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/CSEIT195629

A Comparison of Text Classification Techniques Applied to Indonesian Text Dataset

Umniy Salamah
Faculty of Computer Science, Universitas Mercu Buana, Jakarta Barat, Indonesia

ABSTRACT

In organization, statement contained opinion and complaint to a service or program by it organization. can be proceed using machine learning and the result can be used by organization to improve and enhance their quality. This research attempted to classify the reports from social media based on complaint and non-complaint using machine learning algorithm named Logistic regression (LR) and eXtreme Gradient Boosting (XGBoost). Logistic Regression model using CountVectorizer feature extraction and TfidfVectorizer. Moreover, 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 XGBoost with 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, in wich the computational time is 13870.012468 and the best accuracy that achieved is 0.927943760984. Furthermore, the performance evaluation results for Logistic Regression using TfidfVectorizer and CountVectorizer feature extraction are 0.9181 and 0.9356.

Keywords : Indonesian Text, Logistic Regression, Text Classification, Xgboost

I. INTRODUCTION

Now, information and communication technology development present to the innovative application that helps our daily life in many areas, including cultural heritage [1], [2], government [3], [4], medical [5] and so forth. One of the tremendous popular applications is social media that has been practiced by people for all range of ages. Social media is regularly used to state emotion in judgment or complaint statement regarding an object [6]. The increase of text data contained opinion or complaint in social media showed the importance of social media channels to campaign or published about programs or services that are launched by governments or other organizations [7].

Machine learning algorithm has been used in text classification are Logistic Regression and XGBoost. Logistic Regression can predict output or target in various categories. In this case, the algorithm will classify text data into two classes (1, -1), including complaint and non-complaint. Logistic regression (LR) have been applied to numerous obstacles in data mining and machine learning in which logistic regression (LR) explained between the response variable and predictor variables.

eXtreme Gradient Boosting or XGBoost is a supervised classification method that uses ensemble decision trees. The ensemble boosting technique is used to improve Taylor expansion against loss of function loss. The model constructed by XGBoost is also insensitive to the data imbalance [8].

This research is focused on classifying complaint and non-complaint reports from social media. This study implemented machine learning algorithm named logistic regression (LR) for a text whether classified as complaint or non-complaint based on existing data in social media.

II. RELATED WORK

A. Logistic Regression

The recent works have been proposed Logistic regression (LR) algorithm to solve the research problems, e.g. Cheng and Eyke (2009), Rus et al. (2009), Freyberger et al. (2004), Feng and Back (2009), Kotsiantis et. al (2003), Mittal (2009) and Felix (2014).

Cheng and Eyke (2009) proposed the combination of Instance Based Learning and Logistic Regression to complete the multi-label classification [14]. Freyberger et al. (2004) proposed Logistic regression (LR) algorithm to find the best fitting of transfer model in case student learning data [15].

Rus et al. (2009) attempted to compare the result of data processing using several machine learning method for student mental model detection, e.g. Naïve Bayes, Bayes net, Support Vector Machines (SVM), Logistic Regression and Decision Trees [16].

Feng and Back (2009) proposed logistic regression for construction model of transfer in order to predict student can represent their knowledge [17]. Kotsiantis et. al (2003) attempted to classify student dropout prediction by using Neural Network, Decision Tree, Naïve Bayes, Instance Based Learning, Logistic Regression, and Support Vector Machine (SVM) [18].

B. XGBoost

Some studies have used XGBoost to solve specific problems including [8], 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 [19] 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 [20] 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 research has been done through five research phases. The first phase is data collection, which is followed by crawling and labelling dataset, procession data, Modelling using a) Logistic Regression b) XGBoost, and Classifier Perfomance Comparison.

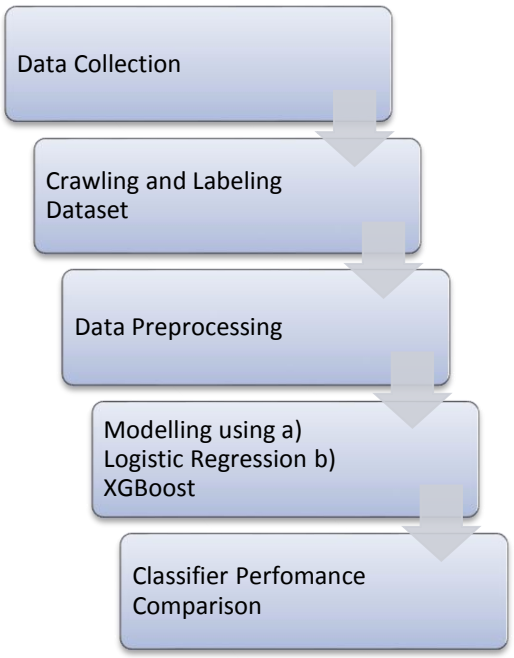

Fig. 1 Research Methodology

In the first phase, we collect datasets from social media. Then, we label the dataset based on the tag label. In the second phase, the collected and marked data is pre-processed to delete unimportant text data (stop-word removal), and then we conducted a feature extraction based algorithm parameter. In the third phase, we model the data applying training data, and then classified the testing data using this model. In the last stage, we compare classifier logistic regression and XGBoost for dataset classification.

IV. EXPERIMENTAL RESULT

A. Logistic Regression

The phase in this work, for instance, pre-processing, classification, validation, and evaluation are conducted in python programming language and the scikit-learn library.

The pre-processing stage is done by removing stop words and extracting features using TfidfVectorizer and CountVectorizer, the feature extraction function in the sklearn library. Classification in this stage is done using the Logistic Regression in the sklearn library. The validation in this stage is done using cross-validation, where the percentage of training sample 70% and testing sample 30%. The feature that we extracted using TfidfVectorizer and CountVectorizer in the previous step, then we used them to compare the best model.

The performance evaluation results for Logistic Regression using TfidfVectorizer and CountVectorizer feature extraction are presented in Table 1.

TABLE I .

Evaluation results for Logistic Regression

Par.	Acc.	F1 score	Precision	Recall	Kappa
Tfidf	0.9181	0.9181	0.9191	0.9181	0.8363
Count	0.9356	0.9355	0.9406	0.9356	0.8715

The result shows that Logistic Regression using CountVectorizer feature extraction achieved better performance than TfidfVectorizer feature extraction. The precision, recall, and f1-score for each class using TfIdf-Vectorizer presented in Table 2.

TABLE II.

Precision, recall, and f1-score for each class using TfIdf Vectorizer

class	precision	recall	F1-score
0	0.94	0.90	0.92
1	0.90	0.94	0.92
average	0.92	0.92	0.92

The precision, recall, and f1-score for each class using Count Vectorizer presented in Table 3.

TABLE III.

Precision, recall, and f1-score for each class using Count Vectorizer

class	precision	recall	F1-score
0	0.99	0.89	0.93
1	0.89	0.99	0.94
average	0.94	0.94	0.92

B. XGBoost

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 GridSearch for tuning parameters with parallel processing.

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

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.

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.

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.

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.

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.

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.

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.

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

TABLE IV
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 below.

TABLE V
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

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's GridSearch in order to tuning parameters with parallel processing. The computational time that required to tune parameters is 13870.012468 and the best accuracy that achieved is 0.927943760984.
2. The performance evaluation results for Logistic Regression using TfidfVectorizer and CountVectorizer feature extraction are 0.9181 and 0.9356.

VI. ACKNOWLEDGEMENT

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

VII. REFERENCES

[1]. I. Nurhaida, A. Noviyanto, M. Manurung, and A. M. Arymurthi, "Automatic Indonesian's Batik Pattern Recognition using SIFT Approach," in ICCSCI - 1st International Conference on Computer Science and Computational Intelligence, Jakarta, 2015.

[2]. H. Noprisson, E. Hidayat, and N. Zulkarnaim, "A Preliminary Study of Modelling Interconnected Systems Initiatives for Preserving Indigenous Knowledge in Indonesia," in 2015 International Conference on Information Technology Systems and Innovation (ICITSI), 2015, pp. 1-6.

[3]. 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.

[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]. M. O. Pratama, R. Meiyanti, H. Noprisson, A. Ramadhan, and A. N. Hidayanto, "Influencing factors of consumer purchase intention based on social commerce paradigm," in Advanced Computer Science and Information Systems (ICACSIS), 2017 International Conference on, 2017, pp. 73-80.

[7]. H. Noprisson, N. Husin, M. Utami, Puji Rahayu, Y. G. Sucahyo, and D. I. Sensuse, "The Use of a Mixed Method Approach to Evaluate m-Government Implementation," in 2016 International Conference on Information Technology Systems and Innovation (ICITSI), 2016.

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

[9]. 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.

[10]. 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.

[11]. 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.

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

[13]. 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.

[14]. W. Cheng and H. Eyke, "Combining Instance-Based Learning and Logistic Regression for Multilabel Classification," pp. 1-15, 2009.

[15]. J. Freyberger, N. T. Heffernan, and C. Ruiz, "Using Association Rules to Guide a Search for Best Fitting Transfer Models of Student Learning," 2004.

[16]. V. Rus, M. Lintean, and R. Azevedo, "Automatic Detection of Student Mental Models During Prior Knowledge Activation in MetaTutor," pp. 161-170, 2009.

[17]. M. Feng and J. Beck, "Back to the future : a non-automated method of constructing transfer models," pp. 240-249, 2009.

[18]. S. B. Kotsiantis, C. J. Pierrakeas, and P. E. Pintelas, “Preventing Student Dropout in Distance Learning Using Machine Learning Techniques,” pp. 267-274, 2003.

[19]. W. Andangsari and M. N. Suprayogi, “Personality Prediction Based on Twitter Information in Bahasa Indonesia,” vol. 11, pp. 367-372, 2017.

[20]. 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.

Cite this article as :

Umniy Salamah, "A Comparison of Text Classification Techniques Applied to Indonesian Text Dataset", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 5 Issue 6, pp. 217-222, November-December 2019.
Available at doi : https://doi.org/10.32628/CSEIT195629
Journal URL : http://ijsrcseit.com/CSEIT195629