RESEARCH ARTICLE OPEN ACCESS

Supervised Classification of Indonesian Text DocumentUsing Extreme Gradient Boosting (XGBoost)

UmniySalamah*, DesiRamayanti**
*Faculty of Computer Science, Universitas Mercu Buana, Jakarta, Indonesia
Email: *umniy.salamah@mercubuana.ac.id, **desi.ramayanti@mercubuana.ac.id

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

Abstract:

One of the machine learning algorithms that recently gained benchmarks in state of the art various problems in machine learning is eXtreme Gradient Boosting (XGBoost). This algorithm has been solved many in various data mining and machine learning challenge, and has been applied in various problems by giving good results. This research will classify text data using XGBoost to predict a text whether classified as complain or non-complaint based on existing data in social media.In this experiment, we installed XGBoost package on our system for used in Python. XGBClassifier is imported to our codes to support in using sklearn's Grid Search for tuning parameters with parallel processing. The XGBoost algorithm uses multiple parameters so that it can be improved by tuning the parameters, i.e. eta or learning rate, gamma, max_depth, min_child_weight, subsample, colsample_bytree and alpha. As the result, the best value for each parameter are 'reg_alpha': 0.01, 'colsample_bytree': 0.9, 'learning_rate': 0.5, 'min_child_weight': 1, 'subsample': 0.8, 'max_depth': 3, 'gamma': 0.0. Moreover, the computational time that required to tune those parameters is 13870.012468 and the best accuracy that achieved is 0.927943760984.

Keywords **—Indonesian text, text classification, marine and fisheries sciences, XGBoost**

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

I. INTRODUCTION

In information technology era, textual data on a particular topic is very easy to find, both on websites and social media. Data in the form of opinions and suggestions are useful continue to increase so difficult to interpret quickly. If interpreted properly, the data o is very useful because it can give a general description of a topic, without having to do a survey manually[1]–[5].

One of the social media where users often interact with giving opinions and suggestions is Twitter. Twitter is a popular micro-blogging medium where users can write status messages or called "tweets". These tweets sometimes express opinions about a topic[6]–[8].

This study proposed an automated method for classifying a tweet whether it is classified as a complaint or not a complaint in a tweet. It is useful, in an organization to save time in order to follow up complaints quickly which is an urgent thing. This research focuses on text classification with object of maritime text data with case study of Ministry of Marine Affairs and Fishery of Republic of Indonesia.

One of the machine learning algorithms that recently gained benchmarks in state of the art various problems in machine learning is eXtreme Gradient Boosting (XGBoost). XGBoost is a

supervised classification technique that uses ensemble decision trees. The ensemble boosting technique is used to increase Taylor expansion against loss of function loss. The model built by XGBoost is also insensitive to the data imbalance[2].

This algorithm has been solved many in various data mining and machine learning challenge, and has been applied in various problems by giving good results. This research will classify text data using XGBoost to predict a text whether classified as complain or non-complaint based on existing data in social media.

II. Literature Review

We explained the Extreme Gradient Boosting (XGBoost) and related works in this section.

A. *Extreme Gradient Boosting (XGBoost)*

Extreme Gradient Boosting (XGBoost) proposed by Chen and Guestrin is an open source project to implement an efficient, fast, and measurable machine learning system called Gradient Tree Boosting for various learning problems[9].XGBoost is developed from Regression Trees (CART)$\{R_1(x_i, y_i) \ldots\ldots R_k(x_i, y_i)\}$andClassification C, where x_i and y_i are trainingdataandclass label. Prediction score is evaluated by using additive function C as follows:

$$\widehat{y_i} = \sum_{k=1}^{C} f_k(x_i), f_k \in F$$

wheref_k is an independent tree structure with leaf scores and F is the total area of the CART. Then the regularization is done to optimize with the following equation:

$$Obj(\Theta) = \sum_{i}^{n} l(y_i, \widehat{y_i}) + \sum_{k}^{C} \Omega(f_k)$$

In the equation above the first term is a loss function (l) which measures the difference between predictions $\widehat{y_i}$ andtarget y_i . The second term is regularization$\Omega$to avoidover-fittingof model.

B. *Related Work*

Some studies have used XGBoost to solve specific problems including[2], proposing XGBoost for human movement recognition. The study was conducted by comparing XGBoost with several other machine-learning algorithms and tested for various types of human movement datasets. The results obtained show that the XGBoost algorithm is better at human movement recognition.

Other research, study by [10] proposed Support Vector Machine algorithm (SVM) and XGBoost to predict personality based on Twitter data. Evaluation with 10-cross validation shows that XGBoost algorithm is superior in predicting personality compared to SVM algorithm.

Research [11]proposed the sign language recognition by applying several machine learning methods, namely XGBoost, SVM, and k-NN. The experimental results show that XGBoost has significant results compared to some other machine learning algorithms such as SVM and k-NN.

III. Methodology

This study has been done through five research phases which depicted in Figure 1.

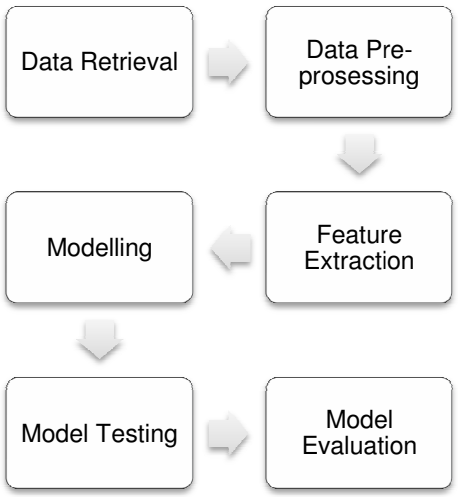

Fig. 1Research methodology

This research will be conducted in 6 main stages:

1. Data retrieval

   The initial stage is the data retrieval stage. The data we use here is tweet data intended for twitter account Ministry of Marine and Fisheries of the Republic of Indonesia (@ kkpgoid). The data consists of 1170 tweets, obtained on November 11, 2017. The data is taken by crawling using the Twitter API and R script.

2. Pre-process data

   After the data obtained next is to pre-process the data. Pre-processes include data cleansing by removing duplicate data (retweet), word reps and or sentences, deletion of '@username', user name, URL (web address), 'RT' (retweet mark), punctuation, and characters special characters, and the removal of Stop Word. After that the data is labelled by labelling for each data whether classified as positive, negative, or neutral sentiment.

3. Feature Extraction

   Having obtained the data that has been clean and has the next label is extracting the data into a feature, so it will produce a Vector Feature. The feature used is TF-IDF feature.

4. Model Development

   After data in Vector form The next feature is data separation for training and testing. Separation of data was done by cross validation with 70% data proportion for training and 30% for testing. Training data will then be used to build the model. Classifier will be trained to use training data so as to produce the most optimal model.

5. Model Testing

   The model produced at the next stage of the training is tested using test data. Testing is done by predicting labels for test data.

6. Model Evaluation

   Evaluation is done by measuring the result of prediction accuracy. It also calculated kappa statistic, precision and recall.

IV. RESULT

In this experiment, we installedXGBoost package on our system for used in Python. XGBClassifier is imported to our codes to support in usingsklearn'sGridSearchfor tuning parameters with parallel processing.

In general, the XGBoost algorithm uses multiple parameters so thatto improve the model, we must conduct the process of tuning the parameters. The XGBoostparameters that are tuned to improve the model elaborated below:

1. Parameter*eta*(or*learning rate*).It used to shrinks the feature weight to make the process of boosting more conservative. The size shrinkage step is used in update to prevent over-fitting. Eta range is 0-1 and default is 0.3.
2. Parameter *gamma*(alias:*min_split_loss*). It is the minimum loss reduction, this valaue is

required to make further partition on a leaf node of the tree. The larger this value, the more conservative the algorithm will be, the range of this value is 0-∞ and the default is 0.

3. Parameter *max_depth*. It is the maximum depth of a tree. Incrase this will make the model more complex and tend to be overfitting. The range is between 0, ∞ and the default is 6.
4. Parameter *min_child_weigth*. It is the minimum sum of instance weight needed in a child. The larger this value, the more conservative the algorithm will be. The range of the value is 0-∞ where the default is 1.
5. Parameter *subsample*. It is the subsample ratio of the training instance. If we set in to 0.5, the XGBoost will be randomly collected half of data instance to grow trees and will prevent over-fitting. The range of the value is 0-1, and the default is 1.
6. Parameter *colsample_bytree*. It is the subsample ratio of columns when constructing each tree. The range of the value is 0-1, and the default is 1.
7. Parameter*alpha*(or *reg_alpha*). It is the L1 regularization term on weights. Increase this value will make model more conservative. The default of this value is 0.

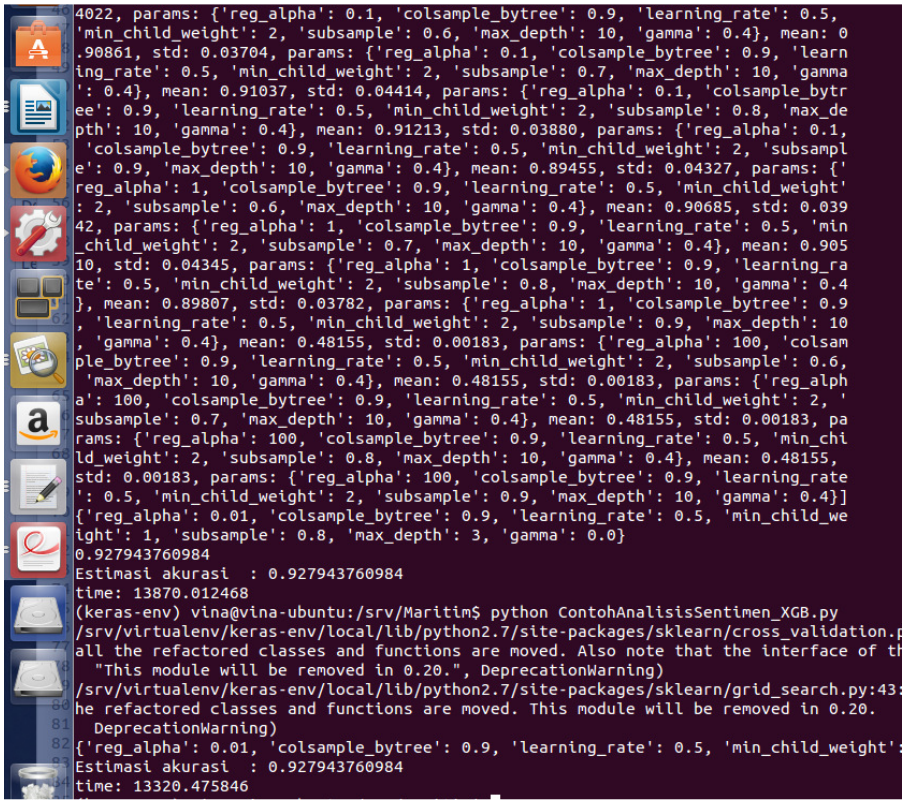

Fig.2Experiment

Based on description above, we conducted experiment to get value of parameters. As the result of the value is presented in Table 1.

TABLE I
VALUE OF PARAMETERS

Parameters	Value
eta/ learning rate	0.01, 0.1, 0.25, 0.5
gamma	0.1, 0.2, 0.3, 0.4, 0.5
max_depth	2, 3, 5, 10
min_child_weight	1, 6, 2
subsample	0.6, 0.7, 0.8, 0.9, 1
colsample_bytree	0.6, 0.7, 0.8, 0.9, 1
alpha	1e-5, 1e-2, 0.1, 1, 100

Moreover, the summary of result regarding the best value for each parameter is presented in Table 2 below.

TABLE II
SUMMARY OF THE BEST PARAMETER VALUES

Parameters	Best value
eta/ learning rate	0.5
gamma	0.0
max_depth	3
min_child_weight	1
subsample	0.8
colsample_bytree	0.9
alpha	0.01

Regarding accuracy and computational time, we have conducted data processing using our codes. The computational time that required to tune those parameters is 13870.012468 and the best accuracy that achieved is 0.927943760984 in which the result is presented in Figure 1.

```
{'reg_alpha': 0.01,
'colsample_bytree': 0.9,
'learning_rate': 0.5,
'min_child_weight': 1, 'subsample':
0.8, 'max_depth': 3, 'gamma': 0.0}
Estimasiakurasi  : 0.927943760984
time: 13320.475846
```

Fig.3Result

V. CONCLUSION

Based on our research, we can conclude some points as follows:

1. XGBoost package can be used in Python by using XGBClassifier for sklearn'sGridSearch in order to tuning parameters with parallel processing.
2. The XGBoost algorithm uses multiple parameters so that it can be improved by tuning the parameters, i.e. eta or learning rate, gamma, max_depth, min_child_weight, subsample, colsample_bytree and alpha.
3. The best value for each parameter are 'reg_alpha': 0.01, 'colsample_bytree': 0.9, 'learning_rate': 0.5, 'min_child_weight': 1, 'subsample': 0.8, 'max_depth': 3, 'gamma': 0.0.
4. The computational time that required to tune those parameters is 13870.012468 and the best accuracy that achieved is 0.927943760984.

ACKNOWLEDGMENT

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

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] 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 and I. Wasito, "Translation and classification algorithm of FDA-Drugs to DOEN2011 class therapy to estimate drug-drug interaction," in *The 2nd International Conference on Information Systems for Business Competitiveness*, 2013.

[6] N. Azizah, M. Ivan, and I. Budi, "Twitter Sentiment to Analyze Net Brand Reputation of Mobile Phone Providers," *Procedia - Procedia Comput. Sci.*, vol. 72, pp. 519–526, 2015.

[7] A. Mittal, "Stock Prediction Using Twitter Sentiment Analysis," no. June, 2009.

[8] B. J. Jansen, M. Zhang, K. Sobel, and A. Chowdury, "Twitter Power: Tweets as Electronic Word of Mouth," *J. Am. Soc. Inf. Sci. Technol.*, vol. 60, no. 11, pp. 2169–2188, 2009.

[9] T. Chen and C. Guestrin, "XGBoost," *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min. - KDD '16*, pp. 785–794, 2016.

[10] W. Andangsari and M. N. Suprayogi, "Personality Prediction Based on Twitter Information in Bahasa Indonesia," vol. 11, pp. 367–372, 2017.

[11] M. Borg and K. P. Camilleri, "Towards a Transcription System of Sign Language Video Resources via Motion Trajectory Factorisation," *Proc. 2017 ACM Symp. Doc. Eng.*, pp. 163–172, 2017.