PERHITUNGAN TF-IDF PADA DATASET SENTIPOL

Skripsi

oleh

ANDAR SETIAWAN POLE

22104906

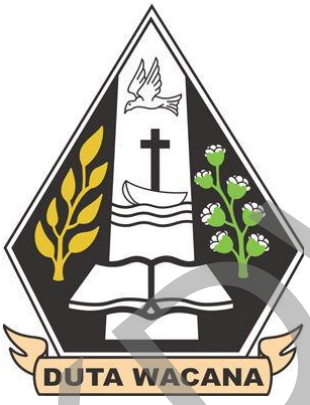
PROGRAM STUDI INFORMATIKA FAKULTAS TEKNOLOGI INFORMASI

UNIVERSITAS KRISTEN DUTA WACANA

2018

PERHITUNGAN TF-IDF PADA DATASET SENTIPOL

Skripsi



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

oleh

ANDAR SETIAWAN POLE

22104906

PROGRAM STUDI INFORMATIKA FAKULTAS TEKNOLOGI INFORMASI

UNIVERSITAS KRISTEN DUTA WACANA

2018

PERNYATAAN KEASLIAN SKRIPSI

Saya menyatakan dengan sesungguhnya bahwa skripsi dengan judul:

PERHITUNGAN TF-IDF PADA DATASET SENTIPOL

yang saya kerjakan untuk melengkapi sebagian persyaratan menjadi Sarjana Komputer pada pendidikan Sarjana Program Studi 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, 16 Januari 2018

ANDAR SETIAWAN POLE
22104906

HALAMAN PERSETUJUAN

Judul Skripsi	:	PERHITUNGAN TF-IDF PADA DATASET SENTIPOL
Nama Mahasiswa	:	ANDAR SETIAWAN POLE
N I M	:	22104906
Matakuliah	:	Skripsi (Tugas Akhir)
Kode	:	TIW276
Semester	:	Gasal
Tahun Akademik	:	2017/2018

Telah diperiksa dan disetujui di
Yogyakarta,
Pada tanggal 16 Januari 2018

Dosen Pembimbing I	Dosen Pembimbing II
Antonius Rachmat C., S.Kom.,M.Cs.	Yuan Lukito, S.Kom., M.Cs.

iv

HALAMAN PENGESAHAN

PERHITUNGAN TF-IDF PADA DATASET SENTIPOL

Oleh: ANDAR SETIAWAN POLE / 22104906

Dipertahankan di depan Dewan Penguji Skripsi
Program Studi Informatika Fakultas Teknologi Informasi
Universitas Kristen Duta Wacana - Yogyakarta
Dan dinyatakan diterima untuk memenuhi salah satu syarat memperoleh gelar
Sarjana Komputer
pada tanggal 21 Desember 2017

Yogyakarta, 16 Januari 2018
Mengesahkan,

Dewan Penguji:

1. Antonius Rachmat C., S.Kom.,M.Cs.
2. Yuan Lukito, S.Kom., M.Cs.
3. R. Gunawan Santosa, Drs. M.Si.
4. Danny Sebastian, S.Kom., M.M., M.T.

Dekan	Ketua Program Studi
(Budi Susanto, S.Kom., M.T.)	(Gloria Virginia, Ph.D.)

v

KATA PENGANTAR

Puji syukur dipanjatkan kepada Tuhan Yang Maha Esa, karena atas berkat karunia-Nya penulis mampu menyelesaikan skripsi dengan juduul ":Klasifikasi Sentimen Dataset PEMILU menggunakan Algoritma Rocchio" sebagai salah satu salah satu syarat untuk memperoleh gelar sarjana pada Program Studi Teknik Informatika Universitas Kristen Duta Wacana Yogyakarta.

Selesainya skripsi ini tidak dapat terlepas dari campur tangan dari berbagai pihak yang tidak henti-hentinya memberikan dukungan. Oleh sebab itu, penulis endak menyampaikan terima kasih kepada

1. Bapak Antonius Rachmat C., S.Kom., M.Cs, selaku koordinator Tugas Akhir dan juga Dosen Pembimbing 1, dan Bapak Yuan Lukito, S.Kom., M.Cs, selaku Dosen Pembimbing 2 yang telah memberikan bimbingan, arahan, masukan, terlebih kesabaran atas segala keterbatasan penulis dalam menyusun skripsi ini.
2. Dekan Fakultas Teknologi Informasi, Ketua Program Studi Teknik Informatika, Wakil Dekan, Dosen, dan Staf Pendukung Akademik di Prodi TI yang telah memberikan bantuan dalam semua proses akademik, dan administrasi selama masa perkuliahan
3. Kepada Ishak Pole, Elna Pelowe, dan Abdi Gunawan Pole yang terus memberikan doa dan dukungan
4. Kepada Deraya Sandika Ratri yang terus memberikan dukungan doa
5. Kepada pihak-pihak yang telah memberikan doa, dukungan, dan semangat untuk menyelesaikan proses skripsi ini:
    a. Teman-teman seperjuangan di proses, menyemangati, dan berbagi selama proses pengerjaan skripsi ini : Adi Atmaja, David Pande, Leo Ramses, Andreas, Karel Kalpito, Yulius Lolos, Ezra.

b. Sahabat-sahabat yang senantiasa rela membantu dalam penyelesaian skripsi ini: Bramasti Pramudyawardani, Yusuf Hari, Kukuh Aldyanto
c. Sahabat-sahabat angkatan 2010 yang selalu memberikan doa dan dukungan dan percaya bahwa penulis mampu menyelesaikan proses ini.
d. Keluarga besar GAPPALA yang telah menjadi rumah dimana penulis menempa karakter diri dan selalu memberikan dukungan, doa, dan menjaga kesehatan penulis selama menyusun skripsi ini
e. Keluarga besar Paduan Suara Mahasiswa "Duta Voice" yang menjadi tempat penulis bertumbuh sejak awal masuk kuliah dan selalu memberikan dukungan, doa untuk penulis.
f. Keluarga besar Staf Fakultas Kedokteran UKDW yang selalu memberikan dukungan, doa untuk penulis

6. Semua pihak yang telah membantu dalam keseluruhan proses yang tidak dapat disebutkan satu-persatu

Penulis menyadari bahwa di waktu yang sangat singkat dalam proses ini, masih ada kekurangan dalam skripsi ini sebagai cerminan keterbatasan diri penulis. Oleh sebab itu, penulis mengharapkan kritik dan saran agar skripsi ini bisa berguna bagi semua pihak.

Yogyakarta, 5 Desember 2017

Andar Setiawan Pole
NIM: 22104906

INTISARI

PERHITUNGAN TF-IDF PADA DATASET SENTIPOL

Saat ini, media sosial cukup mengambil peran dalam kegiatan PEMILU tahun 2014. Setiap kandidat aktif menggunakan media sosial sebagai media kampanye dan para konstituen juga aktif komentar dukungan ataupun kritik mereka bagi para kandidat.

Dalam penelitian ini akan dibangun sistem yang mampu menghitung bobot kata terhadap sebuah dokumen dengan menggunakan metode perhitungan TF-IDF. Sistem akan menghitung bobot kata terhadap komentar yang sudah baku dengan memanfaatkan hasil *text preprocessing* dan *text transformation*. Data penelitian ini akan menggunakan hasil dari penelitian Rachmat dan Lukito (2016) yaitu dataset SentiPol. Dataset SentiPol dikumpulkan dari status dan komentar terhadap calon presiden Indonesia pada masa kampanye pemilu tahun 2014 dari halaman *facebook.*

Hasil implementasi penghitungan yang dibuat oleh penulis dapat menghasilkan nilai-nilai bobot setiap token terhadap suatu dokumen.

Kata Kunci : Pembobotan, TF-IDF, Sentimen, PEMILU, SentiPol

DAFTAR ISI

DAFTAR GAMBAR

DAFTAR TABEL

DAFTAR GRAFIK

INTISARI

PERHITUNGAN TF-IDF PADA DATASET SENTIPOL

Saat ini, media sosial cukup mengambil peran dalam kegiatan PEMILU tahun 2014. Setiap kandidat aktif menggunakan media sosial sebagai media kampanye dan para konstituen juga aktif komentar dukungan ataupun kritik mereka bagi para kandidat.

Dalam penelitian ini akan dibangun sistem yang mampu menghitung bobot kata terhadap sebuah dokumen dengan menggunakan metode perhitungan TF-IDF. Sistem akan menghitung bobot kata terhadap komentar yang sudah baku dengan memanfaatkan hasil *text preprocessing* dan *text transformation*. Data penelitian ini akan menggunakan hasil dari penelitian Rachmat dan Lukito (2016) yaitu dataset SentiPol. Dataset SentiPol dikumpulkan dari status dan komentar terhadap calon presiden Indonesia pada masa kampanye pemilu tahun 2014 dari halaman *facebook.*

Hasil implementasi penghitungan yang dibuat oleh penulis dapat menghasilkan nilai-nilai bobot setiap token terhadap suatu dokumen.

Kata Kunci : Pembobotan, TF-IDF, Sentimen, PEMILU, SentiPol

BAB 1
PENDAHULUAN

1.1 Latar Belakang

Antusiasme masyarakat Indonesia terhadap dunia politik seperti kegiatan PEMILU (Pemilihan Umum) berkembang seiring dengan perkembangan media sosial, yang tidak hanya memberikan kemudahan dalam mendapatkan informasi, tetapi juga memberikan ruang yang luas bagi para penggunanya untuk berpendapat. Dalam masa kampanye PEMILU tahun 2014, tiap kandidat dan tim suksesnya aktif menggunakan akun resmi media sosial mereka sebagai media untuk menawarkan janji-janji dan program-program kepada para konstituen. Bagi para konstituen, media sosial digunakan tidak hanya sebagai media untuk mengetahui informasi, tapi juga sebagai media untuk memberikan opini mengenai kegiatan kampanye kandidat tertentu.

Dalam penelitian yang dilakukan Rachmat dan Lukito (2016), dilakukan pembangunan dataset yang dikumpulkan dari status dan komentar terhadap calon presiden Indonesia pada masa kampanye pemilu tahun 2014 dari halaman *facebook*, dan dinamakan dataset SentiPol. Penelitian ini menghasilkan dataset sejumlah 3400 komentar dari 68 status dalam format CSV dengan label positif lebih dominan dari lebel negatif dan netral sehingga dapat digunakan dalam pembelajaran sistem *supervised learning* lainnya.

Dalam penelitian ini akan dibangun sistem yang mampu menghitung bobot kata terhadap sebuah dokumen dengan menggunakan metode perhitungan TF-IDF. Melalui penelitian ini akan diketahui bobot setiap kata terhadap suatu dokumen yang diimplementasikan menggunakan dataset SentiPol. (Rachmat & Lukito, 2016)

1.2 Perumusan Masalah

Berdasarkan latar belakang yang telah dikemukakan, maka rumusan masalah yang dibahas dalam penelitian ini adalah bagaimana implementasi model pembobotan TF-IDF dalam memberikan nilai bobot kata terhadap suatu komentar berbahasa Indonesia dari dataset? (Rachmat & Lukito, 2016)

1.3 Batasan Masalah

Dalam penelitian ini, permasalahan dibatasi sebagai berikut:

1. Dokumen teks yang digunakan merupakan komentar dari dataset SentiPol (Rachmat & Lukito, 2016)
2. *Stoplist* yang digunakan adalah data *stopwords* yang diambil dari http://www.illc.uva.nl/Research/Reports/MoL-2003-02.text.pdf
3. *Library* yang digunakan untuk *stemming* Bahasa Indonesia diambil dari https://github.com/sastrawi/sastrawi

1.4 Tujuan Penelitian

Tujuan dari penilitian ini adalah untuk mengetahui model pembobotan TF-IDF dalam memberikan nilai bobot kata terhadap suatu komentar berbahasa Indonesia

1.5 Metode Penelitian

Metode penelitian yang akan digunakan adalah

1. Studi pustaka dan literatur

   Mencari literatur dan sumber pustaka yang berhubungan dengan *text preprocessing* dan pembobotan TF-IDF
2. Perancangan Sistem

Dilakukan perancangan sistem untuk melakukan *preprocessing* sesuai dengan kebutuhan dataset dan pembobotan berdasar metode yang sudah ditetapkan

3. Implementasi

   Tahap Implementasi berisi proses pengimplementasian rancangan *preprocessing* dan penetepan nilai bobot suatu kata menggunakan metode penghitungan TF-IDF.

4. Analisis Hasil Percobaan dan Evaluasi

   Berisi Analisa dan evaluasi dari program yang dibangun dalam pembobotan TF-IDF

1.6 Sistematika Penulisan

Secara garis besar, sistematika penulisan yang digunakan dalam penulisan ini terdiri dari 5 bab. Di setiap babnya berisi beberapa sub bab yang berguna untuk menunjang penjelasan pokok dalam tiap babnya. 5 bab tersebut antara lain sebagai berikut:

Bab 1 Pendahuluan, berisi tujuh bagian, yaitu Latar belakang, Perumusan, Batasan Sistem, Hipotesis, Tujuan Penelitian, Metode Penelitian, Sistematika Penulisan.

Bab 2 Landasan Teori, berisi dua bagian, yaitu Tinjauan Pustaka dan Landasan Teori

Bab 3 Analisis dan perancangan sistem, berisi rancangan pembuatan sistem berdasarkan analisa teori yang digunakan

Bab 4 Implementasi dan Analisis sistem, berisi hasil implementasi rancangan dan analisa sistem

Bab 5 Kesimpulan dan saran, berisi penjelasan singkat hasil perancangan dan analisa sistem, serta saran untuk pengembangan sistem untuk penelitian yang sejenis kedepannya.

BAB V
KESIMPULAN DAN SARAN

5.1 Keimpulan

1. Perhitungan TF-IDF pada dataset SentiPol berhasil memberikan nilai pada setiap kata di dalam dataset.
2. Pada proses *cleanse*, tingkat keberhasilan sistem dalam menghapus karakter non huruf pada data *training* 98,57% dan di data *testing* sebesar 98.52%.
3. Sistem perhitungan ini menghapus 9126 token pada data *training* dan 1860 token pada data *testing* karena tidak dianggap sebgai kata baku dalam bahasa Indonesia.

5.2 Saran

1. Perlu ditambahkan fungsi *spell checker* yang akan membantu proses pendeteksian kata asli dari kumpulan kata yang tidak baku agar tidak banyak kata yang harus dihapus
2. Dapat ditambahkan daftar kombinasi token dengan menggunakan perhitungan *bigram* agar daftar token semakin banyak dan hasil lebih optimal.
3. Hasil perhitungan TF-IDF ini dapat dikembangkan utnuk keperluan klasifikasi sentiment opini dengan menggunakan algoritma apapun yang memanfaatkan TF-IDF dalam proses pembobotannya.

DAFTAR PUSTAKA

Februariyanti, H., & Zuliarso, E. (2012). Klasifikasi Dokumen Berita Teks Bahasa Indonesia Menggunakan Ontologi. *Jurnal Teknologi Informasi DINAMIK Volume 17, No.1*, 14-23.

Feldman, R. (2007). *The Text Mining Handbook.* New York: Cambridge University Press.

Joachims, T. (1997). A Probabilistic Analyisis of the Rocchio Algorithm with TFIDF for Text Categorization. *International Conference on Machine Learning*.

Lumbanraja, F. R. (2013). Sistem Pencarian Data Teks dengan Menggunakan Metode Klasifikasi Rocchio(Studi Kasus:Dokumen Teks Skripsi). *Kumpulan Makalah Seminar Semirata* , 217-224.

Manning, C. D. (2009). *An Introduction to Information Retrieval.* Cambridge: Cambridge University Press.

Monarizqa, N., Nugroho, L. E., & Hantono, B. S. (2014). Penerapan Analisis Sentimen pada Twitter Berbahasa Indonesia Sebagai Pemberi Rating. *Jurnal Penelitian Teknik Elektro dan Teknologi Informasi*, 151 -155.

Naradhipa, A. R., & Purwarianti, A. (2012). Sentiment classification for Indonesian message in social media. *IEEE International Conference on Computational Intelligence and Cybernetics*.

Pramana, I. (2011). *Implementasi Metode Rocchio's Relevance Feedback Dalam Perangkingan Dokumen Text.* Yogyakarta: Universitas Kristen Duta Wacana.

Rachmat, A., & Lukito, Y. (2016). SENTIPOL: DATASET SENTIMEN KOMENTAR PADA KAMPANYE PEMILU PRESIDEN INDONESIA 2014 DARI FACEBOOK PAGE. *Buku Abstrak Konferensi Nasional Teknologi Informasi dan Komunikasi*, 37.

Sunni, I., & Widyantoro, D. H. (2012). Analisis Sentimen dan Ekstraksi Topik Penentu Sentimen pada Opini Terhadap Tokoh Publik. *Jurnal Sarjana Institut Teknologi Bandung Bidang Teknik Elektro dan Informatika*, 200-205.

Widjojo, E. A. (2013). *IMPLEMENTASI ROCCHIO'S CLASSIFICATION DALAM MENGKATEGORIKAN RENUNGAN HARIAN KRISTEN.* Yogyakarta: Universitas Kristen Duta Wacana.

Wilson, T., Wiebe, J., & Hoffman, P. (2005). Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis. *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language*, 347-354.

Wusana, N. Y. (2015). *KLASIFIKASI SENTIMEN PENGUNJUNG YOUTUBE BERDASAR KOMENTAR MENGGUNAKAN ALGORITMA ROCCHIO STUDI KASUS : VIDEO COVERING.* Yogyakarta: Universitas Kristen Duta Wacana.

Yugianus, P., Dachlan, H. S., & Hasanah, R. N. (2013). Pengembangan Sistem Penelusuran Katalog Perpustakaan Dengan Metode Rocchio Relevance Feedback. *Jurnal EECCIS Vol. 7*, 47 - 52.