

**STUDI KOMPARATIF PEMANFAATAN VECTOR SPACE MODEL
DENGAN PENERAPAN ALGORITMA NAZIEF ADRIANI,
K-NEAREST NEIGHBOR DAN FUNGSI SIMILARITY COSINE:
STUDI KASUS PROTOTIPE SISTEM KLASIFIKASI TEKS
BAHASA INDONESIA**

TESIS



Oleh:

Diki Susandi
1311601239

**PROGRAM STUDI: MAGISTER ILMU KOMPUTER (MKOM)
PROGRAM PASCASARJANA
UNIVERSITAS BUDI LUHUR**

**JAKARTA
2016**

**STUDI KOMPARATIF PEMANFAATAN VECTOR SPACE MODEL
DENGAN PENERAPAN ALGORITMA NAZIEF ADRIANI,
K-NEAREST NEIGHBOR DAN FUNGSI SIMILARITY COSINE:
STUDI KASUS PROTOTIPE SISTEM KLASIFIKASI TEKS
BAHASA INDONESIA**

TESIS

Diajukan untuk memenuhi salah satu persyaratan memperoleh gelar
Magister Ilmu Komputer (MKOM)



Oleh:

Diki Susandi
1311601239

**PROGRAM STUDI: MAGISTER ILMU KOMPUTER (MKOM)
PROGRAM PASCASARJANA
UNIVERSITAS BUDI LUHUR**

**JAKARTA
2016**



**PROGRAM STUDI: MAGISTER ILMU KOMPUTER
PROGRAM PASCASARJANA
UNIVERSITAS BUDI LUHUR**

LEMBAR PERNYATAAN

Saya yang bertanda tangan dibawah ini:

Nama Mahasiswa : Diki Susandi
Nomor Induk Mahasiswa : 1311601239
Program Studi : Magister Ilmu Komputer
Konsentrasi : Teknologi Sistem Informasi
Fakultas/Program : Magister Ilmu Komputer

Menyatakan bahwa Tesis yang berjudul:

Studi Komparatif Pemanfaatan Vector Space Model
Dengan Penerapan Algoritma Naïf Adriani, K-
Nearest Neighbor Dan fungsi Similarity Cosine : Studi
Kasus Prototipe Sistem Klasifikasi teks Bahasa Indonesia.

1. Merupakan hasil karya tulis ilmiah sendiri dan bukan merupakan karya yang pernah diajukan untuk memperoleh gelar akademik oleh pihak lain,
2. Saya ijin untuk dikelola oleh Universitas Budi Luhur sesuai dengan norma hukum dan etika yang berlaku.

Pernyataan ini saya buat dengan penuh tanggung jawab dan saya bersedia menerima konsekuensi apapun sesuai aturan yang berlaku apabila di kemudian hari pernyataan ini tidak benar.

Jakarta, 29 Januari 2016



(Diki Susandi)



**PROGRAM STUDI: MAGISTER ILMU KOMPUTER
PROGRAM PASCASARJANA
UNIVERSITAS BUDI LUHUR**

LEMBAR PENGESAHAN TESIS

Nama Mahasiswa : Diki Susandi
Nomor Induk Mahasiswa : 1311601239
Konsentrasi : Teknologi Sistem Informasi
Judul Tesis : Studi Komparatif Pemanfaatan Vector Space Model
Dengan Penerapan Algoritma Nazief Adriani, K-
Nearest Neighbor Dan Fungsi Cosine: Studi Kasus
Prototipe Sistem Klasifikasi Teks Bahasa Indonesia

Jakarta, 29 Januari 2016

Tim Penguji:

Ketua,

(Prof. Dr. Moedjiono, M.Sc)

Tanda Tangan:

Anggota,

(Dr. Ir. Nazori AZ, M.T)

Pembimbing,

(Dr. Krisna Adiyarta M., M.Sc)

Ketua Program Studi

(Dr. Ir. Nazori AZ, M.T)

ABSTRAK

Informasi saat ini sangat mudah didapatkan oleh setiap orang dimanapun berada. Informasi bisa mudah didapatkan karena perkembangan teknologi informasi yang semakin cepat dan pengetahuan setiap orang yang terus bertambah. Teknologi informasi khususnya internet sangat mendukung terjadinya pertukaran informasi dengan sangat cepat. Internet menjadi media informasi dan telekomunikasi yang telah dimanfaatkan banyak orang dengan banyak kepentingan. Internet menciptakan banyak informasi dan karena kemudahan dalam akses internet inilah banyak orang menggunakan media online dalam kegiatannya sehari-hari. Setiap orang membutuhkan informasi dan mereka bisa mencari informasi dengan menggunakan mesin pencari yang ada di dunia maya. Begitupun dalam penyimpanan dokumen secara digital semakin meningkat. Kondisi tersebut memunculkan masalah untuk mengakses informasi yang diinginkan secara akurat dan cepat. Seringkali pada web, dimana kita mencari suatu informasi tertentu, banyak hal yang penting justru terlewatkan, malah yang tidak penting banyak terserap. Untuk mengatasi masalah tersebut, salah satu teknik yang dapat digunakan adalah dengan mengklasifikasikan teks tersebut sesuai dengan karakteristik, fitur, maupun kelasnya berdasarkan aturan baku bahasa yang akan diolah, dalam penelitian ini Bahasa Indonesia yang digunakan sebagai sumber acuan. Didasari alternatif tersebut, maka dalam penelitian ini akan dibangun suatu aplikasi perangkat lunak yang dapat melakukan klasifikasi data teks terhadap sumber informasi teks elektronik yang diunggah secara terpandu dan selektif. Metode yang digunakan untuk mendukung proses klasifikasi ini adalah Algoritma *Nazief Adriani*, *K-Nearest Neighbor* dan Fungsi *Cosine*. Hasil penelitian menunjukkan dengan menggunakan pembobotan TF-IDF, rata-rata ketepatan dan kelengkapan sistem dalam melakukan klasifikasi dokumen sebesar 70,7%. Sedangkan dengan menggunakan pembobotan WIDF, rata-rata ketepatan dan kelengkapan sistem dalam melakukan klasifikasi dokumen sebesar 63,1%.

Kata Kunci : Internet, Klasifikasi Dokumen, *Nazief Adriani*, *K-Nearest Neighbor*, *Cosine*

ABSTRACT

Current information is very easily obtained by any person wherever located. Information can be easily obtained due to the development of information technology is growing rapidly and the knowledge of each person who continues to grow. Information technology, especially the Internet strongly supports the exchange of information very quickly. Internet becomes a medium of information and telecommunications that have been used by many people with many interests. Internet creates a lot of information and because of the ease of access to the Internet is a lot of people use online media in their daily activities. Everyone needs information and they can search for information by using a search engine that is the virtual world. Likewise in the storage of digital documents is increasing. The condition raises a problem to access the desired information accurately and quickly. Often on the web, where we look for a specific information, a lot of important things it missed, even that did not matter much absorbed. To overcome these problems, a technique that can be used to classify the text according to the characteristics, features, and class based language standard rules to be processed, in this study Indonesian language that is used as a reference source. The alternative is based on, so in this study will be built a software application that can perform data classification text to text electronic resources uploaded guided and selective. The method used to support the classification process is Nazief Adriani algorithm, K-Nearest Neighbor and Cosine Functions. The results showed using TF - IDF weighting , the average accuracy and completeness of the systems classify documents by 70.7 % . While using the weighting WIDF, the average accuracy and completeness of the systems classify documents by 63,1% .

Keywords: Internet, Document Classification, Nazief Adriani, K-Nearest Neighbor, Cosine

KATA PENGANTAR

Puji dan syukur Penulis panjatkan kepada Allah SWT, Tuhan Yang Maha Esa, karena berkat rahmat dan hidayah-Nya sehingga penulis dapat menyelesaikan tesis yang berjudul “Studi Komparatif Pemanfaatan *Vector Space Model* Dengan Penerapan Algoritma *Nazief Adriani*, *K-Nearest Neighbor* Dan Fungsi *Cosine*: Studi Kasus Prototipe Sistem Klasifikasi Teks Bahasa Indonesia. Tujuan dari penyusunan tesis ini dalam rangka memenuhi salah satu syarat untuk menyelesaikan studi Strata-2 pada program studi Magister Ilmu Komputer (MKOM) Fakultas Pascasarjana Universitas Budi Luhur Jakarta.

Penulis sudah berusaha secara maksimal untuk meneliti dan menyusun tesis ini dan menyadari masih terdapat banyak kekurangan dari tesis ini, oleh karena itu Penulis mengharapkan kritik dan saran yang bersifat membangun demi kesempurnaan penyusunan selanjutnya. Penulis juga menyadari bahwa tanpa kesempatan, bimbingan, bantuan, serta dukungan, tesis ini tidak berarti sama sekali. Maka pada kesempatan ini, Penulis ingin menyampaikan terima kasih yang sebesar-besarnya kepada :

1. Dr. Krisna Adiyarta M., M.Sc selaku Dosen Pembimbing yang telah menyediakan waktu, saran, bantuan, bimbingan, pengarahan, dorongan, kepercayaan dan kesempatan kepada Penulis dalam menyelesaikan masalah-masalah yang dihadapi oleh Penulis selama menyelesaikan tesis ini.
2. Prof. Dr. Moedjiono, M.Sc selaku Direktur Program Pascasarjana sekaligus selaku Dosen yang telah memberikan banyak pengarahan mengenai tata cara penelitian dan penyusunan tesis ini.
3. Dr. Ir. Nazori AZ, M.T selaku Kaprodi Magister Komputer yang telah memberikan kepercayaan dan kesempatan kepada Penulis untuk menyusun tesis ini.
4. Bapak Legirin dan Ibu Eni selaku Orang Tua Penulis tercinta yang terus memberikan doa dan motivasi di dalam penyusunan tesis ini.

5. Dika Priatin dan Mei Mayang Sari selaku Saudara Kembar dan Adik Penulis tercinta yang terus memberikan dukungan motivasi di dalam penyusunan tesis ini.
6. Kelompok belajar tesis, Hasanusi, Sukisno selaku teman satu team dalam persiapan penyusunan tesis ini.
7. Fitri Rusdiany selaku Orang Spesial dan Tersayang yang selalu memberikan dukungan motivasi lebih dalam penyusunan tesis ini.
8. Teman-teman Mahasiswa Magister Ilmu Komputer Universitas Budi Luhur kelompok XC pada Semester 1, kelompok XL pada Semester 2 dan kelompok MI pada Semester 3, terima kasih atas kebersamaan, kerja sama dan dukungan serta motivasinya selama ini.
9. Bapak dan Ibu Dosen Pengampu Mata Kuliah di Program Studi Magister Ilmu Komputer Budi Luhur yang telah dengan ikhlas dan sabar dalam memberikan ilmu pengetahuan serta bimbingan dalam belajar.
10. Semua pihak yang secara langsung maupun tidak langsung yang telah membantu Penulis dalam menyelesaikan tesis ini yang tidak dapat disebutkan satu persatu.

Akhir kata, Penulis berharap agar tesis ini dapat bermanfaat bagi pihak yang memerlukannya. Amiin

Jakarta, 29 Januari 2016



(Diki Susandi)

DAFTAR ISI

	Halaman
LEMBAR HALAMAN SAMPUL	
LEMBAR HALAMAN JUDUL	i
LEMBAR PERNYATAAN	ii
LEMBAR PENGESAHAN TESIS	iii
ABSTRAK	iv
ABSTRACT	v
KATA PENGANTAR	vi
DAFTAR ISI	viii
DAFTAR GAMBAR	x
DAFTAR TABEL	xi
DAFTAR LAMPIRAN	xiii
BAB I PENDAHULUAN	1
1.1 Latar Belakang	1
1.2 Masalah Penelitian	2
1.2.1 Identifikasi Masalah	2
1.2.2 Pembatasan Masalah	3
1.2.3 Rumusan Masalah	3
1.3 Tujuan dan Manfaat Penelitian	4
1.3.1 Tujuan Penelitian	4
1.3.2 Manfaat Penelitian	4
1.4 Tata-Urut Penulisan	6
1.5 Daftar Pengertian	7
BAB II LANDASAN TEORI DAN KERANGKA KONSEP	9
2.1 Tinjauan Pustaka	9
2.2 Tinjauan Studi	44
2.3 Tinjauan Objek Penelitian	49
2.4 Kerangka Konsep/Pola Pikir Pemecahan Masalah	49
2.5 Hipotesis	50

BAB III	METODOLOGI DAN RANCANGAN/DESAIN PENELITIAN....	52
3.1	Metode Penelitian	52
3.2	Sampling/Metode Pemilihan Sampel.....	54
3.3	Instrumentasi	55
3.4	Teknik Analisis, Prototipe Model dan Pengujian	55
3.5	Langkah-langkah Penelitian	64
3.6	Jadwal Penelitian	67
BAB IV	PEMBAHASAN HASIL PENELITIAN	70
4.1	Pengkatagorian Data	70
4.2	Prototipe Sistem Klasifikasi Dokumen	73
4.3	Hasil Perhitungan Kemiripan Dokumen <i>Query</i>	76
4.3.1.	Hasil Perhitungan Dengan Pembobotan TF.IDF	77
4.3.2.	Hasil Perhitungan Dengan Pembobotan WIDF	84
4.4	Pengujian <i>Precision</i> dan <i>Recall</i>	90
4.5	Hasil Pembahasan	91
4.6	Implikasi Penelitian	92
4.7	Tahapan Rencana Implementasi Penelitian	93
BAB V	PENUTUP	96
5.1	Kesimpulan	96
5.2	Saran	96
DAFTAR PUSTAKA		98
LAMPIRAN-LAMPIRAN		102
DAFTAR RIWAYAT HIDUP		109

DAFTAR GAMBAR

Gambar	Halaman
II-1	Langkah-langkah Dalam Proses Data Mining 9
II-2	Bidang Ilmu Data Mining 13
II-3	Ilustrasi Information Retrieval System 14
II-4	Skema <i>Information Retrieval (IR) System</i> 15
II-5	Proses Sistem Temu-Kembali Informasi 18
II-6	Pembelajaran Mesin Dengan Klasifikasi Terpandu 20
II-7	Diagram Proses Klasifikasi Teks 23
II-8	<i>Flowchart Stemming</i> Bahasa Indonesia 28
II-9	Kerangka Fikir 50
III-1	Tahapan Pengideksan Teks 59
III-2	<i>Recall</i> dan <i>Precision</i> Pada Contoh Hasil Temu Kembali Informasi..... 61
III-3	<i>Trade-off</i> Antara <i>Precision</i> Dan <i>Recall</i> 63
III-4	Langkah-langkah Penelitian 65
IV-1	Halaman <i>Login Administrator</i> 73
IV-2	Halaman Utama <i>Administrator</i> 73
IV-3	Halaman Data <i>Administrator</i> 74
IV-4	Halaman Data Kategori 74
IV-5	Halaman <i>Learning Document</i> 75
IV-6	Halaman Klasifikasi Dokumen Dengan Pembobotan TF.IDF 75
IV-7	Halaman Klasifikasi Dokumen Dengan Pembobotan WIDF 76

DAFTAR TABEL

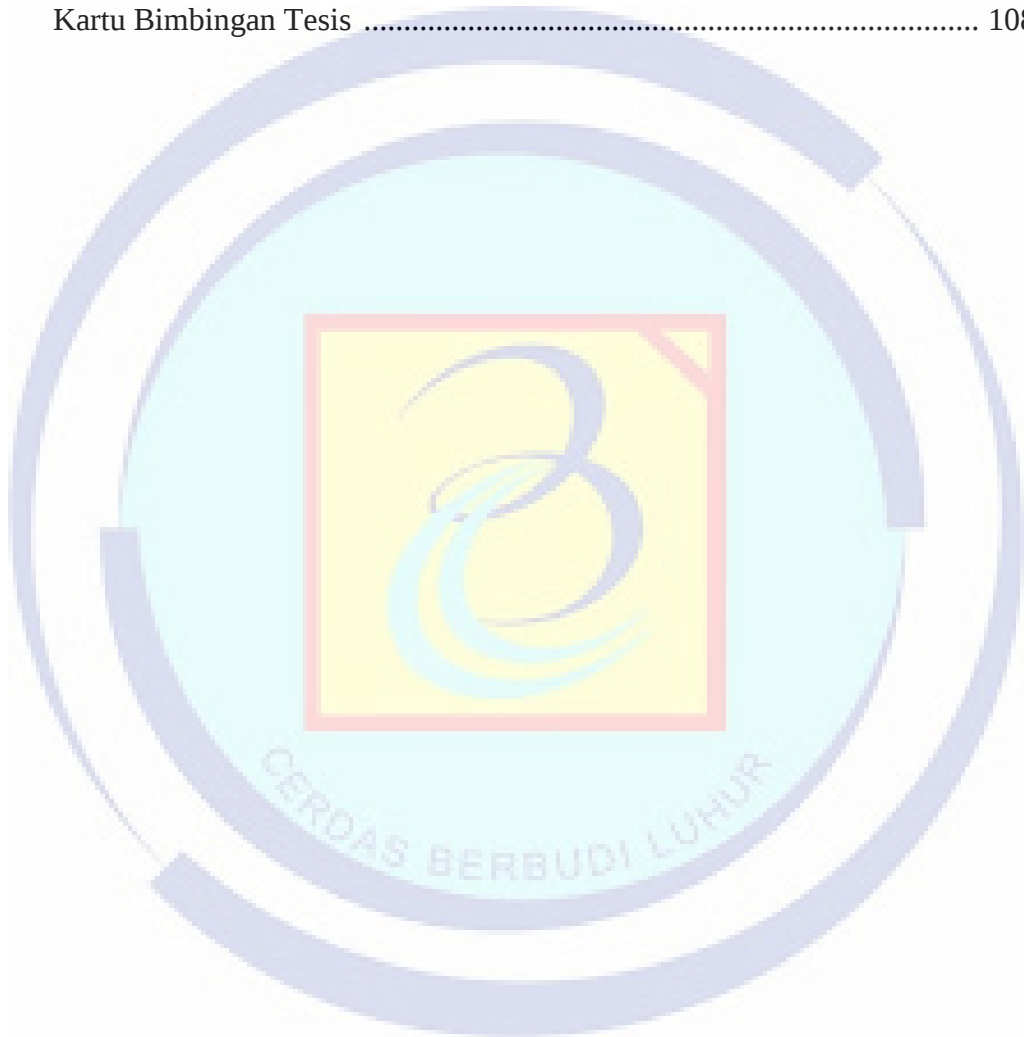
Tabel		Halaman
II-1	Perbandingan Algoritma Klasifikasi Teks	23
II-2	Kombinasi Awalan Akhiran Yang Tidak Diiijinkan	30
II-3	Cara Menentukan Tipe Awalan Untuk awalan “te-”	31
II-4	Jenis Awalan Berdasarkan Tipe Awalannya	31
II-5	<i>Example of Collection</i>	42
II-6	Ringkasan Tinjauan Studi	46
III-1	Jadwal Penelitian	67
IV-1	Hasil <i>Dataset</i> Terhadap Dokumen Sampel	71
IV-2	Hasil Perhitungan Jumlah Kemunculan <i>term</i>	77
IV-3	Hasil Perangkingan Q1 Dengan Masing-Masing Dokumen Sampel ...	82
IV-4	Hasil Perangkingan Q2 Dengan Masing-Masing Dokumen Sampel ...	82
IV-5	Hasil Perangkingan Q3 Dengan Masing-Masing Dokumen Sampel ...	82
IV-6	Hasil Perangkingan Q4 Dengan Masing-Masing Dokumen Sampel ...	82
IV-7	Hasil Perangkingan Q5 Dengan Masing-Masing Dokumen Sampel ...	82
IV-8	Hasil Perangkingan Q6 Dengan Masing-Masing Dokumen Sampel ...	83
IV-9	Hasil Perangkingan Q7 Dengan Masing-Masing Dokumen Sampel ...	83
IV-10	Hasil Perangkingan Q8 Dengan Masing-Masing Dokumen Sampel ...	83
IV-11	Hasil Perangkingan Q9 Dengan Masing-Masing Dokumen Sampel ...	83
IV-12	Hasil Perhitungan Jumlah Kemunculan <i>term</i>	84
IV-13	Hasil Perangkingan Q1 Dengan Masing-Masing Dokumen Sampel ...	88
IV-14	Hasil Perangkingan Q2 Dengan Masing-Masing Dokumen Sampel ...	89
IV-15	Hasil Perangkingan Q3 Dengan Masing-Masing Dokumen Sampel ...	89
IV-16	Hasil Perangkingan Q4 Dengan Masing-Masing Dokumen Sampel ...	89
IV-17	Hasil Perangkingan Q5 Dengan Masing-Masing Dokumen Sampel ...	89
IV-18	Hasil Perangkingan Q6 Dengan Masing-Masing Dokumen Sampel ...	90
IV-19	Hasil Perangkingan Q7 Dengan Masing-Masing Dokumen Sampel ...	90
IV-20	Hasil Perangkingan Q8 Dengan Masing-Masing Dokumen Sampel ...	90
IV-21	Hasil Perangkingan Q9 Dengan Masing-Masing Dokumen Sampel ...	90

IV-22	Hasil Pengujian <i>Precision</i> dan <i>Recall</i> Dengan KNN=3	91
IV-23	Hasil Perhitungan Nilai <i>Recall</i> dan <i>Precision</i>	91
IV-24	Tahapan Rencana Implementasi Penelitian	93



DAFTAR LAMPIRAN

Lampiran	Halaman
1 Indeks	103
2 Tabulasi Hasil Perhitungan <i>Similarity Cosine</i>	107
3 Kartu Bimbingan Tesis	108



DAFTAR PUSTAKA

- [Ari 2015] Andika, Ari. 2015. *Perancangan Aplikasi Pengukuran Similaritas Pada Dokumen Dengan Metode Semantic*. STMIK Budi Darma: Medan
- [Attardi 2004] Attardi, G., *Text Categorization*, Roma: Pisa University Pr.
- [Cahyo 2012] Cahyo Darujati, Agustinus BG. "Pemanfaatan Teknik Supervised untuk Klasifikasi Teks Bahasa Indonesia." Universitas Narotama, 2012.
- [Christopher 2009] Christopher, D. Manning, Prabhakar Raghavan, Hinrich Schütze. 2009]. *An Introduction to Information Retrieval*. Cambridge: Cambridge UP.
- [Fatmawati 2013] Fatmawati, *Studi Komparatif Kecepatan Temu Kembali Informasi di Depo Arsip Koran Suara Merdeka Antara Sistem Simpan Manual dengan Foto Repro*. Diponegoro University.
- [Feldman 2007] Feldman, Ronen & James Sanger. 2007. *The Text Mining Handbook*. Cambridge University Press.
- [Han 2006] Han, J., Kamber, M. 2006. *Data Mining Concept and Tehniques*. San Fransisco: Morgan Kauffman.
- [Henny 2013] Henny Leidiyana. 2013. *Penerapan Algoritma K-Nearest Neighbor Untuk Penentuan Resiko Kredit Kepemilikan Kendaraan Bermotor*. STMIK Nusa Mandiri: Jakarta
- [Kusrini 2009] Kusrini dan Luthfi, E. T., 2009, *Algoritma Data Mining*, Andi Offset, Yogyakarta
- [Larose 2005] Larose, Daniel T, 2005, *Discovering Knowledge in Data: An Introduction to Data Mining*. New Jersey: John Willey & Sons. Inc
- [Lasarus 2015] Lasarus, P. Malese. 2015. *Model Mesin Pencari Dokumen Bahasa Indonesia Studi Efektifitas pada Vektor Space Model Algoritma Stemming Poter Pembobotan Frekuensi Term Berbanding Frekuensi Term Dalam Pencarian dan Fungsi Kesamaan Cosine*. Magister Komputer, Universitas Budi Luhur Jakarta
- [Lu 1999] Lu. 1999. *Multimedia Database Management Systems*. Boston: Artech House, INC