Sentiment classification of hotel service review on traveloka sites using naïve bayes classifier (NBC) and binary logistic regression

Silvia Astri Rahmaningrum¹, Pratnya Paramitha Oktaviana²

¹ Department of Statistics, Institut Teknologi Sepuluh Nopember, Surabaya City, Indonesia
² Department of Actuarial Science, Institut Teknologi Sepuluh Nopember, Surabaya City, Indonesia

astriviaa@gmail.com

Abstract. Utilization of social media sites such as Traveloka can help hotel marketing by providing information as review related to search and hotel bookings online. The given review could be a feedback to the related hotel as well as to assist visitors in choosing the right hotel. Feedback information as a review is important data text, so we need to develop a method to classify it then the right method used for classifying text data is text mining. Sources of data are obtained from web scraping process which aims to obtain online data on website page by collecting visitor review’s of Favehotel and Gunawangsa Hotel from Traveloka sites. Text mining methods used in this study is Naïve Bayes Classifier (NBC) and Binary Logistic Regression which is the process of sentiment labeling based on Lexicon dictionary. Word cloud visualization shows that the highest keywords that lead to the both of the hotel with the largest positive sentiment are ‘clean’ and ‘comfortable’. A comparison methods between NBC and Binary Logistic Regression for Favehotel and Gunawangsa Hotel obtained a decision that Binary Logistic Regression method with SMOTE was better than NBC where AUC value in testing data for Favehotel is 0.84 and Hotel Gunawangsa is 0.82.

1. Introduction

The development of social media on the website today makes someone will using media for decision making. It is undeniable that with such rapid development it has a big impact on life include easy access to information 24 hours. Many convenience and benefit obtained by the internet include service companies related to tourism such as hotel. As the development of online media, the industrial company include the hotel industry is using this media for promotion and marketing. One site that provide online tourism service is Traveloka sites. Traveloka feature’s enable hotel visitors can write a review as feedback when they have stayed at a hotel booking of Traveloka application. Reviews that have been written by visitors can be useful for the hotel regarding the services that have been provided to determine customer satisfaction. Gunawangsa MERR Hotel and Favehotel Rungkut are three star hotel that uses the Traveloka application for marketing. The review from these visitors are important text data that can be used as evaluation material for hotel performances.
Based on the review given, a classification method is needed to classify whether the review tends to be positive or negative. One right method used for classifying text data is text mining. Text mining aims to find words that can represent the contents of the documents so that it can be analyzed the relationship between documents. There are several text classification methods used including Support Vector Machine (SVM), Naïve Bayes Classifier (NBC), K-Nearest Neighbor (KNN), Classification and Regression Trees (CART). Based on several methods available, in this study used NBC method. NBC method is a classification based on the Bayes theorem which aims to calculate the opportunities in each class and it has the advantage of requiring only a small amount of training data to estimate the parameters needed for classification and has high accuracy [1]. While the classic method used to compare the classification accuracy with NBC is Binary Logistic Regression. In this research, web scraping will be implemented to collect visitor review data on Gunawangsa MERR Hotel and Favehotel Rungkut based text mining on Traveloka sites and classify them into positive and negative sentiments based on the lexicon dictionary using NBC and Binary Logistic Regression. After the classification results are obtained, then we compare the classification performance between two methods and visualize with Word Cloud. The variables used in this study consisted of independent variable which is the basics words of the review that have been carried out preprocessing data and dependent variable which is the reviews classification into positive and negative. Through this research, it is expected to provide advice to the Gunawangsa MERR Hotel and Favehotel Rungkut in terms of evaluating services that must be improved and maintained to optimize customer satisfaction.

2. Literature Review

Web Scraping

Web scraping is the process of retrieving a semi structured data from internet, generally in the form of website pages in markup, languages such as HTML or XHTML and analyzing these documents to extract certain data from these pages for use for other purposes [2].

Text Preprocessing

The text that will be the process of text mining in general has not been structured well. Therefore, before determining the features that represent the required preprocessing stage which is done in general in text mining on documents with the following stages such as case folding, cleansing, stemming, stopwords, tokenizing [3].

TF-IDF

Term Frequency – Inverse Document Frequency (TF-IDF) is a feature extraction process by assigning values to each word in the training data. To find out how important a word represents a sentence, weighting or calculation will be done. The scoring in TF-IDF is based on the frequency of words appearing in the document [4].

\[
w_{ji} = tf_{ji} \times idf
\]

\[\text{idf} = \log \left( \frac{n}{df_j} \right)\]  

where,

- \(w_{ji}\): weight of the word \(j\) in the review to \(i\)
- \(tf_{ji}\): number of occurrences of the word \(j\) in the review to \(i\)
- \(n\): total number of reviews
- \(df_j\): the number of review that contain the word \(j\)

K-Fold Cross Validation

Cross validation is one technique to validate the accuracy of a model that is built based on a specific dataset. The model created will aim to predict or classify new data that is not yet in the dataset. The data used in the model development process is called training data, while the data that will be used to validate the models is referred to as testing data. K-Fold Cross Validation is one of the methods of
cross validation used to partition data into training and testing data, where each data has the opportunity to become testing data [5].

**Naïve Bayes Classifier (NBC)**
Naïve Bayes Classifier (NBC) is a simplified model of the Bayes algorithm that is suitable for classifying text or documents. This method is used to find the highest probability value to classify test data in the most appropriate category [6]. The advantage of NBC is that it has relatively high accuracy. The basic concept used in NBC is the Bayes theorem which refers to the conditional probability concept as in equation (3).

\[
P(Y|V) = \frac{P(V|Y)P(Y)}{P(V)}
\]  

(3)

Through the Bayes theorem rule in equation (3), if the review is considered as \( v_i \) then it is assumed to have a collection of reviews with \( V = \{v_i | i = 1, 2, ..., |V|\} = \{v_1, v_2, ..., v_{|V|}\} \) and collection of sentiment categories with \( Y = \{y_j | k = 1, 2\} = \{y_1, y_2\} \) [7]. NBC classification is done by finding the probability of \( P(Y=y_k|V=v_i) \) is the probability of category \( y_k \) if known that \( v_i \) where \( v_i \) seen as words in the review \( x_1, x_2, ..., x_m \) where \( x_1 \) is the first word, \( x_2 \) is the second word and so on. So the NBC formula for classification is presented in equation (4).

\[
V_{MAP} = \arg \max_{y \in Y} \prod_{j=1}^{m} P(x_j | y_k)
\]  

(4)

the value of \( P(y_k) \) was calculated at the time of training, obtained by the equation formula (5).

\[
P(y_k) = \frac{|docs k|}{|contoh|}
\]  

(5)

with, \( |docs k| \) is the number of reviews that have \( k \) category in the training. Whereas \( |contoh| \) is total number of sample reviews used in the training process. Each probability of the word \( x_j \) for each category is calculated at the time of training in equation (6).

\[
P(x_j | y_k) = \frac{m_j + 1}{m + |kosakata|}
\]  

(6)

where \( m_j \) is the number of occurrences of the word \( x_j \) in the categorized review \( y_k \) while \( m \) is the total number of words with the \( y_k \) category and \( |kosakata| \) is the number of words in the training data.

**Binary Logistic Regression**
Logistic regression is a data analysis method to find the relationship between response variables \( y \) which are dichotomous (two categories) or polycotomus (more than two categories) with one or more predictor variables \( x \) on a scale or continuous scale [8]. Suppose a set of \( j \) predictor variables is shown as \( x = (x_1, x_2, ..., x_m) \). Logit form of the logistic regression as in equation (7).

\[
g(x) = \ln \left( \frac{\pi(x)}{1 - \pi(x)} \right) = \beta_0 + \beta_1 x_1 + ... + \beta_n x_n
\]  

(7)

The logistic regression model with \( j \) is the number of predictor variables written in equation (8).

\[
\pi(x) = \frac{e^{g(x)}}{1 + e^{g(x)}}
\]  

(8)

One important assumption that must be fulfilled in logistic regression analysis is the absence of multicollinearity. One way to detect multicollinearity is Variance Inflation Factor (VIF). If the VIF > 10 indicates the presence of multicollinearity in the data.

2.1.1. Parameter Estimation
Parameter estimation can be done using the Maximum Likelihood Estimation (MLE) method. The probability function follows the Bernoulli distribution where each observation \( (x_i, y_i) \) is shown in equation (9).

\[
f(y_i) = \pi(x_i)^{y_i}[(1 - \pi(x_i))]^{1-y_i}
\]  

(9)
If the observation are assumed to be independent then the likelihood function of independent observations is shown in equation (10).

\[
 l(\beta) = \prod_{i=1}^{n} f(y_i) = \prod_{i=1}^{n} \pi(x_i)^{y_i} (1-\pi(x_i))^{1-y_i} \tag{10}
\]

\( \beta \) is a vector that contains \( \beta \). The likelihood (\( L(\beta) \)) function then converted into equation \( ln \) written in equation (11).

\[
 L(\beta) = \ln \left[ l(\beta) \right] = \sum_{i=1}^{n} \left( y_i \ln[\pi(x_i)] + (1-y_i)\ln[1-\pi(x_i)] \right) \tag{11}
\]

To find the value of \( \beta \) that maximizes \( L(\beta) \) next \( L(\beta) \) lowered towards \( \beta_0 \) and \( \beta_1 \) then equation zero.

**Classification Performance**

Measurement of classification accuracy is done to evaluate how much the ability of a method in classifying documents into appropriate classes. Accuracy testing in this test is using confusion matrix contained in Table 1.

**Table 1.** Confusion Matrix

| Actual Class | Prediction Class |
|--------------|------------------|
|              | Positive | Negative |
| Positive     | TP       | FN       |
| Negative     | FP       | TN       |

Measurements that can be used to calculate classification performance include accuracy, precision and recall [9] shown in the following equation (12), (13), (14).

\[
 \text{accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{12}
\]

\[
 \text{precision} = \frac{TP}{TP + FP} \tag{13}
\]

\[
 \text{recall} = \frac{TP}{TP + FN} \tag{14}
\]

In addition the three measures, classification accuracy can be calculated using Area Under Curve (AUC) as equation (15)

\[
 AUC = \frac{1}{2} \left( \frac{TP}{TP + FN} + \frac{TN}{TN + FP} \right) \tag{15}
\]

**3. Research Methodology**

*Research Data Sources and Variable*

The data used in this research from hotel review of the Hotel Gunawangsa MERR and Favehotel Rungkut Surabaya contained on the Traveloka website. The review was recorded from May 2016 to April 2019. The review was carried out using web scraping method.

Based on the source of data obtained in the review then the data must be filtered by taking data that contains sentiment. The research variables used are in Table 2.

**Table 2.** Research Variables

| Variable | Information | Data Scale |
|----------|-------------|------------|
| \( y \)  | Sentiment (Positive/Negative) | Nominal     |
|          | 0 = Negative Sentiment |             |
|          | 1 = Positive Sentiment  |             |
| \( x \)  | Keyword weight obtained from TF-IDF results | Ratio       |
4. Analysis and Discussion

Data Characteristics

Traveloka review’s provided for users in conveying criticism and suggestions for related hotels where in this study is Gunawangsa MERR Hotel and Favehotel Rungkut. Based on data collected then selected and reviews containing sentiments will be further analyzed. Comparison of the number of reviews containing positive and negative sentiments of two hotels using Lexicon dictionary for classification as follows.

![Figure 1. Percentage of Positive and Negative Sentiment](image_url)

Figure 1 shows that the data used consisted of negative and positive sentiments. Based on the classification using the lexicon, the percentage of positive sentiment in Favehotel is 65% (1153 data) and negative sentiment is 35% (623 data). Whereas at Gunawangsa MERR Hotel is 58% (1270 data) for the percentage of positive sentiment and negative sentiment is 42% (902 data).

Visualization of Word Cloud

Visualization with word cloud is used to find out the predictor variables (words) that often appear in data reviews. Here are the word cloud results for positive sentiment on Favehotel and Gunawangsa Hotel.

![Figure 2. Word Cloud](image_url)

Based on Figure 2 (i), the most keywords that lead to the two hotels with the biggest positive sentiment are ‘clean’ and ‘comfortable’. Then the visualization of negative word cloud sentiment based on Figure 2 (ii) a shows that the keywords that lead to Favehotel with the biggest negative sentiment are 'hot' and 'service'. Whereas in Figure 2 (ii) b shows that in Gunawangsa the biggest negative sentiment is the words 'service' and 'toilet'. This is because the services provided are still not(219,484),(777,790)

Naïve Bayes Classification (NBC)

Figure 1 shows that the percentage of positive sentiment reviews is more than negative sentiment, this condition shows the imbalance data. Therefore, data review will be carried with Synthetic Minority Oversampling Technique (SMOTE), where the amount of data containing negative sentiment will be
equated with the amount of positive sentiment data. This research examines impact SMOTE to sentiment analysis while percentage splitting involves into 90% will be training data and 10% will be testing data. Then the results of confusion matrix of Favehotel and Gunawangsa with SMOTE Data is as follows.

### Table 3. Confusion Matrix of Favehotel with NBC method

| Actual Class | Prediction Class | |
|--------------|-----------------|---|
| Positive     | 55              | 7 |
| Negative     | 9               | 106 |

Table 3 shows that the confusion matrix of Favehotel data true classified as positive is 55, whereas for positive data that is classified as negative is 7. In negative data that is classified as positive is 9 while for negative data that is classified as true negative is 106.

### Table 4. Confusion Matrix of Gunawangsa with NBC method

| Actual Class | Prediction Class | |
|--------------|-----------------|---|
| Positive     | 85              | 6 |
| Negative     | 30              | 97 |

Table 4 shows the confusion matrix of Gunawangsa for positive data that is classified as positive is 85, whereas for positive data that is classified as negative is 6. On negative data that is classified as positive is 30 while for negative data that is classified as true negative is 97.

### Binary Logistic Regression

Figure 1 shows that the percentage of positive sentiment reviews is more than negative sentiment, this condition shows the imbalance data. Therefore, data review will be carried with SMOTE while percentage splitting involves into 90% will be training data and 10% will be testing data. Then the results of confusion matrix of Favehotel and Gunawangsa with SMOTE Data is as follows.

### Table 5. Confusion Matrix of Favehotel with Binary Logistic Regression Method

| Actual Class | Prediction Class | |
|--------------|-----------------|---|
| Positive     | 56              | 7 |
| Negative     | 19              | 97 |

Table 5 shows a confusion matrix where the classification for positive data which is classified as positive is 56, whereas for positive data that is classified as negative is 7. In negative data which is classified as positive is 19 while for negative data which is classified as true negative is 97.

### Table 6. Confusion Matrix of Gunawangsa with Binary Logistic Regression Method

| Actual Class | Prediction Class | |
|--------------|-----------------|---|
| Positive     | 69              | 8 |
| Negative     | 13              | 127 |

Table 6 shows a confusion matrix where the classification for positive data that is classified as positive is 69, whereas for positive data which is classified as negative is 8. On negative data that is classified as positive is 13 while for negative data that is classified as true negative is 127.
Selection of The Best Classification Method

Comparison between NBC and Binary Logistic Regression method results that with SMOTE, Binary Logistic Regression is better than NBC. The following are presented with training and testing based on average value of the accuracy, precision, recall and AUC shown in Table 7.

Table 7. Selection of the Best Classification Method

| Data | Hotel     | Accuracy | Precision | Recall | AUC  |
|------|-----------|----------|-----------|--------|------|
|      | Favehotel | 0.89     | 0.87      | 0.92   | 0.89 |
|      | Gunawangsa| 0.88     | 0.85      | 0.91   | 0.88 |

Table 7 shows that Logistic Binary Regression with SMOTE in training and testing is better than NBC method. This can be seen AUC value in Favehotel training data is 0.89 while in Gunawangsa the value of AUC is 0.88. Meanwhile, using testing data, the AUC value at Favehotel is 0.84 and Gunawangsa is 0.82.

5. Conclusion and Suggestion

Classification using lexicon showing that the percentage of positive sentiment in Favehotel is 65% and negative sentiment is 35% while in Hotel Gunawangsa MERR each is 58% for positive sentiment and negative sentiment by 42%. Word cloud visualization shows that the most keywords that lead to the two hotels with the biggest positive sentiment are the words 'clean' and 'comfortable'. While the most keywords with negative sentiment on Favehotel are 'hot' and 'service' while in Gunawangsa are 'service' and 'toilet'. Comparison of the method between NBC and Binary Logistic Regression obtained a decision that the Binary Logistic Regression method with SMOTE is better when compared to the NBC method where the AUC Favehotel value on testing data is 0.84 and AUC Gunawangsa on testing data is 0.82.

Suggestions that can be given from the results of this study are need to pay attention to suggestions and negative reviews from visitors. In addition, expected to make the results of this study as additional information so that the Favehotel and Hotel Gunawangsa made issue's rooms that hot because the air conditioning and another services are less than the maximum so that in the future even better.

References

[1] Dey L, Chakraborty S, Biswas A, Bose B and Tiwari S 2016 Sentiment Analysis of Review Datasets using Naive Bayes and K-NN Classifier. I.J. Information Engineering and Electronic Business, 8(4), 54-62
[2] Turland M 2010 php|architect's Guide to Web Scraping Canada: Marco Tabini & Associates. Inc
[3] Weiss S, Indurkhya N, Zhang T and Damerau F 2005 Text Mining Predictive Methods for Analyzing Unstructured Information New York: Spinger
[4] Pravina A M, Choliissodin I, and Adikara P P 2018 Analisis Sentimen tentang Opini Maskapai Penerbangan pada Dokumen Twitter Menggunakan Algoritme Support Vector Machine (SVM) Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer. 3(3). 2789-2797
[5] Refaeilzadeh P, Tang L and Liu H 2009 Cross-validation. Encyclopedia of Database System. (pg. 532-538) United States: Springer
[6] Feldman R and Sanger J 2007 The Text Mining Handbook : Advanced Approaches in Analyzing Unstructured Data New York: Cambridge University Press
[7] Murnawan and Sinaga A 2017 Pemanfaatan Analisis Sentimen untuk Pemeringkatan Popularitas Tujuan Wisata Jurnal Penelitian Pos dan Informatika. 7(2). 109-120
[8] Hosmer D W and Lemeshow S 2000 Applied Logistic Regression (2nd Edition) Canada: John Wiley & Sons. Inc
[9] Hotho A, Numberger A and Paas G 2005 *A Brief Survey of Text Mining* Kassel: University of Kassel.