IMPLEMENTASI ROCCHIO'S CLASSIFICATION DALAM MENGKATEGORIKAN RENUNGAN HARIAN KRISTEN

Skripsi

oleh
ELISABETH ADELIA WIDJOJO
22094680

PROGRAM STUDI TEKNIK INFORMATIKA FAKULTAS TEKNOLOGI INFORMASI
UNIVERSITAS KRISTEN DUTA WACANA
2013

IMPLEMENTASI ROCCHIO'S CLASSIFICATION DALAM MENGKATEGORIKAN RENUNGAN HARIAN KRISTEN

Skripsi

Diajukan kepada Program Studi Teknik Informatika Fakultas Teknologi Informasi
Universitas Kristen Duta Wacana
Sebagai Salah Satu Syarat dalam Memperoleh Gelar
Sarjana Komputer

Disusun oleh

ELISABETH ADELIA WIDJOJO
22094680

PROGRAM STUDI TEKNIK INFORMATIKA FAKULTAS TEKNOLOGI INFORMASI
UNIVERSITAS KRISTEN DUTA WACANA
2013

PERNYATAAN KEASLIAN SKRIPSI

Saya menyatakan dengan sesungguhnya bahwa skripsi dengan judul:

IMPLEMENTASI ROCCHIO'S CLASSIFICATION DALAM MENGKATEGORIKAN RENUNGAN HARIAN KRISTEN

yang saya kerjakan untuk melengkapi sebagian persyaratan menjadi Sarjana Komputer pada pendidikan Sarjana Program Studi Teknik Informatika Fakultas Teknologi Informasi Universitas Kristen Duta Wacana, bukan merupakan tiruan atau duplikasi dari skripsi kesarjanaan di lingkungan Universitas Kristen Duta Wacana maupun di Perguruan Tinggi atau instansi manapun, kecuali bagian yang sumber informasinya dicantumkan sebagaimana mestinya.

Jika dikemudian hari didapati bahwa hasil skripsi ini adalah hasil plagiasi atau tiruan dari skripsi lain, saya bersedia dikenai sanksi yakni pencabutan gelar kesarjanaan saya.

Yogyakarta, 15 Januari 2013

ELISABETH ADELIA WIDJOJO
22094680

iii

HALAMAN PERSETUJUAN

Judul Skripsi	:	IMPLEMENTASI ROCCHIO'S CLASSIFICATION DALAM MENGKATEGORIKAN RENUNGAN HARIAN KRISTEN
Nama Mahasiswa	:	ELISABETH ADELIA WIDJOJO
N I M	:	22094680
Matakuliah	:	Skripsi (Tugas Akhir)
Kode	:	TIW276
Semester	:	Gasal
Tahun Akademik	:	2012/2013

Telah diperiksa dan disetujui di
Yogyakarta,
Pada tanggal 15 Januari 2013

Dosen Pembimbing I	Dosen Pembimbing II
Antonius Rachmat C., SKom.,M.Cs	Drs. R. Gunawan Santosa, M.Si.

HALAMAN PENGESAHAN

IMPLEMENTASI ROCCHIO'S CLASSIFICATION DALAM MENGKATEGORIKAN RENUNGAN HARIAN KRISTEN

Oleh: ELISABETH ADELIA WIDJOJO / 22094680

Dipertahankan di depan Dewan Penguji Skripsi
Program Studi Teknik Informatika Fakultas Teknologi Informasi
Universitas Kristen Duta Wacana - Yogyakarta
Dan dinyatakan diterima untuk memenuhi salah satu syarat memperoleh gelar
Sarjana Komputer
pada tanggal 10 Januari 2013

Yogyakarta, 15 Januari 2013
Mengesahkan,

Dewan Penguji:

1. Antonius Rachmat C., SKom.,M.Cs
2. Drs. R. Gunawan Santosa, M.Si.
3. Joko Purwadi, M.Kom
4. Lukas Chrisantyo, M.Eng.

Dekan

(Drs. Wimmie Handiwidjojo, MIT.)

Ketua Program Studi

(Nugroho Agus Haryono, M.Si)

v

UCAPAN TERIMA KASIH

Puji syukur penulis panjatkan ke hadirat Tuhan Yesus Kristus yang telah melimpahkan rahmat dan anugerah, sehingga penulis dapat menyelesaikan skripsi yang berjudul "Implementasi Rocchio's Classification dalam Mengkategorikan Renungan Harian Kristen" ini dengan tepat waktu.

Dalam menyelesaikan penelitian ini, penulis menyadari banyak menerima masukan dan saran dari berbagai pihak, baik secara langsung maupun secara tidak langsung. Oleh karena itu, pada kesempatan ini penulis ingin menyampaikan terima kasih kepada:

1. Antonius Rachmat C, S.Kom., M.Cs. dan Drs. R. Gunawan Santosa, M.Si., selaku dosen pembimbing yang telah banyak membimbing penulis dalam menyelesaikan penelitian ini.
2. Budi Susanto, S.Kom., M.T. selaku koordinator Tugas Akhir.
3. Keluarga terkasih yang selalu mendukung dan memberikan motivasi kepada penulis selama ini.
5. Semua teman dan pihak lain yang telah mendukung penulis dalam penyelesaian penelitian ini yang tidak dapat penulis sebutkan satu per satu.

Akhir kata penulis ingin meminta maaf bila ada kesalahan dalam penyusunan laporan ini. Terima kasih.

Yogyakarta, 15 Januari 2013

Penulis

INTISARI

IMPLEMENTASI ROCCHIO'S CLASSIFICATION DALAM MENGKATEGORIKAN RENUNGAN HARIAN KRISTEN

Pada masa sekarang, kebanyakan gereja atau lembaga-lembaga Kristen mulai menggunakan media digital untuk penyimpanan data, seperti teks, gambar, suara, maupun video rohani. Bahkan renungan harian Kristen yang biasanya berupa buku pun sekarang sudah banyak dipublikasikan dalam media digital melalui internet. Dengan semakin banyaknya renungan Kristen yang dipublikasikan di internet, akan sangat sulit untuk menemukan renungan yang sesuai dengan topik yang diinginkan karena data-data renungan yang ada masih belum dikelompokkan.

Untuk mempermudah pengelompokkan tersebut, pada penelitian ini penulis menggunakan *Rocchio's classification*, yang menekankan penggunaan *tf-idf weighting* untuk pembobotan token yang digunakan untuk proses klasifikasi, serta perhitungan *prototype vector* (*centroid*) setiap kategori. Setiap dokumen uji akan diuji kemiripannya dengan setiap *centroid*. Dalam penelitian ini penulis melakukan beberapa pengujian untuk mengetahui pada kondisi apa keakuratan sistem dicapai.

Sistem klasifikasi yang dibangun dapat memberikan akurasi cukup tinggi pada saat *feature selection* 20% yaitu sebesar 73,33%. Nilai *precision* tertinggi jatuh pada kategori hikmat dengan nilai *precision* 1 dalam semua *feature selection*. Sedangkan nilai *recall* tertinggi jatuh pada kategori motivator dengan nilai *recall* 1 dalam semua *feature selection*.

Kata kunci : klasifikasi, pengkategorian, renungan, Rocchio, *centroid*, *similarity*

DAFTAR ISI

DAFTAR TABEL

DAFTAR GAMBAR

DAFTAR LISTING

DAFTAR GRAFIK

BAB 1
PENDAHULUAN

1.1 Latar Belakang Masalah

Pada masa sekarang, kebanyakan gereja atau lembaga-lembaga Kristen mulai menggunakan media digital untuk penyimpanan data, seperti teks, gambar, suara, maupun video rohani. Bahkan renungan harian Kristen yang biasanya berupa buku pun sekarang sudah banyak dipublikasikan dalam media digital melalui internet. Orang seringkali mencari renungan dengan topik tertentu sebagai bahan khotbah atau mungkin sekedar menguatkan hati dalam menghadapi pergumulan hidup. Dengan semakin banyaknya renungan Kristen yang dipublikasikan di internet, akan sangat sulit untuk menemukan renungan yang sesuai dengan topik yang diinginkan karena data-data renungan yang ada masih belum dikelompokkan.

Berbicara mengenai pengelompokkan data, dalam hal ini berupa dokumen teks, ada banyak metode klasifikasi yang bisa diterapkan antara lain Naïve Bayes, k-Nearest Neighbor, Decision Tree, Support Vector Machines, dan lain-lain. Salah satu teknik yang umum digunakan dalam pemrosesan dokumen teks adalah dengan pembobotan kata-kata yang dianggap penting dalam dokumen, yaitu dengan menghitung *term-frequency* (tf). Akan tetapi, jumlah dokumen di mana sebuah token/kata muncul (df) juga harus dipertimbangkan untuk melihat seberapa penting suatu token dalam dokumen tersebut (*scarcity of tokens*). Dalam hal ini akan dihitung *inverse document frequency* untuk tiap token (idf). Dari tf dan idf, kita akan mendapatkan *tf-idf weighting* (Robertson, 2005), yaitu suatu formula untuk menghitung bobot hubungan suatu token di dalam dokumen.

Oleh karena itu, pada penelitian ini penulis menggunakan *Rocchio's classification*, yang menekankan penggunaan *tf-idf weighting* untuk pembobotan token yang digunakan untuk klasifikasi dokumen. Klasifikasi dilakukan sesuai dengan algoritma Rocchio yang menggunakan *prototype vector* (*centroid*) untuk

membantu klasifikasi/pengelompokkan renungan harian. Dengan adanya penelitian ini, maka dapat membantu siapa saja yang ingin mencari renungan harian sesuai dengan topik yang diinginkan.

1.2 Perumusan Masalah

Rumusan masalah yang dapat dibuat dalam tugas akhir ini adalah:

- Seberapa akuratkah *Rocchio's classification* dalam mengkategorikan renungan harian Kristen?
- Kategori manakah yang memiliki nilai precision dan recall yang paling tinggi?
- Apakah pemilihan *feature selection* mempengaruhi nilai precision dan recall pada masing-masing kategori?
- Sumber data manakah yang memberikan nilai precision dan recall yang paling tinggi?

1.3 Batasan Masalah

Batasan-batasan masalah dalam pembuatan aplikasi ini adalah:

- Sistem yang dibangun adalah berupa aplikasi desktop
- Sumber data dari renungan harian Kristen diambil dari :
  - http://www.pelitahidup.com/
  - http://www.renungan-spirit.com/renungan-kristen.html
  - http://gkysydney.org/renungan-gema-2012/index.php
  - http://www.sabdaharian.com
  - http://www.bethanygraha.org
  - http://renungan-harian-online.com
  - Renungan Harian Air Hidup
  - http://renungan-harian-online.com
  - http://www.gkpi.or.id/renungan/

- Renungan harian Kristen akan dibagi dalam 4 kategori yaitu : berkat, iman, hikmat, dan motivator.
- Tidak ada tahap *stemming* dalam tahap *preprocessing* dokumen.
- *Stoplist* yang digunakan adalah data *stopwords* yang bersumber dari http://www.illc.uva.nl/Research/Reports/MoL-2003-02.text.pdf, kata kunci *database* dan tanda baca.
- Dokumen yang digunakan berekstensi .txt dimana :
  - nama dokumen mewakili kode kategori dan judul dari renungan harian Kristen
  - isi dokumen merupakan isi dari satu renungan harian Kristen
- Menggunakan pembobotan TF/IDF untuk merepresentasikan dokumen teks

1.4 Tujuan Penelitian

Tujuan dari penelitian ini adalah melihat keakuratan *Rocchio's classification* dalam mengkategorikan renungan harian Kristen, khususnya dalam topik berkat, iman, hikmat, dan motivator.

1.5 Metode Penelitian

Metodologi yang akan digunakan dalam pembuatan aplikasi ini adalah dengan studi pustaka dan pengumpulan data. Studi pustaka dilakukan dengan mempelajari teori-teori melalui buku, artikel, jurnal, dan bahan lain yang mendukung yang berhubungan dengan klasifikasi dokumen. Pengumpulan data dilakukan dengan mencari referensi resmi (sumber data) untuk setiap renungan harian yang akan digunakan.

1.6 Sistematika Penulisan

Penulisan laporan tugas akhir ini dibagi menjadi lima bab yaitu:

Bab 1, Pendahuluan, yang memberikan gambaran umum mengenai apa yang diteliti dalam tugas akhir ini. Pendahuluan berisi latar belakang masalah, perumusan masalah, batasan masalah, tujuan penellitian, metode penelitian, dan sistematika penulisan.

Bab 2, Tinjauan Pustaka, yang terdiri dari dua bagian utama, yakni tinjauan pustaka dan landasan teori. Tinjauan pustaka menguraikan berbagai teori mengenai klasifikasi dengan metode *Rocchio's classification* yang didapatkan dari berbagai sumber pustaka yang digunakan untuk penyusunan tugas akhir. Landasan teori memuat penjelasan tentang konsep dan prinsip utama yang diperlukan untuk memecahkan masalah dalam penelitian. Hanya penjelasan yang berhubungan dengan penelitian yang dilakukan yang akan dicantumkan di sini.

Bab 3, Analisis dan Perancangan Sistem, yang mencakup perancangan sistem yang akan dibuat, yakni mengenai kebutuhan *hardware* dan *software*, spesifikasi sistem, arsitektur sistem, *use case diagram*, algoritma yang digunakan, skema *database*, dan rancangan antarmuka sistem.

Bab 4, Implementasi dan Analisis Sistem, yang memuat hasil implementasi dan pembahasan mengenai pengujian sistem yang dibuat berdasarkan bab 3, beserta hasil dari sistem yang dijalankan dan analisis dari sistem yang dibuat.

Bab 5, Kesimpulan dan Saran, berisi kesimpulan dari hasil penelitian yang didapatkan dan saran untuk memberikan analisis dan pengembangan yang lebih baik lagi pada penelitian ke depannya dalam topik yang serupa.

BAB 5
KESIMPULAN DAN SARAN

5.1 Kesimpulan

Berdasarkan hasil implementasi dan analisis sistem, maka diperoleh kesimpulan sebagai berikut :

1. Sistem klasifikasi Rocchio memberikan akurasi cukup tinggi untuk *feature selection* 20% yaitu sebesar 73,33%.
2. Nilai precision tertinggi jatuh pada kategori hikmat dengan nilai precision 1 dalam semua *feature selection*. Sedangkan nilai recall tertinggi jatuh pada kategori motivator dengan nilai recall 1 dalam semua *feature selection*.
3. Peningkatan *feature selection* tidak terlalu mempengaruhi nilai precision dan recall pada setiap kategori.
4. Sumber data Renungan Harian Spirit sangat cocok untuk kategori motivator karena memiliki nilai precision dan recall 1 dari total 13 renungan harian yang diambil.

5.2 Saran

Saran yang diajukan penulis untuk perbaikan dan pengembangan sistem adalah sebagai berikut :

1. Diperlukan penggunaan *store procedure* pada VB.NET untuk mempercepat *preprocessing* data dan mengurangi penggunaan memori.
2. Dapat ditambahkan proses *stemming* dalam bahasa Indonesia untuk lebih meningkatkan akurasi sistem.

DAFTAR PUSTAKA

Auvil, Loretta, et all. (2007). VAST to Knowledge : Combining Tools for Exploration and Mining. *IEEE VAST*, 197-198. Diakses pada tanggal 1 September 2012 dari http://webcache.googleusercontent.com/search?q=cache:http://vac.nist.gov/2007/summaries/auvil.pdf

Kurniawan, Erick. (2011). *Cepat Mahir Visual Basic 2010*. Yogyakarta : Penerbit Andi.

Koster, Cornelis H. A. (2003). *Chapter 2 : Document Classification*. Diakses pada tanggal 17 Mei 2012 dari http://www.cs.ru.nl/~kees/ir2/papers/h03.pdf

Feldman, R., dan Sanger, J. (2007). *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data. Cambridge* : Cambridge University Press.

Grossman, David A. dan Frieder, Ophir. (2004). *Information Retrieval Algorithms dan Heuristics, 2nd edition*. New York : Springer.

Han, J. & Kamber, M. (2006). *Data Mining : Concepts and Technique 2nd Edition*. San Fransisco : Morgon Kauffman Publishers.

Intan, Rolly & Defeng, Andrew. (2006). *HARD : Subject Based Search Engine Menggunakan TF-IDF dan Jaccard's Coeffisient*. Diakses pada tanggal 25 Agustus 2012 dari http://puslit.petra.ac.id/files/published/journals/IND/IND060801/IND06080106.pdf

Joachims, T. (1997). *A Probabilistic Analysis of the Rocchio Algorithm with TF/IDF for Text Categorization*. Diakses pada tanggal 20 Oktober 2012 dari http://www.cs.cornell.edu/People/tj/publications/joachims_97a.ps.gz

Manning, Christopher D., et all. (2008). *Introduction to Information Retrieval*. New York : Cambridge University Press.

Salton, G, et all. (1975). A Vector Space Model for Automatic Indexing. *Communications of the ACM, 12*, 613-620. Diakses pada tanggal 1 September 2012 dari http://www.cs.uiuc.edu/class/fa05/cs511/Spring05/other_papers/p613-salton.pdf

Uden, Mark Van. (n.d). *Rocchio : Relevance Feedback in Learning Classification Algorithms*. Diakses pada tanggal 23 Agustus 2012 dari http://www.cs.kun.nl/nsccs/artikelen/markuden.ps.Z.

Weiss, Sholom M., et all. (2005). *Text Mining : Predictive Methods for Analyzing Unstructured Information*. New York : Springer