Vol. 6 No. 2 Desember 2025 Hal : 513 - 523

E-ISSN: 2723-1089
P-ISSN: 2776-7779

Ekstraksi Kata Kunci Pada Portal Jurnal JATISI Menggunakan Metode TF-IDF dan *Cosine Similarity*

Michael Wijaya [1)]*, Hafiz Irsyad [2)]

1,2)Informatika , Fakultas Ilmu Komputer dan Rekayasa ,
Universitas Multi Data Palembang , Indonesia

*Coresponding Email: michaelwijaya_2226250046@mhs.mdp.ac.id, hafizirsyad@mdp.ac.id

Abstrak

Penelitian ini bertujuan untuk mengimplementasikan metode Term Frequency-Inverse Document Frequency (TF-IDF) dan Cosine Similarity dalam mengekstraksi kata kunci dari abstrak artikel jurnal guna mengidentifikasi topik penelitian dalam periode tertentu. Harapannya, pengguna dapat menganalisis tren topik dan memperoleh gambaran komprehensif mengenai dinamika keilmuan di bidang Teknik Informatika dan Sistem Informasi. Penelitian ini menggunakan pendekatan eksperimental kuantitatif dengan tahapan: pengumpulan data dari 11 artikel jurnal JATISI Vol 6 No. 1 (2019), pre-processing data (case folding, penghapusan tanda baca dan angka, tokenisasi, stopword removal, dan stemming), pembobotan menggunakan TF-IDF, serta pengukuran relevansi antardokumen dengan Cosine Similarity. Hasil penelitian berhasil mengekstraksi kata kunci dari dokumen dan memberikan peringkat berdasarkan persentase kemunculannya, serta menghasilkan matriks cosine similarity untuk mengidentifikasi kemiripan antartulisan. Namun, nilai presisi 0.05, recall 0.04, dan F1-score 0.04 menunjukkan bahwa model ini belum mampu memberikan prediksi yang memadai untuk kasus ini. Temuan ini dapat dijadikan acuan bahwa model/metode tersebut tidak direkomendasikan tanpa modifikasi signifikan, sekaligus menjadi dasar untuk eksplorasi solusi alternatif di masa depan.

Kata Kunci: *Term Frequency-Inverse Document* Frequency (TF-IDF), Cosine Similarity

Abstract

This study aims to implement Term Frequency-Inverse Document Frequency (TF-IDF) and Cosine Similarity methods to extract keywords from journal article abstracts, enabling the identification of research topics within a specific period. The expected outcome is that users can analyze topic trends and gain comprehensive insights into the academic dynamics in the fields of Informatics Engineering and Information Systems. This quantitative experimental research follows several stages: data collection from 11 articles in JATISI Vol 6 No. 1 (2019), data pre-processing (case folding, punctuation and number removal, tokenization, stopword removal, and stemming), weighting using TF-IDF, and document relevance measurement using Cosine Similarity. The research successfully extracted keywords from documents and ranked them based on occurrence percentage, while also producing a cosine similarity matrix to identify similarities between articles. However, the precision of 0.05, recall of 0.04, and F1-score of 0.04 indicate that this model is currently inadequate for making useful predictions in this case. These findings serve as a reference that the model/method is not recommended without significant modifications, while also providing a foundation for exploring alternative solutions in the future.

Keywords: Term Frequency-Inverse Document* Frequency (TF-IDF), *Cosine Similarity

PENDAHULUAN

Kata Kunci merupakan sebuah kata atau frasa yang dapat menjelaskan topik inti dari sebuah dokumen [1]. Kata kunci sendiri tidak terbatas pada kata - kata yang tercantum dalam isi dokumen atau judul dokumen saja, tetapi seluruh kata yang dapat menjelaskan dokumen tersebut dalam satu kata atau frasa. Teknologi untuk mengambil kata kunci pada suatu dokumen secara otomatis dinamakan "keyword extraction", dimana dengan menggunakan aturan dan rumus penambangan teks, kita dapat mendapatkan kata kunci dari suatu dokumen.

Perkembangan teknologi informasi juga memudahkan orang untuk mengakses dan mempublikasikan jurnal. Seiring meningkatnya jumlah publikasi, muncul kebutuhan untuk menganalisis tren penelitian yang berkembang di kalangan akademisi. Salah satu cara untuk memahami tren ini adalah dengan mengidentifikasi topik-topik penelitian yang dibahas pada periode waktu tertentu.

Kata kunci seringkali merupakan topik dari suatu artikel, sehingga ekstraksi kata kunci memegang peran penting dalam proses analisis dari tren penelitian. Dengan penggunaan TF-IDF (*Term Frequency-Inverse Document Frequency*), sebuah istilah dapat diukur seberapa istilah tersebut muncul dan seberapa penting istilah tersebut dalam kumpulan dokumen. *Cosine Similarity* akan membantu dalam mengelompokkan dokumen-dokumen yang memiliki topik yang sama / mirip, sehingga dapat terlihat topik dominan yang yang diteliti [2].

Dalam konteks portal jurnal seperti JATISI (Jurnal Teknik Informatika dan Sistem Informasi) yang berasal dari Universitas Multi Data Palembang, analisis tren penelitian melalui ekstraksi kata kunci dapat memberikan gambaran komprehensif mengenai dinamika keilmuan di bidang Teknik Informatika dan Sistem Informasi. Pengelola jurnal dapat memanfaatkan informasi ini untuk mengevaluasi arah kebijakan penerbitan, sementara institusi pendidikan dapat menggunakannya sebagai acuan dalam pengembangan kurikulum. Namun, tantangan utama dalam

analisis ini adalah menentukan representasi kata kunci yang benar-benar mencerminkan inti dari suatu penelitian, terutama ketika menghadapi variasi terminologi dan cakupan topik yang luas [3]

Penelitian ini bertujuan untuk mengimplementasikan metode TF-IDF dan *Cosine Similarity* dalam menganalisis tren topik penelitian yang dominan pada portal Jurnal JATISI. Dengan pendekatan ini, diharapkan pola-pola penelitian yang sedang berkembang dapat diidentifikasi. Hasil analisis ini dapat menjadi bahan pertimbangan bagi peneliti dalam menentukan topik penelitian baru, serta memberikan masukan berharga bagi pengelola jurnal dalam mengembangkan strategi penerbitan yang lebih terarah.

LANDASAN TEORI

Abstrak Artikel Ilmiah

Abstrak pada artikel junal, adalah sebuah paragraph singkat yang berisi maksimal 250 kata, dan dapat memberikan narasi singkat untuk menjelaskan isi jurnal. Abstrak sendiri telah memuat tujuan, metode, hasil, dan kesimpulan artikel [4].

Informational Retrieval

Informational Retrieval / pencarian informasi, proses pencarian, pengambilan, dan pengelolan informasi dari kumpulan dokumen atau sumber data yang tidak terstruktur berdasarkan kebutuhan pengguna. [5][6][7][8].

Term Frequency-Inverse Document Frecuency (TF-IDF)

Term Frequency-Inverse Document Frequency (TF-IDF), adalah metode untuk mengkalkulasi berat dari sebuah kata / fitur yang telah di ambil dari dokumen melalui pencarian informasi dan penambangan teks [9][10][11].

$$TF - IDF(t, d, D) = \left(\frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}}\right) \times \log\left(\frac{N}{n_t}\right) \ldots\ldots(1)$$

(1) Rumus TF-IDF

Keterangan:

t	: Kata (term) yang dihitung
d	: Dokumen spesifik
D	: Seluruh Kumpulan dokumen (korpus)
$f_{t,d}$	: Frekuensi kata t di dokumen d
$\sum_{t' \in d} f_{t',d}$	: Total semua kata di dokumen d
N	: Total jumlah dokumen dalam korpus D
n_t	: jumlah dokumen yang mengandung kata t

Cosine Similarity

Cosine Similarity, adalah metode untuk mengukur kemiripan antara 2 dokumen berbeda menggunakn kemiringan cosine diantara 2 vektor multi dimensi di dalam ruang multi dimensi tanpa terkekang oleh ukuran mereka [12][13][14].

$$\text{CosineSimilarity}(A, B) = \frac{A \cdot B}{\|A\| \times \|B\|} \text{.....(2)}$$

(2) Rumus Cosine Similarity

$$A \cdot B = \sum_{i=1}^{n} A_i \times B_i \text{.....(3)}$$

(3) Rumus Dot Product

$$\|A\| = \sqrt{\sum_{i=1}^{n} A_i^2}, \quad \|B\| = \sqrt{\sum_{i=1}^{n} B_i^2} \text{.....(4)}$$

(4) Rumus magnitude Vektor

Keterangan:

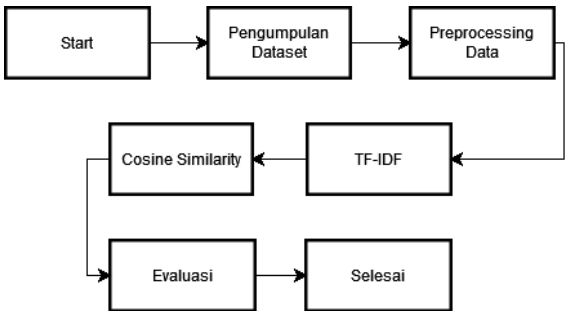
A, B	: Vektor dokumen (misal vector TF-IDF)
A_i, B_i	: Nilai TF-IDF kata ke-I di dokumen A dan B
N	: Jumlah kata unik dalam korpus

Text Mining

Text Mining, adalah kegiatan ekstraksi informasi dan corak yang implisit / teersirat, dari kumpulan banyak data / dokumen text yang tidak terstruktur, contohnya abstrak dari artikel jurnal [15][16].

METODE PENELITIAN

Penelitian ini bersifat kuantitatif eksperimental, dengan pendekatan *Text Mining* untuk mengekstraksi kata kunci secara otomatis dari artikel jurnal pada portal jurnal JATISI. Gambar 1 menunjukkan alur kerja yang akan dilakukan, dari pengumpulan dataset, hingga evaluasi model.



Gambar 1. Diagram Alur Penelitian

Penjelasan tahapan pada gambar 1 dapat dijabarkan sebagai berikut:

1. Pengumpulan Dataset

Pengumpulan dataset dilakukan dengan mengambil abstrak dari artikel jurnal JATISI Volume 6 Nomor 1 (2019). Data dapat dilihat pada Tabel 1.

Tabel 1. Data Abstrak Artikel Jurnal

No.	Judul	Abstrak	Doi
1	Evaluation of Online Portal Success With the DeLone and ...	*XYZ company has used the information system to survive and develop to...*	https://doi.org/10.35957/jatisi.v6i1.174
2	Aplikasi Media Pembelajaran Augmented Reality Pada...	Dalam penulisan kali ini penulis membahas tentang perancangan...	https://doi.org/10.35957/jatisi.v6i1.166
3	Aplikasi Perhitungan Dan Transaksi Penjualan Rumah...	*At present the need for housing is increasing in the community. This...*	https://doi.org/10.35957/jatisi.v6i1.162
4	Project Management System with Website Based On PT XYZ	*Project management is a project processing process that includes...*	https://doi.org/10.35957/jatisi.v6i1.160
5	Identifikasi Potensi Glaukoma dan Diabetes Retinopati...	*Glaucoma and diabetic retinopathy identification conduct form fundus...*	https://doi.org/10.35957/jatisi.v6i1.158

2. Pre-processing data

Tahap selanjutnya yang akan dilakukan adalah pre-processing data, yaitu dilakukan proses case folding, yang akan mengubah semua huruf menjadi *lowercase* atau huruf kecil. Setelah itu akan dilakukan proses penghilangan angka dan tanda baca, dan dilanjutkan dengan proses tokenisasi, *stopword* removal dan proses stemming. Pre-

processing sendiri dilakukan untuk membersihkan data teks dalam dokumen sehingga akan memudahkan proses penarikan kata kunci karena seluruh huruf / simbol yang tidak diperlukan telah dihilangkan.

3. *Term Frequency-Inverse Document Frequency* (TF-IDF)

Setelah tahap pre-processing, dokumen / query akan dilakukan proses TF-IDF Dimana pada proses ini dokumen / query akan direpresentasikan dalam bentuk vector. Proses TF-IDF sendiri akan melakukan perhitungan untuk memberikan bobot sebanding dengan seberapa pentingnya kata (term). Dengan dijalankannya TF-IDF, akan membantu proses untuk mengetahui kata kunci dari artikel jurnal / dokumen.

4. *Cosine Similarity*

Metode *Cosine Similarity* akan digunakan untuk mengukur kemiripan antara dokumen berdasarkan query. *Cosine similarity* akan menentukan kemiripan antar dokumen, dan dapat digunakan untuk melihat kemiripan topik antara 2 artikel jurnal.

HASIL DAN PEMBAHASAN

Berikut akan dibahas hasil dari penelitian yang telah dilakukan.

Pre-processing Data

Pada tahap ini dilakukan proses pre-processing pada dataset yang ada. Proses pertama yang akan dilakukan adalah case folding. Tabel 2 menunjukkan hasil case folding yang dilakukan pada data.

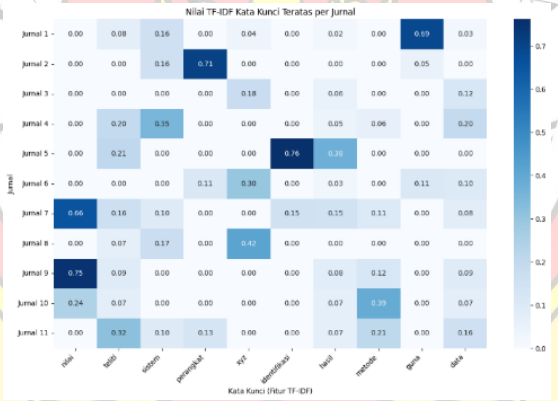
Tabel 2. Perbandingn abstrak sebelum dan sesudah pre-processing

No	Judul	Abstrak Asli	Abstrak Pre-process
1	Evaluation of Online Portal Success With…	Perusahaan XYZ telah memanfaatkan system…	usaha xyz manfaat sistem informasi tahan kembang sesuai kembang…

2	Aplikasi Media Pembelajaran…	Dalam penulisan kali ini penulis membahas tentang…	tulis kali tulis bahas ancang buat augmented reality objek dimensi…
3	Aplikasi Perhitungan Dan Transaksi…	Kebutuhan akan perumahan semakin meningkat di…	butuh rumah tingkat masyarakat peluang cepat ambil developer…
4	Project Management System with Website…	Manajemen proyek adalah suatu proses pengolahan proyek…	manajemen proyek proses olah proyek liput rencana organisasi atur...
5	Identifikasi Potensi Glaukoma dan…	Identifikasi potensi glaukoma dan retinopati diabetes dapat…	identifikasi potensi glaukoma retinopati diabetes citra fundus…

TF-IDF

Pada tahap ini dilakukan proses kalkulasi bobot untuk setiap kata (term) menggunakan TF-IDF, bobot diberikan berdasarkan seberapa sering term muncul pada satu dokumen dan seluruh dokumen. Gambar 2 menampilkan hasil TF-DF pada setiap term (kata)

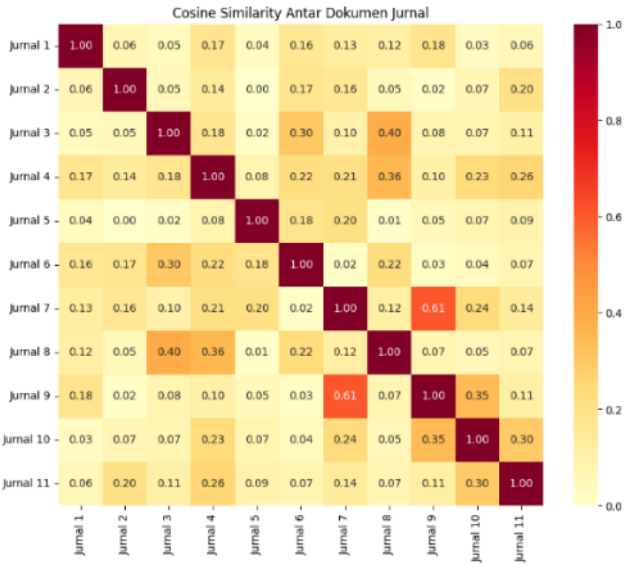


Gambar 2 Tabel Hasil TF-IDF

Nilai pada gambar diatas menunjukkan bobot pada term teratas seluruh dokumen, semakin besar nilai maka semakin penting term tersebut dalam dokumen.

Cosine Similarity

Setelah didapatkan hasil pembobotam TF-IDF, akan dilakukan kalkulasi *cosine similarity* untuk mendapatkan hasil kemiripan antar dokumen.



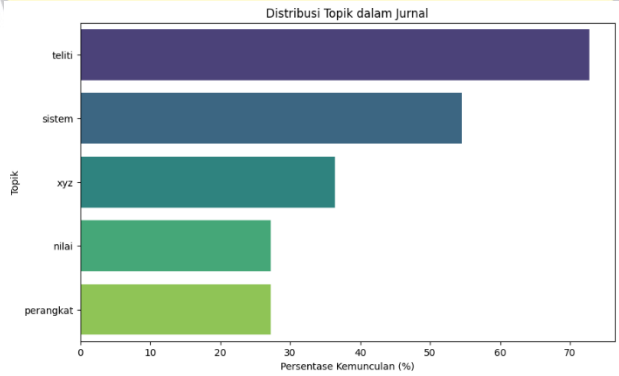
Gambar 3 Matriks *Cosine Similarity* antar dokumen

Hasil Analisis

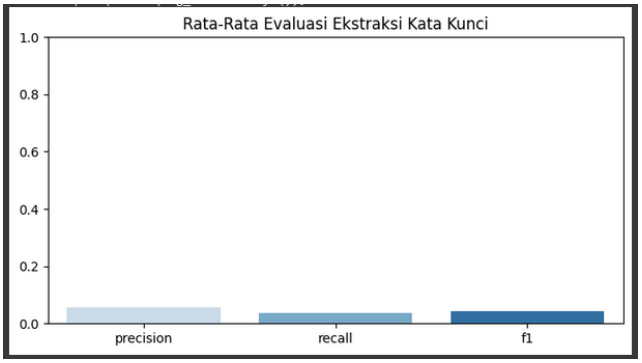
Dari kode yang dijalankan didapatkan hasil sebagai berikut:

Tabel 3 kata kunci teratas yang diambil program

No.	Topik	Persentase
1	Teliti	72.727273%
2	Sistem	54.646466%
3	XYZ	36.363636%
4	Nilai	27.272727%
5	Perangkat	27.272727%



Gambar 4 Distribusi Topik Dalam jurnal



Gambar 5 Rata-Rata Evaluasi Ekstraksi kata kunci

Pada gambar 4,5 dan table 3 diatas ditujukkan 5 kata kunci dengan persentase kemunculan teratas pada seluruh dokumen. Berdasarkan hasil "teliti" merupakan kata kunci yang paling sering muncul, dengan persentase kemunculan 72,72%, diikuti dengan "system" dengan persentase kemunculan 54,64%, "xyz" dengan persentase kemunculan 36,36%, "nilai" dan "perangkat" dengan persentase kemunculan 27,27%. Pada Gambar 5 dapat dilihat bahwa jika dibandingkan dengan kata kunci yang ditambahkan secara manual oleh penulis, program memiliki performa yang sangat buruk, dengan presisi 0.05, recall 0.04, dan F1 0.04, hasil evaluasi yang sangat buruk ini juga dikarenakan kata kunci yang biasanya ditambahkan pada abstrak menggunakan kata-kata yang tidak ada pada abstrak itu sendiri.

SIMPULAN

Penelitian ini berhasil mengimplementasikan metode TF-IDF dan Cosine Similarity untuk mengekstraksi kata kunci dari 11 abstrak jurnal JATISI. Hasilnya menunjukkan kata kunci teratas yaitu "teliti" (72.73%), "system" (54.55%), dan "XYZ" (36.36%), namun model memiliki kinerja rendah dengan presisi 0.05, recall 0.04, dan F1-score 0.04. Analisis cosine similarity mengungkap variasi kemiripan kata kunci antar artikel (nilai 0.1-0.8), tetapi ketidakmampuan model dalam menangkap kata kunci manual membatasi akurasinya. Temuan ini menyimpulkan bahwa pendekatan ini kurang efektif dalam ekstraksi kata kunci berbasis abstrak tanpa modifikasi, seperti integrasi Teknik NLP berbasis semantic atau perluasan dataset. Eksplorasi metode hybrid direkomendasikan untuk penelitian selanjutnya.

DAFTAR PUSTAKA

[1] T. Nomoto, "Keyword Extraction: A Modern Perspective," *SN Comput Sci*, vol. 4, no. 1, hlm. 92, Des 2022, doi: 10.1007/s42979-022-01481-7.

[2] P. Marto Hasugian, J. Manurung, L. Logaraz, dan U. Ram, "IMPLEMENTATION OF TF-IDF AND COSINE SIMILARITY ALGORITHMS FOR CLASSIFICATION OF DOCUMENTS BASED ON ABSTRACT SCIENTIFIC JOURNALS," *INFOKUM*, vol. 9, no. 2, June, hlm. 518–526, Agu 2021, [Daring]. Tersedia pada: https://infor.seaninstitute.org/index.php/infokum/ article/view/201

[3] T. Bin Sarwar, N. M. Noor, dan M. Saef Ullah Miah, "Evaluating keyphrase extraction algorithms for finding similar news articles using lexical similarity calculation and semantic relatedness measurement by word embedding," *PeerJ Comput Sci*, vol. 8, hlm. e1024, Jul 2022, doi: 10.7717/peerj-cs.1024.

[4] Admin, "Jurnal Ilmiah: Pengertian, Fungsi, Jenis, dan Struktur,"SampoernaUniversity, [Online]. Tersedia: https://www.sampoernauniversity.ac.id/id/news/contoh-jurnal-ilmiah. [Diakses: 20 Mei 2025].

[5] R. Ariyansyah, R. Nanda, dan O. Wiranda, "Search Engine Menggunakan Metode Information Retrival," 2022.

[6] R. Mandala, "Evaluasi Efektifitas Metode Machine Learning Pada Search Engine,"Keahlian Inform. Sekol. Tek. Elektro dan Inform., vol. 2006, no. Snati, pp. 11–15, 2006.

[7] I. T. Hapsari, B. S. Andoko, and C. Rahmad, "Aplikasi Information Retrieval UntukPencarian Dokumen Laporan Penelitian," J. Inform. Polinema, vol. 1, no. 3, p. 23, 2017, doi: 10.33795/jip.v1i3.109.

[8] N. M. A. Lestari and M. Sudarma, "Perencanaan Search Engine E-commerce dengan Metode Latent Semantic Indexing Berbasis Multiplatform," Lontar Komput. J. Ilm. Teknol. Inf., vol. 8, no. 1, p. 31, 2017, doi: 10.24843/lkjiti.2017.v08.i01.p04.

[9] H. J. Kim, J. W. Baek, dan K. Chung, "Optimization of associative knowledge graph using TF-IDF based ranking score," *Applied Sciences (Switzerland)*, vol. 10, no. 13, Jul 2020, doi: 10.3390/app10134590.

[10] Paik, J.H. A novel TF-IDF weighting scheme for effective ranking. In Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval, Dublin, reland, 28 July–1 August 2013; pp. 343–352.

[11] Yun-tao, Z.; Ling, G.; Yong-cheng, W. An improved TF-IDF approach for text classification. J. Zhejiang Univ. Sci. A 2005, 6A, 49–55.

[12] M. S. U. Miah, J. Sulaiman, T. Bin Sarwar, K. Z. Zamli, dan R. Jose, "Study of Keyword Extraction Techniques for Electric Double-Layer Capacitor Domain Using Text Similarity Indexes: An Experimental Analysis," *Complexity*, vol. 2021, 2021, doi: 10.1155/2021/8192320.

[13] "Cosine Similarity-understanding the math and how it works? (with python)," https://www.machinelearningplus.com/nlp/cosine-similarity/

[14] 9.5.2. -e Cosine Similarity Algorithm-9.5. Similarity Algorithms, https://neo4j.com/docs/graph-algorithms/current/labs-algorithms/cosine/

[15] H. Hassani, C. Beneki, S. Unger, M. T. Mazinani, dan M. R. Yeganegi, "Text mining in big data analytics," *Big Data and Cognitive Computing*, vol. 4, no. 1, hlm. 1–34, Mar 2020, doi: 10.3390/bdcc4010001.

[16] Dumais, S. Using SVMs for text categorization, Microsoft research. IEEE Intell. Syst. Mag. 1998, 13, 18–28.