Computer Engineering and Applications Vol. 9, No. 1, February 2020

Evaluating Naïve Bayes Automated Classification for GBAORD

Hani Febri Mustika*, Arida Ferti Syafiandini, Lindung Parningotan Manik, Yan Rianto

Research Center for Informatics, Indonesian Institute of Sciences
**Correspondence: hani.febri.mustika@lipi.go.id*

ABSTRACT

The Indonesian Government Budget Appropriations or Outlays for Research and Government (GBAORD) has been analyzed manually every year to measure the government expenditures in research and development. The analysis process involved several experts in making the budget classification. This method, commonly known as manual classification, has its downsides, which are time consumption and inconsistent result. Therefore, a study about implementing the machine learning method in GBAORD budget classification to avoid inconsistency is proposed in the previous research. For further analysis, this paper evaluates the performance of the Naïve Bayes algorithm for the GBAORD budget classification. This paper aims to measure the robustness of the Naïve Bayes to classify GBAORD data taken from 2017 until 2019. This paper uses three models of Naive Bayes with different preprocessing methods and features. This paper concludes that using the Naïve Bayes algorithm in Indonesian GBAORD budget classification is suitable since the robustness of the algorithm is proved to be high with 96.788+-0.185% average accuracy.

Keywords: Classification, Naïve Bayes, GBAORD.

1. INTRODUCTION

Indonesian Government Budget Appropriations or Outlays for Research and Development (GBAORD) is the main part of Gross Expenditure on Research and Development (GERD). The GBAORD is counted for measuring government support for research and development activities [1]. In the GERD report, the GBAORD is calculated by classifying government expenditure every year. Thousands of rows of government expenditure data are classified into six classes. They are non-research and development expenditure (non-R&D), research and development expenditure (R&D), Science and Technology Services (STS), Staff Training Expenditure (STE), Current and Capital. However, the GBAORD classification was done manually by an expert group. Manual classification needs a lot of expert effort, uses high time consumption and produces inconsistent results due to numerous rows of government expenditure data processing. Therefore, the automated classification that applies machine learning algorithms such as Naïve Bayes, Decision Tree, etc is needed as a solution to solve the problems.

The previous study about automated classification for GBAORD [2] was performed by using the Decision Tree and the Naïve Bayes. The study utilized government expenditure data in 2016 and had the conclusion that the Naïve Bayes has a higher accuracy score than the Decision Tree. This study assessed the

GBAORD automated classification with new data (government expenditure data in 2017 until 2019 that have validated by the expert group). This research also evaluated several Naïve Bayes automatic classification models with different features combinations and various data preprocessing to measure the robustness of Naïve Bayes on classifying the GBAORD.

2. RELATED WORK

The development of the classification model starts moving towards machine learning. Machine learning method develops automated classification modeling for many fields and problems. Research about automated classification has been conducted with various algorithms. Some of them were Decision Tree [2] [3], Support Vector Machine [5], Naïve Bayes [6] and other algorithms [7].

Text classification is an example of classification problem which become an active field of research and development nowadays. The solution to the problem is identic with the Naïve Bayes classifier. The previous study [8] conducted supervised machine learning for classifying lyrics text using the Naïve Bayes. The other studies [9] conducted a document classification of DRDO Tender also using the Naïve Bayes. According to [8], the Naïve Bayes classifier has good characteristics such as computational efficiency, low variance, incremental learning, direct prediction of posterior probability, robustness to noise, and robustness on missing value.

Aborisade and Anwar (2018) [10] attempted comparing the Logistic Regression and the Naïve Bayes for classifying authorship of tweets. The study concluded that the accuracy of the Logistic Regression is higher than the Naïve Bayes, but only 1,3%. For GBAORD classification, the previous study [2] attempted automated classification using the Decision Tree and the Naïve Bayes and utilized government expenditure data in 2016 as a dataset. The Naïve Bayes achieved 98,462% accuracy while the Decision Tree only had 90,236%. However, the Naïve Bayes used all features while the Decision Tree only used one feature.

2.1 CONTRIBUTION

This study evaluated GBAORD automated classification modelling with new and validated data namely government expenditure data from 2017 until 2019. Our contribution is to evaluate the robustness of the Naïve Bayes automatic classification models with different features combinations and various data preprocessing.

3. METHODOLOGY

3.1 DATA

Government expenditure data are government ministry or institution expenditure in Indonesian. Data were taken from 2017 until 2019 and validated by the expert group. Fields of data consist of government ministry or institution, unit, program,

ISSN: 2252-4274 (Print)
ISSN: 2252-5459 (Online)

function, subfunction, activity, output, sub output, component and account. There are six classes (0,1,2,3,4 and 5):

- 0 refers to non-R and D expenditure (non-R&D),
- 1 refers to R and D expenditure (R&D),
- 2 refers to Science and Technology Services (STS),
- 3 refers to Staff Training Expenditure (STE),
- 4 refers to Current, and
- 5 refers to Capital.

The sample of government expenditure data is shown in Table 1.

TABLE 1.
Sample of government expenditure data

Government ministry / institution	Unit	Program	Function	Sub function	Activity	Output	Sub output	Component	Account
001 MPR	001 01 SEKRETARIAT JENDERAL	001.01.01 Program Dukungan Manajemen dan Pelaksanaan Tugas Teknis Lainnya MPR	01 PELAYANAN UMUM	01 LEMBAGA EKSEKUTIF DAN LEGISLATIF, MASALAH KEUANGAN DAN FISKAL, SERTA URUSAN LUAR NEGERI	1001 Pengelolaan Administrasi MPR dan Sekretariat Jenderal	1001 001 Layanan Administrasi MPR dan Sekretariat Jenderal	001 Tanpa Sub Output	051 Pembinaan SDM dan Pengelolaan Administrasi Keanggotaan serta Aparatur Sipil Negara	52 BELANJA BARANG
079 LIPI	079 01 LEMBAGA ILMU PENGETAHUAN INDONESIA	079.01.01 Program Dukungan Manajemen dan Pelaksanaan Tugas Teknis Lainnya LIPI	04 EKONOMI	10 LITBANG EKONOMI	3385 Pengembangan Jaringan Kerja Sama Penelitian dan Pemasyarakatan Iptek	3385 001 Layanan Kehumasan dan Pembinaan Ilmiah	001 Hasil Pemasyarakatan IPTEK	051 Diseminasi Hasil Penelitian LIPI dan Science Briefing for Parliament	52 BELANJA BARANG
086 LAN	086 01 LEMBAGA ADMINISTRASI NEGARA	086.01.06 Program Pengkajian Administrasi	10 PENDIDIKAN	05 PENDIDIKAN KEDINASAN	3611 Penyelenggaraan Pendidikan Tinggi	3611 001 Laporan Penelitian dan Penge	001 Dokumen Penelitian Mandiri	051 Penyelenggaraan Penelitian Mandiri	52 BELANJA BARANG

Negara dan Diklat Aparatur Negara	Bidang Ilmu Adminis trasi STIA LAN Jakarta	mbang an Pendidi kan Tinggi Bidang Ilmu Admini strasi

Data preprocessing adjusts data to modelling criteria. This research attempts a combination of data representations, namely code and text that are shown in Table 2. Table 2 illustrate the code representation (CR), which only takes code of field, and text representation (TR), which take texts in the field without the code. Data preprocessing of text representation is performed in two steps. First step is transforming the text into uppercase and the next step is removing punctuation marks, number and stop words. Stop words removed from the text representation are 'yang; untuk; pada; ke; para; seperti; dan; tidak; kepada; oleh; saat; sekitar; bagi; serta; di; dari; telah; sebagai; adalah; dalam; bisa; bahwa; atau; hanya; dengan; ada; terhadap; secara; agar; daripada; lagi; tentang; seterusnya; boleh; dapat; akan; setiap; dsb; dst; dll'

TABLE 2.
Data representation

Representation	Raw Data	Data (after preprocessing)
Code representation (CR)	051 Pembinaan SDM dan Pengelolaan Administrasi Keanggotaan serta Aparatur Sipil Negara.	51
Text representation (TR)	051 Pembinaan SDM dan Pengelolaan Administrasi Keanggotaan serta Aparatur Sipil Negara.	PEMBINAAN SDM PENGELOLAAN ADMINISTRASI KEANGGOTAAN APARATUR SIPIL NEGARA

3.2. NAÏVE BAYES MODELLING

In this research, the GBAORD automated classification implements supervised machine learning algorithm using the Naïve Bayes. It is a simple modelling, yet it is effective for text classification. The Naïve Bayes classifier is a simple classifier based on applying Bayes theorem with independence assumption [11]. The Naïve Bayes is basically represented as [8]:

$$P\ (x) = \frac{P\ (\mathcal{X}|C)\ P(c)}{P\ (x)} \quad (1)$$

ISSN: 2252-4274 (Print)
ISSN: 2252-5459 (Online)

where c is a class, x is a feature, $P\ (c)$ is the prior probability of a class, $P\ (x)$ is the prior probability of feature, $P\ (c)$ is conditional probability of the class for the given feature x (likelihood), $P\ (x|c)$is the conditional probability that feature x belongs to class c (posterior probability).

According to [8], the Naïve Bayes has the possibility of easy parallelization, especially for large datasets. Three different models are evaluated in this study. In Table 3, the first model utilized ten features, the second model used only five features and the third model applied four features. All features were preprocessed to extract the code representation (CR) and/or the text representation (TR).

TABLE 3.
Features of three Naïve Bayes models for GBAORD automated classification

First Model	Second Model	Third Model
Government ministry / institution (CR)	Program (CR)	Program (CR)
Unit (CR)	Sub function (CR)	Sub function (CR)
Program (CR)	Output (CR)	Sub output (TR)
Function (CR)	Sub output (TR)	Component (TR)
Sub function (CR)	Component (TR)	
Activity (CR)		
Output (CR)		
Sub output (TR)		
Component (TR)		
Account (TR)		

Government ministry/institution, unit, program, function, sub function, activity, and output features utilized code representation because all codes have consistent text. For an instance, program with code “001.01.01” equals to “Program Dukungan Manajemen dan Pelaksanaan Tugas Teknis Lainnya MPR” and it is consistent for all rows. Thus, the code representation is enough to represent the data. Meanwhile, in sub output, component, and account feature, the same code could have different text or substance data. Therefore, they used text representation.

4. RESULTS AND DISCUSSION

The result of the evaluation of the three models of the Naïve Bayes algorithm using a combination of features and preprocessing is shown in Table 4. It informs the results of the evaluation of the Naïve Bayes automated classification with training and testing using the 2017 data. The result was measured to calculate error rate, average error rate, standard deviation, and average accuracy. The smallest average error rate is achieved by the second model. Meanwhile, the third model has the best value of standar deviation with 0.171. However, the second model also has 96.788 % average accuracy value. Thus, the second model is chosen as the selected model to evaluate the 2018 and the 2019 data.

TABLE 4.
Evaluation result of three Naïve Bayes models

Test	Error rate		
	First Model	Second Model	Third Model
1	4,172	3,222	3,157
2	4,736	3,415	3,028
3	4,446	3,028	3,093
4	4,430	3,431	3,624
5	4,333	3,334	3,383
6	4,350	3,464	3,029
7	4,463	3,029	3,222
8	5,188	3,238	3,415
9	4,350	3,109	3,206
10	4,753	2,852	3,190
Average Error rate	4,522	**3,212**	3,235
Std. Deviation	0,266	0,185	**0,171**
Average Accuracy	95,478	**96,788**	96,765

According to Hossin and Sulaiman [12], accuracy measures the ratio of correct predictions over the total number of instances evaluated. Meanwhile, the error rate for misclassification error measures the ratio of incorrect predictions over the total number of instances evaluated. Sensitivity is used to measure the fraction of positive patterns that are correctly classified, and specificity is utilized to measure the fraction of negative patterns that are correctly classified. Precision measures the positive patterns that are correctly predicted from the total predicted patterns in a positive class. Recall indicates the fraction of positive patterns that are correctly classified. F-measure represents the harmonic mean between recall and precision values. Average accuracy is used to show average effectiveness of all classes.

Table 5 shows the evaluation value of the second model using the 2017 data as training data and the 2018 data as testing data. The performance metrics for each class was evaluated by measuring recall, precision, sensitivity, specificity and F-measure. In the class 0 or non R&D class, all metric values are more than 0.94 since many rows of training data are classified as class 0. Other classes have recall more than 0.80, but precision values are ranging between 0.50 and 0.79. It means the positive patterns that are correctly predicted from the total predicted patterns in a positive class. However, sensitivity values for class 1 until 5 are between 0.73 and

ISSN: 2252-4274 (Print)
ISSN: 2252-5459 (Online)

0.84. It indicates that the fractions of positive patterns are correctly classified. Moreover, specificity values reaches more than 0.98 for all classes. It concluded that the fractions of negative patterns are also correctly classified. The F-measure values for each class are ranging between 0.62 and 0.82. Additionally, the Cohen's Kappa value that indicates measurement consistency from the second model is 0.8265 and the accuracy achieves 0.9553.

TABLE 5.
Evaluation result of the second model (training data 2017, testing data 2018)

Class	Recall	Precision	Sensitivity	Specificity	F-measure
0	0,9803	0,9909	0,9803	0,9432	0,9856
1	0,7376	0,7980	0,7376	0,9888	0,7666
2	0,8342	0,7112	0,8342	0,9903	0,7678
3	0,8429	0,6541	0,8429	0,9973	0,7366
4	0,8497	0,7965	0,8497	0,9911	0,8222
5	0,8075	0,5051	0,8075	0,9943	0,6215

Table 6 shows the evaluation results using the 2017 data as training data and the 2019 data as testing data. For class 0, the recall, precision, sensitivity, specificity and F-measure values are more than 0.92. Meanwhile, the precision values of class 3 and 5 are quite small, 0.3, and the F-measure is a little bit higher than 0.4. It means that the pattern positive is not classified clearly. In general, the second model has the accuracy value of 0.9302 and the Cohen's Kappa value of 0.7367.

TABLE 6.
Evaluation result of the second model (training data 2017, testing data 2019)

Class	Recall	Precision	Sensitivity	Specificity	F-measure
0	0,9762	0,9880	0,9762	0,9280	0,9821
1	0,6206	0,7076	0,6206	0,9832	0,6612
2	0,6662	0,5172	0,6662	0,9826	0,5823
3	0,5438	0,3233	0,5438	0,9954	0,4055
4	0,6700	0,6842	0,6700	0,9864	0,6771
5	0,8235	0,3491	0,8235	0,9907	0,4903

5. CONCLUSION

Evaluation of the Naïve Bayes classifier for GBAORD in this research concludes that the features combination and the data preprocessing affected the robustness of automated classification. Based on the result, the Naïve Bayes automated classifier using all features in the first model, yields low accuracy. Meanwhile, the second model using only five features, namely program, sub function, output, sub output, and component, with combination of data preprocessing, which is used to extract the data in order to represent the value and the meaning of the data, affected the accuracy of the classifier significantly. The combination of selected features in the modelling process improves the accuracy of automated classification. It achieved the average accuracy of 96.788%, which is the better than the other models. Automated classification using the Naïve Bayes algorithm for Indonesian GBAORD is suitable since the robustness of the algorithm is proved to be high with 96.788+-0.185%.

ACKNOWLEDGEMENTS

This research is supported by Research Center for Informatics (P2I) and Center for Science and Technology Development Studies (PAPPIPTEK) Indonesia Institute of Sciences and Indonesian Ministry of Research, Technology and Higher Education.

REFERENCES

[1] OECD, *Frascati Manual -2015 : Guidelines for Collecting and Reporting Data on Research and Experimental Development*, 2015th ed., no. October. Paris: OECD Publishing, 2015.

[2] A. F. Syafiandini, H. F. Mustika, and Y. Rianto, "Classification of Indonesian Government Budget Appropriations or Outlays for Research and Development (GBAORD) Using Decision Tree and Naive Bayes," in *2018 Third International Conference on Informatics and Computing (ICIC)*, 2018, pp. 1–5.

[3] X. Ma and C. Liu, "Information extraction of mining-land use based on automatic threshold classification of decision tree," in *2011 International Conference on Remote Sensing, Environment and Transportation Engineering, RSETE 2011 - Proceedings*, 2011, pp. 421–424.

[4] L. Hu, Z. Yu, and Y. Liu, "An algorithm of decision-tree generating automatically based on classification," in *Proceedings of the 1st International Workshop on Education Technology and Computer Science, ETCS 2009*, 2009, vol. 1, pp. 823–827.

[5] J. Liu and L. Xie, "SVM-based automatic classification of musical instruments," in *2010 International Conference on Intelligent Computation Technology and Automation, ICICTA 2010*, 2010, vol. 3, pp. 669–673.

[6] Y. Cheng, Z. Li, H. Wang, P. Zhou, and W. Zhou, "The design of automatic classification system of integrated emergency response system," in *Proceedings of the 2009 International Conference on Communication Software and Networks, ICCSN 2009*, 2009, pp. 675–678.

[7] L. Gan, R. Benlamri, and R. Khoury, "Improved sentiment classification by multi-modal fusion," in *Proceedings - 3rd IEEE International Conference on Big Data Computing Service and Applications, BigDataService 2017*, 2017, pp. 11–16.

[8] D. Buzic and J. Dobsa, "Lyrics classification using Naive Bayes - IEEE Conference Publication," in *2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, 2018, pp. 1011–1015.

[9] S. Goswami, P. Bhardwaj, and S. Kapoor, "Naïve bayes classification of DRDO tender documents," in *2014 International Conference on Computing for Sustainable Global Development, INDIACom 2014*, 2014, pp. 593–597.

ISSN: 2252-4274 (Print)
ISSN: 2252-5459 (Online)

[10] O. M. Aborisade and M. Anwar, "Classification for authorship of tweets by comparing logistic regression and naive bayes classifiers," in *Proceedings - 2018 IEEE 19th International Conference on Information Reuse and Integration for Data Science, IRI 2018*, 2018, pp. 269–276.

[11] Y. An, S. Sun, and S. Wang, "Naive Bayes classifiers for music emotion classification based on lyrics," in *Proceedings - 16th IEEE/ACIS International Conference on Computer and Information Science, ICIS 2017*, 2017, no. 1, pp. 635–638.

[12] H. M and S. M.N, "A Review on Evaluation Metrics for Data Classification Evaluations," *Int. J. Data Min. Knowl. Manag. Process*, vol. 5, no. 2, pp. 1–11, 2015.