Penerapan Algoritma K-Nearest Neighbor pada *Information Retrieval* dalam Penentuan Topik Referensi Tugas Akhir

Implementation of K-Nearest Neighbor on Information Retrieval to Determine Topic of Thesis Report

Ramadhan Rakhmat Sani[1], Junta Zeniarja[2], Ardytha Luthfiarta[3]
[1,2,3]Teknik Informatika S1, Fakultas Ilmu Komputer, Universitas Dian Nuswantoro
Jalan Imam Bonjol No. 207, Semarang 50131, Jawa Tengah, Indonesia
Email : [1]ramadhan_rs@dsn.dinus.ac.id, [2]junta@dsn.dinus.ac.id,
[3]ardytha.luthfiarta@dsn.dinus.ac.id

Abstrak

Perpustakan sebagai bagian penting universitas, berperan sebagai sumber pustaka dan referensi laporan Tugas Akhir(TA). Semua koleksi tersebut ditempatkan dalam ruangan tertentu sesuai dengan kategorinya dan data tersebut sudah terkomputerisasi dengan baik di dalam server database. Permasalahan penelitian adalah pencarian dan pengidentifikasian laporan Tugas Akhir berdasarkan topik yang diangkat. Tujuan penelitian ini mendapatkan akurasi klasifikasi yang baik menggunakan algoritma K-Nearest Neighbor (KNN) dan menerapkannya kedalam sebuah program. Sehingga klasifikasi topik TA dapat menjadi solusi untuk menyelesaikan permasalahan tersebut.

Kata kunci — Information Retrieval, K-Nearest Neighbor, Text Mining, Klasifikasi

Abstract

The library is an important part of the university, serve as a source literature and reference Thesis report. All collections are placed in a specific rooms according to its category and the data is already computerized properly in database server. The research problem lies in finding and identifying the Thesis report based on the topics raised. The purpose of this research to get a good classification accuracy using K-Nearest Neighbor (KNN) algorithms and apply it into an assistive program. So that the classification of final project topic can be a solution to solve the problem.

Keywords — Information Retrieval, K-Nearest Neighbor, Text Mining, Classification

1. PENDAHULUAN

Berkembangnya volume informasi dalam berbagai bentuk dokumen sebagai akibat dari majunya teknologi digital saat ini. Diperkirakan lebih dari 80% dokumen digital merupakan data bertipe teks [1]. Sehingga akan memaksa munculnya disiplin baru yaitu *text mining* dimana berperan untuk menganalisis teks dari berlebihnya informasi teks yang tidak tersrtuktur. Cara kerjanya dimulai dari informasi yang digali di suatu teks yang tidak terstruktur akan dicoba untuk menjumpai pola-polanya [1]. Beberapa penelitian yang berkaitan dengan teks mining pernah dilakukan yaitu [2] mengklasifikasikan dokumen berita berbahasa Indonesia menggunakan algoritma single pass clustring dengan menggunakan sampel berita dari media massa berbasis web. Hasil yang didapat dari pengujian dengan pemilihan nilai threshold yang tepat akan

meningkatkan kualitas information retrieval dengan tingkat recall 79% dan precision 88%. Adapun [3] mengklasifikasian dokumen berita berbahasa indonesia menggunakan *Latent Semantic Indexing* (LSI) dan *Support Vector Machine* (SVM) dengan pendekatan *information retrieval* dan *machine learning*. Hasil akurasi yang didapat dengan metode LSI-SVM secara keseluruhan diatas 80% sedangkan akurasi yang diperoleh dengan menggunakan metode SVM murni secara keseluruhan diatas 70%. Sedangkan waktu pemrosesan yang digunakan baik pada saat training maupun testing, metode LSI-SVM lebih cepat dibandingkan dengan SVM murni.

Pada tahun 2009 [4] dokumen teks Berbahasa Indonesia diklasifikasikan dengan menggunakan *Naïve Bayes*. Dari data sampel dokumen teks yang diambil pada media massa menunjukkan bahwa metode tersebut efektif dalam mengklasifikasikan dokumen. Dengan porsi klasifikasi dokumen teks berbahasa Indonesia dengan menggunakan *Naïve Bayes*. Ada pula [5] yang mengklasifikasikan emosi berdasarkan teks bahasa Indonesia menggunakan metode *Naïve Bayes* dan *Naïve Bayes Multinominal*. Dari hasil percobaan yang dilakukan menghasilkan tingkat akurasi sebesar 61.57% dengan rasio data 0.6 dari metode *Naïve Bayes Multinominal* yang mampu mengenali dokumen dengan perlakuan berbeda pada preprosesing data. Sedangkan tahun 2012 [6] program peramban ontologi untuk klasifikasi dokumen dalam berita teks bahasa Indonesia dibuat dengan objek artikel berita berbahasa Indonesia dari situs http://www.google.com. Dengan salah satu metode di dalam data mining ini menghasilkan klasifikasi dokumen dari hasil unduhan dari web tersebut dapat lebih terstruktur sehingga relevan dan cepat dalam mendapatkan informasi. *Naïve Bayes Classifier* juga digunakan [7] dalam klasifikasi text bahasa Indonesia. Dengan hasil akurasi terbaik dari data uji yang bersumber dari situs web dengan data latih yang besar menghasilkan akurasi lebih dari 87% dan berjalan baik untuk data latih lebih dari 150 dokumen. Dengan metode yang sama [1] mengelompokkan teks berita dan abstrak akademis. Akurasi yang dicapai untuk dokumen berita sebesar 91%, sedangkan untuk dokumen akademik sebesar 82% dengan 450 dokumen abstrak akademik dan 1000 dokumen berita. Kelemahan dari metode ini dalam pengasumsian yang sulit dipenuhi, yaitu independensi fitur kata. Adapun penggunaan algoritma *K-Nearest Neighbor* [8] yang digunakan untuk mengklasifikasi data hasil kelapa sawit menghasilkan 6 *cluster* berdasarkan kesamaan hasil produksi. Sehingga pada produksi mendatang dapat diperkirakan dengan tepat.

Dengan masalah yang sama pengolahan informasi digital juga dapat dimanfaatkan oleh perpustakaan. Dimana perpustakaan merupakan bagian terpenting pada universitas dalam menyediakan buku-buku referensi untuk topik tugas akhir. Sering terjadi kesulitan ketika perpustakaan harus mengenali buku-buku referensi tersebut sesuai dengan topik tugas akhir dikarenakan melimpahnya informasi yang tersimpan [1]. Seringkali topik suatu buku tugas akhir dijadikan gambaran umum mengenai isi suatu buku, padahal isi dari buku tugas akhir tesebut dapat jadi menjelaskan hal yang lainnya.

Pendekatan *supervised learning* dengan klasifikasi atau kategorisasi teks menjadi penting untuk menjawab permasalahan tersebut dan mempunyai banyak cara pendekatannya seperti berbasis numeris, misalnya pendekatan probabilistik, *Artificial Neural Network*, *Support Vector Machine*, KNN, dan juga berbasis non numeris salah satunya *Decision Tree*. Kelebihan yang didapat dalam pendekatan berbasis numeris untuk KNN diantranya, cepat, berakurasi tinggi dan sederhana [9]. Dalam metode KNN penggunaan atribut kata yang hadir pada suatu dokumen menjadi dasar klasifikasi untuk penggolongan atau klasifikasi teks. Dari penjelasan tersebut peneliti ingin melakukan pengklasifikasian dengan menggunakan metode KNN untuk mengklasifikasikan buku-

buku laporan tugas akhir berdasarkan topik yang dibahas dengan memakai informasi pada buku berupa abstrak. Meskipun dalam dokumen asumsi kemandirian antar kata tidak sepenuhnya bisa dipenuhi, tetapi klasifikasi relatif sangat bagus dalam kinerjanya.

2. METODE PENELITIAN

2.1. *Information Retrival* (IR)

Ilmu yang mempelajari tentang konsep, prosedur-prosedur dan metode-metode dalam melakukan proses mencari dan mendapatkan informasi relevan yang tersimpan. Data berupa teks, tabel, gambar (image), video, audio merupakan jenis data yang sering dipakai dalam proses pencarian. Salah satunya dengan mengurangi dokumen pencarian yang tidak relevean atau meretrieve dokumen yang relevan yang bertujuan untuk memenuhi informasi pengguna sebagai acuan dalam melakukan panggilan *(searching),* index *(indexing),* pemanggilan data kembali *(recalling)* [2].

2.2. Nearest Neighbor

Nearest Neighbor (NN) merupakan salah satu metode dalam pengklasifikasian yang paling popular dan termasuk dalam klasifikasi yang *lazy learner* karena menunda proses pelatihannya sampai ada data uji yang ingin diketahui label kelasnya, maka metode baru akan menjalankan algoritmanya [10]. Prinsip sederhana dari algoritma *Nearest Neighbor* dengan melakukan klasifikasi berdasarkan kemiripan suatu data dengan data yang lain. Kemiripan antara data uji terhadap data latih semakin dekat lokasi data latih terhadap data uji, maka bisa dikatakan bahwa data latih tersebut yang lebih dipandang mirip oleh data uji. Semakin dekat maka semakin mirip, yang berarti juga semakin kecil jarak maka semakin mirip [11]. Dengan kata lain, semain kecil nilai ketidakmiripan jarak maka semakin miriplah data uji terhadap sejumlah K tetangga data latih yang terdekat.

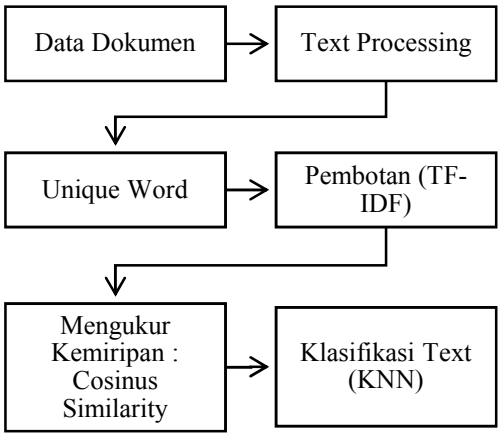
2.3. K-Nearest Neighbor (KNN)

Merupakan salah satu metode berbasis NN yang paling tua dan populer di dalam melakukan pengkategorian teks [12][13][14]. Dalam penentuan prediksi label kelas pada data uji ditentukan dengan nilai k yang menyatakan jumlah tetangga terdekat [15]. Dari k tetangga terdekat yang terpilih dilakukan voting dengan memilih kelas yang jumlahnya paling banyak sebagai label kelas hasil prediksi pada data uji [12]. Klasifikasi dianggap sebagai metode terbaik dalam preses ketika data latih yang berjarak paling dekat dengan objek [16]. Cara kerja dari KNN perlu adanya penetuan inputan berupa data latih, data uji dan nilai k. Kemudian mengurutkan data latih berdasarkan kedekatan jaraknya berdasarkan hitungan dari jarak data yang diuji dengan data latih. Setelah itu diambil dari k data latih teratas untuk menentukan kelas klasifikasi untuk kelas yang dominan dari k data latih yang diambil. Dekat atau jauhnya tetangga biasanya dihitung dari Euclidean Distance yang direpresentasikan dengan rumus sebagai berikut:

$$D(a,b) = \sqrt{\sum_{k=1}^{d}(a_k - b_k)^2} \quad (1)$$

Penelitian ini bertujuan untuk mengetahui pengelompokan referensi topik tugas akhir berdasarkan hasil klasifikasi menggunakan informasi dari buku berupa abstrak

dengan algoritma KNN. Pada gambar 1 memberikan gambaran arsitekstur klasifikasi dan pengelompokan abstrak tugas akhir :



Gambar 1 Arsitektur Klasifikasi Teks

Dimana untuk hasil dari kinerja algoritma ini diperoleh bedasarkan dari penentuan nilai K dalam penetapan jumlah kemiripan dari apa yang ingin dikategorisasikan.

2.4. Jenis Penelitian

Ada empat metode penelitian yang umum digunakan, penelitian tindakan, eksperimen, studi kasus, dan survei [17]. Penelitian eksperimen terdiri dari dua jenis, percobaan mutlak dan komparatif. Penelitian eksperimental umumnya dilakukan dalam pengembangan, evaluasi, dan pemecahan masalah proyek.

Penelitian ini menggunakan metode eksperimental dengan melakukan percobaan pengujian klasifikasian teks dan pengukuran akurasi dengan menggunakan perhitungan serta menguji proses metode dengan program *soft computing*. Implementasi dari KNN untuk klasifikasi beberapa sampel topik tugas akhir mahasiswa menggunakan *soft computing* yang rencana akan di bangun dengan pemrograman web (*PHP&MySQL).*

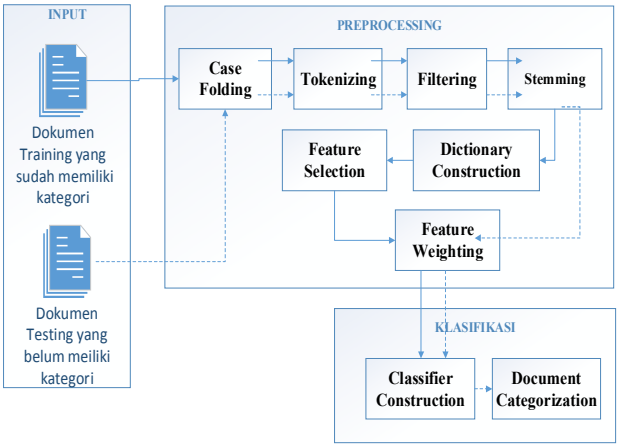
2.5. Pengumpulan Data

Data yang digunakan pada penelitian ini berupa data abstrak tugas akhir mahasiswa yang didapat dari beberapa sumber. Dokumen abstrak tersebut berjumlah 3 data dan dibagi menjadi 3 kategori yaitu abstrak *Information Retrieval*, abstrak Pengolahan Citra Digital, abstrak Jaringan dimana masing-masing kategori berjumlah 3 dokumen abstrak. Dari data dokumen tersebut 9 data abstrak dijadikan sebagai data training dan 1 data abstrak dijadikan sebagai data testing.

2.6. Desain Penelitian

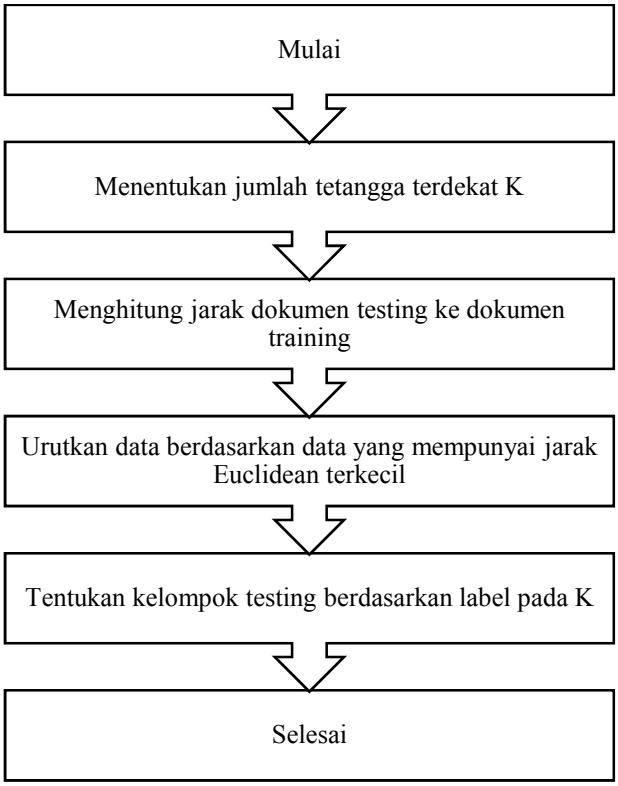
Pada gambar 2. menjelaskan tentang desain penelitian yang dilakukan, dimana ada dua proses utama yaitu *preprocessing data* dan proses klasifikasi. Dokumen *training* yang sudah memiliki kategori dan dokumen testing yang belum memiliki kategori merupakan dua jenis inputan dari sistem ini. Keduanya melalui proses

preprocessing antara lain *case folding*, *tokenizing*, *filtering*, *stemming*, *dictionary construction*, *feature selection*, dan *feature weighting*.



Gambar 2 Desain Penelitian

Pada fase *feature weighting*, *term-term* dari dokumen training dan dokumen testing dihitung bobotnya. Kemudian membangun classifier dari bobot dokumen training dan dokumen testing pada proses klasifikasi. Algoritma KNN digunakan untuk membentuk *Classifier*. *Cosine similarity* digunakan untuk proses menghitung kemiripan antar dokumen. Kelas atau pengelompokan dari dokumen testing merupakan hasil dari proses *classifier*



Gambar 3 Cara Kerja K-Nearest Neighbor

Pada gambar 3 menjelaskan mengenai cara kerja algoritma KNN. Dimulai dengan menentukan jumlah tetangga terdekat dari nilai K yang sudah ditentukan. Kemudian menghitung jarak dokumen *testing* ke dokumen *training*. Data diurutkan berdasarkan data yang mempunyai jarak *euclidean* terkecil yang sebelumnya sudah dihitung terlebih dahulu menggunakan rumus jarak *euclidean*. Setelah itu, menentukan kelompok dokumen testing berdasarkan label pada nilai K.

3. HASIL DAN PEMBAHASAN

Dari penelitian ini mempunyai 2 tahapan yaitu tahapan pertama meliputi pembelajaran atau training yaitu tahap pengklasifikasian terhadap abstrak yang sudah di ketahui kategorinya. Pada pengujian tahap testing hal-hal yang dilakukan adalah dengan melakukan klasifikasi abstrak tugas akhir yang belum diketahui kategorinya. Hasil Penelitian yang dilakukan sebagai berikut:

- Dokumen Teks

Dokumen Teks yang digunakan dibagi menjadi dua bagian penting yaitu *dokumen training* sebanyak 9 dokumen dan *dokumen testing (dokumen uji)* yang berjumlah 1 dokumen abstrak. Dimana label atau kategori yang digunakan terdapat 3 yaitu IMAGE, IR dan JARKOM.

- *Preprocessing*

Dokumen training dan uji dilakukan tokenizing, yaitu pemecahan dokumen menjadi frase atau term (kata), sesuai dengan dokumen masing – masing. Setelah dilakukan tokenizing, tahapan selanjutnya adalah di lakukan *stopword removal* atau penghilangan kata yang dirasa tidak diperlukan. Kemudian tahapan yang terakhir adalah di stemming, yaitu penghilangan imbuhan sehingga menjadi kata dasar. Pada tabel 1 menunjukkan contoh hasil tokenization dokumen dari 9 dokumen abstrak, yang sudah dipisahkan sehingga terbentuk hasil tiap frasenya dalam setiap dokumen tersebut.

Tabel 1 Contoh Hasil Tokenization Dokumen

D1	**D2**	**D3**	**D4**	**D5**	**D6**	**…**	**D9**
IR	**Image**	**Jarkom**	**Image**	**Image**	**Jarkom**	**…**	**IR**
pemilahan	pengolahan	pengamanan	kompresi	purwarupa	simulasi	…	clustering
Artikel	citra	pesan	citra	penghitung	routing	…	berita
Berita	digital	text	berwarna	Jasa	protokol	…	berbahasa
Dengan	dan	dengan	menggunakan	pemakaian	pada	…	indonesia
Text	analisis	metode	metode	Printer	jaringan	…	volume
Mining	kuantitatif	kriptografi	pohon	Menggunakan	sensor	…	berita
Seiring	dalam	advance	biner	penerapan	nirkabel	…	elektronik
pesatnya	karakterisasi	encryption	huffman	pengolahan	dengan	…	berbahasa
perkembangan	citra	standart	makalah	Citra	Menggunakan	…	indonesia
internet	mikroskopik	dan	ini	Digital	metode	…	yang
…	…	…	…	…	…	…	…
pula	mikroskop	metode	algoritma	purwarupa	jaringan	…	merupakan
Bermunculan	elektron	echo	kompresi	penghitung	sensor	…	sumber
situs	walaupun	data	citra	Jasa	nirkabel	…	informasi

Kemudian menentukan bobot untuk setiap term dari 9 dokumen abstrak yang terlibat dengan cara menghitung TF – IDF (Term Frequency – Inverse Document Frequency) terlebih dahulu. Pada tabel 2 pembobotannya dihitung dengan cara nilai TF dikalikan dengan nilai IDF. Menghitung kemiripan vektor dokumen Dtest dengan setiap dokumen yang sudah terklasifikasi (D1, D2, D3, D4, D5, D6, D7, D8, D9). Kemiripan antar dokumen menggunakan rumus Cosine Similarity sebagai berikut:

$$\cos(\Theta_{ij}) = \frac{\sum_k (d_{ik} d_{jk})}{\sqrt{\sum_k d_{ik}^2} \sqrt{\sum_k d_{jk}^2}} \quad (2)$$

Hasil perhitungan skalar antara Dtest dengan kesembilan dokumen yang telah terklasifikasi ditunjukkan pada tabel 3. Hasil perkalian dari setiap dokumen dengan Dtest dijumlahkan.

Pada tabel 4 menampilkan hasil dari menghitung panjang setiap dokumen, termasuk Dtest dengan cara mengkuadratkan bobot setiap term dalam tiap dokumen, kemudian jumlahkan nilai kuadrat tersebut dan diakar dengan menerapkan rumus *cosine similarity*, untuk menghitung kemiripan Dtest dengan D1,D2,D3,D4,D5,D6,D7,D8 dan D9. Pada langkah terakhir dengan mengurutkan hasil perhitungan kemiripan.

Tabel 2 Tabel Contoh Hasil Perhitungan TF-IDF

Wd9 x di (Q x di)							
D1	**D2**	**D3**	**D4**	**D5**	**D6**	**…**	**D9**
0	0	0	0	0	0	..	0
0	0	0	0	0	0	..	0
0	0	0	0	0	0	..	0
0	0	0	0	0	0	..	0
0	0	0	0	0	0	..	0
0	0	0	0	0	0	..	0
. . . .	. . . .	. . . .	. . . .	. . . .	. . . .	..	. . . .
0,00000	0,00014	0,00014	0,00000	0,00014	0,00014	..	0,00014
0,00014	0,00000	0,00014	0,00014	0,00000	0,00014	..	0,00014
0,00001	0,00001	0,00001	0,00000	0,00001	0,00001	..	0,00001
0,00001	0,00001	0,00001	0,00001	0,00000	0,00001	..	0,00001
1,30941	**0,37276**	**0,45382**	**0,20183**	**1,55313**	**0,45117**	..	**1,09090**

Tabel 3 Contoh Hasil Perkalian Skalar D9 dengan (D1,D2,D3,D4,D5,D6,D7,D8)

TF-IDF								
D1	**D2**	**D3**	**D4**	**D5**	**D6**	**D7**	**…**	**D9**
0	0	0,9542	0	0	0	0	...	0
0	0	0,9542	0	0	0	0	...	0
0	0	0	0	0	0	0	...	0,9542
0	0	0	0	0	0	0,9542	...	0
0	0	0,9542	0	0	0	0	...	0
0	0	0	0	0	0	0,9542	...	0

0	0,954	0	0	0	0	0	...	0
0	0	0	0	0	0	0	...	0
0	0	0,9542	0	0	0	0	...	0
...	...	...	...	...	...	...	...	...
0	0	0,9542	0	0	0	0	...	0
0	0	0	0	0,9542	0	0	...	0
0,9542	0	0	0	0	0	0	...	0
0	0	0	0	0	0,9542	0	...	0

Tabel 4 Contoh Hasil Menghitung Panjang Vektor Setiap Dokumen

	Panjang Vektor								
	Dtest atau Q	**D1**	**D2**	**D3**	**D4**	**D5**	**D6**	**…**	**D9**
	0,0000	0,0000	0,0000	0,8292	0,0000	0,0000	0,0000	…	0,0000
	0,0000	0,0000	0,0000	0,8292	0,0000	0,0000	0,0000	…	0,0000
	0,8292	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	…	0,0000
	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	…	0,0000
	0,0000	0,0000	0,0000	0,8292	0,0000	0,0000	0,0000	…	0,0000
	. . .	. . .	. . .	. . .	. . .	. . .	. . .	…	. . .
	0,0001	0,0001	0,0000	0,0001	0,0001	0,0000	0,0001	…	0,0001
	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	…	0,0000
	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	…	0,0000
TOTAL	**44,9666**	**42,8367**	**55,4940**	**62,0197**	**6,6878**	**36,6562**	**37,1703**	…	**23,5323**
AKAR	**6,7057**	**6,5450**	**7,4494**	**7,8753**	**2,5861**	**6,0544**	**6,0967**	…	**4,8510**

Tabel 5 Contoh menghitung kemiripan dokumen uji dengan dokumen training

Cos (Dtest, Di)	**Hasil**	**Urutan**	**Kelas**
Cos (Dtest, D1)	0,029835	3	IR
Cos (Dtest, D2)	0,007462	8	IMAGE
…	…	..	...
Cos (DTest, D9)	0,033536	2	IR

Pada tabel 6 menunjukan pengambilan sebanyak k (k=3) yang paling tinggi tingkat kemiripannya dengan Dtest dan menentukan kelas dari Dtest, hasilnya : D5, D8, dan D1.

Tabel 6 Contoh urutan hasil perhitungan

Dok	**D1**	**D2**	**D3**	**D4**	**D5**	**D6**	**D7**	**D8**
Ranking	**3**	8	7	4	**1**	5	6	**2**

Sehingga dapat dilihat kembali bahwa D5 merupakan kategori **IMAGE**, sedangkan D8 dan D1 merupakan kategori **IR**. Karena yang terbanyak dari **IR**, Oleh karena itu dapat disimpulkan bahwa Dtest atau dokomen testing yang digunakan termasuk kedalam kategori **IR**.

Tabel 7 Confusion Matrix

		Aktual		
		Jarkom	**IR**	**Image**
Predicted	**Jarkom**	0	1	1
	IR	0	4	0
	Image	0	0	4

Dari pengujian menggunakan 10 dokumen testing yang berupa 2 dokumen untuk kelas jarkom, 4 dokument untuk kelas IR dan 4 dokumen untuk kelas IR diperoleh hasil yang ditransformasikan dalam tabel confusion matrix yang ditampilkan pada tabel 7.

$$Akurasi = \frac{0+4+4}{0+1+1+0+4+0+0+0+0} = \frac{8}{10} x\ 100\% = 80\ \%$$

Hasil ketepatan prediksi untuk kelas IR dan kelas Image sebanyak 4 kali. Sedangkan untuk ketepatan prediksi pada kelas jarkom belum semunya tepat. Sehingga perlu adanya penambahan data training data abstrak untuk kelas jarkom. Dari confusion matrix yang didapat ditentukan untuk tingkat akurasinya sebesar 80% berdasarkan perhitungan total prediksi yang benar dibagi jumlah total prediksi benar dan salah.

4. KESIMPULAN

Dari hasil dan pembahasan yang sudah dilakukan pada penelitian ini, dapat diambil kesimpulan bahwa dalam menentukan kategori teks dari sampel dokumen abstrak tugas akhir mahasiswa menggunakan algoritma *K-Nearest Neighbor* sudah terbukti dengan baik, dimana dilakukan dengan cara melakukan perhitungan sesuai dengan kinerja algoritma *K-Nearest Neighbor* untuk menghasilkan model sehingga masuk kedalam kategori klasifikasi yang sudah baik. Dibuktikan dengan hasil akurasi menggunakan diagram confussion matrix dari pengujian model yang dapat menentukan topik tugas akhir dengan akurasi sebesar 80%. Untuk mendapatkan kinerja dan performa yang lebih baik, dapat dilakukan dengan cara penambahan dokumen *training* untuk kelas jarkom, image dan IR, sehingga diharapkan nantinya dapat meningkatkan akurasi menjadi lebih tinggi.

5. SARAN

Penelitian ini nantinya akan terus berlanjut ke jenjang yang lebih baik lagi yaitu dengan mengembangkan model atau algoritma yang digunakan. Selain itu juga dengan memperkaya jumlah dokumen teks yang digunakan dan mengkombinasikannya dengan algoritma klasifikasi beserta beberapa algoritma berbasis pencarian lainnya. Dan pada akhirnya akan dilakukan perancangan software atau perangkat lunak yang nantinya dapat digunakan secara mudah oleh pihak perpustakaan.

DAFTAR PUSTAKA

[1] A. Hamzah, "Klasifikasi teks dengan naïve bayes classifier (nbc) untuk pengelompokan teks berita dan abstract akademis," in *Prosiding Seminar*

Nasional Aplikasi Sains & Teknologi (SNAST) Periode III, 2012, no. 2011, pp. 269–277.

[2] A. Z. Arifin and A. N. Novan, "Klasifikasi Dokumen Berita Kejadian Berbahasa Indonesia dengan Algoritma Single Pass Clustering," *Pros. Semin. Intell. Technol. its Appl. (SITIA), Tek. Elektro, Inst. Teknol. Sepuluh Nop. Surabaya*, 2002.

[3] L. Noviani, A. A. Suryani, and A. P. Kurniati, "PENGKLASIFIKASIAN DOKUMEN BERITA BERBAHASA INDONESIA MENGGUNAKAN LATENT SEMANTIC INDEXING (LSI) DAN SUPPORT VECTOR MACHINE (SVM)," 2008.

[4] J. Samodra, S. Sumpeno, and M. Hariadi, "Klasifikasi Dokumen Teks Berbahasa Indonesia dengan Menggunakan Naïve Bayes," *Semin. Nas. Electr. Informatic, IT's Educ.*, pp. 1–4, 2009.

[5] I. Destuardi and S. Sumpeno, "Klasifikasi Emosi Untuk Teks Bahasa Indonesia Menggunakan Metode Naive Bayes," *Semin. Nas. Pascasarj. Inst. Teknol. Sepuluh Nop.*, no. c, 2009.

[6] H. Februariyanti and E. Zuliarso, "Klasifikasi Dokumen Berita Teks Bahasa Indonesia menggunakan Ontologi," *J. Teknol. Inf. Din.*, vol. 17, no. 1, pp. 14–23, 2012.

[7] C. Darujati and A. B. Gumelar, "Pemanfaatan Teknik Supervised Untuk Klasifikasi Teks Bahasa Indonesia," *J. LINK*, vol. 16, no. 1, pp. 1–8, 2012.

[8] N. Krisandi, Helmi, and B. Prihandono, "Algoritma K-Nearest Neighbor dalam Klasifikasi Data Hasil Produksi Kelapa Sawit pada PT. Minamas Kecamatan Parindu," *Bul. Ilm. Math. Stat. dan Ter.*, vol. 02, no. 1, pp. 33–38, 2013.

[9] C. Z. Charu C. Aggarwal, *Mining Text Data*, 1st ed. Springer-Verlag New York, 2012.

[10] H. E. Huppert, J. C. Tannehill, and F. Mechanics, *The Top Ten Algoritm in Data MIning*, no. 1961. 2009.

[11] O. Kwon and J. Lee, "Text categorization based on k-nearest neighbor approach for Web site classification," *Inf. Process. Manag.*, vol. 39, no. 1, pp. 25–44, 2003.

[12] G. Toker and Ö. Kırmemiş, "TEXT CATEGORIZATION USING k-NEAREST NEIGHBOR CLASSIFICATION."

[13] X. Yan, W. Li, W. Chen, W. Luo, C. Zhang, and Q. Wu, "Weighted K-Nearest Neighbor Classification Algorithm Based on Genetic Algorithm," *TELKOMNIKA*, vol. 11, no. 10, pp. 6173–6178, 2013.

[14] M. Yao and B. Vocational, "Research on Learning Evidence Improvement for k NN Based Classification Algorithm," *Int. J. Database Theory Appl.*, vol. 7, no. 1, pp. 103–110, 2014.

[15] A. D. Arifin, I. Arieshanti, and A. Z. Arifin, "Implementasi Algoritma K-Nearest Neighbour Yang Berdasarkan One Pass Clustering Untuk Kategorisasi Teks," pp. 1–7.

[16] D. Santoso, D. E. Ratnawati, and Indriati, "Perbandingan Kinerja Metode Naive

Bayes, K-Nearest Neighbor, dan Metode Gabungan K-Means dan LVQ dalam Pengkategorian Buku Komputer Berbahasa Indonesia berdasarkan Judul dan Sinopsis," *Repos. J. Mhs. PTIIK UB*, vol. 4, no. 9, 2014.

[17] C. W. Dawson, *Projects in Computing and Information Systems*, vol. 2. 2009.