INTISARI

PERBANDINGAN METODE SELEKSI FITUR *FILTER*, *WRAPPER* DAN *HYBRID* PADA KLASIFIKASI KOMENTAR *SPAM*

Oleh

Amalia Nur Anggraeni
17/418620/PPA/05404

Pertumbuhan internet menyebabkan penggunaan media sosial untuk berbagai kepentingan meningkat. Beberapa pihak yang tidak bertanggung jawab memanfaatkan fitur komentar pada sosial media untuk mengambil keuntungan dengan memberikan komentar yang tidak relevan dengan objek yang dibagikan atau diposting. Komentar tersebut termasuk dalam salah satu jenis *spam*. *Spam* untuk beberapa kasus dapat merugikan pengguna. Salah satu pendekatan untuk menyelesaikan permasalahan *spam* yaitu dengan *content base filtering*. Filterisasi dilakukan menggunakan teknik klasifikasi teks. Klasifikasi dilakukan berdasarkan teks dari komentar. Variasi teks menyebabkan jumlah fitur yang harus diproses besar sehingga dapat memberikan pengaruh terhadap performa suatu algoritma klasifikasi.

Metode yang digunakan untuk mengatasi masalah fitur yang besar adalah *feature selection*. Seleksi fitur dilakukan untuk mendapatkan fitur terbaik (*optimal feature*). Penelitian ini melakukan dan membandingkan hasil seleksi fitur metode *filter* menggunakan *Chi Square*, metode *wrapper* menggunakan *Sequential Feature Selection* (SFS) dan gabungan keduanya (*hybrid*) untuk mengetahui seleksi fitur terbaik dalam melakukan klasifikasi komentar *spam* dan *nonspam* berbahasa Indonesia. Klasifikasi menggunakan *Multinomial Naïve Bayes* dan *Support Vector Machine*.

Berdasarkan hasil pengujian dengan data latih sejumlah 4944 komentar dan data uji sejumlah 100 komentar, akurasi terbaik dicapai menggunakan hasil seleksi fitur metode kombinasi seleksi fitur *Chi Square* dan *Sequential Forward Selection* dengan *subset* 500 fitur terbaik. Peningkatan akurasi pada klasifikasi MNB mencapai 8% sedangkan pada klasifikasi SVM mencapai 4% jika dibandingkan dengan hasil akurasi sebelum menggunakan seleksi fitur. Selain mampu memberikan perbaikan akurasi, seleksi fitur *hybrid* mampu menghemat waktu komputasi.

Kata Kunci: *text classification, feature selection, spam comment, naïve bayes, support vector machine*

ABSTRACT

COMPARISON OF FILTER, WRAPPER AND HYBRID FEATURE SELECTION METHODS IN SPAM COMMENT CLASSIFICATION

By

Amalia Nur Anggraeni
17/418620/PPA/05404

The continuous growth of the internet has led to the use of social media for various purposes increase. Some irresponsible parties take advantage of the comment feature on social media platforms to harm others by providing spam comments on the shared object. One of the approaches to resolve this challenge includes content-based filtering, performed using a comments-based text classification technique. In addition, variations in text requires processing large numbers of features, which potentially influence the classification algorithm performance.

This problem is possibly solved through feature selection, in order to obtain the optimal variants. This study performs and compares the results of feature selection using the Chi Square filter method, the wrapper method using Sequential Feature Selection (SFS) and the combination (hybrid) to determine the best feature selection in classifying spam and non-spam comments in Indonesian. Classification using Multinomial Naïve Bayes and Support Vector Machine.

Based on the test results with 4944 training data and 100 test data comments, the best accuracy was achieved using the outcome from feature selection methods, involving the combination of Chi-Square and Sequential Forward Selection, to generate a subset comprising the best 500. The MNB classification accuracy increased to 8%, while SVM was 4%, compared to the results obtained before adopting feature selection. Besides to provide improved accuracy, hybrid feature selection can save computational time.

Keyword**:** *text classification, feature selection, spam comment, naïve bayes, support vector machine*