International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2018 IJSRCSEIT | Volume 3 | Issue 7 | ISSN : 2456-3307

# Complaint Classification Using Support Vector Machine for Indonesian Text Dataset

**Desi Ramayanti, Umniy Salamah**
Faculty of Computer Science, Universitas Mercu Buana, Jakarta Barat, Indonesia
desi.ramayanti@mercubuana.ac.id, umniy.salamah@mercubuana.ac.id

## ABSTRACT

Text classification is used to classify text data, for example, to find some information from news stories and text from social media that can be used by data owner. Since manual classification is time-consuming and difficult, many studies have been done to this research area. However, the most of studies focused on English text classification. This research attempted to classify Indonesian text dataset by using SVM classifiers. We have conducted research to classify Indonesian text using Python programming language and scikit-learn library. As the result, the experiment without cross validation and tuning parameter for SVM classifier on the dataset achieved the accuracy 0.89473 with value of precision and recall is 0.90289 and 0.89473 respectively. Moreover, value of K for SVM classifier is 0.78992 so that strength of agreement is included into good category. Then, the experiment using cross validation with k-5 and k-10 and tuning parameter with C constant and gamma value. Result of cross validation with k-10 is derived the best accuracy with value 0.9648, however, it spend 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.

**Keywords :** Complaint Classification, Indonesian Text, Support Vector Machine

## I. INTRODUCTION

The study of the machine learning has become increasingly important in recent years as the result of the challenges to process data [1], [2]. With rapid growth of text data, consequently, increase the demand of machine learning for text mining to process text data quickly. The owners of text data forced rethink on how to process text so that it can be used to support business purposes or decision-making.

Text mining related to text classification, which represent the task of assigning text documents to one or more keyword categories [3]. Text classification has become one of the key methods for handling text data. Text classification is used to classify text data, for example, news stories and text from social media, to find some information that can be used by user or data owner. Since manual classification is time-consuming and difficult, many studies have been done to this research area [4]. However, the most of studies focused on English text classification [5]. Every language has different methods or ways to classify text depending on characteristics of its language. The result of classification among languages may be different even though it used same classification algorithm.

Given the importance of text classification, one of text classification algorithms that can be applied is support vector machine (SVM). The excellence of SVM utilization for solving some problems can be showed from the number of publications that used this algorithm. For instances, in the material construction research field, research 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, Ghaddar and Naoum-sawaya (2017) conducted research about *high dimensional* data classification and feature selection using SVM [8]. In other research domain, SVM is still being used to solve existing problems [9]–[12].

In all the research discussed above, SVM is employed by implementing a set of SVM algorithm. Current study in this area is also focused on SVM classifiers that classify Indonesian text dataset simultaneously. Text data used is text data from tweet data taken using program script in R via Twitter API. The data taken are tweets intended for accounts of the Ministry of Marine Affairs and Fisheries of the Republic of Indonesia (@ kkpgoid). Using the SVM method, this research will classify to determine whether a text sentence is either a complaint or a non-complaint. Based on the background above, this research titled Automatic Complaint Classification Using Support Vector Machine Automatically (Dataset: Twitter Data Crawling) is aimed to find out the performance of support vector machine algorithm in classification of Indonesian text.

## II. LITERATURE REVIEW

This section will deliver the literature review about support vector machine (SVM) and its related works. The theory of SVM can be used to understand the fundamental of SVM and related works can be used to know the recent research of SVM.

### A. Support Vector Machine

Support Vector Machines (SVM) is one of supervised learning algorithm that can be used for regression or classification [13]. SVM has been developed by Vladimir Vapnik as machine learning algorithm [14]. SVM is a classier method that separates data optimally by constructing hyper-planes in a multidimensional input space. The best hyper-planes can be defined by measuring boundaries among class memberships. SVM is linear classifier, however, it can be used to linear problem with using kernel trick concept by mapping input space to the high multidimensional input space [14]. Fundamentally, SVM algorithm is designed to classify two classes. But, to solve a problem involved to more than two classes, SVM algorithm can be modified to use in multiclass classification [14].

For classifying two classes, for example data sets $\{(y_1, x_1), (y_2, x_2), (y_3, x_3), \dots, (y_n, x_n)\}$ with $x_i = [x_{i1}, x_{i2}, \dots, x_{im}]$ is input vector in dimension $m$, and $y_i \in \{-1,1\}$ is class target. Value of $N$ is amount of data and $m$ is amount of data attributes or features. With assumption of linearly separable data, SVM will separate data with its hyperlane with equation [14]:

$$x.w + b = 0$$

where $w$ is weight vector and $b$ is scalar. The hyperlane will separate data into two classes, i.e. positive class and negative class with rules as follows [14]:

$x_i.w + b \geq 1$ for positive class $y_i = 1$, dan

$x_i.w + b \geq -1$ for negative class $y_i = -1$,

So that, it can be formulated into [14]:

$y_i(x_i.w + b) \geq 1$, for $\forall_i$

### B. Related Works

The related works of SVM utilization for solving some problems is completed by many researchers, for example, in the material construction research field, research 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].

## III. METHODOLOGY

This research will be completed through five phases, i.e. data acquisition, data pre-process, feature selection, classification and evaluation, as depicted in below.

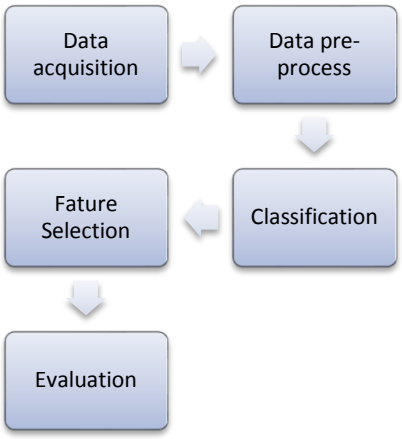


**Figure 1.** Research methodology

The flow of proposed research phase is elaborated below:

1. Data acquisition

At this stage we extract the tweet data automatically using the program script in R via the Twitter API. The data taken are tweets intended for accounts of the Ministry of Marine Affairs and Fisheries of the Republic of Indonesia (@ kkpgoid). Data was taken on November 17, 2017. The data obtained as many as 1170 tweets.

2. Data pre-process

At this stage the data that has been obtained is then done pre-process among others are:

a. Data cleansing by deleting duplicate data (re-tweets), as well as deleting word repetition in tweets.
b. Labeling data by labeling positive, negative, and neutral sentiment data.
c. Case folding, by converting all text into the lower case.
d. Removing special character, by removing URL (web address), @username (user account name), 'RT' (re-tweet), special characters, and punctuation.
e. Removing Stop Word, eliminating Words that are considered to have no meaning. In Indonesian Stop Word like 'yang,' di ',' ke ', and others. While Stop Word in twitter like 'xoxoxo', 'wkwkwk', and others.

3. Feature selection

This phase is the feature selection stages that is choosing the right features and form the Feature Vector. In this study we use TF-IDF Feature to represent data.

4. Classification

After formed Feature Vector, then, we separated data into training and testing data to support classification process. This stage is done by training SVM classifier with training data so as to form the model. Once the model is generated then the model is tested with test data to predict the data label.

5. Evaluation

Furthermore, the test results are evaluated to calculate the model accuracy, and calculated also using other evaluation calculations such as precision, recall and, f-measure.

## IV. EXPERIMENTAL RESULT

Python programming language and *scikit-learn* library. Moreover, the data augmentation has been completed by using R-library. The pre-processing stage is completed 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 experiment conducted in two phases, i.e. experiment without cross validation and tuning parameter, then, experiment with cross validation and tuning parameter. The performance result for SVM classification for experiment without cross validation and tuning parameter is shown in Table 1 with detail of accuracy, f1-score, precision, recall, and kappa value.

TABLE I

RESULT OF EXPERIMENT WITHOUT CROSS VALIDATION AND TUNING PARAMETER

| Accuracy | F1 score | Precision | Recall | Kappa | Computational time |
|---|---|---|---|---|---|
| 0.89473 | 0.89434 | 0.90289 | 0.89473 | 0.78992 | 0.019545 s |

Table 1 shown experiment without cross validation and tuning parameter for SVM classifier on the dataset. Based on experiment, SVM classifier achieved the accuracy 0.89473. It showed that SVM is being able to classify text data based on category: complaint and non-complaint. Moreover, a model can be stated the good model if it is resulted the high value of precision and recall. Based on experiment, it showed that value of precision and recall are 0.90289 and 0.89473 respectively. Then, for interpreting value of Cohen's Kappa can refer to Table 2 below [15]:

TABLE II

INTERPRETATION OF COHEN'S KAPPA

| K Value | Strength of agreement |
|---|---|
| < 0.20 | Poor |
| 0.21 – 0.40 | Fair |
| 0.41 – 0.60 | Moderate |
| 0.61 – 0.80 | Good |
| 0.81 – 1.00 | Very good |

Based on Table 1 and Table 2, we can state that value of K for SVM classifier is 0.78992 so that strength of agreement is included into good category. Moreover, the detail of precision, recall, and f1-score for each class is shown in Table 3 and the confusion matrix is depicted in Figure 2.

TABLE III

PRECISION, RECALL, AND F1-SCORE FOR EACH CLASS

| Class | Precision | Recall | F1-score |
|---|---|---|---|
| 0 | 0.96 | 0.83 | 0.89 |
| 1 | 0.84 | 0.96 | 0.90 |
| average | 0.90 | 0.89 | 0.89 |

| | |
|---|---|
| 72 | 15 |
| 3 | 81 |

**Figure 2.** Confusion matrix for each class

The second experiment is experiment with cross validation and tuning parameter. The result of performance for SVM classification for experiment

with cross validation and tuning parameter is shown in Table 4. In this experiment, we conducted cross validation with k-5 and k-10 and implemented tuning parameter of SVM with C constant and gamma value.

TABLE IV
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 Table 4 above, C constant and gamma is obtained a better result for cross validation. Result of cross validation with k-10 is derived the best accuracy with value 0.9648, however, it spend computational time as much 40.118 second. Then, we conducted experiment to find the best kernel function among Sigmoid, Linear and RBF. The result of experiment can be seen in Table V with detail of accuracy and computational time for each kernel function. Moreover, based on result of experiment, kernel function Sigmoid achieved the best accuracy and computational time.

TABLE V
CROSS VALIDATION

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

## V. CONCLUSION

We have conducted research to classify Indonesian text using Python programming language and *scikit-learn* library and conclude some results:

1. The experiment without cross validation and tuning parameter for SVM classifier on the dataset achieved the accuracy 0.89473 with value of precision and recall is 0.90289 and 0.89473 respectively. Moreover, value of K for SVM classifier is 0.78992 so that strength of agreement is included into good category.
2. The second experiment is experiment using cross validation with k-5 and k-10 and tuning parameter with C constant and gamma value. Result of cross validation with k-10 is derived the best accuracy with value 0.9648, however, it spend 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.

## VI. ACKNOWLEDGEMENT

This research have 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] C. Burges, *A tutorial on support vector machines for pattern recognition*. Boston: Kluwer Academic Publishers, 1998.

[14] V. N. Vapnik, "An Overview of Statistical Learning Theory," vol. 10, no. 5, pp. 988–999, 1999.

[15] D. . Altman, *Practical Statistics for Medical Students*. London: Chapman and Hall, 1991.

[16] R. L. B. Bai, "How do the preferences of online buyers and browsers differ on the design and content of travel websites?," *Int. J. Contemp. Hosp. Manag.*, vol. 20, no. 4, pp. 388–400, 2008.

[17] J. Brooke, "SUS - A quick and dirty usability scale," 1986.