Jurnal Informatika dan Teknologi Interaktif
Volume 3, Nomor 1, April 2026

DOI : 10.63547/jiite.v3i1.109
e-ISSN : 3048-1023

Penerapan Metode Vector Space Model TF-IDF dan Cosine Similarity pada Sistem Temu Balik Informasi Berita

Adinda Pangestu[1], Rias Estriana[2], Rahma wati[3] , Aldrian Firmansyah[4], Muhammad Fahat[5] ,Aulia Safira[6], Toik Zakiyudin[7]

[1,2,3,4,5]Program Studi Informatika, Fakultas Ilmu Komputer, Universitas Amikom Purwokerto, Purwokerto, Indonesia
[6]Program Studi Sistem Informasi, Fakultas Matematika dan Komputer, Universitas Nahdlatul Ulama Al Ghazali Cilacap, Cilacap, Indonesia
surel:[1]adindapangestu17@gmail.com,[2]estrianarias@gmail.com,[3]snozerhma@gmail.com,[4]azamamirul123@gmail.com,[5]ryury345@gmail.com
[6]auliasafirap79@gmail.com, [7]toikzaki5753@gmail.com

Info Artikel

Sejarah artikel:
Diterima 25-01-2026
Revisi 05-02-2026
Diterima 10-02-2026

Kata kunci:
Sistem Temu Balik Informasi
Vector Space Model
TF-IDF
Cosine Similarity
Berita Daring

ABSTRAK

Perkembangan pesat media berita daring menyebabkan peningkatan volume dokumen teks yang signifikan, sehingga menimbulkan permasalahan information overload dalam proses pencarian informasi. Pengguna sering mengalami kesulitan menemukan berita yang relevan karena banyaknya dokumen yang memiliki kemiripan kata, namun tidak selalu sesuai dengan konteks kebutuhan informasi. Oleh karena itu, diperlukan suatu sistem temu balik informasi yang mampu melakukan pencarian dan pemeringkatan dokumen berita secara akurat berdasarkan tingkat relevansi konten. Penelitian ini bertujuan untuk menerapkan metode Vector Space Model (VSM) dengan pembobotan Term Frequency–Inverse Document Frequency (TF-IDF) serta pengukuran kemiripan menggunakan Cosine Similarity pada sistem temu balik informasi berita berbahasa Indonesia. Pendekatan penelitian yang digunakan adalah pendekatan kuantitatif dengan metode content-based information retrieval. Data penelitian berupa kumpulan dokumen berita daring yang diproses melalui tahapan preprocessing teks, meliputi case folding, tokenisasi, stopword removal, dan stemming, untuk menghasilkan data teks yang bersih dan seragam. Setiap dokumen kemudian direpresentasikan dalam bentuk vektor numerik menggunakan VSM dan diberi bobot TF-IDF untuk menonjolkan istilah yang bersifat spesifik terhadap topik dokumen. Tingkat kemiripan antara kueri pengguna dan dokumen berita dihitung menggunakan Cosine Similarity, yang selanjutnya digunakan sebagai dasar pemeringkatan dokumen. Hasil penelitian menunjukkan bahwa integrasi VSM, TF-IDF, dan Cosine Similarity mampu meningkatkan relevansi hasil pencarian dan menyajikan dokumen berita secara terstruktur sesuai dengan kebutuhan pengguna. Dengan demikian, sistem yang dikembangkan dapat menjadi solusi efektif dalam pencarian informasi berita berbasis teks serta berpotensi diterapkan pada koleksi dokumen berskala besar.

Penulis yang sesuai:
Adinda Pangestu
Program Studi Informatika Fakultas Ilmu Komputer Universitas Amikom Purwokerto

Email: adindapangestu17@gmail.com

1. PENDAHULUAN

Kemajuan teknologi digital telah mendorong peningkatan yang signifikan dalam produksi dan distribusi informasi berbasis teks, khususnya pada media berita daring[1]. Setiap portal berita secara kontinu mempublikasikan artikel dalam jumlah besar dengan topik yang beragam, mulai dari politik, ekonomi, hingga sosial dan budaya. Kondisi ini memberikan kemudahan akses informasi bagi masyarakat, namun di sisi lain menimbulkan permasalahan baru berupa *information overload*, di mana pengguna kesulitan menemukan informasi yang benar-benar relevan dengan kebutuhan pencarian mereka dalam waktu yang singkat. Permasalahan utama yang muncul adalah banyaknya dokumen berita yang memiliki kemiripan kata atau frasa, tetapi tidak selalu memiliki kesesuaian makna dengan informasi yang dicari pengguna[2]. Proses pencarian manual menjadi tidak efisien karena pengguna harus menelusuri sejumlah besar dokumen yang relevansinya rendah. Oleh karena itu, dibutuhkan suatu mekanisme pencarian yang tidak hanya mampu menemukan dokumen berdasarkan kecocokan kata kunci, tetapi juga mampu menyaring dan memeringkat dokumen berita berdasarkan tingkat relevansi kontennya secara lebih akurat[1].

Sistem temu balik informasi (*information retrieval*) dirancang untuk menjawab kebutuhan tersebut dengan cara menghubungkan kueri pengguna dan koleksi dokumen melalui suatu model representasi tertentu [3]. Salah satu model representasi dokumen yang banyak digunakan adalah *Vector Space Model* (VSM), di mana dokumen dan kueri direpresentasikan dalam bentuk vektor numerik berdimensi banyak. Keunggulan VSM terletak pada kemampuannya untuk memodelkan hubungan kemiripan antar dokumen secara matematis, sehingga sistem tidak hanya bergantung pada pencocokan kata secara langsung, tetapi juga mempertimbangkan kedekatan makna antar dokumen dalam proses pemeringkatan[4]. Dalam penerapan VSM, proses pembobotan istilah menjadi faktor krusial yang sangat memengaruhi kualitas hasil pencarian. Metode *Term Frequency–Inverse Document Frequency* (*TF-IDF*) digunakan untuk memberikan bobot pada setiap kata berdasarkan tingkat kemunculannya dalam suatu dokumen dan distribusinya pada keseluruhan koleksi dokumen. Pendekatan ini memungkinkan sistem untuk menekan pengaruh kata-kata umum yang sering muncul pada banyak dokumen, sekaligus menonjolkan kata-kata yang lebih spesifik dan mencirikan topik tertentu. Dengan demikian, representasi dokumen yang dihasilkan menjadi lebih informatif dibandingkan metode pembobotan yang hanya mengandalkan frekuensi kata semata[3].

Penelitian terdahulu yang menjadi rujukan dalam studi ini menunjukkan bahwa pengukuran kemiripan dokumen menggunakan *Cosine Similarity* mampu menghasilkan pemeringkatan dokumen yang lebih akurat dibandingkan pendekatan berbasis pencocokan kata secara langsung. Hal ini disebabkan oleh kemampuan *Cosine Similarity* dalam menghitung sudut antara vektor kueri dan vektor dokumen, sehingga tingkat kesamaan yang dihasilkan merepresentasikan kedekatan makna antar dokumen secara lebih proporsional. Pendekatan tersebut memungkinkan sistem untuk menilai kesamaan arah vektor tanpa dipengaruhi oleh perbedaan panjang dokumen maupun jumlah kata yang terkandung di dalamnya, sehingga relevansi dokumen terhadap kueri dapat ditentukan secara lebih objektif dan konsisten[5]. Meskipun demikian, penelitian sebelumnya lebih banyak difokuskan pada konteks pemrosesan teks untuk keperluan peringkasan otomatis, dengan tujuan meningkatkan kualitas ringkasan dokumen melalui pemilihan kalimat atau segmen teks yang paling representatif. Pendekatan tersebut belum diarahkan secara spesifik pada pengembangan sistem temu balik informasi berita berbasis konten yang menekankan proses pencarian dan pemeringkatan dokumen secara menyeluruh[6]. Oleh karena itu, masih terdapat celah penelitian dalam penerapan *Cosine Similarity* yang terintegrasi dengan model representasi dokumen dan pembobotan istilah untuk mendukung sistem pencarian berita yang mampu menyajikan hasil secara relevan, terstruktur, dan sesuai dengan kebutuhan pengguna[7].

Berdasarkan permasalahan dan kajian literatur, penelitian ini mengusulkan solusi berupa penerapan kombinasi Vector Space Model, pembobotan TF-IDF, dan Cosine Similarity untuk membangun sistem temu balik informasi berita berbahasa Indonesia yang berorientasi pada relevansi konten. Pendekatan ini diarahkan pada proses pencarian dan pemeringkatan dokumen secara otomatis berdasarkan kata kunci pengguna, guna menampilkan dokumen dengan tingkat kesesuaian tertinggi [8]. Integrasi ketiga komponen tersebut diharapkan mampu meningkatkan kualitas representasi dokumen serta akurasi pengukuran kemiripan antar dokumen dalam koleksi berita. Nilai kebaruan penelitian ini terletak pada penyesuaian alur sistem temu balik yang menekankan efektivitas pemrosesan teks dan ketepatan hasil pencarian dalam konteks berita daring berbahasa Indonesia. Penyesuaian

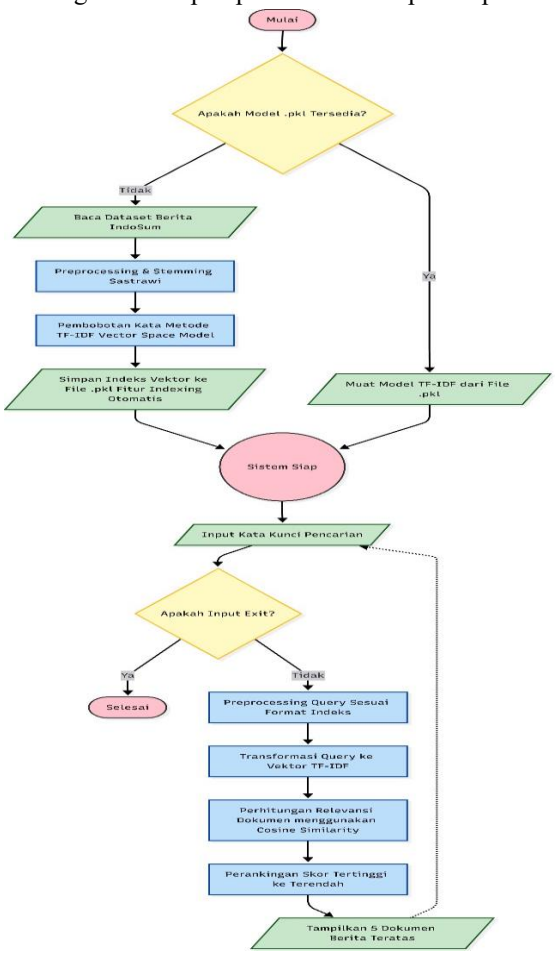
mencakup tahapan pengolahan teks, pembobotan istilah, hingga mekanisme pemeringkatan dokumen yang dirancang sesuai karakteristik data berita[9]. Dengan demikian, tujuan penelitian ini adalah menghasilkan sistem temu balik informasi berita yang lebih relevan, terstruktur, dan mudah digunakan oleh pengguna dari berbagai latar belakang.

2. METODE

Jenis pendekatan yang digunakan dalam penelitian ini adalah pendekatan kuantitatif dengan metode content-based information retrieval[4]. Pendekatan ini dipilih karena mampu merepresentasikan dokumen dan kueri dalam bentuk numerik serta mengukur tingkat kemiripan secara objektif menggunakan perhitungan matematis. Fokus utama penelitian adalah memodelkan hubungan antara kata kunci pencarian pengguna dan kumpulan dokumen berita melalui representasi vektor dan pengukuran kesamaan vector[10].

Kerangka kerja penelitian ini diadaptasi dari jurnal rujukan, yang menerapkan representasi teks berbasis vektor dan pengukuran kemiripan Cosine Similarity dalam konteks pemrosesan teks [5]. Namun, metode pada jurnal rujukan lebih diarahkan pada automatic text summarization, sedangkan penelitian ini melakukan penyesuaian dengan mengubah orientasi sistem menjadi sistem temu balik informasi (information retrieval) yang berfokus pada pencarian dan pemeringkatan dokumen berita secara menyeluruh [11] .

Perbedaan utama penelitian ini dengan jurnal rujukan terletak pada tujuan sistem dan alur pemrosesan. Penelitian ini mengintegrasikan Vector Space Model (VSM) dengan pembobotan TF-IDF dan Cosine Similarity untuk memeringkat dokumen berita berdasarkan relevansi terhadap kueri pengguna. Selain itu, sistem dilengkapi dengan mekanisme penyimpanan indeks vektor (model persistence) untuk meningkatkan efisiensi komputasi pada saat sistem dijalankan kembali[10]. Alur metodologi dan tahapan penelitian ditampilkan pada Gambar 1.



Gambar 1. Flowchart Sistem Temu Balik Informasi Berita

2.1. Desain Penelitian

Desain penelitian yang digunakan adalah desain penelitian eksperimental berbasis sistem, di mana peneliti merancang dan mengimplementasikan suatu sistem temu balik informasi berita berbahasa Indonesia. Sistem ini dibangun dengan mengintegrasikan Vector Space Model (VSM) sebagai model representasi dokumen, TF-IDF sebagai metode pembobotan istilah, serta Cosine Similarity sebagai teknik pengukuran tingkat kemiripan antara dokumen dan kueri[12]. Desain ini bertujuan untuk menguji kemampuan sistem dalam menyajikan dokumen berita yang relevan berdasarkan kata kunci yang dimasukkan pengguna, sebagaimana pendekatan serupa telah digunakan pada penelitian rujukan namun dengan konteks dan tujuan yang berbeda .

2.2. Prosedur Penelitian

Prosedur penelitian dilakukan secara bertahap dan sistematis sebagai berikut.

2.2.1 Pengumpulan Data

Metode pengumpulan data yang digunakan dalam penelitian ini adalah metode arsip, yaitu dengan memanfaatkan kumpulan dokumen berita daring berbahasa Indonesia yang telah tersedia dan tersimpan dalam basis data sistem[13]. Dokumen berita tersebut mencakup berbagai topik dan disusun dalam format teks utuh yang terdiri atas judul dan isi berita[14].

Seluruh dokumen dalam koleksi digunakan sebagai korpus penelitian tanpa proses sampling, karena penelitian ini berfokus pada pengujian kinerja sistem temu balik informasi secara menyeluruh. Penggunaan seluruh data memungkinkan sistem diuji pada kondisi yang mendekati penerapan nyata, di mana pengguna dapat melakukan pencarian pada kumpulan dokumen yang besar dan beragam. Selain itu, pendekatan ini juga bertujuan untuk menghindari bias hasil yang dapat muncul akibat pemilihan sampel tertentu[15].

2.2.2 Preprocessing Teks

Tahapan preprocessing teks dilakukan untuk menghasilkan data yang bersih, seragam, dan siap digunakan dalam proses representasi vektor. Preprocessing merupakan tahap krusial karena kualitas hasil temu balik sangat bergantung pada kualitas data teks yang diolah. Adapun tahapan preprocessing yang diterapkan adalah sebagai berikut:

a. Case folding, yaitu mengubah seluruh karakter dalam dokumen menjadi huruf kecil. Tahap ini bertujuan untuk menghilangkan perbedaan representasi kata akibat variasi huruf besar dan kecil, sehingga kata dengan makna yang sama diperlakukan secara identik[16].
b. Tokenisasi, yaitu proses pemecahan teks menjadi satuan kata (token). Tokenisasi memungkinkan sistem mengidentifikasi kata-kata penyusun dokumen yang selanjutnya digunakan dalam pembentukan kamus istilah.
c. Stopword removal, yaitu penghapusan kata-kata umum yang sering muncul dalam bahasa Indonesia, seperti kata hubung dan kata depan, yang tidak memiliki kontribusi signifikan terhadap makna pencarian. Penghapusan stopword dilakukan untuk mengurangi dimensi vektor dan meningkatkan fokus sistem pada kata-kata yang lebih informatif.
d. Stemming, yaitu proses mengubah kata berimbuhan menjadi kata dasar menggunakan algoritma stemming Bahasa Indonesia. Tahap ini bertujuan untuk menyatukan variasi bentuk kata yang memiliki akar kata yang sama, sehingga mengurangi redundansi istilah dan meningkatkan akurasi representasi dokumen[17].

Secara keseluruhan, tahapan preprocessing bertujuan untuk mengurangi variasi kata yang tidak perlu, menekan noise pada data teks, serta meningkatkan kualitas dan konsistensi representasi dokumen dalam ruang vector.

2.2.3 Pembentukan Representasi Vektor

Setelah melalui proses preprocessing, setiap dokumen direpresentasikan dalam bentuk vektor numerik menggunakan Vector Space Model (VSM). Dalam model ini, setiap dokumen dianggap sebagai sebuah vektor dalam ruang berdimensi banyak, di mana setiap dimensi merepresentasikan satu istilah unik yang terdapat dalam keseluruhan koleksi dokumen[18]. Representasi vektor memungkinkan sistem untuk memodelkan hubungan antar dokumen dan kueri secara matematis. Dengan pendekatan ini, proses pencarian tidak lagi hanya bergantung pada pencocokan kata secara langsung, tetapi juga mempertimbangkan kedekatan posisi vektor dokumen dan kueri dalam ruang vektor[19]

2.2.4 Pembobotan TF-IDF

Pada tahap ini dilakukan pembobotan istilah menggunakan metode Term Frequency–Inverse Document Frequency (TF-IDF). Metode ini mengombinasikan dua komponen utama, yaitu frekuensi kemunculan suatu istilah

dalam dokumen (Term Frequency) dan tingkat keunikan istilah tersebut dalam keseluruhan koleksi dokumen (Inverse Document Frequency)[20]. Pembobotan TF-IDF memberikan bobot yang lebih besar pada kata-kata yang sering muncul dalam suatu dokumen tetapi jarang muncul pada dokumen lain. Sebaliknya, kata-kata umum yang muncul di hampir seluruh dokumen akan memiliki bobot yang lebih kecil. Dengan demikian, TF-IDF mampu menonjolkan kata-kata yang lebih representatif terhadap topik berita dan berperan penting dalam proses penentuan relevansi dokumen terhadap kueri[21].

2.2.5 Penyimpanan Indeks Vektor

Hasil pembobotan TF-IDF beserta kamus istilah yang terbentuk kemudian disimpan dalam bentuk file model (.pkl). Penyimpanan indeks vektor ini bertujuan untuk meningkatkan efisiensi sistem dengan menghindari proses preprocessing dan pembobotan ulang setiap kali sistem dijalankan. Dengan adanya mekanisme penyimpanan model, sistem dapat langsung memuat indeks vektor yang telah terbentuk sebelumnya, sehingga waktu inisialisasi sistem menjadi lebih singkat dan penggunaan sumber daya komputasi dapat dioptimalkan [22]

2.2.6 Pemrosesan Kueri

Kueri yang dimasukkan oleh pengguna diproses menggunakan tahapan preprocessing yang sama seperti dokumen berita. Penerapan preprocessing yang konsisten bertujuan untuk memastikan kesesuaian antara istilah pada kueri dan istilah yang terdapat dalam indeks dokumen. Setelah preprocessing, kueri ditransformasikan ke dalam bentuk vektor berdasarkan kamus TF-IDF yang telah dibentuk sebelumnya. Dengan demikian, kueri dan dokumen berada pada ruang vektor yang sama, sehingga memungkinkan dilakukan pengukuran kemiripan secara langsung.[23]

2.2.7 Perhitungan Cosine Similarity

Tingkat kemiripan antara vektor kueri dan vektor dokumen dihitung menggunakan metode Cosine Similarity. Metode ini mengukur sudut kosinus antara dua vektor, sehingga fokus pada kesamaan arah vektor tanpa dipengaruhi oleh panjang vektor. Nilai Cosine Similarity berada pada rentang 0 hingga 1, di mana nilai mendekati 1 menunjukkan tingkat kemiripan yang tinggi antara dokumen dan kueri, sedangkan nilai mendekati 0 menunjukkan tingkat kemiripan yang rendah. Pendekatan ini memungkinkan sistem untuk menilai relevansi dokumen secara lebih proporsional dan objektif [24]

2.2.8 Pemeringkatan Dokumen

Tahap akhir dalam prosedur penelitian adalah pemeringkatan dokumen berita berdasarkan nilai Cosine Similarity yang diperoleh[25]. Dokumen diurutkan dari nilai kemiripan tertinggi hingga terendah, kemudian sejumlah dokumen teratas ditampilkan sebagai hasil pencarian kepada pengguna. Pemeringkatan ini bertujuan untuk memastikan bahwa dokumen yang paling relevan dengan kebutuhan informasi pengguna muncul pada urutan teratas, sehingga proses pencarian menjadi lebih efisien dan efektif.

2.3. Metode Pengujian

Metode pengujian dilakukan dengan memasukkan beberapa kata kunci sebagai kueri dan mengamati hasil pemeringkatan dokumen yang dihasilkan sistem. Analisis data dilakukan secara deskriptif kuantitatif dengan membandingkan nilai Cosine Similarity antar dokumen untuk menilai tingkat relevansi hasil pencarian. Hasil pengujian ditampilkan dalam bentuk tabel pemeringkatan dokumen. Analisis ini bertujuan untuk mengevaluasi kemampuan sistem dalam menyajikan dokumen berita yang relevan dan konsisten dengan kata kunci pengguna, serta membuktikan efektivitas integrasi VSM, TF-IDF, dan Cosine Similarity dalam sistem temu balik informasi berita [26]

2.4. Spesifikasi Perangkat

Tabel 1. Spesifikasi Perangkat

No.	Komponen	Spesifikasi
1.	Prosesor	Intel Core i5
2.	RAM	16 GB
3.	Sistem Operasi	Windows 11
4.	Bahasa Pemrograman	Python
5.	Platform Pengembangan	Visual Studio Code / Spyder
6.	Library	Pandas, NumPy, Scikit-learn, Sastrawi

3. HASIL DAN PEMBAHASAN

Penelitian ini menghasilkan serangkaian output yang menggambarkan kinerja sistem temu balik informasi berita berbasis Vector Space Model (VSM), pembobotan TF-IDF, dan Cosine Similarity. Setiap tahapan dalam alur sistem dievaluasi berdasarkan keberhasilan pengolahan teks, pembentukan representasi vektor, perhitungan tingkat kemiripan, serta kemampuan sistem dalam memeringkat dokumen berita sesuai dengan kueri pengg [3].

3.1. Hasil Pengumpulan Data

Tahap pengumpulan data menghasilkan sebuah korpus berita daring berbahasa Indonesia yang terdiri dari dokumen teks utuh berupa judul dan isi berita. Seluruh dokumen dalam basis data berhasil dimanfaatkan tanpa proses sampling, sehingga sistem diuji pada kondisi yang mendekati penggunaan nyata. Hasil dari tahap ini menunjukkan bahwa koleksi dokumen memiliki keragaman topik dan kosakata yang tinggi, mencerminkan kompleksitas data berita sesungguhnya. Keberagaman tersebut menjadi tantangan bagi sistem temu balik informasi, namun sekaligus memberikan dasar yang kuat untuk menguji kemampuan metode VSM, TF-IDF, dan Cosine Similarity dalam menilai relevansi dokumen secara objektif [13].

3.2. Hasil Preprocessing Teks

Hasil preprocessing menunjukkan terjadinya penyederhanaan dan penyeragaman teks secara signifikan. Proses case folding berhasil menghilangkan perbedaan penulisan huruf besar dan kecil, sehingga istilah yang sama tidak lagi diperlakukan sebagai token yang berbeda [16]. Tokenisasi menghasilkan daftar kata yang terstruktur dengan baik, sementara proses stopword removal mampu mengurangi dominasi kata-kata umum yang tidak informatif. Dampak paling signifikan terlihat pada tahap stemming, di mana berbagai bentuk kata berimbuhan berhasil direduksi menjadi kata dasar yang sama. Hal ini menurunkan redundansi istilah dan membuat distribusi kata dalam dokumen menjadi lebih representatif. Secara keseluruhan, preprocessing menghasilkan teks yang lebih bersih, ringkas, dan siap digunakan untuk pembentukan representasi vektor dengan kualitas yang lebih baik [3][12].

3.3. Hasil Pembentukan Representasi Vektor

Tahap pembentukan representasi vektor menghasilkan model dokumen berbasis Vector Space Model (VSM), di mana setiap dokumen direpresentasikan sebagai vektor numerik dalam ruang berdimensi banyak [8][18]. Setiap dimensi vektor merepresentasikan istilah unik hasil preprocessing. Hasil ini memungkinkan dokumen berita yang bersifat tekstual untuk diolah secara matematis. Dengan adanya representasi vektor, sistem dapat mengukur kedekatan antar dokumen dan antara dokumen dengan kueri secara kuantitatif. Representasi ini menjadi fondasi utama dalam proses perhitungan kemiripan dan pemeringkatan dokumen [19].

3.4. Hasil Pembobotan TF-IDF

Pembobotan TF-IDF menghasilkan nilai bobot yang bervariasi untuk setiap istilah dalam dokumen. Istilah yang muncul secara spesifik dalam suatu dokumen tetapi jarang ditemukan pada dokumen lain memperoleh bobot yang lebih tinggi, sedangkan istilah umum memiliki bobot yang relatif rendah [21][22]. Hasil ini menunjukkan bahwa TF-IDF berhasil menonjolkan kata-kata kunci yang mencerminkan topik utama dokumen, sehingga perbedaan karakteristik antar dokumen menjadi lebih jelas dalam ruang vektor. Dengan pembobotan ini, sistem tidak hanya mempertimbangkan keberadaan kata, tetapi juga tingkat kepentingannya dalam konteks keseluruhan koleksi dokumen [3][22].

3.5. Hasil Penyimpanan Indeks Vektor

Penyimpanan indeks vektor dalam bentuk file model (.pkl) menghasilkan sistem yang lebih efisien dari sisi waktu dan komputasi. Model yang telah tersimpan dapat dimuat kembali tanpa harus mengulang proses preprocessing dan pembobotan TF-IDF [22]. Hasil implementasi menunjukkan bahwa mekanisme ini mampu mempercepat waktu inisialisasi sistem dan menjaga konsistensi representasi dokumen antar sesi penggunaan. Dengan demikian, sistem lebih siap untuk digunakan dalam skenario pencarian berulang oleh pengguna [10].

3.6. Hasil Pemrosesan Kueri

Hasil pemrosesan kueri menunjukkan bahwa penerapan tahapan preprocessing yang sama antara dokumen dan kueri mampu menghasilkan kesesuaian istilah yang lebih baik. Kueri yang telah diproses dan diubah menjadi vektor berada pada ruang vektor yang sama dengan dokumen berita [23]. Hal ini memungkinkan sistem membandingkan kueri dan dokumen secara langsung dan adil. Kueri yang mengandung istilah relevan terhadap koleksi dokumen menghasilkan vektor yang memiliki arah mendekati vektor dokumen tertentu, sehingga proses pencarian menjadi lebih akurat [3][8].

3.7. Hasil Perhitungan Cosine Similarity

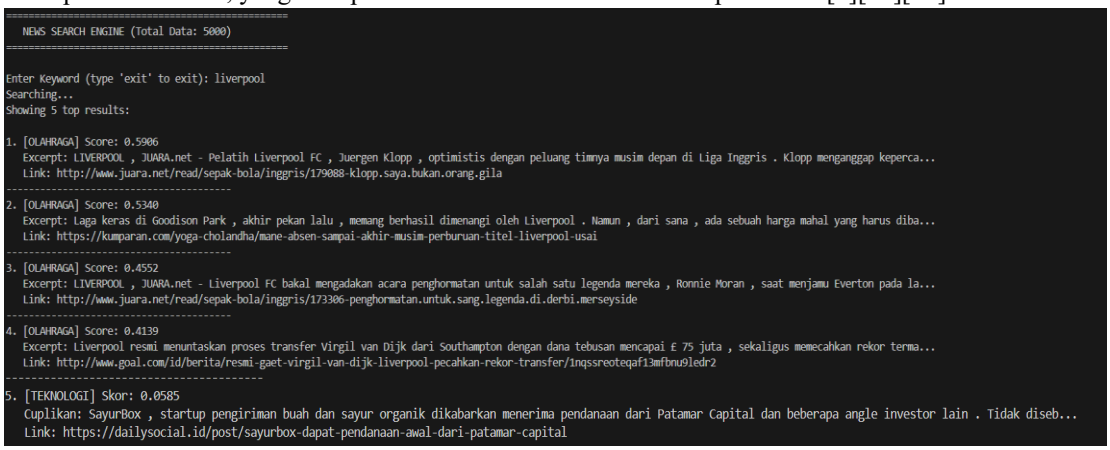
Perhitungan Cosine Similarity menghasilkan nilai kemiripan yang bervariasi antar dokumen terhadap kueri. Dokumen yang memiliki kesesuaian istilah dan konteks dengan kueri memperoleh nilai kemiripan yang lebih tinggi, sedangkan dokumen yang tidak relevan memperoleh nilai mendekati nol [5][24].

Hasil ini menunjukkan bahwa Cosine Similarity mampu membedakan tingkat relevansi dokumen secara proporsional, tanpa dipengaruhi oleh panjang dokumen atau jumlah kata. Pendekatan ini terbukti efektif dalam menilai kedekatan makna antara kueri dan dokumen dalam ruang vector [2][10].

3.8. Hasil Pemeringkatan Dokumen

Tahap pemeringkatan menghasilkan daftar dokumen berita yang tersusun berdasarkan nilai Cosine Similarity dari yang tertinggi hingga terendah. Dokumen dengan nilai kemiripan tertinggi muncul pada urutan teratas sebagai hasil pencarian utama [25].

Hasil pemeringkatan menunjukkan bahwa sistem mampu menyajikan dokumen yang relevan secara konsisten, sehingga membantu pengguna menemukan informasi yang sesuai dengan kebutuhan pencarian. Proses ini membuktikan bahwa integrasi VSM, TF-IDF, dan Cosine Similarity berhasil mendukung sistem temu balik informasi berita yang efektif dan terstruktur. Contoh tampilan hasil pemeringkatan dokumen berdasarkan kueri pengguna ditunjukkan pada Gambar 2, yang memperlihatkan urutan dokumen hasil pencarian [3][10][12].



Gambar 2. Hasil Pemeringkatan Dokumen

4. KESIMPULAN

Berdasarkan hasil penelitian dan pembahasan yang telah dilakukan, dapat disimpulkan bahwa penerapan metode Vector Space Model (VSM) dengan pembobotan TF-IDF serta pengukuran kemiripan menggunakan Cosine Similarity berhasil diimplementasikan secara efektif pada sistem temu balik informasi berita berbahasa Indonesia. Tujuan penelitian sebagaimana dijelaskan pada bagian pendahuluan, yaitu membangun sistem pencarian berita yang mampu menampilkan dokumen relevan berdasarkan kesesuaian isi dengan kueri pengguna, telah tercapai dengan baik.

Hasil pengujian menunjukkan bahwa tahapan preprocessing teks yang meliputi case folding, tokenisasi, stopword removal, dan stemming mampu meningkatkan kualitas representasi dokumen. Proses ini terbukti mengurangi noise teks dan menyatukan variasi kata, sehingga berdampak positif terhadap pembentukan vektor

dokumen dan kueri. Representasi dokumen dalam ruang vektor memungkinkan sistem memodelkan hubungan kemiripan secara matematis dan terukur. Pembobotan menggunakan TF-IDF menghasilkan bobot istilah yang lebih proporsional, di mana kata-kata yang bersifat spesifik terhadap topik berita memiliki kontribusi yang lebih besar dibandingkan kata-kata umum. Selanjutnya, penerapan Cosine Similarity mampu mengukur tingkat relevansi dokumen terhadap kueri secara objektif, tanpa dipengaruhi oleh panjang dokumen. Hasil pemeringkatan dokumen berdasarkan nilai kemiripan menunjukkan bahwa dokumen dengan konten paling sesuai dengan kueri berada pada urutan teratas, sebagaimana ditunjukkan pada hasil pengujian dan visualisasi sistem. Dengan demikian, sistem temu balik informasi yang dikembangkan telah mampu memberikan hasil pencarian yang relevan, terstruktur, dan efisien, serta dapat digunakan sebagai solusi pencarian berita berbasis teks dalam skala koleksi dokumen yang besar dan beragam.

Sebagai prospek pengembangan, penelitian selanjutnya dapat mengintegrasikan metode semantic-based retrieval seperti word embedding (Word2Vec, FastText, atau BERT) untuk menangani permasalahan sinonim dan konteks makna yang belum sepenuhnya dapat ditangkap oleh TF-IDF. Selain itu, sistem dapat dikembangkan lebih lanjut dengan menambahkan evaluasi kuantitatif menggunakan metrik seperti precision, recall, dan F1-score berdasarkan penilaian relevansi pengguna. Pengembangan antarmuka pengguna yang lebih interaktif serta penerapan sistem pada domain berita tertentu juga menjadi peluang penelitian lanjutan untuk meningkatkan manfaat dan penerapan sistem di dunia nyata.

REFERENSI

[1] Renjith R, "The Effect of Information Overload in Digital Media News Content Amrita Vishwa Vidyapeetham The Effect of Information Overload in Digital Media News Content," vol. 6, no. 1, pp. 73–85, 2017.

[2] G. Yunanda, D. Nurjanah, and S. Meliana, "Recommendation System from Microsoft News Data using TF-IDF and Cosine Similarity Methods," *Building of Informatics, Technology and Science (BITS)*, vol. 4, no. 1, pp. 277–284, 2022, doi: 10.47065/bits.v4i1.1670.

[3] K. D. Putung, A. S. M. Lumenta, and A. Jacobus, "Penerapan Sistem Temu Kembali Informasi Pada Kumpulan Dokumen Skripsi," *Jurnal Teknik Informatika*, vol. 8, no. 1, 2016, doi: 10.35793/jti.8.1.2016.12227.

[4] R. E. N. Rongcai, W. U. Guoxiong, and C. A. I. Ming, "No Implementasi Vector Space Model Dan Beberapa Notasi Metode Term Frequency Inverse Document Frequency (TF-IDF) pada Sistem Temu Kembali InformasiTitle".

[5] Nurhaliza and Suendri, "Document Similarity using Term Frequency-Inverse Document Frequency Representation and Cosine Similarity," *Journal.Ittelkom-Pwt.Ac.Id/Index.Php/Dinda*, vol. 5, no. 2, pp. 258–267, 2025.

[6] C. Van Gysel, M. De Rijke, and E. Kanoulas, "Neural vector spaces for unsupervised information retrieval," *ACM Trans. Inf. Syst.*, vol. 36, no. 4, 2018, doi: 10.1145/3196826.

[7] D. S. adillah Maylawati, Y. J. Kumar, and F. Kasmin, "Feature-based approach and sequential pattern mining to enhance quality of Indonesian automatic text summarization," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 30, no. 3, pp. 1795–1804, 2023, doi: 10.11591/ijeecs.v30.i3.pp1795-1804.

[8] G. Salton, A. Wong, and C. S. Yang, "A Vector Space Model for Automatic Indexing," *Commun. ACM*, vol. 18, no. 11, pp. 613–620, 1975, doi: 10.1145/361219.361220.

[9] A. A. J. Raj, "Fluorosis in Relation to Nutrition , Fluoride in Drinking Water and Socio Economic Background in Agastheeswaram Union , India," *International Journal of Innovations in Engineering and Technology*, vol. 2, no. Table 1, pp. 50–52, 2013.

[10] J. M. Hudin and A. Rifai, "Perancangan Sistem Temu Kembali Informasi Menggunakan Metode Vector Space Model Pada Pencarian Dokumen Berbasis Teks Berita," *Agustus*, no. 2, pp. 128–135, 2017.

[11] A. P. Widyassari *et al.*, "Review of automatic text summarization techniques & methods," *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 4, pp. 1029–1046, 2022, doi: 10.1016/j.jksuci.2020.05.006.

[12] I. Irmawati, "Sistem Temu Kembali Informasi Pada Dokumen Dengan Metode Vector Space Model," *Jurnal Ilmiah FIFO*, vol. 9, no. 1, p. 74, 2017, doi: 10.22441/fifo.v9i1.1444.

[13] J. Lin and C. Dyer, "Inverted Indexing for Text Retrieval," pp. 65–83, 2010, doi: 10.1007/978-3-031-02136-7_4.

[14] A. Saeroji, Rizka Andriyati, and M. Muhsin, "Analisis Efektivitas Aplikasi E-Arsip Sebagai Media Temu Kembali Informasi," *Jurnal Efisiensi :Kajian Ilmu Administrasi*, vol. 13, no. 1, pp. 1–100, 2015.

[15] H. Latiar, "Efektifitas Sistem Temu Kembali Arsip Digital Universitas Lancang Kuning Pekanbaru," *Jurnal Pustaka Budaya*, vol. 6, no. 1, pp. 9–15, 2019, doi: 10.31849/pb.v6i1.2131.

[16] D. Alwan and M. A. Ridla, "Jurnal Sistem dan Teknologi Informasi Indonesia Averaged Word2vec sebagai Ekstraksi Fitur pada Analisis Sentimen Ulasan Film di IMDb menggunakan Artificial Neural Network (ANN) Averaged Word2vec as Feature Extraction in Sentiment Analysis of Movie Revie," vol. 9, no. 1, pp. 36–45, 2024.

[17] Rianto, A. B. Mutiara, E. P. Wibowo, and P. I. Santosa, "Improving the accuracy of text classification using stemming method, a case of non-formal Indonesian conversation," *J. Big Data*, vol. 8, no. 1, pp. 1–16, 2021, doi: 10.1186/s40537-021-00413-1.

[18] M. N. Ellyanza, "Penerapan Vector Space Model Untuk Rekomendasi Produk Di E-Commerce," *JTIDIAll RJURNAL TEKNOLOGI DAN INOVASI DIGITA*, vol. 01, no. 01, pp. 50–62, 2024.

[19] A. Fauzi and G. Ginabila, "Information Retrieval System Pada Pencarian File Dokumen Berbasis Teks Dengan Metode Vector Space Model," *Jurnal Pilar Nusa Mandiri*, vol. 15, no. 1, pp. 41–46, 2019, doi: 10.33480/pilar.v15i1.61.

[20] D. Septiani and I. Isabela, "Analisis Term Frequency Inverse Document Frequency (Tf-Idf) Dalam Temu Kembali Informasi Pada Dokumen Teks," *Sintesia*, vol. 1, pp. 81–88, 2022.

[21] H. Christian, M. P. Agus, and D. Suhartono, "Single Document Automatic Text Summarization using Term Frequency-Inverse Document Frequency (TF-IDF)," *ComTech: Computer, Mathematics and Engineering Applications*, vol. 7, no. 4, p. 285, 2016, doi: 10.21512/comtech.v7i4.3746.

[22] N. P. Yunita, "Aplikasi Pencarian Hadis Menggunakan Vector Space Model Dengan Pembobotan TF-IDF Dan Confix-Stripping Stemmer," *Jurnal Teknologi Informasi dan Ilmu Komputer*, vol. 10, no. 3, pp. 665–676, Jul. 2023, doi: 10.25126/jtiik.2023106736.

[23] G. Salton, A. Wong, and C. S. Yang, "A vector space model for automatic indexing," *Commun. ACM*, vol. 18, no. 11, pp. 613–620, Nov. 1975, doi: 10.1145/361219.361220.

[24] H.-J. Goltz, "Functional data term models and semantic unification," 1988, pp. 158–167. doi: 10.1007/3-540-50667-5_68.

[25] R. Aurelia, H. Irsyad, and A. Rahman, "OPTIMASI RANGKING DOKUMEN DENGAN MODIFIKASI TF-IDF BERBASIS WAKTU PUBLIKASI DAN COSINE SIMILARITY," *JISCOM*, vol. 3.

[26] "PENGANTAR REDAKSI."