Journal of Soft Computing Exploration

Optimization of Naïve Bayes Classifier By Implemented Unigram, Bigram, Trigram for Sentiment Analysis of Hotel Review

Ilham Esa Tiffani
Computer Science Department, Faculty of Mathematics and Natural Sciences, Universitas Negeri Semarang, Indonesia

Article Info

Article history:

Received Aug 20, 2020
Revised Sept 6, 2020
Accepted Sept 19, 2020

Keywords:

Text Mining;
Sentiment Analysis;
Naïve Bayes Classifier;
N-Gram;
Hotel Review

ABSTRACT

The information needed in its development requires that proper analysis can provide support in making decisions. Sentiment analysis is a data processing technique that can be completed properly. To make it easy to classify hotels based on sentiment analysis using the Naïve Bayes Classifier algorithm. As a classification tool, Naïve Bayes Classifier is considered efficient and simple. In this study consists of 3 stages of sentiment analysis process. The first stage is text pre-processing which consists of transform case, stopword removal, and stemming. The second stage is the implementation of N-Gram features, namely Unigram, Bigram, Trigram. The N-Gram feature is a feature that contains a collection of words that will be referred to in the next process. Next, the last click is the hotel review classification process using Na menggunakanve Bayes Classifier. OpinRank Hotels Review dataset on Naïve Bayes Classifier using N-Gram namely Unigram, Bigram, Trigram with research results that show Unigram can provide better test results than Bigram and Trigram with an average accuracy of 81.30%.

Corresponding Author:

Ilham Esa Tiffani
Computer Science Departement
Faculty of Mathematich and Natural Sciences, Universitas Negeri Semarang,
Email: tiffaniilham@gmail.com

1. INTRODUCTION

The development of information technology and websites, it allows the managers of the world of tourism to provide more detailed information about the tourism products offered. Currently there are many travel websites that provide facilities for internet users to write their opinions and personal experiences online [1].

A text can consist of only one word or sentence structure [2]. Information in the form of text is important information and is widely obtained from various sources such as books, newspapers, websites, or e-mail messages. Retrieval of information from text (text mining), among others, can include text or document categorization, sentiment analysis, search for more specific topics (search engines), and spam filtering [3]. Text mining is one of the techniques that can be used to do classification where, text mining is a variation of data mining that tries to find interesting patterns from a large collection of textual data [4].

The classification method itself many researchers use the Naïve Bayes Classifier where a text will be classified in machine learning based on probability [5]. Naïve Bayes Classifier is a pre-processing technology in the classification of features, which adds scalability, accuracy and efficiency which is certainly very much in the process of classifying a text. As a classification tool, Naïve Bayes Classifier is considered efficient and simple, and sensitive to feature selection [6].

The data used in this study contains hotel reviews in English so that it can be seen that the grammar used by a person is very diverse in writing the review, diversity makes the features generated through N-Gram will be very much. Therefore, here we will use N-Gram word characters with N = 1, 2, 3 to retrieve features in a review which will then be classified with the Naïve Bayes Classifier Algorithm.

It is expected that the N-Gram Naïve Bayes Classifier Algorithm in this study can be classified correctly and appropriately. So that the main purpose of this study can be fulfilled which is to know the effect of N-Gram features on Naïve Bayes Classifier for sentiment analysis of hotel reviews.

2. METHOD

The step of research include text pre-processing, the application of N-Gram features, and the Naïve Bayes Classifier classification. The research starts by inputting OpinRank Hotels Review datasets. Next, the data will be processed in the text pre-processing stage, namely with a transform case, stopword removal, and stemming. Then the N-Gram feature selection will be carried out, namely unigram, bigram, trigram. Based on the selected features, the classification process will be carried out using the Naïve Bayes Classifier algorithm. Then the classification model is tested using test data and evaluated using a confusion matrix to produce accuracy values. Flowchart of the research method can be seen in Figure 1.

2.1 Dataset

The data used in this study is OpinRank Hotels Review Dataset (in English) obtained from the UCI Machine Learning Repository. The data contains 1000 documents consisting of 500 documents labeled positive and 500 labeled negative. The dataset is obtained in .txt format and then the file is converted into a table with two columns: the first column contains text and the second column contains labels defined by "0" means negative and "1" means positive as shown in Table 1.

Table 1. Data samples in CSV format

Text	Label
Poor location.. This hotel is located in a run down part of the city. The hotel room smelt of ammonia, the toilet would not flush and we could not sleep due to the traffic/street noise. The breakfast was poor and over priced at $12.50. We would not stay there again.	0
I had two nights stay at this hotel, very nice sleep, the bed was fantastic. Staffs' service was good and helpful.	1

2.2 Text Pre-processing

The text pre-processing phase performed in this study is Transform Case, Stopword Removal, Stemming. Text pre-processing is the stage of the initial process of the text to prepare the text into data that will be further processed.

2.2.1 Transform case

The process to change the form of words, in this process the characters are made into lowercase or lower case all. The steps of the transform case process are as follows:

Step 1 : Data input used in the form of hotel reviews.
Step 2 : Hotel review data if there are characters that use capital letters (uppercase), then these characters will be changed to lowercase (lowercase).
Step 3 : Hotel review data becomes lowercase which is then used in the stopword removal process.
Step 4 : The process is complete. The results of the process of the transform case stage can be seen in Table 2

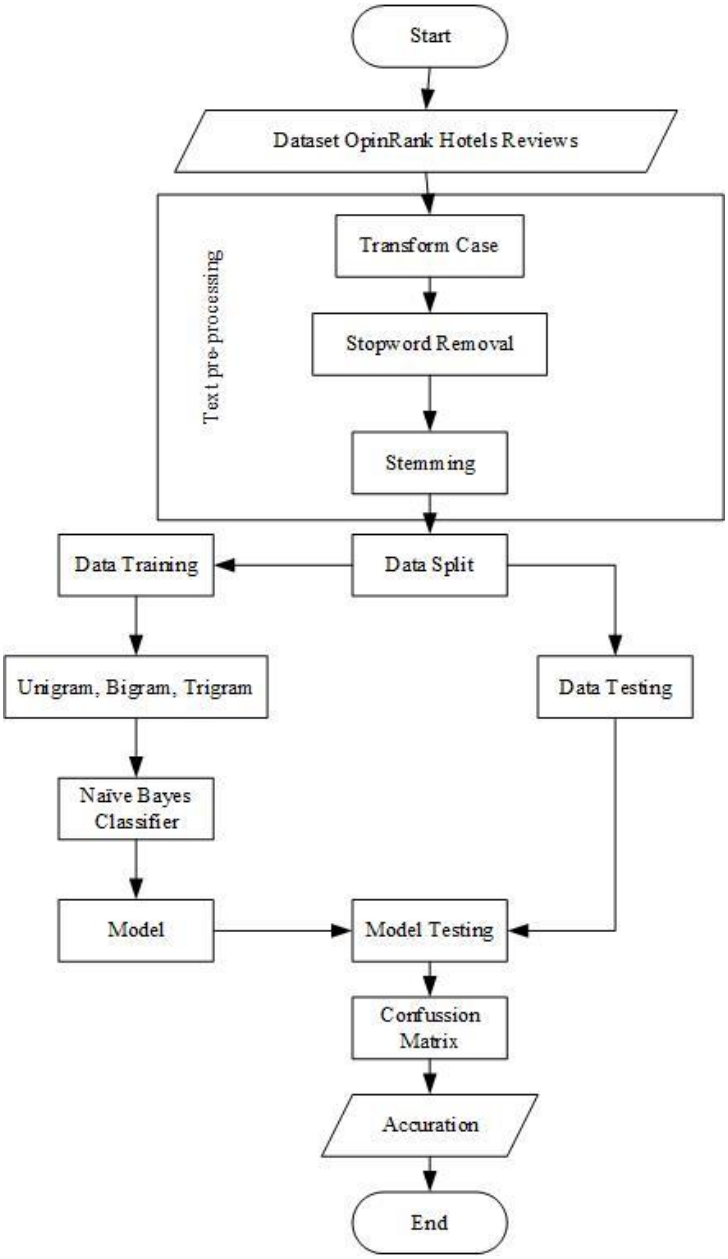

Figure 1. Naive Bayes Classifier algorithm with N-Gram flowchart

Table 2. Results of the Transform Case Process

Review of Data	Transform Case Results
My husband and I stayed at the Chamberlain Hotel for three nights	my husband and i stayed at the chamberlain hotel for three nights

2.2.2 Stopword removal

Stopword Removal is the process of removing words that often appear but do not have any effect in the extraction of text classifications. The steps for the stopword removal process are as follows:

Step 1 : The word from the transform case result will be compared with the word in the stopword list.
Step 2 : Check whether the word is the same as the stopword list or not.
Step 3 : If the word is the same as the stopword list, then the word will be deleted.
Step 4 : The process is complete. The results of the process of the stopword removal stage can be seen in Table 3.

Tabel 3. Results of the Stopword Removal Process

Transform Case Results	Stopword Removal Results
my husband and i stayed at the chamberlain hotel for three nights	husband stayed chamberlain hotel nights

2.2.3 Stemming

The process of mapping and decomposing the shape of a word into basic word forms. The purpose of the stemming process is to eliminate the affixes that exist in each word. The words in the stopword list are pronouns, connectors and pointers. The steps in the stemming process are as follows:

Step 1 : The word from the stopword removal result is checked, whether the word from the stopword removal result is a basic word or not
Step 2 : If the root word then the process has stopped or finished but if it is not a root word then delete the suffix (the affix which is located at the end of the word)
Step 3 : Word resulting from suffix deletion if it is a base word, the process is complete, but if it is not
Step 4 : The process is complete. The results of the process of the stemming stage can be seen in Table 4.

Tabel 4. Results of the Stemming Process

Stopword Removal Results	Stemming Results
husband stayed chamberlain hotel nights	husband stay chamberlain hotel night

2.3 N-Gram

N-Gram is a n-character chunk taken from a string [7]. N-Gram is used in the process of making a model by dividing a sentence into parts of words. In N-Gram, 'N' shows the number of words that will be grouped into one section. On research [8] divide the N-Gram into three types, namely:

a. Unigram: token consisting of only one word.
b. Bigram: a token consisting of two words.
c. Trigram: a token consisting of three words.

The rules used to form the three types of tokens are overlapping tokens. Examples of the process of N-Gram characters generated from the comments results of the Stemming stage can be seen in Table 5.

Table 5. N-Gram Process

N-Gram Results	
Unigram	husband, stay, chamberlain, hotel, night
Bigram	husband stay, stay chamberlain, chamberlain hotel, hotel night
Trigram	husband stay chamberlain, stay chamberlain hotel, chamberlain hotel night

In this study took up to 3 words because in the structure of English phrases with a single meaning have a maximum of 3 words. Phrases are added to a sentence to make the sentence more complex. The advantage of N-Gram is based on the characteristics of N-Gram as part of a string, so errors in some strings will only result in differences in some N-Gram.

2.4 Naïve Bayes Classifier

Naïve Bayes Classifier is a statistical classification that can be used to predict the probability of membership of a class. Naïve Bayes Classifier is based on the Bayes theorem which has the same classification capabilities as the Decision Tree and Neural Network. Naïve Bayes Classifier is proven to have high accuracy and speed when applied to databases with large data [9]. Naïve Bayes Classifier is a popular and good algorithm for high-dimensional data such as text [10].

The flow of the Naïve Bayes Classifier can be seen in Figure 2 as follows:

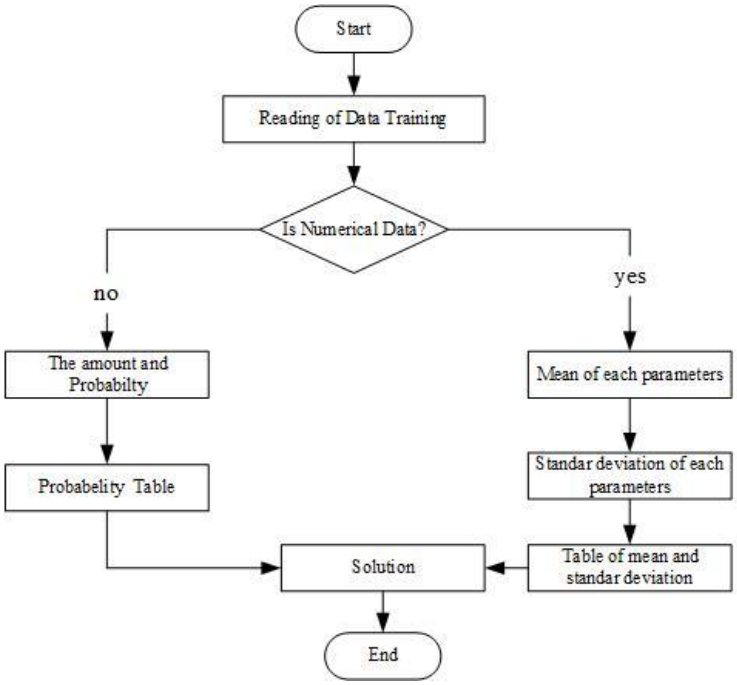

Figure 2. Naïve Bayes Classifier flowchart

Classification is the process of classifying a collection of objects, data or ideas into groups, where each member has one or more of the same characteristics. The classification stage using the Naïve Bayes Classifier is divided into 2 processes, namely training and testing. The training process is carried out to produce a probabilistic model of features that will later be used as a reference calculation for classifying testing data. The stages of sentiment classification use the Naïve Bayes Classifier as follows:

a. Training Process
    1. Count $P(c_i)$
    2. Count $P(w_k \mid c_i)$ for each w_k on the model
b. Testing Process
    1. Count $P(c_i)$ Πk $P(w_k \mid c_i)$ for each category
    2. Decide c^*, i.e. categories with values $P(c_i)$ Πk $P(w_k \mid c_i)$ the highest

3. RESULT AND DISCUSSION

3.1 Result

In research, the proposed algorithm is tested using the python programming language. The classification process in the dataset uses the Naïve Bayes Classifier algorithm, the classification in the dataset with the Naïve Bayes Classifier algorithm applied to Unigram produces 81.30% accuracy, the dataset classification with the Naïve Bayes Classifier algorithm applied to Bigram produces an accuracy of 71.60%, and the classification of the dataset with the Naïve Bayes Classifier algorithm is applied Trigram produces 71.90% accuracy.

3.2 Discussion

3.2.1 Naïve Bayes Classifier algorithm + *Unigram*

This classification stage applies the Naïve Bayes Classifier algorithm with Unigram on the OpinRank hotels review dataset. The training data will be divided into 10 subset data to conduct the learning process. This process takes 10 iterations to then get the classification model. Then the algorithm will be tested with a confusion matrix. From the classification of the Naïve Bayes Classifier algorithm by applying Unigram produces accuracy as shown in Table 6.

Table 6. Accuracy results of Naïve Bayes Classifier + Unigram algorithm

k to	Accuracy
1	68,00%
2	75,00%
3	79,00%
4	85,00%
5	83,00%
6	84,00%
7	86,00%
8	85,00%
9	84,00%
10	84,00%
Average	81,30%

3.2.2 Algoritma *Naïve Bayes Classifier + Trigram*

This classification phase applies the Naïve Bayes Classifier algorithm with Trigram on the OpinRank hotels review dataset. The training data will be divided into 10 subset data to conduct the learning process. This process takes 10 iterations to then get the classification model. Then the algorithm will be tested with a confusion matrix. From the classification of the Naïve Bayes Classifier algorithm by applying Trigram produces accuracy as shown in Table 7.

Table 4.8. Accuracy results of Naïve Bayes Classifier + Trigram algorithm

k to	Accuracy
1	63,00%
2	74,00%
3	75,00%
4	75,00%
5	73,00%
6	72,00%
7	70,00%
8	71,00%
9	74,00%
10	72,00%
Average	**71,90%**

The accuracy of the Naïve Bayes Classifier algorithm using Unigram, Bigram, Trigram compared to related research results in better accuracy. OpinRank hotels review dataset using Naïve Bayes Classifier in related research has an accuracy of 55.00%, the classification of the Naïve Bayes Classifier algorithm with Unigram is able to produce an average accuracy of 81.30%, the classification of the Naïve Bayes Classifier algorithm with Bigram is able to produce an average accuracy of 71.60%, and the classification of the Naïve Bayes Classifier algorithm with the Trigram is capable of producing an average accuracy of 71.90. Comparison of accuracy results can be seen in Figure 3.

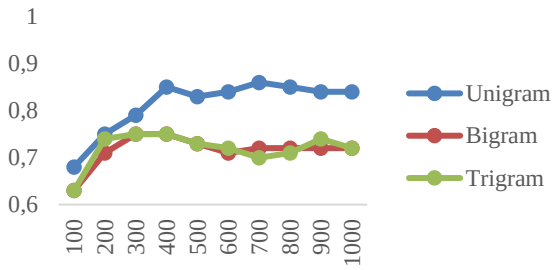

Figure 3. Graph of accuracy results of Naïve Bayes Classifier algorithm with N-Gram

Based on the results of the implementation of Unigram, Bigram, Trigram on the Naïve Bayes Classifier algorithm that has been done, it can be seen that the accuracy for sentiment analysis of hotel reviews using

OpinRank Hotels Review datasets is taken from the UCI Machine Learning Repository after going through text pre-processing and then applying the N-Gram feature and classification using the Naïve Bayes Classifier can improve accuracy so that it can be used by subsequent researchers as a reference in conducting hotel review sentiment analysis research.

4. CONCLUSION

The application of Unigram, Bigram, Trigram on the Naïve Bayes Classifier algorithm for the analysis of hotel review sentiments in this study is the OpinRank hotel reviews dataset that has been done in the text pre-processing stage will be divided into training data and testing data. The training data will be given the features of Unigram, Bigram, Trigram by breaking a sentence into words whose results will be classified with the Naïve Bayes Classifier algorithm so as to produce a more optimal Naïve Bayes Classifier model. The final result of the classification is testing the model of data testing. Based on these tests, an accuracy of the Naïve Bayes Classifier algorithm can be seen using a confusion matrix. The average accuracy of 10 subsets of data obtained using Unigram in the Naïve Bayes Classifier algorithm is 81.30%, the average accuracy of 10 subset data obtained using Bigram in the Naïve Bayes Classifier algorithm is 71.60% and the results of the average accuracy 10 subsets of data obtained using Trigram in the Naïve Bayes Classifier algorithm are 71.90%.

REFERENCES

[1] C. Wang, Y. Zhang, J. Song, Q. Liu and H. Dong, "A novel optimized SVM algorithm based on PSO with saturation and mixed time-delays for classification of oil pipeline leak detection, " *Sys. Sci. & Con. Eng.*, vol. 7, no. 1, pp. 75-88, 2019.

[2] R. Carter and M. McCarthy, Cambridge Grammar of English Paperback with CD ROM: A Comprehensive Guide. Cambridge, UK: Cambridge University Press, 2006, pp.179-185.

[3] N. Buslim, A. E. Putra, and L. K. Wardhani, "Chi-square feature selection effect on naive bayes classifier algorithm performance for sentiment analysis document, " presented at the 7th International Conference on Cyber and IT Service Management, Jakarta, Indonesia, Nov. 6-8. 2019.

[4] R. Feldman and J. Sanger, The Text Mining Handbook . Cambridge, UK: Cambridge University Pres. 2009.

[5] W. Zhang and F. Gao, "An Improvement to Naive Bayes for Text Classification, " *Procedia Engineering*, vol. 15, no. 2, pp. 2160–2164, 2011.

[6] J. Chen, H. Huang, S. Tian, and Y. Qu, "Feature Selection for Text Classification with Naïve Bayes, " *Expert System Application*, vol, 36, no. 3, pp. 5432–5435, 2009.

[7] J. Violos, K. Tserpes, I. Varlamis, and T. Varvarigou, "Text classification using the n-gram graph representation model over high frequency data streams, " *Front. Appl. Math. Stat*. vol. 4, no. 41, pp. 1-19, 2018.

[8] Z. Drus and H. Khalid, " Sentiment Analysis in Social Media and Its Application: Systematic Literature Review, " presented at the fifth Information Systems International Conference, Surabaya, Indonesia, July 23-24, 2019.

[9] R. E. Banchs. Text Mining with MATLAB®. New Delhi, India: Springer-Verlag New York, 2013, pp. 49-75.

[10] I. Y. R. Pratiwi, R. A. Asmara, and F. Rahutomo, "Study of Hoax News Detection using Naïve Bayes Classifier in Indonesian Language," presented at the 11th International Conference Informatics Communication Technology System ICTS, Surabaya, Indonesia, Oct. 30-31, 2017.