Abstrak

Email telah menjadi salah satu alat komunikasi yang murah dan cepat. Masalah utama yang dihadapi adalah meningkatnya jumlah email komersial yang tidak diharapkan atau biasa disebut spam. Spam berdampak negatif seperti penyalahgunaan bandwith koneksi internet, mengurangi ukuran penyimpanan data, meningkatkan waktu komputasi dan sangat mengganggu pengguna.

Pendekatan yang banyak dilakukan untuk mendeteksi spam menggunakan metode representasi sekumpulan kata. Namun isi spam seringkali terdapat kata-kata yang salah secara tatabahasa dan menggunakan variasi tanda baca yang aneh seperti ‘f.r.e.e.’,’f-r-e-e’,’f r e e’. Hal ini mengakibatkan pendekatan ini tidak tangguh pada kondisi tersebut. Selain itu, metode ini juga perlu dilakukan proses pembuangan tanda baca, stemming dan lemmatisasi yang sangat bergantung pada bahasa yang tertentu.

Pada tugas akhir ini, pendekatan yang dilakukan untuk mendeteksi spam menggunakan representasi sekumpulan karakter *n*-grams. Pendekatan ini berusaha mengatasi permasalah yang dihadapi menggunakan metode representasi sekumpulan kata. Namun jumlah feature yang dihasilkan masih sangat besar, sehingga digunakan algoritma klasifikasi Support Vector Machine(SVM) yang mampu mengatasi ruang data dimensi data yang tinggi.

Hasil penelitian menunjukkan bahwa sistem pendeteksian spam menggunakan karakter *n*-grams dapat diterapkan dengan baik. Metode karakter *n*-grams memiliki kelebihan yaitu dapat menghindari penggunaan stop list, stemming dan lemmatisasi yang sangat bergantung pada bahasa tertentu. Hasil penelitian menunjukkan panjang karakter *n* terbaik adalah *n*=4 untuk tipe pembobotan *binary* dan *n*=5 untuk tipe pembobotan *term frequency*(TF).

Kata kunci: *spam detection, character n-grams, support vector machine(SVM).*