Masuk: 10 April 2026 | Revisi: 17 Mei 2026 | Diterima: 29 Juni 2026
e-SSIN: 3031-3082; DOI: https://doi.org/10.26618/a6wkdt68

JUMPTECH

Implementasi Metode TF-IDF dan *Cosine Similarity* pada Sistem Pencarian Artikel yang Relevan

Selfira Madoa[1] | Ida Mulyadi[*2] | Darniati[2]

1 Mahasiswa Program Studi Informatika, Fakultas Teknik, Universitas Muhammadiyah Makassar, Indonesia.
Email: 105841111121@student.unismuh.ac.id

2 Program Studi Informatika, Fakultas Teknik, Universitas Muhammadiyah Makassar, Indonesia.
Email: idamulyadi@unismuh.ac.id; darniati@unismuh.ac.id

Korespondensi:
*Ida Mulyadi
idamulyadi@unismuh.ac.id

ABSTRAK

Perkembangan teknologi informasi menyebabkan peningkatan volume data teks digital, khususnya artikel ilmiah, yang menuntut adanya sistem pencarian informasi yang mampu menyajikan hasil secara relevan dan kontekstual. Pencarian berbasis pencocokan kata kunci secara literal dinilai belum optimal dalam menangani variasi bahasa dan konteks kueri. Oleh karena itu, penelitian ini bertujuan untuk mengimplementasikan dan mengevaluasi metode *Term Frequency–Inverse Document Frequency* (TF-IDF) dan *Cosine Similarity* pada sistem pencarian artikel ilmiah berbahasa Indonesia. Penelitian ini menggunakan pendekatan kuantitatif dengan metode eksperimen, di mana data berupa judul dan abstrak artikel diperoleh dari repositori digital terbuka. Tahapan *preprocessing* teks meliputi *case folding*, *tokenisasi*, *stopword removal*, dan *stemming* untuk meningkatkan kualitas representasi data. Hasil penelitian menunjukkan bahwa sistem mampu menghasilkan nilai *precision* hingga 0,75 dan *F1-score* sebesar 0,67, yang mengindikasikan bahwa metode TF-IDF dan *Cosine Similarity* efektif dalam meningkatkan relevansi hasil pencarian. Dengan demikian, sistem yang dikembangkan mampu memberikan hasil pencarian yang lebih akurat dan kontekstual dibandingkan metode pencarian berbasis kata kunci literal, serta layak diterapkan pada repositori artikel ilmiah berskala kecil hingga menengah.

Kata Kunci:
TF-IDF, *Cosine Similarity*, Sistem Pencarian Informasi, Artikel Ilmiah, *Text Mining*

ABSTRACT

The rapid growth of information technology has led to a significant increase in digital text data, particularly scientific articles, thereby requiring effective information retrieval systems capable of providing relevant and contextual results. Conventional keyword-based search methods are often insufficient in handling linguistic variations and complex query contexts. Therefore, this study aims to implement and evaluate the Term Frequency–Inverse Document Frequency (TF-IDF) and Cosine Similarity methods in an Indonesian scientific article search system. This research adopts a quantitative approach with an experimental method, using article titles and abstracts obtained from open-access digital repositories as the research dataset.Text preprocessing stages include case folding, tokenization, stopword removal, and stemming to improve data consistency and representation quality. The results indicate that the proposed system achieves a precision value of up to 0.75 and an F1-score of 0.67, demonstrating that the combination of TF-IDF and Cosine Similarity effectively enhances the relevance of search results. Thus, the developed system provides more accurate and contextual article retrieval compared to literal keyword matching and is suitable for implementation in small to medium-scale academic repositories.

Keywords:
TF-IDF, Cosine Similarity, Information Retrieval System, Scientific Articles, Text Mining

1 | PENDAHULUAN

Seiring dengan pesatnya pengembangan teknologi informasi dan komunikasi, volume data digital mengalami peningkatan yang sangat signifikan dari waktu ke waktu (Eskandari et al., 2021). Informasi dalam bentuk teks, gambar, maupun multimedia kini tersebar luas di berbagai platform digital, termasuk situs web, repositori ilmiah, dan media sosial (Lyu et al., 2021). Salah satu bentuk data yang paling dominan dan memiliki nilai strategis dalam dunia akademik adalah artikel ilmiah. Artikel tersebut

tidak hanya berfungsi sebagai sarana penyebaran pengetahuan, tetapi juga sebagai referensi utama dalam kegiatan riset dan pengambilan keputusan berbasis data (Bai et al., 2022). Sejalan dengan hal tersebut, sistem informasi juga terus mengalami peningkatan dalam hal kecepatan dan akurasi. Banyak platform telah menerapkan teknologi pencarian yang mampu merespons permintaan pengguna secara cepat dan menyajikan informasi relevan (Njoya et al., 2022). Di era digital, jumlah artikel dan data teks yang tersedia secara *online* meningkat pesat. Tanpa sistem pencarian yang efektif, pengguna akan kesulitan menemukan informasi yang relevan di antara sekian banyak data. Relevansi adalah kunci utama dalam sistem pencarian(Cahyani & Patasik, 2021).

Namun demikian, upaya untuk terus meningkatkan kualitas pencocokan antara kata kunci pengguna dan konten artikel tetap menjadi tantangan tersendiri, terutama ketika berhadapan dengan variasi bahasa, sinonim, atau konteks kueri yang kompleks. Kondisi ini menjadi semakin penting mengingat kebutuhan akan pencarian informasi yang akurat dan kontekstual sangat tinggi, khususnya di kalangan akademisi, mahasiswa, dan peneliti yang bergantung pada artikel ilmiah sebagai dasar pengembangan ilmu pengetahuan dan pengambilan keputusan. Oleh sebab itu, diperlukan pendekatan yang tidak hanya cepat, tetapi juga mampu memahami makna dan relevansi kueri pengguna. Oleh karena itu, pengembangan sistem pencarian artikel yang mampu menyajikan informasi secara relevan dan kontekstual menjadi isu penting yang perlu mendapatkan perhatian dalam bidang pengolahan informasi digital (Mohammed & Rashid, 2023), (Diningrat, 2025).

Dalam bidang pengembalian dan pencarian informasi berbasis teks, beberapa pendekatan telah dikembangkan untuk meningkatkan kualitas pencocokan antara kueri pengguna dan dokumen yang tersedia dalam basis data. Salah satu teknik yang paling banyak digunakan adalah *Term Frequency-Inverse Document Frequency* (TF-IDF), yaitu metode pembobotan kata yang mempertimbangkan seberapa sering suatu kata muncul dalam dokumen tertentu dibandingkan dengan seluruh koleksi dokumen. Bobot yang dihasilkan kemudian direpresentasikan dalam bentuk vektor untuk menggambarkan signifikansi kata terhadap konteks pencarian TF-IDF secara umum digunakan untuk memberikan bobot lebih tinggi pada kata-kata yang bersifat spesifik dan informatif, serta menurunkan bobot kata-kata umum yang kurang bermakna dalam dokumen (Yunanda et al., 2022), (Tiwari et al., 2023), (Widianto et al., 2024).

TF-IDF biasanya dipadukan dengan algoritma *Cosine Similarity* untuk mengukur tingkat kemiripan antara vektor dokumen dan kueri pencarian. *Cosine Similarity* mengukur nilai kosinus dari sudut antara dua vektor, yang mencerminkan seberapa dekat atau seberapa mirip isi dari dokumen terhadap permintaan pengguna. Kombinasi kedua metode ini telah terbukti memberikan hasil yang lebih akurat dan relevan pada berbagai aplikasi, seperti sistem pencarian destinasi wisata (Al Rasyid & Ningsih, 2024), klasifikasi judul artikel ilmiah (Rinjeni et al., 2024), hingga sistem pencarian berbasis *chatbot* layanan informasi (Hidayat et al., 2025).

Dalam pengembangannya, metode TF-IDF dan *Cosine Similarity* juga memerlukan tahap *preprocessing* yang baik, seperti normalisasi teks, *stopword removal*, *stemming,* dan *tokenization* (Abdurrafi & Ningsih, 2023). Tahapan ini bertujuan untuk meningkatkan kualitas representasi data sebelum dilakukan pembobotan dan perhitungan kemiripan. Studi yang menerapkan teknik ini menunjukkan peningkatan performa sistem pencarian yang mencapai rata-rata 83% pada pengujian berbagai jenis kueri (Aohana & Fitri Bimantoro, 2023). Hal ini memperkuat relevansi metode tersebut dalam konteks pencarian artikel yang relevan dan kontekstual (Huang, 2023). Studi menggunakan TF-IDF dalam model *Bag-of-Words* dan mengkombinasikannya dengan *embedding* untuk memperbaiki fungsi *Cosine Similarity* dalam melacak keterkaitan antar level persyaratan (Mohammad et al., 2024).

Berdasarkan uraian tersebut, dapat disimpulkan bahwa pengembangan sistem pencarian artikel yang relevan merupakan langkah strategis dalam menjawab tantangan pencarian informasi di era digital. Penerapan metode *Term Frequency-Inverse Document Frequency* (TF-IDF) dan *Cosine Similarity* dalam sistem pencarian diharapkan mampu meningkatkan akurasi dan relevansi hasil pencarian terhadap kebutuhan pengguna. Penelitian ini difokuskan pada implementasi kedua metode sebagai upaya untuk merancang sistem pencarian artikel yang tidak hanya efisien, tetapi juga responsif terhadap variasi kata kunci dan konteks kueri yang diberikan. TF-IDF berperan menyeleksi kata yang benar-benar penting, sedangkan *Cosine Similarity* mengukur kesesuaian kueri dengan artikel. Implementasi kedua metode tersebut diharapkan mampu menjawab kebutuhan pengguna untuk memperoleh hasil pencarian yang tidak hanya cepat, tetapi juga tepat dan relevan. *Novelty* pada studi ini menawarkan kebaruan berupa peningkatan akurasi hasil pencarian melalui pembobotan kata yang lebih bermakna serta pengukuran kesamaan yang terukur secara matematis. Hal ini menjadi keunggulan dibandingkan sistem pencarian berbasis pencocokan kata kunci yang hanya mencocokkan kata secara langsung. Selain itu, sistem ini berbeda dengan mesin pencari umum atau model *deep learning* yang kompleks, karena bersifat lebih sederhana, transparan, dan efisien sehingga mudah diimplementasikan untuk *dataset* terbatas, khususnya repositori artikel akademik.

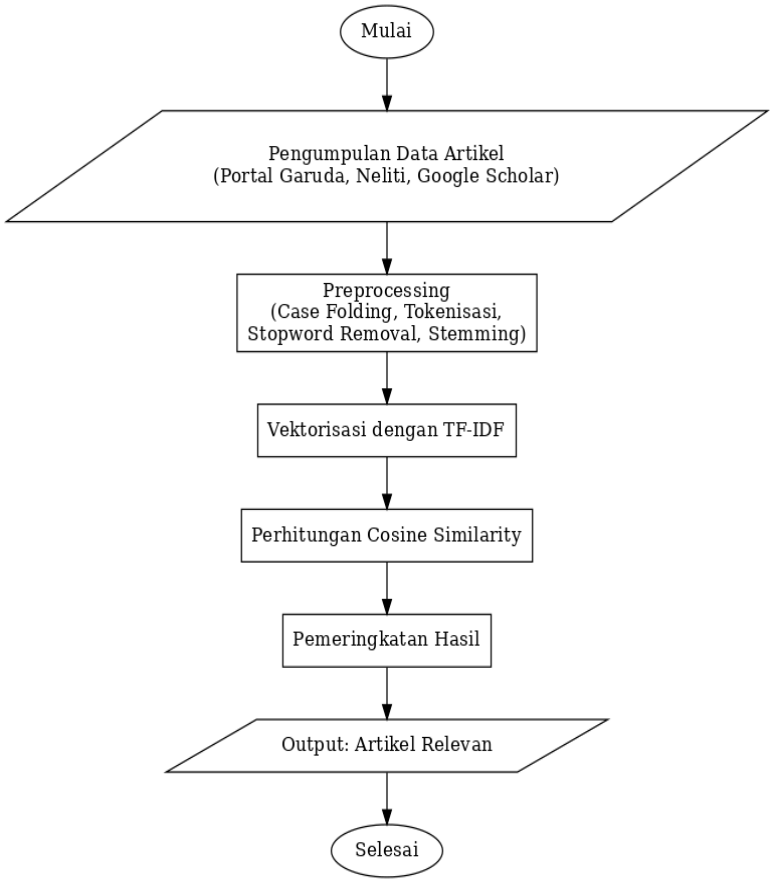
2 | METODE

2.1 | Jenis Pendekatan Penelitian

Penelitian ini merupakan penelitian kuantitatif dengan pendekatan eksperimen, yang bertujuan untuk merancang, mengimplementasikan, dan mengevaluasi sistem pencarian artikel ilmiah berbahasa Indonesia. Sistem dikembangkan menggunakan metode *Term Frequency–Inverse Document Frequency* (TF-IDF) sebagai teknik pembobotan teks dan *Cosine Similarity* sebagai metode pengukuran kemiripan antara kueri dan dokumen. Pendekatan eksperimen digunakan untuk membandingkan kinerja sistem pencarian berbasis metode yang diusulkan dengan pencarian konvensional berbasis padanan kata kunci (exact keyword matching), sehingga efektivitas metode dapat dianalisis secara objektif.

2.2 | Tahapan Metode Penelitian

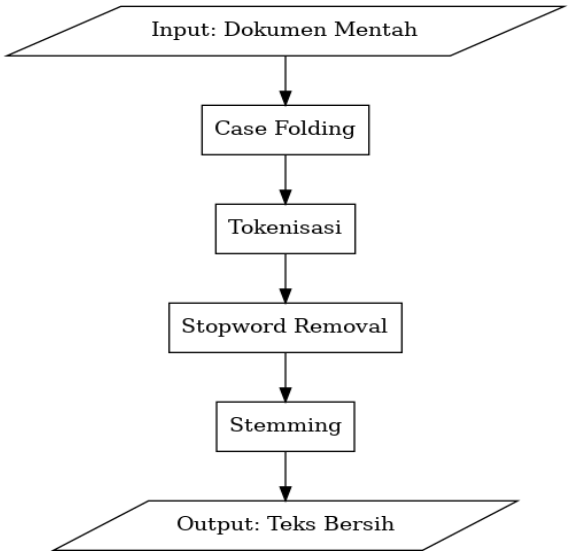
Penelitian ini menggunakan data sekunder berupa artikel ilmiah berbahasa Indonesia yang diperoleh dari repositori digital terbuka, antara lain Portal Garuda, Neliti, *Google Scholar*, dan SINTA. Dataset yang digunakan terdiri dari judul dan abstrak artikel, yang kemudian digabungkan menjadi satu kesatuan teks untuk setiap dokumen. Data tersebut dipilih karena merepresentasikan konten utama artikel dan umum digunakan dalam sistem temu kembali informasi. Metode penelitian menggambarkan tahapan sistematis yang digunakan dalam pengembangan dan evaluasi sistem pencarian artikel ilmiah.



GAMBAR 1. *Flowchart* Perancangan Sistem

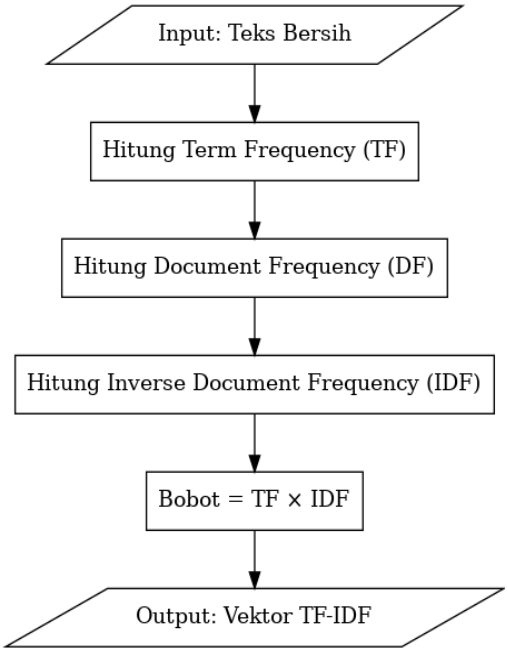
Proses dimulai dari tahap pengumpulan data artikel ilmiah berbahasa Indonesia yang diperoleh dari repositori digital terbuka seperti Portal Garuda, Neliti, *Google Scholar*, dan SINTA. Data yang dikumpulkan berupa judul dan abstrak artikel, yang kemudian diseleksi dan divalidasi untuk memastikan relevansi topik, kualitas akademik sumber, serta menghindari adanya data duplikat. Setelah data dinyatakan layak, dilakukan tahap pra-pemrosesan teks untuk meningkatkan kualitas dan konsistensi data. Teks hasil pra-pemrosesan selanjutnya direpresentasikan dalam bentuk vektor numerik menggunakan metode TF-IDF. Pada tahap berikutnya, kueri yang dimasukkan oleh pengguna juga melalui proses pra-pemrosesan dan vektorisasi yang sama. Vektor kueri kemudian dibandingkan dengan vektor dokumen menggunakan metode *Cosine Similarity* untuk menghitung tingkat kemiripan. Hasil perhitungan tersebut digunakan sebagai dasar pemeringkatan dokumen, sehingga artikel

dengan nilai kemiripan tertinggi ditampilkan sebagai hasil pencarian. Tahap akhir dari *flowchart* ini adalah evaluasi sistem, yang dilakukan dengan membandingkan kinerja pencarian berbasis TF-IDF dan *Cosine Similarity* terhadap pencarian berbasis padanan kata kunci.

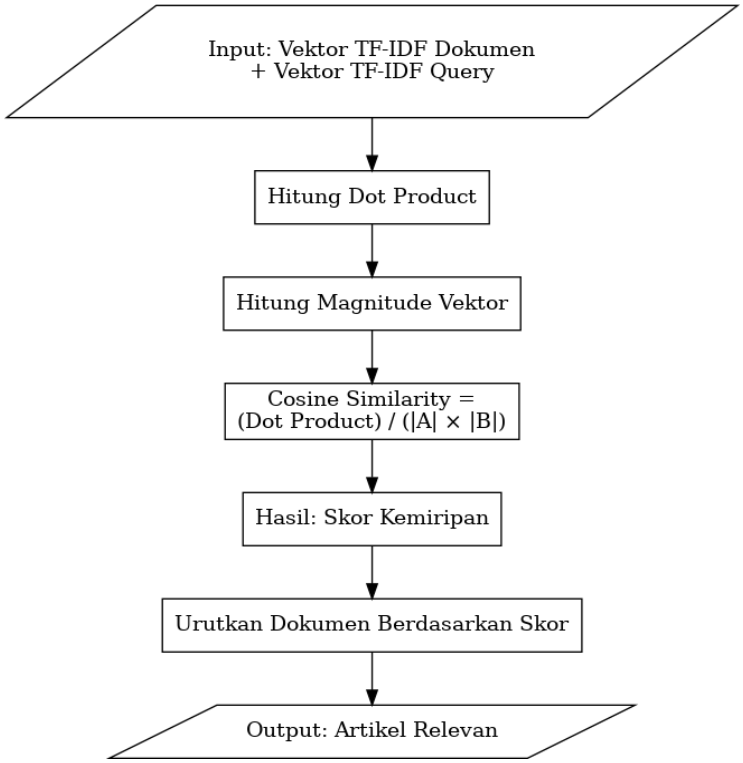


GAMBAR 2. *Flowchart Preprocessing* Teks

Tahapan pra-pemrosesan teks menjelaskan proses transformasi data teks mentah menjadi teks yang bersih, terstruktur, dan siap dianalisis secara matematis. Proses diawali dengan *case folding*, yaitu mengubah seluruh huruf pada teks menjadi huruf kecil untuk menghindari perbedaan makna akibat variasi kapitalisasi. Tahap selanjutnya adalah pembersihan karakter, di mana tanda baca, simbol, dan karakter *non-alfabet* yang tidak relevan dihapus. Setelah itu, dilakukan tokenisasi untuk memecah teks menjadi unit-unit kata yang lebih kecil. Tahap berikutnya adalah *stopword removal*, yaitu penghapusan kata-kata umum yang sering muncul namun memiliki nilai semantik rendah, sehingga dapat mengurangi noise pada data. Proses pra-pemrosesan diakhiri dengan *stemming*, yaitu mengembalikan setiap kata ke bentuk dasarnya agar variasi morfologis dapat disatukan. Hasil akhir dari *flowchart* ini adalah teks yang lebih ringkas, konsisten, dan representatif terhadap makna dokumen, sehingga mampu meningkatkan akurasi pembobotan TF-IDF dan perhitungan kemiripan dokumen.



GAMBAR 3. *Flowchart* Perhitungan TF-IDF



GAMBAR 4. *Flowchart* Perhitungan *Cosine Similarity*

Perhitungan TF-IDF menggambarkan proses pembobotan istilah yang digunakan untuk merepresentasikan dokumen dan kueri dalam bentuk vektor numerik. Tahapan dimulai dengan teks hasil pra-pemrosesan yang kemudian dianalisis untuk menghitung *Term Frequency* (TF), yaitu frekuensi kemunculan suatu istilah dalam sebuah dokumen. Selanjutnya dihitung *Document Frequency* (DF), yang menunjukkan jumlah dokumen dalam korpus yang mengandung istilah tersebut. Berdasarkan nilai DF, dihitung *Inverse Document Frequency* (IDF) untuk menurunkan bobot istilah yang terlalu umum dan meningkatkan bobot istilah yang jarang muncul namun bermakna. Bobot akhir suatu istilah diperoleh melalui perkalian nilai TF dan IDF, sehingga menghasilkan bobot TF-IDF.

Perhitungan TF-IDF menggambarkan proses pembobotan istilah yang digunakan untuk merepresentasikan dokumen dan kueri dalam bentuk vektor numerik. Proses dimulai dengan menghitung *Term Frequency* (TF), yaitu tingkat kemunculan suatu istilah dalam sebuah dokumen, yang dirumuskan sebagai:

$$TF(t,d) = \frac{f(t,d)}{\sum_k f(k,d)} \quad (1)$$

di mana *f*(t,*d)* menyatakan frekuensi istilah *t* dalam dokumen *d*. Selanjutnya dihitung *Document Frequency* (DF), yaitu jumlah dokumen dalam korpus yang mengandung istilah tertentu, yang dirumuskan sebagai:

$$DF(t) = |\{d \in D : t \in d\}| \quad (2)$$

Berdasarkan nilai DF, dihitung *Inverse Document Frequency* (IDF) untuk menurunkan bobot istilah yang sering muncul di banyak dokumen dan meningkatkan bobot istilah yang lebih spesifik, dengan rumus:

$$IDF(t) = \log \vdash(\{N\}\{DF(t)\}\dashv)\} \quad (3)$$

di mana *N* adalah jumlah total dokumen dalam korpus. Bobot akhir TF-IDF diperoleh dari perkalian TF dan IDF, yaitu:

$$TF\text{-}IDF(t,d) = TF(t,d) \times IDF(t) \quad (4)$$

Hasil perhitungan ini berupa vektor TF-IDF yang merepresentasikan tingkat kepentingan setiap istilah dalam dokumen relatif terhadap keseluruhan korpus.

Perhitungan *Cosine Similarity* menjelaskan mekanisme pengukuran tingkat kemiripan antara dokumen dan kueri berdasarkan representasi vektor TF-IDF. Proses diawali dengan menghitung *dot product* antara vektor dokumen dan vektor kueri untuk mengetahui kesesuaian arah kedua vektor tersebut. Selanjutnya, masing-masing vektor dinormalisasi dengan menghitung norma vektor guna menghindari bias akibat perbedaan panjang dokumen. Nilai *Cosine Similarity* kemudian dihitung menggunakan rumus:

$$Cosine\ Similarity(d,q) = \{\sum\_\{i = 1\}\^\{n\}\ d\_i\ \times q\_i\}\sqrt{\sum_{i=1}^{n} d_i^2} \times \sqrt{\sum_{i=1}^{n} q_i^2} \quad \text{.....(5)}$$

di mana d_i dan q_i masing-masing merupakan bobot TF-IDF istilah ke-i pada dokumen dan kueri. Nilai *Cosine Similarity* berada pada rentang 0 hingga 1, di mana nilai yang mendekati 1 menunjukkan tingkat kemiripan semantik yang tinggi. Nilai inilah yang digunakan sebagai dasar pemeringkatan dokumen, sehingga artikel dengan tingkat relevansi tertinggi ditampilkan terlebih dahulu.

3 | HASIL DAN PEMBAHASAN

3.1 | Analisis Kinerja Sistem Berdasarkan Nilai *Precision*, *Recall*, dan *F1-Score*

Untuk mengetahui tingkat kinerja sistem pencarian artikel ilmiah yang dikembangkan, dilakukan evaluasi menggunakan metrik *precision*, *recall*, dan *F1-score*. Evaluasi ini bertujuan untuk mengukur tingkat ketepatan, kelengkapan, serta keseimbangan hasil pencarian terhadap kueri yang diberikan. Hasil pengukuran kinerja sistem untuk setiap kueri uji disajikan pada **TABEL 1**.

TABEL 1. Hasil Evaluasi Kinerja Sistem

Kueri	*Precision*	*Recall*	*F1-Score*
Pendidikan	0.60	0.50	0.55
Analisis	0.67	0.57	0.62
sistem informasi	0.75	0.60	0.67

Berdasarkan hasil evaluasi kinerja yang disajikan pada Tabel 1, sistem pencarian artikel ilmiah menunjukkan performa yang cukup baik dalam menampilkan dokumen yang relevan terhadap kueri pengguna. Pada kueri *"*pendidikan*"*, sistem menghasilkan nilai *precision* sebesar 0,60, recall 0,50, dan F1-score 0,55. Sementara itu, kueri *"*analisis*"* memperoleh nilai precision 0,67, *recall* 0,57, dan *F1-score* 0,62, serta kueri *"*sistem informasi*"* menghasilkan nilai *precision* tertinggi sebesar 0,75, *recall* 0,60, dan *F1-score* 0,67. Nilai *precision* yang lebih tinggi dibandingkan *recall* pada seluruh kueri menunjukkan bahwa sistem cenderung lebih akurat dalam menampilkan artikel yang relevan, meskipun belum sepenuhnya optimal dalam menjangkau seluruh artikel relevan yang tersedia. Nilai *F1-score* yang berada pada rentang 0,55–0,67 mengindikasikan adanya keseimbangan yang cukup baik antara tingkat ketepatan dan kelengkapan hasil pencarian. Dengan demikian, dapat disimpulkan bahwa penerapan metode TF-IDF dan *Cosine Similarity* mampu memberikan kinerja pencarian yang stabil dan cukup efektif dalam konteks pencarian artikel ilmiah berbahasa Indonesia.

3.2 | Analisis Pengaruh Spesifisitas Kueri terhadap Akurasi Pencarian

Hasil pengujian menunjukkan bahwa tingkat spesifisitas kueri memiliki pengaruh signifikan terhadap kualitas hasil pencarian. Kueri yang bersifat spesifik atau berupa frasa, seperti *"*sistem informasi*"*, menghasilkan nilai *precision* 0,75 dan *F1-score* 0,67, yang lebih tinggi dibandingkan kueri tunggal yang bersifat umum, seperti *"*pendidikan*"*, dengan *precision* 0,60 dan *F1-score* 0,55. Perbedaan ini menunjukkan bahwa kueri frasa mampu membatasi ruang pencarian secara lebih efektif, sehingga meningkatkan bobot istilah yang relevan dalam proses pembobotan TF-IDF. Sebaliknya, kueri tunggal cenderung muncul dalam berbagai konteks dokumen yang berbeda, sehingga meningkatkan kemungkinan munculnya dokumen yang hanya relevan secara parsial. Temuan ini mengindikasikan bahwa sistem pencarian berbasis TF-IDF dan *Cosine Similarity* bekerja lebih optimal ketika pengguna memasukkan kueri yang memiliki konteks yang jelas dan terfokus. Oleh karena itu, kualitas

hasil pencarian tidak hanya dipengaruhi oleh metode yang digunakan, tetapi juga oleh karakteristik kueri yang diberikan oleh pengguna.

3.3 | Analisis Keunggulan dan Keterbatasan Sistem Berdasarkan Nilai Akurasi

Sistem pencarian artikel ilmiah yang dibangun menunjukkan tingkat akurasi yang cukup baik, dengan nilai *precision* maksimum mencapai 75% dan nilai *F1-score* tertinggi sebesar 67%. Nilai tersebut menunjukkan bahwa sistem mampu mengidentifikasi dan memprioritaskan artikel yang relevan secara konsisten pada peringkat teratas hasil pencarian. Keunggulan ini mencerminkan efektivitas pendekatan ruang vektor yang mengombinasikan pembobotan TF-IDF dan pengukuran *Cosine Similarity* dalam menangkap kesamaan konteks antara kueri dan dokumen. Namun demikian, nilai *recall* yang masih berada pada rentang 50%–60% menunjukkan bahwa sistem belum sepenuhnya mampu menampilkan seluruh artikel relevan yang terdapat dalam korpus. Keterbatasan ini berkaitan dengan sifat TF-IDF yang bersifat statistik dan belum mampu merepresentasikan hubungan semantik secara mendalam. Dengan demikian, meskipun sistem telah menunjukkan kinerja yang memadai untuk *dataset* berskala kecil hingga menengah, pengembangan lebih lanjut dengan pendekatan berbasis representasi semantik diharapkan dapat meningkatkan nilai *recall* dan akurasi sistem secara keseluruhan.

4 | KESIMPULAN

Berdasarkan hasil penelitian dan analisis yang telah dilakukan, dapat disimpulkan bahwa penerapan metode *Term Frequency–Inverse Document Frequency* (TF-IDF) dan *Cosine Similarity* efektif dalam meningkatkan kualitas sistem pencarian artikel ilmiah. Representasi dokumen dan kueri dalam bentuk vektor TF-IDF serta pengukuran tingkat kemiripan menggunakan *Cosine Similarity* memungkinkan sistem menyajikan artikel yang relevan secara kontekstual dan lebih akurat dibandingkan pencarian berbasis kata kunci literal, dengan nilai *precision* tertinggi mencapai 0,75, *recall* sebesar 0,60, dan *F1-score* maksimum sebesar 0,67. Selain itu, tahapan pra-pemrosesan teks yang meliputi *case folding*, t*okenisasi*, *stopword removal*, dan *stemming* berperan penting dalam meningkatkan konsistensi representasi data, sehingga perhitungan bobot TF-IDF dan nilai *Cosine Similarity* dapat dilakukan secara optimal. Kombinasi antara pra-pemrosesan teks yang tepat dan metode pengukuran kemiripan berbasis vektor terbukti menjadi faktor utama dalam menghasilkan sistem pencarian artikel ilmiah yang relevan, efektif, dan objektif.

Daftar Pustaka

Abdurrafi, M. F., & Ningsih, D. H. U. (2023). Content-based filtering using cosine similarity algorithm for alternative selection on training programs. *Journal of Soft Computing Exploration*, *4*(4), 204–212. https://doi.org/10.52465/joscex.v4i4.232

Al Rasyid, R., & Ningsih, D. H. U. (2024). Penerapan Algoritma TF-IDF dan Cosine Similarity untuk Query Pencarian Pada Dataset Destinasi Wisata. *Jurnal JTIK (Jurnal Teknologi Informasi Dan Komunikasi)*, *8*(1), 170–178. https://doi.org/10.35870/jtik.v8i1.1416

Aohana, M. R., & Fitri Bimantoro. (2023). Tourism Destination Article Search Features using TF-IDF and Cosine similarity. *Dielektrika*, *10*(2), 145–153. https://doi.org/10.29303/dielektrika.v10i2.338

Bai, Y. et al. (2022). A new data mining method for time series in visual analysis of regional economy. *Information Processing & Management*, *59*(1), 102741. https://doi.org/https://doi.org/10.1016/j.ipm.2021.102741

Cahyani, D. E., & Patasik, I. (2021). Performance comparison of tf-idf and word2vec models for emotion text classification. *Bulletin of Electrical Engineering and Informatics*, *10*(5), 2780–2788. https://doi.org/10.11591/eei.v10i5.3157

Diningrat. (2025). *Optimasi Algoritma Pencarian Dokumen Akademik Menggunakan BM25 dan TF-IDF*. *May*, 0–5. https://doi.org/10.13140/RG.2.2.15689.66401

Eskandari, L. et al. (2021). I-Scheduler: Iterative scheduling for distributed stream processing systems. *Future Generation Computer Systems*, *117*, 219–233. https://doi.org/https://doi.org/10.1016/j.future.2020.11.011

Hidayat, L. R. et al. (2025). *Optimalisasi Layanan Sistem Informasi Mahasiswa dengan Integrasi Telegram : Chatbot Retrieval-Augmented-Generation berbasis Large Language Model (Optimization of Student Information System Services with Telegram Integration : Chatbot Retrieval-Augmented* . *7*(1), 121–131.

Huang, R. (2023). Improved content recommendation algorithm integrating semantic information. *Journal of Big Data*, *10*(1). https://doi.org/10.1186/s40537-023-00776-7

Lyu, D. et al. (2021). Robot path planning by leveraging the graph-encoded Floyd algorithm. *Future Generation Computer Systems*, *122*, 204–208. https://doi.org/https://doi.org/10.1016/j.future.2021.03.007

Mohammad, B. et al. (2024). *Cross-level Requirement Traceability: A Novel Approach Integrating Bag-of-Words and Word Embedding for Enhanced Similarity Functionality*. 1–11.

Mohammed, M. T., & Rashid, O. F. (2023). Document retrieval using term frequency inverse sentence frequency weighting scheme. *Indonesian Journal of Electrical Engineering and Computer Science*, *31*(3), 1478–1485. https://doi.org/10.11591/ijeecs.v31.i3.pp1478-1485

Njoya, A. N. et al. (2022). Power-saving system designs for hexagonal cell based wireless sensor networks with directional transmission. *Journal of King Saud University - Computer and Information Sciences*, *34*(10, Part A), 7911–7919. https://doi.org/https://doi.org/10.1016/j.jksuci.2022.07.008

Rinjeni, T. P. et al. (2024). Matching Scientific Article Titles using Cosine Similarity and Jaccard Similarity Algorithm. *Procedia Computer Science*, *234*(2023), 553–560. https://doi.org/10.1016/j.procs.2024.03.039

Tiwari, S. et al. (2023). Innovative Framework for Context-based Recommendation using TF-IDF and Cosine Similarity. *International Journal of Innovative Research in Science,Engineering and Technology*, *12*(06), 9019–9023. https://doi.org/10.15680/ijirset.2023.1206157

Widianto, A. et al. (2024). Document Similarity Using Term Frequency-Inverse Document Frequency Representation and Cosine Similarity. *Journal of Dinda : Data Science, Information Technology, and Data Analytics*, *4*(2), 149–153. https://doi.org/10.20895/dinda.v4i2.1589

Yunanda, G. et al. (2022). Recommendation System from Microsoft News Data using TF-IDF and Cosine Similarity Methods. *Building of Informatics, Technology and Science (BITS)*, *4*(1), 277–284. https://doi.org/10.47065/bits.v4i1.1670