ABSTRACT

Short Message Service (SMS) is one of the communication media that is cheap and easy to use. Because of the cheapness SMS service tariffs, the irresponsible parties commit criminal acts such as SMS spam which aimed at fraud. It causes a loss for people who use SMS service. Therefore, a classification technique is required to distinguish SMS messages between SMS spam and ham.

Naïve Bayes Classifier (NBC) is one of the most widely used classification algorithm for SMS classification because of the high degree accuracy than any other algorithms. However, NBC has a disadvantage that it depends only on the probability distribution. Correlated Naïve Bayes Classifier (CNBC) is a new algorithm that can overcome that weakness from the NBC algorithm. The result of several studies on the CNBC algorithm have proven that CNBC has higher accuracy than NBC. However, CNBC has a weakness in the calculation of the correlated coefficient. If all values of a feature in the training data are zero then it will produce an error value which will cause the decrease performance of CNBC. One method to overcome the number of zero values on a feature in the training data is Principal Component Analysis (PCA). PCA can reduce a feature as well as decompose the value of each feature without compromising the important information in it.

In this study, an experimental classification of SMS data consisting of 43 spam SMS and 57 SMS ham using NBC algorithm, CNBC and PCA integration with CNBC. PCA method is applied to overcome the weakness of CNBC algorithm that is error value on the calculation of correlation coefficient value. Based on the test results, the integration of PCA with CNBC successfully improved the accuracy of CNBC algorithm from 56.63% to 68.23% on the classification of the Indonesian spam SMS dataset.

Keywords – classification, SMS spam, Naïve Bayes Classifier, Correlated Naïve Bayes Classifier, Principal Component Analysis.

INTISARI

Short Message Service (SMS) merupakan salah satu media komunikasi yang murah dan mudah dalam penggunaannya. Dengan murahnya tarif SMS, banyak pihak-pihak yang tidak bertanggungjawab melakukan tindak kejahatan seperti SMS spam yang bertujuan untuk penipuan. Hal ini menyebabkan kerugian bagi masyarakat yang menggunakan layanan SMS. Oleh karena itu, diperlukan teknik klasifikasi untuk membedakan SMS spam dan *ham*.

Naïve Bayes Classifier (NBC) merupakan salah satu algoritme klasifikasi yang banyak digunakan untuk klasifikasi SMS, karena memiliki tingkat akurasi yang tinggi daripada algoritme yang lain. Namun, NBC memiliki kelemahan yaitu hanya bergantung pada distribusi probabilitas. Ada algoritme terbaru yang dapat mengatasi kelemahan dari algoritme NBC yaitu *Correlated Naïve Bayes Classifier* (CNBC). Hasil dari beberapa penelitian tentang algoritme CNBC, telah terbukti bahwa CNBC memiliki tingkat akurasi lebih tinggi daripada NBC. Namun, CNBC memiliki kelemahan pada perhitungan koefisien korelasi yang apabila semua nilai dari suatu fitur pada data latih bernilai nol maka akan menghasilkan nilai *error* yang akan menyebabkan turunnya performa dari CNBC. Salah satu metode untuk mengatasi banyaknya nilai nol pada suatu fitur pada data latih yaitu metode *Principal Component Analysis* (PCA). PCA dapat mereduksi suatu fitur sekaligus mendekomposisi nilai dari setiap fitur tanpa mengurangi informasi penting didalamnya.

Pada penelitian ini, dilakukan percobaan klasifikasi data SMS yang terdiri dari 43 SMS spam dan 57 SMS *ham* dengan menggunakan algoritme NBC, CNBC dan integrasi PCA dengan CNBC. Metode PCA diterapkan untuk mengatasi kelemahan dari algoritme CNBC yaitu nilai *error* pada perhitungan nilai koefisien korelasi. Berdasarkan hasil pengujian, integrasi PCA dengan CNBC berhasil memperbaiki tingkat akurasi algoritme CNBC dari 56,63% menjadi 68,23% pada klasifikasi *dataset* SMS spam berbahasa Indonesia.

Kata kunci --Klasifikasi, SMS spam, *Naïve Bayes Classifier*, *Correlated Naïve Bayes Classifier*, *Principal Component Analysis*.