



**PENERAPAN ALGORITMA *COSINE SIMILARITY* DAN TF-IDF DALAM
MENENTUKAN RUMPUN JABATAN**

SKRIPSI

RANGGA SAPUTRA

2010511036

PROGRAM STUDI INFORMATIKA

FAKULTAS ILMU KOMPUTER

UNIVERSITAS PEMBANGUNAN NASIONAL "VETERAN" JAKARTA

2023



**PENERAPAN ALGORITMA *COSINE SIMILARITY* DAN TF-IDF DALAM
MENENTUKAN RUMPUN JABATAN**

SKRIPSI

**Diajukan Sebagai Salah Satu Syarat Untuk Memperoleh
Gelar Sarjana Komputer**

RANGGA SAPUTRA

2010511036

PROGRAM STUDI INFORMATIKA

FAKULTAS ILMU KOMPUTER

UNIVERSITAS PEMBANGUNAN NASIONAL "VETERAN" JAKARTA

2023

PERNYATAAN ORISINALITAS

Tugas akhir ini adalah hasil karya sendiri, dan semua sumber yang dikutip maupun yang dirujuk telah saya nyatakan dengan benar.

Nama : Rangga Saputra

NIM : 2010511036

Tanggal : 15 Januari 2024

Bilamana dikemudian hari ditemukan ketidaksesuaian dengan pernyataan saya ini, maka saya bersedia dituntut dan diproses sesuai dengan ketentuan yang berlaku.

Jakarta, 15 Januari 2024

Yang menyatakan,



(Rangga Saputra)

**PERNYATAAN PERSETUJUAN PUBLIKASI TUGAS AKHIR UNTUK
KEPENTINGAN AKADEMIS**

Sebagai civitas akademik Universitas Pembangunan Nasional “Veteran” Jakarta, saya yang bertanda tangan di bawah ini :

Nama : Rangga Saputra

NIM : 2010511036

Fakultas : Ilmu Komputer

Program Studi : Informatika

Demi pengembangan ilmu pengetahuan, menyetujui untuk memberikan kepada Universitas Pembangunan Nasional “Veteran” Jakarta Hak Bebas Royalti Non eksklusif (*Non-exclusive Royalty Free Right*) atas karya ilmiah saya yang berjudul :

PENERAPAN ALGORITMA *COSINE SIMILARITY* DAN TF-IDF DALAM
MENENTUKAN RUMPUN JABATAN.

Dengan Hak Bebas Royalti ini Universitas Pembangunan Nasional “Veteran” Jakarta berhak menyimpan, mengalih media/formatkan, mengelola dalam bentuk pangkalan data (database), merawat, dan mempublikasikan Tugas Akhir saya selama tetap mencantumkan nama saya sebagai penulis/pencipta dan sebagai pemilik Hak Cipta. Demikian Pernyataan ini saya buat dengan sebenarnya.

Dibuat di : Jakarta

Pada tanggal : 15 Januari 2024

Yang menyatakan,



(Rangga Saputra)

LEMBAR PENGESAHAN

Skripsi diajukan oleh :

Nama : Rangga Saputra

NIM : 2010511036

Program Studi : S1 Informatika

Judul Skripsi : Penerapan Algoritma *Cosine Similarity* dan TF-IDF Dalam Menentukan Rumpun Jabatan

Telah berhasil dipertahankan di hadapan Tim Penguji dan diterima sebagai bagian persyaratan yang diperlukan untuk memperoleh gelar Sarjana pada program Studi Informatika, Fakultas Ilmu Komputer, Universitas Pembangunan Nasional Veteran Jakarta.



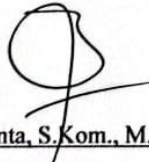
Dr. Didit Widiyanto, S.Kom, M.Si.

Penguji 1



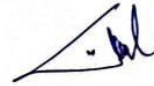
Nindy Irzavika, S.SI., M.T.

Penguji 2



Jayanta, S.Kom., M.Si

Pembimbing 1



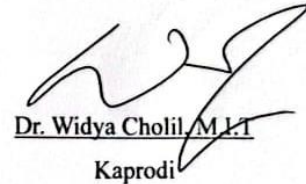
Musthofa Galih Pradana, M.Kom

Pembimbing 2



Prof. Dr. Ir. Suhyanto, S.T., M.Sc., IPM

Dekan



Dr. Widya Cholil, M.I.T

Kaprosdi

Ditetapkan di : Jakarta

Tanggal Ujian : 09 Januari 2024

PENERAPAN ALGORITMA *COSINE SIMILARITY* DAN TF-IDF DALAM MENENTUKAN RUMPUN JABATAN

Rangga Saputra

ABSTRAK

Pada tahun 2019, sebuah instansi pemerintah memperkenalkan sistem rumpun jabatan untuk meningkatkan efisiensi dalam penugasan jabatan pegawai. Tantangan muncul ketika data pegawai tahun sebelumnya tidak memiliki klasifikasi rumpun jabatan, dan data yang tersedia berupa teks dalam jumlah besar. Rumpun jabatan hanya dikelompokkan berdasarkan kesamaan karakteristik, termasuk keluarga jabatan, fungsi jabatan, dan peran jabatan. Rumpun Jabatan adalah kelompok dari jabatan administrasi yang saling terkait dalam tugas, fungsi, dan kompetensi untuk melaksanakan pekerjaan tertentu. Karena jumlah pegawai yang banyak, informasi yang melimpah, dan data yang dikelola merupakan data teks dalam jumlah yang besar, maka pengklasifikasian manual menjadi tidak efisien. Kondisi ini menyulitkan Biro SDM yang bertanggung jawab atas data pegawai. Untuk mengatasi permasalahan ini, diperlukan metode bantu yang dapat memproses data dengan cepat dan akurat. Salah satu pendekatan yang digunakan adalah *Cosine Similarity* menggunakan metode TF-IDF. Evaluasi hasil menunjukkan bahwa metode ini memberikan nilai *precision* untuk keluarga jabatan dan fungsi secara berurutan mencapai 89% dan 81%. Sementara itu, dalam mengklasifikasikan peran, tingkat *precision* yang dihasilkan rendah sebesar 52%. Meskipun terdapat *error rate* sebesar 26%, model ini dianggap cukup akurat dalam mengidentifikasi kesamaan antara data pegawai dan rumpun jabatan. Dengan toleransi *error rate* sebesar 30%, model ini dapat dipertimbangkan untuk implementasi oleh Biro SDM guna otomatisasi penentuan rumpun jabatan. Meskipun terdapat perbedaan kinerja antara kategori rumpun jabatan, model ini dianggap dapat diandalkan terutama untuk keluarga jabatan dan fungsi. Penelitian ini menyoroti potensi positif dalam penerapan algoritma tersebut sebagai solusi dalam konteks pengelolaan data pegawai dan klasifikasi rumpun jabatan.

Kata kunci: Instansi Pemerintahan; Rumpun Jabatan; *Cosine Similarity*

PENERAPAN ALGORITMA *COSINE SIMILARITY* DAN TF-IDF DALAM MENENTUKAN RUMPUN JABATAN

Rangga Saputra

ABSTRACT

In 2019, a government agency introduced a job cluster system to enhance efficiency in assigning employee positions. Challenges arose as the previous year's employee data lacked job cluster classifications, and the available data consisted of extensive text. Job clusters were only grouped based on shared characteristics, including job family, job function, and job role. Job clusters represent a grouping of administrative positions interconnected by tasks, functions, and competencies for specific job execution. Due to a large number of employees and copious textual data managed, manual classification became inefficient. This posed difficulties for the Human Resources Department responsible for employee data. To address this issue, a supporting method capable of processing data quickly and accurately was necessary. One approach employed was Cosine Similarity using the TF-IDF method. Evaluation results revealed that the method provided precision values for job family and function sequentially reaching 89% and 81%. However, in classifying roles, the generated precision level was low at 52%. Despite a 26% error rate, the model was considered sufficiently accurate in identifying similarities between employee data and job clusters. With a 30% error rate tolerance, this model could be considered for implementation by the Human Resources Department to automate job cluster determinations. Despite performance differences between job cluster categories, the model was considered reliable, especially for job family and function. This research highlights the positive potential of implementing this algorithm as a solution in the context of employee data management and job cluster classification.

Keywords: *Government Agencies; Position Family; Cosine Similarity*

KATA PENGANTAR

Puji dan syukur penulis panjatkan kehadirat Tuhan Yang Maha Esa yang telah memberikan banyak kenikmatan, terutama kenikmatan kesehatan dan kesempatan sehingga proses pembuatan Proposal Skripsi berjalan dengan baik. Begitupun atas Rahmat-Nya Proposal Skripsi dengan judul “Penerapan Algoritma *Cosine Similarity* dan TF-IDF dalam Menentukan Rumpun Jabatan” dapat penulis selesaikan dengan baik.

Penulis menyadari banyak pihak yang membantu dan berkontribusi dalam terselesaikannya Proposal Skripsi ini. Segala bentuk bantuan, baik dari segi moril dan materil sangat membantu penulis dalam mengumpulkan semangat serta keinginan untuk menyelesaikan Proposal Skripsi ini. Demikian penulis ucapkan terima kasih dengan ketulusan hati kepada pihak-pihak yang sudah membantu dan membimbing saya selama penyusunan Proposal Skripsi ini, yakni kepada :

1. Instansi Pemerintah terkait yang berperan sebagai sumber dan penyedia data untuk penelitian.
2. Pak Jayanta, S.Kom., M.Si, selaku dosen pembimbing 1 penulis selama penelitian ini dilakukan.
3. Pak Musthofa Galih Pradana, M.Kom selaku dosen pembimbing 2 penulis selama penelitian ini dilakukan.
4. Keluarga, yang selalu mendukung dan menyemangati penulis dengan doa dan dukungan.
5. Teman – teman seperjuangan yang telah memberikan dukungan selama perkuliahan.

Semoga Tuhan YME memberikan pahala yang berlimpah atas segala bentuk bantuan yang sudah diberikan kepada penulis. Penulis juga berharap agar Proposal Skripsi ini dapat bermanfaat bagi para pembaca dari berbagai segi kalangan. Serta penulis ingin mengucapkan permohonan maaf jika selama proses penyusunan Proposal Skripsi ini banyak melakukan kesalahan yang sudah penulis lakukan secara tidak sengaja ataupun disengaja.

DAFTAR ISI

LEMBAR COVER	i
LEMBAR JUDUL	ii
PERNYATAAN ORISINALITAS	iii
PERSETUJUAN PUBLIKASI.....	iv
LEMBAR PENGESAHAN.....	v
ABSTRAK	vi
<i>ABSTRACT</i>	vii
KATA PENGANTAR.....	viii
DAFTAR ISI.....	ix
DAFTAR TABEL	xi
DAFTAR GAMBAR	xii
BAB I PENDAHULUAN	1
1.1. Latar Belakang.....	1
1.2. Rumusan Masalah.....	3
1.3. Batasan Masalah	3
1.4. Tujuan	3
1.5. Manfaat	4
BAB II TINJAUAN PUSTAKA.....	5
2.1. Kajian Literatur.....	5
2.2. <i>Text Mining</i>	7
2.3. Instansi Pemerintahan	10
2.4. Rumpun Jabatan	11
2.5. Metode N-gram.....	12
2.6. TF-IDF	13
2.7. <i>Cosine Similarity</i>	14
2.8. <i>Precision</i>	16
BAB III METODE PENELITIAN	18
3.1. Tahapan Penelitian.....	18
3.1.1. <i>Studi Literatur</i>	18
3.1.2. Identifikasi Masalah.....	18
3.1.3. Pengumpulan Data.....	19
3.1.4. <i>Text Preprocessing</i>	19

3.1.5.	Pembobotan Kata.....	21
3.1.6.	<i>Text Matching</i>	22
3.1.7.	Evaluasi.....	23
3.1.8.	Kesimpulan.....	23
3.2.	Perangkat Penelitian.....	23
3.2.1.	Perangkat Keras.....	23
3.2.2.	Perangkat Lunak.....	24
3.3.	Jadwal Penelitian.....	24
BAB IV HASIL DAN PEMBAHASAN.....		25
4.1	Pengumpulan data.....	25
4.2	<i>Text Preprocessing</i>	27
4.2.1.	Concatenation.....	27
4.2.2.	Case folding.....	27
4.2.3.	Filtering.....	28
4.2.4.	Stemming.....	29
4.2.5.	Tokenizing.....	29
4.3	Pembobotan Kata.....	30
4.4	<i>Text Matching</i>	33
4.5	Evaluasi.....	38
BAB V PENUTUPAN.....		41
5.1	Kesimpulan.....	41
5.2	Saran.....	41
DAFTAR PUSTAKA.....		42
RIWAYAT HIDUP.....		44

DAFTAR TABEL

Tabel 2. 1 Hasil Perhitungan TF dan IDF	13
Tabel 2. 2 Hasil Perhitungan TF-IDF.....	14
Tabel 2. 3 Contoh Hasil Perkalian Vector Dokumen (D) dengan Query (Q).....	15
Tabel 2. 4 Contoh Hasil Perhitungan Sigma Akar Vector Dokumen (D).....	16
Tabel 2. 5 Hasil Perhitungan Cosine Similarity Terhadap Query (Q)	16
Tabel 3. 1 Jadwal Penelitian	24
Tabel 4. 1 Contoh Dataset Riwayat Jabatan Administrasi	25
Tabel 4. 2 Contoh Dataset Rumpun Jabatan.....	26
Tabel 4. 3 Data Hasil Concatenation.....	27
Tabel 4. 4 Data Hasil Case Folding	28
Tabel 4. 5 Data Hasil Filetering	28
Tabel 4. 6 Data Hasil Stemming.....	29
Tabel 4. 7 Data Hasil Tokenizing.....	29
Tabel 4. 8 Sampel Data Riwayat Jabatan Administrasi	30
Tabel 4. 9 Sampel Data Rumpun Jabatan (Setelah Text Preprocessing).....	30
Tabel 4. 10 Data Sampel Setelah Text Preprocessing	31
Tabel 4. 11 Hasil Perhitungan TF dan IDF.....	31
Tabel 4. 12 Hasil Perhitungan TF-IDF.....	32
Tabel 4. 13 Perhitungan Perkalian Vector Dokumen (D) dengan Query (Q).....	33
Tabel 4. 14 Hasil Perhitungan Sigma Akar Vector Dokumen (D).....	34
Tabel 4. 15 Hasil Perhitungan Cosine Similarity.....	35
Tabel 4. 16 Kelompok Rumpun Jabatan	35
Tabel 4. 17 Hasil Pengelompokan Rumpun Jabatan.....	35

DAFTAR GAMBAR

Gambar 1. 1 Proses Pengelompokan	2
Gambar 2. 1 Gambar Diagram Venn Text Mining.....	8
Gambar 2. 2 Gambar Hubungan Berbagai Teknik Text Mining.....	9
Gambar 2. 3 Gambar Tahapan Text Mining	9
Gambar 3. 1 Tahapan Penelitian	18
Gambar 3. 2 Flowchart Concatenation	19
Gambar 3. 3 Flowchart Case Folding	20
Gambar 3. 4 Flowchart Filtering	20
Gambar 3. 5 Flowchart Stemming.....	21
Gambar 3. 6 Flowchart Tokenizing.....	21
Gambar 3. 7 Flowchart Pembobotan Kata (TF-IDF).....	22
Gambar 3. 8 Flowchart Cosine Similarity.....	22
Gambar 4. 1 Hasil Validasi Pegawai	38
Gambar 4. 2 Perbandingan Hasil Precision.....	40