICATION PERFORMANCE PARAMETERS

Performance Parameters	Multinomial Naïve Bayesian Classifier (MNBC)	Propose Method (MNBC + Weight Corpus)
Accuracy	0,86	0,89
Error Rate	0,14	0,11
Precision	0,92	0,99
Negative Predictive Value (NPV)	0,42	0,04
Recall	0,93	0,89
Specificity	0,39	0,50

IV. CONCLUSION.

The results of application performance testing that developed based on the method of corpus development and its utilization for the classification of traveloka.com travel sentiment using Multinomial Naïve Bayesian Classifier, has successfully answered the problem formulation and purpose of this research. A method has been developed that can be used to analyze sentiments and ratings by modifying likelihood variables in the Naïve Bayesian Multinomial classification method by multiplying the variables of corpus weights and the proportion of positive and negative exercise data. The performance of the classification method increases from an accuracy of 0.86 to 0.89.

REFERENCES

[1] Lyu, K., and Kim, H.: 'Sentiment analysis using word polarity of social media', Wireless Personal Communications, 2016, 89, (3), pp. 941-958
[2] Dumbill, E.: 'Big Data Now Current Perspective', in Editor (Ed.)^(Eds.): 'Book Big Data Now Current Perspective' (O'Reilly Media, 2012, edn.), pp.
[3] Putra, S.J., and Khalil, I.: 'Context for the intelligent search of information', in Editor (Ed.)^(Eds.): 'Book Context for the intelligent search of information' (IEEE, 2017, edn.), pp. 1-4
[4] Card, M.: 'Readings in information visualization: using vision to think' (Morgan Kaufmann, 1999. 1999)
[5] Ma, C., Wang, M., and Chen, X.: 'Topic and sentiment unification maximum entropy model for online review analysis', in Editor (Ed.)^(Eds.): 'Book Topic and sentiment unification maximum entropy model for online review analysis' (ACM, 2015, edn.), pp. 649-654
[6] O'Hare, N., Davy, M., Bermingham, A., Ferguson, P., Sheridan, P., Gurrin, C., and Smeaton, A.F.: 'Topic-dependent sentiment analysis of financial blogs', in Editor (Ed.)^(Eds.): 'Book Topic-dependent sentiment analysis of financial blogs' (ACM, 2009, edn.), pp. 9-16
[7] Titov, I., and McDonald, R.: 'Modeling online reviews with multi-grain topic models', in Editor (Ed.)^(Eds.): 'Book Modeling online reviews with multi-grain topic models' (ACM, 2008, edn.), pp. 111-120
[8] Zhang, Z., Ye, Q., Zhang, Z., and Li, Y.: 'Sentiment classification of Internet restaurant reviews written in Cantonese', Expert Systems with Applications, 2011, 38, (6), pp. 7674-7682
[9] Boiy, E., Hens, P., Deschacht, K., and Moens, M.-F.: 'Automatic Sentiment Analysis in On-line Text', in Editor (Ed.)^(Eds.): 'Book Automatic Sentiment Analysis in On-line Text' (2007, edn.), pp. 349-360
[10] Broß, J.: 'Aspect-oriented sentiment analysis of customer reviews using distant supervision techniques', Freie Universität Berlin, 2013
[11] Buntoro, G.A., Adji, T.B., and Purnamasari, A.E.: 'Sentiment Analysis Twitter dengan Kombinasi Lexicon Based dan Double Propagation', CITEE 2014, 2014, pp. 39-43
[12] Haddi, E., Liu, X., and Shi, Y.: 'The role of text pre-processing in sentiment analysis', Procedia Computer Science, 2013, 17, pp. 26-32
[13] Jijkoun, V., de Rijke, M., and Weerkamp, W.: 'Generating focused topic-specific sentiment lexicons', in Editor (Ed.)^(Eds.): 'Book Generating focused topic-specific sentiment lexicons' (Association for Computational Linguistics, 2010, edn.), pp. 585-594
[14] Jo, Y., and Oh, A.H.: 'Aspect and sentiment unification model for online review analysis', in Editor (Ed.)^(Eds.): 'Book Aspect and sentiment unification model for online review analysis' (ACM, 2011, edn.), pp. 815-824
[15] Lu, B., Ott, M., Cardie, C., and Tsou, B.K.: 'Multi-aspect sentiment analysis with topic models', in Editor (Ed.)^(Eds.): 'Book Multi-aspect sentiment analysis with topic models' (IEEE, 2011, edn.), pp. 81-88
[16] Medhat, W., Hassan, A., and Korashy, H.: 'Sentiment analysis algorithms and applications: A survey', Ain Shams Engineering Journal, 2014, 5, (4), pp. 1093-1113
[17] Pak, A., and Paroubek, P.: 'Twitter as a corpus for sentiment analysis and opinion mining', in Editor (Ed.)^(Eds.): 'Book Twitter as a corpus for sentiment analysis and opinion mining' (2010, edn.), pp.
[18] Sianipar, R., and Setiawan, E.B.: 'Pendeteksian Kekuatan Sentimen Pada Teks Tweet Berbahasa Indonesia Menggunakan Sentistrength', eProceedings of Engineering, 2015, 2, (3)
[19] Sutabri, T., and Ardiansyah, M.: 'Framework of sentiment annotation for document specification in Indonesian language base on topic modeling and machine learning', in Editor (Ed.)^(Eds.): 'Book Framework of sentiment annotation for document specification in Indonesian language base on topic modeling and machine learning' (IEEE, 2017, edn.), pp. 1-6
[20] Tang, H., Tan, S., and Cheng, X.: 'A survey on sentiment detection of reviews', Expert Systems with Applications, 2009, 36, (7), pp. 10760-10773
[21] Umamaheswari, K., and Karthiga, R.: 'Sentiment Classification based on Latent Dirichlet Allocation'
[22] Zhang, H.: 'The optimality of Naïve Bayes. American Association for Artificial Intelligence', in Editor (Ed.)^(Eds.): 'Book The optimality of Naïve Bayes. American Association for Artificial Intelligence' (2004, edn.), pp.
[23] Zulfikar, W.B., Irfan, M., Alam, C.N., and Indra, M.: 'The comparation of text mining with Naive Bayes classifier, nearest neighbor, and decision tree to detect Indonesian swear words on Twitter', in Editor (Ed.)^(Eds.): 'Book The comparation of text mining with Naive Bayes classifier, nearest neighbor, and decision tree to detect Indonesian swear words on Twitter' (IEEE, 2017, edn.), pp. 1-5

[24] El-Din, D.M., Mokhtar, H.M., and Ismael, O.: 'Online paper review analysis', International Journal of Advanced Computer Science and Applications (IJACSA), 2015, 6, (9)

[25] Agusta, L.: 'Perbandingan algoritma stemming Porter dengan algoritma Nazief & Adriani untuk stemming dokumen teks bahasa indonesia', Konferensi Nasional Sistem dan Informatika, 2009, pp. 196-201

[26] Manning, C.D., Raghavan, P., and Schutze, H.: 'Introduction to Information Retrievall Cambridge University Press, 2008', Ch, 20, pp. 405-416