BAB 1 PENDAHULUAN

Pada bab ini dijelaskan latar belakang dari penelitian klasifikasi dokumen teks. Tujuan dan ruang lingkup dari tugas akhir memberikan penjelasan mengenai hasil yang ingin diketahui dan batasan-batasan yang terdapat dalam melakukan penelitian. Pada akhir dari bagian ini dijelaskan mengenai metodologi serta sistematika penulisan laporan.

1.1 Latar Belakang

Kebutuhan akan informasi semakin meningkat seiring perkembangan teknologi dalam menyebarkan informasi kepada masyarakat. Informasi yang dibutuhkan mengalami perkembangan mulai dari informasi yang bersifat umum hingga informasi yang bersifat khusus. Banyaknya informasi dan dokumen yang tersedia mendorong manusia untuk mencari cara untuk mendapatkan informasi dan dokumen yang tepat dalam waktu singkat. Apabila dokumen yang akan dicari berada pada kumpulan dokumen yang berjumlah sedikit, pencarian dapat dilakukan secara manual. Namun, apabila jumlah dokumen yang tersedia sangat besar, proses pencarian secara manual akan menghabiskan waktu dan tenaga. Apabila waktu pencarian terlalu lama, maka manfaat dari informasi yang diperoleh dapat berkurang. Hal ini dikarenakan informasi yang telah melewati suatu waktu sudah tidak berguna atau tidak valid. Oleh karena itu, diperlukan sebuah cara untuk memperoleh data secara cepat dan tepat.

Klasifikasi dokumen dapat membantu proses pencarian sebuah dokumen dengan cepat dan tepat. Klasifikasi dokumen mengelompokkan dokumen yang sesuai dengan kategori yang terkandung pada dokumen tersebut. Apabila terdapat permintaan untuk mencari sebuah dokumen yang telah diketahui memiliki kategori tertentu, maka proses pencarian hanya dilakukan pada kumpulan dokumen yang memiliki kategori tersebut, pencarian tidak dilakukan pada semua kumpulan dokumen yang dimiliki sehingga proses pencarian dapat dilakukan lebih cepat.

1.2 Permasalahan

Permasalahan yang sering muncul dalam klasifikasi dokumen teks adalah kumpulan dari dokumen yang digunakan untuk membangun sebuah *classifier* belum diketahui kategorinya, sehingga perlu dilakukan pemberian kategori secara manual terhadap kumpulan dokumen tersebut. Permasalahannya, seberapa banyak dokumen *training* yang diperlukan agar *classifier* yang dihasilkan memberikan hasil yang maksimal. Apabila jumlah dokumen yang diberi kategori secara manual terlalu sedikit, *classifier* yang terbentuk tidak akan menghasilkan tingkat akurasi yang maksimal. Namun, jika jumlah dokumen yang diberi kategori secara manual terlalu banyak akan menghabiskan waktu dan tenaga.

Selain itu, ingin dilihat juga apakah metode-metode yang telah digunakan pada (Nigam, McCallum, & Mitchell, 1999) dapat diterapkan pada dokumen dengan bahasa Indonesia. Beberapa metode tersebut adalah Naïve Bayes dan Expectation Maximization. Permasalahan lain yang dihadapi adalah keterbatasan dokumen dalam bahasa Indonesia.

1.3 Tujuan

Tujuan utama dari tugas akhir ini adalah mengetahui manfaat dari penggunaan *unlabeled documents* dalam membantu meningkatkan akurasi klasifikasi dokumen teks. Selain itu, tugas akhir ini juga bertujuan untuk mengetahui kinerja metode klasifikasi dokumen teks dengan melihat perbandingan nilai akurasi hasil klasifikasi dari aspek penggunaan *stopwords*, jumlah kategori, dan jenis fitur yang digunakan.

1.4 Ruang Lingkup

Ruang lingkup pengerjaan dari tugas akhir ini adalah sebagai berikut:

1. Klasifikasi dokumen dilakukan dengan menggunakan tiga jenis data, yaitu dokumen hukum dari hukumonline.com, artikel media massa dari kompas.com dan kumpulan *email* dari 20Newsgroups *dataset*.

2. Jenis fitur yang digunakan adalah *presence, frequency, frequency normalized*, dan pembobotan *tf-idf*.

3. Klasifikasi dokumen dilakukan dengan metode *machine learning* Naïve Bayes dan Expectation Maximization dengan *tools* yang sudah tersedia. Pekerjaan yang dilakukan mencakup persiapan data, pemilihan fitur, persiapan *input* untuk masing-masing *tools*, dan pemanfaatan *tools* atau *library* yang tersedia.

1.5 Metodologi Penelitian

Metodologi yang digunakan dalam pengerjaan tugas akhir ini adalah metode eksperimental dan dilaksanakan dalam tahapan-tahapan sebagai berikut:

1. Studi literatur – Pencarian informasi mengenai klasifikasi dokumen, metode-metode yang dapat digunakan, penelitian yang telah dilakukan, dan pembelajaran mengenai metode-metode tersebut.

2. Perancangan – Membuat perancangan percobaan dengan mempersiapkan kumpulan dokumen, penentuan variabel percobaan, dan perancangan klasifikasi dokumen menggunakan *machine learning*.

3. Implementasi – Melakukan implementasi dari perancangan yang telah disusun dengan pembuatan program untuk melakukan pengolahan dokumen dan penggunaan metode *machine learning* dalam pengklasifikasian dokumen.

4. Analisis hasil – Melakukan perbandingan nilai akurasi yang didapat dengan menggunakan metode *machine learning* dilihat dari aspek metode, jenis fitur, penggunaan *stopwords*, jumlah kategori, dan jumlah dokumen yang digunakan.

1.6 Sistematika Penulisan

Sistematika penulisan laporan mengikuti tahap-tahap yang dilakukan untuk menyelesaikan tugas akhir, sebagai berikut:

- BAB 1 PENDAHULUAN - Pada bab ini dijelaskan latar belakang dari penelitian klasifikasi dokumen teks. Tujuan dan ruang lingkup dari tugas akhir memberikan penjelasan mengenai hasil yang ingin diketahui dan batasan-batasan yang ada dalam melakukan penelitian. Pada akhir dari bab ini dijelaskan mengenai metodologi serta sistematika penulisan laporan.

- BAB 2 LANDASAN TEORI - Pada bab ini dijelaskan landasan teori dan metode yang digunakan pada tugas akhir ini dalam pengklasifikasian dokumen teks. Pembahasan dimulai dengan penjelasan mengenai klasifikasi dokumen teks. Pada subbab berikutnya dijelaskan metode-metode yang digunakan dalam melakukan klasifikasi dokumen teks.

- BAB 3 PERANCANGAN - Pada bab ini dijelaskan mengenai perancangan untuk melakukan klasifikasi dokumen teks. Klasifikasi dilakukan dengan menentukan kategori dari semua dokumen *testing* yang ada. Perancangan klasifikasi dokumen teks ini meliputi persiapan dokumen, pembuatan *term documents matrix* dan klasifikasi dokumen teks menggunakan *machine learning* yaitu dengan algoritma Naïve Bayes dan Expectation Maximization.

- BAB 4 IMPLEMENTASI - Pada bab ini dijelaskan mengenai implementasi dari perancangan klasifikasi dokumen teks. Penjelasan dimulai dari proses persiapan dokumen yang meliputi *converting* dan *filetering*, hingga modifikasi yang dilakukan pada *framework* yang digunakan untuk melakukan klasifikasi dokumen dengan algoritma Naïve Bayes dan Expectation Maximization. Sebagian besar program dibuat sesuai perancangan dengan menggunakan bahasa pemrograman Java, hanya pada saat *converting* digunakan *tools* yang sudah tersedia untuk sistem operasi Windows.

- BAB 5 HASIL DAN PEMBAHASAN - Pada bab ini dijelaskan hasil percobaan yang telah dilakukan untuk klasifikasi dokumen teks menggunakan metode Naïve Bayes dan Expectation Maximization. Pembahasan dibagi

menurut variabel yang diujikan pada percobaan tugas akhir ini. Pembahasan diawali dengan penjelasan mengenai variabel eksperimen yang digunakan, hasil eksperimen terhadap jumlah fitur dan penghilangan *stopwords*, dilanjutkan pembahasan mengenai hasil klasifikasi dokumen teks untuk masing-masing variabel percobaan yaitu hasil klasifikasi berdasarkan aspek jumlah kategori, penggunaan jenis fitur, dan pengaruh penggunaan *unlabeled documents* terhadap hasil klasifikasi dokumen teks. Pada setiap subbab dibahas klasifikasi untuk setiap jenis data yang digunakan yaitu data dokumen hukum, 20Newsgroups *dataset*, dan data artikel media massa.

- BAB 6 PENUTUP - Bab ini merupakan penutup dari laporan tugas akhir yang berisi kesimpulan dan kendala dari percobaan yang dilakukan dengan metode Naïve Bayes dan Expectation Maximization. Selain itu, pada subbab terakhir juga diberikan saran-saran untuk pengembangan lebih lanjut dalam penelitian klasifikasi dokumen teks.