INTISARI

DETEKSI SPAM PADA SMS BAHASA INDONESIA MENGGUNAKAN SELEKSI FITUR BERBASIS *PARTICLE SWARM OPTIMIZATION*

Oleh

Dyah Mawar Umbulsari
15/378058/PA/16533

Permasalahan spam pada *Short Message Service* (SMS) dapat diatasi salah satunya dengan cara mengklasifikasi atau memisahkan pesan spam dengan pesan bukan spam. SMS yang masuk idealnya dapat langsung dideteksi sebagai pesan spam atau bukan. Permasalahan ini menunjukan bahwa waktu sangat krusial, namun sistem klasifikasi teks memiliki permasalahan tersendiri yaitu besar dimensi data berbanding lurus dengan jumlah kata dalam teks. Besar dimensi data padahal akan mempengaruhi lama waktu yang diperlukan untuk mengenali data tersebut.

Penelitian ini bertujuan untuk mereduksi dimensi data tersebut dengan menerapkan algoritme *Particle Swarm Optimization (PSO)* pada tahap seleksi fitur dan melihat apakah performa klasifikasi dapat dipertahankan walaupun fitur datanya berkurang. PSO adalah algoritme random berbasis populasi untuk menemukan solusi optimal yang terinspirasi dari tingkah laku sekelompok burung. Model klasifikasi pada penelitian ini dibangun berdasarkan algoritme *Gaussian Naive Bayes* dengan beberapa tahapan dalam sistem yaitu pengumpulan data, *preprocessing*, seleksi fitur, klasifikasi dan evaluasi. Validasi yang dilakukan terhadap model menggunakan metode *Stratified K-Fold Cross Validation* dengan nilai k=10 dan paramaeter evaluasi berupa akurasi, presisi, *recall* dan *f1 score.*

Hasil pengujian dan evaluasi model yang dibangun menunjukan terdapat peningkatan pada model dengan penambahan seleksi fitur PSO, yaitu sebesar 0,09 pada rata-rata nilai akurasi dari 0,84 menjadi 0,93 dan peningkatan sebesar 0,034 pada rata-rata nilai *f1-score* dari 0,919 menjadi 0,953 yang disertai pengurangan jumlah fitur sebanyak 324 fitur untuk dataset yang dikumpulkan secara manual dari 822 fitur menjadi 498 fitur. Model yang dibangun berdasarkan dataset publik menunjukkan peningkatan sebesar 0,0026 pada rata-rata nilai akurasi dari 0,908 menjadi 0,9106 dan peningkatan sebesar 0,003 untuk rata-rata nilai *f1-score* dari 0,908 menjadi 0,911 yang disertai pengurangan fitur sebanyak 1427 fitur dari 3382 fitur menjadi 1955 fitur.

Kata kunci: Deteksi spam, klasifikasi, *Particle Swarm Optimization, Gaussian Naive Bayes*, seleksi fitur

ABSTRACT

SPAM DETECTION ON INDONESIAN SMS WITH PARTICLE OPTIMIZATION BASED FEATURE SELECTION

By

Dyah Mawar Umbulsari
15/378058/PA/16533

Spam problem in Short Message Service (SMS) can be resolved with which one of the solution is by classification or classifying text messages into spam or non spam. The classification system ideally should be able to detect spam messages as soon as the text arrives, thus time is very crucial in this problem. Classifying text itself meanwhile also have a problem and that is the dimension size of the data. The bigger size of the data dimension is, the longer it takes for the classifier to learn that data.

The research aims to reduces the data features by applying Particle Swarm Optimization (PSO) method in feature selection. The population based randomized algorithm PSO was inspired by the behaviour of a flock of birds in finding source of food or in this research represented by the optimal solution. The classifier that is used in this research is Gaussian Naive Bayes, the validation is done using the Stratified K-Fold Cross Validation method with the value of k=10. The evaluation of the model is measured using parameters such as accuracy, precision, recall and f1-score.

The results of system implementation and evaluation shows that by implementing PSO in feature selection the classification system could achieve 0,09 uplift average in accuracy from 0,84 to 0,93 and 0,034 uplift average in f1-score from 0,919 to 0,953 with 324 features reduced in manually collected dataset from 822 features to 498 features. The system achieved 0,0026 uplift average in accuracy from 0,908 to 0,9106 and 0,003 uplift average in f1-score from 0,908 to 0,911 with 1427 reduced features from 3382 features to 1955 features while using the public dataset.

Keywords: Spam detection, classification, Particle Swarm Optimization, Gaussian Naive Bayes, feature selection