Improving Naïve Bayes in Sentiment Analysis For Hotel Industry in Indonesia

Conference Paper · October 2018

DOI: 10.1109/IAC.2018.8780444

4 authors, including:

Agung Suryatno
Asia e University AeU
3 PUBLICATIONS **21** CITATIONS
SEE PROFILE

Edi Surya Negara
Universitas Bina Darma
53 PUBLICATIONS **197** CITATIONS
SEE PROFILE

Improving Naïve Bayes in Sentiment Analysis For Hotel Industry in Indonesia

Tata Sutabri
Information Technology Faculty
University of Respati Indonesia
Jakarta, Indonesia
tata.sutabri@urindo.ac.id

Agung Suryatno
Technical Information
STMIK ERESHA
Banten, Indonesia
agungsr@mitgroup.co.id

Dedi Setiadi
Faculty of Computer
MH.Thamrin Univesity
Jakarta, Indonesia
ranggalawededi@gmail.com

Edi Surya Negara
Computer Science
Universitas Bina Darma
Palembang, Indonesia
e.s.negara@binadarma.ac.id

***Abstract*— in the online ordering process, sometimes purchasing services often face problems in determining the services chosen closest to the characteristics of the user. Ratings used by some marketplace are sometimes not objective with the content of reviews provided by users. This will reduce the level of trust the user provides in the ratings provided by the service. Therefore, this study will try to produce a comprehensive analysis, by reading and analyzing any reviews related to certain services. The burden for users is the number of reviews that are not small and the use of very different language styles. This study proposes a method that can provide a rating that is more in line with the content of the review in connection with the sentiments in the review. The method developed using the corpus on the topic model on the hotel management site. Sentiment analysis was obtained using the Naïve Bayesian method and the use of probabilistic values of the corpus. The test results showed the success rate of the method in analyzing sentiment was 89%. The results of sentiment analysis are used as a standard for calculating rating.**

Keywords—analysis sentiment, corpus, naïve bayesian, topic model, hotel riview.

I. INTRODUCTION

The development of online media today, has a positive impact with the emergence of unlimited textual information, resulting in the need to represent that information, without reducing the value of information. Textual information is divided into two, namely facts and opinions. Facts are objective expressions of an entity or an event, whereas opinions are subjective expressions that express one's sentiments or opinions about an entity or event. (Y.Nur & D.Santika, 2011). The amount of information in the form of user testimonials for various items ranging from computer products, smart phones, holiday services, hotel services to movie reviews. At present the valuable source of knowledge will greatly help other users, find the information needed, and make accurate decisions for the various interests needed.

The Tourism Industry is an object that has a great opportunity to be promoted massively and developed online through a website. Most of the tourist destinations currently available make it easy for tourists to provide accommodation and comfort during the holidays. (E.Indrayuni, 2016). Hotels are a very important tourist product to pay attention to in terms of facilities, excellent service or the distance to the hotel. Currently there are many e-traveling sites, such as wisatakita.com, pegi-pegi.com, booking.com, tripadvisor.co.id, traveloka.com, trivago.com, wisatakita.com, misteraladin.com, and so on, which provides facilities for tourists to write testimonials about their opinions and personal experiences online on the site.

The e-traveling site that is the object of this research is the traveloka.com site; the reason is that traveloka is the first tourist travel site in Indonesia based online, since March 2012. In addition, the number of hotel service users or tourists who use traveloka services is very large. There are 19,272 hotels in Indonesia promoted through traveloka.com, so this e-traveling site is the best-selling and trendy, used by domestic tourists. One problem that arises is that tourists or visitors must read all the testimonials in their entirety, so that it takes a long time. In addition, it was also found that the ratings or scores given in the evaluation of testimonials were sometimes not in accordance with testimonials written by tourists or hotel service users.

When a tourist or hotel service user, searches for tourist destinations and hotels in certain tourist destinations, they will usually look for hotel testimonials online in the destination city, to make hotel booking decisions. These testimonials are sometimes doubted by tourists or new users of hotel services, because it is very difficult to read and understand all these testimonials in a short time.

What was done in this study, looked at the limitations of the hotel testimonials and analyzed sentiments, to determine the positive or negative testimonials and ranked the hotel testimonials, by applying the topic model approach, which uses generative techniques to model the topics contained in the testimonial document. The topic of the model was built to fit the satisfaction measurement categories contained in e-traveling sites such as, cleanliness, comfort, food, location and service. The Corpus is built on expert knowledge, which is used to analyze the sentiments of hotel testimonials online using the classification method.

The previous forms of research relating to sentiment analysis have been conducted by J. Samudra, S. Supeno, and M. Hariyadi, in 2009, dividing or classifying text as an alternative to processing digital documents, so as to simplify and accelerate the search for information needed. The method used is Naïve

Bayes. Text documents are represented as a set of words, and each word in the document is considered independent of each other.

Research conducted by TB. Adji, GA. Buntoro, and A.E Purnamasari in 2014, conducted a research on community sentiment analysis on social media issues, especially Twitter, using a combination of Lexicon-based and Double Propagation methods which produced 7 parameters such as very positive, positive, somewhat positive, neutral, somewhat negative, negative and very negative with an accuracy rate of 23.44%.

In addition, other studies conducted by M.El-Din, H.Muktar, and Ismail in 2015, conducted sentiment analysis to automatically detect the subject of information such as emotions and feelings. Online testimonials on paper can be a reference source, where information can save time in reading reviews.

The next explanation, Naïve Bayes method will be described which is the basis for developing the proposed method, by utilizing the corpus that has been formed, then followed by a discussion of the results of the research and concluding with conclusions.

II. LITERATURE REVIEW

The Research related to this sentiment analysis has been carried out by several researchers. The results of the study are summarized in Table 2.1. The summary table describes authors and years, research topics, and theories that have been developed from existing research.

Table 2.1: Summary of Related Research

No	Author Year	Research Topics	Theories Has Been Developed
1	Jo, Y.,and Oh, Alice, 2011.	Aspect and Sentiment Unification Model for Online Review Analysis. Dataset: electronic and restaurant	Combination of S-LDA and ASUM methods.
2.	Boiy, E.; Hens, P.; Deschacht Moens, M.F., 2007	Automatic Sentiment Analysis in On-line Text. Dataset: review film	Test sentiment analysis techniques. SVM and Naive Bayes tools / applications
3.	Rui Xie, Chunping Li, dan Qiang Ding, Li Li, 2014	Integrating Topic, Sentiment and Syntax for Modeling Online Review	Combining Part of Speech in the model. Tag Sentiment Aspect Models (TSA).
4.	Brob, J., 2013	Aspect Oriented Sentiment Analysis of Customer Reviews Using Distant Supervision Techniques	Distant supervision technique to reduce human supervision in the annotation process. 91% accuracy is correct in giving labels / annotations in the process of seeing the corpus.
5.	Changlin Ma, Meng Wang, dan Xuewen Chen, 2015	Topic and Sentiment Unification Maximum Entropy Model for online review analysis Object: restaurant	Sentiment classification with Maximum Entropy. Methodology with Topic Sentiment Unification Maximum Entropy-LDA (TSU Max-LDA).

III. METHODOLOGY

The research conducted by Indrayuni (2016), explained that the world of tourism is one of the attractions that can be promoted and developed through online sites. The increasingly widespread e-traveling sites that exist today, provide comfort and convenience to tourists to get accommodations during tourist trips. In addition, hotel facilities are one of the most important tourism products and must be considered in terms of services and facilities. Currently there are many e-traveling sites such as GoIndonesia.com, Tiket.com, valadoo.com, traveloka.com, burufly.com, bobobobo.com and others, which provide facilities for tourists to use their services, when going to do Tour.

The data sets used in this study were obtained from two different sources. The first source is a review of the popular e-traveling site, the second source from articles relating to the hospitality of some specific sites. Two types This data set, used to build corpus and sentiment analysis. Amount The reviews used to build the corpus are 14,323 reviews from 625 hotels, while related articles are used, there are 10 articles related to cleanliness, comfort, food, location, and hotel services, as an attribute or parameters.

The data acquisition process, to get hotel reviews from popular e-traveling sites, done using web scrapping automatically, using the application UBOT studio, without API (application programmable interface). After the scrapping process reviews of hotels finished, texting the hotel reviews, which consists of tokenizes process, stop word removal and stemming.

In addition, text-processing activities conducted on 10 related articles with cleanliness, comfort, food, location, and hotel services, as an attribute or its parameters. The articles are taken from the URL https://phinemo.com, www.swiss-belhotel.com, www.triadvisor.com, and https://perantara.net/.

The explanation steps can generally be seen in the general overview of the study. This Research, aimed to analyze an online hotel review, was written by the users of the hotel services. Broadly speaking the process of analysis such as presented in Figure 3.1.

Fig. 3.1. Overview of Research

3.1. Sentiment Analysis by Method of Classification

The analysis of sentiment research conducted is a process of classifying textual documents from hotel online testimonials, which are divided into two parts, namely positive and negative sentiment classes, using the Naive Bayes classification method. This classification process begins with preprocessing which consists of tokenization, stop words, filtering and stemming, which is then carried out by the Naïve Bayes classification process.

3.1.1. Sentiment Analysis by Multinomial Naïve Bayes.

The classification method used in this study is Naïve Bayes Classifier, to classify online testimonial data from leading e-traveling sites. The current Naïve Bayes Classifier method has been developed to calculate the probabilistic size of each word and provide an assessment for each class. One of them is the Multinomial Naïve Bayes model developed by Manning et al (2008). This method estimates the conditional probability of a token that has a class, as the relative frequency of the word t in the document belonging to the class c.

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

The Naïve Bayes Multinomial Method takes into account the number of occurrences of the word t in class c training documents, as well as several existing events. The process of training documents with Multinomial Naïve Bayes can be seen in Algorithm_1.

Algorithm_1. Training document by multinomial naïve bayes

Input : Document D, Class C
Output : Vocabulary V, Prior Knowledge*prior*, Likelihood *condprob*

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

The testing phase based on the results of training data can be used Algorithm_2.

Algorithm_2. Testing document by multinomial naïve bayes

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

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

3.1.2. Sentiment Analysis by Multinomial NB+Corpus.

The Naïve Bayes Classifier performance can be improved by using corpus data that has been created and developed in the previous stage. The use of corpus aims to give more weight to the parameters of the probability value, for each token listed in the corpus. The corpus used is the corpus that deals with the topic of hotel parameters, namely comfort, cleanliness, location of the hotel, food, and friendly service.

Corpus value weights are obtained from probabilistic values. The occurrence of the term t on the existing topic, the goal is to normalize the weight. In this study using the proportionality of token numbers for each class c, positive classes p + = 0.65 and negative p- = 0.35 in the data sequence. So that condprob can be calculated by a formula such as,

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

To get a score for each class [c] can use the following formula

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

3.2. Calculation of Rating with scoring

Classification performance in determining sentiment analysis from online testimonials, aims to increase the rating given by tourists or hotel service users on e-traveling sites. Many found that the scoring given was not in accordance with the contents of the testimony. To calculate the rating to fit the testimonial content, the score [c] is obtained by combining Naïve Bayes with the corpus model.

The Positive rating is obtained by multiplying the positive score [c +] by number 5, then adding it to the number 5 as the initial positive value. While the negative rating without being added to number 5. The formula used to search for ratings is as follows,

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

For example, if the score obtained from an online testimonial is 0.75 with a positive sentiment, then the rating obtained is 5 + (0.75 x 5) = 8.75.

IV. DISCUSSION RESULT

According to Tan and Zhang(2008). Data Acquisition in English, Data Acquisition abbreviated DAQ is, the sampling process of physical real-world conditions and conversions of the resulting sample into a digital numerical value that can be manipulated by computer.

In this research, the data acquisition process through several stages of scrapping, labeling, tokenizes, stop word filter, and stemming. The scrapping process aims to get text data from popular e-traveling sites. Each from the review given to the expert to make a positive sentiment label or negative.

In addition to labeling sentiments, it is also possible to label the type of data sets as training data or test data. This process is done randomly without viewing the review content. This is to maintain data independence. The labeling process does not change the structure from the content of the review. To change the review structure to be used as data set then done pre-processing. Processing stages consist of tokenizes, stop word filters, and stemming.

Tokenizes is the process of searching for words from reviews and removing caps lock as well as punctuation. The result of this tokenization process is filtered by eliminating words that are not important terms or words. To perform this filter, use stop word developed by Tala, 2004. The last stage of preprocessing is stemming.

The stemming process works to remove all affixes either which consists of prefixes, infixes, suffixes and confixes (combination of prefix and suffix) to a term or term that has been found. Stemming used to change the shape of a word into the word of the word, which is in accordance with good and correct Indonesian morphology structure. Evaluation of stemming results done manually by making observations directly to the stemming result. To assess whether stemming results are done right or wrong, used Big Indonesian Dictionary (KBBI). The data sets used consisted of 3,919 reviews as training data, and 2,314 reviews as test data. Data sets have a structure consisting of a review field, sentiment (positive and negative), type of data set (test data and test data), rating, result of tokenizes process, result of stop word filter, and result of stemming process. The result of the stemming process used for corpus development process

The Naïve Bayes method in the learning process consists of prior knowledge and possible value. The process of storing the value of previous knowledge can also be done by storing the value of occurrence of Nc tokens in each class, both positive and negative. This process is done to save storage. whereas prior knowledge is the result of the occurrence of tokens with the number of tokens throughout the data set. The result is a floating number that requires more space than an integer.

The Possible values will generate by each token for each class, both positive and negative. For example learning outcomes in training data with the Naïve Bayes multinomial method can be seen in table_1.

Table_1. Outcomes of Learning by Naïve Bayes

No.	Token	Nc Pos	Nc Neg	Likelihood Positive	Likelihood Negative
1	Alat	37	9	-6.70895	-6.81991
2	makanan	25	4	-7.08644	-7.51325
3	Ganti	40	18	-6.63397	-6.17835
4	kompor	8	1	-8.14831	-8.42935
5	lokasi	217	20	-4.96224	-6.07787
6	kamar	11	1	-7.86193	-8.42985
7	lorong	4	0	-8.73720	-9.12289
8	lewat	3	0	-8.96044	-9.12239
9	tingkat	95	7	-5.78229	-7.04365
10	kebersihannya	1	0	-9.65329	-9.12269
etc.	…	…	…	…	…

It looks the same, when the learning process using the Naïve Bayes method, the proposed method (Naive Bayes + Corpus) also produces prior knowledge and possible values. If the results of the learning process with the original Naïve Bayes method compared to the Naive Bayes + Corpus method, it will produce a possible value that has a relatively longer distance between positive and negative tokens. For example, the "holiday" token, in the original method has a positive probability value of -5.64 and negative -7.71. The proposed method has a positive probability value of -5.64 and negative -36.80. This high range value is caused by the use of "holiday" tokens weighing 0.365. The results of learning training data by the Naive Bayes + Corpus method can see in table_2.

Table_2. Result of Learning by Naïve Bayes + Corpus

No.	Token	Nc Pos	Nc Neg	Likelihood Positive	Likelihood Negative
1	Alat	37	9	-6.70895	-6.81891
2	makanan	25	4	-7.06644	-7.51325
3	ganti	40	18	-0.74403	-6.17835
4	kompor	8	1	-8.14631	-8.42955
5	lokasi	217	20	-4.98204	-6.07787
6	kamar	11	1	-7.86163	-8.42985
7	lorong	4	0	-8.78710	-8.75882
8	lewat	3	0	-8.99024	-9.12239
9	tingkat	95	7	-5.76219	-10.5768
10	kebersihan	1	0	-9.65439	-9.12269
etc.	…	…	…	…	…

The proposed Naive Bayes + Corpus method produces better performance than the original Naïve Bayes method. The process results show an increase in accuracy of 0.3 or converted to 3%. The original method has an accuracy of 0.86, while the Naive Bayes + Corpus method has an accuracy of 0.89. Performance calculations for both methods can be seen in table_3.

Table_3. Parameters Performances

Parameters Performance	**Naïve Bayes**	**Naïve Bayes + Corpus**
Accuracy	0,86	0,89
Error Rate	0,15	0,11
Precision	0,91	0,99
Negative Predictive Value	0,40	0,04
Recall	0,92	0,89
Specificity	0,38	0,51

IV. CONCLUSION

The application performance from the test results developed on the corpus development method and its utilization, for the sentiment analysis classification using Naïve Bayes, has succeeded in answering the formulation of the problem and the purpose of this study. The method developed can be used to analyze sentiment and give a rating by modifying the possible variables in the Naïve Bayes method by multiplying the body weight and the proportion of positive and negative training data. The performance of the Naive Bayes + Corpus method has increased, with an accuracy of 0.86 to 0.89.

REFERENCES

[1] A. Josi, L.A. Abdillah, Suryayusra, Penerapan teknik web scraping pada mesin pencari artikel ilmiah., Prosiding, 2

[2] Alexander Pak, Patrick Paroubek., Twitter as a Corpus for Sentiment Analysis and Opinion Mining., Journal ACM, 2010.

[3] Boiy, E.; Hens, P., Deschacht, K., Moens, M.F., Automatic Sentiment Analysis in On-line Text. Proceeding, 2007.

[4] Brob, J., Aspect Oriented Sentiment Analysis of Customer Reviews Using Distant Supervision Techniques, Dissertation, 2013.

[5] Bin Lu, Myle Ott, Claire Cardie, and Benjamin K. Tsou., Multi-aspect Sentiment Analysis with Topic Models, Journal IEEE, 2011.

[6] Card S.K., Mackinlay J.D. Shneiderman B. (eds), 2009, Reading in Information Visualization, Using Vision to Think, San Francisco: Morgan Kaufmann

[7] Changlin Ma, Meng Wang, and Xuewen Chen., Topic and Sentiment Unification Maximum Entropy Model for Online Review Analysis. Journal ACM, 2015.

[8] Dumbill, E. 2012. *Big Data Now Current Perspective.* O'Reilly Media.

[9] Eaton, C., Dirk, D., Tom, D., George, L., & Paul, Z. (n.d.). 2012. *Understanding Big Data.* Mc Graw Hill.

[10] Ghulam Asrofi Buntoro, Teguh Bharata Adji, dan Adhistya Erna P., Twitter Sentiment Analysis with Combination Lexicon Based and Double Propagation. Presiding, 2014.

[11] H. Tang, S., Tan, X. Cheng., A survey on sentiment detection of reviews, Expert Systems with Applications 36(7). Journal, 2009.

[12] Haddi, E., Liu, X., dan Shi, Y. The Role of Text Pre-processing in Sentiment Analysis. Procedia Computer Science., Proceeding, 2013.
[13] Ivan Titov, Ryan McDonald., Modeling Online Reviews with Multi-Grain Topic Models, 2008.
[14] Jo, Y., Oh, Alice, Aspect and Sentiment Unification Model for Online Review Analysis. Journal ACM, 2011.
[15] Liu, Y., Wang, G., Chen, H., Dong, H., Zhu, X., & Wang, S., An Improved Particle Swarm Optimization for Feature Selection. Journal of Bionic Engineering, 8(2), 191– 200. doi:10.1016/S1672-6529(11)60020-6. 2011.
[16] Ledy Agusta., Perbandingan Algoritma Stemming Porter dengan Algoritma Nazief & Adriani untuk Stemming Dokumen Teks Bahasa Indonesia., Proseding, 2009.
[17] Lyu Kigon, Kim Hyeoncheol., Sentiment Analysis Using Wrd Polarity of Social Media., Springer Science , Businee Media, New York, McGinty,L,Smyth,B, "Adaptive selection :analysis, 2016.
[18] Manning, C.D., Raghavan, P., and Schutze, H., Introduction to Information Retrieval. Cambridge University. 2008.
[19] Medhat, W., Hassan, A., dan Korashy, H., Sentiment analysis algorithms and applications: A survey. Ain Shams Engineering Journal., 2014.
[20] Miftah Ardiansyah, Annotated corpus-based topic models for client analysis system supporting consular dialogue. Dissertation, 2015.
[21] Neil O'Hare, Michael Davy, A. Bermingham, Paul F., Páraic Sheridan, Cathal Gurrin, Alan F. Smeaton., Topic Dependent Sentiment Analysis of Financial Blogs. Journal ACM, 2009.
[22] Rui Xie, Chunping Li, dan Qiang Ding, Li Li., Integrating Topic, Sentiment and Syntax for Modeling Online Review. Journal, 2014.
[23] Raja Mohana S.P, Umamaheswari K, and Karthiga R., Sentiment Classification based on Latent Dirichlet Allocation., Journal, 2015.
[24] Sianipar Raisa and Budi Erwin Setiawan., Strength detection sentiment on Indonesian Language Tweet Text Using Sentistrength. Journal, 2015.
[25] V. Jijkoun, M. de Rijke, and W. Weerkamp. Generating focused topic-speci_c sentiment lexicons. In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pages 585{594, Uppsala, Sweden, July 2010. Association for Computational Linguistics.
[26] Zhang, Ye, Q., Zhang, Z., & Li, Y. Sentiment classification of Internet restaurant reviews written in Cantonese. Expert Systems with Applications, 38, 7674–7682., Journal, 2011.
[27] Zhang, Harry., The Optimality of Naive Bayes., American Association for Artificial Intelligence (www.aaai.org). Journal, 2004.