BAB 6 PENUTUP

Bab ini merupakan bab terakhir yang memberikan kesimpulan dan kendala dari percobaan yang dilakukan untuk klasifikasi dokumen teks menggunakan metode Naïve Bayes dan Expectation Maximization. Selain itu, pada subbab terakhir juga diberikan saran untuk pengembangan lebih lanjut dalam penelitian klasifikasi dokumen.

6.1 Kesimpulan

Berdasarkan hasil percobaan yang diperoleh, terdapat beberapa hal yang penting. Pertama, penggunaan jenis fitur tertentu dapat mempengaruhi kinerja *machine learning* untuk klasifikasi dokumen teks. Dari percobaan yang telah dilakukan, secara keseluruhan pembobotan *tf-idf* memberikan hasil yang paling baik dibandingkan ketiga jenis fitur lain yang digunakan. Pembuangan *stopwords* dari daftar fitur juga dapat meningkatkan akurasi klasifikasi dokumen teks. Kedua, jumlah kategori yang ada mempengaruhi kinerja *machine learning* untuk klasifikasi dokumen teks. Secara umum, penambahan jumlah kategori menurunkan tingkat akurasi klasifikasi.

Secara umum, penggunaan *unlabeled documents* dapat membantu proses *learning* dari metode klasifikasi yang digunakan sehingga dapat meningkatkan akurasi klasifikasi. Rata-rata peningkatan akurasi hasil klasifikasi dokumen teks bahkan dapat mencapai sekitar 9,5% pada percobaan dengan konfigurasi tertentu (lihat subbab 5.5.3). Namun yang perlu diperhatikan adalah jumlah penggunaan *labeled documents* untuk membangun *intial classifier*. Dari percobaan yang telah dilakukan, diperlukan sekitar 30 hingga 60 *labeled documents* tiap kategorinya untuk membangun *initial classifier* yang dapat memanfaatkan *unlabeled documents* secara maksimal. Jumlah tersebut masih dipengaruhi jumlah kategori yang digunakan, semakin banyak jumlah kategori yang digunakan, semakin banyak pula *labeled documents* tiap kategori yang diperlukan.

6.2 Kendala

Beberapa kendala yang dihadapi dalam melakukan percobaan klasifikasi dokumen teks dalam tugas akhir ini, antara lain:

- Jumlah dokumen yang terlalu sedikit, khususnya untuk dokumen-dokumen dalam bahasa Indonesia. Hal ini menyebabkan variasi penggunaan jumlah *labeled documents* dan *unlabeled documents* tidak terlalu besar.

- Keterbatasan sumber daya untuk melakukan serangkaian proses klasifikasi dokumen teks dengan jumlah dokumen yang besar. Beberapa hal telah dilakukan untuk mengatasi masalah ini seperti meminimalkan representasi *term documents matrix* serta meminimalkan implementasi penggunaan *dataset* dalam tools yang digunakan, namun hal tersebut hanya mampu mengakomodasi penggunaan 16000 dokumen teks.

- Eksekusi program untuk melakukan klasifikasi dokumen teks memerlukan waktu yang cukup lama, khususnya pada metode Expectation Maximization. Hal ini mengingat bahwa metode Expectation Maximization adalah algoritma yang iteratif dan mencapai konvergensi setelah melalui lebih kurang delapan iterasi (Bing Liu, 2004).

6.3 Saran

Klasifikasi dokumen teks yang dilakukan pada tugas akhir ini masih memiliki kekurangan. Beberapa saran yang mungkin berguna untuk melakukan penelitian klasifikasi dokumen selanjutnya, antara lain:

- Mengumpulkan lebih banyak dokumen, khususnya untuk dokumen-dokumen berbahasa Indonesia sehingga variasi penggunaan jumlah *labeled documents* dan *unlabeled documents* dapat ditingkatkan untuk melihat hasil yang lebih mendetail.

- Meningkatkan atau memperbaiki implementasi dari *tools* yang digunakan untuk melakukan klasifikasi dokumen teks sehingga percobaan dapat dilakukan dengan jumlah dokumen yang besar. Perubahan yang masih dapat dilakukan adalah mengubah struktur data yang digunakan pada *tools* yang digunakan.

- Menambah sumber daya untuk mengakomodasi penggunaan jumlah dokumen yang besar dan mempercepat eksekusi program.

- Melakukan penelitian lebih lanjut mengenai perbedaan hasil *expectation step* setiap iterasi pada metode Expectation Maximization sehingga dapat diketahui jumlah iterasi optimal yang diperlukan untuk memberi kategori perkiraan *unlabeled documents.*