Seminar Nasional Sistem Informasi dan Teknologi Informasi 2018
SENSITEK 2018
STMIK Pontianak, 12 Juli 2018

Analisis Pengklasifikasian Dokumen dengan Pembobotan Frekuensi Kata Berbanding Logaritma Jumlah Kata Serta Fungsi Cosine

Krisna Adiyarta[1)], **Yohana Yohana**[2)]

Program Studi Magister Ilmu Komputer, Program Pascasarjana, Universitas Budi Luhur
Jl. Raya Cileduk, Petukangan Utara, Kebayoran Lama, Jakarta Selatan 12260
Telp. (021)5853753, Fax. (021)5869225
e-mail : hanayoha77@yahoo.co.id

Abstrak

Dokumen merupakan suatu kumpulan data yang berguna sebagai referensi dalam penulisan suatu karya tulis ilmiah maupun non ilmiah yang dapat dimanfaatkan sebagai referensi tulisan. Dengan dukungan referensi yang tepat maka suatu karya dapat dikatakan kredibel dan dapat dipercaya. Namun pertumbuhan yang pesat dari jumlah dokumen informasi maka diperlukan sebuah teknik pencarian yang tepat untuk dapat menemukan dokumen sesuai kebutuhan salah satunya dengan sistem temu kembali informasi (information retrieval system/IRS). Salah satu teknik IRS yang dapat digunakan untuk merepresentasikan profil dokumen adalah model ruang vektor (vector space model). Pembobotan yang didasarkan pada term dengan teknik stemming untuk mendapatkan bentuk kata dasar dari term yang bersangkutan. Penelitian ini akan menguji aplikasi mesin klasifikasi teks bahasa Indonesia yang mengunakan algoritma stemming Nazief Adriani, algoritma K-Nearest Neighbor dan metode Vector Space Model berdasarkan pembobotan frekuensi kata berbanding logaritma jumlah kata serta fungsi cosine. Dengan penelitian ini dapat disimpulkan bahwa proses pengkategorian dokumen teks berbahasa Indonesia yang dilakukan melalui perbandingan antara dokumen uji dengan koleksi sampel pengujian mendapatkan hasil yang sesuai dengan kategori yang telah ditentukan setelah diuji dan juga telah dilakukan pengukuran precision dan recall untuk mengetahui efektifitas proses klasifikasi.

Kata kunci: klasifikasi dokumen, cosine, stemming, K-Nearest Neighbor

1. Pendahuluan

Kebutuhan akan informasi yang selalu jumlahnya berkembang pesat dimana kumpulan data dari berbagai informasi dikumpulkan untuk suatu kebutuhan. Informasi dapat berupa audio, video, dan teks. Salah satu bentuk informasi yang sering digunakan adalah data teks yang tersimpan dalam bentuk dokumen PDF. Setiap dokumen tersebut terdiri dari beberapa kalimat yang merupakan kumpulan kata-kata yang menjadi sebuah kalimat.

Pada setiap dokumen teks terdiri dari sekumpulan kata-kata, dimana tiap kata memiliki ciri khas yang berbeda. Oleh karena itu, pada sebagian besar proses kategorisasi teks, terdapat banyak ciri khas yang mungkin terjadi, baik ciri khas yang relevan dengan tema dokumen maupun yang tidak relevan dari proses kategorisasi. Adapun metode yang mengelompokkan semua ciri khas tersebut cenderung lebih baik daripada metode yang hanya mengelompokkan ciri khas yang relevan [1].
Adanya pertumbuhan yang pesat dari jumlah dokumen informasi maka diperlukan sebuah teknik pencarian yang tepat untuk dapat menemukan dokumen sesuai kebutuhan. Dokumen berisi informasi sesuai pencarian tersebut dapat dimanfaatkan sebagai referensi tulisan.
Klasifikasi adalah salah satu bentuk aplikasi sistem yang memerlukan kemampuan untuk memahami materi atau topik yang terkandung dari sebuah dokumen teks. Kategori adalah bagian dalam suatu sistem klasifikasi. Kategorisasi teks (text categorization) merupakan salah satu tahap pemrosesan dokumen pada information retrieval, dimana dokumen-dokumen yang ada dikelompokkan atau diklasifikasikan ke dalam beberapa topik atau tema [2].

Dalam proses klasifikasi kata dapat digunakan sistem temu kembali informasi (*information retrieval system/IRS*). Salah satu teknik IRS yang dapat digunakan untuk merepresentasikan profil dokumen adalah model ruang vektor (*vector space model*). Model ruang vektor (*vector space model*) yang menekankan kepada teknik pembobotan berdasarkan kata-kata (*term)*. Menurut Zafikri [3], baik *query* maupun dokumen teks yang disimpan dinyatakan dalam bentuk vektor. Pembobotan yang didasarkan pada *term* selalu dikaitkan dengan teknik *stemming* untuk mendapatkan bentuk kata dasar dari *term* yang bersangkutan. *Stemming* adalah proses klasifikasi kata dasar pada suatu kata. Pada proses klasifikasi, imbuhan merupakan bagian dari informasi yang perlu dihilangkan untuk mencapai efektivitas pada proses klasifikasi. Penelitian ini akan menguji aplikasi mesin klasifikasi dokumen teks bahasa Indonesia yang mengunakan algoritma *stemming* Nazief Adriani, algoritma K-*Nearest Neighbor* dan metode *Vector Space Model* berdasarkan pembobotan frekuensi kata berbanding logaritma jumlah kata, fungsi *similarity cosine*. Dengan penelitian ini diharapkan proses

pengkategorian dokumen teks berbahasa Indonesia yang dilakukan secara terkomputerisasi akan mendapatkan hasil yang sesuai dengan pengkategorian secara manual serta dapat diketahui risiko-risiko apa saja yang dapat terjadi dalam pengembangan teknologi informasi.

2. Analisis Pengklasifikasian Dokumen

Proses yang menjadi komponen utama dalam sebuah aplikasi mesin klasifikasi:

1. *Convert* Dokumen

Convert Dokumen adalah proses dimana dokumen *PDF* yang di-*upload* kemudian diubah formatnya menjadi dokumen teks [4], yang kemudian isi dokumen teks tersebut akan diproses lebih lanjut untuk mencari nilai *token* di dokumen tersebut. Dokumen diambil dari situs berita [5].

2. *Tokenizing*

Sebuah dokumen berisi kumpulan kalimat-kalimat. Setelah melalui proses *wordtoken* maka kalimat-kalimat tersebut akan dipecah menjadi kata per kata. Tanda baca dan karakter spesial lainnya akan ikut terbuang dari kumpulan kata tersebut [6].

3. Proses *Stopword Removal*

Stopword adalah kumpulan kata yang sering muncul pada dokumen dan dianggap tidak memiliki arti. *Stopword Removal* merupakan proses yang dilakukan untuk menghilangkan kata yang sering ditampilkan dalam berbagai kategori dokumen.

4. Proses *Stemming*

Merupakan proses mengubah sebuah kata menjadi kata dasar dengan menghilangkan imbuhan yang dapat berupa awalan dan akhiran. Proses ini dilakukan setelah melalui proses filtrasi / *stopword*. Dalam penelitian ini, menggunakan algoritma *stemming Nazief Adriani* [7].

5. Proses pembobotan *term* (*Term Weight*)

Term Weight merupakan proses dimana setiap *term* yang ada dalam dokumen diberikan nilai bobot. Proses pembobotan kata yaitu proses pemberian bobot tiap kata yang dihitung dari jumlah kemunculan kata dalam sebuah *query* maupun isi dari sebuah dokumen [8].

6. Proses penghitungan nilai *Similarity (Similarity Measurement)*

Similarity Measurement merupakan proses pengukuran kemiripan dokumen yang dimiliki dengan *query* yang dimasukan. *Similarity measurement* ini digunakan setelah nilai *term weight* ditemukan. Algoritma yang digunakan pada aplikasi ini adalah fungsi *Cosine* sebagai algoritma untuk menghitung nilai *similarity*.

7. Proses perangkingan dokumen

Proses perangkingan dokumen menggunakan Algoritma *K-Nearest Neighbor*. *K-Nearest Neighbor (K-NN)* adalah sebuah metode untuk melakukan klasifikasi terhadap objek berdasarkan data *learning* yang jaraknya paling dekat dengan objek tersebut. K-NN termasuk algoritma *supervised learning* dimana *query instance* yang baru diklasifikasikan berdasarkan mayoritas dari kategori pada K-NN. Kelas yang paling banyak muncul-lah yang akan menjadi hasil klasifikasi. Proses ini mengurutkan dokumen dari yang tertinggi ke rendah berdasarkan nilai *similarity*.

Tabel 1. *Koleksi Sampel Pengujian*

No	Nama Dokumen	Kategori
1	100%Desa_di_Bangka_Belitung_Sudah_Berlistrik.pdf	Energi
2	Belajar_Soal_Listrik_Tenaga_Surya,Jonan_Akan_ke_Uni_Emirat_Arab.pdf	Energi
3	Dirut_PLN_Heran,Subsidi_Listrik_Orang_Kaya_Lebih_Besar_dari_Warga_Desa_Papua.pdf	Energi
4	ESDM-Harga_Solar_dan_Premium_Tetap,Hanya_Pertamax_Cs_yang_Naik.pdf	Energi
5	Harga_Pertamax_Cs_Naik_Rp300_per_Liter_Hari_Ini.pdf	Energi
6	Jokowi_Masih_Pasang_Target_35.000MW_Meski_Realisasi_Hanya_22.000MW.pdf	Energi
7	Jokowi_Tidak_Ingin_RI_Hanya_Bergantung_ke_BBM.pdf	Energi
8	Jokowi-Jangan_Biarkan_Rakyat_Daerah_Terpencil_Dapat_BBM_Harga_Berlipat.pdf	Energi
9	Penerima_Subsidi_Listrik_450_VA_Bertambah_4_Juta_Tahun_Ini.pdf	Energi
10	PGN_Alirkan_Gas_ke_Pemasok_Suku_Cadang_Grup_Astra.pdf	Energi
11	PGN_Bangun_Pipa_Gas_Bumi_Gresik-Lamongan-Tuban_Sepanjang-141Km.pdf	Energi
12	PLN_Siapkan_Rp7_Triliun_Terangi_2.700_Desa_Tahun_Ini.pdf	Energi
13	Akhir_Tahun_industri.pdf	Industri
14	Darmin_Kunjungi_Pabrik.pdf	Industri
15	Industri_Alas_kaki.pdf	Industri
16	Ini_Alasan_Ekspor_Tekstil_RI_Kalah.pdf	Industri
17	Investasi_Industri_Semen.pdf	Industri
18	Kalah_dengan_impor.pdf	Industri
19	Nasib_Industri_Tekstil_RI.pdf	Industri
20	Pembangunan_3_Kawasan_Industri.pdf	Industri
21	Perajin_Sepatu_Lokal.pdf	Industri
22	RI_Kembangkan_Industri_Ramah.pdf	Industri
23	RI_Masih_Impor_Kelapa_untuk_Kebutuhan_Industri.pdf	Industri
24	RI_Masih_Impor_Teh.pdf	Industri
25	Sektor_Industri_Berkontribusi_17.pdf	Industri
26	jababeka_pp_properti.pdf	Properti
27	Jitu_memilih_lahan.pdf	Properti
28	Jokowi_Teken_PP64.pdf	Properti
29	Keuntungan_lain.pdf	Properti
30	Lippo_Karawaci_Resmi.pdf	Properti
31	Masuk_Bisnis_Properti.pdf	Properti
32	Menteri_PUPR.pdf	Properti
33	Pameran_Rumah_Murah.pdf	Properti
34	Pemerintah_Sudah_Bangun_410.pdf	Properti
35	Retaknya_jembatan.pdf	Properti
36	RI_Mau_Mulai_Tabungan_Perumahan_Rakyat.pdf	Properti
37	Selain_Disubsidi_KPR.pdf	Properti
38	Usai_tax_amnesty.pdf	Properti
39	Begini_Cara_Jonan_Bikin_Listrik_di_Sumsel_dan_Kalimantan_Jadi_Murah.pdf	Energi

Sampel pengujian berisi atas 39 dokumen baik berupa artikel berita yang diambil secara acak dari situs berita detik.com dengan kategori energi, industri dan properti. Semua dokumen yang dijadikan sampel telah diubah ke dalam bentuk .pdf. Tabel 1 menunjukkan daftar koleksi pengujian yang digunakan.
Kumpulan data-data tersebut disebut sebagai *data set* yang digunakan untuk learning dokumen ke database aplikasi. Hal ini ditujukan supaya aplikasi atau mesin mengenal dokumen-dokumen sesuai dengan kategori yang dinput oleh pengguna (*user*)
Berdasarkan bentuk Tabel dapat diketahui ada 3 baris berisikan nomor urut, nama dokumen dan kategori. Tabel 2 menunjukkan rincian dokumen berdasarkan kategori.

Tabel 2. *Jumlah Koleksi Pengujian Berdasarkan Kategori*

No	Kategori	Jumlah Dokumen
1	Properti	13
2	Industri	13
3	Energi	13

4.2 Tahapan *TextPreprocessing*

Textpreprocessing adalah suatu tahapan mempersiapkan teks-teks sebelum dilakukan proses klasifikasi. Untuk menjelaskan tahapan *text preprocessing* maka akan digunakan contoh kalimat untuk menjelaskan tiap tahapan dengan lebih efektif. Diumpamakan dokumen sudah dikonversi dari bentuk pdf ke dalam bentuk teks.
Contoh :

a. Dokumen 1 (Kategori Energi)
Energi yang dibutuhkan masyarakat harus bisa menjangkau semua wilayah geografis.

b. Dokumen 2 (Kategori Properti)
Pemerintah menargetkan pembangunan 410 rumah bagi masyarakat berpenghasilan rendah.

c. Dokumen 3 (Kategori Industri)
Industri kreatif mulai mendapat perhatian lebih dari pemerintah

Dokumen a,b,c merupakan dokumen *learning* yang sudah didaftarkan pada basis data aplikasi klasifikasi. Setiap dokumen sudah memiliki kategori masing-masing. Setelah mengumpulkan dokumen learning, maka akan dilakukan klasifikasi pada sebuah dokumen uji dalam kasus ini kita sebut *query*.

Query (Kategori Tujuan : Properti)
Perumahan bagi masyarakat ekonomi rendah akan dibangun sesuai dengan rencana pemerintah.

Berdasarkan pada kalimat yang dikandung dalam *query* di atas, peneliti mengasumsikan bahwa kategori dokumen tersebut adalah properti. Untuk itu perlu dibuktikan dengan aplikasi, sebelum diklasifikasi maka perlu dilakukan tahapan *preprocessing* dengan urutan sebagai berikut.

1. *Tokenizing*

Proses ini memecah kalimat menjadi kata-kata seperti yang ditunjukkan pada Tabel 3.

Tabel 3. *Hasil Proses Tokenizing*

Q1	D1	D2	D3
perumahan	energi	pemerintah	industri
bagi	yang	menargetkan	kreatif
masyarakat	dibutuhkan	pembangunan	mulai
ekonomi	masyarakat	410	mendapat
rendah	harus	rumah	perhatian
akan	bisa	bagi	lebih
dibangun	menjangkau	masyarakat	dari
sesuai	semua	berpenghasilan	pemerintah
dengan	wilayah	rendah	
rencana	geografis		
pemerintah			

2. *Stopword Removal*

Setelah dokumen dipecah menjadi kata per kata maka selanjutnya dilakukan proses membuang kata yang tidak memiliki arti seperti yang ditandai dengan warna merah pada Tabel 4.

Tabel 4. *Sebelum Proses Stopword Removal*

Q1	D1	D2	D3
perumahan	energi	pemerintah	industri
bagi	yang	menargetkan	kreatif
masyarakat	dibutuhkan	pembangunan	mulai
ekonomi	masyarakat	410	mendapat
rendah	harus	rumah	perhatian
akan	bisa	bagi	lebih
dibangun	menjangkau	masyarakat	dari
sesuai	semua	berpenghasilan	pemerintah
dengan	wilayah	rendah	
rencana	geografis		
pemerintah			

Kata-kata yang berwarna merah pada Tabel 4 merupakan kata yang tidak memiliki arti jika tidak dipasangkan dengan kata lain atau disebut juga kata penghubung. Kata-kata penghubung tersebut dibuang untuk mempermudah dan mmbuat proses klasifikasi lebih efisien.

Tabel 5. *Hasil Proses Stopword Removal*

Q1	D1	D2	D3
perumahan	energi	pemerintah	industri
masyarakat	dibutuhkan	menargetkan	kreatif
ekonomi	masyarakat	pembangunan	mulai
rendah	menjangkau	410	mendapat
dibangun	wilayah	rumah	perhatian
rencana	geografis	masyarakat	pemerintah
pemerintah		berpenghasilan	
		rendah	

3. *Stemming*

Setelah didapat kata-kata yang sudah dilakukan proses pembuangan kata penghubung seperti yang ditunjukkan pada Tabel 5 maka selanjutnya dilakukan proses stemming. Proses stemming adalah proses membuang imbuhan pada kata yang berimbuhan dan mengubahnya menjadi kata dasar. Pada penelitian ini algoritma yang digunakan adalah algoritma Nazief Adriani.

Tabel 6. *Daftar Kata Sebelum Proses Stemming*

Q1	D1	D2	D3
perumahan	energi	pemerintah	industri
masyarakat	dibutuhkan	menargetkan	kreatif
ekonomi	masyarakat	pembangunan	mulai
rendah	menjangkau	410	mendapat
dibangun	wilayah	rumah	perhatian
rencana	geografis	masyarakat	pemerintah
pemerintah		berpenghasilan	
		rendah	

Pada Tabel 6 terdapat beberapa kata yang berwarna merah pada setiap dokumen. Kata-kata tersebut memiliki imbuhan dan bukan merupakan kata dasar sehingga perlu dilakukan proses stemming untuk mendapatkan kata dasarnya. Kata-kata yang terdapat pada Tabel 7 merupakan daftar kata yang siap melalui proses klasifikasi.

Tabel 7. *Daftar Kata Setelah Proses Stemming*

Q1	D1	D2	D3
rumah	energi	pemerintah	industri
masyarakat	butuh	target	kreatif
ekonomi	masyarakat	bangun	mulai
rendah	jangkau	410	dapat
bangun	wilayah	rumah	perhatian
sesuai	geografis	masyarakat	pemerintah
rencana		hasil	
pemerintah		rendah	

Hasil pengujian adalah hasil *similarity* antara dokumen pengujian dengan koleksi sampel pengujian yang dihitung dengan fungsi *cosine*, dimana *term* pada setiap dokumen sudah dihitung dengan *term weighing* berdasarkan pada persamaan (1) dan (2).

$$\textbf{\textit{Term Weighing}}: w_{in} = \frac{f_{in}}{\log k_n} \quad (1)$$

$$\textbf{Fungsi \textit{Cosine}}: S_{xy} = \frac{\sum x_i y_i}{\left(\sum x_i^2 \sum y_i^2\right)^{1/2}} \quad (2)$$

Hasil dokumen yang diujikan masih berbentuk acak atau tidak berurutan maka dilakukan proses perangkingan untuk mempermudah *user* melakukan pembacaan hasil pengujian dengan menggunakan metode K-*Nearest Neighbor* (KNN). Hasil perangkingan diatur dengan nilai K=7. Perangkingan tersebut diurutkan dari hasil *similarity* tertinggi ke *similarity* terendah.

Hasil Perangkingan *Query* 1

Dokumen : 100% Desa_di_Bangka_Belitung_Sudah _Berlistrik.pdf

Tujuan Kategori :Energi

Tabel 8. *Hasil Perangkingan Q1 dengan Koleksi Sampel*

No. Dok	Hasil Similarity	KNN (K=7)	Kategori	Ket.
1	0,664613115	1	Energi	Relevan
39	0,390884774	2	Energi	Relevan
6	0,373340606	3	Energi	Relevan
3	0,342605026	4	Energi	Relevan
12	0,321990982	5	Energi	Relevan
9	0,317435878	6	Energi	Relevan
19	0,101048814	7	Industri	Tidak Relevan

Berdasarkan jumlah kategori yang memilik kemiripan pada Tabel 8 dapat disimpulkan bahwa query 1 merupakan dokumen dengan kategori **energi.**

Hasil Perangkingan *Query* 13

Dokumen : Akhir_Tahun_industri.pdf

Tujuan Kategori : Industri

Tabel 9. *Hasil Perangkingan Q13 dengan Koleksi Sampel*

No. Dok	Hasil Similarity	KNN (K=7)	Kategori	Ket.
13	0,820028265	1	Industri	Relevan
27	0,096307746	2	Properti	Tidak Relevan
6	0,071942981	3	Energi	Tidak Relevan
1	0,071039597	4	Energi	Tidak Relevan
3	0,069753501	5	Energi	Tidak Relevan
7	0,069272688	6	Energi	Tidak Relevan
10	0,068787559	7	Energi	Tidak Relevan

Berdasarkan jumlah kategori yang memilik kemiripan pada Tabel 9 dapat disimpulkan bahwa query 13 merupakan dokumen dengan kategori **Energi.**

Hasil Perangkingan *Query* 26

Dokumen : jababeka_pp_properti.pdf

Tujuan Kategori : Properti

Berdasarkan jumlah kategori yang memilik kemiripan pada Tabel 10 dapat disimpulkan bahwa query 26 merupakan dokumen dengan kategori **properti.**

Tabel 10. *Hasil Perangkingan Q26 dengan Koleksi Sampel*

No. Dok	Hasil Similarity	KNN (K=7)	Kategori	Ket.
26	0,779272742	1	Properti	Relevan
28	0,152239982	2	Properti	Relevan
30	0,114105667	3	Properti	Relevan
24	0,103683513	4	Industri	Tidak Relevan
15	0,102644741	5	Industri	Tidak Relevan
35	0,100701763	6	Properti	Relevan
31	0,087419761	7	Properti	Relevan

Tabel 11. *Hasil Precision dan Recall*

Kategori	Query	Precision	Recall
Energi	1	6/7 = **85.71%**	6/13 = **46.15%**
Energi	2	2/7 = **28.57%**	2/13 = **15.38%**
Energi	3	6/7 = **85.71%**	6/13 = **46.15%**
Energi	4	5/7 = **71.43%**	5/13 = **38.46%**
Industri	13	1/7 = **14.29%**	1/13 = **7.69%**
Industri	14	5/7 = **71.43%**	5/13 = **38.46%**
Industri	15	6/7 = **85.71%**	6/13 =**46.15%**
Industri	16	3/7 = **42.86%**	3/13 = **23.08%**
Properti	26	5/7 = **71.43%**	5/13 = **38.46%**
Properti	27	3/7 = **42.86%**	3/13 = **23.08%**
Properti	28	4/7 = **57.14%**	4/13 = **30.77%**
Properti	29	5/7 = **71.43%**	5/13 = **38.46%**
Properti	30	6/7 = **85.71%**	6/13 = **46.15%**
Total		**814.28**	**438.44**
Mean		**62.64**	**33.73**

Kinerja model klasifikasi dievaluasi berdasarkan kemampuan prediksinya digunakan perhitungan *precision dan recall.* Pengujian ketepatan *(precision)* adalah perbandingan antara jumlah relevan yang didapatkan sistem dengan jumlah dokumen seluruh dokumen yang terambil oleh sistem baik relevan maupun tidak. Nilai *precision* diperoleh dari hasil bagi antara jumlah dokumen yang relevan dengan jumlah semua dokumen yang relevan dan tidak relevan pada tabel perangkingan. Pengujian kelengkapan *(recall)* adalah perbandingan antara jumlah dokumen relevan yang ditetapkan sistem dengan jumlah seluruh dokumen relevan yang ada dalam koleksi dokumen (terambil ataupun tidak terambil sistem). Nilai *recall* diperoleh dari hasil bagi antara jumlah dokumen yang relevan pada tabel perangkingan dengan jumlah dokumen pada masing-masing kategori. Dalam penelitian ini, jumlah masing-masing kategori adalah 13 dokumen teks. Dari hasil pengujian diatas, kategori hasil prediksi ditabulasikan dengan kategori aktual menghasilkan Tabel 11 (jumlah query yang ditampilkan pada jurnal hanya berupa sampling dari keseluruhan query untuk meringkas penulisan jurnal ini).

Berdasarkan Tabel 11 menunjukkan bahwa hasil pengukuran *precision* tidak mencapai 65% dan *recall*tidak mencapai 35% namun dapat dikatakan metode ini cukup mampu mengklasifikasikan dokumen ke dalam kategori yang tepat.

3. Kesimpulan

Penelitian ini bertujuan untuk mengetahui efektifitas proses pengklasifikasian dokumen teks berbahasa Indonesia dengan menerapkan algoritma *stemming* Nazief Adriani dengan *vector space model* berdasarkan pembobotan frekuensi kata berbanding logaritma jumlah kata dengan similarity fungsi *cosine* pada dokumen .pdf. Aplikasi yang dibuat dapat mengklasifikasikan dokumen teks berbahasa Indonesia dengan fungsi pembobotan frekuensi kata berbanding logaritma jumlah kata dengan similarity fungsi *cosine*.
Pengukuran tingkat efektifitas pengklafisikasian menggunakan metode pengukuran *precision* dan *recall*.
Dari hasil pengujian dalam penelitian ini yang menguji proses klasifikasi dengan membandingkan dokumen dengan koleksi sampel pengujian diketahui bahwa penelitian ini dikatakan cukup mampu mengkategorikan dokumen uji pada kategori yang tepat

Daftar Pustaka

[1] R. Mooney, *Intelligent Information Retrieval and Web Search*. Austin: Texas University, 2001.
[2] G. Attardi, *Text Categorization*. Roma: Pisa University, 2004.
[3] A. Zafikri, *Implementasi Metode Term Frequency Inverse Document Frequency (TF-IDF) Pada Sistem Temu Kembali Informasi*. Universitas Sumatera Utara, 2008.
[4] A. K. Mann, "Survey Paper on Clustering Techniques," *Int. J. Sci. Eng. Technol. Res.*, vol. 2, no. 4, 2013.
[5] W. Pangestu, J. A.P, S. A., and A. R., "Klasifikasi Kategori Dokumen Berbahasa Indonesia dengan Metode Kategorisasi Multi Label Berbasis Domain Specific Ontology," *J. Ilm. Teknol. Inf. Terap.*, vol. 11, no. 2–15, 2016.
[6] B. Widodo, *Artificial Intelligence*. Yogyakarta: Andi, 2014.
[7] G. Karyono and F. S. Utomo, "Temu Balik Informasi Pada Dokumen Teks Berbahasa Indonesia dengan Metode Vector Space Retrieval Model," *Semin. Nas. Teknol. Inf. Komun. Terap.*, pp. 282–289, 2012.
[8] T. Noreault, M. Mcgill, and M. B. Koli, "A Performance Evaluation of Similarity Measures , Document Term Weighting Schemes and Representations in a Boolean Environment," no. 1976, 1980.