Review Sentiment Analysis of World Class Hotel Using Naive Bayes Classifier And Particle Swarm Optimization Method

Sopian Aji[1], Warjiyono[2], Dany Pratmanto[3], Angga Ardiansyah[4], Andrian Eko Widodo[5], Husni Faqih[6], Suleman[7], Fandhilah[8]
{sopian.sop@bsi.ac.id[1], warjiyono.wrj@bsi.ac.id[2], dany.dto@bsi.ac.id[3]}

STMIK Nusa Mandiri Jakarta[1345678], Universitas Bina Sarana Informatika[2]

Abstract. Hotel service website is currently growing very rapidly, along with the development of tourism worldwide. The progress of world-class tourism is influenced by the provision of the international class accommodation. Selecting hotels considers generally expensive facilities, services and other supporting infrastructure. Currently the hotel booking service providers already provide facilities for tourists to write reviews and experiences of staying in the hotel room for other travel recommendations. With so many reviews displayed, it is necessary to perform a classification analysis of the review into a positive or negative grade. The method used for sentiment analysis of the hotel review is Naive Bayes Algorithm and Particle Swarm Optimization, This research examined data from sentiment hotel reviews on several hotel booking websites of 100 positive reviews and 100 negative reviews. The resulting combining Naive Bayes with Particle Swarm Optimization and obtains the best value with accuracy of 85.00%.

Keywords: sentiment analysis, hotel review, naive bayes classifier, paticle swarm optimization.

1   Introduction

Nowadays, tourism has become a very lucrative business area for each district, regions and even countries. Tourism is very helpful in growing the economy of a country. Tourism is necessary to be promoted online in order to be known by the tourists of the world. The hotel has currently an important role in facilitating the tourists to stay. Hotel should be able to provide all facilities and services, which pamper the tourists for the long stay at hotel. To be able to attract tourists to stay in the best hotel needs review of other tourist experiences. Currently there are many reviews of the hotel booking services company like www.expedia.com, www.booking.com, www.traveloka.com, www.pegi-pegi.com. Positive reviews illustrate that the hotel is nice and the negative reviews illustrate that there are still shortcomings which can be an evaluation to be better [1].

Based on the publication of The Top 100 Hotels In The World 2018 website www.travelandleisure.com, that the Four Seasons Resort Bali at Sayan, Indonesia occupies the first position of the 100 world's best hotels. This is very encouraging for the nation of Indonesia, where the 5 best Indonesia hotels come into the Top 100 Hotels In The World 2018 [2]. Four Seasons Resort Bali at Sayan, Indonesia has the highest review using English. To

help hotel management sees tourist's sentiment towards its hotel services, it is necessary for methods of analysis to summarize the positive or negative reviews.

Sentiment classification aims to classify the user review be positive or negative opinion. In sentiment analysis classification, it is divided into three levels, namely the document level, the sentence level, and the aspect level [3]. In this research, analysis of sentiment take the sentence level, using Naïve Bayes Classifier and Particle Swarm Optimization.Sentiment mainly refers to feelings, emotions, opinions or attitudes. With increasing World Wide Web, people often express their feelings through the internet via social media, blogs, ratings and reviews. Due to this increase textual data, there is a need to analyze the concept of expressing sentiments and count to explore business insight [4]. Sentiment analysis, is a process aimed at establishing whether the polarity of the textual corpus (documents, sentences, paragraphs, etc.) tends toward the positive, negative or neutral [5]. Mining opinion does not pay attention to the topic of the text but it is more focus on expression described from the text of the opinion. It determines the comments in online forums, blogs, or comments related to particular topics (products, books, movies, etc.) including a positive, negative or neutral opinion [6].

This research will identify the problems related to the number of reviews of Four Seasons Resort Bali at Sayan, Indonesia on hotel booking websites as well as what level of accuracy of sentiment analysis generated by algortma Naïve Bayes and Particle Swarm Optimization. While the benefits of this research is to help the tourists and the hotel management to the positive or negative opinion rating and can provide empirical evidence for relevant theories so that it can be a conceptual contribution to the development of the next theory.

2 Literature Review

Many ideas have emerged over the years about how to achieve quality results of the web classification system, so there are different approaches that can be used for sentiment analysis issues, such as Naïve Bayes and Bayesian Networks, NNS, DTs, Support Vector Machines (SVM) etc. Model (Naïve Bayes) is popular in machine learning applications, due to their simplicity in allowing each attribute to contribute against the same final decision, independent of other attributes [7]. Pre-Processing at its testing, Sentiment Analysis uses Naïve Bayes classifier Method with Chi Square Feature Selection [8] as follows:

1) Tokenization
   It is the task of cutting sequences of characters and a set of documents provided into chunks of words or characters that match the needs of the system. The pieces are known as tokens.
2) Stemming
   Is one of the processes of changing token having affixes to be a root word, by eliminating all existing affixes on the token. Stemming importance in the system manufacturing process is to eliminate the affixes on the prefix and suffix. Based on the results of the process, we will get an information about the many features that appear in a document.
3) Stopwords
   Can be interpreted as eliminating the characters, punctuation, and common words that have no meaning or information needed. Stopwords is commonly used in information

retrieval, one example is the Google search engine. Reducing the size of the index in the text with removal of some verbs, adjectives and other adverbs can be included in the list of stopwords.

Particle Swarm Optimization (PSO) is a population-based optimization technique developed by Eberhart and Kennedy in 1995, inspired by the social behavior of flocks of birds or fish [9]. Characteristics of particle swarm optimization is the social interaction that promotes the sharing of information between particles that would assist in the search for solutionsoptimal [9]. Particle swarm optimization has several parameters such as position, velocity, speed maximum, constant acceleration and inertia weight. In Particle swarm optimization technique there are several ways to do optimizing such as: increasing the attribute weight against all attributes or variables used, attribute selection, and feature selection. Particle swarm optimization has a better comparison of the search or even superior performance to many of the problems of optimization with faster and more stable level of convergence [9].

3   Methodology

This research uses an experimental research. Data collected will be through the selection phase of the data that is not in accordance with the research. After that the data is manually filtered to group them into positive and negative reviews. Research done on sentiment analysis using Naive Bayes algorithm and Naive Bayes based on Particle Swarm Optimization is the first position in the Top 100 Hotels In The World 2018 (www.travelandleisure.com). Research done on sentiment analysis using Naive Bayes algorithm and Naive Bayes based on Particle Swarm Optimization is the Four Seasons Resort Bali at Sayan, Indonesia by measuring accuracy of experiment data conducted in research using an application of RapidMiner Studio 8.2.001.

To perform the experiment researcher used comment review data of www.expedia.com, www.tripadvisor.com, en.tripadvisor.com.hk, booking.com, then it obtained 200 review data consisting of 100 positive reviews and 100 negative reviews. Positive review data is together in a folder with the name of Positive. While the negative review data is put together in a folder with the name of Negative. Each document has extension .txt that can be opened using Notepad application.

Then, after receiving the positive and negative data in the form of .txt with the same amount of data, namely 100 positive data reviews, the negative must be obtained 100 negative data reviews, the data will be processed using the application on RapidMiner. In RapidMiner application for text processing the researcher used tokenize, Transform Cases, stopwords (dictionar). Test is performed by choosing the feature selection based on Naive Bayes Particle Swarm Optimization with the aim to get the value of accuracy. From the accuracy value of the model equations of confusion matrix, proportion of the predicted value and ROC Curve are obtained so that we can see a ROC graph with the value of AUC (Area Under the Curve).

Figure 1, explains that the framework in this research begins by taking the hotel review of some of the hotel's website, from the review then the text processing is carried out by using tokenize, transform cases, stopwords (directory). After that, the feature selection phase is done by using Particle Swarm Optimization (PSO) and Naive Bayes. Then the accuracy of the results is obtained from processing the data.
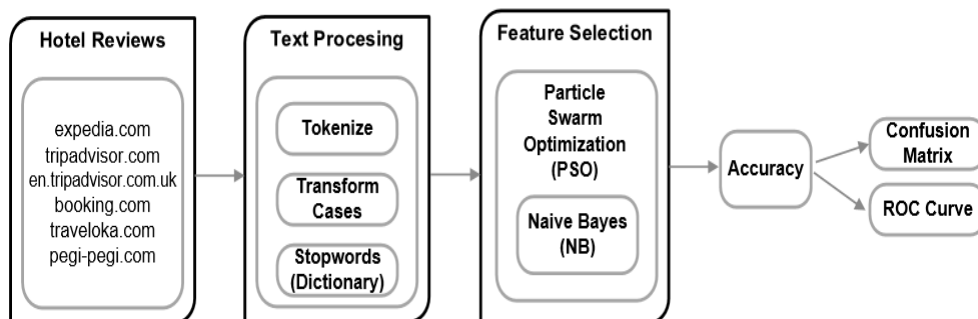
Fig. 1. Research Framework.

4   Result & Discussion

Research conducted using application of RapidMiner Studio 6.5.002 by using the computer's specifications CPU Intel Core i3 1.90GHz, RAM 10GB, and operating system Microsoft Windows 10 Professional 64-bit. Results of research use Text Processing in table 1 passed consisting of tokenize, Transform Cases, stopwords (dictionar). The processing results are as follows:

Table 1. Text Processing using Tokenize, Transform Cases, Stopwords (Dictionar)

Text Review	Tokenize	Transform Cases	Stopwords (Dictionary)
Worst four season i ever stay. The staff are rude and arrogant. The breakfast is far away for the level of 5 stars around Bali. The room is good clean, nice view, but old. The spa is nice, but too price compare with other amazing 5 stars ...this hotel is not amazing...was a decpecion...far away from my expectation. Soo sad.	Worst four season i ever stay The staff are rude and arrogant The breakfast is far away for the level of stars around Bali The room is good clean nice view but old The spa is nice but too price compare with other amazing stars this hotel is not amazing was a decpecion far away from my expectation Soo sad	worst four season i ever stay the staff are rude and arrogant the breakfast is far away for the level of stars around bali the room is good clean nice view but old the spa is nice but too price compare with other amazing stars this hotel is not amazing was a decpecion far away from my expectation soo sad	worst four season i ever stay the staff are and the breakfast is far away for the level of stars around bali the room is good clean nice view but old the spa is nice but too price compare with other amazing stars this hotel is not amazing was a decpecion far away from my expectation soo sad

In table 1 it appears that Text Processing of text reviews that are negative will turn into text reviews that are positive. In the text review through tokenize the entire punctuation will be deleted like punctuation of dots, exclamation marks and others, then Transform Cases, the entire text reviews that use uppercase letters will be changed to lowercase and the last

stopwords (Dictionar) negative vocabulary or word will be eliminated as rude, arrogant, bad and others.

The first test of research was carried out using Naive Bayes algorithm to get the accuracy, we can see in table 2 with a accuracy value of 66.67%, as follows:

Table 2. Accuracy Algoritma Naive Bayes

	true Negatif	true Positif
pred. Negatif	27	17
pred. Positif	3	13

In table 2, The results of the first test of accuracy or accuracy using a Naive Bayes algorithm with 66.67% accuracy value, then the second test is done by using a Naive Bayes algorithm based on Particle Swarm Optimization, we can see in Table 4 by obtaining the best results of 85.00% accuracy. In determining the results of research using Naive Bayes algorithm on RapidMiner Studio as follows:
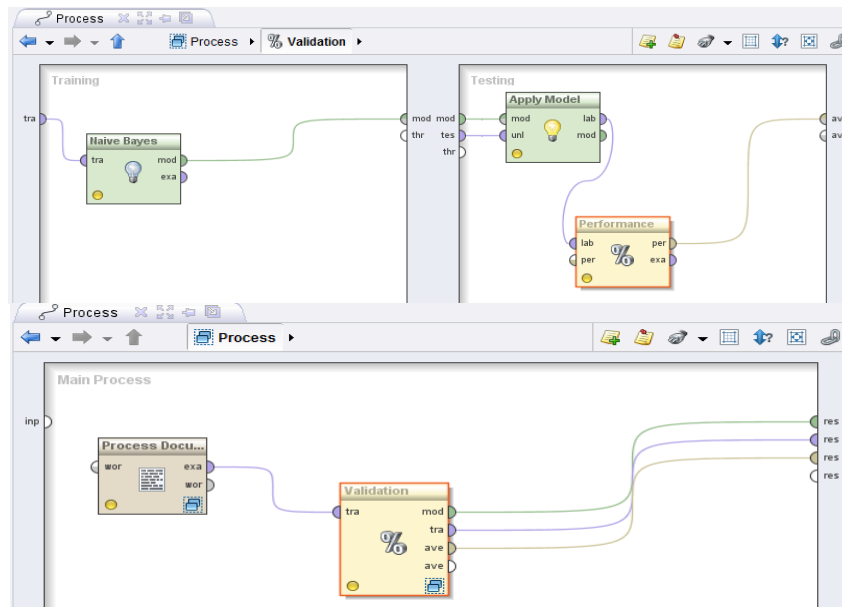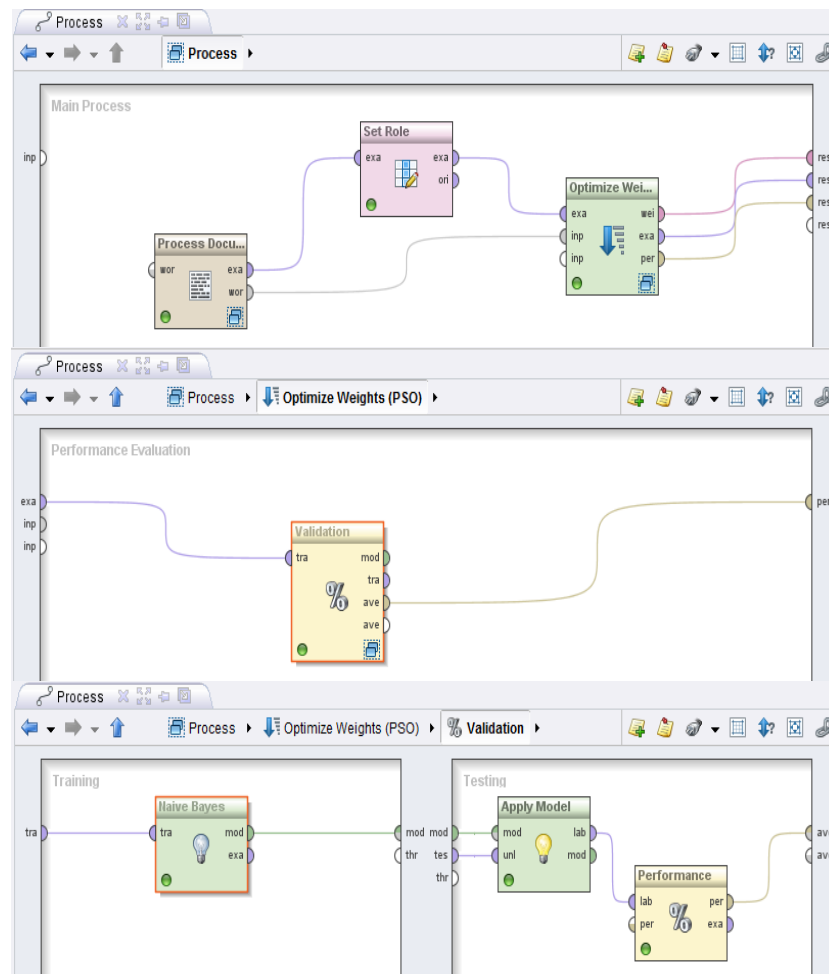

Fig. 2. Modelling Validation Testing Naive Bayes.

Table 3. Accuracy of Naive Bayes Algorithm based on Particle Swarm Optimization

	true Negatif	true Positif
pred. Negatif	28	7
pred. Positif	2	23

In table 3, The results of this second test namely Particle Swarm Optimization-based Naive Bayes algorithm based on observation accuracy gets better with a value of 85.00% accuracy. Thus the results of review of Naïve Bayes sentiment analysis can improve accuracy value after combining with Particle Swarm Optimization.

In determining the results of research using a merger between Naive Bayes with Particle Swarm Optimization in RapidMiner Studio as follows:

Fig. 3. Modelling Validation Testing Naive Bayes and Particle Swarm Optimization.

5   Conclusions

Testing of the model is by using Naive Bayes and Naive Bayes based on Particle Swarm Optimization by using review data of sentiment analysis of ethics in the social media surfing from 200 review data in Indonesian text consisting of 100 positive reviews and 100 negative reviews. The resulting model gets the accuracy result of first test done by using a Naive Bayes namely 66.67% accuracy values, then tested again with the second test by combining Naive Bayes with Particle Swarm Optimization so that it is obtained the best values with 85.00% accuracy. it can be concluded that a sentiment analysis test using a Naive Bayes will get better result if we combine Naive Bayes with Particle Swarm Optimization so that we can solve the problem of sentiment analysis by choosing best hotels to be more accurate.

References

[1] M. H. Azhar, P. P. Adikara, and Y. A. Sari, "Analisis Sentimen pada Ulasan Hotel dengan Fitur Score Representation dan Identifikasi Aspek pada Ulasan Menggunakan K-Modes," vol. 2, no. 9, pp. 2777–2782, 2018.

[2] "The Top 100 Hotels In The World 2018." [Online]. Available: https://www.travelandleisure.com/worlds-best/hotels-top-100-overall. [Accessed: 24-Jul-2018].

[3] K. Schouten, F. F.-I. T. on Knowledge, and U. 2016, "Survey on aspect-level sentiment analysis," *Ieeexplore.Ieee.Org*, vol. 28, no. 3, pp. 813–830, 2016.

[4] A. Tripathy, A. Agrawal, and S. K. Rath, "Classification of Sentimental Reviews Using Machine Learning Techniques," *Procedia Comput. Sci.*, vol. 57, pp. 821–829, 2015.

[5] E. Kontopoulos, C. Berberidis, T. Dergiades, and N. Bassiliades, "Ontology-based sentiment analysis of twitter posts," *Expert Syst. Appl.*, vol. 40, no. 10, pp. 4065–4074, 2013.

[6] I. Peñalver-Martinez *et al.*, "Feature-based opinion mining through ontologies," *Expert Syst. Appl.*, vol. 41, no. 13, pp. 5995–6008, 2014.

[7] D. Xhemali, C. J. Hinde, and R. G. Stone, "Naive Bayes vs. Decision Trees vs. Neural Networks in the Classification of Training Web Pages," *Int. J. Comput. Sci.*, vol. 4, no. 1, pp. 16–23, 2009.

[8] J. Ling, I. putu E. N. Kencana, and T. B. Oka, "Analisis Sentimen Menggunakan Metode Naïve Bayes Classifier Dengan Seleksi Fitur Chi Square," *E-Jurnal Mat.*, vol. 3, no. 3, pp. 92–99, 2014.

[9] T. S. Park, J. H. Lee, and B. Choi, "Optimization for Artificial Neural Network with Adaptive Inertial Weight of Particle Swarm Optimization," in *IEEE International Conference*, 2009, pp. 481–485.