

**DETEKSI KOMENTAR SPAM MENGGUNAKAN
EKSTRAKSI FITUR DAN METODE *SUPPORT
VECTOR MECHINE* (SVM) PADA TEKS BERBAHASA
INDONESIA**

TUGAS AKHIR

**Sebagai Persyaratan Guna Meraih Gelar Sarjana Strata 1
Teknik Informatika Universitas Muhammadiyah Malang**



Oleh:

**DILA AISYAH RIMA WIDOWATI
201010370311425**

**JURUSAN TEKNIK INFORMATIKA
FAKULTAS TEKNIK
UNIVERSITAS MUHAMMADYAH MALANG**

2015

LEMBAR PERSETUJUAN

DETEKSI KOMENTAR SPAM MENGGUNAKAN EKSTRAKSI FITUR DAN METODE *SUPPORT VECTOR MECHINE* (SVM) PADA TEKS BERBAHASA INDONESIA

TUGAS AKHIR

Sebagai Persyaratan Guna Meraih Gelar Sarjana Strata 1
Teknik Informatika Universitas Muhammadiyah Malang

Menyetujui

Pembimbing I



Yufis Azhar, M.Kom
NIDN : 0728088701

Pembimbing II



Nur Hayatin, S.ST, M.Kom
NIDN : 0726038402

LEMBAR PENGESAHAN

DETEKSI KOMENTAR SPAM MENGGUNAKAN EKSTRAKSI FITUR DAN METODE SUPPORT VECTOR MECHINE (SVM) PADA TEKS BERBAHASA INDONESIA

TUGAS AKHIR

Sebagai Persyaratan Guna Meraih Gelar Sarjana Strata 1
Teknik Informatika Universitas Muhammadiyah Malang

Disusun Oleh:

Dila Aisyah Rima Widowati

201010370311425

Tugas Akhir ini telah di uji dan dinyatakan lulus melalui sidang majelis penguji

Menyetujui,

Penguji I



Galih Wasis Wicaksono, S.Kom, M.Cs

NIP : 108.1410.0541

Penguji II



Hyas Nuryasin, S.Kom, M.Kom

NIP : 108.1410.0561

Mengetahui,

Ketua Jurusan Teknik Informatika



Yuda Munarko, S.Kom, M.Sc

NIP : 108.0611.0443

LEMBAR PERNYATAAN

Yang bertanda tangan dibawah ini :

NAMA : DILA AISYAH RIMA WIDOWATI

NIM : 201010370311425

FAK./JUR. : TEKNIK / INFORMATIKA

Dengan ini saya menyatakan bahwa Tugas Akhir dengan judul **“DETEKSI KOMENTAR SPAM MENGGUNAKAN EKSTRAKSI FITUR DAN METODE SUPPORT VECTOR MECHINE (SVM) PADA TEKS BERBAHASA INDONESIA”** beserta seluruh isinya adalah karya saya sendiri dan bukan merupakan karya tulis orang lain, baik sebagian maupun seluruhnya, kecuali dalam bentuk kutipan yang telah disebutkan sumbernya.

Demikian surat pernyataan ini saya buat dengan sebenar-benarnya. Apabila kemudian ditemukan adanya pelanggaran terhadap etika keilmuan dalam karya saya ini, atau ada klaim dari pihak lain terhadap keaslian karya saya ini maka saya siap menanggung segala bentuk resiko/sanksi yang berlaku.

Mengetahui,

Dosen Pembimbing



Yufis Azhar, M.Kom
NIP : 0728088701

Malang, 16 April 2015

Yang Membuat Pernyataan



Dila Aisyah Rima Widowati
NIM : 201010370311425

KATA PENGANTAR

Dengan memanjatkan puji syukur kehadirat Allah SWT. Atas segala limpahan rahmat dan hidayah-NYA sehingga saya dapat menyelesaikan Tugas Akhir yang berjudul :

“DETEKSI KOMENTAR SPAM MENGGUNAKAN EKSTRAKSI FITUR DAN METODE SUPPORT VECTOR MACHINE (SVM) PADA TEKS BERBAHASA INDONESIA”

Di dalam tulisan ini disajikan pokok-pokok bahasan yang meliputi :

1. Perancangan dan implementasi pendeteksi komentar *spam* dengan mengimplementasi metode text mining dan algoritma SUPPORT VECTOR MACHINE (SVM).
2. Melakukan seleksi fitur yang akan digunakan untuk klasifikasi.
3. Melakukan pengujian berdasarkan penggunaan beberapa kombinasi fitur yang ada.
4. Membandingkan hasil klasifikasi berdasarkan beberapa kombinasi fitur dan melakukan pengamatan dari hasil klasifikasi yang bertujuan untuk menarik kesimpulan dari seluruh kegiatan yang ada.

Saya menyadari sepenuhnya bahwa dalam penulisan Tugas Akhir ini masih banyak kekurangan dan keterbatasan. Oleh karena itu saya mengharapkan saran yang membangun agar tulisan ini bermanfaat bagi perkembangan ilmu pengetahuan kedepan.

Malang, 16 April 2015

Penulis

DAFTAR ISI

ABSTRAK	i
ABSTRACT	ii
LEMBAR PERSETUJUAN	iii
LEMBAR PENGESAHAN	iv
LEMBAR PERNYATAAN	v
LEMBAR PERSEMBAHAN	vi
KATA PENGANTAR	vii
DAFTAR ISI	viii
DAFTAR GAMBAR	xi
DAFTAR TABEL	xiii
BAB I PENDAHULUAN	1
1.1 LATAR BELAKANG	1
1.2 RUMUSAN MASALAH	2
1.3 TUJUAN	2
1.4 BATASAN MASALAH	2
1.5 METODOLOGI	3
1.5.1 Studi Pustaka	3
1.5.2 Analisa Sistem	3
1.5.3 Perancangan Sistem	3
1.5.4 Implementasi	3
1.5.5 Pengujian Perangkat Lunak	4
1.5.6 Pembuatan Laporan	4
1.6 SISTEMATIKA PENULISAN	4
1.6.1 Bab I : Pendahuluan	4
1.6.2 Bab II : Landasan Teori	4
1.6.3 Bab III : Analisa dan Perancangan	4
1.6.4 Bab IV : Implementasi dan Pengujian	5
1.6.5 Bab V : Penutup	5
BAB II LANDASAN TEORI	6

2.1.1	<u>Pengertian Blog</u>	6
2.1.2	<u>Struktur Blog</u>	6
2.2	<u>SPAM</u>	7
2.2.1.	<u>Pengertian Spam</u>	7
2.2.2.	<u>Macam-Macam Spam</u>	7
2.3	<u>DATA MINING</u>	9
2.3.1	<u>Pengertian Data mining</u>	9
2.3.2	<u>Pekerjaan dalam Data mining</u>	10
2.4	<u>PERHITUNGAN SIMILARITY PADA BLOG POST DAN KOMENTAR</u>	11
2.5	<u>SUPPORT VECTOR MACHINE (SVM)</u>	12
2.6	<u>LIBSVM</u>	18
2.7	<u>PENGUJIAN</u>	19
<u>BAB III ANALISIS DAN PERANCANGAN</u>		21
3.1	<u>ANALISA MASALAH DAN GAMBARAN UMUM</u>	21
3.2	<u>PERANCANGAN SISTEM</u>	22
3.2.1	<u>Flowchart sistem</u>	22
3.2.2	<u>Tahap Preprosesing dan Seleksi Fitur</u>	23
3.3	<u>DESAIN ANTARMUKA</u>	33
3.3.1	<u>Tampilan Antarmuka Home</u>	33
3.3.2	<u>Tampilan Antarmuka Klasifikasi</u>	34
<u>BAB IV IMPLEMENTASI DAN PENGUJIAN</u>		35
4.1	<u>IMPLEMENTASI SISTEM</u>	35
4.1.1	<u>Implementasi Preprosesing</u>	35
4.1.2	<u>Implementasi Pembobotan Fitur</u>	38
4.1.3	<u>Implementasi Algoritma Support Vector Machine (SVM) menggunakan LIBSVM</u>	41
4.2	<u>PENGUJIAN SISTEM</u>	43
4.2.1	<u>Pengujian Fungsionalitas Sistem</u>	44
4.2.2	<u>Pengujian Keberhasilan Sistem</u>	46
<u>BAB V PENUTUPAN</u>		51

<u>5.2</u> <u>SARAN</u>	52
<u>DAFTAR PUSTAKA</u>	53

DAFTAR GAMBAR

BAB II

Gambar 2. 1 bussiness intelligence (diambil dari Buku “Konsep Data Mining Konsep dan Aplikasi Menggunakan Matlab)	9
Gambar 2. 2 Margin Hyperplane (diambil dari Buku “Konsep Data Mining Konsep dan Aplikasi Menggunakan Matlab)	12
Gambar 2. 3 Mencari fungsi pemisah yang optimal untuk obyek yang bisa dipisahkan secara linie	13
Gambar 2. 4 Memperbesar margin bisa meningkatkan probabilitas pengelompokkan suatu data secara benar	14

BAB III

Gambar 3. 1 Flowchart Sistem.....	23
Gambar 3. 2 Contoh kasus	23
Gambar 3. 3 Hasil dari proses casefolding	24
Gambar 3. 4 Hasil dari proses tokenizing.....	25
Gambar 3. 5 Hasil dari proses perubahan kata baku.....	25
Gambar 3. 6 Hasil dari proses <i>stopword removal</i>	26
Gambar 3. 7 Hasil dari proses stemming	27
Gambar 3. 8 Perhitungan TF.....	27
Gambar 3. 9 Pendekteksian link aktif	28
Gambar 3. 10 Pendekteksian anonim.....	28
Gambar 3. 11 Pendekteksian perbedaan waktu komentar dan <i>posting</i>	29
Gambar 3. 12 Pendekteksian kalimat promosi atau ajakan	32
Gambar 3. 13 Tampilan antarmuka home – data <i>train</i>	33
Gambar 3. 14 Tampilan antarmuka home – data <i>test</i>	33
Gambar 3. 15 Tampilan antarmuka klasifikasi – hasil klasifikasi	34

BAB IV

Gambar 4. 1 Proses casefolding.....	36
Gambar 4. 2 Proses tokenizing	36
Gambar 4. 3 Proses perubahan kata baku	36
Gambar 4. 4 Proses <i>stopword removal</i>	37
Gambar 4. 5 Proses stemming	38
Gambar 4. 6 Proses TF – mengambil kata dan menghitung jumlah masing-masing kata	38
Gambar 4. 7 Proses TF – menyimpan hasil perhitungan kata ke database.....	38
Gambar 4. 8 Proses pengecekan url.....	39
Gambar 4. 9 Proses pengecekan nama author	39
Gambar 4. 10 Proses pengecekan waktu <i>post</i> dan komentar	39
Gambar 4. 11 Proses pengecekan <i>post similarity</i>	40
Gambar 4. 12 Proses pengecekan duplikasi kata	40
Gambar 4. 13 Proses pengecekan <i>stopword ratio</i>	41
Gambar 4. 14 Proses pengecekan kalimat promosi	41
Gambar 4. 15 Proses perubahan format.....	42
Gambar 4. 16 Proses pengaturan parameter	42
Gambar 4. 17 Proses menyimpan model	43
Gambar 4. 18 Proses klasifikasi pada data <i>test</i>	43
Gambar 4. 19 Halaman Home	44
Gambar 4. 20 Halaman Klasifikasi	45
Gambar 4. 21 Halaman Evaluasi	46

DAFTAR TABEL

Tabel 4. 1 Contoh hasil uji untuk kategori kelas <i>spam</i>	46
Tabel 4. 2 Contoh hasil uji untuk kategori kelas <i>non-spam</i>	48
Tabel 4. 3 Pengujian dengan melakukan kombinasi fitur.....	49
Tabel 4. 4 Hasil Pengujian	50

Daftar Pustaka

- [1] A. Rajadesingan and A. Mahendran, "Comment Spam Classification in Blogs through Comment Analysis and Comment-Blog Post Relationships," *Comment Spam Classif. Blogs through Comment Anal. Comment-Blog Post Relationships*, pp. 490–501, 2012.
- [2] P. Kolari, A. Java, T. Finin, and J. Mayfield, "Blog track open task: Spam blog classification," *TREC Blog Track ...*, 2006.
- [3] A. Bhattarai, V. Rus, and D. Dasgupta, "Characterizing comment spam in the blogosphere through content analysis," *Comput. Intell. ...*, 2009.
- [4] E. Prasetyo, *Data Mining: Konsep dan Aplikasi Menggunakan Matlab*. Yogyakarta, Indonesia: Penerbit ANDI, 2012.
- [5] R. Ferdig and K. Trammell, "Content delivery in the 'blogosphere,'" *Journal*, no. February, 2004.
- [6] S. C. Herring and E. Wright, "Bridging the Gap: A Genre Analysis of Weblogs," pp. 1–11, 2004.
- [7] T. Y. Huann, O. Eu, G. John, J. Marie, and H. Pau, "Weblogs in Education," in *IT Literature Review*, 2005, pp. 1–10.
- [8] A. Thomason, "Blog Spam: A Review.," *CEAS*, pp. 2–5, 2007.
- [9] C. Chang and C. Lin, "LIBSVM: A Library for Support Vector Machines," pp. 1–39, 2013.
- [10] C. Chang and C. Lin, "LIBSVM: a Library for Support Vector Machines," pp. 1–26, 2003.
- [11] M. Powers, "Evaluation: from Precision, Recall and F-measure to ROC, Informedness, Markedness & Correlation," pp. 37–63, 2011.