AICOMS
Applied Information Technology and Computer Science
e-ISSN: 2964-7703
Vol. 4 No. 1, Juni-2025, pp. 25-32

Pemrosesan Query dan Pemeringkatan Judul Berita Terkait Gubernur Jawa Barat Menggunakan TF-IDF dan Cosine Similarity

Chandra Saputra [1,*], Wilcent [2], Hafiz Irsyad [3], and Abdul Rahman [4]

[1] Universitas Multi Data Palembang; chandrasaputra_2226250093@mhs.mdp.ac.id
[2] Universitas Multi Data Palembang; wilcent_2226250120@mhs.mdp.ac.id
[3] Universitas Multi Data Palembang; hafizirsyad@mdp.ac.id
[4] Universitas Multi Data Palembang; arahman@mdp.ac.id
* Universitas Multi Data Palembang; chandrasaputra_2226250093@mhs.mdp.ac.id

Info Artikel:
Dikirim: 16 Mei 2025
Direvisi: 17 Mei 2025
Diterima: 19 Mei 2025

Abstract: Increasing efficiency and relevance in searching for news information is a pressing need in the digital era. This study aims to develop a news title ranking system based on keywords (queries) by combining the Term Frequency-Inverse Document Frequency (TF-IDF) and cosine similarity methods. The data used are 2,507 news titles from four of the most popular news sites in Indonesia, namely Kompas.com, Detik.com, CNNIndonesia.com, and Tempo.com in the last one year. The stages carried out include web scraping, pre-processing (case folding, tokenizing, stopwords removal, and stemming), word weighting using TF-IDF, similarity calculation using cosine similarity, to system performance evaluation with accuracy, precision, recall, and f1-score metrics. The test results on three different queries show that the system is able to provide very good results with an average accuracy of 99.75%, precision 96.67%, recall 100%, and f1-score 98.33%. This study shows that the combination of TF-IDF and cosine similarity is effective in optimizing the search for news titles that are relevant to the entered query.

Keywords: Cosine Similarity; Document Ranking; News Title Search; TF-IDF; Web Scraping.

Intisari: Peningkatan efisiensi dan relevansi dalam pencarian informasi berita menjadi kebutuhan penying di era digital. Penelitian ini bertujuan untuk mengembangkan sistem pemeringkatan judul berita berdasarkan kata kunci (*query*) dengan menggabungkan metode *Term Frequency-Inverse Document Frequency* (TF-IDF) dan c*osine similarity*. Data yang digunakan berupa 2.507 judul berita dari empat situs berita terpopuler di Indonesia, yaitu Kompas.com, Detik.com, CNNIndonesia.com, dan Tempo.com dalam rentang waktu satu tahun terakhir. Tahapan yang dilakukan meliputi *web scraping*, *pre-processing* (*case folding, tokenizing, stopwords removal, dan stemming*), pembobotan kata menggunakan TF-IDF, perhitungan kemiripan menggunakan *cosine similarity*, hingga evaluasi kinerja sistem dengan metrik akurasi, presisi, *recall*, dan *f1-score*. Hasil pengujian terhadap tiga *query* yang berbeda menunjukkan bahwa sistem mampu memberikan hasil yang sangat baik dengan rata-rata akurasi sebesar 99.75%, presisi 96.67%, *recall* 100%, dan *f1-score* 98.33%. Penelitian ini menunjukkan bahwa kombinasi TF-IDF dan *cosine similarity* efektif dalam mengoptimalkan pencarian judul berita yang relevan terhadap *query* yang dimasukkan.

Kata Kunci: *Cosine Similarity*; Pemeringkatan Dokumen; Pencarian Judul Berita; TF-IDF; *Web Scraping*.

1. Pendahuluan

Berita adalah laporan atau informasi yang tersebar dengan cepat mengenai fakta-fakta yang benar dan sah. Hampir semua lapisan masyarakat mengharapkan informasi yang benar-benar sah dan akurat [1]. Berita berfungsi sebagai sarana penyampaian informasi bagi banyak orang, dengan menyajikan beragam informasi. Dalam sebuah berita terdapat judul yang merupakan sebuah gambaran tentang topik yang digunakan untuk memberitahu isi berita yang akan

diuraikan [2]. Pada saat ini, berita dapat diakses dengan mudah melalui internet oleh para pembaca dari penyedia berita digital yang biasanya dikenal dengan situs web berita [3].

Belakangan ini banyak muncul berita salah satu pejabat publik yang aktif menggunakan media sosial Instagram. Beliau telah menjabat sebagai bupati Purwakarta Provinsi Jawa Barat sebanyak dua kali [4]. Beliau bernama Dedi Mulyadi yang kerap disapa dengan sebutan Kang Dedi [5]. Sekar ang mantan pejabat bupati Purwakarta Provinsi Jawa Barat menjabat sebagai Gubernur Provinsi Jawa Barat untuk periode 2025-2030.

Penelitian oleh Rizqa, Ema, dan Suwanto tahun 2020 melakukan analisis komentar perdagangan sosial instagram menggunakan TF-IDF [6]. Algoritma *cosine similarity* dan TF-IDF juga digunakan pada penelitian yang dilakukan Rangga Saputra, Jayanta, dan Musthofa Galih Pradana tahun 2024 yang diyakini dapat membantu Biro SDM dalam mengotomatiskan proses penentuan rumpun jabatan, meningkatkan efisiensi dan mengurangi intervensi manual [7]. Dalam penelitian yang dilakukan oleh Venny Meida Hersianty, Eka Larasati Amalia, Dwi Puspitasari, Dimas Wahyu Wibowo untuk sistem lowongan pekerjaaan menunjukkan bahwa algoritma TF-IDF dan *cosine similarity* mampu menghasilkan tingkat akurasi sebesar 80% [8].

Dalam penelitian ini menggunakan empat situs berita terpopuler di Indonesia yaitu, Kompas.com, Detik.com, CNNIndonesia.com, dan Tempo.com. Empat situs tersebut merupakan sumber yang populer, terpercaya, dan aktual. Judul berita dipilih dari setahun terakhir agar dapat melihat isu dan tren yang terjadi saat ini. Pemilihan judul satu tahun terakhir ini bertujuan untuk merangkum dan mencerminkan peristiwa terbaru terhadap isu-isu yang berkembang [9]. Pengumpulan data dari situs berita didapatkan sebanyak 2507 judul berita pada rentang waktu 8 Mei 2024 sampai dengan 8 Mei 2025.

Teknik optimasi pencarian berperan penting dalam meningkatkan pengalaman pencarian [10]. Teknik-teknik ini bertujuan untuk mengoptimalkan proses pencarian supaya lebih efektif yang berbasis kata kunci (*query)*. Dalam pemrosesan *query* dan pemeringkatan dibutuhkan algoritma dan metode yang lebih canggih dibanding dengan cara tradisional untuk meningkatkan akurasi dan presisi dalam pencarian.

Ada banyak algoritma dan metode yang dapat digunakan, namun penelitian ini akan berfokus pada penggunaan *Term Frequency-Inverse Document Frequency* (TF-IDF) dan *cosine similarity*. TF-IDF adalah algoritma yang dipakai untuk memberikan bobot kata dengan memperhatikan jumlah frekuensi atau banyaknya kata yang muncul dalam judul atau dokumen [11]. Metode *cosine similarity* digunakan untuk menghitung tingkat kemiripan antara dua kalimat atau dokumen tetapi tidak menyertakan frekuensi term (kata) [12]. Dengan memanfaatkan pembobotan dari TF-IDF yang dapat memberikan nilai bobot pada setiap kata dalam judul atau dokumen akan mendukung *cosine similarity* dalam memproses kata dengan maksimal.

Penelitian ini bertujuan untuk meningkatkan ketepatan atau relevansi dalam pencarian judul berita dengan memasukkan kata kunci atau *query* tertentu. Penelitian ini juga diharapkan dapat membantu pecinta berita dalam menemukan judul berita yang diinginkan. Penelitian ini dimaksudkan untuk memberikan kontribusi dalam sektor pengambilan informasi terkait judul berita dengan metode yang kreatif untuk meningkatkan akurasi dan presisi dalam pencarian judul berita.

2. Tinjauan Pustaka

Dalam pemrosesan dan pemeringkatan *query* memuat teori-teori yang mendukung dari metode yang diusulkan untuk menyelesaikan masalah dan pengembangan dari metode tersebut, yang didasari dari referensi yang jelas, seperti jurnal.

2.1. *Web Scraping*

Teknik ini digunakan untuk mengambil informasi-informasi tertentu yang ada dalam situs *web* secara langsung dan otomatis. Kegiatan pengambilan data atau informasi dari situs *web* ini hanya mengambil data spesifik dari situs yang diinginkan sesuai dengan kebutuhan [13]. Hasil dari *web scraping* ini nantinya akan dapat digunakan lagi oleh sistem lain atau dianalisis lebih lanjut [14].

2.2. *TF-IDF*

Pembobotan TF-IDF adalah suatu metode yang digunakan untuk memberikan bobot setiap kata [11]. Bobot pada TF-IDF dapat dipahami sebagai ukuran relevansi yang merupakan hasil dasar dari model keputusan probabilistik. Dengan kata lain, tingkat relevansi akan meningkat sebanding dengan seberapa sering kata tersebut muncul dalam dokumen dan disesuaikan dengan frekuensi kata dalam korpus [15].

2.3. *Cosine Similarity*

Untuk mengukur kemiripan antar dokumen tanpa mempertimbangkan frekuensi kata secara langsung, pendekatan Cosine Similarity dapat digunakan. Teknik ini menjadi lebih efektif apabila

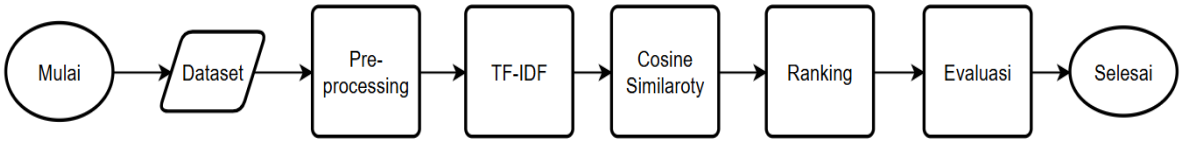
dikombinasikan dengan pembobotan TF-IDF (Term Frequency – Inverse Document Frequency), karena TF-IDF memberikan nilai penting pada setiap istilah dalam teks, sehingga proses perbandingan antar kata dapat dilakukan dengan lebih optimal [12].

3. Metode dan Hasil Penelitian

Pada penelitian ini menggunakan metode dan proses yang tergolong tradisional. Metode-metode yang digunakan akan saling barkaitan supaya menjadi satu penelitian yang baik.

3.1. Metode Penelitian

Dalam penelitian ini, rancangan proses sistematis yang terdiri dari beberapa tahapan utama yang dimulai dari tahapan pengambilan dan pengumpulan dataset berupa judul berita yang menjadi objek penelitian penulis. Selanjutnya tahapan *pre-processing* untuk menghasilkan dataset yang lebih terstruktur. Lalu tahapan untuk menghitung bobot dari tiap kata menggunakan TF-IDF. Untuk melakukan pemeringkatan digunakan juga pertihungan dengan metode *cosine similarity*. Dan terakhir dilakukan tahapan evaluasi untuk melihat tingkat akurasi, presisi, *recall* dan *f1-score*. Proses ini dapat dilihat pada Gambar 1.



Gambar 1. *Flowchart* tahapan sistem

Dataset yang digunakan dalam penelitian ini adalah judul berita yang diambil dari empat situs berita online di Indonesia dalam satu tahun terakhir. Selama rentang waktu satu tahun total terdapat sebanyak 2.507 judul berita. Empat situs tersebut adalah Kompas.com, Detik.com, CNNIndonesia.com, dan Tempo.com. Judul diambil menggunakan teknik *web scraping* dimana proses ini menggunakan program berbasis bahasa *Python*. Menurut teori, *web scraping* adalah teknik untuk mengambil data dengan cara yang berbeda dengan penggunaan API (*Application Programming Interface*) [16].

3.2. *Pre-processing*

Pada tahapan ini, akan melibatkan serangkaian proses dasar, antara lain *case folding*, tokenizing, stopword, dan stemming [17] [18]. Tahapan ini sangat penting untuk mempersiapkan data supaya menjadi terstruktur.

3.2.1. *Case Folding*

Case Folding merupakan langkah untuk melakukan pengubahan semua teks yang ada menjadi huruf kecil serta menghapus katakter yang berada di luar rentang abjad [19].

3.2.2. Tokenizing

Tokenizing merupakan langkah untuk memisahkan antar kata menjadi kata-kata tunggal. Pemisah yang digunakan adalah delimeter seperti spasi (" ") yang bertujuan untuk mempermudah proses *stemming* [20].

3.2.3. *Stopwords*

Stopwords merupakan langkah untuk menghapus kata-kata yang tidak memiliki makna atau kurang berarti serta sering muncul dalam kumpulan kata. Kata-kata tersebut sering dihilangkan dari proses analisis agar tidak mengganggu keakuratan hasil. *Stopwords* meliputi kata penghubung, kata ganti, preposisi, dan lain sebagainya, seperti: yang, dan, di, ke, dari, atau, dia, dan semacamnya [20].

3.2.4. *Stemming*

Stemming merupakan langkah untuk mengubah kata-kata setelah *stopwords* ke bentuk kata dasar sesuai dengan aturan yang ada, sehingga kata-kata tersebut mempunyai representasi yang sama [20].

3.3. Metode TF-IDF

TF-IDF adalah metode yang tergolong tradisional dalam mengektrak kata kunci yang digunakan untuk mengukur nilai suatu kata dalam teks [9]. Perhitungan TF-IDF pada dasarnya terbagi jadi dua bagian yaitu, TF (*Term Frequency*) dan IDF (*Inverse Document Frequency*) yang masing-masing memiliki rumus dan cara kerja yang berbeda, lalu digabungkan pada akhir perhitungan [21].

3.3.1 TF (*Term Frequency*)

Menghitung frekuensi total kemunculan kata pada dokumen [21]. Masing-masing dokumen memiliki panjang yang beragam maka nilai TF akan dibagi dengan panjang dokumen. Persamaan 1 adalah rumus menghitung TF.

$$tf(t,d) = \frac{n_{t,d}}{\sum_t i} \quad (1)$$

Ket:
$tf(t,d)$ = frekuensi kemunculan kata pada dokumen
$n_{t,d}$ = frekuensi kemunculan *term* di dokumen
$\sum_t i$ = total *term* dalam dokumen

3.3.2 IDF (*Inverse Document Frequency*)

Mengukur seberapa penting suatu kata, apabila nilai IDF semakin rendah maka semakin tidak penting kata tersebut, dan sebaliknya [21]. Persamaan 2 adalah rumus menghitung IDF.

$$idf_d = \log \frac{N}{df(t)} \quad (2)$$

Ket:
idf_d = mengukur penting/tidak kata dalam dokumen
N = jumlah dokumen yang ada kata kunci
$df(t)$ = jumlah dokumen yang ada *term*

3.3.3 TF-IDF

Menghitung TF-IDF dengan mengalikan TF dengan IDF. Persamaan 3 adalah rumus menghitung TF-IDF.

$$tf(t,d)idf_d = tf(t,d) \times idf_d \quad (3)$$

Ket:
$tf(t,d)idf_d$ = hasil perkalian TF dan IDF

3.4. *Cosine Similarity*

Cosine similarity merupakan salah satu metode yang sangat populer dalam analisis teks atau *text mining* yang digunakan untuk mengukur sejauh mana dua teks memiliki kemiripan [22]. Inti dari *cosine similarity* ialah menghitung nilai cosinus dari sudut antara dua vektor, yang mencerminkan arah atau orientasi masing-masing vektor tanpa mempertimbangkan panjangnya. Dengan demikian, meskipun dua dokumen mempunyai panjang berbeda, kedekatan kontennya tetap dapat diukur secara proporsional. Metode ini sangat efektif dalam berbagai aplikasi seperti pencarian informasi, klasifikasi dokumen dan sistem rekomendasi. Nilai *cosine similarity* yang mendekati 1 maka tergolong memiliki distribusi kata yang mirip. Persamaan 4 adalah rumus menghitung *cosine similarity*.

$$C_S = \frac{A.B}{\|A\|\|B\|} \quad (4)$$

Ket:
A = *Query* TF-IDF *Weight Vector*
B = *Document* TF-IDF *Weight Vector*

3.5. Evaluasi

Pada tahapan evaluasi diperlukan *confusion matrix* yang merupakan sebuah tabel yang digunakan untuk mengilustrasikan kinerja model atau algoritma dengan lebih detail. Pada proses ini akan membandingkan antara kelas yang diprediksi oleh model dengan kelas yang sebenarnya [23]. Tabel *confusion matrix* dapat dilihat di Tabel 1.

Tabel 1. *Confusion matrix*

	Prediksi Positif	**Prediksi Negatif**
Aktual Positif	*True Positive* (TP)	*Fasle Negative* (FN)
Aktual Negatif	*False Positive* (FP)	*True Negative* (TN)

1. *True Positive* = Data aktual yang berkelas positif dan model memprediksi positif;
2. *False Positive* = Data aktual yang berkelas negatif dan model memprediksi positif;
3. *False Negatif* = Data aktual yang berkelas positif dan model memprediksi negatif;
4. *True Negartif* = Data aktual yang berkelas negatif dan model memprediksi negatif;

Accuracy

Dengan *confusion matrix,* akurasi dapat dihitung. Semakin besar nilai akurasi yang dicapai, maka semakin efektif metode yang digunakan [24]. Persamaan 5 adalah rumus *accuracy.*

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN} \times 100\% \quad (5)$$

Precision

Precision adalah ukuran yang membandingkan jumlah prediksi positif yang benar (TP) dengan jumlah seluruh positif [24]. Persamaan 6 adalah rumus *precision*.

$$Precision = \frac{TP}{TP+FP} \times 100\% \quad (6)$$

Recall

Recall adalah ukuran yang membandingkan jumlah prediksi positif yang benar (TP) dengan jumlah data benar positif [24]. Persamaan 7 adalah rumus *recall*.

$$Recall = \frac{TP}{TP+FN} \times 100\% \quad (7)$$

F1-Score

F1-Score adalah hasil nilai rata-rata dari perbandingan *precision* dan *recall* [24]. Persamaan 8 adalah rumus *f1-Score*.

$$F1 - Score = 2 \times \frac{precision \times recall}{(precision+recall)} \times 100\% \quad (8)$$

Tahapan sistem yang digunakan dalam penelitian ini terdiri dari judul-judul berita yang ada pada situs *web* dalam rentang waktu satu tahun terakhir. Pada penelitian ini digunakan satu *query* sebagai contoh pada tahap *pre-processing*. *Query*: "Larangan wisuda sekolah". Tabel 2 adalah hasil dari tahap *pre-processing* dokumen dalam hal ini adalah judul.

Tabel 2. Hasil *pre-processing* dokumen

Dokumen	***Pre-processing***
Beda Dedi Mulyadi dan Mendikdasmen Soal Larangan Wisuda Sekolah	"beda" "dedi" "mulyadi" "mendikdasmen" "larangan" "wisuda" "sekolah"

Setelah melakukan tahapan *pre-processing* yang terdiri dari *case folding,* tokenizing, stopword, dan stemming, akan dilakukan tahapan menghitung bobot kata menggunakan metode TF-IDF. Metode ini bertujuan untuk mengukur tingkat kepentingan sebuah kata dalam suatu dokumen dengan mempertimbangkan kemunculannya di keseluruhan kumpulan dokumen. Kata yang muncul secara dominan di satu dokumen namun jarang ditemukan di dokumen lain akan

diberikan bobot yang lebih besar, karena dianggap lebih mewakili isi dari dokumen tersebut. Tabel 3 adalah hasil TF-IDF vektor dokumen.

Tabel 3. Hasil TF-IDF vektor dokumen

Dokumen	**Pembobotan Dokumen (TF-IDF)**			
	Query	**TF**	**IDF**	**TF-IDF**
Beda Dedi Mulyadi dan Mendikdasmen Soal Larangan Wisuda Sekolah	beda	0	6.5243	0
	dedi	0	1.2224	0
	mulyadi	0	1.2513	0
	mendikdasmen	0	5.7823	0
	larangan	0.5838	5.0656	2.9571
	wisuda	0.6417	5.5687	3.5737
	sekolah	0.4974	4.3160	2.1467

Untuk memenuhi kriteria dalam pengukuran kesamaan antara *query* dan dokumen, metode yang sama digunakan untuk menghitung vektor TF-IDF pada *query*. Jika pada tahap sebelumnya perhitungan vektor TF-IDF dilakukan untuk membandingkan antara *query* dengan dokumen contoh, maka pada tahap ini pembobotan dilakukan secara khusus terhadap *query* itu sendiri. Langkah ini bertujuan untuk mengubah *query* menjadi bentuk representasi numerik berdasarkan frekuensi kemunculan dan distribusi kata-kata yang dikandungnya. Dengan representasi tersebut, *query* dapat dianalisis dalam ruang vektor yang sama seperti dokumen lainnya, sehingga memungkinkan proses perbandingan atau pencocokan dilakukan secara lebih akurat dan objektif. Hasil dari perhitungan ini disajikan pada Tabel 4 adalah hasil vektor TF-IDF vektor *query*.

Tabel 4. Hasil TF-IDF vektor *query*

Query	***Query keyword***	**TF-IDF**
Larangan Wisuda Sekolah	larangan	0.5838
	wisuda	0.4974
	sekolah	0.6417

Tahapan berikutnya adalah menghitung *cosine similarity* untuk mendapatkan skor pada setiap dokumen yang nantinya skor akan digunakan untuk pemeringkatan, melihat kemiripan dokumen dengan *query* yang di cari. Dari *query* "Larangan Wisuda Sekolah" akan mendapatkan hasil *cosine similarity* yang terbaik. Tabel 5 adalah hasil *cosine similarity*.

Tabel 5. Hasil *cosine similarity*

Query	**Dokumen**	***Cosine similarity***
Larangan Wisuda Sekolah	Beda Dedi Mulyadi dan Mendikdasmen Soal Larangan Wisuda Sekolah	0.6984

Tahapan pemeringkatan merupakan proses lanjutan yang dilakukan setelah diperoleh nilai *cosine similarity* antara masing-masing dokumen dengan *query*. Pada tahap ini, dokumen-dokumen akan diurutkan berdasarkan nilai *cosine similarity*, di mana semakin besar nilai *cosine similarity*, maka semakin tinggi tingkat relevansi dokumen tersebut terhadap *query*, dan semakin tinggi pula peringkatnya (dengan angka ranking yang lebih kecil). Hal ini menunjukkan bahwa nilai *cosine similarity* yang tinggi mencerminkan kesamaan yang kuat antara isi dokumen dan *query* yang dimasukkan, yaitu 'Larangan Wisuda Sekolah'. Oleh karena itu, pemeringkatan ini menjadi langkah penting dalam menentukan dokumen mana yang paling relevan untuk ditampilkan. Tabel 6 adalah hasil pemeringkatan dokumen.

Tabel 6. Hasil pemeringkatan dokumen

Top-N	Dokumen	*cosine similarity*
Dokumen-329	Beda Dedi Mulyadi dan Mendikdas-men Soal Larangan Wisuda Sekolah	0.6984
Dokumen-399	Sekali Lagi, Dedi Mulyadi Tegaskan Larangan "Study Tour" dan Wisuda Berbayar di Sekolah	0.5975
Dokumen-340	Tanggapi Kebijakan Larangan Wisuda di Sekolah, Mendikdasmen: Ya Masa Sih Tidak Boleh	0.5194

Dapat dilihat bahwa dokumen-329 merupakan dokumen yang paling relevan dari 2507 dokumen lainnya, dengan tingkat kemiripan 0.6984 (70%). Sehingga dapat disimpulkan bahwa dokumen-329 dengan tingkat kemiripan sekitar 70% memiliki kesamaan dengan *query* dilihat dari pemeringkatan yang menempati peringkat 1. Tahapan terakhir ialah evaluasi, dimana melakukan pengukuran relevansi dan kualitas dari penelitan yang dilakukan dengan menggunakan metode pengujian akurasi, presisi, *recall*, dan *f1-score.* Pengujian akan dilakukan menggunakan 3 *query* berbeda untuk dilihat rata-rata dari pengujian yang ada. Tabel 7 adalah hasil pengujian evaluasi.

Tabel 7. Hasil pengujian evaluasi

Query	Akurasi (%)	Persisi	*Recall*	*F1-score*
Larangan Wisuda Sekolah	99.80	0.96	1.00	0.98
Tanah Negara	100.00	1.00	1.00	1.00
Anak Nakal Kirim Barak Militer	99.44	0.94	1.00	0.97
Total	299.24	2.9	3	2.95
Rata-rata (%)	99.75	96.67	100	98.33

Dari pengujian terhadap sistem dengan memasukkan *query* yang berbeda-beda menghasilkan hasil yang menarik. Hasil dari setiap pengujian ditemukan bervariasi antar *query* dengan skor rata-rata akurasi (99.75%), presisi (96.67%), *recall* (100%), dan *f1-score* (98.33%).

5. Kesimpulan

Pada penelitian ini peneliti menggunakan TF-IDF yang digabungkan dengan *cosine similarity*, dimana peneliti berhasil mengoptimalkan pencarian judul di dataset menggunakan *query*. Papa sistem pemeringkatan juga telah dievaluasi dengan menggunakan keempat metrik evaluasi. Berdasarkan pengujian yang dilakukan dengan 3 *query* yang berbeda, sistem pemeringkatan berhasil memberikan evaluasi yang sangat baik dimana mendapatkan nilai rata-rata untuk akurasi sebesar 99.75%, persisi sebesar 96.67%, *recall* sebesar 100%, dan *f1-score* sebesar 98.33%.

Daftar Pustaka

[1] M. A. H. Erwan Effendy, Forsaktinahot Hasugian, "Menulis Isi Berita Dan Feature," *J. Pendidik. dan Konseling*, vol. 4, pp. 1349–1358, 2022.

[2] Oktamia Anggraini Putri, "Jurnal Pendidikan dan Konseling," *J. Pendidik. dan Konseling*, vol. 4, no. 20, pp. 1349–1358, 2022.

[3] N. Tahir *et al.*, "FNG-IE: an improved graph-based method for keyword extraction from scholarly big-data," *PeerJ Comput. Sci.*, vol. 7, pp. 1–24, 2021, doi: 10.7717/peerj-cs.389.

[4] M. Fikri, "BRANDING POLITIK CALON GUBERNUR JAWA BARAT PADA AKUN INSTAGRAM @ dedimulyadi71 BRANDING POLITICS OF WEST JAVA GUBERNOR CANDIDATE ON @ dedimulyadi71 INSTAGRAM ACCOUNT," vol. 3, no. 1, pp. 1–11, 2024.

[5] R. Adolph, "済無No Title No Title No Title," vol. 3, no. 2, pp. 1–23, 2016.

[6] R. L. Musyarofah, E. U. Utami, and S. R. Raharjo, "Analisis Komentar Potensial pada Social Commerce Instagram

Menggunakan TF-IDF," *J. Eksplora Inform.*, vol. 9, no. 2, pp. 130–139, 2020, doi: 10.30864/eksplora.v9i2.360.

[7] R. Saputra and M. Galih Pradana, "Implementasi Algoritma Cosine Similarity dan TF-IDF dalam Menentukan Rumpun Jabatan," vol. 12, no. 1, pp. 1–11, 2024, doi: 10.32832/kreatif.v12i1.15470.

[8] V. M. Hersianty *et al.*, "PENERAPAN ALGORITMA TF-IDF DAN COSINE SIMILARITY," vol. 9, no. 1, pp. 1619–1625, 2025.

[9] I. S. Wibowo, A. Witanti, and I. Susilawati, "Keyword Extraction Judul Berita Online Di Indonesia Menggunakan Metode TF-IDF," *J. Tek. Inform. dan Sist. Inf.*, vol. 11, no. 1, pp. 99–111, 2024, [Online]. Available: http://jurnal.mdp.ac.id

[10] S. Patulus *et al.*, "IMPLEMENTASI TEKNIK QUERY OPTIMIZATION UNTUK MENINGKATKAN," vol. 9, no. 2, pp. 2437–2442, 2025.

[11] I. Widaningrum, D. Mustikasari, R. Arifin, S. L. Tsaqila, and D. Fatmawati, "Algoritma Term Frequency-Inverse Document Frequency (TF-IDF) dan K-Means Clustering Untuk Menentukan Kategori Dokumen," *Pros. Semin. Nas. Sist. Inf. dan Teknol.*, pp. 145–149, 2022.

[12] H. Zakiyudin and K. Marzuki, "Penerapan Algoritma Cosine Similarity dan Pembobotan TF-IDF System Penerimaan Mahasiswa Baru pada Kampus Swasta Application of the Cosine Similarity Algorithm and Weighting of the TF-IDF System for New Student Admissions on Private Campuses," vol. 3, no. 1, pp. 19–27, doi: 10.30812/bite.v3i1.1110.

[13] A. Z. Rizquina and C. I. Ratnasari, "Implementasi Web Scraping untuk Pengambilan Data Pada Website E-Commerce," *J. Teknol. Dan Sist. Inf. Bisnis*, vol. 5, no. 4, pp. 377–383, 2023, doi: 10.47233/jteksis.v5i4.913.

[14] A. Rahmatulloh and R. Gunawan, "Web Scraping with HTML DOM Method for Data Collection of Scientific Articles from Google Scholar," *Indones. J. Inf. Syst.*, vol. 2, no. 2, pp. 95–104, 2020, doi: 10.24002/ijis.v2i2.3029.

[15] E. Prayitno, T. Suprawoto, and ..., "Optimasi Hasil Pencarian Pada Web Scrapping Menggunakan Pembobotan Kata Tf-Idf," *J. Innov. Res. Knowl.*, vol. 1, no. 7, pp. 241–246, 2021, [Online]. Available: https://bajangjournal.com/index.php/JIRK/article/view/822

[16] Y. Sahria, "Implementasi Teknik Web Scraping pada Jurnal SINTA Untuk Analisis Topik Penelitian Kesehatan Indonesia," *URECOL (Unversity Res. Colloqium)*, pp. 297–306, 2020, [Online]. Available: http://repository.urecol.org/index.php/proceeding/article/view/1079

[17] S. Data *et al.*, "TF-IDF DAN K-MEANS DENGAN MEMANFAATKAN," vol. 3, no. 1, 2022.

[18] R. Wati, S. Ernawati, and H. Rachmi, "Pembobotan TF-IDF Menggunakan Naïve Bayes pada Sentimen Masyarakat Mengenai Isu Kenaikan BIPIH," *J. Manaj. Inform.*, vol. 13, no. 1, pp. 84–93, 2023, doi: 10.34010/jamika.v13i1.9424.

[19] I. P. Wibina, K. Gumi, and A. Syafrianto, "Perbandingan Algoritma Naïve Bayes dan Decision Tree Pada Sentimen Analisis," vol. 1, pp. 1–15, 2022.

[20] D. F. AL-Hafiidh, I. F. Rozi, and I. K. Putri, "Peringkasan Teks Otomatis pada Portal Berita Olahraga menggunakan metode Maximum Marginal Relevance.," *J. Inform. Polinema*, vol. 8, no. 3, pp. 21–30, 2022, doi: 10.33795/jip.v8i3.519.

[21] K. Tri Putra, M. Amin Hariyadi, and C. Crysdian, "Perbandingan Feature Extraction Tf-Idf Dan Bow Untuk Analisis Sentimen Berbasis Svm," *J. Cahaya MAndalika*, p. 1449, 2023.

[22] P. C. Siswipraptini, "Klasifikasi Pekerjaan Bidang Teknologi Informasi Menggunakan Algoritma Cosine Similarity," *Kilat*, vol. 12, no. 1, pp. 38–48, 2023, doi: 10.33322/kilat.v12i1.2001.

[23] Y. S. Cendikia *et al.*, "MENENTUKAN WARNA MAKE UP YANG COCOK BERDASARKAN JENIS SKINTONE PADA CITRA WAJAH MENGGUNAKAN NAIVE BAYES CLASSIFIER," vol. 9, no. 1, pp. 816–823, 2025.

[24] M. Maulidah, Windu Gata, Rizki Aulianita, and Cucu Ika Agustyaningrum, "Algoritma Klasifikasi Decision Tree Untuk Rekomendasi Buku Berdasarkan Kategori Buku," *E-Bisnis J. Ilm. Ekon. dan Bisnis*, vol. 13, no. 2, pp. 89–96, 2020, doi: 10.51903/e-bisnis.v13i2.251.