1. Pendahuluan

1.1 Latar belakang

Adanya berita elektronik berbahasa Inggris merupakan salah satu penyajian informasi pada *digital library* yang mempunyai peranan sangat penting terutama dalam meningkatkan kegunaan dari sekumpulan dokumen. Dokumen-dokumen tersebut tentunya mengandung berbagai informasi yang berharga yang dapat dimanfaatkan oleh banyak pihak. Suatu dokumen merupakan sumber informasi yang memiliki nilai yang berharga. Sebenarnya secara eksplisit, bisa saja mengetahui informasi pada suatu dokumen dengan mudah, salah satu contohnya dengan melihat judul yang ada pada bagian dokumen tersebut. Namun dokumen-dokumen yang ada mempunyai nilai informasi yang bersifat implisit dan pada umumnya tidak pernah diperhatikan secara seksama. Sedangkan informasi inilah yang sebenarnya merupakan pengetahuan yang mempunyai informasi sangat berguna yang perlu kita peroleh.

Text mining merupakan bagian dari data mining yang memfokuskan bidangnya pada pengambilan informasi berharga dari basis data yang berupa *text*. Kategorisasi adalah salah satu teknik dari *text mining* yang bertujuan untuk menentukan topik dari suatu artikel atau *text document* berdasarkan atribut kelasnya, kategorisasi bersifat *supervised*. Untuk memperoleh hasil kategorisasi yang baik maka diperlukan suatu preproses *stopwords removal* yaitu penghapusan kata-kata yang sering muncul tapi tidak memiliki kontribusi terhadap informasi suatu dokumen.

Dengan terus bertambahnya jumlah dokumen pada *digital library*, pengkategorian secara manual tentu saja akan menjadi suatu masalah baru untuk pengguna. Hal tersebut akan memakan banyak waktu dan menimbulkan kejenuhan. Oleh karena itu, akan digunakan teknik *text mining* yaitu kategorisasi untuk mengkategorikan dokumen berbahasa Inggris. Sebelum proses kategorisasi dilakukan, perlu dilakukan proses *stopwords removal* terlebih dahulu pada dokumen. Maka diperlukan daftar *stopword* yang baik yang digunakan pada proses *stopword removal* ini, agar hasil pengkategorisasian yang dihasilkan bisa lebih baik.

Pada tugas akhir ini akan dibahas tentang pembangunan daftar stopword yang akan digunakan dalam proses *stopword removal* yang akan mempengaruhi hasil kategorisasi. Salah satu pendekatan yang digunakan untuk memperoleh daftar stopword adalah pendekatan *Term Based Random Sampling*. Pendekatan *Term-Based Random Sampling* ini bisa mengukur seberapa pentingnya suatu *term* dalam suatu dokumen yaitu dengan menghitung bobot. Bobot dari suatu *term* bisa dihitung dengan menggunakan *Kullback-Leibler Divergence Measure*. Semakin kecil bobot suatu *term* maka semakin cocok *term* tersebut disebut sebagai *stopword*. Dengan menggunakan pendekatan *Term-Based Random Sampling* ini maka bisa diperoleh suatu daftar *stopword* yang lebih efektif.

1.2 Perumusan masalah

Berdasarkan latar belakang, permasalahan yang dijadikan objek penelitian dan pengembangan tugas akhir ini adalah sebagai berikut :

1. Bagaimana menghasilkan daftar stopword pada dokumen yang digunakan dengan menggunakan pendekatan *Term-Based Random Sampling* untuk kategorisasi dokumen berbahasa Inggris.
2. Bagaimana mengaplikasikan pre-proses dengan *stopwords removal* dengan daftar *stopword* yang diperoleh dengan pendekatan *Term-Based Random Sampling* dan daftar *stopword* bahasa Inggris yang sudah didefenisikan.
3. Bagaimana mengukur dan menganalisis performansi kategorisasi setelah dilakukan *stopwords removal* dengan *Term-Basd Random Sampling* untuk kategorisasi dokumen berbahasa Inggris berdasarkan F-Measure..

1.3 Tujuan

Dalam tugas akhir ini, diharapkan tercapai hal-hal berikut :

1. Menghasilkan daftar *stopword* pada dokumen yang digunakan dengan menggunakan pendekatan *Term-Based Random Sampling* untuk kategorisasi dokumen berbahasa Inggris.
2. Mengaplikasikan pre-proses dengan *stopwords removal* dengan daftar *stopword* yang diperoleh dengan pendekatan Term-Based Random Sampling dan daftar *stopword* bahasa Inggris yang sudah didefenisikan.
3. Mengukur dan menganalisis performansi kategorisasi setelah dilakukan *stopwords removal* untuk kategorisasi dokumen berbahasa Inggris berdasarkan F-Measure.

1.4 Batasan masalah

Untuk memfokuskan penulisan tugas akhir ini, penulis mengambil studi kasus kategorisasi untuk berita berbahasa Inggris. Masalah yang akan dibahas memiliki batasan-batasan sebagai berikut:

1. Metode kategorisasi yang digunakan adalah Naïve Bayes dan J48 yang ada pada *Software* WEKA.
2. Dokumen yang digunakan adalah dokumen berbahasa Inggris yang bersifat offline yaitu dokumen Reuters dengan jumlah 200 dokumen.
3. Dokumen yang akan digunakan dilakukan proses *stemming* terlebih dahulu.

1.5 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, stopword, text categorization* dari berbagai jurnal, buku, Internet dan referensi lainnya yang mendukung.

- **Pengumpulan Data**
  Melakukan pengumpulan artikel berita yang berbentuk file txt 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 hasil daftar *stopword*, kemudian melihat hasil kategorisasi dari dokumen atau artikel web berita dengan mengevaluasi hasil keakuratan pengkategorian dokumen. Proses pengukuran keakuratan dari hasil kategorisasi 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.