1. Pendahuluan

1.1 Latar belakang

Semakin pesatnya pengunaan Internet telah memacu pertumbuhan dan pertukaran informasi yang sangat pesat dibandingkan era sebelumnya. Sebagai akibatnya, jumlah informasi terus meningkat secara exponensial. Sumber informasi tidak hanya dalam bentuk dokumen atau teks tetapi di Internet jumlah informasi juga di hasilkan, dikoleksi dan dikonsumsi dalam jumlah yang tidak kalah besar di berbagai institusi dan perusahaan.

Sumber informasi di Internet yang dibawa oleh suatu website merupakan sumber informasi yang sangat beraneka ragam. Ada berbagai macam jenis web dengan berbagai macam informasi yang dibawa. Secara umum tidak seperti kebanyakan data atau teks konvensional lainnya suatu halaman web selain mengandung konten utama juga mengandung banyak blok informasi yang tidak berhubungan dengan konten utama misalnya, panel navigasi, *copyright*, *user guide*, kumpulan *link*, sinopsis suatu berita, berbagai macam iklan dan lain-lain yang secara keseluruhan adanya hal tersebut secara tidak langsung bertujuan untuk memudahkan user dalam mengakses web dan atau untuk tujuan bisnis tertentu.

Blok-blok informasi yang tidak relevan dengan konten utama pada suatu halaman web disebut sebagai *web pages noise*[5]. Adanya informasi yang tidak relevan dalam suatu sumber data *(web pages)* dapat menjadi permasalahan yang serius dalam proses data mining misalnya dalam proses klasifikasi atau klasterisasi.

Tugas akhir ini akan menggunakan teknik *feature weighting* dengan membobotkan tiap *feature* pada halaman web untuk kemudian dilakukan proses klasifikasi dengan tujuan meningkatkan performasi pada proses datamining. Dengan metoda ini suatu halaman web akan dimodelkan dengan struktur tree *Dokumen Object Model (DOM)* tree dan *Compressed Structure Tree (CST)* untuk mendapatkan struktur umum dan membandingkan blok-blok informasi dalam suatu website. Informasi yang didapatkan digunakan untuk melakukan pengukuran dan mengevaluasi tingkat kepentingan dari masing-masing node. Proses pengambilan data dari Internet digunakan *webcrawler* yang ada yaitu Teleport.

Hasil akhir dari tugas akhir ini adalah didapatkan suatu sumber data yang baik yang telah dideteksi noisenya dan diujikan dalam proses datamining yaitu proses klasifikasi. C*lassifier* yang digunakan adalah c*lassifier* yang terdapat dalam *tools* Weka. Kemudian akan dilakukan analisis perbandingan antara data yang diberi bobot dengan *feature weighting* dengan data yang diberi bobot dengan TFIDF dengan *clasifier* tersebut, serta akan menganalisis metoda pembobotan antara *feature weighting* dan TFIDF. Parameter yang digunakan untuk mengevaluasi proses deteksi noise adalah dengan menghitung nilai *F-measure*..

1.2 Perumusan masalah

Berdasarkan latar belakang yang dijelaskan diatas, maka permasalahan yang akan diangkat dalam tugas akhir ini adalah :

1. Bagaimana mendeteksi, dan mengurangi noise yang ada dalam suatu halaman web yang akan digunakan dalam proses klasifikasi.
2. Bagaimana menghitung bobot untuk setiap *feature* yang ada pada halaman website.
3. Bagaimana mengevaluasi proses *deteksi dan flagging* yang dilakukan dengan menghitung nilai akurasi dan *F-measure* dari proses klasifikasi.

Batasan masalah dalam tugas akhir ini adalah :

1. Menggunakan teknik *feature weighting* untuk melakukan pembobotan tiap feature ada dalam suatu halaman web.
2. Dataset yang diambil dalam betuk html *pages* secara *offline* yang merupakan berita berbahasa Indonesia.
3. Tiap halaman web berita sudah mempunyai layout standar dalam arti tiap halaman yang satu dengan yang lain memiliki kesesuaian bentuk umum, misalnya dibangun dengan CMS(*Content Managemen System*).
4. Proses pengambilan data menggunakan tools teleport
5. Tiap berita hanya akan dikelompokkan ke dalam satu kelompok berita.

1.3 Tujuan

Berdasarkan rumusan masalah diatas, tujuan yang ingin dicapai dari tugas akhir ini adalah :

1. Mengelola data pada halaman web agar siap dimining dengan teknik *feature weighting*.
2. Membandingkan dan menganalisis data sebelum dan setelah pendeteksian noise dengan mengevaluasi hasil klasifikasi.
3. Membandingkan dan menganalisis metoda pembobotan yang digunakan antara *feature weight dan TFIDF* dengan melihat nilai akurasi dan *f-measure* yang dihasilkan dari proses klasifikasi.

1.4 Metodologi penyelesaian masalah

Metodologi penyelesaian masalah yang digunakan pada Tugas Akhir ini adalah:

- **Studi Literatur**
  a. Pencarian referensi
     Mencari referensi yang layak dan berhubungan dengan topik tugas akhir, memahami dan memperlajari tentang *text mining, web mining, text categorization, web page cleaning,* teknik eliminasi *web pages Noise* dari berbagai jurnal, buku, Internet dan referensi lainnya yang mendukung.

- **Pengumpulan Data**
  Melakukan pengumpulan artikel berita yang berbentuk html yang akan digunakan sebagai data *collection*.

- **Perancangan perangkat lunak yang meliputi :**
  a. Implementasi perangkat lunak
     Mengimplementasikan perancangan menjadi perangkat lunak. Proses implementasi perangkat lunak dilakukan berdasarkan dari proses analisis dan perancangan yang telah dibangun.
  b. Pengujian
     Memeriksa error handling yang ada pada perangkat lunak yang dibangun misalnya kesalahan perhitungan, kesalahan dalam penginputan data, human error dan lain sebagainya.

- **Analisis hasil**
  Melakukan analisis terhadap tiap feature yang berpotensi sebagai noise, kemudian melihat hasil klasifikasi dari dokumen atau artikel web berita dengan mengevaluasi hasil keakuratan pengklasifikasian dokumen. Proses pengukuran keakuratan dari hasil klasifikasi dengan menggunakan konfusion matrik yang ada pada tools WEKA.

- **Pengambilan Kesimpulan dan Pembuatan Laporan**
  Pembuatan laporan Tugas Akhir yang mendokumentasikan tahap-tahap kegiatan dan hasil dalam Tugas Akhir ini.