SENTIMENT ANALYSIS OF HOTEL REVIEW USING NAÏVE BAYES ALGORITHM AND INTEGRATION OF INFORMATION GAIN AND GENETIC ALGORITHM AS FEATURE SELECTION METHODS

Dinda Ayu Muthia
Informatics Management, Academy of Informatics Management and Computer Bina Sarana Informatika
Cut Mutiah Rd No. 88, Bekasi
email: dinda.dam@bsi.ac.id

Abstract - Nowadays consumers are increasingly making their opinions and experiences online. Reading those reviews are time-consuming, but, if only few reviews were read, the evaluation would be biased. Sentiment analysis aims to solve this problem by automatically classifying user reviews into positive or negative opinions. Naïve Bayes classifier is a popular machine learning technique for text classification, because it is so simple, efficient and it has a great performance in many domains. However, it has a lack that it is highly sensitive to the high number of feature. Therefore, in this research the concatenation of feature selection methods is used, that is Information Gain and Genetic Algorithm that could increase the accuracy of Naïve Bayes classifier. This research turns out text classification in the form of positive or negative from hotel reviews. The measurement is based on the accuracy of Naïve Bayes before and after adding the feature selection method. Validation was performed using 10 fold cross validation. Whereas the measurement of accuracy was measured by using confusion matrix and ROC curve. The result of this research is the improvement of accuracy of Naïve Bayes from 78.50% to 83%.

Keywords: Sentiment analysis, Review, Hotel, Naïve Bayes, Text Classification

I. INTRODUCTION

1.1. Background

Nowadays consumers are increasingly making their opinions and experiences online. Reading those reviews are time-consuming, but, if only few reviews were read, the evaluation would be biased. Sentiment analysis aims to solve this problem by automatically classifying user reviews into positive or negative opinions [1].

There are some researches on classifying sentiment of online reviews, such as sentiment analysis on movie review comments and other multi-domain reviews including books, DVDs, electronics and many more on Amazon.com using different classifier, such as Lexicon Labeling, Heuristic Labeling, Self-labeled Instance, Self-learned Features, Oracle Labeling, Naïve Bayes, Support Vector Machine, and Maximum Entropy that have been researched by He & Zhou in 2011. Sentiment analysis on movie review from IMDB (Internet Movie Data Base), product review such as GPS, books, and cameras from Amazon.com using Support Vector Machine and Artificial Neural Network that have been researched by Moraes in 2013. Sentiment analysis on movie review opinion using Support Vector Machine and Particle Swarm Optimization which was done by Basari in 2013. Sentiment classification on restaurant review in internet which is written in Cantonese by using Naïve Bayes dan Support Vector Machine classifier which was done by Z. Zhang in 2011. Sentiment analysis on reuters and Chinese texts by using Naïve Bayes and two metrics of feature evaluation which is Multi-class Odds Ratio (MOR) and Class Discriminating Measure (CDM) which was done by Chen in 2009. Sentiment classification of online review on travel destination by using Naïve Bayes, Support Vector Machine, and Character Based N-gram Model which was done by Ye, Zhang, and Law in 2009.

Naïve Bayes classifier is very simple and efficient, [2]. Besides its simplicity, Naïve Bayes classifier is a popular machine learning technique for text classification, and has good performance in many domains [3]. However, Naïve Bayes has minus which is very sensitive in feature selection [2]. Too many amount of feature, not only to increase the time of computation but also to decrease classification accuracy [4].

The other level which generally found in approach of sentiment classification is feature selection. Feature selection can make either good classification more efficient/effective by reducing the number of analyzed data or feature identification which is suitable to be considered in learning process. There are two main kinds of feature selection methods in machine learning: wrapper and filter. Wrapper uses classification accuracy from some algorithm as its evaluation function. Filter method contains Document Frequency, Mutual Information, Information Gain,
and Chi-Square. None of such four methods which is accepted broadly as the best feature selection method for sentiment classification or text categorization, however, Information Gain mostly more excellent compared to others. Wrapper evaluates feature repeatedly and produces high classification accuracy. One of wrapper method which can be used in feature selection is Genetic Algorithm (GA).

Generally, the method of feature selection which is preferred is filter due to the processing time that is relatively short. Information Gain measures how much information of presence and absence of one word that roles to make a decision of right classification in any class. Information Gain is one of filter approach that is successful in text classification [4]. Information gain as feature selection method was used to reduce the computation complexity [1].

In this research, Naïve Bayes classifier with the combination of Information Gain and Genetic Algorithm as feature selection methods will be implemented to classify text on hotel review to increase the accuracy of sentiment analysis.

1.2. Problem Identification

Naïve Bayes classifier is simple, efficient, and is a popular machine learning techniques for text classification, and has performed well in many domains. However, Naïve Bayes has the disadvantage that is very sensitive to too many amount of feature, which resulted in the classification accuracy is low.

1.3. Problem Definition

How big is the effect of Information Gain and Genetic Algorithm as feature selection methods on the accuracy of sentiment analysis on a hotel review using Naïve Bayes classifiers?

1.4. Research Objective

The purpose of this research is to integrate Information Gain and the Genetic Algorithm as feature selection methods in analyzing the sentiment in a hotel review using Naïve Bayes classifier.

1.5. Research Benefit

1. To help the review readers in making a decision when they want to find a hotel to stay while travelling in order to reduce the time in reading the reviews and comments of a hotel.
2. To help system developers with regard to hotel reviews, both from source TripAdvisor.com or from social media such as Twitter, Blog, and others.

1.6. Research Contribution

Classifying text of sentiment analysis on a hotel review by using Naïve Bayes classifiers and the integration of Information Gain and Genetic Algorithm as feature selection methods.

II. THEORY

2.1. Review

According to Reddy [5] in many cases the decisions we make are influenced by the opinions of others. Prior awareness of the Internet became widespread, many of us who usually ask a friend or neighbor’s opinion regarding electronic devices or movie before actually buying it. With the growing availability and popularity of opinion would be a rich source such as online review websites and personal blogs, new opportunities and challenges arise since people are now able to actively and use information and technology to find and understand the opinions of others.

According to Yessenov [6] there are some examples of websites that can review products, such as Amazon, or sites such as Rotten Tomatoes movie reviews are allowed to rate the product, usually within the same scale determined by personal reviews made.

2.2. Sentiment Analysis

According to Feldman [7] sentiment analysis (Opinion mining) is defined as the task of finding opinion of the author of a particular entity. According to Tang in Haddi [8], sentiment analysis is the process of investigating on review product reviews on the internet to determine the opinion or feelings of a product as a whole.

According to Thelwall in Haddi, sentiment analysis is treated as a classification task that classifies the orientation of a text into positive or negative. According to Mejova in Basari [9], the purpose of the analysis is to determine the behavior sentiment or opinion of an author with attention to a particular topic. Behavior may indicate the reason, opinion or judgment, conditions inclination (how the author would like to influence the reader).

2.3. Feature Selection

According to Gorunescu [10] feature selection is used to eliminate irrelevant features and repetitive, which may lead to chaos, using certain methods. According to John, Kohavi, and Pfleger in Chen [2] there are two types of feature selection methods in machine learning, namely the wrapper and filter.

According to Chen [2] wrapper using the classification accuracy of several algorithms as a function evaluation. According to Gunal [11] one of the wrapper method that can be used in the selection of features is a Genetic Algorithm (GA).

1. Genetic Algorithm

According to Han [12] Genetic algorithm seeks to combine the ideas of natural evolution. In general, genetic learning starts as follows:

a. An initial population is created consisting of a random rule. Each rule can be represented by a string of bits. As a simple example, suppose that a sample in a given training set is
described by two Boolean attributes, A1 and A2, and that there are two classes, C1 and C2. Rule “If A1 And A2 Not Then C2” can be encoded as a string of bits "001," where the two leftmost bits represent attributes A1 and A2, respectively, and the rightmost bit represents the class. Similarly, the rule of "If Not A1 And A2 Not Then C1" can be encoded as "001". If the attribute has k values, where k > 2, then k bits may be used to encode the attribute values. Classes can be encoded in the same way.

b. Based on the idea of the most appropriate resistance, a new population is formed consisting of the rules most suitable in the current population, as well as the descendants of this rule. Typically, the rule fitness was assessed by the classification accuracy on a set of training samples.

c. Offspring are created by applying genetic operators such as crossover and mutation. In crossover, substrings from a pair of rules exchanged to form new pairs of rules. In mutation, randomly selected bits in a rule string is reversed.

d. The process of generating new populations based on prior populations of rules continues until a population, P, developed in which each rule in P satisfy the fitness threshold that has been determined.

Genetic Algorithm is easy to be aligned and have been used for classification as other optimization problems. In data mining, Genetic Algorithm can be used to evaluate the fitness of other algorithms.

According to Chen [2] filter method consists of Information Gain, Term Frequency, Chi-Square, Expected Cross Entropy, Odds ratio, the weight of evidence of the text, Mutual information, and the Gini index. According to Moraes [13], there is also a method of Document Frequency. None of the four methods that are widely accepted as the best method of feature selectors for sentiment classification or text categorization, however, Information Gain is often superior to the others.

2. Information Gain

Stages in the process of calculating Information Gain as follows:

a. Find the value of entropy before splitting with the following formula:

$$Entropi(y) = - \sum_{i} P_i \log_2 P_i$$  \hspace{1cm} (1)

P_i : data proportion of y on the class i

b. Find the value of entropy after separation by attribute A with the following formula:

$$\text{Total entropi} = \sum_{c \in nitial(A)} \frac{Yc}{Y} \cdot \text{entropi}(y_c)$$  \hspace{1cm} (2)

c. Find information gain value with the following formula:

$$\text{gain}(y, A) = \text{entropi}(y) - \sum_{c \in nitial(A)} \frac{Yc}{Y} \cdot \text{entropi}(y_c)$$  \hspace{1cm} (3)

2.4. Naive Bayes Algorithm

According to Markov [14] Naive Bayes algorithm stages in:

1. Calculate the conditional probability / likelihood:

$$P(x \mid C) = P(x_1, x_2, ..., x_n \mid C)$$  \hspace{1cm} (4)

C = class

x = vektor from attribute value of n

$$P(x|C) = \text{document proportion of class C contains attribute value of } x_i$$

2. Calculate the prior probability for each class:

$$P(C) = \frac{N_j}{N}$$  \hspace{1cm} (5)

N_j = number of documents in a class

N = total number of documents

3. Calculate the posterior probability with the formula:

$$P(C \mid x) = \frac{P(x \mid C) \cdot P(C)}{P(x)}$$  \hspace{1cm} (6)

Bayes formula can be given as follows:

$$\text{Posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{evidence}}$$  \hspace{1cm} (7)

2.5. Framework

This study starts from a problem in the text classification using Naive Bayes hotel reviews, in which the classifier has the disadvantage that it is very sensitive to too many amount of feature, which resulted in the classification accuracy is low. The dataset used in this research is a review of Hotel Royal at Queens in Singapore housed derived from hotel reviews on the particular site www.tripadvisor.com consisting of 100 positive reviews and 100 negative reviews. Preprocessing is performed with tokenization, stopword removal, stemming, and generate N-grams. Feature selection method used is information gain and genetic algorithm, whereas the classifier used is Naive Bayes. 10 fold cross validation testing will be performed, the accuracy of the algorithm will be measured using the confusion matrix and the processed data in the form of ROC curves. RapidMiner version 5.3 is used as a tool to measure the accuracy of experimental data. Figure 1 illustrates the framework that the authors have proposed in this research.
III. RESULTS AND DISCUSSION

3.1. Result

3.1.1. Text Classification Using Naïve Bayes Algorithm

Training data used in text classification consists of 100 hotel reviews positive and 100 negative hotel reviews. The data is still a bunch of separate text in the form of documents. Before classified, the data must go through several stages of the process to be classified in the next process, the following are the stages of the process:

1. Collecting Data
   Data positive reviews together in a folder with the name of the positive. While the data is stored in the negative reviews put together a folder with the name of negative. Txt extension for each document that can be opened using Notepad application.

2. Preprocessing
   The process through which consists of tokenization, stopword removal, stemming, and generate N-grams. Initial data processing result can be seen in Table 2.

Table 2. Initial Processing of Data

<table>
<thead>
<tr>
<th>Review</th>
<th>Tokenization</th>
<th>Stopwords removal</th>
<th>Stemming</th>
<th>Generate N-grams</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wou...</td>
<td>Would...</td>
<td>would...</td>
<td>recommend...</td>
<td>recommend...</td>
</tr>
<tr>
<td>d not...</td>
<td>not...</td>
<td>not...</td>
<td>especially...</td>
<td>especially...</td>
</tr>
<tr>
<td>recom...</td>
<td>recommend...</td>
<td>recommend...</td>
<td>think...</td>
<td>think...</td>
</tr>
<tr>
<td>mend...</td>
<td>especially...</td>
<td>especially...</td>
<td>think...</td>
<td>think...</td>
</tr>
<tr>
<td>end...</td>
<td>this...</td>
<td>this...</td>
<td>hotel...</td>
<td>hotel...</td>
</tr>
<tr>
<td>this...</td>
<td>if...</td>
<td>if...</td>
<td>is...</td>
<td>is...</td>
</tr>
<tr>
<td>is...</td>
<td>you...</td>
<td>you...</td>
<td>a...</td>
<td>a...</td>
</tr>
<tr>
<td>this...</td>
<td>this...</td>
<td>this...</td>
<td>is...</td>
<td>is...</td>
</tr>
<tr>
<td>is...</td>
<td>star...</td>
<td>star...</td>
<td>hotel...</td>
<td>hotel...</td>
</tr>
<tr>
<td>a...</td>
<td>4...</td>
<td>4...</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

3. Classification

Classification process is to determine a sentence as a member of a class of positive or negative class probability calculation based on the value of the larger Bayes formula. If the results of the sentence probability for the positive class is larger than the negative class, then the sentence belongs to the positive class. If the probability for the positive class is smaller than the negative class, then the sentence belongs to the negative class. The author get 6 words appears most frequently, such as Clean, Comfort, Quiet, Shop, Travel, and View that related to hotel review.

3.1.2. Optimization Model with Combination of Feature Selection Methods

By combining filter and wrapper as feature selection methods, where in this study the method used is the Information Gain of the filter and Genetic Algorithm of the wrapper. The data will be processed given the weight of the Information Gain to improve Naïve Bayes classifier accuracy.

3.1.3. Experiments on Indicator Model

To get a good model, some indicator adjusted value to obtain high accuracy results. In adjustment indicator on Genetic Algorithm, the highest accuracy is obtained with the combination of population size = 50, p initialize = 0.8, p = 0.8 crossovers, and generate p = 1.0. The results achieve 83% accuracy. If other indicators also changed its value, may lead to the data processing becomes increasingly longer.

3.2. Discussion

By having text classification models on the review, the reader can easily identify where the review is positive or negative. From a review of data that already exists, is separated into words, then given a score on each of these words. Can be seen any word related to the sentiment that frequently arise and have the highest weights. Thus it can be seen that a positive or negative review.

In this study, the results of testing of the model will be discussed through the confusion matrix to show how well the model is formed. Without the use of feature selection methods, Naïve Bayes algorithm itself has resulted in an accuracy of 78.50%. Accuracy is still less accurate, so it needs to be improved using feature selection methods. After using the feature selection method of the filter and wrapper are combined, Naïve Bayes algorithm accuracy increases to 83% as can be seen in Table 3.

Table 3. Naïve Bayes algorithm model before and after using the feature selection method

<table>
<thead>
<tr>
<th>Naive Bayes Algorithm</th>
<th>Naive Bayes Algorithm +</th>
</tr>
</thead>
</table>

Algorithm | Information Gain & Genetic Algorithm
--- | ---
Successful classification of positive reviews | 72 | 86
Successful classification of negative reviews | 85 | 80
Model Accuracy | 78.50% | 83%

3.2.1. Measurements by Using Confusion Matrix

Measurement of the confusion matrix here will show a comparison of the results of the accuracy of Naïve Bayes models before being added feature selection methods which can be seen in Table 4 and the following added feature selection methods, namely the incorporation of Information Gain and Genetic Algorithm can be seen in Table 5.

<table>
<thead>
<tr>
<th>True negative</th>
<th>True positive</th>
<th>Class precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pred. negative</td>
<td>85</td>
<td>28</td>
</tr>
<tr>
<td>Pred. positif</td>
<td>15</td>
<td>72</td>
</tr>
<tr>
<td>Class recall</td>
<td>85.00%</td>
<td>72.00%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>True negative</th>
<th>True positive</th>
<th>Class precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pred. negative</td>
<td>80</td>
<td>14</td>
</tr>
<tr>
<td>Pred. positif</td>
<td>20</td>
<td>86</td>
</tr>
<tr>
<td>Class recall</td>
<td>80.00%</td>
<td>86.00%</td>
</tr>
</tbody>
</table>

Here is a view of a ROC curve test results data. Figure 3 is the ROC curve for Naïve Bayes models before using the feature selection methods and Figure 4 is the ROC curve for Naïve Bayes models after using the feature selection methods.

IV. CONCLUSION

From data processing which is done, the combination of feature selection method which are filter and wrapper, it is proven it can increase classification accuracy of Naïve Bayes. Data review of hotel can be classified well into the form of positive and negative. The accuracy of Naïve Bayes before using the combination of feature selection methods reaches 78.50% While after using the combination of feature selection methods, which is Information Gain and Genetic Algorithm, accuracy increased up to 83%. The accuracy improvement up to 4.5%.

The model that has been built could be implemented to all of hotel reviews, so that we could see the result directly in the form of positive and negative. This could help people to minimize the time when they are searching for hotel reviews while planning to go travelling and to avoid them from inappropriate rating given on some reviews.

REFERENCES


**Dinda Ayu Muthia**, is currently a lecturer of the Study Program of Informatics Management, AMIK BSI. She received a Master Degree in Computer Science from STMIK Nusa Mandiri in 2013 on “Information Retrieval”. Dinda Ayu Muthia, M. Kom research interests are in Information Retrieval and Sentiment Analysis.