Implementation of Naïve Bayes updateable with modified absolute discount smoothing on Pamekasan Regent SMS center data classification

Implementation of Naïve Bayes updateable with modified absolute discount smoothing on Pamekasan Regent SMS center data classification

B Said* and N P Dewi

Department of Informatics Engineering, Universitas Madura, Jl. Panglegur Km 3,5 Pamekasan, Indonesia

*badarsaid@unira.ac.id

Abstract. Classification is a grouping process based on a predetermined class. Previous research has classified Regent Pamekasan SMS Center using Naïve Bayes and Modified Absolute Discounting (MAD) Smoothing, but the average classification accuracy is still equal to 76.83%. to improve the accuracy of classification then in this study applied Naïve Bayes Updateable by using MAD Smoothing. The classes used remain 15 classes: Education, Health, Infrastructure, Crime, Administrative Services, Sports, Government, Agriculture, Small and Medium Enterprises, Order, Weak Economy, Religion, Art and Culture, Natural Disasters, and Others. Before doing the classification process first done pre-processing such as equating characters, deletion of punctuation, restore abbreviation, translation of the local language (Madura), deletion of numbers, deletion of words that are not important in SMS, and stemming to convert into a basic word. Results Some experiments obtained an average accuracy of 78.89%, with the accuracy of one test reached 87.65%. And Naïve Bayes Updateable can increase accuracy by 2.07% with the addition of 0.47-minute classification time.

1. Introduction

Text classification is often done to bring up hidden knowledge. To classify data there are several methods that can be used such as SVM, Naïve Bayes, KNN and so on. The method often used for text classification is Naïve Bayes, because this technique is known as the best technique in terms of computational time compared to other data mining algorithm techniques [1], namely classification with probability methods by predicting future opportunities based on experience in the future previous. In the implementation of this method, there are several smoothing methods that can be done including Jelinek-Mercer (JM), Dirichlet (Dir), Absolute discounting (AD) and Two-stage (TS) [2].

In a previous study, the classification was done using Naïve Bayes with several smoothing methods, the results proved that Absolute Discounting (AD) Smoothing is better than others in improving classification accuracy. In addition, classification is also done using Naïve Bayes with Modified Absolute Discount (MAD) Smoothing, the results prove an accuracy increase of 4% when compared to the use of Absolute Discount Smoothing [3].

The classification of the Pamekasan Regent SMS Center data had previously been carried out using Naïve Bayes with MAD Smoothing, but the average classification accuracy still reached 76.83% [4]. Therefore, it is necessary to conduct further research using Naïve Bayes Updateable which has been shown to improve classification accuracy but still does not use the smoothing method [5]. So that in

this study a combination of Naïve Bayes Updateable was carried out with MAD Smoothing to improve classification accuracy.

2. Literature review

2.1. Naïve Bayes

In this study, the Naïve Bayes method was used. This method was chosen because it has been tested in several studies to produce good accuracy [6]. Naïve Bayes is a machine learning method that has a model in the form of probability or opportunity. Training data in the form of SMS and class pairs are used as sources of formation of the analysis model. Every feature that represents an SMS has calculated the probability of its occurrence in each class.

2.2. Modified absolute discount smoothing

To maximize the performance of Naïve Bayes in a previous study, the authors used Modified Absolute Discount (MAD) Smoothing which was proven to improve the accuracy of Naïve Bayes accuracy in classification [4]. The formula for determining the class of an SMS is shown by the following equation [3]:

$$P(w_k|C_i) = \frac{\max(count(w_k, C_i) - delta, 0) + delta\left(N_{uC_i}\right)f(w_k)}{\sum_{w \in V} count(w, C_i)} \quad (1)$$

N_{uC_i}= the number of unique words in *Ci*

$$f(w_k) = P_{unif}(w_k)\sum_{j=1}^{m} count(w_k, C_j)$$

$$P_{unif}(w_k) = \frac{1}{|V|}$$

2.3. Naïve Bayes updateable

To further improve the classification accuracy, the Naïve Bayes Updateable method is applied. This method is an incremental form of the Naive Bayes Classifier by learning one instance or one batch at a time. The incremental learning process, in this case, is that training data gradually increases, each batch of testing data that has been processed will be used as knowledge for the next classification.

2.4. Enhanced Confix Stripping

Eliminating affixes to words to get basic words is also done to increase the accuracy of the classification results. in this case, the author uses the Enhanced Confix Stripping method because it is the best-performing word stemming algorithm in Indonesian [7].

3. Methodology

The number of classes is still adapted to previous researches namely Education, Health, Infrastructure, Crime, Administrative Services, Sports, Government, Agriculture, SMEs, Order, Weak Economy, Religion, Art and Culture, Natural Disasters, and Others.

The preprocessing process that is done is equalizing the characters, removing punctuation, returning abbreviations, translating regional languages (Madurese), deleting numbers, removing words that are not important in SMS, and stemming to change the word to basic words.

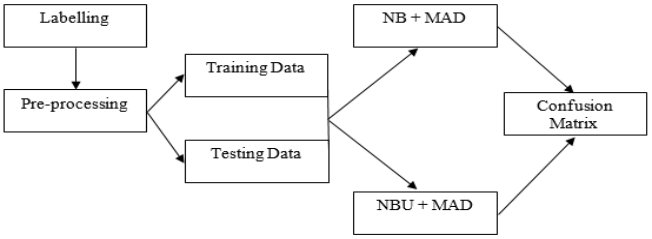

Figure 1. Research flow.

To improve the classification accuracy, Naïve Bayes Updateable is applied and still uses MAD Smoothing. Variation of the test is done based on the month of SMS received. Evaluation uses Confusion Matrix to get classification accuracy, that is, taking into account incorrect classification and correct classification. The time needed in the classification process is also calculated from each trial.

4. Results and discussion

The labeling process is carried out one by one for all data, namely 2134 SMS. In this labeling stage sometimes SMS is found that can be classified more than one category or class, for that case one class remains selected which is more dominant than the other class based on the number of words in the text content.

Preprocessing stages are also carried out by the system. Classification test process with six variations of training data and test data with five-month provisions as training data and one month as test data. Evaluation is carried out on all the results of the classification test by taking into account the accuracy and time required.

evaluation of classification accuracy is done by paying attention to the results in the confusion matrix for each variation of the classification test. Table 1 shows the comparison of accuracy and average generated.

Table 1. The Accuracy of each classification test.

Method	Accuracy (%)						Average
	1	2	3	4	5	6	
NB MAD Smoothing	82,68	75,87	73,76	77,64	76,33	74,69	76,83
NB Updateable MAD *Smoothing*	87,65	76,91	77,59	78,20	80,12	74,58	78,89
Difference							2,07

In Table 1 it can be seen that the average accuracy of the classification trials using Naïve Bayes with MAD Smoothing is 76.83%, while the average accuracy of classification trials using Naïve Bayes Updateable with MAD Smoothing is 78.89%. Proven classification using Naïve Bayes Updateable with MAD Smoothing can improve classification accuracy by 2.7%. In addition, the analysis of the time required in the classification trial process is shown in Table 2 below:

Table 2. Time needed for classification.

Method	Time (s)						Average
	1	2	3	4	5	6	
NB MAD Smoothing	1,2	3,23	3,35	2,46	3,09	1,51	2,47
NB Updateable MAD *Smoothing*	1,7	3,73	3,85	2,86	3,59	1,91	2,94
Difference							0,47

In table 2 it can be seen that the average time required for the classification process using Naïve Bayes with MAD Smoothing is 2.47 minutes, while the average time required for the classification test process using Naïve Bayes Updateable with MAD Smoothing is 2.94 minutes. So the time needed to process the Naïve Bayes classification is updated with MAD Smoothing 0.47 minutes longer.

5. Conclusion

Based on the facts that occurred during the research and analysis of the results of the study it can be concluded that Naive Bayes Updateable with Modified Absolute Discounting Smoothing proven to increase classification accuracy by 2.7%. This is because training data increases continuously and allows the addition of unique words from the classification results. The Naive Bayes Updateable implementation requires an additional time of 0.47 minutes because the longer it takes to process training data that increases continuously. The drawback in implementing Naive Bayes Updateable with MAD Smoothing is that it has the possibility of reducing accuracy. This is done if the results of the classification are wrong and are immediately used as training data for the next classification process. This happened at the 6th trial.

References

[1] Dwi W 2011 *Analisa Perbandingan Algoritma SVM, Naive Bayes, Dan Decision Tree Dalam Mengklasifikasikan Serangan (Attacks) Pada Sistem Pendeteksi Intrusi Jurusan Sistem Informasi Universitas Gunadarma*

[2] Q Yuan, G Cong and N M Thalmann 2012 *Enhancing Naive Bayes with Various Smoothing Methods for Short Text Classification WWW Companion*

[3] A Chharia and R K Gupta 2013 *Enhancing Naïve Bayes Performance with Modified Absolute Discount Smoothing Method in Spam Classification IJARCSSE*

[4] Said B, Gunawan and Pranoto Y M 2015 *Klasifikasi Data SMS Center Bupati Pamekasan menggunakan Naïve Bayes dengan MAD Smoothing Ideatech*

[5] Saptono R, Sulistyo M E and Trihabsari N S 2016 *Text Classification Using Naive Bayes Updateable Algorithm in SBMPTN Test Questions TELEMATIKA*

[6] Ahmed I, Guan D and Choong T 2014 *SMS Classification Based on Naïve Bayes Classifier and Apriori Algorithm Frequent Itemset International Journal of Machine Learning and Computing*

[7] Arifin A Z, I P A K Mahendra and H T Ciptaningtyas 2009 *Enhanced Confix Stripping Stemmer and Ants Algorithm for Classifying News Document in Indonesian Language ICTS Komputindo*