



UNIVERSITAS INDONESIA

**KLASIFIKASI TOPIK MENGGUNAKAN METODE NAÏVE
BAYES DAN MAXIMUM ENTROPY PADA ARTIKEL MEDIA
MASSA DAN ABSTRAK TULISAN**

SKRIPSI

Dyta Anggraeni

1205000304

PROGRAM : ILMU KOMPUTER

FAKULTAS : ILMU KOMPUTER

DEPOK

JANUARI, 2008



UNIVERSITAS INDONESIA

**KLASIFIKASI TOPIK MENGGUNAKAN METODE NAÏVE
BAYES DAN MAXIMUM ENTROPY PADA ARTIKEL MEDIA
MASSA DAN ABSTRAK TULISAN**

SKRIPSI

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

Dyta Anggraeni

1205000304

PROGRAM : ILMU KOMPUTER

FAKULTAS : ILMU KOMPUTER

DEPOK

JANUARI, 2008

HALAMAN PERNYATAAN ORISINALITAS

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

Nama : Dyta Anggraeni

NPM : 1205000304

Tanda Tangan :

Tanggal : 24 Desember 2008

HALAMAN PENGESAHAN

Skripsi ini diajukan oleh :

Nama : Dyta Anggraeni
NPM : 1205000304
Program Studi : Ilmu Komputer
Judul Skripsi : Klasifikasi Topik Menggunakan Metode Naïve Bayes dan
Maximum Entropy pada Artikel Media Massa dan
Abstrak Tulisan Ilmiah

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 : Hisar Maruli Manurung (.....)
Penguji : Mirna Adriani (.....)
Penguji : Indra Budi (.....)

Ditetapkan di : Fakultas Ilmu Komputer
Tanggal : 24 Desember 2008

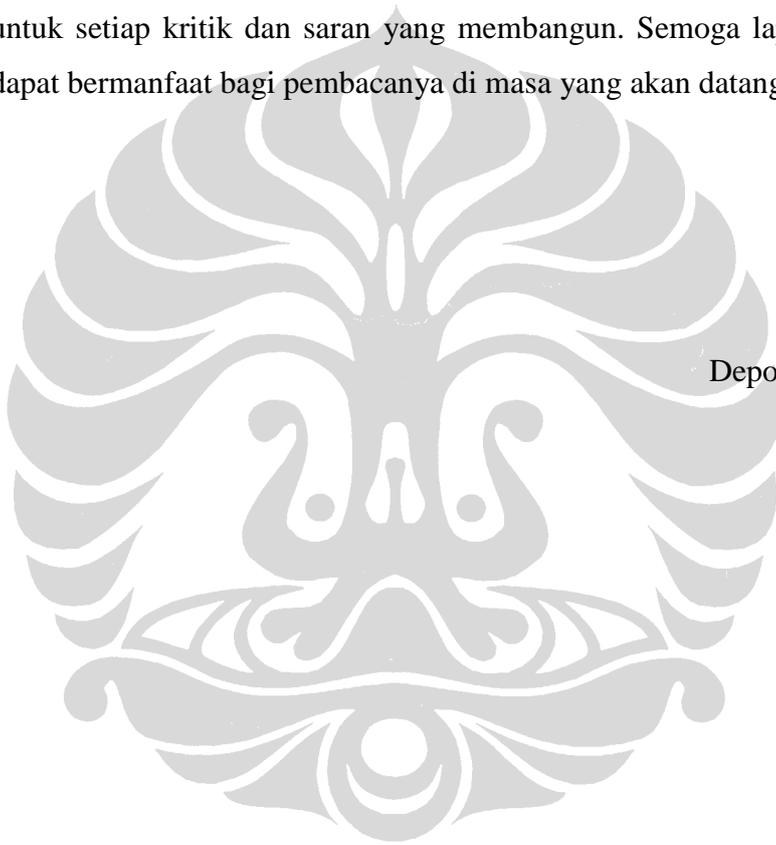
KATA PENGANTAR

Puji syukur penulis panjatkan kepada Tuhan Yang Maha Esa karena atas kasih karunia dan penyertaan-Nya akhirnya penulis dapat menyelesaikan tugas akhir dalam jangka waktu yang telah ditentukan dan menghasilkan laporan tugas akhir ini. Selama penulis 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 atas semua perhatian, kasih sayang, dukungan baik moral ataupun material, dan semangat yang penulis dapatkan selama penulis kuliah.
- (2) Bapak Ruli Manurung selaku dosen pembimbing tugas akhir.
- (3) Ibu Dina Chahyati selaku pembimbing akademis.
- (4) Fabian Sulaiman, sebagai orang yang selalu memberikan semangat dan dukungan terhadap penulis untuk dapat menyelesaikan tugas dan laporan ini dengan baik dan tepat pada waktunya.
- (5) 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) Clara Vania dan Bernadia Puspasari, yang telah menjadi teman baik bagi penulis dan telah selalu menemani penulis.
- (7) Anak-anak angkatan 2005, yang telah menemani penulis selama 4 tahun di Fakultas Ilmu Komputer. Terima kasih karena telah membuat kenangan indah selama 4 tahun ini.

- (8) Anak-anak KUKSACS UI, yang telah menemani dalam suka dan duka.
- (9) Teman-teman lain yang tidak dapat disebutkan oleh penulis. Terima kasih untuk semuanya

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. Semoga laporan tugas akhir ini dapat bermanfaat bagi pembacanya di masa yang akan datang.



Depok, 24 Desember 2008

Dyta Anggraeni

HALAMAN PERNYATAAN PERSETUJUAN PUBLIKASI TUGAS AKHIR UNTUK KEPENTINGAN AKADEMIS

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

Nama : Dyta Anggraeni
NPM : 1205000304
Program Studi : Ilmu Komputer
Fakultas : Ilmu Komputer
Jenis karya : Skripsi

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 :

Klasifikasi Topik Menggunakan Metode Naïve Bayes dan Maximum Entropy pada Artikel Media Massa dan Abstrak Tulisan Ilmiah

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 : 24 Desember 2008

Yang menyatakan

(Dyta Anggraeni)

DAFTAR ISI

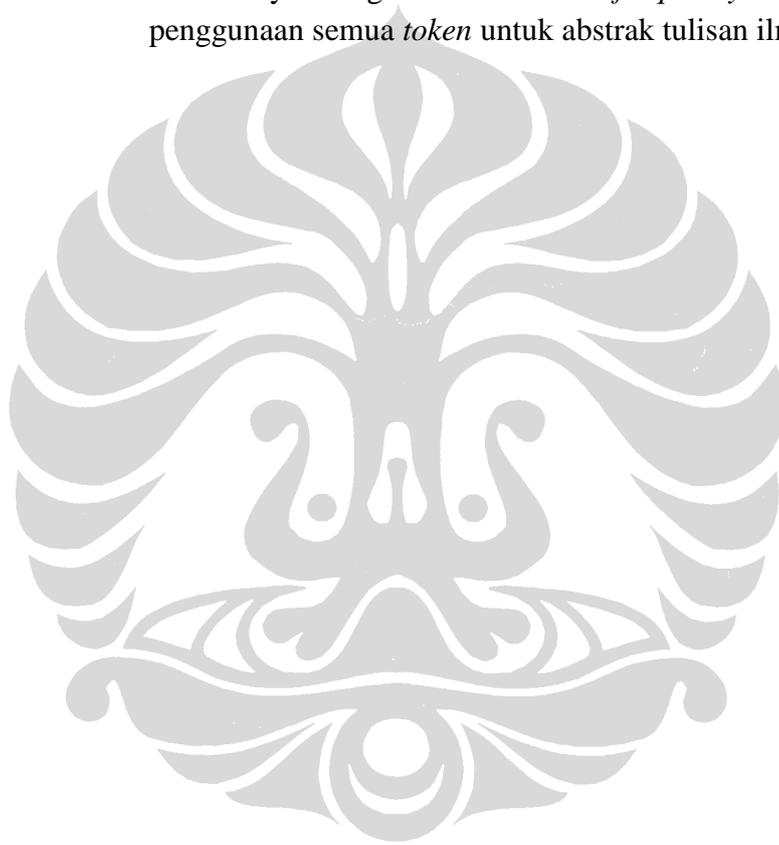
HALAMAN PERNYATAAN ORISINALITAS.....	ii
HALAMAN PENGESAHAN.....	iii
KATA PENGANTAR	iv
HALAMAN PERNYATAAN PERSETUJUAN PUBLIKASI TUGAS AKHIR UNTUK KEPENTINGAN AKADEMIS	vi
ABSTRAK	vii
ABSTRACT.....	viii
DAFTAR TABEL.....	xi
DAFTAR GAMBAR	xiii
BAB 1 PENDAHULUAN	1
1.1 Latar Belakang	1
1.2 Permasalahan.....	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 Topik.....	6
2.2 <i>Machine learning</i> untuk klasifikasi topik	8
2.3 Naïve Bayes	9
2.3.1 Model Naïve Bayes	10
2.3.2 Naïve Bayes Multinomial	12
2.4 Maximum Entropy	12
2.4.1. Entropy	13
2.4.2 Model Maximum Entropy	14
2.4.3 Model Parametrik Maximum Entropy.....	16
BAB 3 PERANCANGAN	19
3.1 Gambaran umum proses klasifikasi topik	19
3.2 Data	20
3.2.1. Persiapan Data	21
3.2.2 Analisis data	21
3.3. Pemilihan Fitur.....	22
3.4 <i>K-fold Cross Validation</i>	23

3.5. Matriks Pasangan Fitur Dokumen.....	24
3.6 Metode Klasifikasi Topik.....	25
3.6.1 Naïve Bayes	26
3.6.2 Maximum Entropy.....	26
Bab 4 IMPLEMENTASI	28
4.1 Persiapan data.....	28
4.1.1. Pengambilan data.....	28
4.1.2. Prapemrosesan data	32
4.1.3 Pembagian <i>Fold</i> Data	34
4.2 Implementasi Pemilihan Fitur	36
4.3 Pembuatan Matriks Pasangan Fitur-Dokumen.....	38
4.4. implementasi klasifikasi topik dengan <i>machine learning</i>	41
4.4.1 Analisis Sentiment dengan Naïve Bayes	42
4.4.2 Klasifikasi Topik dengan Maximum Entropy	45
BAB 5 HASIL DAN PEMBAHASAN.....	48
5.1 Hasil Klasifikasi Topik	48
5.2 Hasil Klasifikasi Topik dari Aspek Metode dan Fitur yang Digunakan	49
5.2.1 Hasil Klasifikasi dari Aspek Metode.....	51
5.2.2 Hasil Klasifikasi dari Aspek Jumlah <i>token</i>	53
5.2.3 Hasil Klasifikasi dari Aspek Informasi Fitur.....	55
5.3 Hasil Klasifikasi dari Banyaknya Topik yang Digunakan.....	58
5.4 Hasil Klasifikasi Topik dari Aspek Data yang Digunakan	64
5.4.1 Hasil Klasifikasi dari Aspek Keseragaman Jumlah Data untuk Tiap Topik.....	64
5.4.2 Hasil Klasifikasi dari Jumlah Data yang Digunakan.....	66
5.4.3 Hasil Klasifikasi dari Aspek Kemiripan Data	67
5.5 Kesalahan Klasifikasi Pada Beberapa Percobaan	68
5.6 Rangkuman Hasil	72
BAB 6 PENUTUP	74
6.1 Kesimpulan	74
6.2 Kendala	75
6.3 Saran.....	76
DAFTAR PUSTAKA	77

DAFTAR TABEL

Tabel 4. 1 Variasi Pemilihan Fitur	36
Tabel 4. 2 Variasi informasi nilai fitur	38
Tabel 5. 1 Keterangan mengenai variabel <i>input</i> yang digunakan pada percobaan.	48
Tabel 5. 2 Hasil nilai akurasi dari percobaan dengan menggunakan berbagai metode dan variasi pada jumlah <i>token</i> dan informasi yang terkandung dalam <i>token</i>	50
Tabel 5. 3 Hasil akurasi berbagai metode dengan variasi jumlah <i>tokens</i> dan memakai informasi <i>presence</i> untuk nilai fiturnya	51
Tabel 5. 4 Hasil nilai akurasi dari metode Naïve Bayes dan Maximum Entropy dengan menggunakan variasi jumlah <i>token</i> dan nilai informasi yang dimiliki fitur.....	53
Tabel 5. 5 Tabel Pengaruh informasi fitur pada nilai akurasi yang dihasilkan pada percobaan yang menggunakan metode Naïve Bayes (keterangan: p adalah <i>presence</i> , f adalah <i>frequency</i> , dan f-n adalah <i>frequency-normalized</i>).....	55
Tabel 5. 6 Tabel pengaruh informasi fitur pada nilai akurasi yang dihasilkan pada percobaan yang menggunakan metode Naïve Bayes Multinomial (keterangan: p adalah <i>presence</i> , f adalah <i>frequency</i> , dan f-n adalah <i>frequency-normalized</i>).....	57
Tabel 5. 7 Tabel nilai akurasi yang dihasilkan pada percobaan yang menggunakan metode Naïve Bayes dengan informasi <i>frequency</i> normalized dan Maximum Entropy untuk perubahan jumlah topik dari 2 hingga 5 pada artikel media massa.....	59
Tabel 5. 8 Tabel nilai akurasi yang dihasilkan pada percobaan yang menggunakan metode Naïve Bayes dengan informasi <i>frequency</i> normalized untuk perubahan jumlah topik 2 dan 3 pada abstrak tulisan ilmiah.	62
Tabel 5. 9 Hasil nilai akurasi dengan menggunakan metode Naïve Bayes dengan informasi <i>frequency-normalized</i> dan Maximum Entropy pada jumlah data 3 topik yang tidak seragam jumlahnya.	65
Tabel 5. 10 Tabel pengaruh keseragaman data untuk tiap topik pada nilai akurasi yang dihasilkan dengan menggunakan metode Naïve Bayes <i>frequency-normalized</i> dan Maximum Entropy.....	66
Tabel 5. 11 Tabel pengaruh tingkat kemiripan data untuk tiap topik pada nilai akurasi yang dihasilkan dengan menggunakan metode Naïve Bayes <i>frequency-normalized</i> dan Maximum Entropy.....	67

Tabel 5. 12	Tabel hasil klasifikasi pada percobaan yang menggunakan metode Maximum Entropy pada penggunaan semua <i>token untuk artikel media massa</i>	69
Tabel 5. 13	Tabel hasil klasifikasi pada percobaan yang menggunakan metode Naïve Bayes dengan informasi fitur <i>frequency-normalized</i> pada penggunaan semua <i>token untuk artikel media massa</i>	70
Tabel 5. 14	Tabel hasil klasifikasi pada percobaan yang menggunakan metode Maximum Entropy pada penggunaan semua <i>token</i> untuk abstrak tulisan ilmiah.	71
Tabel 5. 15	Tabel hasil klasifikasi pada percobaan yang menggunakan metode Naïve Bayes dengan informasi fitur <i>frequency-normalized</i> pada penggunaan semua <i>token</i> untuk abstrak tulisan ilmiah.	71



DAFTAR GAMBAR

Gambar 3. 1 Alur klasifikasi topik dengan <i>machine learning</i>	20
Gambar 3. 2 Matriks Pasangan Fitur-Dokumen	25
Gambar 4. 1 pseudocode pengambilan artikel secara otomatis	29
Gambar 4. 2 Tampilan yang ditunjukkan oleh <i>link</i> http://www.kompas.com/getrss/olahraga	30
Gambar 4. 3 Tampilan halaman dari salah satu link yang dirujuk pada halaman http://www.kompas.com/getrss/olahraga	31
Gambar 4. 4 <i>pseudocode</i> untuk melakukan proses randomisasi artikel.....	32
Gambar 4. 5 Pseudocode Pemilihan Fitur.....	37
Gambar 4. 6 <i>Pseudocode</i> pembuatan matriks pasangan fitur-dokumen untuk satu <i>fold</i>	39
Gambar 4. 7 Format Penyimpanan matriks pasangan fitur-dokumen dengan informasi <i>presence</i>	41
Gambar 4. 8 Format data ARFF.....	43
Gambar 4. 9 <i>Pseudocode</i> klasifikasi topik menggunakan metode Naïve Bayes ..	45
Gambar 4. 10 Contoh format data untuk OpenNLP Maxent	46
Gambar 4. 11 <i>Pseudocode</i> klasifikasi topik dengan menggunakan metode Maxent Entropy	47
Gambar 5. 1 Grafik pengaruh pemilihan metode pada nilai akurasi yang dihasilkan.	52
Gambar 5. 2 Grafik pengaruh jumlah <i>token</i> yang digunakan dengan nilai akurasi yang dihasilkan pada beberapa percobaan.	54
Gambar 5. 3 grafik pengaruh jenis informasi fitur pada nilai akurasi percobaan menggunakan metode Naïve Bayes.	56
Gambar 5. 4 grafik pengaruh jenis informasi fitur pada nilai akurasi percobaan menggunakan metode Naïve Bayes Multinomial.	58
Gambar 5. 5 Grafik pengaruh jumlah topik pada nilai akurasi percobaan menggunakan Naïve Bayes <i>Frequency-normalized</i> (f-n) dan Maximum Entropy untuk artikel media massa.	61
Gambar 5. 6 Grafik pengaruh jumlah topik pada nilai akurasi percobaan menggunakan Naïve Bayes <i>Frequency-normalized</i> (f-n) dan Maximum Entropy untuk abstrak tulisan ilmiah.....	63