Comparison of Naïve Bayes Algorithm with Genetic Algorithm and Particle Swarm Optimization as Feature Selection for Sentiment Analysis Review of Digital Learning Application

Comparison of Naïve Bayes Algorithm with Genetic Algorithm and Particle Swarm Optimization as Feature Selection for Sentiment Analysis Review of Digital Learning Application

Siti Ernawati[1*], Risa Wati[2], Nuzuliarini Nuris[2], Lita Sari Marita[2], Eka Rini Yulia[1]

[1]Sistem Informasi, Sekolah Tinggi Manajemen Informatika dan Komputer Nusa Mandiri
[2]Sistem Informasi, Universitas Bina Sarana Informatika

E-mail: siti.ste@nusamandiri.ac.id

Abstract. The problem examined in this study is about the user's trust in using digital learning applications that are downloaded on playstore. Many reviews are given by the public about the application that has been downloaded on playstore. This review is very influential on their trust in using the application. The purpose of this study is to classify data according to labels and find out the best choice between the classification method and the proposed selection feature as a consideration in determining the use of digital learning applications.This study compares the classification method, the Naïve Bayes algorithm and the genetic algorithm (GA) as feature selection with the Naïve Bayes algorithm classification method and the particle swarm optimization (PSO) as feature selection to categorize the reviews in the playstore. The experimental results show that the Naïve Bayes algorithm and PSO as feature selection is the best model between the two models proposed in this study. Reviews can be classified into positive and negative labels well. The accuracy is 98.00%. The results of the classification are expected to help in making decisions when going to use digital learning application.

1. Introduction

Indonesia has entered the industrial revolution 4.0, the reason can be seen from the increasingly sophisticated technological advancements that bring many conveniences in various fields. The examples that are most felt by the people of Indonesia are in the areas of buying and selling, transportation, education to the ease of making payments, all these activities are carried out digitally. All of that is one of the impacts felt by the people of Indonesia in the development of the industrial revolution 4.0. As technology develops, there are many digital-based startups that focus on education. Now technology-based learning services, can be accessed easily in the form of applications that can be downloaded via android-based smartphones. The number of digital learning applications that can be downloaded in Playstore will cause problems about the user's trust in using the application.

Categorize the review is not easy because the number of reviews that are generally published in social media is very large, so it requires a special technique or method that can categorize reviews in positive or negative reviews without us having to sort them out manually. This is one

of the problems in the process of classifying review. In determining the method for sentiment analysis machine learning based usually has a very large feature space, so there will be problems that can cause some tasks to be solved[1]. One technique or method for classifying reviews using the Naïve Bayes algorithm. Naïve Bayes is widely used for the classification of texts based on the probability/likelihood requirements of each class, each class feature is selected using the feature selection method[2]. Categorization involves identifying the main themes of a document by including the document in a series of topics that have been predetermined. When categorizing documents, a computer program will often treat the document as a collection of words. Categorization only counts the words that appear and identifies the main topic. Categorization often depends on a list of rare words whose topics have been predetermined[3]. This study compares the classification method, the Naïve Bayes algorithm and the genetic algorithm as feature selection with the Naïve Bayes algorithm classification method and PSO as feature selection to categorize the reviews in the playstore.

Feature selection is one of the factors that can improve classification accuracy[8]. Four examples of feature selection that are often used in text mining are: (1) genetic algorithms, (2) evolutionary programming, (3) evolutionary strategies and genetic programming, (4) PSO[9]. For this reason, researchers choose and compare GA and PSO and then find the best one between the two feature choices. The main reason why we chose GA and PSO is because both of these techniques are widely used among researchers and have been successfully applied in many fields. Previous research stated that overall GA and PSO are good solutions as a feature selection technique but PSO is much better than GA because PSO has succeeded in reducing the number of features[6]. PSO is also Easier to implement and can find the optimal point quickly[11].

Research conducted by Serkan Gunal shows that text classification combined with genetic algorithms as feature selection is proven to be relatively capable and fast among the many algorithms used in the process of text classification[10]. One study conducted by Ernawati, et al on sentiment analysis explains that the improvement of the Naive Bayes algorithm when using GA as a feature selection[4]. The study, entitled Sentiment Analysis of Movie Reviews using the Hybrid Method of Naive Bayes and Genetic Algorithm, concluded that the proposed method, the hybrid NB-GA, showed a significant increase[5]. The purpose of this study is to classify data according to labels and find out the best choice between the classification method and the proposed selection feature as a consideration in determining the use of digital learning applications.

2. Reseach Method

In the classification process using training data and random test data using a cross validation dataset to get the best accuracy.

Figure 1. Proposed Research Model.

The following are the steps of the research method:
a. Data Collection : Type of data used in this research is primary data. This research used data from the Play Store. The data used are 200 reviews consisting of 100 positive reviews and 100 negative reviews.
b. Initial Data Processing : The dataset used is training data. In the initial data processing,

this dataset must pass the preprocessing stage. Preprocessing method has a very important role in text mining techniques[3]. The method used in pre-processing is as follows:
1) Case Folding: converts the entire text in a document into a standard form. This process usually converts uppercase to lowercase.
2) Tokenization: This method is used to tokenize which is to separate words or letters from punctuation and symbols.
3) Stopwards Removal: used to eliminate unnecessary words in processing data reviews.
4) N-gram: N-gram is obtained by reading each line of text and grouping strings into different sizes, the string moves forward character by character[7].
c. Proposed Method : The method proposed by the researchers is the classification method, the Naïve Bayes algorithm and GA as feature selection with the Naïve Bayes algorithm classification and PSO as feature selection. This comparison is done to find out the best feature selection to improve accuracy.
d. Experimentation and Testing Methods : Experimentation and testing methods are measured using a confusion matrix while the results of data processing using rapidminer to get the maximum accuracy value.
e. Evaluation of Results : The evaluation was done after the data is processed by comparing the value of the accuracy of each experiment. The higher accuracy and the better proposed model.

3. Result and Discussion

There are two models proposed in this research, namely the naïve bayes algorithm and GA as feature selection model with the naïve bayes algorithm and PSO as a feature selection model. Data in this study are a collection of reviews of digital learning applications taken from Playstore. Results of testing the model will be discussed through a confusion matrix to show the best model of the proposed model.

3.1. Naïve Bayes Algorithm and GA as Feature Selection

The first model to be processed is the naïve bayes algorithm and GA as feature selection. In the GA to get the highest accuracy results required parameters that require adjusment. In table 1 is the parameter that will be evaluated in this study. After the parameters are evaluated, the highest accuracy value can be taken which can be seen in table 2.

Table 1. Experiment Plans for Naïve Bayes Algorithm and GA as Feature Selection.

Population Size	P Initialize	P Crossover	P Generate	Accuracy
5-10	0.5-0.6	0.5-0.7	0.1-0.2	?

The proposed model uses naïve bayes algorithm and GA produces an accuracy of 95.50% with population size value is 9, initialize 0.5, crossover=0.5 and generate=0.1. The accuracy will be compared using naïve bayes and PSO as feature selection. Table 3 is the confusion matrix of the first proposed model.

Table 2. Experiment Results of Naïve Bayes Algorithm and GA as Feature Selection.

Population Size	P Initialize	P Crossover	P Generate	Accuracy
8	0.5	0.5	0.1	95.00%
9	0.5	0.5	0.1	95.50%
9	0.5	0.6	0.1	95.50%
9	0.5	0.7	0.1	95.50%
9	0.5	0.5	0.2	95.50%

Table 3. Confusion Matrix Naïve Bayes Algorithm and GA as Feature Selection.

Accuracy: 95.50%			
	true review positif	true review negatif	class precision
pred. review positif	97	6	94.17%
pred. review negatif	3	94	96.91%
class recall	97.00%	94.00%	

3.2. Naïve Bayes Algorithm and PSO as Feature Selection

The second model to be processed is the naïve bayes algorithm and PSO as feature selection model. To get the highest accuracy results required also adjusment parameters. Table 4 is the parameter to be evaluated in this study. Table 5 is the accuracy value after the parameters in the PSO are evaluated.

Table 4. Experiment Plans for the Naïve Bayes Algorithm and PSO as Feature Selection.

Population Size	Inertia Weight	Accuracy
10-15	0.1-1.0	?

Table 5. Experiment Results of Naïve Bayes Algorithm and PSO as Feature Selection.

Population Size	Inertia Weight	Accuracy
11	1.0	96.50%
13	1.0	97.00%
14	0.9	96.50%
14	1.0	98.00%
15	1.0	97.50%

The proposed model using naïve bayes and PSO as feature selection produces an accuracy is 98.00% with population size=14 and intertia weight=1.0. Table 6 show the confusion matrix generated from the naïve bayes algorithm and PSO as feature selection.

Table 6. Confusion Matrix Naïve Bayes Algorithm and PSO as Feature Selection.

Accuracy: 98.00%			
	true review positif	true review negatif	class precision
pred. review positif	100	4	96.15%
pred. review negatif	0	96	100.00%
class recall	100.00%	96.00%	

3.3. Model Comparison

Based on the proposed model, the accuracy values of each model are compared so that the best accuracy will be obtained. The results of the comparison are the Naïve Bayes algorithm and GA as feature selection is 95.50%. The accuracy of Naïve Bayes algorithm and PSO as feature selection is 98.00%. The results of the comparison can be seen in table 7. Comparison graphs of the accuracy values can be seen in Figure 2.

Table 7. Comparison Results between Naïve Bayes Algorithm and GA with Naïve Bayes Algorithm and PSO as Feature Selection.

Experimentation	Naïve Bayes Algorithm + GA	Naïve Bayes Algorithm + PSO
Success classification positive review	97	100
Success classification negative review	94	96
Accuracy	95.50%	98.00%

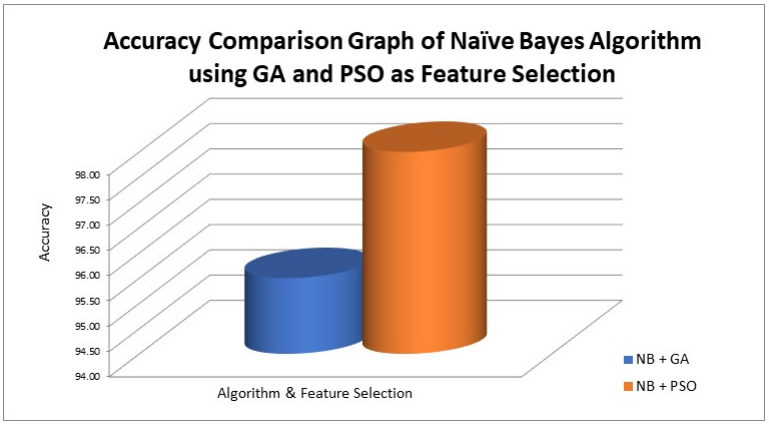

Figure 2. Accuracy Comparison Graph.

4. Conclusion

Researchers have applied two proposed models namely Naive Bayes algorithm and GA as feature selection with Naive Bayes algorithm and PSO as feature selection into digital learning application review data. From the results of data processing that has been done, reviews of digital learning service applications can be correctly classified into positive and negative labels. Comparison between the two proposed models is proven that the Naïve Bayes algorithm and PSO as feature selection produce the highest accuracy value. The accuracy of naïve bayes algorithm and GA as feature selection is 95.50%, while the accuracy of naïve bayes and PSO algorithms as feature selection is 98.00%. The difference in accuracy is 2.50%. PSO proved better than GA. PSO has succeeded in reducing the number of features of the data in the form of reviews. Future studies are expected to use more data and use newer classification methods so as to obtain higher accuracy.

Acknowledgments

We thank all those who have offered prayers and support to researchers in completing this research. Also the editorial team who have checked our paper to be published in the international journals.

References

[1] Koncz P and Paralic J, 2011 An approach to feature selection for sentiment analysis 2011 15th IEEE Int. Conf. Intell. Eng. Syst. p. 357–362.

[2] Zhang W and Gao F, 2011 An Improvement to Naive Bayes for Text Classification Elsevier 15 p. 2160–2164.

[3] Vijayarani S Ilamathi M J and Nithya M, 2015 Preprocessing Techniques for Text Mining - An Overview Int. J. Comput. Sci. Commun. Networks 5, 1 p. 7–16.

[4] Ernawati S Yulia E Frieyadie and Samudi, 2018 Implementation of The Naïve Bayes Algorithm with Feature Selection using Genetic Algorithm for Sentiment Review Analysis of Fashion Online Companies 2018 6th Int. Conf. Cyber IT Serv. Manag. Citsm p. 1–5.

[5] Ghareb A S Bakar A A and Hamdan A R, 2015 Hybrid Feature Selection Based On Enhanced Genetic Algorithm For Text Categorization Expert Syst. Appl. 49 p. 31–47.

[6] Syarif I, 2016 Feature Selection of Network Intrusion Data using Genetic Algorithm and Particle Swarm Optimization Emit. Int. J. Eng. Technol. 4, 2 p. 277–290.

[7] Ahmed B Cha S and Tappert C, 2004 Language Identification from Text Using N-gram Based Cumulative Frequency Addition Proc. Student/Faculty Res. Day, CSIS, Pace Univ. p. 12.1-12.8.

[8] Liu Y Wang G Chen H Dong H Zhu X and Wang S, 2011 An Improved Particle Swarm Optimization for Feature Selection J. Bionic Eng. 8, 2 p. 191–200.

[9] Shi Y and Eberhart R C, 1945 Empirical Study of Particle Swarm Optimization IEEE p. 1945–1950.

[10] Günal S, 2012 Hybrid feature selection for text classification Turkish J. Electr. Eng. Comput. Sci. 20, SUPPL.2 p. 1296–1311.

[11] Yonghe L Minghui L Zeyuan Y and Lichao C, 2015 Improved particle swarm optimization algorithm and its application in text feature selection Appl. Soft Comput. J. 35 p. 629–636.