Implementasi Algoritma Term Frequency – Inverse Document Frequency dan Vector Space Model untuk Klasifikasi Dokumen Naskah Dinas

A. Achmad[1], A. A. Ilham[2], Herman[3]
[1]Program Studi Teknik Elektro, Jurusan Teknik Elektro, Universitas Hasanuddin, Makassar
[2] Program Studi Teknik Informatika, Jurusan Teknik Elektro, Universitas Hasanuddin, Makassar
[3]Balai Besar Pengkajian dan Pengembangan Komunikasi dan Informatika, Kementerian Kominfo, Makassar

Abstrak— **Pada kenyataannya dokumen naskah dinas diinstansi masih disimpan dan dicari secara manual. Penelitian ini bertujuan untuk merancang dan mengimplementasikan sistem klasifikasi dokumen naskah dinas secara otomatis dengan banyak kategori sehingga dapat mempermudah dalam penyimpanan dan pencarian dokumen naskah dinas.**

Penelitian ini menerapkan metode text mining dengan supervised learning menggunakan algoritma *term frequency – inverse document frequency* (TF-IDF) dan *vector space model*. Metode text mining digunakan untuk menentukan kata kunci dokumen secara otomatis. Algoritma TF-IDF melakukan pemberian bobot pada setiap kata kunci disetiap kategori dan *vector space model* untuk mencari kemiripan kata kunci dengan kategori yang tersedia. Implementasi dari sistem ini menghasilkan vektor pada setiap kategori sebagai data pembelajaran. sehingga nilai vektor tersebut akan dibandingkan dengan nilai dari kata kunci dokumen yang diuji untuk mencari kemiripan / *similiarity*.

Hasil penelitian menunjukkan bahwa algoritma TF-IDF dan *Vector Space Model* dapat mengklasifikasikan dokumen naskah dinas dengan banyak kategori dengan akurasi hasil klasifikasi 70%-75%.

Kata Kunci— **Klasifikasi dokumen, naskah dinas, TF-IDF, vector space model**

I. PENDAHULUAN

Dalam puluhan tahun terakhir, jumlah dokumen semakin lama semakin bertambah banyak dan beragam. Jika jumlah dokumen semakin bertambah banyak maka proses pencarian dan penyajian dokumen menjadi lebih sukar / sulit, sehingga akan lebih mudah jika dokumen tersebut sudah tersedia sesuai dengan kategorinya masing-masing. Klasifikasi dokumen teks adalah permasalahan yang mendasar dan penting. Didalam dokumen teks, tulisan yang terkandung adalah bahasa alami manusia, yang merupakan bahasa dengan struktur yang kompleks dan jumlah kata yang sangat banyak.
Oleh karenanya, sangatlah penting untuk bisa mengorganisir dan mengklasifikasi dokumen secara otomatis.

Pada kenyataannya masih banyaknya instansi pemerintah baik lembaga negara, pemerintah pusat dan daerah, perguruan tinggi negeri serta BUMN/D yang belum sepenuhnya melaksanakan pedoman tata naskah dinas khususnya dalam mengklasifikasikan naskah dinas sesuai dengan kategori yang secara umum telah diatur pada Peraturan Menteri Negara Pemberdayaan Aparatur Negara (PERMENPAN) nomor 22 tahun 2008 tentang Pedoman Umum Tata Naskah Dinas.

II. LATAR BELAKANG

A. *Text Mining*

Text mining merupakan penerapan konsep dan teknik data mining untuk mencari pola dalam teks, yaitu proses penganalisisan teks guna menyarikan informasi yang bermanfaat untuk tujuan tertentu. Adapun tugas khusus dari text mining antara lain yaitu pengkategorisasian teks (text categorization) dan pengelompokan teks (text clustering). [3]

Berdasarkan ketidakteraturan struktur data teks, maka proses text mining memerlukan beberapa tahap awal yang pada intinya adalah mempersiapkan agar teks dapat diubah menjadi lebih terstruktur. Tahapan text mining yang dilakukan secara umum adalah tahap *case folding* dan *tokenizing, filtering, stemming, tagging* dan *analyzing*.

B. *Term frequency – Inversed document frequency Algoritm (TF-IDF)*

Term frequency – inverse document frequency atau biasa sering disebut TF-IDF adalah metode pembobotan kata dengan menghitung nilai TF dan juga menghitung kemunculan sebuah kata pada koleksi dokumen teks secara keseluruhan.

Metode ini menggabungkan 2 konsep perhitungan bobot yaitu frekuensi kemunculan sebuah kata di dalam sebuah dokumen tertentu dan inverse frekuensi dokumen yang mengandung kata tersebut. *Inverse document frequency* (IDF) adalah jumlah dokumen yang mengandung sebuah term didasarkan pada seluruh dokumen yang ada pada data set.

$$idf = \log [n/dfi] \quad \text{..... (1)}$$

Keterangan :

- Nilai N adalah jumlah dokumen yang terdapat pada kumpulan dokumen yang diamati.
- Nilai dfi adalah jumlah dokumen yang mengandung term i.

Kemudian untuk proses pembobotan dari term yang ada menggunakan rumus sebagai berikut :

$$w_{t.d} = tf_{t.,d} \,.\, log\frac{n}{dfi} \quad \text{..... (2)}$$

Keterangan :
w_(t.d) = frekuensi term t pada dokumen d
log n/dfi = *Inverse document frequency* (idf)
n = banyaknya dokumen
dfi = banyaknya dokumen yang memiliki term t.

C. *Algoritma Model Ruang Vektor / Vector Space Model Algoritm.*

Model ruang vektor adalah suatu model yang digunakan untuk mengukur kemiripan antara suatu dokumen dengan suatu query. Pada model ini, query dan dokumen dianggap sebagai vektor-vektor pada ruang n-dimensi, dimana n adalah jumlah dari seluruh term yang ada dalam leksikon. Leksikon adalah daftar semua term yang ada dalam indeks.

Pada algoritma vector space model digunakan rumus untuk mencari nilai cosines sudut antara dua vector dari setiap bobot dokumen (WD) dan bobot dari kata kunci (WK). Rumus yang digunakan untuk mencari similiarity adalah sebagai berikut :

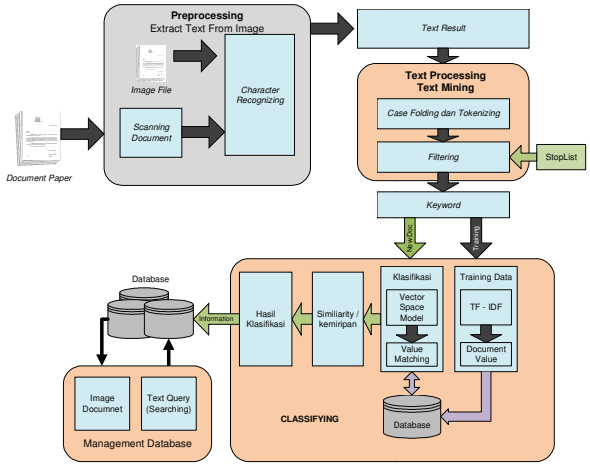
$$\cos ine(D_j, Q) = \frac{\langle D_j \cdot Q \rangle}{\|D_j\| \times \|Q\|} = \frac{\sum_{i=1}^{n} w_{ij} \times w_{iq}}{\sqrt{\sum_{i=1}^{n} w_{ij}^2} \times \sqrt{\sum_{i=1}^{n} w_{iq}^2}} \quad \ldots\ldots \quad (3)$$

Keterangan :

D = Vektor dokumen ke – j,
Q = Vektor query
Wij = Bobot dari kata kunci ke – i pada dokumen ke – j
Wiq = Bobot dari kata kunci ke – i pada query.
n = Banyaknya dimensi

Similiarity antar query dan dokumen dilihat dari besar sudut cosines yang dimiliki oleh vektor query dan vector dokumen yang akan dibandingkan. Jika kedua vector tersebut memiliki similiarity yang tinggi, maka query dan dokumen tersebut dianggap memiliki hubungan (Grossman, 2004).

III. METODOLOGI



Gambar 1. Blok Diagram Sistem

Adapun penjelasan fungsi dari masing-masing blok adalah sebagai berikut :

a) Blok Preprocessing

Preprocessing dalam sistem ini merupakan proses mengekstrak teks dari gambar. Fungsi dari proses ini adalah mengambil teks dari suatu dokumen. Namun, proses ekstraksi teks yang dapat dilakukan masih terbatas pada teks hasil ketikan dan masih belum dapat mendeteksi teks tulisan tangan.

Proses ini membutuhkan masukan file image (jpeg, bmp, png, tiff), dimana outputnya berupa teks. Output ini nantinya akan digunakan pada text processing.

b) Blok Text Processing

Text processing pada sistem ini berfungsi untuk mengolah kata-kata yang ada pada dokumen yang merupakan hasil ekstraksi sebelum digunakan dalam proses klasifikasi maupun pencarian dokumen. Text processing pada sistem ini terdiri dari beberapa tahap antara lain tahap case folding, tokenizing dan filtering.

Adapun yang menjadi input untuk proses ini berupa teks yang merupakan hasil keluaran dari blok preprocessing. Sedangkan outputnya berupa keyword / kata kunci yang pada akhir prosesnya membantu untuk menggambarkan kategori dari dokumen yang diklasifikasikan.

c) Blok *Classifying*

Classifying merupakan blok klasifikasi dokumen naskah dinas. Berfungsi untuk mengklasifikasikan dokumen naskah dinas ke dalam kategori yang diperoleh dari hasil pembelajaran / training. Menggunakan algoritma TF-IDF untuk memberi bobot pada keyword dan algoritma Vector space model untuk mencari kemiripan dengan kategori yang ada.

Inputan yang digunakan adalah kata kunci / keyword hasil text processing sedangkan output yang diharapkan berupa hasil vector yang akan menentukan kemiripan dokumen terhadap kategori yang ada.

d) Database Management

Database management berfungsi untuk mengelola file image. Proses pengelolaan yang dilakukan adalah menyimpan file image dan melakukan pencarian berdasarkan keyword.

Inputan yang digunakan berupa keyword / kata kunci untuk mencari dokumen yang dibutuhkan dan outputnya berupa dokumen yang relevan sesuai dengan kata kunci yang diinputkan.

IV. HASIL DAN PEMBAHASAN

Implementasi

Dalam implementasi metode klasifikasi dokumen naskah dinas terdiri dari 2 (dua) tugas utama yaitu pembelajaran / training dokumen dan klasifikasi dokumen.

Tahap Pembelajaran / Training dokumen

Pada tahap ini sistem akan membangun model yang berfungsi untuk menentukan kelas secara manual dari dokumen yang belum diketahui kelasnya. Tahap ini menggunakan data yang telah diketahui kelas datanya

(manual) yang kemudian akan dilatih agar membentuk model yang representasikan melalui vector dari setiap dokumen, dapat dilihat pada table berikut :

Tabel 1
Tahapan Proses *Training*

No	Tahap	Input	Output
1	Teks Mining	Teks (Hasil ekstraksi gambar)	Kata kunci / Keyword
2	Pembuatan Matriks	Kata kunci / Keyword	Matriks J
3	Hitung TF*IDF (W_{ij})	Matriks J	Bobot J pada n dimensi
4	Hitung Cosines Vektor	Bobot J pada n dimensi	Vektor dari setiap kategori

Berikut ini merupakan contoh proses text mining. Dimana digunakan satu dokumen sebagai contoh untuk menampilkan proses pada tahap training. Adapun hasil ekstraksi teks pada Dokumen1.jpeg sebagai berikut :
"BERITA ACARA SEMINAR USULAN PENELITIAN Pada hari ini Selasa tanggal 22 mei 2012 jam 10.00 bertempat Ruang Sidang Jurusan Teknik Sipil di Ruang Fak. Teknik UNHAS telah diadakan evaluasi nilai seminar usulan penelitian".

Tabel 2.
Hasil filtering dokumen 1.jpeg

Hasil Case Folding dan Tokenizing	*Filtering*	Hasil (Kata Kunci)
berita		berita
acara		acara
seminar		seminar
usulan		usulan
penelitian		penelitian
pada	dihapus (*stoplist*)	
selasa	dihapus (*stoplist*)	
bertempat		bertempat
ruang		ruang
sidang		sidang
jurusan		jurusan
di	dihapus (*stoplist*)	
ruang		ruang
fak		fak
teknik		teknik
unhas	dihapus (*stoplist*)	
telah	dihapus (*stoplist*)	
diadakan		diadakan
evaluasi		evaluasi
nilai		nilai
seminar		seminar
usulan		usulan
penelitian		penelitian

Langkah awal untuk melakukan pelatihan klasifikasi dokumen adalah dengan membangun sebuah matriks A berukuran M x N, dimana M adalah kata-kata kunci dan N adalah kategori.

Oleh karena dilakukan ke 5 kategori dengan masing-masing 10 dokumen. Kemudian menghitung TF, DF, IDF dan bobot (W) setiap kata kunci pada setiap kategori berdasarkan persamaan (1), (2) dan (3). Sehingga hasil perhitungan tersebut menghasilkan matriks sebagai berikut :

Tabel 3.

Matriks Perhitungan TF-IDF dengan 5 kategori

TERM	TF					DF	IDF	W = tf * idf				
	K1	*K2*	*K3*	*K4*	*K5*			*K1*	*K2*	*K3*	*K4*	*K5*
berita	1	1	2	1	2	5	0	0	0	0	0	0
acara	1	2	1	1	0	4	0.097	0.097	0.194	0.097	0.097	0
seminar	2	1	1	2	1	5	0	0	0	0	0	0
usulan	2	0	2	2	2	4	0.097	0.194	0	0.194	0.194	0.194
penelitian	2	2	2	2	2	5	0	0	0	0	0	0
bertempat	1	1	1	1	1	5	0	0	0	0	0	0
ruang	2	1	2	2	2	5	0	0	0	0	0	0
sidang	1	0	1	0	1	3	0.222	0.222	0	0.222	0	0.222
jurusan	1	2	0	1	0	3	0.222	0.222	0.444	0	0.222	0
fak	1	0	0	0	0	1	0.699	0.699	0	0	0	0
teknik	1	1	1	1	0	4	0.097	0.097	0.097	0.097	0.097	0
diadakan	1	1	0	0	0	2	0.398	0.398	0.398	0	0	0
evaluasi	1	1	1	0	2	4	0.097	0.097	0.097	0.097	0	0.194
nilai	1	1	0	0	0	2	0.398	0.398	0.398	0	0	0
diberlakukan	1	1	0	1	0	3	0.222	0.222	0.222	0	0.222	0
mengikuti	0	1	1	2	1	4	0.097	0	0.097	0.097	0.194	0.097
peraturan	0	1	2	1	2	4	0.097	0	0.097	0.194	0.097	0.194
pemerintah	0	1	0	2	1	3	0.222	0	0.222	0	0.444	0.222
pegawai	0	1	1	1	1	4	0.097	0	0.097	0.097	0.097	0.097

Berdasarkan hasil pembobotan (Wij) pada tabel 3, maka akan dihasilkan vektor dari setiap kategori sebagai berikut :

K1 =(0, 0.097, 0, 0.097, 0, 0, 0, 0.222, 0.222, 0.699, 0.097, 0.398, 0.097, 0.398, 0.222, 0, 0, 0, 0)

K2 = (0, 0.194, 0, 0, 0, 0, 0, 0, 0.444, 0, 0.097, 0.398, 0.097, 0.398, 0.222, 0.097, 0.097, 0.222, 0.097)

K3 = (0, 0.097, 0, 0.194, 0, 0, 0, 0, 0, 0, 0.444, 0.097, 0.398, 0.097, 0.398, 0.222, 0.097, 0.097, 0.222, 0.097)

K4 = (0, 0.097, 0, 0.194, 0, 0, 0, 0, 0.222, 0. 0.097, 0, 0, 0, 0.222, 0.194, 0.097, 0.444, 0.097)

K5 = (0, 0, 0, 0.194, 0, 0, 0, 0.222, 0, 0, 0, 0, 0.194, 0, 0, 0.097, 0.194, 0.222, 0.097)

1) Tahap Klasifikasi Dokumen

Tahap selanjutnya adalah melakukan pengujian implementasi klasifikasi dokumen. Pada tahap ini sistem melakukan proses yang sama seperti pada tahap training tetapi setelah memperoleh hasil vektor kemudian dibandingkan dengan vector pada data training yang kemudian dihitung similiarity / kemiripan. Adapun alur proses dalam tahap ini sebagai berikut :

Tabel 4

Tahapan Proses Klasifikasi

NO	TAHAP	INPUT	OUTPUT
1	Teks Mining	Teks (Hasil ekstraksi gambar)	Kata kunci / Keyword
2	Pembuatan Matriks	Kata kunci / Keyword	Matriks Q
3	Melakukan penentuan Q dan perhitungan matriks	Matriks Q	Vektor Q dari n dimensi
4	Hitung Cosines Measure	Vektor Q dari n dimensi	Similiarity (Q,D)
5	Menentukan Nilai CosinesTertinggi	Similiarity (Q,D)	Kategori dari Q

Setelah proses text mining dilakukan maka selanjutnya melakukan proses perhitungan matriks dilakukan dengan terlebih dahulu melakukan pencarian kata kunci yang sama pada database dengan kata kunci yang diperoleh dari dokumen yang akan diklasifikasikan, apabila kata kunci ditemukan maka Q bernilai 1 sebaliknya jika kata kunci tidak sama maka Q bernilai 0, sehingga dapat ditentukan bobot Q sebagai berikut :

Tabel 5.

Penentuan bobot Q (Wiq).

TERM	W = tf * idf					Q
	K1	K2	K3	K4	K5	
berita	0	0	0	0	0	0
acara	0.097	0.194	0.097	0.097	0	0
seminar	0	0	0	0	0	0
usulan	0.194	0	0.194	0.194	0.194	0
penelitian	0	0	0	0	0	0
bertempat	0	0	0	0	0	0
ruang	0	0	0	0	0	0
sidang	0.222	0	0.222	0	0.222	0
jurusan	0.222	0.444	0	0.222	0	0
fak	0.699	0	0	0	0	0
teknik	0.097	0.097	0.097	0.097	0	0
diadakan	0.398	0.398	0	0	0	0
evaluasi	0.097	0.097	0.097	0	0.194	0
nilai	0.398	0.398	0	0	0	0
diberlakukan	0.222	0.222	0	0.222	0	1
mengikuti	0	0.097	0.097	0.194	0.097	0
peraturan	0	0.097	0.194	0.097	0.194	1
pemerintah	0	0.222	0	0.444	0.222	1
pegawai	0	0.097	0.097	0.097	0.097	1
Jumlah	**2.645**	**2.362**	**1.094**	**1.663**	**1.219**	

Keterangan :

Q = Kata Kunci yang sama dengan term pada matriks Wij

Wiq = Bobot kata kunci Q dari matriks Wij

Matriks untuk perhitungan Wiq dapat dilihat pada Tabel 6.

Tabel 6. Perhitungan bobot Q (Wiq).

TERM	W = tf * idf					Q
	K1	K2	K3	K4	K5	
diberlakukan	0.222	0.222	0	0.222	0	1
peraturan	0	0.097	0.194	0.097	0.194	1
pemerintah	0	0.222	0	0.444	0.222	1
pegawai	0	0.097	0.097	0.097	0.097	1
Jumlah	**0.222**	**0.638**	**0.291**	**0.859**	**0.513**	

Untuk menghitung nilai cosine, terlebih dahulu dihitung total bobot setiap kategori, kemudian hitung akarnya (sqrt), kemudian kalikan antara kueri (Q) dengan bobot setiap katakunci disetiap kategori (W), kemudian hitung nilai cosinenya sebagai berikut :

Cosine. Kj = QKj * Wkj

Cosine. K1 = QK1 * Wk1 = 0,002 * 1,019 = 0,002

Cosine. K2 = 0,005 * 0,697 = 0,005

Cosine. K3 = 0,0001 * 0,171 = 0,003

Cosine. K4 = 0,012 * 0,408 = 0,030

Cosine. K5 = 0,003 * 0,230 = 0,012

Sehingga matriks hasil perhitungan cosine measure untuk menghitung kemiripan antara kata kunci dengan term pada database adalah sebagai berikut :

Tabel 7

Perhitungan Cosine measure

TERM	Q^2	W = tf * idf					QK				
		$K1^2$	$K2^2$	$K3^2$	$K4^2$	$K5^2$	Q x K1	Q x K2	Q x K3	Q x K4	Q x K5
berita	0	0	0	0	0	0	0	0	0	0	0
acara	0	0.009392	0.037566	0.009392	0.009392	0	0	0	0	0	0
seminar	0	0	0	0	0	0	0	0	0	0	0
usulan	0	0.037566	0	0.037566	0.037566	0.037566	0	0	0	0	0
penelitian	0	0	0	0	0	0	0	0	0	0	0
bertempat	0	0	0	0	0	0	0	0	0	0	0
ruang	0	0	0	0	0	0	0	0	0	0	0
sidang	0	0.049217	0	0.049217	0	0.049217	0	0	0	0	0
jurusan	0	0.049217	0.196867	0	0.049217	0	0	0	0	0	0
fak	0	0.488559	0	0	0	0	0	0	0	0	0
teknik	0	0.009392	0.009392	0.009392	0.009392	0	0	0	0	0	0
diadakan	0	0.158356	0.158356	0	0	0	0	0	0	0	0
evaluasi	0	0.009392	0.009392	0.009392	0	0.037566	0	0	0	0	0
nilai	0	0.158356	0.158356	0	0	0	0	0	0	0	0
diberlakukan	0.049	0.049217	0.049217	0	0.049217	0	0.0024	0.002	0	0.0024	0
mengikuti	0	0	0.009392	0.009392	0.037566	0.009392	0	0	0	0	0
peraturan	0.009	0	0.009392	0.037566	0.009392	0.037566	0	9E-05	0.0004	9E-05	0.0004
pemerintah	0.049	0	0.049217	0	0.196867	0.049217	0	0.002	0	0.0097	0.0024
pegawai	0.009	0	0.009392	0.009392	0.009392	0.009392	0	9E-05	9E-05	9E-05	9E-05
SUM	**0.117**	**1.019**	**0.697**	**0.171**	**0.408**	**0.230**	**0.002**	**0.005**	**0.000**	**0.012**	**0.003**
SQRT	**0.342**	**1.009**	**0.835**	**0.414**	**0.639**	**0.479**	**0.049**	**0.071**	**0.021**	**0.111**	**0.054**
Cosine		**0.002**	**0.007**	**0.003**	**0.030**	**0.012**					

Setelah perhitungan cosine measure maka didapatkan nilai cosine dokumen terhadap kategori. Dari hasil perhitungan pada tabel 7, diperoleh nilai kosinus tertinggi adalah kategori 4 sebesar 0,030 sehingga K4 dinyatakan sebagai kategori dari dokumen1.jpeg, dapat dilihat sebagai berikut :

Tabel 8

Peringkat Kategori berdasarkan kemiripannya

Rangking	Kategori	Similiarity
I	K_4	0,030
II	K_5	0,012
III	K_2	0,007
IV	K_3	0,003
V	K_1	0,002

A. Pengujian

Dari hasil pengujian yang telah dilakukan berdasarkan jumlah data training yang ada dengan tingkatan jumlah data latih / data training sehingga diperoleh hasil sebagai berikut :

Tabel 9.

Jumlah data latih yang akan digunakan

KATEGORI	JUMLAH DATA LATIH (P1)	JUMLAH DATA LATIH (P2)	JUMLAH DATA LATIH (P3)
Surat Edaran	10	15	20
Surat Perintah / Tugas	10	15	20
Surat Pengantar	10	15	20
Nota Dinas	10	15	20
Berita Acara	10	15	20
JUMLAH	**50**	**75**	**100**

Keterangan :

P1 = Jumlah data latih / training yang digunakan pada percobaan 1.

P2 = Jumlah data latih / training yang digunakan pada skenario 3 dengan menambahkan 25 data latih,

dimana setiap kategori ditambahkan masing-masing 5 data latih.

P3 = Jumlah data latih / training yang digunakan pada skenario 3 dengan menambahkan 25 data latih lagiberdasarkan Percobaan 2 (P2), dimana setiap kategori ditambahkan masing-masing 5 data latih.

Hasilnya sebagai berikut:

Tabel 10

Rekapitulasi Hasil pengujian

DATA LATIH	DATA UJI	KLASIFIKASI BENAR	PERSENTASE
50	20	14	70 %
75	20	15	75 %
100	20	15	75 %

Hasil pengujian terhadap 20 dokumen uji dengan jumlah data latih yang berbeda, dimana jumlah data latih terus ditambah menghasilkan adanya peningkatan hasil klasifikasi dari 70% menjadi 75%..

Namun disamping itu juga terdapat faktor lain yang turut mempengaruhi hasil klasifikasi yaitu dokumen fisik naskah dinas yang dapat menghasilkan karakter-karakter yang tidak jelas setelah proses ekstraksi teks sehingga bukan hanya menghilangkan kata kunci yang dibutuhkan untuk proses klasifikasi tetapi menambah kata kunci baru yang tidak dibutuhkan yang hanya menambah panjang waktu komputasi / perhitungan

IV. KESIMPULAN DAN SARAN

A. Kesimpulan

Berdasarkan hasil penelitian sebagaimana telah diuraikan dalam bab sebelumnya, maka dapat disimpulkan sebagai berikut:

1) Klasifikasi dokumen menggunakan algoritma TF-IDF dan vector space model dapat mengklasifikasikan dokumen naskah dinas.
2) Hasil pengujian dengan menggunakan dokumen yang belum pernah dilatih / training sebelumnya dapat menghasilkan akurasi hasil klasifikasi dengan kisaran 70 – 75%.
3) Hasil pengujian menunjukkan bahwa ada peningkatan akurasi hasil klasifikasi menggunakan 75 dan 100 dokumen sebagai data training dibandingkan dengan menggunakan 50 dokumen sebagai data training.
4) Bentuk fisik asli dokumen yang sudah rusak / usang / kabur mempengaruhi hasil pembacaan dokumen sehingga dapat menghasilkan kata kunci yang berbeda atau baru yang tidak konsisten yang akhirnya dapat merubah hasil klasifikasi yang diperoleh.
5) Penentuan kategori dokumen secara manual pada saat pembelajaran / training berhasil untuk mendapatkan pemodelan yang benar mengingat banyaknya kategori dokumen naskah dinas.

B. Saran

Hasil penelitian ini belum sempurna, oleh karenanya untuk meningkatkan hasil yang dicapai dapat dilakukan hal-hal sebagai berikut:

1) Kekurangan sistem ini adalah hanya menggunakan teknik *Optical Character Recognition* (OCR) yang hanya mendeteksi karakter yang berasal dari hasil ketikan tetapi belum dapat mendeteksi karakter tulisan tangan sehingga belum dapat menjadikan tulisan tangan sebagai teks dalam proses klasifikasi dokumen. Sehingga diharapkan agar dapat dikembangkan sistem OCR yang dapat mendeteksi tulisan tangan yang dapat diintegrasikan ke dalam sistem ini.
2) Belum adanya sistem yang melakukan perbaikan pada hasil ekstraksi teks dokumen sehingga perlunya dikembangkan suatu sistem untuk perbaikan hasil pembacaan secara otomatis yang dapat diintegrasikan dengan sistem ini sehingga menghasilkan tingkat akurasi klasifikasi yang lebih baik.
3) Untuk menambah keakuratan dan kecepatan waktu perhitungan komputasi sistem maka sebaiknya perlu dilakukan penelusuran mengenai daftar stoplist yang relevan untuk dokumen naskah dinas

REFERENCES

[1] Arief, Achmad Fauzi. (2010). *Perangkat Lunak Pengkonversi Teks Tulisan Tangan Menjadi Teks Digital.*
[2] Aunurokhman, Ahmad Hatta. Digital Documents Management System Using Text mining; 2010
[3] Aziz, M. I. (2010). Development Program Application To The Measurement Of Documents Resemblance Text mining, TF-IDF, And Vector space model Algoritm.
[4] Basnur, P. W., & Sensuse, D. I. (April 2010). Pengklasifikasian Otomatis Berbasis Ontologi Untuk Artikel Berita Berbahasa Indonesia. Makara, Teknologi, Vol. 14, No.2 , 29-35.
[5] Chenometh, Megan, Song, Min (2009) Text Categorization, dalam Encyclopedia of Data Warehouse & Data Mining, IGI Global, hal. 1936-1941
[6] Hasibuan, Z. A. (2007). Metodologi Penelitian Pada Bidang Ilmu Komputer dan Teknologi Informasi. Makassar.
[7] Oktanty, Rhizzajian. (2010). Design Structure Of Information System Decree In Faculty Of Information Techonology.
[8] Permenpan. (2008). Pedoman Umum Tata Naskah Dinas.
[9] Umar, Husein. (2008) Metode Penelitian untuk Skripsi dan Tesis Bisnis. PT. Rajagrafindo Persada.