ADLN - Perpustakaan Universitas Airlangga

SISTEM DETEKSI BAHASA MENGGUNAKAN *N-GRAM* UNTUK DOKUMEN BAHASA INDONESIA DAN BAHASA INGGRIS

SKRIPSI

KHARISMA

PROGRAM STUDI S1 SISTEM INFORMASI
FAKULTAS SAINS DAN TEKNOLOGI
UNIVERSITAS AIRLANGGA
SURABAYA
2015

SISTEM DETEKSI BAHASA MENGGUNAKAN *N-GRAM* UNTUK DOKUMEN BAHASA INDONESIA DAN BAHASA INGGRIS

SKRIPSI

Sebagai Salah Satu Syarat untuk Memperoleh Gelar Sarjana Komputer

Bidang Sistem Informasi di Fakultas Sains dan Teknologi

Universitas Airlangga

Oleh :

KHARISMA
NIM. 080810705

Tanggal Ujian : 27 Januari 2015

Disetujui Oleh :

Pembimbing I	Pembimbing II
Badrus Zaman, S.Kom., M.Cs	**Eva Hariyanti, S.Si., MT**
NIP. 19780126 200604 1 001	**NIP. 19810508 200501 2 001**

LEMBAR PENGESAHAN NASKAH SKRIPSI

Judul	**: SISTEM DETEKSI BAHASA MENGGUNAKAN *N-GRAM* UNTUK DOKUMEN BAHASA INDONESIA DAN BAHASA INGGRIS**
Penyusun	**: KHARISMA**
NIM	**: 080810705**
Tanggal Ujian	**: 27 Januari 2015**
1. Pembimbing I	**: Badrus Zaman, S.Kom., M.Cs**
2. Pembimbing II	**: Eva Hariyanti, S.Si., MT**

Disetujui Oleh :

Pembimbing I	**Pembimbing II**
Badrus Zaman, S.Kom., M.Cs	**Eva Hariyanti, S.Si., MT**
NIP. 19780126 200604 1 001	**NIP. 19810508 200501 2 001**

Mengetahui,

Ketua Departemen Matematika	**Ketua Program Studi S1 Sistem Informasi**
Fakultas Sains dan Teknologi	**Fakultas Sains dan Teknologi**
Universitas Airlangga	**Universitas Airlangga**
Dr. Miswanto, M.Si	**Drs. Eto Wuryanto, DEA**
NIP. 19680204 199303 1 002	**NIP. 19660928 199102 1 001**

PEDOMAN PENGGUNAAN SKRIPSI

Skripsi ini tidak dipublikasikan, namun tersedia di perpustakaan dalam lingkungan Universitas Airlangga, diperkenankan untuk dipakai sebagai referensi kepustakaan, tetapi pengutipan harus seijin penyusun dan harus menyebutkan sumbernya sesuai kebiasaan ilmiah.

Dokumen skripsi ini merupakan hak milik Universitas Airlangga

KATA PENGANTAR

Puji syukur kehadirat Tuhan Yang Maha Esa, yang telah melimpahkan anugerah-Nya, sehingga penulis dapat menyelesaikan skripsi yang berjudul "Sistem Deteksi Bahasa dengan Menggunakan *N-Gram* untuk Dokumen Bahasa Indonesia dan Bahasa Inggris" dengan lancar dan baik.

Tak lupa penulis mengucapkan banyak terima kasih kepada Badrus Zaman, S.Kom., M.Cs selaku dosen pembimbing I dan Eva Hariyanti, S.Si., M.T selaku dosen pembimbing II, yang dengan sabar dan ikhlas membimbing dan memberikan ilmunya kepada penulis. Penulis juga berterima kasih kepada keluarga dan teman-teman S1 Sistem Informasi, yang telah memberikan semangat dan dukungannya kepada penulis.

Penulis mengharapkan kritik dan saran yang bersifat membangun demi kesempurnaan proposal ini. Semoga skripsi ini dapat memberikan manfaat dan wawasan yang berguna. Amin.

Surabaya, Januari 2015

Penulis

UCAPAN TERIMA KASIH

Syukur Alhamdulillah kehadirat Allah SWT karena hanya dengan rahmat dan hidayah-Nya, sehingga penulisan naskah skripsi dengan judul **SISTEM DETEKSI BAHASA MENGGUNAKAN *N-GRAM* UNTUK DOKUMEN BAHASA INDONESIA DAN BAHASA INGGRIS** ini dapat diselesaikan dengan baik.

Dalam pelaksanaan dan penyusunan skripsi ini, banyak menemui kendala, namun atas bantuan dari berbagai pihak, akhirnya laporan ini bisa diselesaikan. Maka tidak lupa penulis mengucapkan banyak terima kasih kepada :

1. Allah SWT yang senantiasa memberikan rahmat, hidayah, dan karunia-Nya.
2. Rasulullah Muhammad SAW yang selalu menjadi panutan dan suri tauladan terbaik dalam kehidupan penulis sehingga penulisan skripsi ini dapat terselesaikan dengan baik..
3. Kedua orang tua dan adik tercinta yang tidak pernah lelah memberi dukungan materil maupun inmateril, doa, semangat, kepercayaan dan kasih sayang kepada penulis dari awal hingga dapat menyelesaikan penulisan skripsi ini.
4. Badrus Zaman, S.kom., M.Cs selaku dosen pembimbing I yang senantiasa, membimbing, membantu dan memberikan ilmu selama masa penulisan hingga skripsi ini terselesaikan.

5. Eva Hariyanti S.Si., MT selaku dosen pembimbing II yang telah membimbing, membantu dan berbagi ilmu dalam penulisan skripsi ini.
6. Taufik, S.T, M.Kom selaku dosen wali yang dengan sabar memahami, membimbing, membantu dan memberi pengarahan selama masa perkuliahan.
7. Farid Azhari, Gatut Rachmadi, Mega Gunawan, Teguh Yuwono atas bantuan, motivasi dan tim sukses dalam pengerjaan skripsi.
8. Seluruh keluarga besar yang ada di ketapang atas motivasi dan dukungan yang diberikan selama masa perkuliahan.
9. Seluruh Dosen program studi sistem informasi, yang telah banyak memberikan ilmunya, sehingga penulisan skripsi ini dapat terselesaikan dengan baik.
10. Teman – teman keluarga besar S1 sistem informasi yang telah banyak membantu dalam berbagi ilmu dan pengalaman selama penulisan skripsi.
11. *Staff* TU program studi sistem informasi yang telah banyak membantu dalam keperluan administrasi.
12. Serta teman – teman dan pihak lain yang telah memberikan bantuan dan informasi dalam proses penulisan skripsi ini.

Penulis

DAFTAR ISI

Halaman

HALAMAN JUDUL i
LEMBAR PERNYATAAN ii
LEMBAR PENGESAHAN iii
LEMBAR PEDOMAN PENGGUNAAN SKRIPSI iv
KATA PENGANTAR v
UCAPAN TERIMA KASIH vi
ABSTRAK viii
ABSTRACT ix
DAFTAR ISI x
DAFTAR GAMBAR xii
DAFTAR GRAFIK xiii
DAFTAR TABEL xiv
DAFTAR LAMPIRAN xvi

BAB I PENDAHULUAN 1
1.1 Latar Belakang 1
1.2 Rumusan Masalah 3
1.3 Tujuan 4
1.4 Manfaat 4
1.5 Batasan Masalah 4

BAB II TINJAUAN PUSTAKA 5
2.1 *Information Retrieval* 5
2.2 Deteksi Bahasa 5
2.3 Metode Deteksi Bahasa 6
2.4 *N-Gram* 7
2.5 Profil frekuensi *N-gram* 8

2.6 *Recall, precision* dan *F-measure* 11

BAB III METODE PENELITIAN 13
3.1 Waktu Penelitian 13
3.2 Objek Penelitian 13
3.3 Studi Kepustakaan 14
3.4 Pengumpulan Data dan Informasi 14
3.5 Teknik Pengolahan Data 14
3.6 Perancangan Sistem 15
3.7 Pembangunan Sistem 20
3.8 Evaluasi Sistem 20

BAB IV HASIL DAN PEMBAHASAN 21
4.1 Pengumpulan Data 21
4.2 Pengolahan Data 21
4.3 Analisis Sistem 22
4.4 Perancangan Sistem 47
4.5 Pembangunan Sistem 51
4.6 Uji Coba dan Evaluasi Sistem 57

BAB V KESIMPULAN DAN SARAN 72
5.1 Kesimpulan 72
5.2 Saran 73

DAFTAR PUSTAKA 74

DAFTAR GAMBAR

Nomor **Judul Gambar** **Halaman**

2.1 Gambaran Umum Proses Deteksi Bahasa Metode *N-gram* 9
4.1 *Pseudocode* Ekstraksi *N-Gram* 48
4.2 *Pseudocode* Proses menghitung profil frekuensi *unigram* 49
4.3 *Pseudocode* Proses menghitung profil frekuensi *bigram* dan *trigram* 49
4.4 *Pseudocode* menentukan kecenderungan bahasa dokumen 50
4.5 *Pseudocode* Proses membandingkan dengan *threshold* 51
4.6 Halaman *Training* 53
4.7 Halaman *Testing* 54
4.8 Halaman Data 54
4.9 Halaman Data *Train* 55
4.10 Halaman Data *Test* 56
4.11 Halaman *Threshold* 56

DAFTAR GRAFIK

Nomor	Judul Grafik	Halaman
4.1	Nilai *unigram* dokumen bahasa Indonesia	62
4.2	Nilai *unigram* dokumen bahasa Inggris	63
4.3	Nilai *unigram* dokumen bahasa Lainnya	64
4.4	Nilai *bigram* dokumen bahasa Indonesia	64
4.5	Nilai *bigram* dokumen bahasa Inggris	65
4.6	Nilai *bigram* dokumen bahasa Lainnya	66
4.7	Nilai *trigram* dokumen bahasa Indonesia	66
4.8	Nilai *trigram* dokumen bahasa Inggris	67
4.9	Nilai *trigram* dokumen bahasa Lainnya	68

DAFTAR TABEL

Nomor	Judul Tabel	Halaman
2.1	*Contingency table*	12
4.1	Tabel Rekapitulasi dokumen	22
4.2	Tabel Ektraksi *N-gram* kalimat bahasa Indonesia dokumen biologi	24
4.3	Tabel Ektraksi *N-gram* kalimat bahasa Indonesia dokumen sepakbola	25
4.4	Tabel Data *training set* untuk bahasa Indonesia dari 2 kalimat	26
4.5	Tabel jumlah frekuensi kemunculan *unigram* kalimat yang diuji	28
4.6	Tabel jumlah frekuensi kemunculan *bigram* kalimat yang diuji	28
4.7	Tabel jumlah frekuensi kemunculan *trigram* kalimat yang diuji	29
4.8	Tabel jumlah frekuensi kemunculan *unigram* kalimat Data *training set* bahasa Indonesia	32
4.9	Tabel jumlah frekuensi kemunculan *bigram* kalimat Data *training set* bahasa Indonesia	32
4.10	Tabel jumlah frekuensi kemunculan *trigram* kalimat Data *training set* bahasa Indonesia	33
4.11	Tabel jumlah frekuensi kemunculan *unigram* kalimat Data *training set* bahasa Inggris	34
4.12	Tabel jumlah frekuensi kemunculan *bigram* kalimat Data *training set* bahasa Inggris	34
4.13	Tabel jumlah frekuensi kemunculan *trigram* kalimat Data *training set* bahasa Inggris	35
4.14	Tabel Hasil perhitungan nilai simQ dan h(d) kalimat yang diuji dengan data *training set* bahasa Indonesia dan bahasa Inggris	42
4.15	Tabel Daftar dokumen percobaan pertama bahasa Indonesia	42
4.16	Tabel Daftar dokumen percobaan kedua bahasa Indonesia	44
4.17	Tabel Nilai *Threshold* bahasa Indonesia yang didapatkan dari percobaan	45
4.18	Tabel Nilai *Threshold*	48

4.19 Tabel Hasil pengujian dokumen bahasa Indonesia 59
4.20 Tabel Hasil pengujian dokumen bahasa Inggris 60
4.21 Tabel Hasil pengujian dokumen bahasa Lainnya 61
4.22 Tabel Kontingensi *unigram* 69
4.23 Tabel Kontingensi *bigram* 69
4.24 Tabel Kontingensi *trigram* 70

Daftar Lampiran

Nomor **Judul Lampiran**

1. Daftar dokumen *training* bahasa Indonesia
2. Daftar dokumen *training* bahasa Inggris
3. Daftar dokumen *testing* bahasa Indonesia
4. Daftar dokumen *testing* bahasa Inggris
5. Daftar dokumen *testing* bahasa Lainnya