



**UNIVERSITAS INDONESIA**

**PEMANFAATAN DOKUMEN *UNLABELED* PADA  
KLASIFIKASI TOPIK BERBASIS NAÏVE BAYES DENGAN  
ALGORITMA EXPECTATION MAXIMIZATION**

**SKRIPSI**

**Bayu Distiawan Trisedya**

**1205000215**

**PROGRAM: ILMU KOMPUTER**

**FAKULTAS: ILMU KOMPUTER**

**DEPOK**

**JULI, 2009**



**UNIVERSITAS INDONESIA**

**PEMANFAATAN DOKUMEN *UNLABELED* PADA  
KLASIFIKASI TOPIK BERBASIS NAÏVE BAYES DENGAN  
ALGORITMA EXPECTATION MAXIMIZATION**

**SKRIPSI**

**Diajukan sebagai salah satu syarat untuk memperoleh gelar S.Kom**

**Bayu Distiawan Trisedya**

**1205000215**

**PROGRAM: ILMU KOMPUTER**

**FAKULTAS: ILMU KOMPUTER**

**DEPOK**

**JULI, 2009**

## HALAMAN PERNYATAAN ORISINALITAS

**Skripsi ini adalah hasil karya saya sendiri,  
dan semua sumber baik yang dikutip maupun dirujuk  
telah saya nyatakan dengan benar.**

Nama : Bayu Distiawan Trisedya

NPM : 1205000215

Tanda Tangan : .....

Tanggal : 27 Juli 2009

## HALAMAN PENGESAHAN

Skripsi ini diajukan oleh :

Nama : Bayu Distiawan Trisedya

NPM : 1205000215

Program Studi : Ilmu Komputer

Judul Skripsi : Pemanfaatan Dokumen *Unlabeled* pada Klasifikasi Topik Berbasis Naïve Bayes dengan Algoritma Expectation Maximization

Telah berhasil dipertahankan di hadapan Dewan Penguji dan diterima sebagai bagian persyaratan yang diperlukan untuk memperoleh gelar Sarjana S.Kom pada Program Studi Ilmu Komputer, Fakultas Ilmu Komputer, Universitas Indonesia

### DEWAN PENGUJI

Pembimbing : Dr. Hisar Maruli Manurung, S.Kom (.....)

Penguji : Dra. Mirna Adriani, Ph.D (.....)

Penguji : Dr. Achmad Nizar Hidayanto (.....)

Ditetapkan di : Fakultas Ilmu Komputer

Tanggal : 27 Juli 2009

## KATA PENGANTAR

Puji syukur penulis panjatkan kepada Tuhan Yang Maha Esa karena atas karunia dan rahmat-Nya akhirnya penulis dapat menyelesaikan tugas akhir dalam jangka waktu yang telah ditentukan dan menghasilkan laporan tugas akhir ini. Selama melakukan tugas akhir, penulis mendapatkan bantuan dari berbagai pihak yang sangat berarti bagi penulis. Oleh sebab itu, penulis hendak menyampaikan ungkapan terima kasih kepada berbagai pihak sebagai berikut:

1. Orang tua, kakak, dan seluruh anggota keluarga lainnya yang selalu memberi dukungan, semangat, dan motivasi kepada penulis dalam menyelesaikan tugas akhir ini;
2. Bapak Ruli Manurung selaku dosen pembimbing tugas akhir;
3. Ibu Dina Cahyati selaku dosen pembimbing akademis;
4. Mr. Kamal Nigam yang telah bersedia menjawab pertanyaan-pertanyaan seputar klasifikasi dokumen serta memperkenalkan MinorThird sebagai *tools* dalam melakukan eksperimen klasifikasi dokumen;
5. Armando, Bernadia, Clara, Darwin, Denny, Hansel, Heninggar, Metti, Mulyandra, Refly, Rizal, Suryanto, Teddy, Vinky, dan rekan-rekan yang ada di Lab IR selama pengerjaan tugas akhir. Terima kasih karena telah membantu, mendukung, dan menemani penulis selama pengerjaan tugas akhir;
6. Achmad Rohman dan Mursal Rais yang telah memberi kritik dan saran selama penulisan laporan tugas akhir;
7. Adit, Bambang, Haryadi, Prajna, Rizky, Rio, dan rekan-rekan ekstrakurikuler badminton 2005 serta rekan-rekan tim futsal KTTK yang telah membantu, mendukung, dan menemani penulis selama pengerjaan tugas akhir;

8. Teman-teman Fakultas Ilmu Komputer Universitas Indonesia angkatan 2005 yang namanya tidak dapat disebutkan satu per satu di sini;
9. Semua staf pengajar, administrasi, dan keluarga besar Fakultas Ilmu Komputer Universitas Indonesia;
10. Semua pihak yang belum disebutkan di atas, yang secara langsung maupun tidak langsung telah memberikan kontribusi dan bantuannya atas kelancaran dan kesuksesan tugas akhir hingga penyusunan laporan ini.

Penulis sangat sadar bahwa dalam melakukan penelitian ini banyak kekurangan dan kesalahan yang telah penulis lakukan. Oleh karena itu, penulis sangat terbuka untuk setiap kritik dan saran yang membangun. Akhir kata, penulis berharap semoga laporan tugas akhir ini dapat memberi manfaat kepada para pembaca.

Depok, 27 Juli 2009

**Bayu Distiawan Trisedya**

## HALAMAN PERNYATAAN PERSETUJUAN PUBLIKASI KARYA ILMIAH UNTUK KEPENTINGAN AKADEMIS

Sebagai sivitas akademik Universitas Indonesia, saya yang bertanda tangan di bawah ini:

Nama : Bayu Distiawan Trisedya  
NPM : 1205000215  
Program Studi : Ilmu Komputer  
Fakultas : Ilmu Komputer  
Jenis karya : Skripsi

demikian demi pengembangan ilmu pengetahuan, menyetujui untuk memberikan kepada Universitas Indonesia **Hak Bebas Royalti Noneksklusif (*Non-exclusive Royalty-Free Right*)** atas karya ilmiah saya yang berjudul :  
Pemanfaatan Dokumen *Unlabeled* pada Klasifikasi Topik Berbasis Naïve Bayes dengan Algoritma Expectation Maximization

beserta perangkat yang ada (jika diperlukan). Dengan Hak Bebas Royalti Noneksklusif ini Universitas Indonesia berhak menyimpan, mengalihmedia/formatkan, mengelola dalam bentuk pangkalan data (*database*), merawat, dan memublikasikan tugas akhir saya selama tetap mencantumkan nama saya sebagai penulis/pencipta dan sebagai pemilik Hak Cipta.

Demikian pernyataan ini saya buat dengan sebenarnya.

Dibuat di : Depok  
Pada tanggal : 27 Juli 2009  
Yang menyatakan

( Bayu Distiawan Trisedya )

## DAFTAR ISI

HALAMAN JUDUL.....	i
HALAMAN PERNYATAAN ORISINALITAS.....	ii
HALAMAN PENGESAHAN.....	iii
KATA PENGANTAR .....	iv
HALAMAN PERNYATAAN PERSETUJUAN PUBLIKASI KARYA ILMIAH UNTUK KEPENTINGAN AKADEMIS .....	vi
ABSTRAK .....	vii
ABSTRACT .....	viii
DAFTAR ISI.....	ix
DAFTAR TABEL.....	xii
DAFTAR GAMBAR .....	xiii
DAFTAR PERSAMAAN.....	xv
DAFTAR LAMPIRAN.....	xvi
BAB 1    PENDAHULUAN .....	1
1.1    Latar Belakang .....	1
1.2    Permasalahan.....	2
1.3    Tujuan.....	2
1.4    Ruang Lingkup.....	2
1.5    Metodologi Penelitian .....	3
1.6    Sistematika Penulisan.....	3
BAB 2    LANDASAN TEORI.....	6
2.1    Klasifikasi Dokumen Teks .....	6
2.2 <i>Machine Learning</i> untuk Klasifikasi Dokumen Teks .....	7
2.3    Naïve Bayes.....	9
2.4    Expectation Maximization .....	11



2.5	Klasifikasi Dokumen Bahasa Indonesia.....	15
BAB 3	PERANCANGAN .....	17
3.1	Gambaran Umum Proses Klasifikasi Dokumen Teks .....	17
3.2	Data .....	19
3.3	Persiapan Dokumen.....	21
3.3.1	<i>Converting</i> .....	21
3.3.2	<i>Filtering</i> .....	21
3.4	<i>K-fold Cross Validation</i> .....	22
3.5	Pemilihan Fitur.....	22
3.6	<i>Term Documents Matrix</i> .....	24
3.7	Metode Klasifikasi Dokumen Teks.....	26
3.7.1	Naïve Bayes .....	26
3.7.2	Expectation Maximization .....	29
BAB 4	IMPLEMENTASI.....	38
4.1	Persiapan Dokumen.....	38
4.1.1	<i>Converting</i> .....	38
4.1.2	<i>Filtering</i> .....	40
4.2	Pemilihan Fitur.....	42
4.3	<i>K-fold Cross Validation</i> .....	43
4.4	Pembuatan <i>Term Documents Matrix</i> .....	44
4.5	Klasifikasi Dokumen Teks.....	46
4.5.1	Naïve Bayes .....	47
4.5.2	Expectation Maximization .....	50
BAB 5	HASIL DAN PEMBAHASAN .....	53
5.1	Variabel Eksperimen .....	53
5.2	Hasil Eksperimen terhadap Jumlah Fitur dan Penghilangan <i>Stopwords</i> .....	55

5.2.1	Hasil Klasifikasi untuk Data Dokumen Hukum.....	56
5.2.2	Hasil Klasifikasi untuk Data Artikel Media Massa.....	59
5.2.3	Hasil Klasifikasi untuk Data 20Newsgroups <i>Dataset</i> .....	61
5.3	Hasil Klasifikasi dari Aspek Jenis Fitur .....	63
5.3.1	Hasil Klasifikasi untuk Data Dokumen Hukum.....	63
5.3.2	Hasil Klasifikasi untuk Data Artikel Media Massa.....	64
5.3.3	Hasil Klasifikasi untuk Data 20Newsgroups <i>Dataset</i> .....	65
5.4	Hasil Klasifikasi dari Aspek Jumlah Kategori .....	66
5.4.1	Hasil Klasifikasi untuk Data Dokumen Hukum.....	66
5.4.2	Hasil Klasifikasi untuk Data Artikel Media Massa.....	67
5.4.3	Hasil Klasifikasi untuk Data 20Newsgroups <i>Dataset</i> .....	70
5.5	Pengaruh <i>Unlabeled Documents</i> terhadap Akurasi Klasifikasi.....	73
5.5.1	Hasil Klasifikasi untuk Data Dokumen Hukum.....	73
5.5.2	Hasil Klasifikasi untuk Data Artikel Media Massa.....	75
5.5.3	Hasil Klasifikasi untuk Data 20Newsgroups <i>Dataset</i> .....	78
5.6	Rangkuman Hasil Percobaan.....	80
BAB 6	PENUTUP .....	83
6.1	Kesimpulan.....	83
6.2	Kendala.....	84
6.3	Saran.....	84
	DAFTAR PUSTAKA .....	86

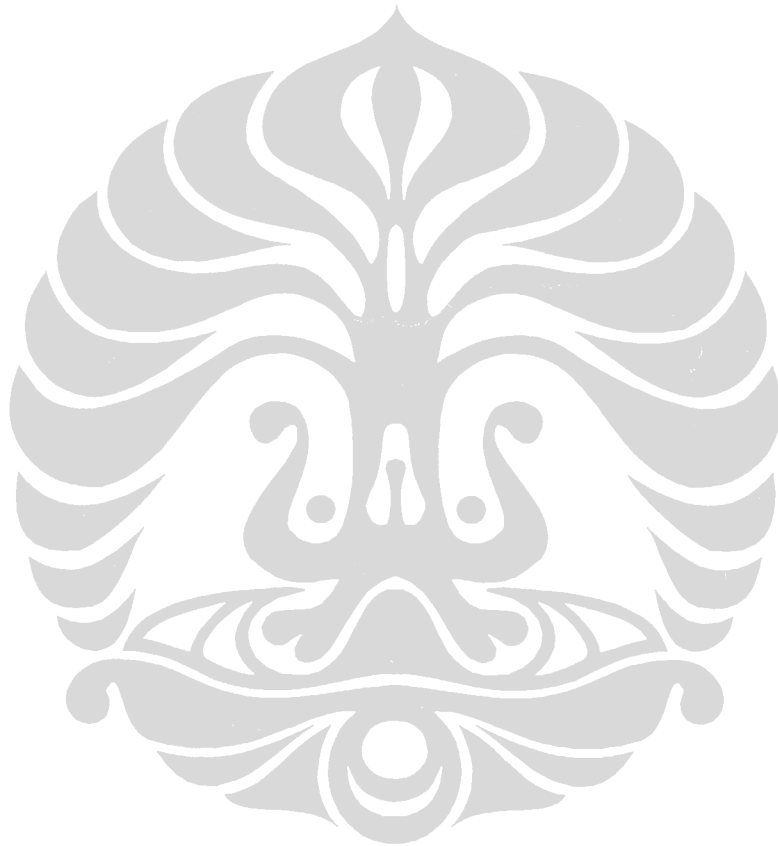
## DAFTAR TABEL

Tabel 3.1 Daftar Kategori dan Jumlah Dokumen yang Digunakan .....	20
Tabel 4.1 Variasi Jenis Fitur .....	44
Tabel 5.1 Variabel Percobaan .....	53
Tabel 5.2 Hasil Pemilihan Fitur pada Dokumen Hukum.....	57
Tabel 5.3 Hasil Pemilihan Fitur pada Artikel Media Massa.....	59
Tabel 5.4 Hasil Pemilihan Fitur pada 20Newsgroups <i>Dataset</i> .....	61
Tabel 5.5 Kesalahan Klasifikasi pada Dokumen Hukum .....	67
Tabel 5.6 Kesalahan Klasifikasi pada Artikel Media Massa .....	69
Tabel 5.7 Kesalahan Klasifikasi dengan Metode Naïve Bayes pada 20Newsgroups <i>Dataset</i> .....	72
Tabel 5.8 Pengaruh <i>Unlabeled Documents</i> pada Dokumen Hukum.....	74
Tabel 5.9 Pengaruh <i>Unlabeled Documents</i> pada Artikel Media Massa.....	76
Tabel 5.10 Pengaruh <i>Unlabeled Documents</i> pada 20Newsgroups <i>Dataset</i> .....	78

## DAFTAR GAMBAR

Gambar 2.1 Tahapan Algoritma Klasifikasi Naïve Bayes .....	11
Gambar 2.2 Tahapan Algoritma Klasifikasi Expectation Maximization.....	14
Gambar 3.1 Perancangan Percobaan Klasifikas Dokumen Teks .....	18
Gambar 3.2 <i>Term Documents Matrix</i> .....	24
Gambar 4.1 <i>Pseudocode</i> Proses <i>Converting</i> Dokumen .....	39
Gambar 4.2 Hasil Keluaran dari Text Mining Tools 1.1.42 .....	40
Gambar 4.3 <i>Pseudocode</i> Proses <i>Filtering</i> Dokumen .....	41
Gambar 4.4 <i>Pseudocode</i> Pemilihan Fitur.....	43
Gambar 4.5 <i>Pseudocode Folding</i> Dokumen .....	44
Gambar 4.6 Format Penyimpanan <i>Term Documents Matrix</i> dengan Informasi <i>Presence</i> .....	45
Gambar 4.7 <i>Pseudocode</i> Pembuatan <i>Term Documents Matrix</i> .....	46
Gambar 4.8 Format Berkas ARFF .....	47
Gambar 4.9 <i>Pseudocode</i> Klasifikasi Dokumen Teks Menggunakan Naïve Bayes ....	49
Gambar 4.10 Hasil Keluaran Klasifikasi Dokumen Teks dengan Naïve Bayes .....	49
Gambar 4.11 <i>Pseudocode</i> Konversi ARFF ke <i>Dataset</i> MinorThird.....	50
Gambar 4.12 <i>Pseudocode</i> Klasifikasi Dokumen Teks Menggunakan Expectation Maximization.....	51
Gambar 4.13 Hasil Keluaran Klasifikasi Dokumen Teks dengan Expectation Maximization .....	52
Gambar 5.1 Hasil Pemilihan Fitur pada Dokumen Hukum .....	58
Gambar 5.2 Hasil Pemilihan Fitur pada Artikel Media Massa .....	60
Gambar 5.3 Hasil Pemilihan Fitur pada 20Newsgroups <i>Dataset</i> .....	62
Gambar 5.4 Pengaruh Penggunaan Jenis Fitur pada Dokumen Hukum .....	63
Gambar 5.5 Pengaruh Penggunaan Jenis Fitur pada Artikel Media Massa .....	64
Gambar 5.6 Pengaruh Penggunaan Jenis Fitur pada 20Newsgroups <i>Dataset</i> .....	65
Gambar 5.7 Pengaruh Jumlah Kategori pada Dokumen Hukum.....	66

Gambar 5.8 Pengaruh Jumlah Kategori pada Artikel Media Massa.....	69
Gambar 5.9 Pengaruh Jumlah Kategori pada 20Newsgroups <i>Dataset</i> .....	71
Gambar 5.10 Pengaruh <i>Unlabeled Documents</i> pada Dokumen Hukum.....	75
Gambar 5.11 Pengaruh <i>Unlabeled Documents</i> pada Artikel Media Massa.....	77
Gambar 5.12 Pengaruh <i>Unlabeled Documents</i> pada 20Newsgroups <i>Dataset</i> .....	79



## DAFTAR PERSAMAAN

Persamaan (2.1) Probabilitas Kelas $c$ jika diketahui Dokumen $d$ .....	9
Persamaan (2.2) Probabilitas Kelas $c$ jika diketahui Dokumen $d$ dengan Perhitungan Hasil Perkalian dari Probabilitas Kemunculan Semua Kata pada Dokumen $d$ .....	10
Persamaan (2.3) Probabilitas Kata $w$ jika diketahui Kategori $c$ .....	10
Persamaan (2.4) Probabilitas Kategori $c$ .....	10
Persamaan (2.5) Probabilitas Kata $w$ jika diketahui Kategori $c$ dengan <i>Laplacian Smoothing</i> .....	10
Persamaan (2.6) Penentuan Kategori Hasil Klasifikasi .....	11
Persamaan (2.7) Probabilitas Keseluruhan Dokumen <i>Training</i> .....	12
Persamaan (2.8) Perhitungan <i>Log Likelihood</i> .....	12
Persamaan (2.9) Perhitungan <i>Complete Log Likelihood</i> .....	13
Persamaan (2.10) Probabilitas Kategori $c$ jika diketahui Dokumen $d$ pada <i>Expectation Step EM</i> .....	13
Persamaan (2.11) Probabilitas Kata $w$ jika diketahui Kategori $c$ pada <i>Maximization Step EM</i> .....	14
Persamaan (2.12) Probabilitas Kategori $c$ pada <i>Maximization Step EM</i> .....	14

## DAFTAR LAMPIRAN

LAMPIRAN 1 HASIL KLASIFIKASI DOKUMEN .....	87
LAMPIRAN 2 CONTOH DATA YANG DIPAKAI .....	102
LAMPIRAN 3 DAFTAR <i>STOPWORDS</i> .....	116

