International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2019 IJSRCSEIT | Volume 5 | Issue 6 | ISSN : 2456-3307
DOI : https://doi.org/10.32628/CSEIT195628

Comparison of Random Forest and Support Vector Machine for Indonesian Tweet Complaint Classification

Desi Ramayanti
Faculty of Computer Science, Universitas Mercu Buana, Jakarta Barat, Indonesia

ABSTRACT

In digital business, the managerial commonly need to process text so that it can be used to support decision-making. The number of text documents contained ideas and opinions is progressing and challenging to understand one by one. Whereas if the data are processed and correctly rendered using machine learning, it can present a general overview of a particular case, organization, or object quickly. Numerous researches have been accomplished in this research area, nevertheless, most of the studies concentrated on English text classification. Every language has various techniques or methods to classify text depending on the characteristics of its grammar. The result of classification among languages may be different even though it used the same algorithm. Given the greatness of text classification, text classification algorithms that can be implemented is the support vector machine (SVM) and Random Forest (RF). Based on the background above, this research is aimed to find out the performance of support vector machine algorithm and random forest in classification of Indonesian text. 1. Result of SVM classifier with cross validation k-10 is derived the best accuracy with value 0.9648, however, it spends computational time as long as 40.118 second. Then, result of RF classifier with values, i.e. 'bootstrap': False, 'min_samples_leaf': 1, 'n_estimators': 10, 'min_samples_split': 3, 'criterion': 'entropy', 'max_features': 3, 'max_depth': None is achieved accuracy is 0.9561 and computational time 109.399 second.

Keywords : Complaint Classification, Indonesian Text, Support Vector Machine, Random Forest

I. INTRODUCTION

Texts on an appropriate topic are straightforward to discover, especially in the current era of technology, either through social media or websites. The number of text documents contained ideas and opinions is progressing and challenging to understand one by one. Whereas if the data are processed and correctly rendered using machine learning, it can present a general overview of a particular case, organization, or object quickly [1]–[5].

The subject of machine learning has frequently become well-known in recent years to treat data [1], [2]. It also supports by increasing the demand for machine learning for text mining to processing text data quickly. In digital business, the managerial commonly need to process text so that it can be used to support decision-making.

Text mining relevant to text classification, which represents the task of tag text documents to one or more classes based on a keyword. [3]. Text classification has grown one of the key methods for managing text data. Text classification is applied to analyse text data, for example, report or 'tweet' from social media, to obtain any information that can be utilized by the data owner. Since the old-fashioned

method is time-consuming and difficult, numerous researches have been accomplished in this research area [4]. Nevertheless, most of the studies concentrated on English text classification [5]. Every language has various techniques or methods to classify text depending on the characteristics of its grammar. The result of classification among languages may be different even though it used the same algorithm. Given the greatness of text classification, text classification algorithms that can be implemented is the support vector machine (SVM) and Random Forest (RF).

The distinction of SVM utilization for solving some problems can be shown from the number of papers that practiced this algorithm. For examples, in the material construction research field, study by [6] used SVM to build a model of strength level of lightweight foamed concrete material. Then, research by [7] applied SVM for gesture phase segmentation. In 2017, Ghaddar and Naoum-sawaya (2017) completed research about high dimensional data classification and feature selection using SVM [8]. In other research domain, SVM is still being applied to solve existing problems [9]–[12]. Moreover, Random forest is an algorithm based on bagging and random subspace which is consisted of multi-way or binary decision trees. Random forest (RF) algorithm consisted of two procedures. The first procedure is training sets are designed and constructed using random bootstrap method with replacement [13][14].

Based on the background above, this research titled Comparison of Random Forest and Support Vector Machine for Indonesian Tweet Complaint Classification is aimed to find out the performance of support vector machine algorithm and random forest in classification of Indonesian text.

II. LITERATURE REVIEW

A. Support Vector Machine

The similar works of SVM utilization for resolving some obstacles are completed by many researchers, for example, in the material construction research field, a study by [6] used SVM to build a model of strength level of lightweight foamed concrete material. Then, research by [7] employeed SVM for *gesture phase* segmentation. In 2017, [8] conducted research about *high dimensional* data classification and feature selection using SVM. In other research domain, SVM is still being used to solve existing problems [9]–[12].

B. Random Forest

Relevance work of random forest algorithm included Bosch, Zisserman, and Muoz (2007); Schroff, Criminisi, and Zisserman (2008); Kuznetsova, Leal-Taixé, and Rosenhahn (2013); Shotton et al., (2013); Joshi, Monnier, Betke, and Sclaroff (2017) has been used as references for this research [15]–[19].

Research by Bosch, Zisserman, and Muoz (2007) conducted images classification using random forest algorithm and SVM algorithm. The research attempted to compare the result of image classification between those algorithms. [15]. Joshi, Monnier, Betke, and Sclaroff (2017) completed research work about random forest algorithm by implementing it into gesture recognition system [18].

Kuznetsova, Leal-Taixé, and Rosenhahn (2013) conducted work in human computer interaction and computer vision research field, especially hand gesture recognition [16]. Shotton et al., (2013) proposed an approach by utilizing random forest algorithm. This research attempted to use single depth image without temporal information for predicting 3D positions of body joints [17]. Schroff, Criminisi, and Zisserman (2008) observed performance of random forest in pixel-wise images segmentation [19].

III. METHODOLOGY

This research will be completed through four phases, i.e. data acquisition, data preprocess, classification and comparison, as depicted in below.

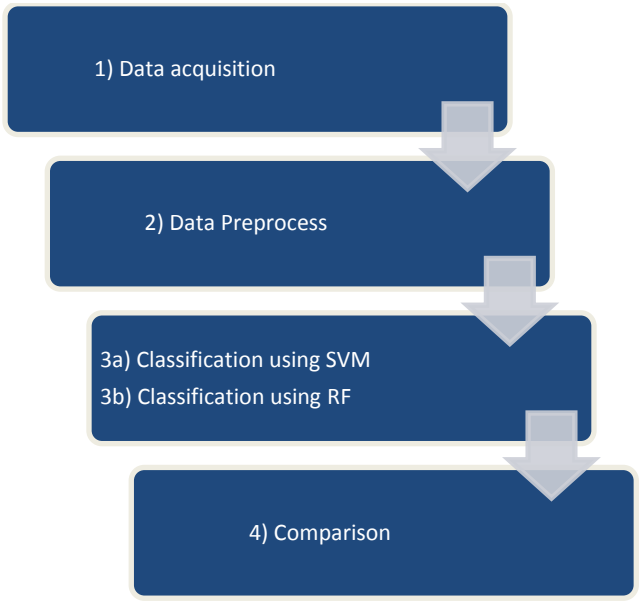

Figure 1. Research methodology

The flow of proposed research phase is elaborated below:

1. Data acquisition is completed by extracting the tweet data using the program script in R via the Twitter API. The data obtained for this research as many as 1170 tweets, which contained about the complaint and non-complaint issues.
2. Data preprocess has been done by conducting sub-phases including data cleansing, data labeling, case folding, special character removing, stop word removing.
3. Classification is done by using SVM classifier and RF classifier.
4. Comparison of the algorithm results are evaluated based on its accuracy and computational time.

IV. EXPERIMENTAL RESULT

A. Support Vector Machine

This research used Python programming language and scikit-learn library and R-library. The pre-processing stage is achieved by using TfidfVectorizer, one of feature extraction functions in sklearn library. Classification is completed using the Support Vector Machine in sklearn library. The validation is done using cross-validation, where the percentage of training sample 70% and testing sample 30%, respectively.

The performance for SVM classification for experiment with cross-validation and tuning parameter is shown in Table 1. In this experiment, we conducted cross-validation with k-5 and k-10 and implemented the tuning parameter of SVM with C constant and gamma value.

TABLE I. CROSS VALIDATION

Variable	k-5	k-10
C constant	32.0	128.0
gamma	0.000122	3.0517578125e-05
accuracy	0.9507	0.9648
time	18.976 s	40.118 s

Based on the Table above, C constant and gamma is obtained a better result for cross-validation. The result of cross-validation with k-10 is derived the best accuracy with value 0.9648; however, it spends computational time as many 40.118 seconds. Then, we experimented to find the best kernel function among Sigmoid, Linear, and RBF. The result of the experiment can be seen in Table V with detail of accuracy and computational time for each kernel function. Moreover, based on the result of the investigation, kernel function Sigmoid achieved the best accuracy and computational time.

TABLE II. ACCURACY AND TIME

Kernel	Accuracy	Time
Sigmoid	0.9683	30.90
Linear	0.9666	35.36
RBF	0.9648	40.67

B. Random Forest

In this study, we practiced Python (python-sklearn) for running and processing datasets that gathered on Twitter using R script based on Twitter API. The tuning parameter process in scikit-learn used GridSearchCV. we used k-10 cross validation with the percentage of training dataset are 70% and the testing dataset are 30% in this experiment. The classifier used in this research is Random Forest Classifier. Moreover, we tuned seven parameters on RF classifier, which are elaborated as follows:

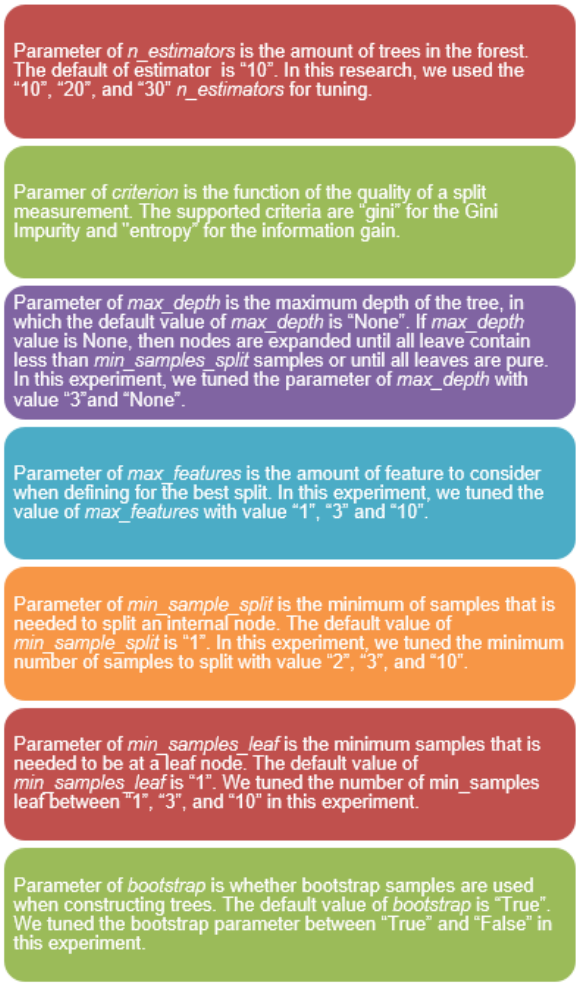

Fig. 2 Paramater for RF classifier

For short, the result of the best value for each parameter which are presented in the Table 1. The best score that achieved in this experiment is 0.956 and computational time required to tune parameters is 109.399434 second.

TABLE III

THE BEST VALUE OF EACH PARAMETER

Parameter	Best value of parameter
max_depth	None
max_features	3
criterion	entropy
min_samples_split	3
n_estimators	10
min_samples_leaf	1
bootstrap	False

V. CONCLUSION

We have conducted research to classify Indonesian text and conclude some results:

1. Result of SVM classifier with cross validation k-10 is derived the best accuracy with value 0.9648, however, it spends computational time as long as 40.118 second. Then, we conducted experiment to find the best kernel function among Sigmoid, Linear and RBF. Moreover, based on result of experiment, kernel function Sigmoid achieved the best accuracy and computational time.
2. Result of RF classifier with values, i.e. 'bootstrap': False, 'min_samples_leaf': 1, 'n_estimators': 10, 'min_samples_split': 3, 'criterion': 'entropy', 'max_features': 3, 'max_depth': None is achieved accuracy is 0.9561 and computational time 109.399 second.

VI. ACKNOWLEDGEMENT

This research has been funded by an internal research grant (named penelitian internal) from Universitas Mercu Buana.

VII. REFERENCES

[1]. W. P. Sari, E. Cahyaningsih, D. I. Sensuse, and H. Noprisson, "The welfare classification of Indonesian national civil servant using TOPSIS and k-Nearest Neighbour (KNN)," in Research and Development (SCOReD), 2016 IEEE Student Conference on, 2016, pp. 1-5.

[2]. V. Ayumi, "Pose-based Human Action Recognition with Extreme Gradient Boosting," 2016.

[3]. J. Dai and X. Liu, "Approach for Text Classification Based on the Similarity Measurement between Normal Cloud Models," Sci. World J., 2014.

[4]. T. Joachims, "Text categorization with support vector machines: Learning with many relevant features," in In Proceedings of the 10th European Conference on Machine Learning.

[5]. N. Boudad, R. Faizi, R. O. H. Thami, and R. Chiheb, "Sentiment analysis in Arabic: A review of the literature," Ain Shams Eng. J., 2017.

[6]. A. M. Abd and S. M. Abd, "Case Studies in Construction Materials Modelling the strength of lightweight foamed concrete using support vector machine (SVM)," Case Stud. Constr. Mater., vol. 6, pp. 8-15, 2017.

[7]. R. Cristina, B. Madeo, and S. M. Peres, "Gesture phase segmentation using support vector machines," Expert Syst. Appl., vol. 56, pp. 100-115, 2016.

[8]. B. Ghaddar and J. Naoum-sawaya, "High dimensional data classification and feature selection using support vector machines," Eur. J. Oper. Res., vol. 0, pp. 1-12, 2017.

[9]. L. Martí, N. Sanchez-pi, J. Manuel, and M. López, "On the combination of support vector machines and segmentation algorithms for anomaly detection : A petroleum industry comparative study," J. Appl. Log., vol. 24, pp. 71-84, 2017.

[10]. T. Pinto, T. M. Sousa, I. Praça, Z. Vale, and H. Morais, "Neurocomputing Support Vector Machines for decision support in electricity markets ' strategic bidding," Neurocomputing, vol. 172, pp. 438-445, 2016.

[11]. S. Shabani, P. Yousefi, and G. Naser, "Support vector machines in urban water demand forecasting using phase space reconstruction," Procedia Eng., vol. 186, pp. 537-543, 2017.

[12]. V. Ayumi and M. I. Fanany, "A comparison of SVM and RVM for human action recognition," Internetworking Indones. J., vol. 8, no. 1, pp. 29-33, 2016.

[13]. B. Efron and R. Tibshirani, An introduction to the Boostrap. New York: Chapman & Hall, 1993.

[14]. L. Breiman, "Bagging predictors," Mach Learn., vol. 24, no. 2, pp. 123-40, 1996.

[15]. A. Bosch, A. Zisserman, and X. Muoz, "Image classification using random forests and ferns," in IEEE 11th International Conference on Computer Vision ICCV, 2007, pp. 1-8.

[16]. A. Kuznetsova, L. Leal-Taixé, and B. Rosenhahn, "Real-time sign language recognition using a consumer depth camera," in Computer Vision Workshops (ICCVW), 2013 IEEE International Conference on, 2013, pp. 83-90.

[17]. J. Shotton et al., "Real-time human pose recognition in parts from single depth images," Commun. ACM, vol. 56, no. 1, pp. 116-124, 2013.

[18]. A. Joshi, C. Monnier, M. Betke, and S. Sclaroff, "Comparing random forest approaches to segmenting and classifying gestures ☙," Image Vis. Comput., vol. 58, pp. 86-95, 2017.

[19]. F. Schroff, A. Criminisi, and A. Zisserman, "Object class segmentation using random forests,"

in Proceedings of the British Machine Vision Conference, 2008.

Cite this article as :

Desi Ramayanti, "Comparison of Random Forest and Support Vector Machine for Indonesian Tweet Complaint Classification ", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 5 Issue 6, pp. 202-207, November-December 2019. Available at doi : https://doi.org/10.32628/CSEIT195628
Journal URL : http://ijsrcseit.com/CSEIT195628