1. Pendahuluan

1.1 Latar belakang

Dengan era teknologi sekarang, internet menjadi sumber informasi yang paling banyak digunakan. Internet dengan *HTTP*-nya dapat dikatakan sebagai keajaiban dunia dalam bidang teknologi informasi. Namun berlimpahnya informasi, justru membuat para pengguna internet mengalami kesulitan untuk mendapatkan halaman web yang mereka inginkan.

Salah satu solusi untuk permasalahan ini adalah *Search Engine*. *Search Engine* merupakan salah satu aplikasi yang paling banyak digunakan saat ini untuk melakukan pencarian terhadap suatu dokumen. *Search Engine* bekerja dengan mencari halaman-halaman web yang dianggap paling relevan dengan permintaan (*query*) pengguna. Selain *Search Engine*, aplikasi lainnya adalah *aggregator*. *Aggregator* merupakan sebuah aplikasi portal yang secara otomatis mengelompokkan suatu informasi berdasarkan kategori-kategorinya. Salah satu agregator yang populer adalah *news aggregator* yang mengelompokkan berita berdasarkan kategori-kategorinya.

News aggregator hanyalah salah satu contoh aplikasi *aggregator*. *Aggregator* di bidang lain masih sangat banyak, misalnya untuk artikel-artikel ilmiah bidang ilmu computer, keislaman, informasi seputar perguruan tinggi dengan event-eventnya di Indonesia dan lain-lain. Salah satu *News aggregator* yang dikenal masyarakat adalah Google News (http://news.google.com). *Google News* merupakan *aggregator* pencarian berita dengan sumber data dari berbagai sumber berita namun *Google News* tidak dapat memproses sumber berita berbahasa Indonesia. Untuk memenuhi kebutuhan akan berita dalam berita berbahasa Indonesia dan kemudahan mendapatkannya maka diperlukan suatu aplikasi yang sejenis dengan *Google News* yang dapat mengelompokkan berita yang berasal dari berita berbahasa Indonesia.

Pengelompokkan berita dapat dilakukan dengan berbagai macam cara, salah satunya melalui kategorisasi. Kategorisasi dapat dibedakan menjadi dua jenis yaitu kategorisasi *single-label* dan kategorisasi multi-label. Kategorisasi berita digolongkan kategorisasi multi-label karena suatu berita bisa memiliki lebih dari satu kategori. Contohnya adalah berita ”Roy Marten tertangkap polisi saat pesta Narkoba”. Berita tersebut dapat dikategorisasikan ke dalam berita kriminal dan *entertainment*. Penanganan kategorisasi multi-label ini dapat dilakukan melalui metode Boosting. Metode Boosting merupakan salah satu metode yang cukup handal[16]. Boosting menggunakan serangkaian *classifier* di dalam membuat modelnya dan secara bertahap merubah distribusi *training* data dengan fokus pada data yang sukar untuk diklasifikasikan sehingga penggabungan rule pada setiap iterasinya akan menghasilkan satu *hypothesis* dengan tingkat akurasi yang lebih tinggi. Metode Boosting memiliki banyak varian, salah satunya adalah BoosTexter. BoosTexter merupakan metode Boosting yang secara khusus menangani kategorisasi teks.

1.2 Perumusan masalah

Dengan mengacu latar belakang di atas, maka permasalahan yang dibahas dan diteliti adalah :

1. Bagaimana menerapkan BoosTexter untuk kategorisasi berita berbahasa Indonesia yang multi-label .
2. Bagaimana melakukan pengujian dan analisis dari implementasi BoosTexter.

Sedangkan batasan masalah dalam tugas akhir ini adalah :
1. Berita yang digunakan adalah berita berbahasa Indonesia.
2. Pengambilan data diambil dari beberapa portal berita berbahasa Indonesia.
3. Tidak melakukan kategorisasi secara online.
4. Data input untuk proses *preprocessing* berupa file .txt sedangkan untuk proses *training* dan *testing* berupa fie .arff
5. Hanya menangani kategori data multi-label dengan tiga label.
6. Pemisahan data *training* dan data *testing* dari *dataset* dilakukan manual.
7. *Text* preprocessing diimplementasikan dalam tugas akhir ini tapi tidak menjadi fokus permasalahan dalam tugas akhir ini.
8. Data *training* dan data *testing* berupa bobot dari masing-masing *term* yang diperoleh melalui *text* preprocessing.
9. Hanya mengimplementasikan BoosTexter dengan AdaBoost.MH prediksi dan kehadiran bernilai *real* sebagai *weak hypotheses*

1.3 Tujuan

Berdasarkan pada masalah yang telah didefinisikan di atas, maka tujuan tugas akhir ini adalah :
1. Menerapkan metode BoosTexter untuk kategorisasi berita berbahasa Indonesia yang multi-label.
2. Melakukan analisis performansi metode BoosTexter berdasarkan Hamming Loss, One Error, Coverage.

Hipotesis awal dari tugas akhir ini :
1. BoosTexter memiliki performansi yang baik dalam melakukan kategorisasi berita berbahasa Indonesia yang multi-label.
2. *Error* BoosTexter semakin kecil pada setiap kenaikan iterasi.

1.4 Metodologi penyelesaian masalah

Metode penyelesaian masalah yang digunakan sebagai berikut :
1. Studi literatur
   Mencari referensi dan sumber-sumber lain yang berhubungan dengan Data Mining khususnya Web Mining, multi-label, dan metode BoosTexter.
2. Pengumpulan data
   Mencari data dari *website* berita Indonesia, dan *Data Understanding.*.
3. Analisis dan perancangan perangkat lunak
   Menganalisis permasalahan yang akan diselesaikan dan menganalisis tahapan-tahapan yang digunakan untuk menyelesaikan permasalahan dengan metode *Object Oriented.*
4. Implementasi sistem
   Melakukan implementasi sistem dengan membangun perangkat lunak sesuai dengan perancangan yang telah di lakukan.

5. Pengujian Sistem dan Analisis Hasil
   Pengujian dilakukan pada metode BoosTexter terhadap parameter-parameter Hamming Loss, One Error, dan Coverage dan analisis dilakukan terhadap hasil yang diperoleh dari pengujian tersebut.
6. Pengambilan kesimpulan dan pembuatan laporan tugas akhir.