Journal of Information Technology and Computer Science
Volume 3, Number 2, 2018, pp. 194-201
Journal Homepage: www.jitecs.ub.ac.id

High Performance of Polynomial Kernel at SVM Algorithm for Sentiment Analysis

Lailil Muflikhah[1], Dimas Joko Haryanto[2], Arief Andy Soebroto[3], Edy Santoso[4]

Department of Computer Science
Brawijaya University
Malang, Indonesia
[1]lailil@ub.ac.id, [2]dimasjokoh@gmail.com, [3]ariefas@ub.ac.id, [4]edy144@ub.ac.id

Received 17 August 2018; accepted 5 November 2018

Abstract. Sentiment analysis is a text mining based on the opinion collection towards the review of online product. Support Vector Machine (SVM) is an algorithm of classification that applicable to review the analysis of product. The hyperplane kernel function of SVM has importance role to classify the certain category. Therefore, this research is address to investigate the performance between Polynomial and Radial Basis Function (RBF) kernel functions for sentiment analysis of review product. They are examined to 200 comments using 10-fold validation and various parameter values (learning rate, lambda, c value, epsilon and iteration). As general, the performance for polynomial kernel of 88.75% is slightly higher than RBF kernel of 83.25%.

Keywords—sentiment analysis, SVM, kernel, RBF, polynomial, performance

1 Introduction

Recently, online shopping is issue trend for customer which has been growing fast. It is virtual transaction that is very easy and unlimited sale or purchase. It can be accessed through internet application anytime and anywhere. However, the weakness of this transaction is unknown the quality of product and credibility of seller. Therefore, consumers get those information through navigating online product reviews or customer's feedback before making decision to purchase the online product. The volume of reviews is increase continuously so that it is time-consuming to get the related information. Sentiment analysis is a method of classification that is addressed for opinion mining to study the review of the online product. It can give information positive, negative or neutral based on the reviews collection.

There are many classification methods that is applicable to text or opinion mining in many domain, such as K-Nearest Neighbor, Naïve Bayes, Deep Learning, and Support Vector Machine (SVM). The SVM is relative robust algorithm of the performance. It is not sensitive to the number of training data and proportional data in each class. The time and space computation is consuming relative low [1]. This method is also implemented to several recommendation domains, including method development such as involving AHP and TOPSIS method for selection and recommendation scholarship as is conducted by Putra, et.al. [2]. Also, the development of SVM algorithm in forecasting (SVR) is applied to estimate software effort [3]. However, there are various parameters involved to determine the quality of the output data for classification including hyperplane kernel function. The kernel function can

determine the class label from spread data collection based on the provided training data.

In the previous research, the study comparison of kernel function such as RBF, linear and polynomial has been conducted and as a result showed that the RBF kernel function has the highest performance for text or image document categorization[4][5]. However, classification of sentiment is a special case of document categorization with two classes such as positive and negative. In this kind classification, it is involved to analyze sentiment for comment of the product review. The other research on classification document on software review has been applied using ontology approach in order to reduce the dimension and it has been combined to SVM algorithm. The research output is the product detail, not to analyze sentiment review with positive or negative [6]. Therefore, this research is purposed to investigate the performance between Polynomial and RBF Kernel for sentiment analysis in Indonesian product review.

2 Proposed Method

There are three main steps in this research as shown in Fig. 1. The first step is preprocessing data (training and testing set) including tokenization, stop word removal, stemming and normalization of informal language. Then, next step is feature representation using TF-IDF weighting term. The last step is applied to classification method using two kernel functions for comparison such as Radial Basis Function (RBF) and Polynomial – 2nd degree. This research is addressed to know both performance of the different kernel function.

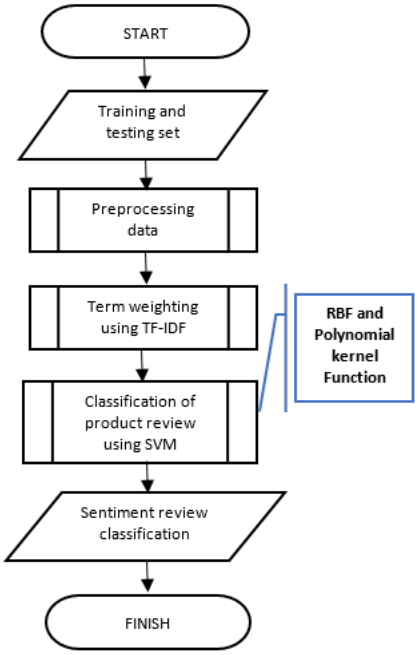

Fig. 1. Block Diagram of General System

2.1 Preprocessing Data

The first step of this research is preprocessing data. It involves tokenization, stop words removal, stemming and normalization including query expansion. Tokenization is a

process removing punctuation, numbers, and characters other than the alphabet [7]. It is also conducted case folding, which is changing all capital letters into lowercase. Then, stop words removal or filtering is removing uninformative words referring to the existing stop word dictionary. Meanwhile, stemming is a process to convert every words to its root. This process is done by removing affixes such as prefix, infix and suffix. Normalization is applied to change the words into their formal form, such as the word "ga" become "tidak" and the word "bisaaaa" become "bisa".

2.2 Term Weighting

Term weighting is a feature representation of text document. It is an important aspect of modern text retrieval systems [8]. Terms are words, phrases, or any other indexing units used to identify the contents of a text. Since different terms have different importance in a text, an important indicator - the term weight - is associated with every term [9].

TF-IDF term weighting is a statistical measure used to evaluate how important a word is to a document in a collection or corpus. The importance increases proportionally to the number of times a word appears in the document but is offset by the frequency of the word in the corpus. Typically, the tf – idf weight is composed by two terms: the first computes the normalized Term Frequency (TF), divided by the total number of words in that document; the second term is the Inverse Document Frequency (IDF). It is computed as the logarithm of the number of the documents in the corpus divided by the number of documents where the specific term appears. Given a document collection D, a word w, and an individual document d ϵ D, we calculate as in Equation (1).

$$w_d = f_{w,d} * \log(\frac{|D|}{f_{w,D}}) \quad (1)$$

where f_(w,d) equals the number of times w appears in d, |D| is the size of the corpus, and f_(w,D)equals the number of documents in which w appears in D [10][11].

2.3 Sentiment Analysis

Sentiment analysis is one of the prominent fields of data mining that deals with the identification and analysis of sentimental contents generally available at social media [12]. In the sentiment analysis, the raw data is the online text that is exchanged by users through social media. Shopping online is a social media which provides the forum to give feedback from customer of product and service. We implement SVM method due to the highest performance for sentiment analysis problem [13][14].

2.4 Classification using Support Vector Machine (SVM)

Classification is a supervised method in machine learning-based approach. Basically, this method consists of two processes. The first is to construct a classification model by learning on a training corpus with previously labeled classes, i.e. positive and negative. The second is to apply the obtained model to classify documents that were not used in the construction of the classifier. Support Vector Machine (SVM) is used in this research due to the most robust algorithm. It represents documents as points in a vector space, which dimensions are selected features. The basic concept of SVM is to find the optimal hyperplane that separates the previously classified data with the largest margin of separation between the two classes as shown in Fig.2.

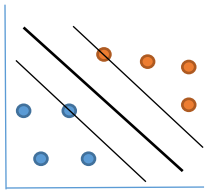

Fig. 2. Support Vector Machine

2.4.1 Kernal Function

In the classification process, the data is spread of information, so that SVM introduces the kernel function [15], K(xn,xi), which transforms the original data space into a new space with a higher dimension; this process includes the transformation function with dot product Φ(x) as in Equation (2). The aim is the data, which already transformed into a higher dimension, can be separated easily. Thus the hyperplane function can be written in Equation (3).

$$K(x_i.x_j) = \Phi(x_i).\Phi(x_j) \quad (2)$$

$$f(x_i) = \sum_{n=1}^{N} \propto_n y_n K(x_n, x_i) + b \quad (3)$$

In this study, we investigate the comparison of using two kernel functions of SVM algorithm, such as Radial Basis Function (RBF) and 2nd degree of polynomial function. The detail formula is shown in Table 1.

TABLE 1. TWO COMMON KERNELS FUNCTIONS

Kernel Function	Formula
RBF	$K(X_i \cdot X_j) = \exp\left(-\left(\frac{\|X_i - X_j\|^2}{2\sigma^2}\right)\right)$
Polynomial	$K(X_i \cdot X_j) = (x_i.x_j + 1)^P$

2.4.2 Sequential Support Vector Machine

The sequential method is modified SVM method to provide optimal solutions and to speed up the iteration process rather than using conventional solutions [16]. The steps of sequential method are as follows:

1. Initialize parameter λ (*lambda*), γ (*learning rate*), *C complexity*), ε (*epsilon*) and maximum iteration.
2. Initialize $\alpha_i = 0$, and compute matrix *D* for *i*, *j*=1,2,…,n
   $Dij = y_i y_j\left(K(x_1, x_j) + \lambda^2\right)$
3. Then, for each pattern, i=1,2,,n, compute:
   a. $E_i = \sum_{j=1}^{n} \alpha_j D_{ij}$
   b. $\delta\alpha_i = \min\{\max[\gamma(1 - E_i), -\alpha_i], C - \alpha_i\}$
   c. $\alpha_i = \alpha_i + \delta\alpha_i$
4. The iteration will be stopped, if it is achieved maximum iteration or $Max(|\delta\alpha|) < \varepsilon$, else go to step 2.

After the above process finished, then it will be obtained the α value and *Support Vector*. So that, the formula of sentiment analysis in this research is as Equation (4).

$$f(x) = \sum_{i=0}^{n} \alpha_i y_i K(x, x_i) + b \quad (4)$$

Where b,,bias value is

$b = -\frac{1}{2}(\sum_{i=0}^{n} \alpha_i y_i K(x_i, x^-) + \sum_{i=0}^{n} \alpha_i y_i K(x_i, x^+))$.

3 Result and Analysis

The data set used to implement the classification method consists of 400 comments. They are taken from tokopedia.com with 200 positive comments and 200 negative comments. Each experiment is used 360 training sets and 40 testing sets by 10-fold cross validation. In order to know the accuracy rate, there are four testing scenarios of SVM parameter including learning rate (γ), lambda (λ), complexity (C) and epsilon (ε).

The first tested parameter value is learning rate (γ) of training process. It effects to accuracy result on the both term representation as shown in Fig. 3. The best accuracy is obtained at γ value =0.0001 for the both kernel function (82.75% of RBF and 82.25% of polynomial). The accuracy is decreased when learning rate is too high. It is used for the calculation of δα to stop iteration conditions. The low value of δ will cause the value of Max (| δα |) to be less than ε. If the value of Max (| δα |) is below of ε, then iteration has stopped due to the value of α has converged.

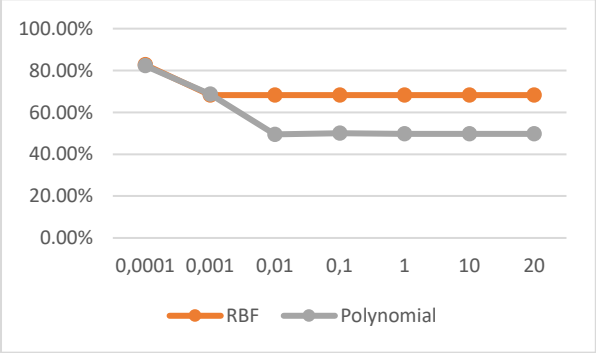

Fig. 3. Accuracy of testing result in different learning rate (γ)

Next parameter is lambda (λ). This is regularization parameter which provide a degree of miss-classification. It looks for maximizing margin between both classes and minimizing miss-classification. The testing result as on Fig. 4 shows that the best parameter lambda (λ) value that can achieves the highest accuracy for RBF of 83.25% at λ =3 and for polynomial of 86.75% at λ =4 . The accuracy of polynomial kernel function is slightly higher than the RBF kernel. This value involves to calculate the Hessian Matrix. This effects to the speed to reach convergence in the learning process.

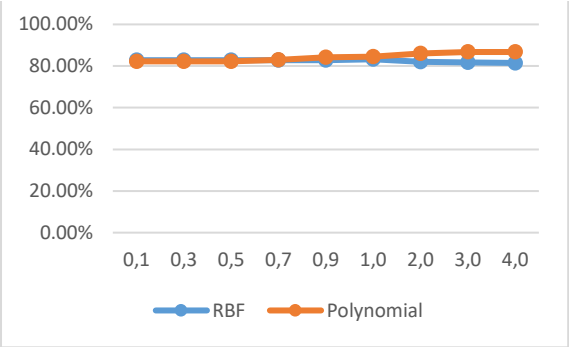

Fig. 4. Accuracy of testing result in different *lambda* (λ)

Complexity factor (C) is also effect on the accuracy rate as shown at Fig. 5 this coefficient is affect to trade-off between complexity and proportion of no separable samples. If it too large, then it has high penalty of no separable data and perhaps, it is overfitting. Otherwise, it may has been under fitting. Based on experiment result that the best parameter C value for the both kernel is stable starting at C=0.01 of 83.25% (RBF kernel function) and 86.75% (Polynomial kernel function). This is due to the complexity is involved on the calculation of δα which influences to search on support vector data and computation time of this opinion analysis process.

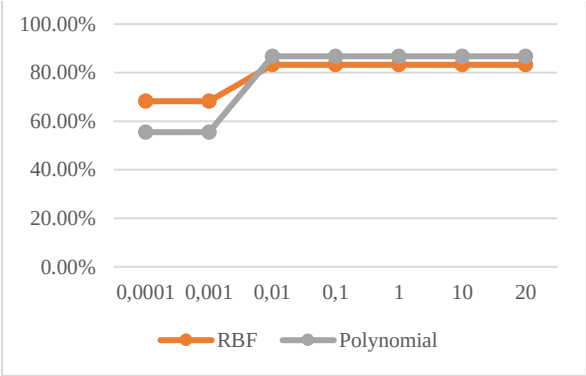

Figure 5. Accuracy of testing result in different C value

Then, another parameter is epsilon (ε). This parameter is used to fit the training data. It is impact to the number of support vector which used to construct the regression function. If the value of ε (epsilon) is too high then the accuracy result is low due to an early convergence. It means that iteration will stop when the α value obtained is not optimal.In this research, the best accuracy rate is obtained at ε= 0.0001 (83.25% for RBF kernel and 86.75% for Polynomial kernel) as it is shown at Fig.6.

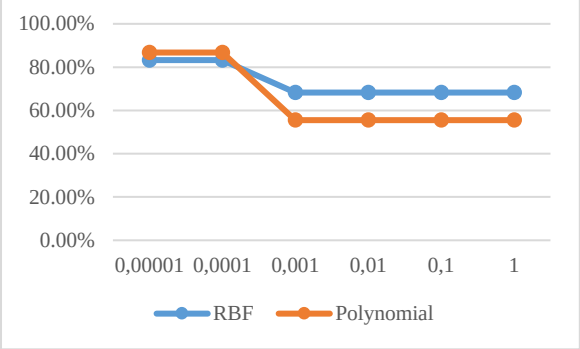

Fig. 6. Accuracy of testing result in different epsilon (ε)

The latest parameter value is iteration for learning process of training data. The best performance is at iteration = 50 of 83.25% (RBF) and at iteration =100 of 88.75% (polynomial) as at Fig.7.

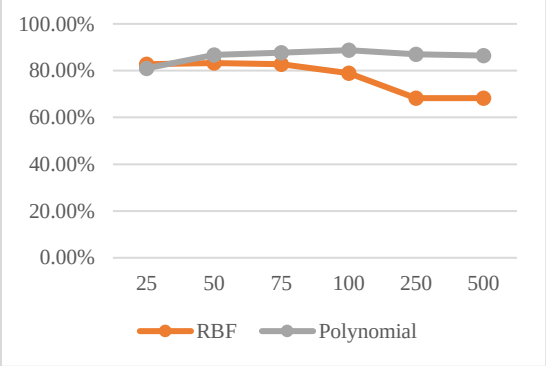

Fig. 7. Accuracy of testing result in different number of iteration

Finally, the both kernel functions are implemented using the best parameter value i.e. RBF kernel (gamma=0.0001; lambda=1; c=0.01; epsilon=0.00001; iteration=50) and polynomial kernel (gamma=0.0001; lambda-3l c=0.01; epsilon=0.00001; iteration=100). And as a result that the accuracy rate for polynomial of 88.75% is higher than the RBF kernel of 83.25%. There is difference of 5.5% as shown in Fig. 8.

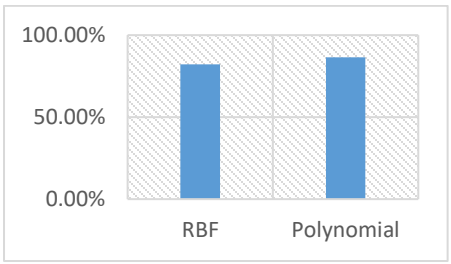

Fig. 8. Comparison of testing result at the best parameter value

4. Conclusion

Sentiment analysis of review shopping online can be applied by SVM algorithm with kernel RBF or polynomial functions. However, the polynomial kernel function has slightly higher performance than the RBF kernel. At the optimal parameter values, the accuracy of polynomial kernel is obtained of 88.75% as the RBF kernel of 83.25%.

References

1. D. Meyer, F. Leisch, and K. Hornik: The support vector machine under test. Neurocomputing, vol. 55, no. 1-2, pp. 169–186 (2003).
2. Putra, M.Gilvy L., Ariyanti, W., Cholissodin, I.: Selection and Recommendation Sholarship Using AHP-SVM-TOPSIS. Journal of Information Technology and Computer Science (JITECS) Vol 1, no.1, pp 1-13 (2016)
3. Novirasari, D., Cholissodin, I., Mahmudy, Wayan F.: Optimizing SVR using Local Best PSO for Software Effort Estimation. . Journal of Information Technology and Computer Science (JITECS) Vol 1, no.1, pp 28-37 (2016)
4. Kaur, G. and Kaur, E.P. " Novel approach to text classification by SVM-RBF kernel and linear SVC". International Journal of Advance Research, Ideas and Innovation in Technology. Vol.3. No.3 (2017)
5. Yekkekhany, G., Safari, A., Homayouni, S., and Hasanlou, M.: A Comparison study of different kernel functions for SVM-based classification of multi-temporal polametry SAR data. In 1st ISPRS International Conference on Geospatial Information Research. International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume XL-2W3 (2014)
6. Khin Phyu Phyu Shein, Thi Thi Soe Nyunt: Sentiment Classification based on Ontology and SVM Classifier. Second International Conference on Communication Software and Networks (2010).
7. Fauzi, M.A., Arifin, A.Z. and Yuniarti A. : Arabic Book Retrieval using Class and Book Index Based Term Weighting. International Journal of Electrical and Computer Engineering (IJECE). Vol. 7, No. 6 (2017)
8. Chris Buckley: The importance of proper weighting methods. In M. Bates, editor, Human Language Technology. Morgan Kaufman (1993).
9. Gerald Salton and Chris Buckley: Term weighting approaches in automatic text retrieval. Information Processing and Management, 24(5):513-523, Issue 5 (1988).
10. Salton, G. & Buckley, C.: Term-weighting approaches in automatic text retrieval. In Information Processing & Management, 24(5): 513-523 (1988).
11. Zachary G. Ives: Information Retrieval. University of Pennsylvania, CSE 455 - Internet and Web Systems (2007)
12. Basari, A.S.H., Hussin, B., Ananta, I.G.P., Zeniarja, J.: Opinion Mining of Movie Review using Hybrid Method of Support Vector Machine and Particle Swarm Optimization. In Procedia Engineering 53 pp.453-462 (2013)
13. Dave K,, Lawrence S,, and Pen-nock D: Opinion extraction and semantic classification of product reviews.Mining the peanut gallery: 2003.
14. Pang B., and Lee L. : Opinion Mining and Sentiment Analysis. Foundations and Trends in Information Retrieval 2-1 (2008).
15. Berger, A et al: Bridging the Lexical Chasm: Statistical Approaches to Answer Finding. In Proc. Int. Conf. Research and Development in Information Retrieval, 192-199(2000).
16. Vijayakumar, S., Wu, S.: Sequential Support Vector Classification and Regression. In Proceeding of International Conference on Soft Computing (1999).