1. Pendahuluan

1.1 Latar belakang masalah

Pada saat ini, 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. Pesan spam berdampak pada penyalahgunaan bandwith koneksi internet, mengurangi ukuran penyimpanan data, meningkatkan waktu komputasi dan sangat mengganggu pengguna. Spam biasanya juga sering terkait dengan material yang sangat merusak seperti iklan situs pornografi atau pendistribusian virus komputer. Untuk mengatasi masalah tersebut, dibutuhkan anti-spam yang dapat mendeteksi dan membuang email spam atau memberi informasi kepada pengguna apabila suatu email memiliki potensial pesan spam.

Metode yang paling sederhana untuk mendeteksi email spam yaitu mendaftarkan frasa yang telah diketahui sebagai pesan spam lalu dilakukan pencarian dan pencocokan frasa yang identik pada suatu email seperti *"free picture"* sehingga dapat dilakukan pemblokiran email atau tidak. Namun metode ini sangat menghabiskan waktu eksekusi dan dapat dikelabui dengan memvariasikan frasa menjadi frasa-frasa yang masih dapat dibaca oleh manusia seperti f*r*e*e. Selain itu, variasi frasa yang sangat beragam dan terus bertambah mengakibatkan pemeliharaan menjadi tidak efektif.

Penerapan teknik teks kategorisasi pada *content-based* deteksi spam telah banyak dilakukan. Data koleksi yang telah diketahui sebagai email *legitimate* dan email spam dilakukan proses pembelajaran*(training)* untuk mendapatkan model. Model diterapkan untuk mengklasifikasi email baru, apakah termasuk dalam kelas spam atau tidak. Metode ini termasuk dalam deteksi spam menggunakan representasi sekumpulan kata(*bag of words*) yang terkandung dalam pesan. Namun isi pesan spam seringkali terdapat kata-kata yang salah secara tatabahasa dan penggunaan variasi tanda baca yang aneh. Hal ini mengakibatkan metode ini tidak tangguh pada kondisi tersebut. Selain itu, metode ini juga perlu dilakukan proses tokenisasi dan lemmatisasi terlebih dahulu sehingga sangat bergantung pada bahasa yang tertentu.

Memperbaiki kekurangan metode teks kategorisasi *content-based*, dibutuhkan representasi yang berbeda yaitu sekumpulan karakter *n*-grams*(bag of character n-grams)*. Sebagai contoh kata '*informatika*' jika menggunakan karakter 4-Grams menjadi '*info*', '*nfor*','*form*','*orma*','*rmat*','*mati*','*atik*','*tika*'. Ciri utama representasi sekumpulan karakater *n*-grams adalah tahan terhadap error secara tata bahasa atau penggunaan singkatan dan tanda bahasa asing, tidak bergantung pada bahasa tertentu dan tidak perlu melakukan praprosesing teks. Kombinasi kata yang dihasilkan sangat banyak sehingga digunakan algoritma Support Vector Machine(SVM) sebagai algoritma klasifikasi pembelajaran*(training)* yang mampu mengatasi ruang data dimensi yang tinggi[1].

Pada tugas akhir ini, membahas tentang deteksi email spam menggunakan karakter *n*-grams dan kombinasi algoritma klasifikasi SVM sebagai alat bantu untuk mengatasi ruang data dimensi yang ditinggi.

1.2 Perumusan masalah

Berdasarkan latar belakang yang telah dijelaskan sebelumnya, maka dirumuskan permasalahan sebagai berikut:

1. Bagaimana menerapkan representasi karakter *n*-grams untuk mendeteksi email spam.
2. Bagaimana keakuratan deteksi email spam menggunakan karakter *n*-grams dan kombinasi algoritma klasifikasi SVM.

Batasan Masalah

1. Sistem deteksi email spam yang dibangun adalah aplikasi yang berdiri sendiri *(stand alone application*) dan tidak diimplementasikan dalam email server.
2. Informasi email yang di proses hanya isi email, tanpa memperhatikan *attachments*, alamat email pengirim maupun penerima.
3. Data set yang digunakan adalah dataset Ling-Spam, terdiri dari 481 email spams dan 2.412 email asli.
4. Panjang *n* pada n-grams yang digunakan adalah 3,4,5.

1.3 Tujuan

Tujuan yang akan dicapai dari tugas akhir ini adalah:

1. Merancang dan mengimplementasikan representasi karakter *n*-grams untuk deteksi email spam.
2. Melakukan analisis apakah representasi karakter *n*-grams menunjukkan hasil yang akurat untuk dapat diterapkan mendeteksi email spam yang ditinjau dari ukuran presisi, *recall* dan *Total Cost Ratio* (TCR).

1.4 Metodologi penyelesaian masalah

Metodologi yang akan digunakan untuk menyelesaikan tugas akhir ini adalah:

1. **Studi Literatur**
   Pada tahap ini dilakukan pencarian sumber referensi yang berhubungan dengan karakter *n*-grams dan algoritma klasifikasi SVM.
2. **Analisis dan Perancangan Sistem**
   Pada tahap ini dilakukan perancangan sistem dari studi literatur dan data penunjang, serta analisis terhadap rancangan yang dikembangkan.
3. **Implementasi Sistem**
   Pada tahap ini dilakukan implementasi sistem dari rancangan yang dikembangkan. Sistem direalisasikan menggunakan program aplikasi berbasis Java dan *tools* Weka.
4. **Evaluasi unjuk kerja sistem**
   Pada tahap ini dilakukan evaluasi dari implementasi sistem yang dikembangkan. Menganalisis spam presisi,spam *recall* dan *Total Cost Ratio*(TCR)
5. **Pengambilan kesimpulan dan penyusanan laporan tugas akhir**
   Pada tahap ini dilakukan pengambilan kesimpulan dari hasil analisis yang telah dilakukan pada tahap sebelumnya untuk kemudian disusun laporan terhadap analisis yang telah dilakukan.