ABSTRACT

Email is one of the communication media that is easy, cheap, fast, and rapidly growing along with the development of internet technology. With the continuing increase of email users, the amount of unwanted email–often called as spam–is also increasing. Email spam can be unwanted promotional products, causing inconvenience in accessing email or it can also contain security harmful content such as viruses or malware.

Email spam problem can be solved by separating email spam and email nonspam based on text content. In this research, email spam and email nonspam are separated using feature weighting method and machine learning method. Feature weighting method is needed to determine the keywords that have relevance to email spam, while the method of machine learning is required to classify email based on keywords with given weight.

This research builds a model for the classification of spam email using a wide variety of feature weighting methods, namely TF-IDF, TF-RF, Delta TF-IDF, ICF-Based and TF-LRR. The machine learning method uses SVM and Naïve Bayes. Based on testing, the TF-LRR feature weighting method produces the best result compared to other method in term of accuracy, precision, and recall value. In the case of classification method, SVM gives better and more stable performance compared to that of Naïve Bayes.

Keywords—classification, email, spam, feature weighting

INTISARI

Email merupakan salah satu media komunikasi yang mudah, murah, dan cepat serta berkembang dengan pesat seiring dengan perkembangan teknologi internet. Dengan terus bertambahnya pengguna layanan *email*, jumlah *email* yang tidak diinginkan yang sering disebut *spam* juga semakin bertambah. *Email spam* dapat berupa promosi produk yang tidak diinginkan pengguna sehingga menimbulkan ketidaknyamanan dalam mengakses *email*. Selain itu, *email* juga dapat memuat konten yang membahayakan keamananan pengguna seperti virus ataupun *malware*.

Masalah *email spam* dapat diatasi dengan memisahkan antara *email spam* dan *email nonspam* berdasarkan konten teks yang dimiliki. Pada penelitian ini, pemisahan antara *email spam* dan *email nonspam* dilakukan dengan menggunakan metode pembobotan fitur dan *machine learning*. Metode pembobotan fitur diperlukan untuk menentukan kata kunci yang memiliki relevansi terhadap *email spam*. Sedangkan metode *machine learning* diperlukan untuk mengklasifikasikan *email* berdasarkan kata kunci yang telah diberi bobot.

Penelitian ini membangun model untuk klasifikasi *email spam* menggunakan berbagai macam variasi dari metode pembobotan fitur yaitu TF-IDF, TF-RF, Delta TF-IDF, ICF-Based, dan TF-LRR. Sedangkan metode *machine learning* yang digunakan yaitu SVM dan *Naïve Bayes*. Berdasarkan pengujian, metode pembobotan fitur TF-LRR memberikan performa paling baik dibandingkan metode pembobotan fitur lainnya pada nilai akurasi, presisi, dan sensitivitas. Sedangkan untuk metode klasifikasi, SVM memberikan performa yang lebih baik dan stabil dibandingan *Naïve Bayes*.

Kata Kunci : klasifikasi, *email*, *spam*, pembobotan fitur