Krea-TIF: Jurnal Teknik Informatika
http://ejournal.uika-bogor.ac.id/index.php/krea-tif
p-ISSN: 2338-2910, e-ISSN: 2658-5836
Vol. 12, No. 1, 2024, pp. 1 – 11
DOI: 10.32832/kreatif.v12i1.15470

Implementasi Algoritma Cosine Similarity dan TF-IDF dalam Menentukan Rumpun Jabatan

Rangga Saputra*, Jayanta, Musthofa Galih Pradana
Program Studi Informatika, Fakultas Ilmu Komputer, Universitas Pembangunan Nasional Veteran Jakarta, Indonesia
*E-mail koresponden: 2010511036@mahasiswa.upnvj.ac.id

Diserahkan *20 November 2023;* ***Direview*** *17 Februari 2024;****Dipublikasikan*** *30 Mei 2024*

Abstrak

Pada tahun 2019, sebuah instansi pemerintah memperkenalkan sistem rumpun jabatan untuk meningkatkan efisiensi dalam penugasan jabatan pegawai. Namun, tantangan muncul ketika data pegawai tahun sebelumnya tidak memiliki klasifikasi rumpun jabatan, dan data yang tersedia berupa teks dalam jumlah besar. Dikarenakan jumlah pegawai yang banyak, informasi yang melimpah, dan data yang dikelola merupakan data teks dalam jumlah yang besar. Proses pengklasifikasian manual menjadi tidak efisien. Salah satu pendekatan yang digunakan untuk mengatasi pemrosesan data yang cepat dan akurat adalah Cosine Similarity menggunakan metode TF-IDF. Evaluasi hasil menunjukkan bahwa metode ini memberikan rata – rata precision sebesar 74%. Lebih rinci, nilai precision untuk kelompok keluarga jabatan dan fungsi secara berurutan mencapai 89% dan 81%. Namun, dalam mengklasifikasikan kelompok peran, tingkat precision yang dihasilkan rendah sebesar 52%. 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 kelompok keluarga jabatan dan fungsi.

Kata kunci: *Cosine Similarity,* Instansi Pemerintahan, Rumpun Jabatan.

Abstract

In 2019, a government agency introduced a position cluster system to improve the efficiency of employee job assignments. However, challenges arise when the previous year's employee data do not have a classification of job groups and the data available are in the form of large amounts of text. Owing to the large number of employees, abundant information, and data management, there is a large amount of text data. Thus, the manual classification process is inefficient. One approach used to overcome fast and accurate data processing is Cosine Similarity using the TF-IDF method. The evaluation of the results shows that this method provides an average precision of 74%. In more detail, the precision value for the family of positions and functions reached 89% and 81 %, respectively. However, in classifying roles, the resulting precision rate was low at 52%. This model can be considered for implementation by the HR Bureau to automate the determination of position clusters. Although there are differences in performance between job family categories, this model is considered reliable, especially for families of positions and functions.

Keywords: *Cosine Similarity, Government Agency*, *Position Agency*.

PENDAHULUAN

Berbagai tantangan dan situasi yang kompleks dalam administrasi pemerintahan perlu dihadapi sehingga dibutuhkan suatu pendekatan manajemen yang memfokuskan pada keahlian dan pencapaian[1]. Pendekatan ini menyoroti signifikansi evaluasi dan penghargaan terhadap kemampuan serta kontribusi para staf, daripada hanya mempertimbangkan pengalaman atau masa kerja. Dengan landasan kompetensi dan hasil kerja sebagai pijakan, diharapkan birokrasi publik dapat mengemban tugasnya dengan lebih profesional [2].

Di samping itu, transformasi dan peningkatan dalam bidang profesionalisme tidak hanya berlaku bagi individu, melainkan juga berfokus pada kerangka organisasi dan sistem manajemen yang berperan mendukung. Penyusunan struktur organisasi yang sesuai akan berkontribusi pada peningkatan keterpaduan, produktivitas, dan kerjasama di antara berbagai unit kerja [3]. Praktik tata kelola yang kuat juga menjadi dasar bagi pegawai negeri dalam menjalankan tugas-tugasnya dengan jujur, tanggung jawab, dan integritas.

Pada tahun 2019, diperkenalkan sistem rumpun jabatan di dalam salah satu instansi pemerintahan, namun data pegawai sebelum tahun tersebut tidak memiliki klasifikasi rumpun jabatan. Kondisi ini menyulitkan Biro Sumber Daya Manusia (SDM) yang bertanggung jawab atas data pegawai dimana jumlah data yang banyak berupa teks akan memiliki informasi yang melimpah. Rumpun jabatan dikelompokkan berdasarkan kesamaan karakteristik, termasuk kelompok 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. Rumpun jabatan terdiri dari 12 keluarga jabatan, 32 fungsi jabatan, dan 109 peran jabatan [4]. Situasi ini menyebabkan pengelompokan rumpun jabatan yang sesuai dengan data pegawai yang ada membutuhkan waktu yang sangat lama sehingga dibutuhkan upaya untuk merapikan dan mengelola data secara lebih efisien. Penerapan teknologi dan sistem informasi yang tepat dapat membantu biro SDM untuk mengatasi permasalah ini. Teknik analisis teks dan pemrosesan bahasa alami bisa dipakai untuk mengelompokkan jabatan sesuai dengan rumpun jabatan. Tidak hanya itu, sistem manajemen SDM yang maju bisa dipakai untuk mencocokan rumpun jabatan dengan profil pegawai, sehingga proses penentuan rumpun jabatan dari pegawai lebih mudah untuk dilakukan.

Penelitian sebelumnya yang membahas tentang rumpun jabatan menggunakan metode *Multiclass Support Vector Machine* menghasilkan akurasi 98,5%. Namun, model gagal mengklasifikasikan rumpun jabatan penetapan keputusan yang kewenangannya pada Presiden dikarenakan *dataset* yang dimiliki tidak seimbang [5]. Penelitian lainnya mengembangkan platform informasi lowongan kerja berbasis website dengan sistem rekomendasi menggunakan algoritma *Cosine Similarity* dan *Rabin Karp* K-Gram. Kelemahan penelitian ini dalam pencocokan teks yang tidak selalu akurat, sehingga perlu variasi atau kombinasi algoritma [6]. Penelitian sistem rekomendasi untuk mahasiswa informatika memilih Program Studi dan Magang Mandiri Bersertifikat (MSIB) yang menggunakan metode *Content-Based Filtering* mencapai hasil presisi rata-rata 89,4%, mempunyai kelemahan yaitu adanya keterbatasan data [7]. Sistem otomatis klasifikasi dokumen pengaduan masyarakat menggunakan algoritma *Cosine Similarity* dan TF-IDF dimana hasilnya menunjukkan tingkat akurasi sebesar 84%, dan sistem ini dapat ditingkatkan dengan penambahan kategori klasifikasi serta data pelatihan [8].

Metode *Multiclass Support Vector Machine* memiliki kinerja yang baik disesuaikan dari hasil akurasi beberapa penelitian yang baik. Meski demikian, metode ini mengalami

kesulitan dalam mengklasifikasikan rumpun jabatan jika memiliki ketidakseimbangan dataset. Di sisi lain, untuk topik penelitian yang serupa menggunakan metode *Cosine Similarity* menghasilkan performa yang memuaskan, dengan tingkat akurasi di atas 80%. Hasil positif dari metode *Cosine Similarity* menunjukkan potensi penggunaan metode tersebut dalam konteks penelitian ini, dan dapat menjadi alternatif atau pelengkap untuk meningkatkan ketepatan klasifikasi.

Fokus dari penelitian ini adalah untuk mengimplementasikan algoritma *Cosine Similarity* dan *Term Frequency-Inverse Document Frequency* (TF-IDF) dalam mengelompokkan rumpun jabatan. Algoritma *Cosine Similarity* adalah teknik yang jalah dua *vector*, maka *Cosine Similarity* (cos θ). Hasil *Cosine Similarity* ini memberikan ukuran sejauh mana dua *vector* tersebut sejajar dalam ruang *vector*.

Term Frequency-Inverse Document Frequency

Pembobotan *Term Frequency-Inverse Document Frequency* (TF-IDF) banyak digunakan karena efisiensinya, kemudahan dalam penerapannya, dan hasil yang akurat [11]. Konteks perhitungan TF-IDF, kata – kata yang sering muncul akan memiliki bobot yang lebih besar, kecuali jika jumlah dokumen yang mengandung kata tersebut juga tinggi. Hal ini dikenal sebagai *inverse document frequency* (IDF). Sebagai contoh, jika kata "*please*" sering muncul namun tersebar di berbagai dokumen, maka kata tersebut akan memiliki bobot yang rendah. Skema persamaan TF – IDF ditunjukkan Persamaan 2.

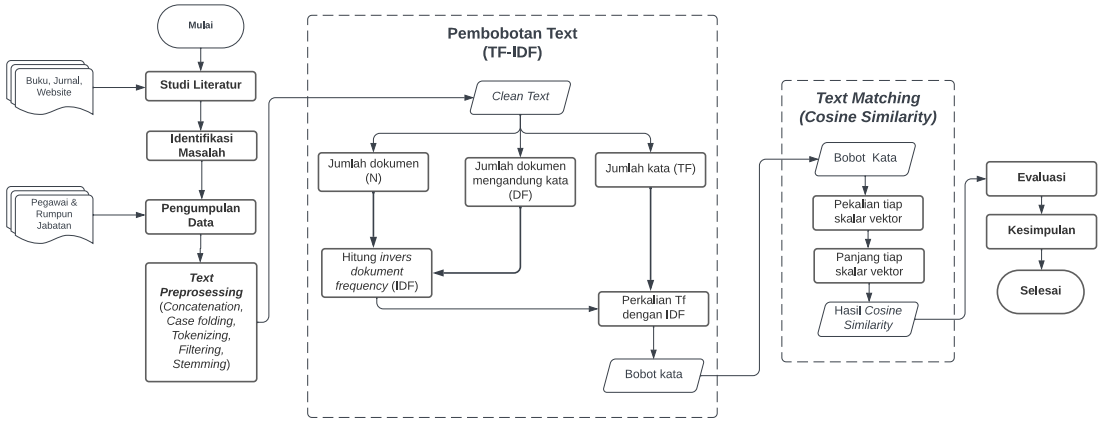
$$tf - idf\ (w) = tf\ x\ \log\frac{N}{df\ (w)} \tag{2}$$

Term frequency (tf) mengacu pada jumlah kemunculan suatu kata dalam suatu dokumen. Dengan kata lain, $tf(w)$ adalah frekuensi kata w dalam suatu dokumen. Sementara itu, *document frequency* (df) mengukur seberapa sering suatu kata muncul dalam seluruh *corpus* dokumen. Jika suatu kata muncul dalam banyak dokumen, *document frequency* (df) akan tinggi. Jumlah dokumen dalam *corpus* tersebut direpresentasikan oleh N.

METODE PENELITIAN

Tahapan Penelitian

Penelitian diawali dengan melakukan studi literatur untuk mengkaji dan memahami langkah – langkah serta pemanfaatan lain yang berkaitan dengan topik penelitian. Tahapan selanjutnya adalah mengidentifikasi masalah yang terjadi pada objek penelitian. Setelah itu dilakukan pengumpulan data berupa data riwayat jabatan pegawai dan data rumpun jabatan dari instansi pemerintahan yang telah dikumpulkan dari *database*. Sebelum dilakukan pencocokan teks dengan *Cosine Similarity* data – data tersebut melewati tahap *text preprocessing*. Setelah data melalui tahap *text preprocessing*, langkah berikutnya adalah melakukan pembobotan kata. Tujuan dari pembobotan kata ini adalah memberikan nilai vektor pada setiap kata atau *term* yang digunakan sebagai *input* dalam proses pencocokan kata. Robertson menyatakan bahwa untuk membangun model *vector*, diperlukan proses pembobotan yang dapat dilakukan dengan skema *term frequency - inverse document frequency* (TF-IDF) [11]. Setelah mendapatkan *vector* dari setiap kata langkah selanjutnya, adalah melakukan *text matching* antara data pegawai dengan data rumpun jabatan untuk menentukan kecocokan data pegawai tersebut dengan rumpun jabatan tertentu. Metode yang diterapkan adalah *cosine similarity*.



Gambar 1 Tahapan Penelitian

Text Preprocessing

Text preprocessing merupakan langkah pertama dalam mempersiapkan teks agar bisa diolah lebih lanjut [13]. Tujuan dari *text preprocessing* adalah untuk membersihkan dan menormalisasikan teks agar terhindar dari *noise* [14]. Penelitian ini memiliki langkah – langkah *text preprocessing* yang dilakukan adalah sebagai berikut : *concatenation* adalah sebuah proses untuk menggabungkan dua atau lebih teks atau rangkaian karakter menjadi suatu kesatuan. *Text mining* dilakukan untuk menggabungkan informasi dari beberapa kalimat atau dokumen agar menjadi satu teks. Sehingga mendapatkan lebih banyak informasi setelah dilakukan analisis lebih lanjut [12]. *Case folding* merupakan tahap dimana dilakukan proses penyamaan huruf besar atau kecil dalam sebuah dokumen teks [15]. *Filtering* merupakan tahapan dokumen diproses untuk menghilangkan karakter yang tidak penting dan tidak bermakna, seperti kata hubung dan konjungsi, tanda baca, tanda hubung, *stopwords* [16]. *Stopwords* merupakan tahap menghilangkan kata – kata yang tidak memiliki makna atau tidak informatif yang sering terdapat di dalam dokumen [17]. Kata – kata tersebut biasanya terdiri dari kata penghubung, kata ganti orang, atau kata- kata lain yang tidak bermakna dalam penentuan topik suatu dokumen [18]. *Stemming* merupakan tahapan dilakukannya Pengolahan berbagai variasi kata ke dalam bentuk tunggal, yakni kata baku [18]. Tahapan ini bertujuan untuk mengubah kata-kata yang mengandung imbuhan dan telah melewati seleksi menjadi kata dasar [19]. Tahapan *tokenizing* proses yang digunakan untuk memecah teks berdasarkan setiap kata yang membentuknya, yang disebut sebagai *term* atau token [20].

Evaluasi

Model yang telah dibuat akan dievaluasi untuk mengukur kinerja. Evaluasi dilakukan dengan hasil model akan diperiksa oleh pegawai instansi pemerintahan untuk memvalidasi kesesuaian antara data riwayat jabatan dan data rumpun jabatan. Setelah validasi dilakukan, langkah selanjutnya adalah menggunakan hasil validasi untuk menghitung tingkat ketepatan prediksi data yang benar, atau disebut juga *precision* [21]. Akurasi sistem dihitung berdasarkan prediksi skor *similarity* konten pada dataset terhadap setiap konten acuan yang digunakan. Adapun perhitungan *precision* digunakan untuk mengevaluasi kinerja model (Persamaan 3).

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

True Positive (TP) merujuk pada jumlah prediksi yang benar-benar relevan atau mirip dengan kelas yang diinginkan. Sebaliknya, *False Positive* (FP) mengindikasikan jumlah prediksi yang sebenarnya tidak relevan, namun model memprediksi bahwa hasilnya relevan atau serupa.

HASIL DAN PEMBAHASAN

Pengumpulan Data

Dalam penelitian ini, *dataset* yang digunakan berasal dari riwayat jabatan administrasi pegawai dan rumpun jabatan di salah satu instansi pemerintahan. Data riwayat jabatan administrasi pegawai diperoleh langsung dari *database* instansi tersebut oleh salah satu pegawai, kemudian disampaikan kepada peneliti dalam format file *.csv* atau *.xlsx*. Pada Tabel 1 merupakan contoh *dataset* dari riwayat jabatan administrasi pegawai.

Tabel 1 Data Jabatan Pegawai

iddiklatteknis	jabatan	nama_jabatan
1	Struktural	kepala subbagian asia dan pasifik bagian antar negara biro hubungan internasional deputi sekretaris wakil presiden bidang politik
10	Struktural	kepala bagian pendidikan biro pendidikan kebudayaan dan olah raga deputi sekretaris wakil presiden bidang kesejahteraan rakyat
20	Struktural	kepala bagian politik hukum dan keamanan negara biro data dan informasi sekretariat dewan pertimbangan presiden x d es. iiia tmt.
50	Struktural	kepala subbagian pemberdayaan masyarakat pada bagian pengentasan kemiskinan biro kesejahteraan rakyat deputi sekretaris kabinet bidang pemerintahan
……	……	……
1547	Struktural	kepala subbidang hubungan mpr dan dpd bidang hubungan mpr dpr dan dpd asisten deputi hubungan lembaga negara dan lembaga non struktural deputi bidang hubungan kelembagaan dan kemasyarakatan kementerian sekretariat negara

Data rumpun jabatan yang digunakan dalam penelitian ini diperoleh dari "Peraturan Menteri Sekretaris Negara Republik Indonesia Nomor 2 Tahun 2019 tentang Rumpun Jabatan di Lingkungan Kementerian Sekretariat Negara" yang mencakup 109 data rumpun jabatan. Tabel 2 merupakan data rumpun jabatan.

Tabel 2 Data Rumpun Jabatan

idrumpun	Keluargajabatan	fungsi	peran
1	Sumber Daya Manusia, Kelembagaan, dan Tata Laksana	Pengelolaan Sumber Daya Manusia	Perencanaan Sumber Daya Manusia
2	Sumber Daya Manusia, Kelembagaan, dan Tata Laksana	Pengelolaan Sumber Daya Manusia	Pengadaan, Pengangkatan, dan Pemberhentian Pegawai
3	Sumber Daya Manusia, Kelembagaan, dan Tata Laksana	Pengelolaan Sumber Daya Manusia	Pengelolaan Kinerja Pegawai
4	Sumber Daya Manusia, Kelembagaan, dan Tata Laksana	Pengelolaan Sumber Daya Manusia	Penilaian dan Pemetaan Kompetensi Pegawai
……..	……..	……..	……..
109	Penetapan keputusan yang kewenangannya pada Presiden	Perjalanan Dinas Luar Negeri	Evaluasi Pelaksanaan Perjalanan Dinas Luar Negeri

Text Preprocessing

Tahapan *text preprocessing* dilakukan untuk mengurangi noise pada data. Metode – metode yang digunakan untuk melakukan *text preprocessing* pada penelitian ini, antara lain :

Concatenation

Pada tahap ini kolom 'Keluargajabatan', 'fungsi', 'peran' di dalam data rumpun jabatan digabungkan menjadi satu kesatuan, memungkinkan terbentuknya informasi yang lebih komprehensif dan terinci. Untuk contoh hasil *Concatenation* dapat dilihat di Tabel 3.

Tabel 3 Sampel Output Concatenation

Output Concatenation
Sumber Daya Manusia, Kelembagaan, dan Tata Laksana Pengelolaan Sumber Daya Manusia Perencanaan Sumber Daya Manusia

Case Folding

Setelah dilakukan penggabungan makan tahap selanjutanya adalah tahap dimana dilakukan proses penyamaan huruf dalam sebuah dokumen teks. Penyamaan ini dilaksanakan sebagai langkah untuk mencegah perbedaan makna ketika huruf besar dan huruf kecil terdeteksi. Penelitian ini menggunakan fungsi yang tersedia untuk memanipulasi string di python yaitu *'.lower()*' agar menjadi huruf kecil.

Filtering

Setelah dilakukan penyamaan ukuran huruf, maka masuk ke dalam tahapan *filtering*. Dokumen akan diproses untuk menghilangkan karakter yang tidak penting dan tidak bermakna. *List* yang dibuat dari *stopwords* menggunakan *library sastrawi*, sedangkan untuk menghilangkan karakter selain huruf saja menggunakan *regex*.

Stemming

Setelah teks melewati tahapan *filtering* selanjutnya akan dilakukan proses *stemming*. Tahapan ini akan membalikan teks menjadi bentuk dasarnya atau dalam bentuk bakunya. Dalam tahap ini peneliti juga menggunakan *library* sastrawi.

Tokenizing

Proses akhir di dalam tahap *text preprocessing* ini adalah *tokenizing* untuk mempermudah perhitungan kata-kata tersebut dalam dokumen maupun dalam perhitungan *frekuensi* kemunculan data tersebut dalam *corpus*. Proses ini dilakukan pada masing-masing data seperti data rumpun jabatan (Tabel 4) dan data jabatan (Tabel 5).

Tabel 4 Sampel Data Rumpun Jabatan Setelah *Text Preprocessing*

Keluargajabatan	**fungsi**	**peran**	**Data Setelah *Text Preprocessing***
Pelayanan Umum	Pelayanan Medis	Administrasi Medis/Kesehatan (Medical Record)	[layan, umum, layan, medis, administrasi, medis, sehat, medical, record]
Analisis Kebijakan	Analisis Kebijakan	Analisis Kebijakan Bidang Pembangunan Manusia, Masyarakat, dan Kebudayaan	[analisis, bijak, analisis, bijak, analisis, bijak, bidang, bangun, manusia, masyarakat, budaya]

Tabel 5 Sampel Data Jabatan Setelah Text Preprocessing

nama_jabatan	**nama_jabatan setelah t*ext preprocessing***
kepala subbagian pemberdayaan masyarakat pada bagian pengentasan kemiskinan biro kesejahteraan rakyat deputi sekretaris kabinet bidang pemerintahan	[daya, masyarakat, entas, miskin, biro, sejahtera, rakyat, bidang, perintah]

Pembobotan Kata

Pembobotan kata menggunakan *Term Frequency – Inverse Document Frequency* (TF-IDF) dilakukan setelah *pre-processing*. Proses TF-IDF dimulai dengan menghitung TF di setiap dokumen, diikuti oleh perhitungan IDF untuk setiap dokumen menggunakan rumus di Persamaan 2. Contoh perhitungan menggunakan sampel dari riwayat jabatan dari pegawai diwakilkan dengan notasi 'Q' dan sampel dari rumpun jabatan diwakilkan dengan notasi 'D1' dan 'D2' yang sebelumnya sudah diproses. Sampel untuk perhitungan dapat dilihat pada Tabel 6.

Tabel 6 Sampel Data untuk Perhitungan TF-IDF

Dokumen	Teks (Setelah T*ext Preprocessing*)
Q	[daya, masyarakat, entas, miskin, biro, sejahtera, rakyat, bidang, perintah]
D1	[layan, umum, layan, medis, administrasi, medis, sehat, medical, record]
D2	[analisis, bijak, analisis, bijak, analisis, bijak, bidang, bangun, manusia, masyarakat, budaya]

Langkah pertama yang akan dilakukan adalah menghitung *Term Frequency* (TF) dan *Inverse Document Frequency* (IDF) dari setiap dokumen. Hasil perhitungan untuk mencari TF dan IDF dapat dilihat pada Tabel 7.

Setelah mendapatkan TF dan IDF langkah selanjutnya adalah melakukan perkalian antara TF dan IDF sehingga akan mendapatkan nilai dari TF-IDF. Tabel 8 merupakan hasil perhitungan TF-IDF.

Tabel 7 Hasil Perhitungan TF dan IDF

Term	TF					DF	IDF
	Q	D1	D2	D3	D4		$\log\frac{N}{DF}$
Daya	1	0	0	0	0	1	0.70
Masyarakat	1	0	1	1	0	2	0.40
Entas	1	0	0	0	0	1	0.70
Miskin	1	0	0	0	0	1	0.70
Biro	1	0	0	0	0	1	0.70
Sejahtera	1	0	0	0	0	1	0.70
Bidang	1	0	1	0	0	1	0.70
Perintah	1	0	0	1	1	3	0.22
Layan	0	2	0	0	0	1	0.70
Umum	0	1	0	0	0	2	0.40
Medis	0	2	0	0	0	1	0.70
Sehat	0	1	0	0	0	2	0.40
Medical	0	1	0	0	0	1	0.70
Record	0	1	0	0	0	1	0.70
Analisis	0	0	3	0	0	1	0.70
Bijak	0	0	3	0	0	1	0.70
Bangun	0	0	1	0	0	1	0.70
Manusia	0	0	1	0	0	2	0.40
Budaya	0	0	1	0	0	1	0.70

Tabel 8 Hasil Perhitungan TF-IDF

Term	Q	D1	D2
Daya	0.47	0	0
Masyarakat	0.17	0	0.17
Entas	0.47	0	0
Miskin	0.47	0	0
Biro	0.47	0	0
Sejahtera	0.47	0	0
Bidang	0.17	0	0.17
Perintah	0.47	0	0
Layan	0	0.34	0
Umum	0	0.47	0
Medis	0	0.34	0
Sehat	0	0.47	0
Medical	0	0.47	0
Record	0	0.47	0
Analisis	0	0	0
Bijak	0	0	0
Bangun	0	0	0.47
Manusia	0	0	0.47
Budaya	0	0	0.47

Text Matching

Text matching digunakan untuk menghitung tingkat kesamaan (*similarity*) antara *query* dengan setiap dokumen menggunakan metode *cosine similarity*. Proses simulasi perhitungan nilai *Cosine Similarity* menggunakan hasil dari perhitungan dari Tabel 8. Adapun tahapan yang pertama kali dilakukan adalah melakukan perkalian *scalar vector* masing – masing dokumen (D) terhadap *scalar vector query* (Q), *scalar vector* yang dimaksudkan adalah nilai TF-IDF. Sehingga akan diperoleh nilai yang dapat dilihat pada Tabel 9.

Tabel 9 Hasil Perkalian Antar *Scalar*

Term	D1	D2
Daya	0	0
Masyarakat	0	0.0289
Entas	0	0
Miskin	0	0
Biro	0	0
Sejahtera	0	0
Bidang	0	0.0289
Perintah	0	0
Layan	0	0
Umum	0	0
Medis	0	0
Sehat	0	0
Medical	0	0
Record	0	0
Analisis	0	0
Bijak	0	0
Bangun	0	0
Manusia	0	0
Budaya	0	0
Total	**0**	**0.0578**

Tabel 10 Hasil Perhitungan Panjang Setiap Dokumen

Term	Q	D1	D2
Daya	0.2209	0	0
Masyarakat	0.0289	0	0.0289
Entas	0.2209	0	0
Miskin	0.2209	0	0
Biro	0.2209	0	0
Sejahtera	0.2209	0	0
Bidang	0.0289	0	0.0289
Perintah	0.2209	0	0
Layan	0	0	0
Umum	0	0.2209	0
Medis	0	0	0
Sehat	0	0.2209	0
Medical	0	0.2209	0
Record	0	0.2209	0
Analisis	0	0	0
Bijak	0	0	0
Bangun	0	0	0.2209
Manusia	0	0	0.2209
Budaya	0	0	0.2209
Total	**1.3832**	**0.8836**	**0.7205**
Hasil Akar Total	**1.1760**	**0.94**	**0.8488**

Selanjutnya adalah menghitung Panjang setiap dokumen (D) termasuk *query* (Q), dengan mengkuadratkan bobot (TF-IDF) dari setiap *term* dalam setiap dokumen lalu jumlahkan nilainya dan terakhir kuadratkan. Untuk hasil perhitungan dapat dilihat pada Tabel 10.

Langkah terakhir hitung nilai *Cosine Similarity* terhadap *query* (Q) menggunakan rumus di Persamaan 1. Sehingga akan diperoleh hasil yang dapat dilihat pada Tabel 11.

Tabel 11 Hasil Perhitungan Cosine Similarity

Dokumen (D)	Teks	Nilai *Similarity*	Bentuk Persentase
1	Pelayanan Umum Pelayanan Medis Administrasi Medis/Kesehatan (Medical Record)	0.0	0%
2	Analisis Kebijakan Analisis Kebijakan Analisis Kebijakan Bidang Pembangunan Manusia, Masyarakat, dan Kebudayaan	0.0579	6%

Berdasarkan hasil perhitungan *Cosine Similarity* untuk jabatan administrasi "kepala subbagian pemberdayaan masyarakat pada bagian pengentasan kemiskinan biro kesejahteraan rakyat deputi sekretaris kabinet bidang pemerintahan", ditemukan bahwa jabatan tersebut memiliki tingkat kemiripan sebesar 15% dengan salah satu rumpun jabatan tertentu. Oleh karena itu, jabatan ini dapat dikelompokkan ke dalam rumpun jabatan yang berada didalam Tabel 12.

Tabel 12 Kelompok Rumpun Jabatan

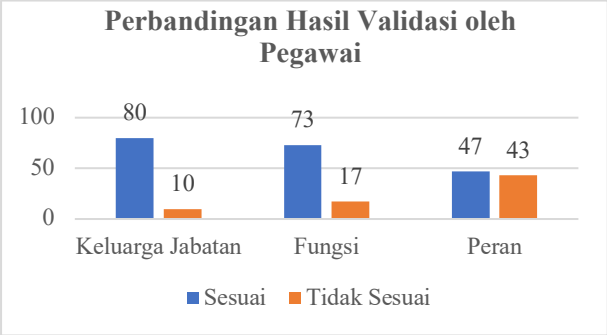
Keluarga jabatan	Fungsi jabatan	Peran jabatan
Analisis Kebijakan	Analisis Kebijakan	Analisis Kebijakan Bidang Pembangunan Manusia, Masyarakat, dan Kebudayaan

Evaluasi

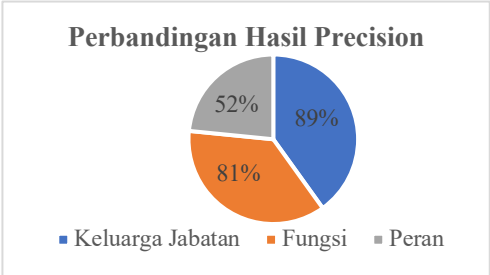
Setelah dilakukan validasi oleh pegawai Biro SDM di instansi pemerintahan tersebut menggunakan teknik *random sampling*. Hasil validasi dapat dilihat pada diagram batang yang ada pada Gambar 2.

Hasil validasi oleh pegawai Biro SDM di instansi pemerintahan tersebut menunjukkan bahwa untuk kelompok keluarga jabatan, terdapat 80 data yang sesuai dan 10 data yang tidak sesuai. Sementara untuk kelompok fungsi, terdapat 73 data yang sesuai dan 17 data yang tidak sesuai. Sedangkan untuk kelompok peran, terdapat 47 data yang sesuai dan 43 data yang tidak sesuai. Dari hasil validasi ini, langkah selanjutnya adalah melakukan evaluasi model. Evaluasi ini dilakukan dengan menghitung tingkat *precision*. *Precision* dipilih karena fokus pada seberapa banyak data yang dianggap positif oleh model yang sebenarnya relevan atau benar positif. Dalam konteks similarity, *precision* mencerminkan seberapa baik model dalam mengidentifikasi objek yang sebenarnya mirip atau serupa.

Perhitungan *precision* menggunakan rumus di Persamaan 3 menghasilkan rata – rata *precision* sebesar 74%. Dengan rincian *precision* kelompok keluarga jabatan, fungsi, dan peran secara berurutan 89%, 81%, 52%. Hasil visualisasi *pie chart* nilai *precision* dapat dilihat di Gambar 3. Terdapat perbedaan untuk setiap kategori rumpun jabatan, perbedaan ini dapat disebabkan karena kelompok peran merupakan bagian dari kelompok fungsi jabatan administrasi yang memiliki tugas serta spesialisasi kompetensi umum dan teknis. Sehingga informasi dari data menjadi terbatas dan juga representasi teks dari data pegawai dan data rumpun jabatan sangat berbeda. Dengan demikian, hasil penelitian menunjukkan bahwa model memiliki kinerja yang lebih tinggi dalam menentukan kelompok keluarga jabatan dan fungsi dibandingkan dengan kelompok peran. Hasil evaluasi precision yang tinggi menunjukkan bahwa model tersebut cukup akurat dalam mengidentifikasi kesamaan antara data pegawai dan data rumpun jabatan. Penggunaan metode ini dapat membantu Biro SDM mengotomatiskan proses penentuan rumpun jabatan, meningkatkan efisiensi, dan mengurangi kebutuhan intervensi manual, meskipun terdapat toleransi *error rate* hingga 30% yaitu dalam menentukan kelompok keluarga jabatan dan peran.



Gambar 2 Hasil Validasi oleh Pegawai



Gambar 3 Perbandingan Hasil *Precision*

KESIMPULAN

Hasil evaluasi *precision* dalam menentukan rumpun jabatan pegawai di instansi pemerintahan menggunakan metode *Cosine Similarity* menunjukkan bahwa model memiliki nilai rata – rata *precision* sebesar 74%. Hasil lebih rinci mengenai nilai *precision* untuk kelompok keluarga jabatan dan fungsi memiliki nilai lebih besar dari 80%, namun untuk kelompok peran memiliki akurasi yang cukup yaitu lebih kecil dari 60%. Meskipun terdapat *error rate* model ini dinilai cukup akurat dalam mengidentifikasi kesamaan antara data pegawai dan data rumpun jabatan. Toleransi *error rate* memiliki nilai yang menjadi batas agar dapat mempertimbangkan saat implementasi model dalam memperhitungkan keakuratan ketika menentukan rumpun jabatan. Penggunaan algoritma *cosine similarity* dan TF-IDF diyakini dapat membantu Biro SDM dalam mengotomatiskan proses penentuan rumpun jabatan, meningkatkan efisiensi, dan mengurangi intervensi manual. Meskipun

terdapat perbedaan kinerja antara kategori rumpun jabatan, model ini dapat diandalkan terutama untuk kelompok keluarga jabatan dan fungsi.

Penelitian ini memberikan beberapa rekomendasi yaitu menjelajahi alternatif – alternatif lain dalam pemilihan algoritma dapat menjadi langkah yang bijaksana. Terdapat berbagai metode lain selain *cosine similarity*, seperti *Jaccard similarity* atau algoritma pembelajaran mesin seperti *Random Forest* atau *Deep Learning* yang mungkin dapat menghasilkan hasil yang lebih baik. Selain itu dapat dijadikan penelitian lanjut untuk mengganti metode TF-IDF dengan Word2Vec.

DAFTAR PUSTAKA

[1] Kamaruddin Sellang, "ADMINISTRASI DAN PELAYANAN PUBLIK Antara Teori dan Aplikasinya," *Ombak*, no. September, pp. 1–229, 2016, [Online]. Available: https://www.mendeley.com/viewer/?fileId=349a0ada-0d19-cc5f-2776-e90886da1735&documentId=e4a8153f-e14a-3a02-a647-dfbbb59f5582

[2] P. A. Belinda and N. Costari, "Pentingnya Implementasi Akuntansi Sektor Publik Dalam Suatu Instansi Pemerintahan," *Jamanta J. Mhs. Akunt. Unita*, vol. 1, no. 1, pp. 58–77, 2021, doi: 10.36563/jamanta_unita.v1i1.421.

[3] D. Kusmayadi, D. Rudiana, and J. Badruzaman, "Good Coorporate Governance," p. 249, 2015.

[4] "PERATURAN MENTERI SEKRETARIS NEGARA REPUBLIK INDONESIA NOMOR 2 TAHUN 2019".

[5] A. N. Arifah, J. Suprijadi, and I. Ginanjar, "Klasifikasi Rumpun Jabatan ASN Berdasarkan Riwayat Pelatihan Menggunakan Multiclass Support Vector Machine," *J. Stat. Teor. dan Apl.*, vol. 1, no. 1, pp. 191–197, 2022, [Online]. Available: http://prosiding.statistics.unpad.ac.id

[6] B. C. Kharisma, B. S. W. P, and A. Widiharini, *Information Recommendation System For Jobs With Cosine Similarity & Rabin Karp K-Gram Metode*, vol. 1. 2021.

[7] D. B. Elnursa, V. Nofriana, A. Syamsuri, and L. Cahyani, "Sistem Rekomendasi Pemilihan Program MSIB Bagi Mahasiswa Pendidikan Informatika," *J. SHIFT*, vol. 2, no. April, pp. 1–4, 2023.

[8] R. R. Anugrah, "PENERAPAN COSINE SIMILARITY DAN PEMBOBOTAN TF-IDF UNTUK KLASIFIKASI PENGADUAN MASYARAKAT BERBASIS WEB (Studi Kasus : BAGWASSIDIK DITRESKRIMUM POLDA KALBAR)," *Coding J. Komput. dan Apl.*, vol. 11, no. 01, pp. 100–109, 2023.

[9] K. N. Artawan, R. S. Hartati, Y. Divayana, and M. Sudarma, "Perancangan Fitur Deteksi Kemiripan Dokumen Jawaban Tugas Mahasiswa pada Sistem Manajemen Pembelajaran dengan Metode K-Shingling dan Cosine Similarity," *Maj. Ilm. Teknol. Elektro*, vol. 22, no. 1, p. 45, 2023, doi: 10.24843/mite.2023.v22i01.p06.

[10] A. P. Muria, H. Sujaini, and H. S. Pratiwi, "Sistem Rekomendasi Artikel sebagai Acuan Studi Literatur Menggunakan Metode N-Gram," *JURISTI (Jurnal Ris. Sains dan Teknol. Inform.*, vol. 01, no. 1, pp. 69–84, 2023, doi: 10.26418/juristi.v1i1.61170.

[11] R. N. Dewi, "Model Text Mining Untuk Identifikasi Keluhan Pelanggan Produk Perusahaan Perangkat Lunak," *Univ. Islam Indones.*, 2018, [Online]. Available: https://dspace.uii.ac.id/handle/123456789/10239

[12] J. S. S. Supriyanto, "Sistem Informasi Gempa Bumi Menggunakan Data Xml Berbasis Pengolahan Teks Parsing Dan Concatenation," *Telematika*, vol. 16, no. 1, p. 36, 2019, doi: 10.31315/telematika.v16i1.3022.

[13] M. F. Riyadhi, "Aplikasi Text Mining Untuk Automasi Penentuan Tren Topik Skripsi Dengan Metode K-Means Clustering (Studi Kasus: Prodi Sistem Komputer)," *Komputika J. Sist. Komput.*, vol. 8, no. 2, pp. 59–64, 2019.

[14] P. C. Siswipraptini, M. I. Wafit, and K. Djunaidi, "Klasifikasi Pekerjaan Bidang Teknologi Informasi Menggunakan Algoritma Cosine Similarity," *KILAT*, vol. 12, no. 1, pp. 38–48, 2023.

[15] M. R. Arifuddin, I. Ar Rafiq, R. Mubarok, and P. H. Susilo, "Sistem Cerdas Penilaian Ujian Essay Menggunakan Metode Cosine Similarity," *Gener. J.*, vol. 7, no. 1, pp. 31–38, 2023, doi: 10.29407/gj.v7i1.18318.

[16] R. T. Wahyuni, D. Prastiyanto, and E. Supraptono, "Penerapan Algoritma Cosine Similarity dan Pembobotan TF-IDF pada Sistem Klasifikasi Dokumen Skripsi," *J. Tek. Elektro Univ. Negeri Semarang*, vol. 9, no. 1, pp. 18–23, 2017, [Online]. Available: https://journal.unnes.ac.id/nju/index.php/jte/article/download/10955/6659

[17] B. Darmawan, A. Dwi Laksito, M. Resa Arif Yudianto, and A. Sidauruk, "Analisis Perbandingan Ekstraksi Fitur Teks pada Sentimen Analisis Kenaikan Harga BBM," *Krea-TIF J. Tek. Inform.* , vol. 11, no. 1, pp. 53–63, 2023, doi: 10.32832/krea-tif.v11i1.13819.

[18] D. Marta, G. L. Ginting, and A. . H. Sihite, "Deteksi Berita Palsu Tentang Vaksinasi Covid-19 Dengan Menggunakan Text Mining Dan Algoritma Cosine Similarity," *KOMIK (Konferensi Nas. Teknol. Inf. dan Komputer)*, vol. 6, no. November, pp. 129–139, 2022, doi: 10.30865/komik.v6i1.5738.

[19] A. Kurnianti, P. Pahlevi, and I. Mufidah, "Recommendation System for Prospective Bride and Groom Using Cosine Similarity Algorithm," *Emerg. Inf. Sci. Technol.*, vol. 4, no. 1, pp. 8–15, 2023, doi: 10.18196/eist.v4i1.18683.

[20] N. I. Zakiyyatunisah, "ANALISIS SENTIMEN PADA MEDIA SOSIAL TWITTER TERHADAP KINERJA KABINET INDONESIA MAJU MENGGUNAKAN K-MEANS," *STMIK AKAKOM YOGYAKARTA.*, 2020.

[21] M. Martin and L. Nilawati, "Recall dan Precision Pada Sistem Temu Kembali Informasi Online Public Access Catalogue (OPAC) di Perpustakaan," *Paradig. - J. Komput. dan Inform.*, vol. 21, no. 1, pp. 77–84, 2019, doi: 10.31294/p.v21i1.5064.

[22] Habyba, A. N., Djatna, T., & Anggraeni, E. (2021). Positioning E-commerce Produk UKM berdasarkan Kebutuhan Afektif Pengguna. Krea-TIF: Jurnal Teknik Informatika, 9(1), 21–28. https://doi.org/10.32832/kreatif.v9i1.3590