

**IMPLEMENTASI METODE IMPROVED K-NEAREST NEIGHBOR
DALAM PENGKLASIFIKASIAN ARTIKEL BAHASA INGGRIS**

Tugas Akhir



Oleh

**Fanny Pritami Widodo
22074208**

Program Studi Teknik Informatika Fakultas Teknik

Universitas Kristen Duta Wacana

2010

**IMPLEMENTASI METODE IMPROVED K-NEAREST NEIGHBOR DALAM
PENGKLASIFIKASIAN ARTIKEL BAHASA INGGRIS**

Tugas Akhir



Diajukan kepada Fakultas Teknik Informatika
Universitas Kristen Duta Wacana
Sebagai salah satu syarat dalam memperoleh gelar
Sarjana Komputer

Disusun Oleh :

Fanny Pritami Widodo

22074208

**Program Studi Teknik Informatika Fakultas Teknik
Universitas Kristen Duta Wacana
2010**

PERNYATAAN KEASLIAN TUGAS AKHIR

Saya menyatakan dengan sesungguhnya bahwa tugas akhir dengan judul :

**Implementasi Metode Improved k-Nearest Neighbor Dalam
Pengklasifikasian Artikel Bahasa Inggris**

Yang saya kerjakan untuk melengkapi sebagian persyaratan menjadi Sarjana Komputer pada pendidikan sarjana Program Studi Teknik Informatika, Fakultas Teknik Universitas Kristen Duta Wacana, bukan merupakan tiruan atau duplikasi dari skripsi kesarjanaan di lingkungan Universitas Kristen Duta Wacana maupun di Perguruan Tinggi atau instansi manapun, kecuali bagian yang sumber informasinya dicantumkan sebagaimana mestinya.

Jika kemudian hari didapati bahwa hasil skripsi ini adalah hasil plagiasi atau tiruan dari skripsi lain, saya bersedia dikenai sanksi yakni pencabutan gelar kesarjanaan saya.

Yogyakarta, 30 November 2010



Fanny Pritami Widodo
(22074208)

INTISARI

IMPLEMENTASI METODE IMPROVED K-NEAREST NEIGHBOR DALAM PENGKLASIFIKASIAN ARTIKEL BAHASA INGGRIS

Semakin hari semakin banyak inovasi, perkembangan, dan temuan-temuan yang terkait dengan bidang Teknologi Informasi dan Komputer. Hal ini menyebabkan semakin banyaknya artikel-artikel ilmiah yang diterbitkan dalam proses mengembangkan dan menemukan hal-hal baru dalam bidang tersebut dan semakin sulitnya untuk mengorganisasi artikel-artikel tersebut secara efisien.

Untuk mempermudah pengorganisasian artikel tersebut, penulis membangun sebuah sistem pengelompokan artikel yang dikenal dengan sistem klasifikasi. Penulis menggunakan metode klasifikasi yang mudah dan efektif yaitu *Improved k-Nearest Neighbor* yang dimodifikasi dari metode *k-Nearest Neighbor*. Modifikasi dilakukan dengan tujuan untuk meningkatkan performa dalam melakukan klasifikasi. Metode ini didasari oleh penelitian yang dilakukan oleh Muhammed Miah (2009). Dalam metode ini, tidak semua dokumen pelatihan akan dihitung kesamaannya dengan dokumen uji dan tidak menggunakan semua kata yang ada dalam dokumen pelatihan sehingga metode ini menjadi lebih efektif. Penulis menguji keakuratan yang dihasilkan oleh metode ini dan melakukan beberapa pengujian untuk mengetahui kondisi yang dapat meningkatkan keakuratan hasil dari sistem.

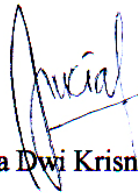
Sistem yang dibangun oleh dapat menghasilkan tingkat akurasi yang cukup tinggi dengan menggunakan nilai k sebesar 20, yaitu 83,33%. Modifikasi yang dilakukan tidak memberi manfaat yang signifikan terhadap performa sistem.

HALAMAN PERSETUJUAN

Judul : Implementasi Metode Improved k-Nearest Neighbor
Dalam Pengklasifikasian Artikel Bahasa Inggris
Nama : Fanny Pritami Widodo
NIM : 22074208
Mata Kuliah : Tugas Akhir
Kode : TI2126
Semester : Gasal
Tahun Akademik : 2010/2011

Telah diperiksa dan disetujui
Di Yogyakarta,
Pada Tanggal 30-11-2010

Dosen Pembimbing I



Lucia Dwi Krisnawati, S.S, M.A

Dosen Pembimbing II



Antonius Rachmat, S.Kom, M.Cs

HALAMAN PENGESAHAN

SKRIPSI

**IMPLEMENTASI METODE K-NEAREST NEIGHBOR DALAM
PENGKLASIFIKASIAN ARTIKEL BAHASA INGGRIS**

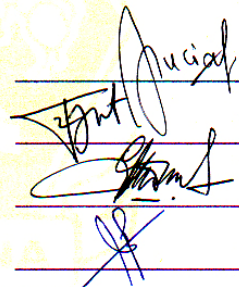
Oleh : Fanny Pritami Widodo / 22074208

Dipertahankan di depan dewan Penguji Tugas Akhir/Skripsi
Program Studi Teknik Informatika Fakultas Teknik
Universitas Kristen Duta Wacana – Yogyakarta
Dan dinyatakan diterima untuk memenuhi salah satu
Syarat memperoleh gelar
Sarjana Komputer
Pada tanggal
22 Desember 2010

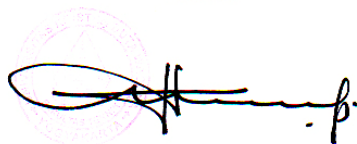
Yogyakarta, 22 Desember 2010
Mengesahkan,

Dewan Penguji :

1. Lucia Dwi Krisnawati, S.S, M.A.
2. Antonius Rachmat C,S.Kom., M.Cs.
3. Drs. R. Gunawan Santosa, M.Si.
4. Drs. Jong Jek Siang, M.Sc.



Dekan



(Ir. Henry Feriadi, Ms.Sc., Ph.D)

Ketua Program Studi



(Restyandito, S.Kom, M.SIS.)

UCAPAN TERIMA KASIH

Puji dan syukur penulis panjatkan ke hadirat Tuhan Yang Maha Esa yang telah melimpahkan rahmat dan anugerah, sehingga penulis dapat menyelesaikan Tugas Akhir dengan judul Implementasi Metode Improved k-Nearest Neighbor Dalam Pengklasifikasian Artikel Bahasa Inggris.

Penulisan laporan ini merupakan kelengkapan dan pemenuhan dari salah satu syarat dalam memperoleh gelar Sarjana Komputer. Selain itu bertujuan melatih mahasiswa untuk dapat menghasilkan suatu karya yang dapat dipertanggungjawabkan secara ilmiah, sehingga dapat bermanfaat bagi penggunaannya.

Dalam menyelesaikan pembuatan program dan laporan Tugas Akhir ini, penulis telah banyak menerima bimbingan, saran, dan masukan dari berbagai pihak, baik secara langsung maupun secara tidak langsung. Untuk itu dengan segala kerendahan hati, pada kesempatan ini penulis menyampaikan ucapan terimakasih kepada :

1. Ibu Lucia Dwi Krisnawati, S.S, M.A. selaku pembimbing I yang telah memberukan bimbingannya dengan sabar dan baik kepada penulis, juga kepada
2. Bpk Antonius Rachmat, S.Kom, M.Cs. selaku dosen pembimbing II atas bimbingannya, petunjuk, dan masukan yang diberikan selama pengerjaan tugas akhir ini sejak awal hingga akhir, juga kepada
3. Deddy Wijaya Suliantoro, S.Kom. atas bimbingan dan masukannya yang diberikan saat akan mengajukan Tugas Akhir ini.
4. Keluarga tercinta yang member dukungan dan semangat.

5. Evarisma Wulandari, Danny Sebastian, Yohanes Septian, Shinta Marzelina, Lidya Agnes, dan Albert Michael yang selalu memberikan dukungan dan semangat.
6. Sahabat dan teman-teman yang telah memberikan masukan dan semangat.
7. Pihak lain yang tidak dapat penulis sebutkan satu per satu, sehingga Tugas Akhir ini dapat terselesaikan dengan baik.

Penulis menyadari bahwa program dan laporan Tugas Akhir ini masih jauh dari sempurna. Oleh karena itu, penulis sangat mengharapkan kritik dan saran yang membangun dari pembaca sekalian. Sehingga suatu saat penulis dapat memberikan karya yang lebih baik lagi.

Akhir kata penulis ingin meminta maaf bila ada kesalahan baik dalam penyusunan laporan maupun yang pernah penulis lakukan sewaktu membuat program Tugas Akhir. Sekali lagi penulis memohon maaf yang sebesar-besarnya. Dan semoga dapat berguna bagi kita semua.

Yogyakarta,

Penulis

DAFTAR ISI

HALAMAN JUDUL.....	
PERNYATAAN KEASLIAN SKRIPSI	i
INTISARI	ii
HALAMAN PERSETUJUAN	iii
HALAMAN PENGESAHAN	iv
UCAPAN TERIMA KASIH	v
DAFTAR ISI.....	vii
DAFTAR TABEL	xi
DAFTAR GAMBAR	xiii
DAFTAR LISTING	xv
BAB 1 PENDAHULUAN	
1.1 Latar Belakang Masalah	1
1.2 Perumusan Masalah	2
1.3 Batasan Masalah	2
1.4 Hipotesis	4
1.5 Tujuan Penelitian	4
1.6 Metode Penelitian	4
1.7 Sistematika Penulisan	5
BAB 2 TINJAUAN PUSTAKA	

2.1	Tinjauan Pustaka.....	6
2.2	Landasan Teori	8
2.2.1	<i>Data Mining</i>	8
2.2.2	<i>Text Mining</i>	9
2.2.3	Klasifikasi	10
2.2.3.1	Prapemrosesan Data.....	11
2.2.3.1.1	Tokenisasi	12
2.2.3.1.2	Penghapusan <i>Stopword</i>	12
2.2.3.1.3	Perhitungan TF.....	13
2.2.3.2	Metode <i>k-Nearest Neighbor</i>	13
2.2.3.3	Metode <i>Improved k-Nearest Neighbor</i>	13
2.2.3.4	Evaluasi Sistem	13
BAB 3 ANALISIS DAN PERANCANGAN SISTEM		
3.1	Kebutuhan <i>Hardware</i> dan <i>Software</i>	23
3.1.1	Kebutuhan <i>Hardware</i>	23
3.1.2	Kebutuhan <i>Software</i>	23
3.2	Spesifikasi Sistem.....	24
3.3	Arsitektur Sistem	26
3.4	Diagram <i>Use Case</i>	27
3.5	Algoritma dan <i>Flowchart</i>	28
3.5.1	<i>Text Pre-processing</i>	28
3.5.2	<i>Update Database</i> Pelatihan.....	29
3.5.3	Hapus Artikel Pelatihan	31

3.5.4	Klasifikasi	31
3.6	Kamus Data.....	33
3.6.1	Tabel Kategori.....	33
3.6.2	Tabel Pelatihan.....	34
3.6.3	Tabel <i>Stoplist</i>	34
3.6.4	Tabel Token	35
3.6.5	Tabel Metadata.....	35
3.6.6	Tabel Metadata_uji	36
3.6.7	Tabel Token_uji	36
3.7	Diagram Skema.....	37
3.8	Rancangan Antarmuka Sistem	38
3.8.1	Rancangan Antarmuka Sistem Klasifikasi.....	38
3.8.2	Rancangan Antarmuka Daftar dan Hapus Artikel Pelatihan.....	38
3.8.3	Rancangan Antarmuka Tambah Dokumen Pelatihan	39
3.9	Rancangan Evaluasi Sistem Klasifikasi	39
3.10	Contoh Kasus Klasifikasi.....	42
BAB 4 IMPLEMENTASI DAN ANALISIS SISTEM		
4.1	Implementasi Sistem.....	48
4.1.1	Konfigurasi <i>Code Igniter</i>	48
4.1.2	Antar Muka Sistem	50
4.1.3	Pengumpulan Dokumen	53
4.1.4	Data <i>Pre-processing</i>	54
4.1.5	Klasifikasi Dokumen.....	57

4.2	Evaluasi Sistem	59
4.2.1	Evaluasi Keakuratan Sistem.....	60
4.2.2	Evaluasi Kecenderungan Nilai k Dalam Keakuratan Sistem.....	61
4.2.3	Evaluasi Pengaruh <i>Keyword</i> dan Judul Dalam Keakuratan Sistem	70
4.2.4	Evaluasi Pengaruh Penghapusan Token dengan <i>Term Frequency</i> (TF) Tinggi dan <i>Document Frequency</i> (DF) Tinggi Dalam Keakuratan Sistem.....	72
BAB 5 KESIMPULAN DAN SARAN		
5.1	Kesimpulan	79
5.2	Saran	80
DAFTAR PUSTAKA		81
LAMPIRAN.....		82

DAFTAR TABEL

Tabel 2.1	DOKUMEN PELATIHAN	17
Tabel 2.2	DOKUMEN UJI	17
Tabel 2.3	PERHITUNGAN TF DOKUMEN PELATIHAN	18
Tabel 2.4	PERHITUNGAN TF DOKUMEN UJI	18
Tabel 2.5	PERBEDAAN METODE K-NN DENGAN K-NN_I	20
Tabel 2.6	HASIL PERCOBAAN	21
Tabel 2.7	CONTOH <i>CONFUSION MATRIX</i>	22
Tabel 2.2	DOKUMEN UJI	17
Tabel 3.1	TABEL KATEGORI	34
Tabel 3.2	TABEL PELATIHAN	34
Tabel 3.3	TABEL <i>STOPLIST</i>	35
Tabel 3.4	TABEL TOKEN	35
Tabel 3.5	TABEL METADATA	36
Tabel 3.6	TABEL METADATA_UJI	36
Tabel 3.7	TABEL TOKEN_UJI	37
Tabel 3.8	DOKUMEN PELATIHAN	42
Tabel 3.9	DOKUMEN UJI	44
Tabel 4.1	<i>CONFUSION MATRIX</i> HASIL PENGUJIAN SISTEM DENGAN $k = 20$	60
Tabel 4.2	<i>CONFUSION MATRIX</i> HASIL PENGUJIAN SISTEM DENGAN $k = 15$	62
Tabel 4.3	<i>CONFUSION MATRIX</i> HASIL PENGUJIAN SISTEM DENGAN $k = 10$	63
Tabel 4.4	<i>CONFUSION MATRIX</i> HASIL PENGUJIAN SISTEM DENGAN $k = 5$	64

Tabel 4.5	<i>CONFUSION MATRIX</i> HASIL PENGUJIAN SISTEM DENGAN $k = 25$	65
Tabel 4.6	<i>CONFUSION MATRIX</i> HASIL PENGUJIAN SISTEM DENGAN $k = 30$	66
Tabel 4.7	<i>CONFUSION MATRIX</i> HASIL PENGUJIAN SISTEM DENGAN $k = 35$	67
Tabel 4.8	<i>CONFUSION MATRIX</i> HASIL PENGUJIAN SISTEM DENGAN PERTIMBANGAN DOKUMEN METADATA	71
Tabel 4.9	TABEL HASIL PENGUJIAN SISTEM DENGAN PENGHILANGAN TOKEN DENGAN TF DAN DF TINGGI	74

DAFTAR GAMBAR

Gambar 3.1	Arsitektur Sistem	26
Gambar 3.2	Diagram <i>Use Case</i>	27
Gambar 3.3	<i>Flowchart Text Pre-Processing</i>	29
Gambar 3.4	<i>Flowchart Update Database</i> Pelatihan	30
Gambar 3.5	<i>Flowchart</i> Hapus Artikel Pelatihan	31
Gambar 3.6	<i>Flowchart</i> Klasifikasi	32
Gambar 3.7	<i>Flowchart</i> Pemilihan Kategori	33
Gambar 3.8	Diagram Skema	37
Gambar 3.9	Rancangan Antarmuka Sistem Klasifikasi	38
Gambar 3.10	Rancangan Antarmuka Daftar dan Hapus Artikel Pelatihan	38
Gambar 3.11	Rancangan Antarmuka Tambah Artikel Pelatihan	39
Gambar 3.10	Rancangan Antarmuka Daftar dan Hapus Artikel Pelatihan	38
Gambar 4.1	Struktur Direktori <i>CodeIgniter</i>	49
Gambar 4.2	<i>Form</i> Klasifikasi (Pengguna)	50
Gambar 4.3	<i>Form</i> Login	51
Gambar 4.4	<i>Form</i> Klasifikasi (Admin)	51
Gambar 4.5	<i>Form</i> Daftar Pelatihan	52
Gambar 4.6	<i>Form</i> Tambah Pelatihan	52

Gambar 4.7	<i>Form</i> Ubah Artikel	53
Gambar 4.8	Grafik Hasil Pengujian Sistem dengan $k = 20$	60
Gambar 4.9	Grafik Hasil Pengujian Sistem dengan $k = 15$	62
Gambar 4.10	Grafik Hasil Pengujian Sistem dengan $k = 10$	63
Gambar 4.11	Grafik Hasil Pengujian Sistem dengan $k = 5$	64
Gambar 4.12	Grafik Hasil Pengujian Sistem dengan $k = 25$	65
Gambar 4.13	Grafik Hasil Pengujian Sistem dengan $k = 30$	66
Gambar 4.14	Grafik Hasil Pengujian Sistem dengan $k = 35$	67
Gambar 4.15	Grafik Perbandingan Nilai k Terhadap Keakuratan Sistem	68
Gambar 4.16	Grafik Perbandingan Nilai $k=1$ hingga $k=5$ Terhadap Keakuratan Sistem	69
Gambar 4.17	Grafik Hasil Pengujian Sistem dengan Pertimbangan Metadata Dokumen	71
Gambar 4.18	Nilai <i>Similarity</i> Dalam Proses Klasifikasi	76
Gambar 4.19	Nilai <i>Similarity</i> Dalam Proses Klasifikasi Korpus Muhammed Miah	78

DAFTAR LISTING

Listing 4.1	<i>Pseudocode</i> Tokenisasi	54
Listing 4.2	<i>Pseudocode</i> Perhitungan Nilai TF	54
Listing 4.3	<i>Pseudocode</i> Tokenisasi Metadata.....	55
Listing 4.4	<i>Pseudocode</i> Penghapusan <i>Stopword</i>	56
Listing 4.5	<i>Pseudocode</i> Perhitungan Perbandingan TF	56
Listing 4.6	<i>Pseudocode</i> Proses Klasifikasi	57
Listing 4.7	<i>Pseudocode</i> Proses Pemilihan Kategori	58

Bab 1

PENDAHULUAN

1.1 Latar Belakang Masalah

Semakin hari semakin banyak inovasi, perkembangan, dan temuan-temuan yang terkait dengan bidang Teknologi Informasi dan Komputer. Hal ini menyebabkan semakin banyaknya artikel-artikel ilmiah yang diterbitkan dalam proses mengembangkan dan menemukan hal-hal baru dalam bidang tersebut dan semakin sulitnya untuk mengorganisasi artikel-artikel tersebut secara efisien. Dengan mengelompokkan artikel-artikel tersebut maka akan memudahkan dalam mencari artikel yang diinginkan. Salah satu cara mendapatkan efisiensi dalam mengorganisasi artikel tersebut adalah dengan mengelompokkannya berdasarkan kategori tertentu. Pengelompokkan artikel ini dikenal dengan pengklasifikasian. Sejumlah metode klasifikasi telah banyak diterapkan untuk melakukan pengkategorian seperti *Naïve Bayes*, pohon keputusan, *k-Nearest Neighbor*, *Support Vector Machine*, dan sebagainya.

Salah satu metode klasifikasi yang mudah dan efektif adalah metode *k-Nearest Neighbor*. Dalam prosesnya, *k-Nearest Neighbor* memeriksa semua kata dalam dokumen pelatihan untuk menghitung kesamaannya dengan dokumen yang akan diklasifikasikan (dokumen uji). Hal ini mengakibatkan lamanya waktu untuk melakukan klasifikasi artikel ilmiah jika jumlah dokumen pelatihan sangat banyak. Selain itu, jika korpus didominasi oleh satu atau beberapa kategori (label), maksudnya satu atau beberapa kategori memiliki jumlah dokumen yang sangat banyak, sedangkan kategori lain hanya memiliki sedikit dokumen, penggunaan

metode *k-Nearest Neighbor* yang menentukan kategori berdasarkan jumlah dokumen terbanyak akan menghasilkan kategori yang salah. Untuk itu, digunakan metode *Improved k-Nearest Neighbor* (k-NN_I) yang melakukan modifikasi pada metode *k-Nearest Neighbor* biasa sehingga dapat meningkatkan performa dalam melakukan klasifikasi. Metode ini didasari oleh penelitian yang dilakukan oleh Muhammed Miah (2009). Dalam metode ini, tidak semua dokumen pelatihan akan dihitung kesamaannya dengan dokumen uji dan tidak menggunakan semua kata yang ada dalam dokumen pelatihan sehingga metode ini menjadi lebih efektif. Dalam penelitian ini, penulis akan merancang sebuah sistem klasifikasi menggunakan metode *Improved k-Nearest Neighbor*.

1.2 Perumusan Masalah

Berdasarkan latar belakang masalah diatas, penulis akan merancang dan membangun sebuah sistem yang akan melakukan proses klasifikasi artikel ilmiah. Masalah yang akan diteliti adalah :

- Seberapa besar akurasi yang dihasilkan sistem dalam melakukan klasifikasi dengan menggunakan metode *Improved k-Nearest Neighbor*.
- Bagaimana pengaruh nilai k terhadap akurasi sistem.
- Bagaimana pengaruh kemunculan keyword dan judul terhadap akurasi sistem.
- Bagaimana pengaruh penghapusan kata yang memiliki TF dan DF tinggi terhadap akurasi sistem.

1.3 Batasan Masalah

Batasan dalam sistem ini adalah sebagai berikut:

- Sistem yang dibangun berupa aplikasi *web* yang diuji pada jaringan lokal.
- Data yang digunakan adalah artikel ilmiah dalam bentuk artikel berbahasa Inggris yang bersumber pada <http://portal.acm.org>.
- Artikel ilmiah diklasifikasikan dalam 5 kategori di bidang *IT*, yaitu *Software Engineering* (Rekayasa Perangkat Lunak), *Information Technology* (Teknologi Informasi), *Communication and Networking* (Komunikasi dan Jaringan), *Intelligent Technology* (Sistem Cerdas), dan *Computer Science* (Ilmu Komputer).
- Pembobotan yang digunakan adalah perhitungan *TF* dimana *TF* ini merupakan perbandingan dari jumlah banyaknya suatu kata terhadap jumlah semua kata yang terdapat dalam suatu dokumen.
- *Stoplist* yang digunakan dalam sistem klasifikasi ini terdiri dari 570 *stopwords* yang diperoleh dari <http://jmlr.csail.mit.edu/papers/volume5/lewis04a/a11-smart-stop-list/english.stop>, 7 nama hari (dalam bahasa inggris), 12 nama bulan (dalam bahasa inggris), kata-kata yang sudah dipesan dalam MySQL sebanyak 131 kata yang diperoleh dari <http://dev.mysql.com/doc/refirman/5.0/en/reserved-words.html>.
- Pemilihan nilai *k* yang digunakan ditentukan oleh user sebelum melakukan proses klasifikasi.
- Tidak adanya tahap *stemming* dalam pre-processing dokumen.
- Bentuk masukan ke sistem berupa *single plained text* yang mengandung isi dari artikel ilmiah dan nama *file* berupa judul dari artikel tersebut, *keyword* artikel serta kategori dari artikel tersebut untuk dokumen pelatihan.
- Bentuk keluaran dari sistem berupa label kategori dari artikel yang dimasukkan untuk pengklasifikasian.

1.4 Hipotesis

- Implementasi metode *Improved k-Nearest Neighbor* dalam pengklasifikasian atikel ilmiah memberikan tingkat akurasi yang cukup tinggi ditinjau dari persentase jumlah dokumen yang diklasifikasian dengan benar.
- Nilai k yang semakin kecil menghasilkan tingkat akurasi yang semakin tinggi.
- Dengan mempertimbangkan kemunculan *keyword* dan judul dari artikel, dapat menghasilkan tingkat akurasi yang lebih tinggi tanpa mempertimbangkan kemunculan *keyword* dan judul.
- Penghapusan kata yang mempunyai TF dan DF tinggi menghasilkan tingkat akurasi yang lebih tinggi.

1.5 Tujuan Penelitian

Tujuan dari penelitian ini adalah menghasilkan sistem klasifikasi artikel yang akurat menggunakan metode *k-Nearest Neighbor* untuk kategori di bidang Teknologi Informasi dan Komputer, khususnya untuk bidang *Software Engineering, Information Technology, Intelligent Technology, Communication and Networking, dan Computer Science*.

1.6 Metode Penelitian

Metode yang digunakan dalam penelitian ini adalah metode *Improved k-Nearest Neighbor*, sedangkan metode yang digunakan untuk mengumpulkan data adalah Studi Pustaka. Studi Pustaka yang dilakukan dengan mencari dan mempelajari buku-buku dan sumber dari internet yang berkaitan dengan klasifikasi dokumen.

1.7 Sistematika Penulisan

Penulisan laporan tugas akhir ini dibagi menjadi lima (5) bab, yaitu :

Bab 1 Pendahuluan, yang memberikan gambaran umum mengenai apa yang diteliti dalam penulisan tugas akhir ini. Pendahuluan memuat latar belakang masalah, perumusan masalah, batasan masalah, hipotesis, tujuan penelitian, metode penelitian, dan sistematika penulisan laporan.

Bab 2 Tinjauan Pustaka, yang terdiri dari tinjauan pustaka dan landasan teori. Tinjauan pustaka menguraikan berbagai teori mengenai klasifikasi dengan metode *k-Nearest Neighbor* yang didapatkan dari berbagai sumber pustaka yang digunakan dalam melakukan penelitian. Landasan teori berisi konsep dan prinsip utama yang digunakan untuk memecahkan masalah penelitian.

Bab 3 Analisis dan Perancangan Sistem, mencakup tahap perancangan sistem yang akan dibuat seperti kebutuhan *hardware* dan *software*, spesifikasi sistem, arsitektur sistem, diagram *use case*, algoritma yang digunakan dalam membuat sistem, kamus data, skema *database*, rancangan antarmuka, dan rancangan pengujian sistem.

Bab 4 Implementasi dan Analisis Sistem, membahas implementasi dan pengujian sistem yang dibuat berdasarkan bab 3, beserta hasil dari sistem yang dijalankan dan analisis dari sistem yang dibuat.

Bab 5 Kesimpulan dan Saran, berisi kesimpulan dari hasil penelitian yang dilakukan dan saran untuk memberikan hasil yang lebih baik lagi dalam penelitian yang sejenis.

Bab 5

KESIMPULAN DAN SARAN

5.1 Kesimpulan

Berdasarkan hasil implementasi dan analisis sistem, maka diperoleh kesimpulan sebagai berikut :

1. Sistem klasifikasi dengan metode *Improved k-Nearest Neighbor* memberikan tingkat akurasi yang cukup tinggi dengan menggunakan nilai k sebesar 20, yaitu 83,33% sesuai pengujian yang telah dilakukan.
2. Semakin kecil nilai k yang digunakan dalam melakukan proses klasifikasi, maka semakin akurat hasil dari sistem. Pernyataan ini tidak berlaku untuk $k=1$ hingga $k=4$.
3. Dari hasil *confusion matrix* selama evaluasi sistem, diperoleh 7,69% *Computer and Networking* terjadi kebingungan ke *Information Technology*, 10,99% *Computer Science* terjadi kebingungan ke *Intelligent Technology*, 18,57% *Information Technology* terjadi kebingungan ke *Intelligent Technology*, 2,38% *Intelligent Technology* terjadi kebingungan ke *Computer Science*, dan 14,29% *Software Engineering* terjadi kebingungan ke *Computer Science*.
4. Memperbesar nilai perbandingan tf untuk metadata tidak memberi pengaruh yang berarti dalam proses klasifikasi.
5. Penghilangan token dengan nilai perbandingan tf dan df yang tinggi, setelah penghapusan *stopword*, dapat meningkatkan keakuratan sistem, namun tidak terlalu signifikan.

6. Algoritma pengecekan nilai jumlah perbandingan tf dengan nilai *similarity* terkecil dalam *buffer* pada metode *Improved k-Nearest Neighbor* tidak pernah berfungsi.

5.2 Saran

Sistem yang dibuat dapat dikembangkan lebih lanjut untuk mencapai hasil yang optimal. Saran yang diajukan penulis untuk pengembangan dan perbaikan sistem adalah :

1. Diperlukan perbaikan struktur data untuk mempercepat proses *text-preprocessing* karena semakin banyaknya kata dalam dokumen maka semakin lama waktu yang dibutuhkan untuk memprosesnya, dengan cara penggunaan *indexing* pada *database* dan lebih banyak penggunaan *stored procedure* untuk melakukan perhitungan sehingga mengurangi penggunaan *memory* dan mempercepat waktu yang dibutuhkan.
2. Dapat ditambahkan kemampuan untuk mem-*parsing keyword* dari artikel secara otomatis.
3. Dapat ditambahkan kemampuan untuk pengelolaan kategori dan *stoplist*.

DAFTAR PUSTAKA

- Feldman R., & Sanger J. (2007). *The Text Mining Handbook Advanced Approaches in Analyzing Unstructured Data*. New York: Cambridge University Press.
- Han E., Karypis G., Kumar V. (2001). *Text Categorization Using Weight Adjusted k-Nearest Neighbor Classification*. Diakses 13 Februari 2010, dari <http://www.springerlink.com/index/25gnd0jb6nklffhh.pdf>.
- Han J. & Kamber M. (2006). *Data Mining : Concepts and Technique 2nd Edition*. San Francisco: Morgan Kaufmann Publishers.
- Intan R. & Defeng A. (2006). HARD: Subject-Based Search Engine Menggunakan TF-IDF dan Jaccard's Coefficient. *Jurnal Teknik Industri*. 8(1). 61-72.
- Manning C. D., Raghavan P., & Schütze H. (2008). *An Introduction to Information Retrieval*. New York: Cambridge University Press.
- Miah M. (2009). *Improved k-NN Algorithm for Text Classification*. Diakses 27 Februari 2010, dari <http://dbxlab.uta.edu/dbxlab/Muhammed/DMIN2009.pdf>.
- Ridok A., Furqon M.T. (2009). *Pengelompokan Dokumen Berbahasa Indonesia Menggunakan Metode K-NN*. Diakses 13 Februari 2010, dari http://matematika.brawijaya.ac.id/web/cms/index2.php?option=com_docman&task=doc_view&gid=319&Itemid=99999999.