PERBANDINGAN METODE COSINE DAN JACCARD SIMILARITY DALAM PENGUKURAN TINGKAT KEMIRIPAN DOKUMEN RENCANA KERJA PEMERINTAH DAERAH

Muhammad Haris Al Baihaqi, Arief Jananto
Sistem Informasi, Universitas Stikubank
Jalan Kendeng No 5 Bendanduwur, Kec Gajahmungkur Kota Semarang
haqihome@gmail.com

ABSTRAK

Evaluasi keselarasan dokumen Rencana Kerja Pemerintah Daerah (RKPD) antar provinsi masih dilakukan secara manual sehingga tidak efisien dan rentan inkonsistensi. Penelitian ini bertujuan membandingkan kinerja metode Jaccard Similarity dan Cosine Similarity berbasis TF-IDF dalam mengukur tingkat kemiripan dokumen RKPD, ditinjau dari skor kemiripan dan efisiensi waktu komputasi. Data penelitian berupa dua dokumen RKPD Tahun 2025 dari Provinsi Jawa Tengah dan Jawa Barat yang diekstraksi dari format PDF. Teks diproses melalui tahapan preprocessing meliputi case folding, tokenizing, stopword removal, dan stemming menggunakan Sastrawi, kemudian diukur kemiripannya dengan kedua metode. Hasil menunjukkan Cosine Similarity menghasilkan skor kemiripan sebesar 58,81% dengan waktu komputasi 0,2501 detik, sedangkan Jaccard Similarity menghasilkan skor 28,57% dengan waktu komputasi 0,0227 detik. Cosine Similarity berbasis TF-IDF lebih representatif untuk mengukur keselarasan tematik RKPD karena memperhitungkan bobot dan distribusi term, sementara Jaccard Similarity lebih efisien untuk pemeriksaan cepat kesamaan leksikal.

Kata kunci : *text mining, RKPD, TF-IDF, cosine similarity, jaccard similarity*

1. PENDAHULUAN

Transformasi digital dalam administrasi publik mendorong ketersediaan dokumen kebijakan dan perencanaan dalam format elektronik, sehingga membuka peluang analisis berbasis data pada korpus dokumen yang besar [1],[2]. Rencana Kerja Pemerintah Daerah (RKPD) merupakan dokumen perencanaan pembangunan tahunan yang memuat prioritas, program, dan arah kebijakan lintas sektor. Dokumen ini umumnya sangat panjang dan kaya terminologi birokratis sehingga evaluasi keselarasan antar dokumen RKPD secara manual sering tidak efisien dan rentan inkonsistensi penilaian [2],[3]. Pendekatan text mining telah banyak digunakan untuk menganalisis dokumen kebijakan dan perencanaan, termasuk untuk mengekstraksi pola tematik dan mendukung evaluasi berbasis bukti [1],[2],[3].

Salah satu tahap krusial dalam text mining adalah pengukuran kemiripan dokumen (document similarity) untuk mengidentifikasi tingkat kesamaan konten antar dokumen [10],[11]. Berbagai metrik dapat digunakan untuk menghitung kemiripan teks, namun dua metode yang umum dibandingkan adalah Jaccard Similarity dan Cosine Similarity karena perbedaan paradigma representasi, yaitu berbasis himpunan (set-based) dan berbasis vektor (vector-based) [4],[7],[8]. Jaccard Similarity menghitung rasio irisan terhadap gabungan himpunan kata unik tanpa mempertimbangkan frekuensi kemunculan kata [4]. Sebaliknya, Cosine Similarity menghitung kedekatan arah vektor dokumen dan sering dikombinasikan dengan pembobotan TF-IDF agar istilah yang informatif memperoleh bobot lebih tinggi [5],[9],[11]. Penelitian mutakhir pada teks berbahasa Indonesia menunjukkan bahwa Cosine Similarity berbasis TF-IDF sering memberikan hasil yang lebih baik untuk dokumen dengan variasi kosakata tinggi dibandingkan metode berbasis himpunan [7],[8],[9]. Namun, kajian spesifik pada dokumen perencanaan daerah yang panjang dan bergaya birokratis masih terbatas sehingga pemilihan metode untuk konteks kebijakan belum memiliki acuan yang kuat [1],[3].

Berdasarkan uraian tersebut, permasalahan yang diangkat dalam penelitian ini adalah belum diketahuinya metode pengukuran kemiripan yang lebih representatif antara Jaccard Similarity dan Cosine Similarity berbasis TF-IDF untuk dokumen RKPD yang bersifat formal, naratif, dan kaya terminologi teknis. Oleh karena itu, penelitian ini bertujuan membandingkan kinerja Jaccard Similarity dan Cosine Similarity berbasis TF-IDF dalam mengukur tingkat kemiripan dokumen RKPD, ditinjau dari skor kemiripan dan efisiensi waktu komputasi. Untuk mencapai tujuan tersebut, dilakukan implementasi pipeline text mining yang mencakup ekstraksi teks dari PDF, tahapan preprocessing (case folding, tokenizing, stopword removal, dan stemming menggunakan Sastrawi), serta perhitungan kemiripan dengan kedua metode pada dua dokumen RKPD Tahun 2025 dari Provinsi Jawa Tengah dan Jawa Barat [3],[11],[12].

2. TINJAUAN PUSTAKA

2.1. Rencana Kerja Pemerintah Daerah [RKPD]

RKPD dapat dipahami sebagai dokumen perencanaan pembangunan tahunan pemerintah daerah yang berisi prioritas, program, serta arah kebijakan lintas sektor, dan umumnya hadir sebagai dokumen naratif yang panjang sehingga menantang untuk ditelaah manual secara konsisten [1],[2],[3].

Dokumen perencanaan/urban planning yang panjang dan kaya terminologi seperti ini sering dianalisis menggunakan pendekatan text mining untuk mengekstraksi pola tematik dan membantu evaluasi berbasis data [1],[3].

2.2. Text Mining

Text mining adalah serangkaian teknik untuk mengubah teks tidak terstruktur menjadi representasi terukur agar bisa dianalisis secara komputasional [misalnya untuk menemukan pola, tema dominan, atau kesamaan antar dokumen] [1],[2],[3]. Dalam konteks dokumen kebijakan/perencanaan, text mining sering dipakai untuk memproses korpus dokumen skala besar dan menghasilkan indikator berbasis teks yang mendukung proses evaluasi dan pengambilan keputusan [1],[3].

2.3. Natural Language Processing [NLP]

NLP merupakan bidang yang memfokuskan pada pemrosesan bahasa alami sehingga teks dapat dibersihkan, direpresentasikan, dan dianalisis menggunakan metode komputasional untuk tugas seperti document similarity [4],[5],[6]. Untuk dokumen formal panjang, pendekatan "klasik" seperti preprocessing + pembobotan TF-IDF + cosine similarity masih sering dipakai karena relatif efisien, mudah diinterpretasi, dan kompetitif pada banyak skenario retrieval/kemiripan dokumen [4],[5],[9].

2.4. Text Preprocessing

Pada penelitian ini, text preprocessing dilakukan setelah teks RKPD diekstraksi dari PDF, dengan tujuan menormalkan dan membersihkan teks agar siap dihitung kemiripannya. [1][2] Tahapannya mencakup case folding [mengubah semua huruf menjadi lowercase agar bentuk kata konsisten], tokenizing [memecah teks menjadi token/kata], lalu filtering/stopword removal [menghapus kata-kata umum yang kurang informatif], sehingga noise berkurang dan fitur yang dipakai algoritma lebih "bersih" [1],[2],[3].

Setelah itu dilakukan stemming untuk mengubah kata berimbuhan ke bentuk dasar agar variasi morfologi bahasa Indonesia tidak membuat dua dokumen tampak "beda" padahal temanya sama [1],[4]. Secara keseluruhan, rangkaian preprocessing ini dipakai karena kualitas preprocessing berpengaruh langsung pada kualitas representasi dokumen dan hasil pengukuran kemiripan, baik untuk pendekatan berbasis himpunan [Jaccard] maupun berbasis vektor (TF-IDF + Cosine) [2],[5].

2.5. TF-IDF

TF-IDF adalah skema pembobotan yang menilai kepentingan term berdasarkan (i) seberapa sering term muncul di dokumen (TF) dan (ii) seberapa jarang term muncul di seluruh korpus [IDF], sehingga term yang spesifik memperoleh bobot lebih besar. [11][9] Pembobotan ini membantu representasi vektor menangkap sinyal tematik yang lebih kuat daripada hanya menghitung kemunculan kata secara biner [11],[6].

2.6. Jacccard Similarity

Jaccard Similarity mengukur kesamaan dua dokumen sebagai kesamaan dua himpunan token, yaitu rasio ukuran irisan terhadap ukuran gabungan token unik [6],[10]. Karena berbasis himpunan, Jaccard mengabaikan frekuensi kemunculan term sehingga cenderung lebih cocok untuk "cek cepat" kesamaan leksikal daripada menangkap intensitas topik pada dokumen panjang. [6],[10]. Nilai kemiripan Jaccard antara dua dokumen A dan B dihitung menggunakan persamaan [1]

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} \quad (1)$$

2.7. Cosine Similarity

Cosine Similarity mengukur kemiripan dua dokumen berdasarkan sudut antar vektor representasi dokumen, sehingga fokus pada arah [pola distribusi term] bukan sekadar panjang dokumen [6],[11]. Ketika dipadukan dengan TF-IDF, cosine similarity banyak dipakai untuk document similarity karena mempertimbangkan bobot term dan relatif robust terhadap perbedaan panjang dokumen [9],[11]. Nilai Cosine Similarity antara dua vektor dokumen D1 dan D2 dihitung menggunakan Persamaan [2].

$$\cos(\theta) = \frac{A \cdot B}{\|A\| \times \|B\|} \quad (2)$$

2.8. Penelitian Terdahulu

Beberapa penelitian sebelumnya telah membahas perbandingan metode similarity pada berbagai domain teks. Rinjeni et al. membandingkan Cosine Similarity dan Jaccard Similarity untuk pencocokan judul artikel ilmiah dengan 16 skenario eksperimen berbeda, dan menemukan bahwa Cosine Similarity secara konsisten menghasilkan nilai kemiripan tertinggi baik dengan maupun tanpa stemming, sehingga lebih unggul untuk matching judul akademik [13].

Pawestri dan Suyanto melakukan analisis komparatif kinerja Doc2Vec, Jaccard Coefficient, Cosine Similarity, dan Euclidean Distance untuk deteksi kemiripan dokumen berbahasa Indonesia menggunakan tiga dataset berbeda. Hasil penelitian menunjukkan Cosine Similarity memiliki performa terbaik secara keseluruhan dengan akurasi 0,98, presisi 0,84, recall 0,95, dan F-1 score 0,89 pada dataset Google News [7].

Henderi et al. membandingkan tiga metode similarity yaitu Cosine Similarity, Jaccard Similarity, dan Dice's Coefficient dalam konteks Automatic Short Answer Grading (ASAG). Penelitian tersebut menyimpulkan bahwa Cosine Similarity menghasilkan nilai kemiripan yang lebih tinggi dan lebih stabil untuk penilaian jawaban singkat dibandingkan kedua metode lainnya [8].

Widianto et al. mengimplementasikan TF-IDF dan Cosine Similarity untuk mendeteksi kemiripan dokumen dan menunjukkan bahwa kombinasi pembobotan TF-IDF dengan Cosine Similarity efektif dalam menangkap kemiripan konten pada dokumen teks [3].

Berbeda dari penelitian-penelitian tersebut, penelitian ini secara spesifik mengaplikasikan perbandingan Jaccard Similarity dan Cosine Similarity berbasis TF-IDF pada domain dokumen perencanaan pemerintah daerah (RKPD) yang memiliki karakteristik unik berupa teks formal, naratif panjang, dan kaya terminologi birokratis, sehingga memberikan kontribusi pada pemilihan metode similarity yang tepat untuk konteks kebijakan publik.

3. METODE PENELITIAN

3.1. Jenis dan Pendekatan Penelitian

Penelitian ini merupakan penelitian eksperimental dengan pendekatan kuantitatif yang bertujuan membandingkan kinerja dua metode pengukuran kemiripan dokumen, yaitu Jaccard Similarity dan Cosine Similarity, pada dokumen perencanaan pemerintah daerah (RKPD). Pendekatan kuantitatif digunakan karena keluaran utama yang dianalisis berupa data numerik, seperti skor similarity, waktu komputasi, dan statistik hasil preprocessing sehingga dapat dibandingkan secara objektif. Penelitian ini juga bersifat deskriptif–komparatif karena tidak hanya memaparkan hasil dari masing-masing metode, tetapi juga membandingkan keduanya untuk menentukan metode yang lebih sesuai pada karakter dokumen RKPD yang formal, terminologis, dan berukuran besar.

3.2. Instrumen Penelitian

Instrumen penelitian mencakup perangkat keras dan perangkat lunak yang digunakan untuk implementasi algoritma, pengolahan teks, perhitungan similarity, serta visualisasi hasil. Perangkat keras yang digunakan adalah:

Tabel 1. Hardware

Laptop	Axioo Hype 5 AMD
Processor	AMD Ryzen5 5500U
RAM	16 GB
Storage	SSD 256 GB
Operating System	Windows 11 Pro

Perangkat keras yang digunakan adalah laptop Axioo Hype 5 AMD dengan spesifikasi utama: prosesor AMD Ryzen 5 5500U, RAM 16 GB, penyimpanan SSD 256 GB, dan sistem operasi Windows 11 Pro, yang berfungsi sebagai platform utama untuk menjalankan seluruh tahapan penelitian mulai dari ekstraksi teks dokumen PDF, pemrosesan data (preprocessing), perhitungan similarity (Jaccard dan Cosine TF-IDF), hingga pembuatan visualisasi serta ekspor hasil analisis. Spesifikasi prosesor dan RAM digunakan untuk mendukung komputasi saat membangun representasi vektor TF-IDF dan menjalankan proses preprocessing pada dokumen berukuran besar, sedangkan SSD membantu mempercepat akses baca/tulis file [PDF, hasil tokenisasi, output tabel/grafik] sehingga alur kerja lebih efisien dan stabil. Sedangkan perangkat lunak yang digunakan yaitu:

Tabel 2. Software

Bahasa Pemrograman	Python
Development Environment	Jupyter Notebook
Package Manager	Anaconda

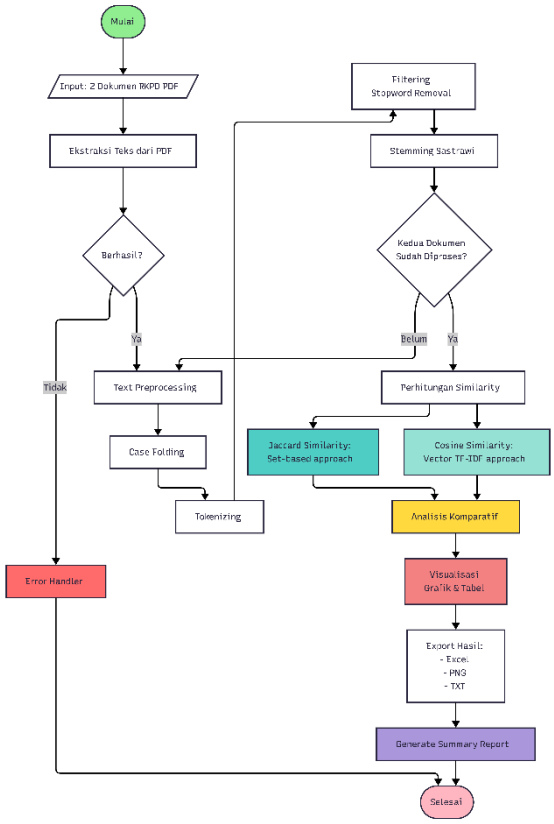
Perangkat lunak yang digunakan meliputi bahasa pemrograman, environment pengembangan, dan package manager, yaitu Python, Jupyter Notebook, dan Anaconda, yang berfungsi sebagai sarana implementasi algoritma serta pengolahan data penelitian secara terstruktur. Python digunakan sebagai bahasa pemrograman inti untuk membangun pipeline ekstraksi–preprocessing–similarity dan mengotomatisasi perhitungan skor kemiripan serta waktu komputasi, sementara Jupyter Notebook digunakan sebagai lingkungan kerja interaktif untuk menulis, menjalankan, dan mendokumentasikan kode beserta output [tabel/grafik] dalam satu naskah kerja. Anaconda berfungsi sebagai pengelola environment dan dependensi agar instalasi library yang diperlukan konsisten, meminimalkan konflik versi paket, serta memudahkan replikasi eksperimen ketika kode dijalankan ulang pada perangkat/lingkungan yang berbeda.

3.3. Pengumpulan Data

Data penelitian berupa dua dokumen RKPD Tahun 2025, yaitu RKPD Provinsi Jawa Tengah dan RKPD Provinsi Jawa Barat dalam format PDF yang diakses dari website resmi Bappeda setiap provinsi. Data yang sudah dikumpulkan selanjutnya masuk ke pipeline ekstraksi-preprocessing-similarity untuk memastikan kedua metode dibandingkan pada korpus yang sama.

3.4. Alur Proses Algoritma

Alur proses algoritma pada penelitian ini divisualisasikan dalam flowchart, dimulai dari input dua dokumen RKPD PDF, ekstraksi teks, kemudian preprocessing, dilanjutkan perhitungan similarity [dua jalur], lalu analisis komparatif dan visualisasi/ekspor hasil. Untuk kebutuhan perbandingan, jalur Jaccard menghitung kemiripan berbasis himpunan token unik (set-based), sedangkan jalur Cosine menghitung kemiripan berbasis vektor TF-IDF vector-based [6],[10],[11].



Gambar 1. Flowchart sistem

Flowchart di atas menunjukkan alur sistem untuk mengukur kemiripan dua dokumen RKPD berbentuk PDF, dimulai dari input dokumen lalu ekstraksi teks dari PDF agar konten dapat diproses sebagai data teks. Jika ekstraksi tidak berhasil, proses dialihkan ke error handler, sedangkan jika berhasil sistem masuk ke tahap text preprocessing untuk menyiapkan teks sebelum analisis.

Pada preprocessing, teks dinormalisasi melalui case folding, tokenizing, filtering/stopword removal, dan stemming menggunakan Sastrawi agar noise berkurang dan bentuk kata lebih seragam [4],[5]. Setelah kedua dokumen selesai diproses, sistem menghitung similarity menggunakan dua metode—Jaccard [set-based] dan Cosine dengan TF-IDF [vector-based]—kemudian melakukan analisis komparatif, visualisasi [grafik/tabel], ekspor hasil, dan menghasilkan laporan ringkas hingga proses selesai [6],[10],[11].

4. HASIL DAN PEMBAHASAN

4.1. Hasil Ekstraksi Awal Dokumen

Tahap awal penelitian adalah ekstraksi teks dari kedua dokumen RKPD (PDF) untuk memperoleh data teks mentah yang akan diproses lebih lanjut pada pipeline text mining. Proses ekstraksi berhasil mengambil konten naratif utama dari kedua dokumen, meskipun sebagian elemen non-teks seperti tabel dan gambar tidak selalu terambil sempurna.

Tabel 3. Statistik awal dokumen

Karakteristik	RKPD Jawa Tengah	RKPD Jawa Barat
Jumlah Halaman	1078	553
Jumlah Karakter	3,281,015	1,130,487
Jumlah kata [raw text]	388,288	142,347
Waktu ekstraksi [detik]	127.5	36.9

Tabel di atas menunjukkan bahwa RKPD Jawa Tengah memiliki volume dokumen dan jumlah kata lebih besar dibanding RKPD Jawa Barat, sehingga berpotensi memengaruhi waktu preprocessing serta kompleksitas analisis similarity pada tahap berikutnya.

4.2. Hasil Text Preprocessing

Penelitian Setelah ekstraksi, teks diproses melalui tahapan preprocessing untuk mengurangi noise dan menstandardisasi bentuk teks sebelum dihitung kemiripannya. [4],[5]. Tahapan yang digunakan meliputi case folding, tokenizing, filtering/stopword removal, dan stemming, sehingga dokumen lebih representatif ketika direpresentasikan sebagai himpunan token [untuk Jaccard] maupun vektor TF-IDF [untuk Cosine] [4],[5].

Untuk memperjelas pola perubahan ukuran data pada tiap tahap preprocessing, hasilnya divisualisasikan dalam grafik tahapan preprocessing pada Tabel 4

Tabel 4. Statistik hasil preprocessing

Tahap	RKPD Jateng	RKPD Jabar
Jumlah Kata Asli	388,288	142,347
Setelah Case Folding	388,288	142,347
Setelah Tokenizing	332,162	121,475
Setelah Filtering	298,774	105,105
Setelah Stemming	298,774	298,774
Reduksi Total [Persen]	23.05%	26.16%
Waktu Total [Detik]	593.7	236.9

Berdasarkan Tabel 4, preprocessing menghasilkan reduksi data sebesar 23,05% [RKPD Jateng] dan 26,16% [RKPD Jabar], yang menunjukkan efektivitas preprocessing dalam menghilangkan kata-kata tidak bermakna serta menekan variasi bentuk kata. [4],[5].

4.3. Hasil Jaccard Similarity

Pengujian kemiripan dokumen menggunakan Jaccard Similarity menghasilkan nilai kemiripan yang merepresentasikan kesamaan leksikal berdasarkan token unik yang sama pada kedua dokumen [10],[6]. Karena berbasis himpunan token unik, metode ini tidak memperhitungkan frekuensi kemunculan kata, sehingga cenderung lebih cocok untuk skrining kesamaan kosakata secara cepat [10],[6].

Tabel 5. Hasil perhitungan jaccard

Metrik	Nilai
Skor Jaccard	0.2857 (28.57%)
Kata unik RKPD Jateng	4,754
Kata unik RKPD Jabar	3,868
Irisan (A ∩ B)	1,916
Gabungan (A ∪ B)	6,706
Waktu komputasi (detik)	0.0227

Nilai Jaccard 28,57% menunjukkan bahwa proporsi kesamaan kata unik antar dokumen relatif rendah, yang dapat terjadi pada dokumen panjang dengan variasi redaksi tinggi meskipun memiliki tema pembangunan yang mirip [10],[6].

4.4. Hasil Cosine dengan TF-IDF

Pengujian Cosine Similarity menggunakan representasi vektor TF-IDF menghasilkan skor kemiripan yang lebih tinggi karena memperhitungkan bobot dan distribusi istilah penting dalam dokumen [11],[6],[9]. Pendekatan berbasis vektor seperti TF-IDF + Cosine umumnya lebih representatif untuk menangkap kemiripan tematik pada dokumen panjang dibanding pendekatan berbasis himpunan token [6],[11].

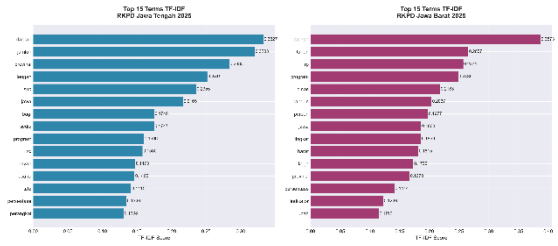
Tabel 6. Hasil perhitungan cosine

Metrik	Nilai
Skor Cosine Similarity	0.5881 (58.81%)
Waktu Komputasi [Detik]	0.2501
Jumlah Dimensi Vektor	5,000

Skor Cosine sebesar 58,81% mengindikasikan adanya kesamaan tematik yang cukup kuat antara RKPD Jawa Tengah dan Jawa Barat ketika istilah dipertimbangkan berdasarkan bobot TF-IDF, meskipun tidak selalu identik secara leksikal [11],[6].

4.5. Top 15 Terms TF-IDF

Untuk membantu interpretasi skor Cosine TF-IDF, dilakukan identifikasi istilah dengan bobot TF-IDF tertinggi pada masing-masing dokumen yang divisualisasikan pada Gambar 2. [11]

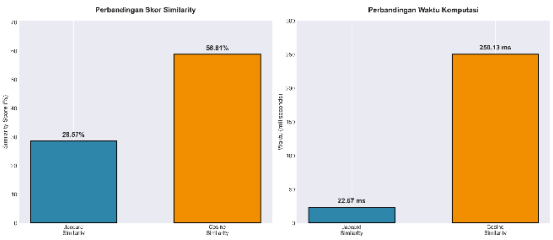


Gambar 2. Top terms tf-idf

Gambar 2 memperlihatkan bahwa RKPD Jawa Tengah lebih menonjolkan istilah seperti “daerah”, “jumlah”, dan “provinsi” yang menunjukkan fokus pada aspek kewilayahan dan kuantifikasi program, sedangkan RKPD Jawa Barat menonjolkan istilah seperti “daerah”, “tahun”, dan “rp” yang mengindikasikan penekanan pada horizon waktu dan alokasi anggaran. Visualisasi ini membantu menjelaskan mengapa Cosine TF-IDF cenderung menangkap kesamaan tematik melalui istilah berbobot tinggi, bukan sekadar menghitung kata unik yang sama [11],[6].

4.6. Analisis Perbandingan Kinerja Kedua Metode

Untuk merangkum perbedaan kinerja, skor similarity dan waktu komputasi kedua metode dibandingkan dalam satu visualisasi pada Gambar 3.



Gambar 3. Perbandingan skor

Gambar 3 menunjukkan bahwa Cosine TF-IDF menghasilkan skor kemiripan lebih tinggi dibanding Jaccard, namun membutuhkan waktu komputasi lebih lama karena proses vektorisasi dan perhitungan pada ruang berdimensi tinggi [11],[6],[10]. Sebaliknya, Jaccard lebih cepat karena hanya menghitung operasi himpunan token unik, sehingga cocok untuk pemeriksaan awal, sedangkan Cosine TF-IDF lebih sesuai untuk screening keselarasan tematik yang lebih representatif pada dokumen RKPD [10],[11],[8].

5. KESIMPULAN DAN SARAN

Berdasarkan hasil implementasi dan analisis perbandingan metode Jaccard Similarity dan Cosine Similarity pada dokumen RKPD Provinsi Jawa Tengah dan Jawa Barat Tahun 2025, dapat disimpulkan bahwa tahapan text preprocessing (case folding, tokenizing, filtering dengan Sastrawi, dan stemming) terbukti krusial karena mampu mereduksi volume kata sebesar 23,05% pada RKPD Jateng dan 26,16% pada RKPD Jabar sehingga noise berkurang dan fitur lebih informatif; pengukuran dengan Jaccard menghasilkan skor 0,2857 (28,57%) dengan waktu komputasi sangat cepat 0,0227 detik karena bersifat set-based dan tidak mempertimbangkan frekuensi, sedangkan Cosine Similarity berbasis TF-IDF menghasilkan skor lebih tinggi 0,5881 (58,81%) namun lebih lambat 0,2501 detik karena memperhitungkan bobot dan distribusi term pada ruang vektor berdimensi tinggi, sehingga untuk karakter RKPD yang formal, naratif, dan kaya terminologi teknis, Cosine TF-IDF lebih representatif untuk menangkap kemiripan tematik, sementara Jaccard cocok untuk skrining cepat; adapun saran pengembangan meliputi perbaikan cleaning/filtering yang lebih spesifik karena masih muncul token noise seperti “xx” (mis. dengan custom stopword dan pembersihan berbasis Regex untuk pola angka Romawi/artefak header-footer), pengujian pendekatan

semantik (mis. embedding atau model bahasa seperti IndoBERT) agar kemiripan substantif tetap tertangkap meski redaksi berbeda, analisis yang lebih granular per bab/per bidang urusan agar tidak bias perbandingan global, serta pengembangan sistem menjadi early warning system bagi Bappeda untuk mendeteksi inkonsistensi/duplikasi antar dokumen (mis. RKPD tahun berjalan vs tahun sebelumnya) guna meningkatkan kualitas penyusunan dokumen perencanaan.

DAFTAR PUSTAKA

[1] L. Hakim, D. Suryadi, and R. Fitriani, " Meninjau Peranan Text Mining sebagai Alat Strategis dalam Industri Kreatif melalui Sajian Webinar," *PIMAS: Jurnal Pengabdian Kepada Masyarakat*, vol. 4, no. 3, pp. 225–233, 2025, doi: 10.35960/pimas.v4i3.1868.

[2] M. Treviso et al., "Efficient Methods for Natural Language Processing: A Survey," *Transactions of the Association for Computational Linguistics*, vol. 11, pp. 826–860, 2023, doi: 10.1162/tacl_a_00577.

[3] A. Widianto, E. Pebriyanto, F. Fitriyanti, and M. Marna, "Document Similarity Using Term Frequency-Inverse Document Frequency Representation and Cosine Similarity," *J. Dinda: Data Sci. Inf. Technol. Data Anal*., vol. 4, no. 2, pp. 149-153, Aug. 2024, doi: 10.20895/dinda.v4i2.1589

[4] Y. Zhang, M. Qiao, Z. Peng, and D. Deng, "Near-Duplicate Text Alignment under Weighted Jaccard Similarity," *arXiv Prepr.*, vol. arXiv:2509.00627, Aug. 2025. [Online]. Available: https://arxiv.org/abs/2509.00627

[5] J. Halim and D. Lasut, "Document Plagiarism Detection Application Using Web-Based TF-IDF and Cosine Similarity Methods," *bit-Tech*, vol. 7, no. 2, pp. 202-213, 2024, doi: 10.32877/bt.v7i2.1697

[6] M. Akbar As Shiddiqi and A. Sanmarino, "Vector Space Model-based Information Retrieval Systems at South Sumatera Regional Libraries," *Journal of Computer Science Application and Engineering (JOSAPEN)*, vol. 1, no. 2, pp. 49–53, Jul. 2023, doi: 10.70356/josapen.v1i2.15

[7] S. Pawestri and Y. Suyanto, "Analisis Perbandingan Metode Similarity untuk Kemiripan Teks Bahasa Indonesia," *J. Media Inform. Budidarma*, vol. 8, no. 3, pp. 1540-1549, Jul. 2024, doi: 10.30865/mib.v8i3.7648

[8] H. Henderi, S. Wahyuningsih, and M. Rahardja, "Text Mining an Automatic Short Answer Grading (ASAG), Comparison of Three Methods of Cosine Similarity, Jaccard Similarity and Dice's Coefficient," Journal of Applied Data Sciences, vol. 2, no. 2, pp. 46–54, May 2021.

[9] M. Ulil Albab, Y. Karuniawati P., and M. N. Fawaiq, "Optimization of the Stemming Technique on Text Preprocessing President 3 Periods Topic," *Jurnal Transformatika*, vol. 20, no. 2, pp. 1–10, 2023, doi: 10.26623/transformatika.v20i2.5374

[10] M. Siino, I. Tinnirello, and M. La Cascia, "Is Text Preprocessing Still Worth the Time? A Comparative Survey on the Influence of Popular Preprocessing Methods on Transformers and Traditional Classifiers," *Inf. Syst*., vol. 121, p. 102342, 2024, doi: 10.1016/j.is.2023.102342

[11] J. Pardede and D. Darmawan, "Perbandingan Algoritma Stemming Porter, Sastrawi, Idris, Dan Arifin Setiono Pada Dokumen Teks Bahasa Indonesia," *Jurnal Teknologi Informasi dan Ilmu Komputer*, vol. 12, no. 1, pp. 69-76, 2025, doi: 10.25126/jtiik.2025121695

[12] M. Rosidin, M. F. Gustafi, and S. A. Pratiwi, "Optimizing nazief adriani's stemmer algorithm in detecting indonesian word errors using sastrawi," *Jurnal Mantik*, vol. 7, no. 3, pp. 1733–1742, Nov. 2023, doi: 10.35335/mantik.v7i3.4020.

[13] T. P. Rinjeni *et al*., "Matching Scientific Article Titles using Cosine Similarity and Jaccard Similarity," *Procedia Computer Science*, vol. 234, pp. 553–560, 2024, doi: 10.1016/j.procs.2024.03.039