

LAPORAN TUGAS AKHIR

**Analisis Sentimen Menggunakan Metode Naive Bayes,
Maximum Entropy, dan Support Vector Machine pada
Dokumen Berbahasa Inggris dan Dokumen Berbahasa
Indonesia Hasil Penerjemahan Otomatis**



Disusun oleh:

Franky

1204000343

Fakultas Ilmu Komputer

Universitas Indonesia

Depok, 2008

HALAMAN PERSETUJUAN

Judul Tugas Akhir:

Analisis Sentimen Menggunakan Metode Naive Bayes, Maximum Entropy, dan Support Vector Machine pada Dokumen Berbahasa Inggris dan Dokumen Berbahasa Indonesia Hasil Penerjemahan Otomatis

Nama: Franky

NPM: 1204000343

Laporan tugas akhir ini telah diperiksa dan disetujui.

Depok, Juli 2008

Pembimbing Tugas Akhir

Dr. Hisar Maruli Manurung

ABSTRAK

Sentimen merupakan opini atau penilaian penulis dokumen mengenai topik yang dibahas dalam dokumen tersebut. Analisis sentimen merupakan suatu tugas yang melakukan polarisasi dokumen berupa pengklasifikasian dokumen ke dalam sentimen positif dan negatif. Penggunaan metode Naive Bayes, Maximum Entropy, dan Support Vector Machine telah ditunjukkan mampu untuk menangkap informasi sentimen dari dokumen *review* film pada domain bahasa Inggris (Pang, Lee, & Vaithyanathan, 2002).

Laporan tugas akhir ini menjelaskan percobaan yang mengaplikasikan kembali metode Naive Bayes, Maximum Entropy, dan Support Vector Machine untuk analisis sentimen pada dokumen berbahasa Indonesia hasil penerjemahan otomatis menggunakan kamus bilingual dan program penerjemah, pada dokumen *review* film. Hasil analisis sentimen yang didapat dibandingkan dengan hasil analisis sentimen pada dokumen berbahasa Inggris. Percobaan analisis sentimen dilakukan dengan memvariasikan metode penerjemahan dan pengolahan data, fitur yang digunakan, dan informasi nilai fitur berupa nilai kemunculan fitur (*presence*), frekuensi, normalisasi nilai frekuensi, dan pembobotan menggunakan *tf-idf*. *Baseline* untuk analisis sentimen pada bahasa Indonesia dibuat dengan metode klasifikasi yang sederhana.

Hasil yang didapat menunjukkan bahwa analisis sentimen menggunakan *machine learning* untuk dokumen berbahasa Indonesia hasil penerjemahan otomatis dapat dilakukan, dengan akurasi tertinggi sebesar 78.82%. Hasil ini lebih baik dari akurasi yang didapat dari *baseline* sebesar 52.43% tetapi tidak melebihi akurasi tertinggi pada dokumen berbahasa Inggris sebesar 80.09%, namun cukup dekat. Penggunaan fitur yang diambil dari 25% bagian terakhir dokumen memberikan hasil yang lebih baik dari penggunaan fitur yang diambil dari keseluruhan dokumen. Sementara, metode Support Vector Machine secara umum memberikan hasil analisis sentimen dengan akurasi yang lebih baik dari metode *machine learning* lain yang digunakan.

KATA PENGANTAR

Puji syukur kepada Tuhan Yang Maha Esa atas penyertaannya untuk kehidupan yang dilalui dalam pengerjaan tugas akhir ini. Terima kasih yang sebesar-besarnya untuk Pak Ruli sebagai pembimbing tugas akhir yang memperhatikan dan membimbing dalam tiap tahapan pengerjaan tugas akhir serta memberikan ide-ide dalam menyelesaikan masalah yang ditemui. Terima kasih juga untuk sumber daya yang diberikan terutama untuk komputer yang memadai.

Terima kasih pada keluarga di rumah yang telah memberikan dukungan berupa penghidupan dan sumber daya untuk tugas akhir ini. Terima kasih juga untuk Mimi yang selalu ada terutama saat makan siang dan sebagai teman bicara. Selain itu, penulis juga ingin mengucapkan terima kasih kepada Desmond dan Eliza sebagai teman yang membantu dan memberikan pendapat yang membangun dalam penyusunan laporan tugas akhir. Ucapan terima kasih juga diberikan untuk teman-teman di laboratorium *information retrieval* (IR) yang cukup ramai dan membantu secara langsung atau tidak langsung lewat kata-kata, tindakan atau pengaruh yang dipancarkan. Terima kasih juga untuk Ibu Mirna yang selalu membawa sesuatu yang luar biasa untuk dikonsumsi saat pertemuan anggota lab pada hari Rabu. Terima kasih untuk penguji, Ibu Mirna dan Pak Stef, yang memberikan saran dan kritik yang positif untuk tugas akhir ini.

Akhir kata, terima kasih kepada teman-teman angkatan 2004 dan kepada seluruh elemen Fakultas Ilmu Komputer Universitas Indonesia yang turut serta mendukung dan mewarnai kehidupan selama berada di Fasilkom ini. Kritik dan saran yang membangun akan diterima dengan tulus. Semoga penelitian yang dilakukan ini membantu dalam pengembangan ilmu komputer di Indonesia.

Jakarta, 27 Juni 2008

Franky

DAFTAR ISI

HALAMAN PERSETUJUAN.....	ii
ABSTRAK	iii
KATA PENGANTAR	iv
DAFTAR ISI.....	v
DAFTAR GAMBAR	viii
DAFTAR TABEL.....	ix
BAB 1 PENDAHULUAN	1
1. 1 Latar Belakang	1
1. 2 Permasalahan.....	2
1. 3 Tujuan	2
1. 4 Ruang Lingkup.....	3
1. 5 Metodologi Penelitian	3
1. 6 Sistematika Penulisan	4
BAB 2 LANDASAN TEORI.....	6
2. 1 Analisis Sentimen	6
2. 1. 1 Penelitian dalam Analisis Sentimen.....	7
2. 1. 2 Machine Learning untuk Analisis Sentimen	8
2. 2 Naive Bayes	9
2. 2. 1 Model Naive Bayes	10
2. 2. 2 Naive Bayes Multinomial	12
2. 3 Maximum Entropy	13
2. 3. 1 Entropy	13
2. 3. 2 Model Maximum Entropy	14
2. 3. 3 Model Parametrik Maximum Entropy	16
2. 4 Support Vector Machine	18
2. 4. 1 Klasifikasi Linear dan Maximal Margin Classifier.....	19
2. 4. 2 Fungsi Kernel	22
2. 4. 3 Soft Margin	24
BAB 3 PERANCANGAN	26

3. 1	Gambaran Umum Proses Analisis Sentimen	26
3. 2	Data	28
3. 2. 1	Data Analisis	28
3. 2. 2	Kamus Bilingual Inggris-Indonesia	29
3. 3	Pemilihan Fitur.....	29
3. 4	K-fold Cross Validation	31
3. 5	Matriks Pasangan Fitur-Dokumen	31
3. 6	Metode Analisis Sentimen	33
3. 6. 1	Naive Bayes	33
3. 6. 2	Maximum Entropy	34
3. 6. 3	Support Vector Machine	35
3. 7	Perancangan Baseline.....	36
3. 7. 1	<i>Baseline</i> dengan Pelabelan Fitur secara Otomatis	36
3. 7. 2	<i>Baseline</i> dengan Pelabelan Fitur secara Manual.....	38
BAB 4 IMPLEMENTASI.....		40
4. 1	Persiapan Data.....	40
4. 1. 1	Kamus Bilingual	40
4. 1. 2	Data Analisis	41
4. 1. 3	Pembagian <i>Fold</i> Data.....	49
4. 2	Implementasi Pemilihan fitur.....	49
4. 3	Pembuatan Matriks Pasangan Fitur-Dokumen.....	51
4. 4	Implementasi Analisis Sentimen dengan Machine Learning.....	55
4. 4. 1	Analisis Sentimen dengan Naive Bayes.....	56
4. 4. 2	Analisis Sentimen dengan Maximum Entropy	58
4. 4. 3	Analisis Sentimen dengan Support Vector Machine	60
4. 5	Implementasi Baseline	61
4. 5. 1	Implementasi <i>Baseline</i> dengan Pelabelan Fitur secara Otomatis.....	62
4. 5. 2	Implementasi <i>Baseline</i> dengan Pelabelan Fitur secara Manual	66
BAB 5 HASIL DAN PEMBAHASAN.....		67
5. 1	Hasil Analisis Sentimen Baseline	67
5. 2	Hasil Analisis Sentimen dengan Machine Learning	71
5. 2. 1	Hasil Analisis Sentimen dari Aspek Metode dan Fitur.....	72

5. 2. 2 Hasil Analisis Sentimen dari Aspek Metode dan Bahasa	76
5. 3 Rangkuman Hasil	80
BAB 6 PENUTUP	83
6. 1 Kesimpulan	83
6. 2 Kendala	84
6. 3 Saran.....	85
DAFTAR PUSTAKA	87
LAMPIRAN A DATA REVIEW	90
A. 1 Data Review English.....	90
A. 2 Data Review English NOT Tag	90
A. 3 Data Review Transtool.....	91
A. 4 Data Review All-Trans Include dan All-Trans Ignore.....	91
A. 5 Data Review First-Trans Include dan First-Trans Ignore	93
A. 6 Data Review Last-Trans Include dan Last-Trans Ignore	94
LAMPIRAN B HASIL ANALISIS SENTIMEN	96
B. 1 Hasil Analisis Sentimen dengan Naive Bayes.....	96
B. 2 Hasil Analisis Sentimen dengan Naive Bayes Multinomial.....	97
B. 3 Hasil Analisis Sentimen dengan Maximum Entropy	98
B. 4 Hasil Analisis Sentimen dengan Support Vector Machine	99

DAFTAR GAMBAR

Gambar 2.1 Metode SVM yang Memisahkan Data ke Kelas +1 dan -1 pada Ruang Dua Dimensi.....	18
Gambar 2.2 <i>Hyperplane</i> (w, b) pada Ruang Dua Dimensi.	19
Gambar 2.3 Ilustrasi dari Data yang <i>Non-linearly Separable</i> pada Ruang Dua Dimensi.	22
Gambar 2.4 Ilustrasi dari <i>Soft Margin</i> Menggunakan <i>Slack Variable</i> ξ	24
Gambar 3.1 Alur Analisis Sentimen dengan <i>Machine Learning</i> pada Data <i>Review</i> Bahasa Inggris	27
Gambar 3.2 Matriks Pasangan Fitur-Dokumen	32
Gambar 3.3 Alur <i>Baseline</i> Analisis Sentimen dengan Pelabelan Fitur Otomatis	37
Gambar 3.4 Alur <i>Baseline</i> Analisis Sentimen dengan Pelabelan Fitur secara Manual.....	39
Gambar 4.1 Contoh Entri Kamus Bilingual Inggris-Indonesia.....	41
Gambar 4.2 <i>Pseudocode</i> Pemberian <i>Tag</i> NOT pada Data <i>Review</i> Bahasa Inggris	43
Gambar 4.3 <i>Pseudocode</i> Penerjemahan Otomatis Data <i>Review</i> dengan Kamus Bilingual	46
Gambar 4.4 <i>Pseudocode</i> Pemilihan Fitur	51
Gambar 4.5 <i>Pseudocode</i> Pembuatan Matriks Pasangan Fitur-Dokumen untuk Satu <i>Fold</i> dengan Informasi <i>Presence</i> , <i>Frequency</i> , dan <i>Frequency-Normalized</i>	53
Gambar 4.6 Format Penyimpanan Matriks Pasangan Fitur-Dokumen dengan Informasi <i>Presence</i>	54
Gambar 4.7 <i>Pseudocode</i> Pembobotan Tf-idf.....	55
Gambar 4.8 Format Data ARFF.....	56
Gambar 4.9 <i>Pseudocode</i> Analisis Sentimen dengan Naive Bayes untuk Satu Pasangan Data <i>Training</i> dan <i>Testing</i>	58
Gambar 4.10 Contoh Format Data untuk OpenNLP Maxent	59
Gambar 4.11 <i>Pseudocode</i> Analisis Sentimen Maximum Entropy untuk Satu Pasangan Data <i>Training</i> dan <i>Testing</i>	59
Gambar 4.12 Format <i>Input</i> Data untuk SVMLight.....	60
Gambar 4.13 Contoh Keluaran SVMLight untuk untuk Satu Pasangan Data <i>Training</i> dan <i>Testing</i>	61
Gambar 4.14 <i>Pseudocode</i> Pelabelan Fitur Otomatis untuk Satu Percobaan.....	64
Gambar 4.15 <i>Pseudocode</i> Analisis Sentimen <i>Baseline</i> untuk Satu Percobaan.....	65

DAFTAR TABEL

Tabel 4.1 Data <i>Review</i> untuk Analisis Sentimen	41
Tabel 4.2 Kata-kata Negasi Bahasa Inggris pada Data <i>Review</i>	44
Tabel 4.3 Variasi Pemilihan Fitur	50
Tabel 4.4 Variasi Informasi Nilai Fitur	52
Tabel 4.5 Variasi Informasi Fitur pada Pelabelan Fitur Otomatis	62
Tabel 4.6 Fitur Positif dan Negatif yang Dipilih Manual	66
Tabel 5.1 Hasil Analisis Sentimen untuk <i>Baseline</i> dengan Pelabelan Fitur Otomatis.....	68
Tabel 5.2 Penghitungan Total Nilai Fitur Positif dan Negatif pada Klasifikasi <i>Fold 3</i> dengan Fitur <i>Presence</i> Mengambil Semua Fitur tanpa <i>Threshold</i>	69
Tabel 5.3 Hasil Analisis Sentimen untuk <i>Baseline</i> dengan Pelabelan Fitur Manual	70
Tabel 5.4 Variabel Percobaan	71
Tabel 5.5 Jumlah Fitur Tiap Variasi Data untuk Penggunaan Fitur All-Features dan End- 25	72
Tabel 5.6 Hasil Analisis Sentimen dari Aspek Metode Terhadap Fitur dengan Informasi Nilai <i>Presence</i>	73
Tabel 5.7 Hasil Analisis Sentimen Maximum Entropy pada Semua Variasi Data dengan Informasi Nilai <i>Presence</i>	73
Tabel 5.8 Hasil Analisis Sentimen dari Aspek Metode Terhadap Fitur dengan Informasi Nilai <i>Frequency</i> dan <i>Frequency-Normalized</i>	74
Tabel 5.9 Hasil Analisis Sentimen dari Aspek Metode Terhadap Fitur dengan Informasi Nilai <i>TF-IDF</i> dan <i>TF-IDF-Normalized</i>	75
Tabel 5.10 Rata-rata Informasi Nilai Fitur Keseluruhan	76
Tabel 5.11 Rata-rata Variasi Fitur Keseluruhan	76
Tabel 5.12 Hasil Analisis Sentimen dari Aspek Bahasa Terhadap Metode.....	77