

**PERBANDINGAN METODE *COSINE SIMILARITY* DAN *JARO-WINKLER*
DISTANCE PADA *INFORMATION RETRIEVAL SYSTEM* UNDANG-UNDANG
PEMILU**

Kusuma Gigih Prakoso

(180411100082)

Pembimbing 1 : Dr. Fika Hastarita Rachman, S.T., M.Eng.

Pembimbing 2 : Firdaus Solihin, S.Kom., M.Kom

ABSTRAK

Republik Indonesia memiliki banyak macam peraturan yang berlaku. Undang-undang yang berlaku di Indonesia memiliki banyak sekali jenis, serta ayat dan pasal yang tercantum di dalamnya. Penelitian ini bertujuan untuk mengembangkan sebuah *Information Retrieval System* yang dapat efektif dan efisien dalam melakukan pencarian informasi dalam *dataset* undang-undang pemilu. Sistem ini dirancang untuk membantu pengguna, seperti penegak hukum, pakar pemilu, dan masyarakat umum, dalam menemukan informasi yang relevan dengan cepat. Dalam penelitian ini, dikembangkan suatu *Information Retrieval System* menggunakan perbandingan metode *Cosine Similarity* dan *Jaro-Winkler Distance* yang diimplementasikan ke dalam pencarian isi undang-undang pada dokumen UU No 07 Tahun 2017 tentang Pemilu. Perbandingan antara dua metode yang dilakukan bertujuan untuk mengetahui metode apa yang lebih baik untuk digunakan pada penelitian kali ini. Dokumen yang digunakan sebagai *dataset* memiliki total 573 pasal dan 1.616 ayat didalamnya. Pada penelitian [7], *Information Retrieval System* yang dibuat menggunakan metode *Synonym Recognition* dan *Cosine Similarity*, dokumen yang digunakan oleh peneliti adalah pasal-pasal yang tercantum didalam KUHP. Nilai *precision* yang dihasilkan dalam penelitian [7] adalah 0.393. Dalam penelitian kali ini digunakan 5 contoh *query* untuk menguji *Information Retrieval System* pada metode *Cosine Similarity* dan *Information Retrieval System* pada metode *Jaro-Winkler Distance*. Pada proses penelitian kali ini, contoh *query* yang digunakan akan berkaitan dengan *dataset*. Hasil uji coba yang dilakukan menunjukkan bahwa metode *Cosine Similarity* lebih unggul yaitu menghasilkan rata-rata nilai *f-measure*, *precision*, dan *recall*, senilai 0.655462, 0.4875, dan 1 dibandingkan metode *Jaro-Winkler Distance* yang menghasilkan nilai 0.007617, 0.00382, dan 1. Hasil penelitian menunjukkan bahwa kedua metode memiliki efektivitas yang berbeda dalam mencari informasi yang relevan.

Kata kunci : *Information Retrieval System*, Undang-undang, *Cosine similarity*, *Jaro-Winkler Distance*

KATA PENGANTAR

Puji syukur kehadiran Allah SWT berkat limpahan rahmat dan hidayah-Nya, penulis mampu menuntaskan skripsi berjudul **“Perbandingan Metode *Cosine Similarity* dan *Jaro-Winkler Distance* pada *Information Retrieval System* Undang-Undang Pemilu”**. Penulis juga menyampaikan terimakasih yang sebesar-besarnya kepada:

1. Allah SWT atas limpahan Rahmat dan berkah-Nya bagi penulis hingga saat ini. Selain itu, Nabi Muhammad SAW adalah pemimpin umat manusia dari era jahiliyah ke era Islamiyah.
2. Kedua orang tua penulis yang selalu menyayangi dan mendukung penulis secara emosional dan materil.
3. Keluarga dan kerabat yang selalu mendukung penuh penulis untuk menyelesaikan studi. Semoga Allah SWT selalu memberikan perlindungan-Nya
4. Ibu Dr. Fika Hastarita Rachman, S.T., M.Eng sebagai dosen pembimbing 1 telah memberi masukan serta motivasi untuk penulis selama penelitian berlangsung
5. Bapak Firdaus Solihin, S.Kom., M.Kom sebagai dosen pembimbing 2 telah memberi kritik beserta saran untuk penulis selama penelitian berlangsung.
6. Bapak dan Ibu dosen Teknik Informatika UTM yang sudah berbagi pengetahuan mereka selama kuliah yang singkat ini.
7. Teman-teman seperjuangan Muhammad Syahrullah, Wildan Mubarak, Fikri Jamaludin, Julius Agung Saputro, dan Prita Dyah Anindhita. Terimakasih atas Upaya dan dorongannya selama ini.
8. Terimakasih pada Nur Aini yang selalu memberikan semangat agar penulis menyelesaikan studinya
9. Raka Pratama Putra, Zulvan Baihaqi, Hidayatur Rohman, Enggar Nur Wicaksana, Deden Nur Eka Abdi, dan Yusuf Al-Farizi yang selalu memberikan hiburan kepada penulis selama penelitian skripsi berlangsung
10. Seluruh keluarga besar HIMATIF UTM dan UKM-FT Tofatek yang tidak bisa penulis sebutkan satu-persatu. Terimakasih telah membersamai penulis dari awal hingga akhir perkuliahan.

Penulis sadar bahwa pasti terdapat kesalahan dalam pengerjaan skripsi ini. Maka dari itu, penulis memohon maaf atas segala kesalahan yang ada. Penulis, dengan segala kerendahan hati, menerima komentar dan saran dari pembaca untuk membantu memperbaiki penelitian di masa mendatang.

Bangkalan, 06 July 2023

Kusuma Gigih Prakoso

DAFTAR ISI

HALAMAN JUDUL.....	i
HALAMAN PENGESAHAN.....	ii
HALAMAN PERNYATAAN ORISINALITAS.....	iii
ABSTRAK.....	iv
KATA PENGANTAR.....	vi
DAFTAR ISI.....	viii
Daftar Gambar.....	xi
Daftar Tabel.....	xii
BAB I PENDAHULUAN.....	1
1.1 Latar Belakang.....	1
1.2 Rumusan Masalah.....	2
1.2.1 Permasalahan.....	2
1.2.2 Solusi Permasalahan.....	2
1.2.3 Pertanyaan Penelitian.....	2
1.3 Batasan Masalah.....	2
1.4 Tujuan dan Manfaat.....	2
1.4.1 Tujuan.....	3
1.4.2 Manfaat.....	3
1.5 Sistematika Proposal.....	3
1.6 Jadwal Penelitian.....	4
BAB II Tinjauan Pustaka.....	5
2.1 Landasan Teori.....	5
2.1.1 <i>Information Retrieval System</i>	5
2.1.2 <i>Text Preprocessing</i>	5
2.1.3 <i>Jaro-Winkler Distance</i>	6

2.1.4 Pembobotan <i>TF-IDF</i>	7
2.1.5 <i>Cosine similarity</i>	8
2.2 Penelitian Terdahulu.....	9
BAB III Metode Penelitian.....	13
3.1 Pengumpulan Dokumen.....	14
3.2 <i>Text Preprocessing</i>	14
3.2.1 <i>Case folding</i>	14
3.2.2 <i>Stemming</i>	15
3.2.3 <i>Tokenization</i>	16
3.3 Proses <i>Cosine similarity</i>	17
3.3.1 Perhitungan Nilai <i>Term Frequency</i> dan <i>Document Frequency</i>	17
3.3.2 Perhitungan Nilai IDF.....	18
3.3.3 Pembobotan <i>TF-IDF</i>	19
3.3.4 <i>Cosine similarity</i>	20
3.4 Proses <i>Jaro-Winkler Distance</i>	22
3.4.1 Perhitungan Nilai <i>Jaro</i>	22
3.4.3 Perhitungan Nilai <i>Jaro-Winkler Distance</i>	24
3.5 Evaluasi Sistem.....	25
3.6 Skenario Pengujian.....	25
BAB IV HASIL DAN PEMBAHASAN.....	27
4.1 Ruang Lingkup.....	27
4.2 Pembuatan <i>Database</i>	27
4.3 Preprocessing.....	33
4.3.1 <i>Case folding</i>	33
4.3.2 <i>Stemming</i>	33
4.3.3 <i>Tokenization</i>	33
4.4 Proses <i>Cosine similarity</i>	34

4.4.1 Menghitung Nilai <i>Vector TF-IDF</i> antara <i>Query</i> dengan Dokumen	34
4.4.2 Proses <i>Cosine similarity</i>	35
4.5 Implementasi Jaro-Winkler Distance.....	36
4.5.1 Memanggil Dokumen dalam <i>Database</i>	36
4.5.2 Proses <i>Jaro-Winkler Distance</i>	36
4.6 Evaluasi Sistem.....	37
4.7 Implementasi Antar Muka.....	38
4.7.1 Halaman Utama.....	38
4.7.2 Halaman Hasil.....	39
4.8 Hasil Pengujian.....	39
4.8.1 Hasil Skenario Uji Coba ke-1.....	40
4.8.2 Hasil Skenario Uji Coba ke-2.....	41
4.8.3 Hasil Skenario Uji Coba ke-3.....	42
4.8.4 Hasil Skenario Uji Coba 4.....	43
4.8.5 Hasil Skenario Uji Coba 5.....	43
4.9 Analisa Hasil Uji Coba.....	44
BAB V PENUTUP.....	45
5.1 Kesimpulan.....	45
5.2 Saran.....	45
Daftar Pustaka.....	46

Daftar Gambar

Gambar 3. 1 Arsitektur Sistem.....	13
Gambar 4. 1 Code txt to csv.....	28
Gambar 4. 2 Hasil konversi file txt to csv.....	28
Gambar 4. 3 Code preprocessing dokumen.....	29
Gambar 4. 4 Code untuk menginputkan dokumen ke database.....	30
Gambar 4. 5 Code untuk menghitung nilai TF-IDF.....	30

Gambar 4. 6 Hasil didalam database pada tabel dokumen.....	31
Gambar 4. 7 Hasil dalam database pada tabel dok_kata.....	31
Gambar 4. 8 Hasil dalam database pada tabel kata.....	32
Gambar 4. 9 Code Casefolding.....	33
Gambar 4. 10 Code Menghapus tanda baca dan angka.....	33
Gambar 4. 11 Code Stemming.....	33
Gambar 4. 12 Code Tokenization.....	33
Gambar 4. 13 Code memanggil nilai TF-IDF didalam database.....	34
Gambar 4. 14 Code untuk menghitung nilai vektor TF-IDF antara Query dengan Dokumen.....	34
Gambar 4. 15 Rumus Cosine similarity.....	35
Gambar 4. 16 Code Menghitung nilai Cosine similarity.....	35
Gambar 4. 17 Code memanggil dokumen didalam database.....	36
Gambar 4. 18 Code untuk menghitung Jaro-Winkler Distance.....	37
Gambar 4. 19 Code untuk memanggil dokumen.....	37
Gambar 4. 20 Tampilan halaman awal.....	38
Gambar 4. 21 Tampilan halaman hasil.....	39

Daftar Tabel

Tabel 1. 1 Jadwal Penelitian.....	4
Tabel 2. 1 Penelitian Terdahulu.....	11
Tabel 3. 1 Contoh Proses <i>Case folding</i>	14
Tabel 3. 2 Contoh Proses Stemming.....	16
Tabel 3. 3 Contoh Proses Tokenisasi.....	16
Tabel 3. 4 Contoh Query.....	17
Tabel 3. 5 Contoh Dokumen.....	17
Tabel 3. 6 Tabel Perhitungan Nilai TF dan DF.....	17
Tabel 3. 7 Proses Perhitungan Nilai IDF.....	18
Tabel 3. 8 Proses Perhitungan Nilai TF-IDF.....	19
Tabel 3. 9 Menghitung $W_{ij} * W_{iq}$	20
Tabel 3. 10 $\sum$ Nilai $W_{ij} * W_{iq}$	21
Tabel 3. 11 Menghitung Nilai $\sqrt{\sum W_{ij}} * \sqrt{\sum W_{iq}}$	21
Tabel 3. 12 Nilai Cosine similarity.....	22
Tabel 3. 13 Contoh Query.....	22
Tabel 3. 14 Perbandingan Query dengan Dokumen 1.....	22
Tabel 3. 15 Perbandingan Query dengan Dokumen 2.....	23
Tabel 3. 16 Perbandingan Query dengan Dokumen 3.....	23
Tabel 3. 17 Kata yang Sama di Dokumen 1.....	23
Tabel 3. 18 Kata yang Sama di Dokumen 2.....	23
Tabel 3. 19 Kata yang Sama di Dokumen 3.....	23
Tabel 3. 20 Skenario Pengujian.....	26
Tabel 4. 1 Hasil Uji Coba Dari Dua Metode.....	39
Tabel 4. 2 Uji Coba ke-1 Cosine similarity.....	41
Tabel 4. 3 Uji Coba ke-1 Jaro-Winkler Distance.....	41
Tabel 4. 4 Uji Coba ke-2 Cosine similarity.....	41
Tabel 4. 5 Uji Coba ke-2 Jaro-Winkler Distance.....	42
Tabel 4. 6 Uji Coba ke-3 Cosine similarity.....	42
Tabel 4. 7 Uji Coba ke-3 Jaro-Winkler Distance.....	42

Tabel 4. 8 Uji Coba ke-4 Cosine similarity.....	43
Tabel 4. 9 Uji Coba ke-4 Jaro-Winkler Distance.....	43
Tabel 4. 10 Uji Coba ke-5 Cosine similarity.....	43
Tabel 4. 11 Uji Coba ke-5 Jaro-Winkler Distance.....	44