

BAB I

LATAR BELAKANG

1.1 Latar Belakang Masalah

Perkembangan teknologi informasi yang melibatkan banyak dokumen semakin meningkat. Penyimpanan dokumen yang berisi tentang segala sumber informasi tersebut tersebar di beberapa lokasi. Penyebaran informasi tersebut banyak dilakukan dengan media berupa halaman web. Menurut penelitian dari Hearst (2003), ukuran data di media internet tahun 2002 mencapai 532897 Terabytes dengan sekitar 41,7%-nya adalah teks. Berdasarkan *Netcraft Web Server Survey*, jumlah halaman yang aktif pada Mei 2008 adalah 168 milyar situs web. Volume yang besar membuat masyarakat semakin sulit mencari informasi yang diinginkan khususnya mahasiswa yang ingin mencari referensi Tugas Akhir. Banyak mahasiswa yang kebingungan mencari referensi Tugas Akhir yang sesuai dengan bidang minat. Untuk itu diperlukan teknik pengolahan teks yang mengordinasikan dokumen sesuai dengan katagorinya, sehingga informasi yang tersedia dapat dikordinasikan dengan baik dan mudah diakses oleh pengguna. Salah satu metode yang dapat digunakan adalah klasifikasi dokumen. Klasifikasi dokumen adalah proses penggolongan suatu dokumen ke dalam suatu katagori tertentu.

Klasifikasi termasuk teknik pembelajaran mesin atau bisa disebut *supervised learning*. Menurut Manning et al (2008), *supervised learning* adalah proses pembelajaran mengenai ciri dari tiap-tiap katagori yang ada. Teknik ini membangun sebuah *classifier* yang mempelajari tipa katagori berdasarkan dokumen latih yang dimiliki. Beberapa metode klasifikasi yang bisa digunakan dalam proses pembelajaran yaitu *multinomial naïve bayes*, *multivariate Bernoulli model*, *Racchio classification*, *k-Nearest Neighbor* dan *support vector machine* (SVM)

Peningkatan dokumen akan mempengaruhi kinerja klasifikasi yang menyebabkan kinerja sistem *classifier* semakin berat. Hal tersebut dikarenakan sistem klasifikasi mengambil isi dari uraian setiap dokumen. Salah satu cara untuk meningkatkan kinerja dari sistem klasifikasi adalah dengan menerapkan teknik pemilihan fitur dokumen. Pemilihan fitur adalah suatu metode yang bertujuan untuk mengurangi jumlah kata yang digunakan untuk menjadi penciri dan meningkatkan akurasi hasil klasifikasi. Ada beberapa teknik yang digunakan untuk pemilihan fitur dokumen antara lain *document frequency thresholding*, *information gain*, *mutual information*, *term strenght* dan *chi-square testing*. Penelitian klasifikasi teks menggunakan pemilihan fitur ciri yang telah dilakukan sebelumnya antara lain Tierawan (2011) menggunakan metode *naïve bayes* dengan ekstraksi ciri dari *chi-square* dan Saputra (2012) menggunakan metode *semantic smooting* dengan menggunakan ekstraksi ciri dari *chi-square*. Akurasi yang diperoleh menggunakan *naïve bayes* adalah 93,26% dan *semantic smoothing* adalah 95,55%. Hal ini membuktikan bahwa kedua penelitian tersebut dapat digunakan untuk melakukan klasifikasi dokumen teks.

Penelitian ini menggunakan metode *support vector mechine* dengan pemilihan fitur *chi-square* yang diharapkan memiliki tingkat akurasi yang lebih tinggi sehingga memudahkan mahasiswa mencari referensi Tugas Akhir dengan lebih mudah.

1.2 Rumusan Masalah

Dalam pembuatan sebuah sistem tentu tidak akan terlepas dari beberapa permasalahan. Dari latar belakang permasalahan diatas maka, dapat disimpulkan permasalahan yang ada yaitu sebagai berikut:

1. Apakah *support vector machine* dengan pemilihan *fitur chi-square* mampu mengklasifikasikan teks?
2. Seberapa besar akurasi yang dihasilkan *support vector machine* dalam mengklasifikasikan teks dengan menggunakan fitur *chi-square*?

1.3 Batasan Masalah

Batasan masalah penelitian ini meliputi :

1. Dokumen yang digunakan adalah adalah dokumen abstrak Tugas Akhir Univeritas Muhamadiyah Jember dalam format txt.
2. Klasifikasi dokumen diambil dari dokemen Tugas Akhir dan dibagi menjadi menjadi tiga kelas yaitu rekayasa perangkat lunak, bisnis cerdas dan jaringan.
3. Penelitian difokuskan kepada klasifikasi dokumen menggunakan metode klasifikasi *support vector machine* dengan pemilihan fitur *chi-square*.
4. Jumlah file yang diklasifikasikan berjumlah 200 file dokumen. Data yang diambil adalah data tahun 2014.

1.4 Tujuan Penelitian

Tujuan penelitian ini adalah menerapkan dan mengevaluasi metode *support vector mechine* menggunakan pemilihan fitur *chi-square* yang dapat meningkatkan kinerja fungsi klasifikasi dokumen teks serta mengukur akurasi algoritma *support vector mechine* dengan melihat pengaruh pemilihan fitur *chi-square* dalam proses komputasi.

1.5 Manfaat Penelitian

Penelitian ini diharapkan dapat menambah metode klasifikasi dokumen dan membantu dalam mengorganisasikan dokemen secara cepat, efisien dan memiliki kinerja yang baik. Sehingga memudahkan mahasiswa melakukan pencarian referensi tugas akhir.