1. Pendahuluan

1.1 Latar belakang

Perkembangan aplikasi internet, salah satunya adalah email, sangat pesat dikarenakan sifatnya yang sangat cepat, tepat, dan murah sehingga banyak pemakai *e-mail* (selanjutnya disebut pemakai) terutama *salesperson* memanfaatkannya untuk mengirimkan pesan-pesan tersebut ke banyak orang. Pesan tersebut dinamakan "*unsolicited bulk email"*, "*junk mail"*, atau "*spam"*. *Spam e-mail* adalah *e-mail* yang tidak diharapkan bagi penerima *e-mail* dikarenakan *e-mail* jenis ini dapat membanjiri *inbox* pemakai dengan *e-mail* yang dianggap tidak berguna. Dengan semakin banyaknya pihak-pihak yang mengirimkan informasi komersial mereka lewat *email*, jumlah *email spam* semakin lama semakin bertambah banyak jumlah dan tipenya. Dan hasilnya, banyak pemilik *email* yang harus menghabiskan waktunya untuk memilih secara manual dan bahkan harus membuka email yang tidak berguna bahkan ada yang sampai melabeli *e-mail – e-mail* tersebut. Sedang di sisi *server email*, kerugian utamanya adalah akan memenuhi media penyimpanan *email* pada *email server* tersebut.

Berdasarkan survei di Amerika pada tahun 2002 terdapat 2200 jenis *spam e-mail* dan bertambah 2% pada tiap bulannya dan diperkirakan akan terdapat 3600 jenis *spam e-mail* pada tahun 2007. Sedangkan penelitian pada *European Comumunity*, memperkirakan biaya yang harus dikeluarkan dalam menerima spam pada rata-rata pengguna internet mencapai 30 euro/tahun. Tapi jumlah biaya untuk spam terus meningkat melebihi dari biaya keseluruhan dari seluruh penerima spam. Dari aspek bisnis tentu saja hal ini sangat merugikan bagi pihak perusahaan yang menjadikan *e-mail* sebagai sarana untuk menjalankan bisnisnya.

Spam filtering merupakan solusi untuk dapat membantu mengenali adanya *spam e-mail*. Dengan menggunakan metode *text classification* dapat dikenali adanya *spam* atau tidak, sesuai dengan *class* yang telah didefinisikan sebelumnya. *Text classification* menjadikan sebuah sistem mampu melakukan pembelajaran terhadap semua *e-mail*, baik *e-mail* yang merupakan *spam* atau tidak, namun demikian untuk menjalankan *text classification* diperlukan juga metode - metode *preprocessing* data.

Dalam kenyataannya, tidak semua orang mempunyai pandangan yang sama tentang *email spam*. Ada yang menganggap bahwa suatu *email* tertentu adalah sebagai *email spam* tetapi di pihak lain menganggap bahwa *email* tersebut adalah *email non-spam*. Ini tergantung dari karakteristik orang itu sendiri, pembuat perangkat lunak *spam filtering* tidak dapat memaksakan hal tersebut kepada mereka. Sehingga diperlukan adanya *personalifikasi spam filtering* yang dapat mengatasi masalah perbedaan karakteristik orang yang menggunakan *email* tersebut hanya dengan *data training* email yang tersedia secara umum baik itu yang berupa *email spam* dan *email non-spam*.

Jaringan Syaraf Tiruan (JST) menawarkan suatu teknik klasifikasi yang menirukan cara pembelajaran kerja otak. Pada Tugas Akhir ini akan dikembangkan suatu implementasi JST sebagai teknik *text classification* untuk *spam filtering* dengan menggunakan teknik *Information Gain (IG)* sebagai *feature selection* pada *data preprocessing*-nya dan menggunakan algoritma genetika (AG) dalam melatih tiap – tiap bobot yang ada pada arsitektur JST

1.2 Perumusan masalah

Berdasarkan latar belakang yang dikemukakan di atas, maka masalah pokok yang akan diteliti adalah :

a. Bagaimana menerapkan *IG* sebagai *feature selection* pada kata dalam *inbox e-mail* (data *training*).
b. Bagaimana merancang arsitektur JST sebagai *text classification* dalam kasus ini.
c. Bagaimana menentukan susunan Algoritma Genetika yang tepat untuk menetukan bobot – bobot yang terhubung dalam arsitektur JST.

Dalam penyusunan tugas akhir ini permasalahan dibatasi dalam beberapa hal yaitu:

a. Perangkat lunak yang akan dihasilkan untuk menangani studi kasus ECML PKDD 2006 Discovery Challenge Data Mining Competition.
b. Data email yang digunakan adalah data email dalam bentuk file dan sudah diubah menjadi bentuk yang berupa id dari setiap kata sehingga tidak diketahui bentuk dan isi *email* yang sebenarnya baik itu urutan kata maupun kata bahkan gambar yang digunakan dalam *e-mail* yang sebenarnya.
c. *Feature selection* dilakukan pada setiap id kata dengan menggunakan *Information Gain*.
d. Perangkat lunak ini menggunakan JST *supervised* sebagai metoda klasifikasi.
e. Pembuatan program tidak diimplementasikan pada salah satu software *e-mail client* seperti *outlook* ataupun *e-mail* server melainkan suatu perangkat lunak yang berdiri sendiri.
f. Metoda yang digunakan untuk melatih bobot – bobot pada JST adalah metoda Algoritma Genetika.

1.3 Tujuan

Berdasarkan rumusan masalah di atas, maka tujuan dari tugas akhir ini adalah mengimplementasikan JST dengan menggunakan algoritma genetika untuk melatih bobot – bobot yang terhubung pada JST dan menerapkan *information gain* pada *feature selection* sebagai teknik untuk *spam filtering* serta menghitung tingkat akurasinya.

1.4 Metodologi penyelesaian masalah

Pendekatan sistematis/metodologi yang akan digunakan dalam merealisasikan tujuan dan pemecahan masalah di atas adalah dengan menggunakan langkah-langkah berikut:

1. Studi Literatur :

   a. Pencarian referensi, mencari referensi dan sumber-sumber lain yang layak yang berhubungan dengan *data mining*, JST, AG, dan pengaruh *Information Gain* sebagai *feature selection*.
   b. Pendalaman materi, mempelajari dan memahami materi yang berhubungan dengan tugas akhir.
2. Pengumpulan data email *spam* maupun *non-spam*.
3. Mempelajari konsep dari JST dan AG serta IG yang akan digunakan dalam implementasi perangkat lunak.
4. Melakukan analisis terhadap parameter - parameter JST dan AG dalam perancangan perangkat lunak
5. Melakukan implementasi perancangan perangkat lunak
6. Melakukan pengujian perangkat lunak dengan memasukkan data yang akan dievaluasi serta mencatat hasil keluaran program.
7. Pengambilan kesimpulan dan penyusunan laporan tugas akhir.