Abstrak

Email filtering (pemfilteran email) dilakukan untuk memisahkan email yang merugikan dari email yang tidak merugikan. Masalah pemfilteran email adalah masalah *Text Categorization* (TC), dimana hanya ada 2 kelas yaitu kelas spam dan kelas legitimate/ham/nonspam. Nonspam mail adalah email yang tidak merugikan si penerima email. Sedangkan spam (*Stupid Pointless Annoying Messages*) adalah email yang merugikan, karena selain memakai banyak ruang memori pada komputer juga menyebabkan penerima di bawah umur mengakses situs-situs yang tidak seharusnya. Salah satu metode untuk menangani spam mail adalah *Statictical Filtering*.

Model yang menerapkan *statistical filtering* adalah *Markov Random Field*, namun tidak hanya memperhitungkan kata melainkan juga frasa. Hubungan antar kata diperhatikan dan nantinya akan membentuk kata dan frasa berdasarkan ukuran *neighborhood*-nya. Pembentukan kata dan frasa adalah menggunakan teknik *Sparse Binary Polynomial Hashing* (SBPH). Kata dan frasa yang terbentuk ini sering disebut fitur. Setiap fitur akan diberi bobot dengan menggunakan teknik pembobotan *Exponential Weighting Sequences* atau sering disebut *Exponential Series* (ES) dan *Minimum Weighting Sequences* (MWS). Penerapan model MRF ini dilakukan dengan memperhatikan hubungan neighborhood antar fitur yang bertetangga. Fitur yang termasuk dalam suatu *neighborhood* disebut sebagai *cliques*.

Ukuran neighborhood yang menghasilkan keakurasian yang tinggi didapati pada ukuran window 5 dan 6 dengan pembobotan ES yang mencapai 86.67% yang berada pada threshold 0.9090. Parameter selain akurasi yang diuji adalah spam *precision*, nonspam *precision*, spam *recall*, nonspam *recall*, akurasi, dan *f-measure*. Berdasarkan hasil pengujian, MRF terbukti dapat meningkatkan hasil klasifikasi dan menghasilkan nilai akurasi yang baik.

Kata Kunci: *email filtering, statistical filtering,* fitur, *Markov Random Field* (MRF), SBPH, ES, MWS, *neighborhood, cliques.*