International Journal of Computer Techniques-Volume 5, Issue 5 Sep - Oct 2018

RESEARCH ARTICLE OPEN ACCESS

Text Classification on Dataset of Marine and Fisheries Sciences Domain using Random Forest Classifier

Desi Ramayanti*, Umniy Salamah**

*Faculty of Computer Science, Universitas Mercu Buana, Jakarta, Indonesia

Email: *desi.ramayanti@mercubuana.ac.id, **umniy.salamah@mercubuana.ac.id

----------**********----------

Abstract:

The number of text documents contained opinions and suggestions are increasing and challenging to interpret one by one. Whereas if the text documents are processed and properly interpreted, this text document can provide a general overview of a particular case, organization, or object. This research focused on text classification on marine and fisheries domain by analyzing the Twitter data related to Ministry of Marine Affairs and Fisheries, Republic of Indonesia. By using random forest algorithm, this research will classify text documents whether classified as complaint or non-complaint based on existing data in social media in order to support follow up to the complaint. Related 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 to completed this research. The phase of this research including data acquisition, data pre-processing, feature selection, classification and classifier evaluation. As the result, we foundbest performance is achieved when we use parameters 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. The best score is achieved in this experiment is 0.956063268893 using those parameter values with computational time required to tune parameters is 109.399434 second.

Keywords **—text classification, social media, marine and fisheries sciences, e-government, text classifier, random forest algorithm**

----------**********----------

I. INTRODUCTION

Text documents on a particular topic are straightforward to find especially in the current era of technology, either through social media or websites. The number of text documents contained opinions and suggestions are increasing and challenging to interpret one by one. Whereas if the dataare processed and properly interpreted using machine learning, the can provide a general overview of a particular case, organization, or object[1]–[5].

This research focused on text classification on marine and fisheries domain by analysing the Twitter data related to Ministry of Marine Affairs and Fisheries, Republic of Indonesia. By using random forest algorithm, this research will classify text documents whether classified as complaint or

non-complaint based on existing data in social media in order to support follow up to the complaint.

The recent research about random forest has been done on several dataset. 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. The image dataset used in this research are the Caltech-101 and Caltech-256 data sets[6].

Joshi, Monnier, Betke, and Sclaroff (2017) completed research work about random forest algorithm by implementing it into gesture recognition system. In that system, every gesture image within continuous video stream will be classified based on its given vocabulary of gesture image[7].

Kuznetsova, Leal-Taixé, and Rosenhahn (2013) conducted work in human computer interaction and computer vision research field, especially hand gesture recognition. At the beginning this research became complicated because the given dataset is only in RGB image format, however, in recent years the development of sensing technologies leads to research of hand gesture recognition more feasible by providing structured-light and time-of-flight cameras.

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. This research aim is to generate confidence-scored of several body joints in 3D form by re-projecting the result of classification and finding local modes[9].

Schroff, Criminisi, and Zisserman (2008) observed performance of random forest in pixel-wise images segmentation. As the result, performance of random forest in combining multiple features leads to performance improvement if colour, textons, filterbanks, and Histogram of Oriented Gradient (HOG) features are employed simultaneously[10].

II. Literature Review

We deliver about theory of random forest algorithm and its related works in this section.

A. Random Forest Algorithm

Random forest is an algorithm based on Bagging and Random Subspace which is consisted of multi-way or binary decision trees $h1(x), h2(x), \ldots h_{nTree}(x)$ as depicted in Fig. 1. The aggregation of all decision trees prediction based on majority votes is used to make the final decision [11].

The dataset $T = \{(x_{i1}, x_{i2}, .., x_{iM}, y_i)\}_{i=1}^{N}$ consist of N samples, the vector $x_{i1}, x_{i2}, .., x_{iM}$ represents the M-dimension features or attributes, $Y = \{y_i\}\frac{N}{i}$ represents label of classification and a sample is concluded as label c by $y_i = c$ [11].

Random forest (RF) algorithm consisted of two procedures. The first procedure is training sets are designed and constructed using random bootstrap method with replacement as shown in Figure 1 [12][13].

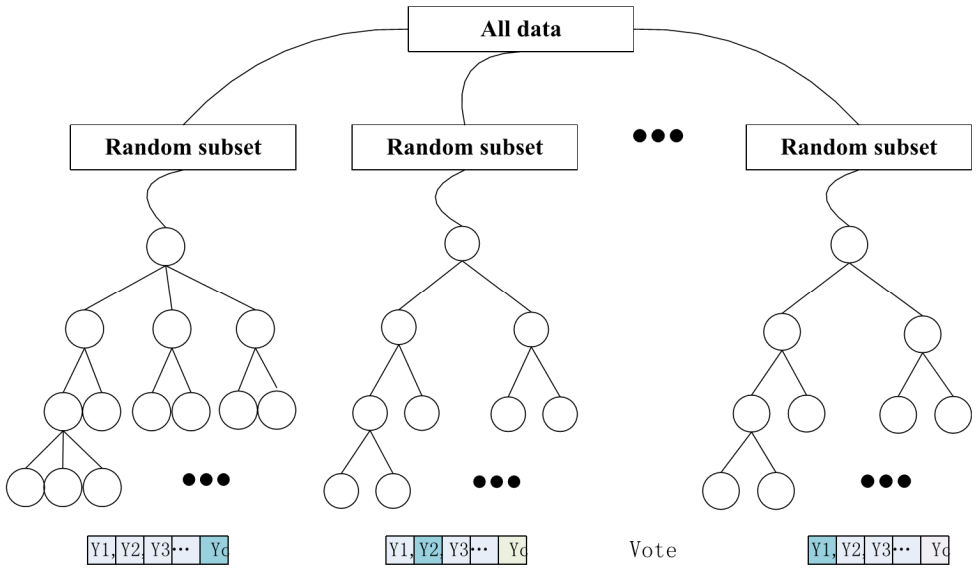

Fig. 1 Overview of random forest algorithm[11]

The second procedure, based on the total features, the random features are defined with non-replacement. The size k of subset feature is commonly far less than the total features size, M.

To select κ features randomly, we must calculate information gain of κ split and define the best features. Then, candidate of features size becomes M – κ as shown in Figure 1 [11].

To find the best attribute, it can be calculated by using some methods, i.e. information gain rate, information gain, and Gini coefficient in which correspond to ID3, C4.5[14]and CART [15], respectively. The best split point must be defined while the attribute value is continuous. One of methods that can be used is CART, whereas a better classification result indicated from smaller Gini coefficient. The P_i is the sample proportion of sample I in the total of sample size with assumption that sample T divided into k parts after separating by attribute A that formulated as follows [11]:

$$Gini(T) = 1 - \sum_{i}^{c} P_i^2$$

$$Gini(T, A) = \sum_{j=1}^{k} \frac{|T_j|}{|T|} Gini(T_j)$$

B. *Related Work*

Related 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 to completed this research[6]–[10]

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. The image dataset used in this research are the Caltech-101 and Caltech-256 data sets[6].

Joshi, Monnier, Betke, and Sclaroff (2017) completed research work about random forest algorithm by implementing it into gesture recognition system. In that system, every gesture image within continuous video stream will be classified based on its given vocabulary of gesture image[7].

In this research, Joshi, Monnier, Betke, and Sclaroff (2017) utilized two methods which are completed simultaneous classification and temporal segmentation respectively. To classify gestures based its given vocabulary which is provided in a 3D video of joint location dataset for training process as well as out-of-vocabulary objects, this research employed a single random forest model. A cascaded approach as the second method is used to train a model of binary random forest for distinguishing gestures from background and a model of multi- class random forest for classifying segmented gestures. The datasets used in this research are the ChaLearn dataset contained 7754 gesture instances (from 20 Italian communication gestures) and the NATOPS dataset contained 9600 gesture instances (from 24 aircraft handling signals)[7].

Kuznetsova, Leal-Taixé, and Rosenhahn (2013) conducted work in human computer interaction and computer vision research field, especially hand gesture recognition. At the beginning this research became complicated because the given dataset is only in RGB image format, however, in recent years the development of sensing technologies leads to research of hand gesture recognition more feasible by providing structured-light and time-of-flight cameras[8].

In this research, Kuznetsova, Leal-Taixé, and Rosenhahn (2013)proposed the precise approach to recognize static gestures by using depth images from sensing technology devices. The dataset is trained by using a multi-layered random forest (MLRF) to classify its feature vectors, which yields to the recognition of the hand gestures. To benchmark of multi-layered random forest (MLRF) utilization, this work also used other datasets, i.e. dataset of 24 signs from American Sign Language (ASL) which is collected by using Intel Creative Gesture Camera[8].

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. This research aim is to generate confidence-scored of several body joints in 3D form by re-projecting the result of classification and finding local modes[9].

Schroff, Criminisi, and Zisserman (2008) observed performance of random forest in pixel-wise images segmentation. As the result, performance of random forest in combining multiple features leads to performance improvement if color, textons, filterbanks, and Histogram of Oriented Gradient (HOG) features are employed simultaneously[10].

III. METHODOLOGY

This research has been done through five research phases.

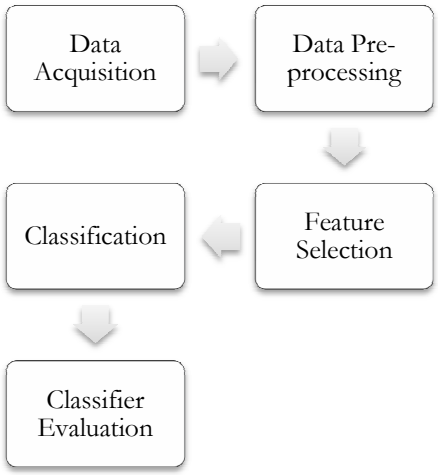

Fig. 2Research flow

The first phase is data acquisition. In this phase, we gathered data on Twitter using R script based on Twitter API. The relevance data is data that included Kementrian Kelautan dan Perikanan Republik Indonesia (@kkpgoid). As many 1170 Twitter data has been collected in this phase.

The second phase is data pre-processing including:

a. Data cleansing is done by deleting duplicate tweets (re-tweets), as well as deleting repeated words (reps).
b. Data labeling by label justification for each data is classified as positive, negative, or neutral sentiment.
c. Converting all text into lower case (case folding).
d. Removing URL, @username (user account name), 'RT' (re-tweet), special characters, and punctuation.
e. Removing Stop Word. Stop Word is a word that is considered to have no meaning. In Indonesian stop word like 'whereas,' will ',' yang ', and others. While stop word in twitter like 'wkwkwk', 'hmm', and others.

Third phase is feature selection. Data that has been clean and ready for processing then converted into a feature vector form. The feature selected in this process is the TF-IDF feature.

Fourth phase is Classification. Feature Vector which has been formed next is done classification process. Classification is done by training the classifier using training data, so model is formed. The model is then used to predict the test data.

Fifth phase is Classifier Evaluation. Prediction results in subsequent test data in the evaluation to determine the performance of the model built.

IV. EXPERIMENTAL RESULT

In this research, we used Python (python-sklearn) for running and processing dataset that gathered on Twitter using R script based on Twitter APIincluded Kementrian Kelautan dan Perikanan Republik Indonesia (@kkpgoid).The tuning parameter process in Scikit Learn used GridSearchCV. The process of running data can be shown in Figure 1.

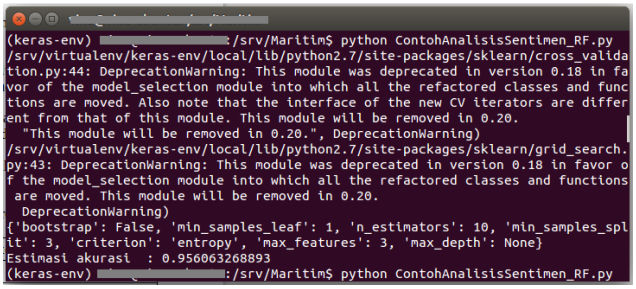

Fig. 3 Running data using Python

Based on Figure 2 above, 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:

1. Parameter of *n_estimators* is the amount of trees in the forest. The default of estimator is "10". In this research, we used the "10", "20", and "30"*n_estimators* for tuning.
2. Paramer of *criterion* is the function of the quality of a split measurement. The supported criteria are "gini" for the Gini Impurity and "entropy" for the information gain.
3. Parameter of *max_depth* is the maximum depth of the tree, in which the default value of *max_depth* is "None". If *max_depth*value is None, then nodes are expanded until all leave contain less than *min_samples_split* samples or until all leaves are pure. In this experiment, we tuned the parameter of *max_depth*with value "3"and "None".
4. Parameter of *max_features* is the amount of feature to consider when defining for the best split. In this experiment, we tuned the value of *max_features*with value "1", "3" and "10".
5. Parameter of *min_sample_split*is the minimum of samples that isneeded to split an internal node. The default value of *min_sample_split* is

"1". In this experiment, we tuned the minimum number of samples to split with value "2", "3", and "10".

6. Parameter of *min_samples_leaf*is the minimum samples that is needed to be at a leaf node. The default value of *min_samples_leaf* is "1". We tuned the number of min_samples leaf between "1", "3", and "10" in this experiment.
7. Parameter of *bootstrap* is whether bootstrap samples are used when constructing trees.The default value of *bootstrap*is "True". We tuned the bootstrap parameter between "True" and "False" in this experiment.

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 0.956063268893.

TABLE I
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

The computational time required to tune parameters is 109.399434 second which are resulted as follows:

```
{'bootstrap': False, 'min_samples_leaf':
1, 'n_estimators': 10,
'min_samples_split': 3, 'criterion':
'entropy', 'max_features': 3,
'max_depth': None}

Estimasi akurasi  : 0.956063268893
```

Fig. 4 Result of running data

V. CONCLUSION

Based on research result can be conclude that the best performance is achieved when we use parameters 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. The best score is achieved in this experiment is 0.956063268893 using those parameter values with computational time required to tune parameters is 109.399434 second.

ACKNOWLEDGMENT

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

REFERENCES

[1] V. Ayumi and M. I. Fanany, "Multimodal Decomposable Models by Superpixel Segmentation and Point-in-Time Cheating Detection," in *2016 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, 2016, pp. 391–396.

[2] 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.

[3] I. Nurhaida, R. Manurung, and A. M. Arymurthy, "Performance comparison analysis features extraction methods for batik recognition," in *International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, 2012.

[4] D. Fitrianah, A. N. Hidayanto, R. A. Zen, and A. M. Arymurthy, "APDATI: E-Fishing Logbook for Integrated Tuna Fishing Data Management," *J. Theor. Appl. Inf. Technol.*, vol. 75, no. 2, 2015.

[5] M. Sadikin, M. I. Fanany, and T. Basaruddin, "A New Data Representation Based on Training Data Characteristics to Extract Drug Name Entity in Medical Text," *Comput. Intell. Neurosci.*, vol. 2016, 2016.

[6] 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.

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

[8] 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.

[9] 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.

[10] F. Schroff, A. Criminisi, and A. Zisserman, "Object class segmentation using random forests," in *Proceedings of the British Machine Vision Conference*, 2008.

[11] L. Ma and S. Fan, "CURE-SMOTE algorithm and hybrid algorithm for feature selection and parameter optimization based on random forests," *BMC Bioinformatics*, vol. 18, pp. 1–19, 2017.

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

[13] L. Breiman, "Bagging predictors," *Mach Learn.*, vol. 24, no. 2, pp. 123–40, 1996.

[14] J. R. Quinaln, *C4.5: programs for machine learning [M]*. Morgan kuafmann, 1993.

[15] L. Breiman, J. Friedman, R. Olshen, and C. Stone, *Classification and Regression Trees*. Boca Raton, FL: CRC Press, 1984.