BAB I

PENDAHULUAN

1.1 Latar Belakang Masalah

Pada sekarang ini ketersediaan informasi berbentuk dokumen teks sebagian besar sudah berbentuk elektronik *(softcopy)*. Kemungkinan penyimpanan media teks ke dalam bentuk elektronik tersebut akan mengalami perkembangan yang sangat besar pada masa yang mendatang. Salah satu hal yang perlu dilakukan adalah penggolongan dokumen-dokumen yang berada dalam satu kumpulan dokumen (*corpus)* ke dalam kelompok-kelompok kategori yang sasuai dengan isi dokumen-dokumen yang berada didalam *corpus***.**

Proses penggolongan dokumen yang berasal dari suatu *corpus* ke dalam kategori-kategori yang telah ditentukan tersebut disebut juga dengan proses dokumen klasifikasi. Tujuan dari pengelompokkan dokumen adalah untuk mempermudah pencarian informasi sesuai dengan kategori yang dimiliki oleh setiap dokumen.

Proses pengklasifikasian dokumen sulit dilakukan jika menggunakan *query* biasa, karena dengan menggunakan *query* yang kurang spesifik dapat mengakibatkan membanjirnya beberapa dokumen yang tidak relevan.

Feature selection adalah suatu bentuk upaya peningkatan algoritma pembelajaran yang digunakan untuk menggolongkan dokumen ke dalam kategori-kategori tertentu dengan cara menemukan suatu bentuk pola yang relevan (minimal satu buah pola). Tujuan dari dilakukannya proses feature selection di dalam proses dokumen kelasifikasi adalah untuk meningkatkan skalabilitas, efisiensi dan akurasi.

Feature adalah seluruh kata yang muncul dalam training set. Set ini biasanya sangat besar yaitu satu dimensi untuk setiap kata unik. Hal inilah yang membuat klasifikasi dokumen menjadi sulit, karena dimensi yang dimiliki oleh *feature space* sangat besar. Penyeleksian seluruh kata yang muncul di dalam training set dapat dilakukan dengan cara mereduksi dimensi pada *feature space* dengan jalan memilih kata-kata yang paling *informative* bagi dokumen yang akan diklasifikasikan. Informasi yang berkualitas merupakan salah satu ciri dari bentuk penurunan berdasarkan pola dan kecenderungan tertentu yang dapat diperoleh melalui *statistical pattern learning*. Menghadapi permasalahan demikian, maka dibutuhkan suatu metode yang efisien untuk melakukan proses *feature selection*.

Metode *feature selection* yang digunakan pada penelitian ini adalah ***Chi Squared***. Sedangkan metode klasifikasi dokumen yang digunakan adalah metode *Naïve Bayes* (NBC) yang memang telah sering digunakan untuk memecahkan permasalahan yang berhubungan dengan proses klasifikasi.

1.2 Rumusan Masalah

Permasalahan yang menjadi fokus penelitian ini adalah :

- Bagaimana *system* dapat melakukan proses pengenalan pola dokumen dari *corpus* yang umumnya tidak terstruktur menjadi data yang terstruktur ?.
- Bagaimana *system* dapat mengkategorikan *corpus* kedalam kategori-kategori yang sudah ditentukan yang bertujuan untuk mempermudah pencarian informasi sesuai dengan kategori yang dimiliki oleh setiap dokumen ?.

1.3 Batasan Masalah

Permasalahan yang akan dibahas di dalam penulisan tugas akhir ini dibatasi sebagai berikut :

1. Dokumen yang digunakan sebagai inputan adalah dokumen berbahasa Indonesia dengan ekstensi *.txt dan bersifat plain text.
2. Proses *Stemming* dan *Stopword* hanya berlaku pada kata-kata ber-Bahasa Indoensia saja.
3. Proses Stemming hanya dilakukan pada proses prefiks dan sufiks.
4. Parameter yang digunakan untuk melakukan pengujian hanya menggunakan parameter *precision*.

5. Metode yang digunakan di dalam melakukan pengujian *proses Feature Selection* adalah metode *Chi Squared.*
6. Metode yang digunakan di dalam proses dokumen klasifikasi adalah metode *Naïve Bayes Classifier*.
7. Dokumen yang digunakan diambil dari www.bolanews.com, www.kompas.com, www.detik.com dan bahan yang digunakan untuk penelitian hanya digolongkan ke dalam tiga kategori, yaitu olahraga, ekonomi dan komputer. Dengan adanya hal ini, maka akan dimungkinkan munculnya *outlier* (dokumen yang tidak masuk ke dalam kategori manapun).

1.4 Tujuan Penelitian

Tujuan dari penulisan Tugas Akhir ini adalah

1. Untuk memudahkan dalam melakukan pengarsipan dokumen dalam skala yang besar.
2. Meningkatkan skalabilitas, efisiensi dan akurasi dalam pengkategorian dokumen dari kumpulan dokumen teks yang besar *(corpus)* kedalam kategori-kategori yang telah ditentukan.
3. Untuk mempermudah pencarian informasi sesuai dengan kategori yang dimiliki oleh setiap dokumen.

1.5 Metodologi / Pendekatan

Adapun metode yang digunakan dalam melakukan penelitian ini adalah sebagai berikut :

1. **Studi Pustaka**

   Dengan memperlajari buku-buku literatur yang berkaitan dengan penelitian dengan tujuan mendapatkan sumber pemahaman yang membantu mengatasi masalah-masalah yang muncul selama penelitian.

2. **Pengumpulan data lewat Internet**

   Data yang dikumpulkan melalui internet berupa artikel, jurnal ilmiah, dan data-data lainnya yang mendukung penelitian .

3. **Perancangan**

   - Perancangan *Database* dalam penelitian dengan menggunakan SQL Server 2000.
   - Perancangan *User Interface* untuk aplikasi ini menggunakan Visual Basic 6.0

1.6 Sistematika Penulisan

Sistematika penulisan dari skripsi ini adalah sebagai berikut :

Bab I : Pendahuluan yang berisi Latar Belakang Masalah, Perumusan Masalah, Batasan Masalah, Tujuan Penelitian, Metode/Pendekatan, Sistematika Penulisan

Bab II : Landasan Teori yang berisi teori-teori yang mendasari penelitian.

Bab III : Sistem yang berisi tahap perancangan lengkap dari program yang akan dibuat.

Bab IV : Implementasi dan Analisis Sistem yang akan menjelaskan informasi tentang implementasi sistem dari perancangan sistem yang telah dibuat pada bab 3, meliputi cara kerja program, input dan output, realisasi sistem, kelebihan dan kekurangannya.

Bab V : Kesimpulan dan Saran berisikan kesimpulan akhir dan saran-saran untuk pengembangan sistem.