Feature Selection Based on Genetic Algorithm, Particle Swarm Optimization and Principal Component Analysis for Opinion Mining Cosmetic Product Review

by Mochamad Wahyudi

Submission date: 23-Apr-2019 11:06PM (UTC+0700)

Submission ID: 1117694365

File name: 2017_-_Feature_Selection_Based_on_Genetic_Algorithm...-20-25.pdf (220.65K)

Word count: 5483

Character count: 28122

Feature Selection Based on Genetic Algorithm, Particle Swarm Optimization and Principal Component Analysis for Opinion Mining Cosmetic Product Review

Dinar Ajeng Kristiyanti
STMIK Nusa Mandiri Jakarta
Jakarta, Indonesia
dinar@nusamandiri.ac.id

Mochamad Wahyudi
STMIK Nusa Mandiri Jakarta
Jakarta, Indonesia
wahyudi@nusamandiri.ac.id

***Abstract*- Opinion mining is an automation technique of textual data from opinion sentence that produce sentiment information. It is also called sentiment analysis that involves the construction of a system for collecting and classifying opinions about a product review done by understanding, extracting and processing the text in an opinion sentence become positive, negative, and neutral. One of the techniques mostly used by data classification is Support Vector Machine (SVM). SVM is able to identify the separated hyper plane that maximizes the margin between two different classes. However, SVM has a weakness for parameter selection or suitable feature. In this research, the researchers made an improvement toward the previous research using combined method of feature selection in SVM through comparing three-feature selection; Genetic Algorithm, Particle Swarm Optimization, and Principal Component Analysis. It can be determined which one of the best feature selections that improve the classification accuracy of SVM. The dataset was cosmetic products review downloaded from www.amazon.com. Measurement is based on SVM accuracy by adding the feature selection method. While the evaluation used 10 Fold Cross Validation and the accuracy measurement used the Confusion Matrix and ROC Curve. The result of the measurement accuracy of SVM accuracy is obtained with average 82.00% and the average AUC 0.988. After the integration of SVM algorithm and feature selection, Genetic algorithm shows the best results with average accuracy 94.00% and the average AUC 0.984. Particle Swarm Optimization indicates the best results with average accuracy 97.00% and the average AUC 0.988. While Principal Component Analysis indicates the best results with average accuracy 83.00% and the average AUC 0.809. As conclusion, the research of SVM Algorithm showed the best accuracy improvement toward the feature selection of Particle Swarm Optimization integrated with the increased accuracy from 82.00% to 97.00%.**

Keyword: Cosmetic Product Review; Feature Selection; Opinion Mining.

I. Introduction

Nowadays consumers who write opinions and experiences online are increasing. Reading the review as a whole can be time consuming, but if only a few reviews that read evaluations will be biased. Some reviews of cosmetic products can help consumers to know the quality of the cosmetics brand whether this is feasible or not to use. Cosmetic products in today marketplace vary, both in terms of type and brand. However, not all cosmetics are good quality according to customer needs and the consumer should be aware of this one. Before consumers decide to buy cosmetics they should know the detail product, it can be learned from the testimony and opinions or results of a review of the consumers who have purchased and used the product before. Sentiment classification aims to overcome this problem by automatically classifying the user review become positive or negative opinion [22]. For that we need the reexamination of the cosmetic product review by classifying these reviews into positive and negative class so that ultimately the consumers can find consumer feedback about the products quickly and accurately.

Opinion mining or sentiment analysis is a computational study of the opinions, behaviors, and emotions over the entity. Entities may describe individuals, events or topics. The topic is likely to be a review [20]. Classification techniques commonly used for such sentiment analysis review Naïve Bayes, Support Vector Machine (SVM) and K-Nearest Neighbor (KNN) [15]. There are several studies that have been conducted within the classification sentiment towards the review available online including sentiment analysis on the review of the mobile phone users [10]. Sentiment analysis on the opinions of movie reviews using Support Vector Machine classifier and Particle Swarm Optimization [1]. Sentiment classification on review online travel destination using Naïve Bayes classifiers, Support Vector Machine and Character Based N-gram model [14]. Sentiment analysis on movie reviews and some products from Amazon.com using Support Vector Machine classifier and Artificial Neural Network [17].

Among of these techniques which are most often used for data classification is Support Vector Machine (SVM). SVM is a supervised learning methods that analyze the data and recognize patterns that are used for classification [1]. Support Vector Machine (SVM) is a special case of the family of algorithms called the regularized linear classification methods and a powerful method to minimize the risk [19]. SVM has the advantage of being able to identify separate hyperplane that maximizes the margin between the two different classes [7]. However Support Vector Machine has shortcomings on election of parameters or suitable features [1]. Selection of features at once setup parameters in SVM significantly affect the results of the classification accuracy [13].

Particle Swarm Optimization (PSO) is an optimization technique that is very simple to implement and modify some parameters [1]. One of wrapper method that can be used in the future selection is a Genetic Algorithm (GA). Genetic algorithm easily aligned and have been used for the classification of such other optimization problems [4]. PCA manages the entire data for the principal components analysis without taking into consideration the fundamental class structure [9].

In this study, the algorithm Support Vector Machine and Particle Swarm Optimization algorithm, Genetic Algorithm and Principal Component Analysis will be applied as a method of feature selection. Researchers will compare these three methods to classify text on cosmetic products review in order to improve the accuracy of sentiment analysis.

II. LITERATURE REVIEW

A. Review Product

Acoding to (Weiss, Indurkhya, & Zhang, 2010) in [11] The website is a forum for diverse opinions. One form of opinions has credibility is product review. Web sites such as amazon.com encourage users to provide reviews (review). The results of the review text mining can be classified in three categories: positive, negative, and neutral.

B. Opinion Mining

According to Tang in [3], opinion mining on the review is the process of investigation of product reviews on the internet to determine opinions or feelings to a product as a whole. According to Thelwall [3], opinion mining is treated as a task of classification that classifies text orientation into a positive or negative. According to Mejova in [1], the purpose of the analysis is to determine the behavior of sentiment or opinion of a writer with regard to a particular topic. Behavior may indicate a reason, opinions or judgments, the tendency of conditions (how the author wants to influence the reader).

C. Feature Selection

Feature Selection is optimization process to reduce a large set of the original great features in order to feature subset that relatively small and significantly improve the accuracy of classification for fast and effective.

a. Particle Swarm Optimization (PSO)

Particle Swarm Optimization (PSO) is widely used to solve optimization problems as well as a feature selection problem [21]. Optimization is the process of adjusting to the input or the device characteristics, a mathematical process, or experiment to find the minimum or maximum output results. Input consists of variables, process or function is known as a cost function, the objective function or the functional capability and the output is the cost or purpose, if the process is a trial, then the variable is the physical input to trial [16].

b. Genetic Algorithm (GA)

One of wrapper method that can be used in the feature selection is a Genetic Algorithm (GA). Genetic algorithm easily aligned and has been used for the classification of such other optimization problems. According to (Zukhri, 2014) in [12] Genetic Algorithm is a heuristic method developed based on the principles of genetics and the process of natural selection from Darwin's theory of evolution.

c. Principal Component Analysis (PCA)

The principal component analysis (PCA) is a kind of algorithms in biometrics. It is a statistics technical and used orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables. PCA also is a tool to reduce multidimensional data to lower dimensions while retaining most of the information. It covers standard deviation, covariance, and eigenvectors [18].

D. Support Vector Machine Algorithm

Support Vector Machines (SVM) is a set of methods related to a method of learning, for both classification and regression problems. With task-oriented, strong, tractable nature of computing, SVM has achieved great success and is considered a state of-the art classifier today [8].

SVM is a supervised learning method that analyze the data and recognize patterns that are used for classification [1]. SVM has the advantage of being able to identify separated hyperplane that maximizes the margin between the two different classes [7]. However Support Vector Machine has shortcomings on paramater election issues or suitable features [1]. Selection of features at once setup parameters in SVM significantly influence the results of the classification accuracy [13].

E. Validation and Evaluation of Data Mining Algorithms

There are many methods used to validate a model based on existing data, such as the holdout, random sub-sampling, cross-validation, stratified sampling, bootstrap and others. According to [6] confusion matrix is a very useful tool to analyze how well the classifier to recognize bias tuple of a different class. In the confusion matrix known terms such as True positives refers to the positive tuple is correctly labeled by the classifier, while True negative is negative tuple is correctly labeled by the classifier. There is also False positive as negative tuple which is incorrectly labelled by the classifier and False negative as positive tuple is incorrectly labeled by the classifier.

K-fold cross-validation is a validation technique with initial data randomly split into k sections mutually exclusive or "fold" [6]. ROC curves will be used to measure the AUC (Area Under the Curve). ROC curve divides a positive result in the y-axis and a negative result in the x-axis [5]. Graph curve ROC (Receiver operating characteristic) is used to evaluate the accuracy classifier and to compare the different classification models [2]. So the larger the area under the curve, the better the prediction results.

III. RESEARCH METHOD

A. *Research Framework*

This study starts from the problem in text classification on review of the product using a classifier Support Vector Machine (SVM) in which the classification has as shortage toward the suitable parameter selection problem. The incompatibility of a parameter setting can cause the classification results to be low. The data used in this research is cosmetic product review data obtained from www.amazon.com consisting of 100 positive reviews and negative reviews. Preprocessing is performed with tokenization, Generate N-Gram and Stemming. Feature weighting method to be used is the Term Frequency Inverse Document Frequency (TF-IDF) and the selection of feature selection using the Particle Swarm Optimization (PSO), Genetic Algorithm (GA) and Principal Component Analysis (PCA). While the classifier used is Support Vector Machine. 10 Fold Cross Validation testing will be done, accuracy 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 in measuring the accuracy of the experimental data is done in research.

B. *Research Methodology*

Research method used by researchers is experimental research method with the following stages:

1. Research Design
   a. Data Collection: Data for this experiment were collected, and then selected from the data that does not fit.
   b. Data Initial Processing: Model selected based on the suitability of the data with the best method of some text classification method that has been used by previous researchers. The model used is the algorithm of Support Vector Machines (SVM).
   c. Proposed method: To improve the accuracy of the algorithm of Support Vector Machines (SVM), then the addition of the improved method of optimization that combines PSO, GA and PCA.
   d. Experiment and Testing Methods: For experimental research data, researchers used Rapid Miner 5.3 to process the data and as an aid in assessing the accuracy of the data of experiments conducted in the study.
   e. Evaluation and Validation Results: The evaluation was conducted to determine the accuracy of the model algorithm Support Vector Machines. Validation is used to compare the results of the accuracy of the model used by the results that have been there before. Validation technique used is Cross Validation, accuracy algorithm will be measured using the Confusion Matrix and the processed data in the form of the ROC curve.
2. Data Collection

Researchers used cosmetic product review data collected from the site www.amazon.com. The data consists of 100 positive reviews and 100 negative reviews. Researchers download the data from http://www.amazon.com.

3. Data Initial Processing

For reducing the time of data processing, researchers only use 100 positive reviews and 100 negative reviews as data training. This dataset that in preprocessing should pass 3 steps, they are:

   a. Tokenization: collect all the words that appear and remove any punctuation or symbols that are not letters.
   b. Stopwords Removal: deletion of the words that are not relevant, such as the, of, for, with, and so on.
   c. Stemming: grouping words into several groups that have the same root. As for the phase transformation by TF-IDF weighting on each word. Where the process calculates the presence or absence of a word in the document.

4. Proposed Model

The method of research that the authors propose is the use of three (3) types of feature selection methods, namely Particle Swarm Optimization, Genetic Algorithm and Principal Component Analysis is used as a method of feature selection so that the accuracy of the classifier Support Vector Machine (SVM) can be increased. Researcher using Support Vector Machine classifier for a machine learning technique that is popular text classification, and has not performed well in many domains.

5. Evaluation and Validation Results

The model proposed in the study on cosmetic product review is by applying Support Vector Machines (SVM), Support Vector Machine (SVM) based Particle Swarm optimization (PSO), Support Vector Machine (SVM) based Genetic Algorithm (GA) and Support Vector Machine (SVM) based Genetic Algorithm (GA) & Principal Component Analysis (PCA). Application of SVM algorithm to determine the type of kernel first. Then determine the selection of parameter C and epsilon right. Having obtained the AUC values of accuracy and the greatest, that value will be the value that will be used to find the value of accuracy and the highest AUC.

IV. RESULT & DISCUSSION

A. *Text Classification Using Support Vector Machine Algorithm*

Training data used in this text classification consists of 100 cosmetic product reviews positive and 100 negative cosmetic product 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 in order to be classified in the next process, the following are the stages of the process:

   a. Data Collection: Positive review of data together in a folder with the name of the post. While the data is negative review unified storage in the folder with the name neg. Each extension.txt document that can be opened using Notepad application.
   b. Initial Data Processing: The process through which consists of tokenization, stopwords removal and stemming.
   c. Classification

   The classification process here is to determine a sentence as a member of a class of positive or negative class based on the calculation of the probability of a larger SVM formula. If the results of the probability of the sentence for a class positives outweigh the negatives class, then the sentence is included in the positive class. If the probability of a positive class is smaller than the negative, then the sentence is included in the negative class.

B. *Experimental Results Test Methods*

1. Support Vector Machine

Value of training cycles in this study was determined by testing inserting C, epsilon. The best results in the experiments above SVM is with C=0.1 and epsilon=0.1 resulting accuracy 82.00% and AUC 0.988.

2. Support Vector Machine Based on Particle Swarm Optimization, Support Vector Machine Based on Genetic Algorithm and Support Vector Machine Based on Genetic Algorithm & Principal Component Analysis

Value of training cycles in this study was determined by testing inserting C, epsilon and population size. Here are the results of the experiments have been conducted to determine the value of training cycles:

Table 1. Experimental Determination of Value Training Cycles SVM-Based PSO, SVM-Based GA, SVM-Based GA & PCA

C	Epsilon	*Population Size*	SVM		SVM+PSO		SVM+GA		SVM+ GA+PCA	
			Accuracy	AUC	*Accuracy*	AUC	*Accuracy*	AUC	*Accuracy*	AUC
0.0	**0.0**	**5**	80.50 %	0.980	82.50 %	0.985	84.50 %	0.943	**83.00 %**	**0.809**
0.1	**0.1**	**5**	**82.00 %**	**0.988**	94.00 %	0.989	93.00 %	0.975	82.50 %	0.800
0.2	**0.2**	**5**	81.50 %	0.987	**97.00 %**	**0.988**	93.00 %	0.980	82.50 %	0.801
0.3	**0.3**	**5**	81.00 %	0.987	95.00 %	0.991	94.00 %	0.975	81.00 %	0.799
0.4	**0.4**	**5**	80.50 %	0.987	96.00 %	0.991	**94.00 %**	**0.984**	80.50 %	0.802
0.5	**0.5**	**5**	80.00 %	0.985	96.50 %	0.984	93.50 %	0.985	81.75 %	0.800
0.6	**0.6**	**5**	80.00 %	0.985	96.50 %	0.994	93.00 %	0.985	81.25 %	0.805
0.7	**0.7**	**5**	80.00 %	0.985	94.00 %	0.985	93.50 %	0.987	80.75 %	0.805
0.8	**0.8**	**5**	80.00 %	0.985	96.50 %	0.995	93.50 %	0.984	80.00 %	0.801
0.9	**0.9**	**5**	80.00 %	0.985	95.00 %	0.994	93.00 %	0.972	81.00 %	0.800
0.0	**1.0**	**5**	50.00 %	0.500	50.00 %	0.500	50.00 %	0.500	82.00 %	0.800
1.0	**1.0**	**5**	50.00 %	0.500	50.00 %	0.500	50.00 %	0.500	50.00 %	0.500
1.0	**0.0**	**5**	81.00 %	0.987	96.00 %	0.984	93.50 %	0.987	82.75 %	0.807

The best results in the experiments above SVM-based PSO is with C=0.2 and Epsilon=0.2 and the population size=5 generated 97.00% accuracy and AUC 0.988. Then the best results in the experiments above SVM-based GA is with C=0.4 and Epsilon = 0.4 and the population size=5 generated 94.00% accuracy and AUC 0.984. While the best results in the experiments above SVM-based GA and PCA is with C=0.0 and Epsilon=0.0 and the population size=5 generated 83.00% accuracy and AUC 0.809.

C. *Results of Testing Model Support Vector Machine (SVM)*

Value of accuracy, precision and recall of training data can be calculated by using Rapid Miner. Test results using Support Vector Machine models showed in Table 2.

1. Confusion Matrix

Table 2 training data used consist of 100 positive reviews of data cosmetics and 100 negative review of data cosmetics. For data of the positive cosmetics reviews, 88 are classified into a positive review in accordance with the predictions made by the method of data predicted SVM and 12 positive reviews but it turns out the prediction results of negative reviews. For data negative reviews cosmetics, 76 negative reviews are classified according to the predictions made by the method of data predicted SVM and 24 negative reviews prediction result is positive review.

Table 2. Confusion Matrix Model to Support Vector Machine Method

***Accuracy*: 82.00%, +/- 8.12% (Micro: 82.00%)**			
	True Positive	**True Negative**	**Class Precission**
Predictions Positive	88	24	78.57%
Predictions Negative	12	76	86.36%
Class Recall	88.00%	76.00%	

Based on Table 2 shows that, the level of accuracy by using the SVM algorithm is equal to 82.00%, and conclusion The results of the calculation of the above equation shown in Table 3 below:

Table 3. Value Accuracy, Sensitivity, Specificity, PPV and NPV for Support Vector Machine Method

	Nilai %
Accuracy	**82.00%**
Sensitivity	**78.57%**
Specifity	**86.36%**
PPV	**88.00%**
NPV	**76.00%**

2. ROC Curve

The calculation result is visualized by ROC curve with the value of AUC (Area Under the Curve) is 0.988 in which diagnosis classification result was very good (excellent classification). To achieve accuracy AUC values close to 1 (perfect) needed a method to improve diagnosis classification results formed. In this case the researchers used the Particle Swarm Optimization (PSO), Genetic Algorithm (GA) and Principal Component Analysis (PCA) as the feature selection algorithm to improve the accuracy of classification.

D. *Results of Testing Model Support Vector Machine (SVM)-Based Particle Swarm Optimization (PSO)*

Accuracy value, precision and recall of training data can be calculated by using Rapid Miner. Test results using Support Vector Machine-based model of PSO is obtained in Table 4.

1. Confusion Matrix

Table 4 training data used consist of 100 positive reviews of data cosmetics and 100 negative review of data cosmetics. For data of the positive cosmetics reviews, 95 are classified into a positive review in accordance with the predictions made by the method of data predicted SVM and 5 positive reviews but it turns out the prediction results of negative reviews. For data negative reviews cosmetics, 99 negative reviews are classified according to the predictions made by the method of data predicted SVM and 1 negative reviews prediction result is positive review.

Table 4. Confusion Matrix Model to Support Vector Machine Method-Based PSO

Accuracy : 97.00%, +/- 3.32% (Mikro : 97.00%)			
	True Positive	True Negative	Class Precission
Predictions Positive	95	1	98.96%
Predictions Negative	5	99	95.19%
Class Recall	95.00%	99.00%	

Based on Table 4 shows that, the level of accuracy by using the SVM algorithm based PSO is equal to 97.00%, and conclusion the results of the calculation of the above equation shown in Table 5 below:

Table 5. Value Accuracy, Sensitivity, Specificity, PPV and NPV for Support Vector Machine Method-Based PSO

	Nilai %
Accuracy	97.00%
Sensitivity	98.96%
Specifity	95.19%
PPV	95.00%
NPV	99.00%

2. ROC Curve

The calculation result is visualized by ROC curve with the value of AUC (Area Under the Curve) is 0.988 in which the diagnosis result excellent classification. The result is still the same as just using SVM alone.

E. Results of Testing Model Support Vector Machine (SVM)-Based Genetic Algorithm (GA)

Value of accuracy, precision and recall of training data can be calculated by using Rapid Miner. Test results using Support Vector Machine-based model of GA is obtained in Table 6.

1. Confusion Matrix

Table 6 training data used consist of 100 positive reviews of data cosmetics and 100 negative review of data cosmetics. For data of the positive cosmetics reviews, 90 are classified into a positive review in accordance with the predictions made by the method of data predicted SVM and 10 positive reviews but it turns out the prediction results of negative reviews. For data negative reviews cosmetics, 98 negative reviews are classified according to the predictions made by the method of data predicted SVM and 2 negative reviews prediction result is positive review.

Table 6. Confusion Matrix Model to Support Vector Machine Method-Based GA

Accuracy : 94.00%, +/- 4.90% (Mikro : 94.00%)			
	True Positive	True Negative	Class Precission
Predictions Positive	90	2	97.83%
Predictions Negative	10	98	90.74%
Class Recall	90.00%	98.00%	

Based on Table 6 shows that, the level of accuracy by using the SVM algorithm based GA is equal to 94.00%, and conclusion the results of the calculation of the above equation shown in Table 7 below:

Table 7. Value Accuracy, Sensitivity, Specificity, PPV and NPV for Support Vector Machine Method-Based GA

	Nilai %
Accuracy	94.00%
Sensitivity	97.83%
Specifity	90.74%
PPV	90.00%
NPV	98.00%

2. ROC Curve

The calculation result is visualized by ROC curve with the value of AUC (Area Under the Curve) is 0.984 in which diagnosis classification result was very good (excellent classification).

F. Results of Testing Model Support Vector Machine (SVM)-Based Genetic Algorithm (GA) and Principal Component Analysis (PCA)

Value of accuracy, precision and recall of training data can be calculated by using Rapid Miner. Test results using Support Vector Machine-based GA & PCA is obtained in Table 8.

1. Confusion Matrix

Table 8 training data used consist of 100 positive reviews of data cosmetics and 100 negative review of data cosmetics. For data of the positive cosmetics reviews, 93 are classified into a positive review in accordance with the predictions made by the method of data predicted SVM and 7 positive reviews but it turns out the prediction results of negative reviews. For data negative reviews cosmetics, 73 negative reviews are classified according to the predictions made by the method of data predicted SVM and 27 negative reviews prediction result is positive review.

Table 8. Confusion Matrix Model to Support Vector Machine Method-Based GA and PCA

Accuracy : 83.00%, +/- 9.27% (Mikro : 83.00%)			
	True Positive	True Negative	Class Precission
Predictions Positive	93	27	77.50%
Predictions Negative	7	73	91.25%
Class Recall	93.00%	73.00%	

Based on Table 8 shows that, the level of accuracy by using the SVM algorithm based GA and PCA is equal to 83.00%, and conclusion the results of the calculation of the above equation shown in Table 9 below:

Table 9. Value Accuracy, Sensitivity, Specificity, PPV and NPV for Support Vector Machine Method-Based GA and PCA

	Nilai %
Accuracy	83.00%
Sensitivity	77.75%
Specifity	73.00%
PPV	93.00%
NPV	73.00%

2. ROC Curve

The calculation result is visualized by ROC curve with the value of AUC (Area Under the Curve) is 0.809 in which the diagnosis result good classification.

G. Evaluation and Validation Analysis Model

From the above test results, the measurement accuracy using the confusion matrix and the ROC curve proved that the test results based SVM algorithm PSO has a value higher accuracy compared with SVM algorithm. Values for the accuracy of SVM algorithm model of 82.00%, the value of accuracy for SVM-based PSO algorithm model by 97.00%, with the difference in accuracy with its own SVM 8.00%. Then the accuracy of the model-based SVM algorithm GA at 94.00%, with the difference in accuracy with its own SVM 5.00%. While the value of accuracy for the model-based SVM algorithm GA and PCA amounted to 83.00%, with the difference in accuracy with its own SVM 1.00%. Can be seen in Table 10 below:

Table 10. Testing Algorithm SVM, SVM-based PSO, GA-based SVM and SVM-based GA & PCA

	Successful Classification of Possitive Review	Successful Classification of Negative Review	Accuracy	AUC
SVM	**88**	**76**	**82.00%**	**0.988**
SVM –Based PSO	**95**	**99**	**97.00%**	**0.988**
SVM-Based GA	**90**	**98**	**94.00%**	**0.984**
SVM-Based GA & PCA	**93**	**73**	**83.00%**	**0.809**

For the evaluation using the curve ROC resulting value of AUC (Area Under the Curve) to the model algorithm SVM generate value 0.988 with a value of diagnosis Excellent Classification, then to the algorithm SVM-based PSO (Particle Swarm Optimization) to the accuracy of the total of 97.00% turned out to produce AUC values similar with SVM algorithm model (fixed) that is equal to 0.988 with the diagnosis Excellent value Classification, and the difference in value both at 0.00. Then for the SVM algorithm based on GA (Genetic Algorithm) with the greatest accuracy is 94.00% turned out to produce AUC values smaller than both the 0984. As for the SVM algorithm based on GA (Genetic Algorithm) and PCA (Principal Component Analysis) with the greatest accuracy is 83.00% turned out to produce AUC values were smaller than the others, namely 0.809. Thus the PSO-based SVM algorithm can provide solutions to the problems in the classification review of cosmetic products.

V. Conclusion

In this study tested the model using Support Vector Machine and Support Vector Machine-based Particle Swarm Optimization using review data of cosmetic products with an overall positive or negative 200 review data. The resulting model was tested to get the value of accuracy, precision, recall and AUC of each algorithm to obtain the test by using Support Vector Machine value obtained accuracy is 82.00%. Later the testing using Support Vector Machine-based Particle Swarm Optimization (PSO) values obtained 97.00% accuracy. Later testing by using Support Vector Machine-based Genetic Algorithm (GA) values obtained 94.00% accuracy. While the testing by using Support Vector Machine-based Genetic Algorithm (GA) and Principal Component Analysis (PCA) values obtained 83.00% accuracy. It can be concluded that the testing of cosmetic product review data using Support Vector Machine-based Particle Swarm Optimization (PSO) is better than on Support Vector Machine-based Genetic Algorithm (GA), Principal Component Analysis (PCA) and Support Vector Machine itself.

Thus the results of testing the model above it can be concluded that Support Vector Machine-based Particle Swarm Optimization provide solutions to review problems of cosmetic products classification become more accurate.

References

[1] A. Samad, H. Basari, B. Hussin, I. G. Pramudya, and J. Zeniarja, "Opinion Mining of Movie Review using Hybrid Method of Support Vector Machine and Particle Swarm Optimization," *Procedia Eng.*, vol. 53, pp. 453–462, 2013.

[2] C. Vercellis, "Business Intelligence: Data Mining and Optimization for Decision Making." United Kingdom: A John Wiley And Sons, Ltd.,Publication, 2009.

[3] E. Haddi, E. Haddi, X. Liu, and Y. Shi, "The Role of Text Pre-processing in Sentiment Analysis The Role of Text Pre-processing in Sentiment Analysis," Procedia Comput. Sci., vol. 17, no. December, pp. 26–32, 2013.

[4] I. Habernal, T. Ptáček, and J. Steinberger, "Sentiment Analysis in Czech Social Media Using Supervised Machine Learning," 50(5), 693–707, 2014.

[5] I. H. Witten, E. Frank, and M.A. Hall, "Data Mining Practical Machine Learning Tools And Technique", Burlington: Elsevier Inc, 2011.

[6] J. Han and M. Kamber, "Data Mining Concepts and Techniques," San Francisco: Diane Cerra, 2007.

[7] J. S. Chou, M. Y. Cheng, Y. W. Wu and A. D. Pham, "Optimizing parameters of support vector machine using fast messy genetic algorithm for dispute classification," Expert Syst. Appl., vol. 41, no. 8, pp. 3955–3964, 2014.

[8] K. Huang, H. Yang, I. King, and M. Lyu, "Machine Learning Modeling Data Locally And Globally," Berlin Heidelberg: Zhejiang University Press, Hangzhou And Springer-Verlag Gmbh, 2008.

[9] K. S. Sodhi and M. Lal, "Face Recognition Using PCA, LDA and Various Distance Classifier," Journal of Global Research in Computer Science, Vol. 4, 2013, pp. 30-35.

[10] L. Zhang, K. Hua, H. Wang, G. Qian, and L. Zhang, "Sentiment Analysis on Reviews of Mobile Users," *Procedia - Procedia Comput. Sci.*, vol. 34, pp. 458–465, 2014.

[11] M. Wahyudi and D. A. Kristiyanti, "Sentiment Analysis of Smartphone Product Review using Support Vector Machine Algorithm-Based Particle Swarm Optimization," vol. 91, no. 1, pp. 189–201, 2016.

[12] M. Wahyudi and D. W. I. A. Putri, "Algorithm Application Support Vector Machine with Genetic Algorithm Optimization Technique for Selection Features for The Analysis of Sentiment on Twitter," vol. 84, no. 3, 2016.

[13] M. Zhao, C. Fu, L. Ji, K. Tang, and M. Zhou, "Expert Systems with Applications Feature selection and parameter optimization for support vector machines : A new approach based on genetic algorithm with feature chromosomes," Expert Syst. Appl., vol. 38, no. 5, pp. 5197–5204, 2011.

[14] Q. Ye, Z. Zhang and R. Law, "Sentiment classification of online reviews to travel destinations by supervised machine learning approaches". Expert Systems with Applications, 36(3), 6527–6535, 2009.

[15] R. Dehkharghani., H. Mercan, A. Javeed and Y. Saygin, "Sentimental causal rule discovery from Twitter. Expert Systems with Applications," 41(10), 4950–4958, 2014.

[16] R. L. Haupt and S. E. Haupt, "Practical Genetic Algorithms," Untied States Of America: A John Wiley & Sons Inc Publication, 2004.

[17] R. Moraes, W. P. G. Neto, R. Moraes, J. F. Valiati, and W. P. G. Neto, "Document-level sentiment classification : An empirical comparison between SVM and ANN Article in Expert Systems with Applications • February 2013," Expert Syst. Appl., no. February, 2013.

[18] S. Karamizadeh, S. M. Abdullah, A. A. Manaf, M. Zamani, and A. Hooman, "An Overview of Principal Component Analysis," vol. 2013, no. August, pp. 173–175, 2013.

[19] S. M. Weiss, N. Indurkhya and T. Zhang, "Fundamentals of Predictive Text Mining", London: Springer-Verlag, 2010.

[20] W. Medhat, A. Hassan, and H. Korashy, "Sentiment analysis algorithms and applications : A survey," Ain Shams Eng. J., vol. 5, no. 4, pp. 1093–1113, 2014.

[21] Y. Liu, G. Wang, H. Chen, H. Dong, X. Zhu, and S. Wang, "An Improved Particle Swarm Optimization for Feature Selection," vol. 8, 191–200, 2011.

[22] Z. Zhang, Q. Ye, Z. Zhang and Y. Li, "Sentiment classification of Internet restaurant reviews written in Cantonese", Expert Systems with Applications, 38(6), 7674–7682, 2011.

Feature Selection Based on Genetic Algorithm, Particle Swarm Optimization and Principal Component Analysis for Opinion Mining Cosmetic Product Review

ORIGINALITY REPORT

93%	66%	93%	49%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Dinar Ajeng Kristiyanti, Mochamad Wahyudi. "Feature selection based on Genetic algorithm, particle swarm optimization and principal component analysis for opinion mining cosmetic product review", 2017 5th International Conference on Cyber and IT Service Management (CITSM), 2017 Publication	93%

Exclude quotes	Off	Exclude matches	Off
Exclude bibliography	Off		