KLASIFIKASI DOKUMEN TEKS MENGGUNAKAN METODE SUPPORT VECTOR MACHINE DENGAN PEMILIHAN FITUR CIRI CHI-SQUARE

[1]Muhammad Isa Hidayatullah (1210651073), [2]Lutfi Ali Muharom, S.Si., M.Si
Jurusan Teknik Informatika Fakultas Teknik Universitas Muhammadiyah Jember
Email : muhisahidayat@gmail.com

ABSTRAK

Peningkatan jumlah dokumen membuat mahasiswa semakin sulit memperoleh informasi sesuai dengan apa yang diinginkan khusunya referensi tugas akhir. Masalah ini memerlukan teknik pengolahan teks yang mengorganisasikan dokumen sesuai dengan katagorinya. Salah satunya klasifikasi teks. Klasifikasi teks dapat mengordinasikan dokumen sesuai dengan yang telah ditentukan sebelumnya secara otomatis. Salah satu klasifikasi vector teks yang terkenal adalah *support vector machine* (SVM) yang berusaha mencari bidang pemisah terbaik pada *input space*. Menurut Chenometh et al (209) *support vector machine* (SVM) merupakan algoritma klasifikasi terbaik dibanding dengan metode klasifikasi metode vektor yang lain yaitu Rocchio, k-Nearest Neighbor dan decision tree. Melalui serangkaian pengujian terhadap sejumlah dokumen yang telah diolah menjadi suatu representasi data berupa matriks vektor, SVM berhasil mengklasifikasikan dokumen dengan akurasi tertinggi sebesar 82,14% dalam waktu 23 detik.

Kata kunci: penjadwalan, algoritma genetika, optimasi.

ABSTRACT

The increase in the number of documents to make students more difficult to obtain information specifically thesis. This problem reuires a text processing techniques to organize documents according to category. Text classification can be coordinated the documents according to predefined categories automatically. One well known text classification is a support vector machine (SVM) is trying to find the best interface on the input space. According Chenomoth et al (2009) support vector machine (SVM) is the best classification alghorithm compared with other vector classification that Rocchio, K-Nearest Neightbor and Decision Tree. Through a series of test on a number of documents that have been processed into a data representation in the form of a matrix vector, SVM managed to classify document with the highest accuracy of 82,14& with 23 seconds.

Keyword : text document classification, text mining, support vector mechine, chi-square

1. PENDAHULUAN

1.1 Latar Belakang

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 *multinominal naïve bayes, multivariative 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 tresholding, 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 penilitian 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 peragkat 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, efisian dan memiliki kinerja yang baik. Sehingga memudahkan mahasiswa melakukan pencarian referensi tugas akhir.

2. TINJAUAN PUSTAKA

2.1 Klasifikasi Dokumen

Klasifikasi dibedakan menjadi dua jenis yaitu klasifikasi berbasis peluang dan klasifikasi ruang vektor. Manning et al. (2008) menyatakan ada beberapa algoritma yang dapat dilakukan untuk melakukan klasifikasi dokumen berbasis vektor yaitu *Rocchio, K-Nearest Neighbour* (KNN), *decision tree* (DT) dan *support vector mechine* (SVM).

Chenometh et al. (2009) merangkum perbandingan antara empat klasifikasi berbasis ruang vektor yang sering digunakan dalam kategori teks yaitu *Rocchio*, KNN, DT, dan SVM. Chenometh et al. (2009) menyatakan bahwa SVM merupakan yang merupakan algoritma klasifikasi terbaik dibandingkan dengan lainnya, meskipun sangat mudah *error* dalam

data *training*. Sedangkan Kaiser et al. (2005) menyatakan teknik-teknik tersebut berbeda dalam mekanisme pembelajaran dan representasi model yang dipelajari. KNN dan SVM merupakan algoritma yang memberikan hasil klasifikasi terbaik dengan presisi di atas 85%..

2.2 Preprosesing Dokumen

Preprosesing dokumen merupakan tahapan dari *text mining* yang harus dilakukan jika ingin menambang informasi berupa teks. *Text Mining* merupakan suatu proses yang bertujuan untuk menemukan informasi atau tren terbaru yang sebelumnya tidak terungkap, dengan memproses dan menganalisa data dalam jumlah besar. Dalam menganalisa sebagian atau keseluruhan *unstructured text*, *text mining* mencoba untuk mengasosiasikan satu bagian teks dengan yang lainnya berdasarkan aturan-aturan tertentu. Selain itu *text mining* juga diartikansebagai kegiatan menambang data yang berupa teks atau dokumen, dengan tujuan mencari kata-kata yang dapat mewakili apa yang ada dalam dokumen sehingga dapat dilakukan analisa keterhubungan antar dokumen. Tahap-tahap preprosesing yang dilakukan dalam penilitian ini adalah:

2.2.1 Tokenisasi

Tokenisasi adalah pemisahan kata dari dokumen dengan menggunakan karakter spasi sebagai pemisahnya (Wibowo 2010). Proses ini diawali dari mengambil isi dokumen dengan tabel *corpus*, selanjutnya dilakukan proses pembacaan seluruh karakter yang terdapat pada dokumen, baik karakter huruf, angka, tanda baca dan karakter yang tidak terlihat.

2.2.2 Stopword

Stopwords adalah kata umum (*common words*) yang biasanya muncul dalam jumlah besar dan dianggap tidak memiliki makna. Stop words umumnya dimanfaatkan dalam *task information retrieval*. Contoh stop words untuk bahasa Inggris diantaranya "*of*", "*the*". Sedangkan untuk bahasa Indonesia diantaranya "yang", "di", "ke". Kata-kata yang termasuk dalam *stopwords* pada umumnya merupakan kata-kata yang sering muncul di setiap dokumen sehingga kata tersebut tidak dapat digunakan sebagai penciri suatu dokumen (Herawan 2011).

2.2.3 Pembobotan

Proses pembobotan dari suatu kata yang terpilih dengan menggabungkan aspek lokal dan global pada setiap *term*, yaitu menghitung *term frequency* (tf) dari setiap dokumen yang ada di koleksi dokumen dikalikan dengan bobot global *inverse document frequency* (idf)

2.2.4 Pemilihan Fitur Ciri

Dalam penilitian ini, fitur ciri yang digunakan adalah *chi-square*. *Chi-square* merupakan pengujian hipotesis mengenai perbandingan antara frekuensi contoh yang benar-benar terjadi dengan frekuensi harapan yang didasarkan atas hipotesis tertentu pada setiap kasus atau data (Herawan 2011). Perhitungan *nilai chi-square* yang digunakan untuk melakukan pengujian perbedaan antara pola frekuensi observasi (o_i) dengan frekuensi harapan (e_i)

2.3 SVM (*support vector machine*)

Menurut Santoso (2007) *support vector machine* (SVM) adalah suatu teknik untuk melakukan prediksi, baik dalam kasus klasifikasi maupun regresi. SVM berada dalam satu kelas dengan Artificial Neural Network (ANN) dalam hal fungsi dan kondisi permasalahan yang bisa diselesaikan. Keduanya masuk dalam kelas *supervised learning*. *Supervised learning* adalah teknik pembelajaran mesin dengan membuat suatu fungsi dari data latihan. Data latihan terdiri dari pasangan nilai *input* dan *output* yang diharapkan dari *input* yang bersangkutan. Tugas dari *supervised learning* adalah untuk memprediksi nilai fungsi untuk nilai semua input yang ada.

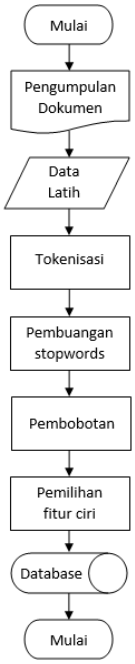
3. METODOLOGI PENELITIAN

Merupakan bentuk kegiatan identifikasi terhadap perancangan yang dilakukan. Pada tahap ini penulis mengelompokkan ke dalam 3 metode yaitu :

1. Pengumpulan data dengan cara mengumpulkan literatur, jurnal, paper, dan bacaan-bacaan yang akan dibahas dengan bersumber buku-buku yang ada kaitannya dengan judul penelitian untuk membantu menyelesaikan pembangunan dalam sistem ini.
2. Pengumpulan data dokumen tugas akhir dengan format .txt yang akan digunakan sebagai data uji dan data latih.
3. Perancangan sistem klasifikasi dengan metode *support vector mechine* dengan pemilihan fitur ciri *chi-square*.
4. Implementasi klasifikasi dengan metode *support vector mechine* dan pemilihan fitur ciri *chi-square.*

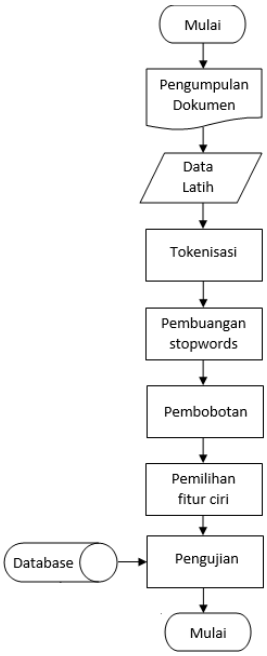
3.2 Arsitektur Sistem

Diagram alur dibagi menjadi 2 yaitu diagram alur data latih dan digram alur data uji. Alur data latih digambarkan dalam gambar 3.1. Tahap alur digram data latih yaitu pengumpulan dokumen, pembagian data, tokenisasi, pembuangan *stopwords*, pemilihan fitur ciri, pembobotan dan setelah itu data disimpan ke *database*.



Gambar 3.1. Flowchart Data Latih

Alur data uji digambarkan dalam gambar 3.2.



Gambar 3.2 Flowchart Data Uji

Tahap alur digram data uji yaitu pengumpulan dokumen, pembagian data, tokenisasi, pembuangan *stopwords*, pemilihan fitur ciri, pembobotan, pengambilan data dari *database* dan pengujian hasil klasifikasi.

Pada penelitian ini, data yang diproses merupakan koleksi dokumen yang dibagi menjadi dua kategori yaitu data latih dan data uji. Kedua kategori data tersebut akan digunakan pada tahapan praproses yang terdiri atas tokenisasi, *stopword*, pemilihan fitur ciri, dan pembobotan.

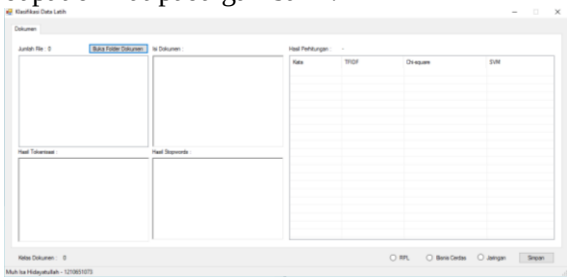
Tahap selanjutnya adalah pengujian dengan menggunakan metode klasifikasi SVM pada data latih dan hasilnya digunakan sebagai dasar pembuatan model SVM. Setelah itu dilakukan pengujian model klasifikasi terhadap dokumen uji yang sudah diketahui kelasnya dan dilakukan proses perhitungan hasil klasifikasi.

4. HASIL dan PEMBAHASAN

Pada tahap ini akan dijelaskan tentang proses implementasi metode *support vector machine* dengan pemilihan fitur ciri *chi-square*, sesuai perancangan sistem yang telah dibahas pada bab 3 serta melakukan pengujian sistem yang telah dibangun.

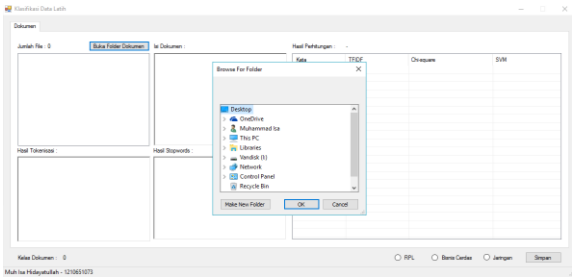
4.3.1. Data Latih

Halaman utama adalah halaman yang akan muncul ketika anda membuka aplikasi pertama kali, pada halaman utama ini terdapat informasi berupa hasil tokenisasi, hasil, stopwords hasil perhitungan dan tombol simpan yang berguna menyimpan data ke database sehingga bisa diproses nantinya. Halaman utama dapat dilihat pada gambar 4.1



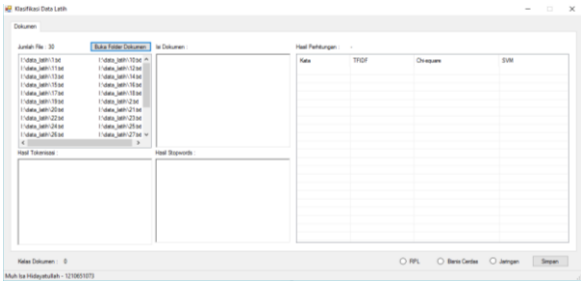
Gambar 4.1 Halaman Utama Data Latih

Halaman buka folder dokumen akan tampil jika user memilih tombol buka folder dokumen pada halaman utama. Halaman ini berfungsi meload folder yang berisikan file dokumen yang akan dihitung. Halaman ini dapat dilihat pada gambar 4.2



Gambar 4.2 Proses Buka Folder Dokumen

Setelah itu jika user telah membuka folder yang berisi dokumen, aplikasi akan kembali ke halaman utama. User bisa memilih folder mana yang akan dilatih seperti yang terlihat pada gambar 4.3.

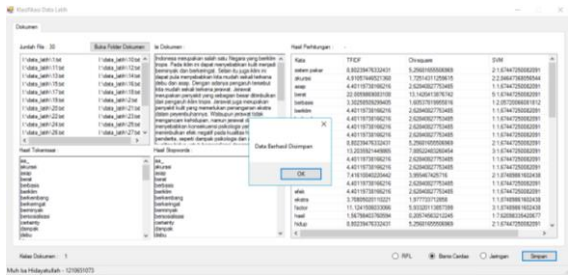


Gambar 4.3 Hasil Buka Folder Dokumen

Setelah user memilih file yang akan diuji, aplikasi akan melakukan perhitungan secara otomatis, meliputi hasil tokenisasi, hasil stopwords dan hasil perhitungan seperti yang terlihat pada gambar 4.4.

Gambar 4.4 Proses Perhitungan Data Latih

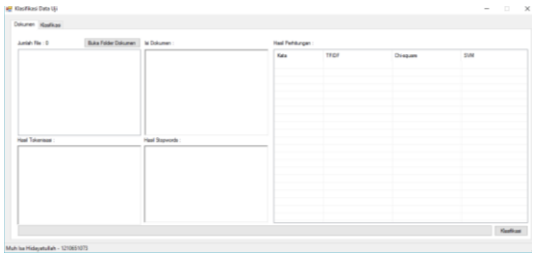
User dapat memilih sendiri dokumen tersebut masuk ntk katagori mana. Setelah itu dokumen tersebut akan disimpan di database yang bertujuan sebagai acuan untuk dibandngkan dengan data uji.



Gambar 4.5 Proses Simpan Data

4.3.2. Data Uji

Halaman ini mirip dengan halaman data latih tapi memiliki fungsi yang berbeda. Halaman data uji berfungsi untuk mengkalsifikasikan dokumen yang tidak diketahui katagorinya.

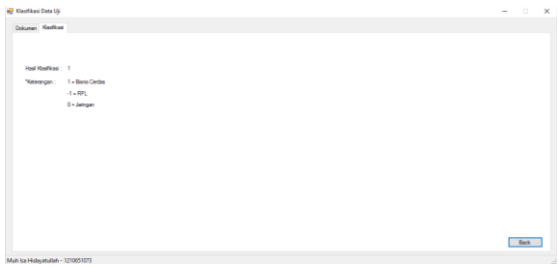


Gambar 4.6 Halaman Utama Data Uji

Setelah membuka folder, aplikasi secara otomatis melakukan perhitungan yang meliputi hasil tokenisasi, hasil stopwords dan hasil perhitungan.

Gambar 4.7 Proses Perhitungan Data Uji

User bisa memilih file yang akan diklasifikasikan dengan memilih tombol klasifikasi. Klasifikasi dilakukan secara otomatis oleh aplkasi sehingga diketahui katagorinya.



Gambar 4.8 Hasil Klasifikasi

BAB V

KESIMPULAN DAN SARAN

5.1 Kesimpulan

Berdasarkan hasil implementasi dan pengujian dalam penerapan metode *support vector mechine* dengan pemilihan fitur ciri *chi-square* untuk klasifikasi dokumen, maka dapat diambil beberapa kesimpulan sebagai berikut :

1. Metode *support vector mechine* dengan pemilhan fitur ciri *chi-square* dapat diterapkan untuk menyelesaikan kasus klasifikasi dokumen.
2. Rata-rata waktu yang dibutuhkan untuk mengklasifikasikan dokumen menggunakan metode *support vector mechine* dengan pemilhan fitur ciri *chi-square* adalah 19 detik.
3. Skenario 1 dengan data latih sebanyak 30 memiliki akurasi sebesar 70,58%, skenario 2 dengan data latih sebanyak 39 memiliki akurasi sebesar 71,42%, skenario 3 dengan data latih sebanyak 45 memiliki akurasi sebesar 74,19%, skenario 4 dengan data latih sebanyak 51 memiliki akurasi sebesar 77,18%, dan Skenario 5 dengan data latih sebanyak 60 memiliki akurasi sebesar 82,14%. Sehingga dapat disimpulkan bahwa semakin banyak data latih maka semakin tinggi akurasi yang dihasilkan.

.

5.2 Saran

Penelitian ini masih memiliki banyak kekurangan yang memerlukan pengembangan lebih lanjut. Berdasarkan penelitian, pengujian ini dilakukan pada tabel yang jumlahnya relatif sedikit dan belum dapat dikatakan valid jika dibandingkan dengan metode lain. SVM diharapkan mampu diuji cobakan pada penelitian dengan data skala besar dan multikelas sehingga hasil akurasi pada penelitian selanjutnya tidak diragukan validitasnya.

DAFTAR PUSTAKA

Arini Daribri Putri. (2013). Klasifikasi Dokumen Teks Menggunakan Metode Support Vector Mechine. Bogor: Institut Pertanian Bogor

Chenometh Megan Song (2009) Text categorization. Encyclopedia of Data Warehouse and Data Mining

Dokument online, https://en.wikipedia.org/wiki/Chi-squared_test, di- akses pada Maret 2016

Dokument online, https://en.wikipedia.org/wiki/Supervised_learning, di- akses pada Maret 2016

Gijsberts A (2007) Evolutionary optimatization of kernel. Delft University of Technology

Hearst Marti (2003) What is text minning? SIMS UC Berkeley. Tersedia pada http://www.sims.berkeley.edu/~hearst/text.mining.html. di- akses pada Juni 2016

Kaiser, Katharina, Miksch, Silvia. 2005. Information extraction: a survey. Tersedia pada: http://ieg.ifs.tuwien.ac.at. di- akses pada Juni 2016

Husnawi. (2010). Teknik document object model (DOM) untuk manipulasi dokumen XML. J Dasi.

Manning CD, Raghavan P, Schütze H. (2008). An Introduction to Information Retrieval. Cambridge (GB): Cambridge Univ Pr.

Mesleh AA. (2007). Chi square feature extraction based SVM arabic language text categorization systems. J Computer Sci.

Netcraft 1995. How many active sites are there?. Tersedia pada: http://news.netcraft.com/active-sites/. di-akses pada Juni 2016.

Sari PD. (2012). Metode Pembobotan kata berbasis sebaran untuk temu kembali informasi dokumen Bahasa Indonesia. Bogor: Institut Pertanian Bogor

Saputra. (2012). Klasifikasi dokumen Bahasa Indonesia menggunakan *simantic smoothing* dengan ekstraksi ciri *ch*ekstraksi ciri *chi-square*. Bogor: Institut Pertanian Bogor