Jurnal Informatika: Jurnal pengembangan IT

Vol. 10, No. 4, 2025
ISSN : 2477-5126, DOI:10.30591/jpit.v10i4.8773

Perbandingan *Cosine Similarity* dan *Weighted Jaccard Similarity* dalam Pengembangan Mesin Pencari Perpustakaan Digital

Jessicha Putrianingsih Pamput [1], Aindri Rizky Muthmainnah [2], Dewi Fatmarani Surianto [3], Nur Fadilah [4]*
[1,2,3] Program Studi Teknik Komputer, Universitas Negeri Makassar, Jl. A.P.Pettarani, Makassar, 90222, Indonesia
[4]Program Studi Pendidikan Teknologi Informasi, Universitas Megarezky Makassar, Jl. Antang Raya, Makassar, 90234, Indonesia

Info Artikel

Riwayat Artikel:

Received 2025-05-16
Revised 2025-08-29
Accepted 2025-08-30

Corresponding Author:

Dewi Fatmarani Surianto

Email: dewifatmaranis@unm.ac.id

Abstract *– This study addressed the problem of low relevance in search results within the digital library system of the Department of Informatics and Computer Engineering (JTIK), Universitas Negeri Makassar. The purpose of this research was to improve the accuracy and relevance of search outcomes, enabling users, particularly students, to access academic materials and research references more efficiently. A search engine system was developed using a term-weighting method based on term frequency and document distribution. The system incorporated similarity measurement techniques to evaluate the degree of match between user queries and document content. An experimental approach was applied, which involved observation, data collection, text preprocessing, implementation of term weighting, and the comparison of cosine similarity and Weighted Jaccard similarity for ranking search results. The The evaluation was conducted using the Precision@K metric and a paired t-test to measure the significance of performance differences between methods. The test results showed that Weighted Jaccard obtained an average Precision@K value of 0.933, slightly higher than Cosine Similarity with an average of 0.9. However, Cosine Similarity produced a higher average similarity value. In addition, system testing was conducted in two stages, namely assessing user satisfaction with search results and assessing system performance. These findings confirmed that the combination of term-weighting and cosine similarity effectively enhanced the relevance and performance of digital library search systems.*

Keywords: *Comparison; Cosine Similarity; Digital Library; Search Engine; Weighted Jaccard Similarity*

Abstrak *– Penelitian ini membahas masalah rendahnya relevansi hasil pencarian dalam sistem perpustakaan digital Jurusan Teknik Informatika dan Komputer (JTIK) Universitas Negeri Makassar. Tujuan dari penelitian ini adalah untuk meningkatkan akurasi dan relevansi hasil pencarian, sehingga memungkinkan pengguna, terutama mahasiswa, untuk mengakses materi akademik dan referensi penelitian dengan lebih efisien. Sebuah sistem mesin pencari dikembangkan dengan menggunakan metode pembobotan term berdasarkan frekuensi term dan distribusi dokumen. Sistem ini menggabungkan teknik pengukuran kemiripan untuk mengevaluasi tingkat kecocokan antara pertanyaan pengguna dan konten dokumen. Pendekatan eksperimental diterapkan, yang melibatkan observasi, pengumpulan data, prapemrosesan teks, implementasi pembobotan term, dan perbandingan Cosine Similarity dan Weighted Jaccard Similarity untuk menentukan peringkat hasil pencarian. Evaluasi dilakukan menggunakan metrik Precision@K serta uji statistik paired t-test untuk mengukur signifikansi perbedaan kinerja antar metode. Hasil pengujian menunjukkan bahwa Weighted Jaccard memperoleh nilai Precision@K rata-rata 0.933, sedikit lebih tinggi dibandingkan Cosine Similarity dengan rata-rata 0.9. Namun, Cosine Similarity menghasilkan rata-rata nilai kemiripan yang lebih tinggi. Selain itu, pengujian sistem dilakukan melalui dua tahap, yaitu penilaian kepuasan pengguna terhadap hasil pencarian dan penilaian terhadap performa sistem. Temuan ini menegaskan bahwa kombinasi pembobotan istilah dan kemiripan kosinus secara efektif meningkatkan relevansi dan kinerja sistem pencarian perpustakaan digital.*

Kata Kunci*: Cosine Similarity;* Mesin Pencari; Perbandingan; Perpustakaan Digital; *Weighted Jaccard Similarity*

I. PENDAHULUAN

Berdasarkan Undang-Undang Nomor 43 Tahun 2007 tentang Perpustakaan, "Perpustakaan adalah institusi pengelola koleksi karya tulis, karya cetak, dan/atau karya rekam secara profesional dengan sistem yang baku guna memenuhi kebutuhan pendidikan, penelitian, informasi, dan rekreasi para pemustaka" [1]. Perpustakaan berfungsi sebagai pusat informasi yang menyediakan berbagai sumber daya untuk mendukung kegiatan belajar dan penelitian bagi berbagai kalangan, termasuk siswa, mahasiswa, peneliti, dan masyarakat umum. Pada era modern ini, akses informasi di perpustakaan kebanyakan masih memerlukan kehadiran fisik, di mana pengguna harus datang untuk mencari dan meminjam bahan bacaan [2]. Layanan manual ini sering kali tidak efisien dalam penggunaan waktu [3]. Namun, dengan perkembangan teknologi informasi, konsep perpustakaan beralih ke digitalisasi.

Perpustakaan digital merupakan bentuk modern dari perpustakaan yang menyajikan seluruh koleksi informasi dan proses pengelolaannya dalam format digital, memungkinkan akses kapan saja dan di mana saja. Fasilitas ini mempermudah mahasiswa dalam mencari, mengunduh, dan membaca referensi akademik melalui berbagai perangkat seperti komputer atau ponsel [4], [5], [6]. Namun, pada kenyataannya, perpustakaan digital sering kali menghadapi tantangan dalam memberikan hasil pencarian yang kurang relevan dengan kebutuhan pengguna.

Masalah utama dalam implementasi perpustakaan digital adalah meningkatkan relevansi hasil pencarian, karena banyak sistem pencarian hanya menggunakan kata kunci sederhana tanpa mempertimbangkan konteks atau tujuan pengguna [7][8]. Akibatnya, hasil pencarian sering kali tidak sesuai harapan, memaksa mahasiswa untuk menyaring informasi secara manual [7]. Hal ini memperlambat proses pencarian dan mengurangi keefektifan perpustakaan digital dalam menunjang aktivitas akademik.

Penelitian sebelumnya menunjukkan bahwa perpustakaan digital masih menghadapi berbagai tantangan untuk meningkatkan relevansi dan aksesibilitas informasi. Salah satu penelitian mengkaji efektivitas pencarian buku dan akurasi sistem pada e-library menggunakan algoritma *binary search* dan *hamming distance*. Hasil penelitian tersebut mengindikasikan bahwa kedua metode ini hanya mencapai akurasi antara 72% hingga 78% [9]. Penelitian lain membandingkan kinerja pencarian dokumen menggunakan metode *word embedding* pada USU *Repository*, yang memperlihatkan rata-rata presisi hingga 73% dengan dataset berjumlah 664 dokumen [10]. Ada juga penelitian yang berfokus pada pengembangan fasilitas perpustakaan digital dan manajemen koleksi buku serta arsip untuk meningkatkan aksesibilitas informasi bagi masyarakat Provinsi Jawa Barat dengan meggunakan metode kualitatif deskriptif dengan pendekatan studi literatur [11]. Selain itu, pengelolaan pengembangan perpustakaan digital untuk mempermudah akses informasi di era informasi dengan metode analisis elemen dalam pengembangan perpustakaan digital, seperti infrastruktur teknologi informasi, metadata, dan sistem temu kembali informasi [4].

Penelitian oleh [12] mengembangkan sistem manajemen perpustakaan digital dengan mengintegrasikan teknik *Named Entity Recognition* (NER) dan *Word Sense Disambiguation* (WSD), yang diaplikasikan secara khusus untuk pemrosesan bahasa Arab dan Inggris. Lebih lanjut, penelitian oleh [13] menunjukkan bahwa LaBSE embedding secara signifikan lebih unggul daripada Latin BERT dalam pencarian informasi di Perpustakaan Digital Latin. Hasil penelitian juga menunjukkan bahwa kueri dalam Bahasa Inggris berkinerja lebih baik, karena bias pelatihan LaBSE terhadap bahasa Inggris. Sementara itu, temuan dalam [14] berkontribusi pada pengembangan layanan perpustakaan yang lebih efektif serta tinjauan kurikulum pendidikan di wilayah Sindh, Pakistan, khususnya yang berkaitan dengan keterampilan literasi informasi digital untuk para profesional perpustakaan (LIS). Sementara itu, penelitian oleh [15] mengusulkan model pencarian dokumen di perpustakaan digital yang mengintegrasikan koreksi ejaan dan perluasan kueri dengan menggunakan algoritma *Levenshtein Distance* untuk mengoreksi kesalahan pengetikan dalam kueri pengguna.

Selain itu, penelitian lain telah mengeksplorasi pengembangan mesin pencari berbasis komputasi awan yang bertujuan untuk meningkatkan efisiensi pencarian informasi dalam sistem perpustakaan digital, dengan menggabungkan teori ekologi dan menekankan perlunya memahami karakteristik pengguna untuk memenuhi kebutuhan informasi [16]. Dalam konteks yang berbeda, [17] mengembangkan sistem pendeteksi berita palsu otomatis menggunakan model berbasis BERT dan GPT untuk meningkatkan pemahaman konten, sementara [18] membuat mesin pencari terkait COVID-19 dengan menggunakan model berbasis kata kunci TF-IDF dan BM25. Berdasarkan tinjauan terhadap penelitian-penelitian sebelumnya, dapat disimpulkan bahwa relevansi hasil pencarian pada sistem perpustakaan digital masih belum optimal. Selain itu, masih kurangnya studi komparatif yang mengevaluasi keefektifan pengukuran relevansi menggunakan *cosine similarity* dan *jaccard similarity*. Kesenjangan ini menyoroti peluang untuk melakukan penelitian lebih lanjut dengan menerapkan model berbasis kata kunci seperti TF-IDF sambil mengintegrasikan dan membandingkan kedua ukuran kemiripan tersebut. Penelitian ini berusaha untuk berkontribusi dalam meningkatkan relevansi pencarian dan menangani preferensi pengguna dalam sistem perpustakaan digital.

Jurusan Teknik Informatika dan Komputer (JTIK) di Universitas Negeri Makassar (UNM), yang menjadi lokasi studi kasus ini, juga menghadapi keterbatasan dalam sistem temu balik informasi untuk perpustakaan. Berdasarkan observasi dan wawancara yang dilakukan dengan staf pengelola perpustakaan di JTIK, ditemukan bahwa perpustakaan menghadapi tantangan dalam pencarian informasi yang relevan, terutama ketika mahasiswa mencari buku tanpa memberikan spesifikasi yang jelas. Dalam kasus seperti ini, pencarian masih dilakukan secara manual oleh staf perpustakaan, yang harus menelusuri berbagai referensi buku tanpa informasi yang rinci. Pendekatan manual ini sering kali menghasilkan referensi yang kurang relevan dengan kebutuhan spesifik siswa, sehingga menghambat efisiensi dan efektivitas proses pencarian informasi.

Berdasarkan permasalahan yang ada, pengembangan sistem pencarian pada perpustakaan digital diharapkan dapat meningkatkan kenyamanan serta efektivitas penggunaannya, khususnya bagi mahasiswa [15], [16], [17]. Penelitian ini bertujuan untuk memberikan solusi terhadap tantangan yang dihadapi oleh pengguna

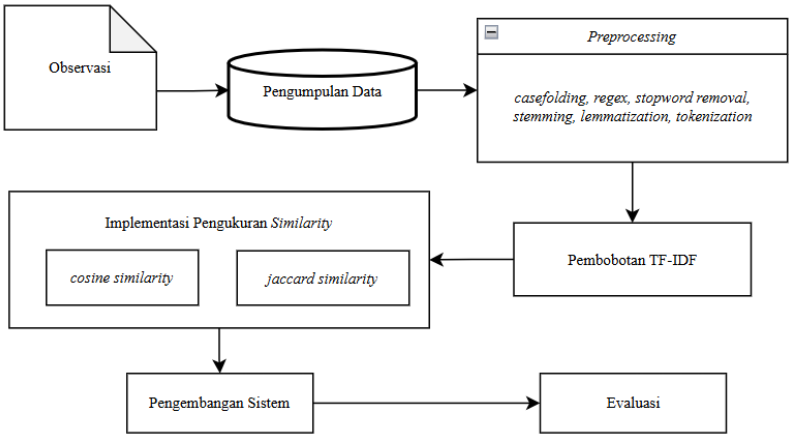
perpustakaan digital, khususnya di JTIK, sehingga perpustakaan dapat lebih mendukung kebutuhan akademik mahasiswa.

Untuk mengatasi masalah pencarian yang ada, implementasi sistem pencarian yang lebih canggih diperlukan untuk meningkatkan akurasi dan relevansi. Salah satu solusi yang dapat diterapkan adalah algoritma *Term Frequency-Inverse Document Frequency* (TF-IDF), yang menilai tingkat kepentingan term dalam sebuah dokumen berdasarkan frekuensi kemunculan term tersebut dalam dokumen dan tingkat kelangkaan term tersebut di seluruh koleksi dokumen, di mana metode ini memprioritaskan kata-kata umum, sedangkan kata-kata khusus yang jarang muncul tetapi relevan diprioritaskan [19]. Selain itu, perhitungan jarak dokumen menggunakan *Vector Space Model* (VSM), khususnya dengan metode *cosine similarity*, juga dapat diterapkan untuk mengukur kemiripan antar dokumen dan memastikan bahwa hasil pencarian yang ditampilkan relevan dengan kueri pengguna [19], [20]. Sementara VSM menunjukkan dokumen dan kueri sebagai vektor dalam ruang multidimensi, *cosine similarity* menghitung sudut kosinus antar vektor, sehingga sudut yang lebih kecil berarti tingkat kemiripan yang lebih tinggi.

Dengan algoritma tersebut, penelitian ini diharapkan dapat memberikan kontribusi dalam meningkatkan relevansi dan akurasi hasil pencarian, sehingga memudahkan pengguna dalam menemukan informasi yang dibutuhkan di perpustakaan digital, khususnya di lingkungan akademis JTIK. Dengan menerapkan model pencarian ini, sistem diharapkan dapat memenuhi kebutuhan penggunanya, sehingga dapat mendukung proses pembelajaran dan penelitian mahasiswa. Selain itu, sistem ini tidak hanya menghilangkan kebutuhan pencarian manual, namun juga mampu meningkatkan efisiensi waktu dan akurasi hasil pencarian, sehingga pengguna dalam lingkungan akademis JTIK dapat menemukan informasi yang relevan secara cepat dan efektif.

II. METODE

Tahapan-tahapan penelitian ini direncanakan dengan maksud untuk mencapai pemahaman yang sistematis dan berperan sebagai arahan dalam menyelesaikan penelitian ini. Berikut ini adalah Gambar 1 struktur urutan tahapan.



Gambar 1. Tahapan Penelitian

A. Observasi

Tahap awal penelitian ini melibatkan proses pengamatan langsung terhadap perpustakaan JTIK UNM sebagai lokasi studi kasus untuk memahami kebutuhan akan sistem pencarian yang efektif. Observasi dilakukan untuk menggali bagaimana pengguna saat ini mencari informasi atau koleksi di perpustakaan tanpa dukungan sistem pencarian digital. Hal ini mencakup pengamatan terhadap proses manual seperti pencatatan katalog fisik, pertanyaan langsung kepada petugas, atau metode lainnya yang digunakan. Dari pengamatan ini, diharapkan dapat diidentifikasi tantangan utama yang dihadapi pengguna, seperti waktu pencarian yang lama atau kesulitan menemukan koleksi yang relevan, sehingga dapat menjadi landasan untuk merancang sistem pencarian yang sesuai dengan kebutuhan perpustakaan dan penggunanya.

B. Pengumpulan Data

Proses pengumpulan data dalam penelitian ini dilakukan melalui observasi dan wawancara dengan pihak pengelola perpustakaan untuk memahami secara mendalam proses pencarian buku yang dilakukan baik oleh pengelola maupun pengunjung perpustakaan. Data utama yang digunakan diperoleh langsung dari database perpustakaan terkait, mencakup informasi tentang 701 buku yang tersedia, yang akan dianalisis untuk mengembangkan dan menguji sistem pencarian yang lebih efektif dan relevan.

C. Preprocessing Data

Tahap *preprocessing* data dalam penelitian ini bertujuan untuk membersihkan dan menyamaratakan seluruh dataset agar dapat digunakan secara efektif dalam sistem pencarian. Pada tahapan ini, dilakukan beberapa proses, di antaranya *casefolding* untuk menyamakan huruf kapital, penggunaan *regular expressions* (regex) untuk menghapus karakter yang tidak relevan, *tokenization* untuk memecah teks menjadi unit-unit kata, *stopword removal* untuk menghapus kata-kata umum yang tidak memberikan makna signifikan, serta *stemming* dan *lemmatization* untuk mengubah kata ke bentuk dasarnya menggunakan beberapa toolkit, seperti Sastrawi, NLTK, dan LangDetect untuk identifikasi Bahasa Indonesia ataupun Bahasa Inggris.

D. Pembobotan TF-IDF

Proses pembobotan TF-IDF bertujuan untuk memberikan bobot ke setiap *terms* yang ada pada dataset, sehingga istilah yang lebih relevan dan penting bagi konteks pencarian mendapatkan bobot yang lebih tinggi [20]. TF-IDF memberikan bobot yang lebih tinggi pada istilah yang sering muncul dalam suatu dokumen, namun jarang muncul di dokumen lain, sehingga meningkatkan akurasi dalam proses pencarian [21].

$$TF(t,d) = \frac{f(t,d)}{\sum_{k \in d} f(k,d)} \tag{1}$$

$$IDF(t) = log\,\frac{N}{1+n(t)} \tag{2}$$

$$TF - IDF(t,d) = TF(t,d) \times IDF(t) \tag{3}$$

E. Implementasi Metode Pengukuran Similarity

Pada penelitian ini, diimplementasikan dua model pengukuran *similarity* untuk membandingkan efektivitas metode pengukuran dalam sistem pencarian perpustakaan, yaitu *cosine similarity* dan *weighted jaccard similarity*. Kedua metode ini digunakan untuk menilai tingkat kemiripan antara *query* pengguna dan dokumen dalam koleksi perpustakaan.

1) *Cosine Similarity*: Metode *cosine similarity* mengukur kesamaan antara dua vektor dengan menghitung nilai kosinus dari sudut di antara keduanya dalam ruang multidimensi. Semakin kecil sudut antara vektor, semakin tinggi nilai *cosine similarity*, yang menunjukkan tingkat kemiripan yang lebih besar [22], [23].

$$Cosine\ Similarity\ (A,B) = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} {A_i}^2} \times \sqrt{\sum_{i=1}^{n} {B_i}^2}} \tag{4}$$

2) *Weighted Jaccard Similarity*: Metode *weighted jaccard similarity* berfungsi untuk mengukur kesamaan antar vektor dengan nilai riil, di mana elemen-elemen tersebut memiliki bobot minimum yang dibagi dengan jumlah bobot maksimum untuk setiap elemen [24].

$$Weighted\ Jaccard\ Similarity(A,B) = \frac{\sum_{i=1}^{n} \min(w_{iA}, w_{iB})}{\sum_{i=1}^{n} \max(w_{iA}, w_{iB})} \tag{5}$$

F. Penerapan Sistem

Setelah tahap pengukuran *similarity*, kedua metode yaitu *cosine similarity* dan *weighted jaccard similarity* akan dievaluasi untuk menentukan mana yang lebih optimal dalam konteks sistem pencarian perpustakaan. Metode yang terpilih sebagai yang paling efektif akan dikombinasikan dengan metode pembobotan TF-IDF dalam tahap penerapan sistem ini. Kombinasi ini bertujuan untuk meningkatkan relevansi hasil pencarian dengan mempertimbangkan baik bobot istilah dalam dokumen maupun tingkat kemiripan antara *query* pengguna dan dokumen dalam koleksi perpustakaan.

G. Evaluasi

Tahap evaluasi merupakan proses untuk mengevaluasi efektivitas relevansi dari sistem pencarian perpustakaan yang telah diterapkan. Pada tahap ini, dilakukan evaluasi model dan evaluasi sistem.

1) *Evaluasi Model*: Evaluasi model bertujuan untuk mengukur kinerja algoritma atau metode yang digunakan dalam sistem pencarian, khususnya dalam menilai tingkat akurasi relevansi hasil pencarian. Pada penelitian ini, digunakan metode evaluasi *precision@K* yang berfokus pada menghitung proporsi dokumen relevan di antara K dokumen teratas yang dihasilkan oleh sistem, dengan nilai K yang ditetapkan

adalah 5. Pemilihan nilai K =5 dianggap proporsional untuk menjaga konsistensi hasil evaluasi karena penetapan nilai ini didasarkan pada keterbatasan jumlah data yang dimiliki, terutama pada beberapa kategori yang masih sangat kecil. Pendekatan ini memungkinkan pengukuran efektivitas algoritma dalam menghasilkan hasil pencarian yang sesuai dengan kebutuhan pengguna pada posisi prioritas [25].

$$Precision@K = \frac{Jumlah\ dokumen\ relevan\ dalam\ K\ teratas}{K} \quad (5)$$

2) *Evaluasi Sistem*: Evaluasi sistem bertujuan untuk menilai performa keseluruhan sistem dalam memenuhi kebutuhan pengguna. Aspek yang dianalisis meliputi relevansi hasil pencarian, tingkat informatif dari hasil yang ditampilkan, kecepatan respon sistem, dan tingkat kepuasan pengguna terhadap kinerja sistem. Penilaian ini melibatkan 36 responden, yang memberikan umpan balik berdasarkan pengalaman mereka selama menggunakan sistem, sehingga menghasilkan data yang komprehensif untuk mengevaluasi efektivitas sistem secara menyeluruh.

Selain itu, pengujian fungsional juga dilakukan dengan metode black box testing yang melibatkan 30 responden. *Black box testing* digunakan untuk mengevaluasi apakah fungsi-fungsi utama sistem telah berjalan sesuai dengan yang diharapkan, seperti penerimaan input *query*, keakuratan hasil pencarian, serta kinerja tombol pencarian informasi detail dokumen.

III. HASIL DAN PEMBAHASAN

Pada tahap ini, pembahasan akan berfokus pada proses pengumpulan dan pengolahan data yang dilakukan untuk mencapai hasil dalam meningkatkan relevansi hasil pencarian pada perpustakaan JTIK. Proses ini sangat penting untuk memastikan bahwa data yang ditemukan sesuai dengan *query* pencarian dan dapat diolah dengan metode yang tepat, sehingga menghasilkan hasil pencarian yang relevan dan sesuai dengan tujuan pengguna dalam sistem pencarian berbasis perpustakaan.

A. Pengumpulan Data

Pengumpukan data pada penelitian ini dilakukan melalui metode observasi dan wawancara terhadap pengurus perpustakaan JTIK untuk memperoleh informasi mendalam terkait sistem dan proses pencarian buku pada perpustakaan JTIK. Selain itu, data utama yang digunakan pada penelitian ini diperoleh langsung dari database buku yang terdapat pada perpustakaan dengan jumlah 701 buku, sehingga memastikan bahwa data yang digunakan bersifat aktual dan relevan. Data yang diperoleh direpresentasikan dalam Tabel 1 di bawah ini.

TABEL 1
HASIL PENGUMPULAN DATA

Kode	Judul	Penulis	Tempat dan Tahun	Deskripsi
TI-001	Integrasi Teknologi Informasi Dengan Strategi	Dr. Ike Janita Dewi, MBA	Yogyakarta, 2005	Model "the five competitive forces" dan "the three generic strategies"dari Porter menjadi framework yang sangat mendasar, dengan mana setiap kebijakan manajemen IT (Information Technology) bisa direfleksikan manfaat strategisnya…
…	…	…	…	…
GM-008	Aplikasi Android Game Pembelajaran AUGMENTED REALITY BERBASIS UNITY	Wahyu Hari Kristiyanto	Yogyakarta, 2020	Buku berjudul A to Z Pembuatan dengan Mudah Aplikasi Android Game Pembelajaran Augmented Reality Berbasis Unity ini merupakan karya kolaborasi lintas generasi melalui kerja online/dalam jaringan (daring) dari rumah masing-masing maupun tatap muka/luar jaringan (luring) dengan mematuhi Protokol Kesehatan sesuai instruksi Pemerintah Republik Indonesia terkait serangan virus Covid- 19….

B. Pengolahan Data

Dari 701 data buku yang terkumpul, tahap pengolahan data dimulai dengan menggabungkan kolom judul dan deskripsi buku menjadi satu kolom teks. Penggabungan ini dilakukan karena judul saja sering kali kurang memberikan konteks lengkap mengenai isi buku, sedangkan deskripsi memuat informasi tambahan yang memperkaya representasi dokumen. Maka, dengan kombinasi judul dan deskripsi buku menjadi *field* indeks untuk memastikan bahwa sistem pencarian dapat memanfaatkan informasi yang lebih komprehensif. Setelah itu, dilakukan tahap *preprocessing* pada kolom gabungan judul dan deskripsi tersebut, yang terdiri dari beberapa langkah, seperti *lower case folding* untuk mengubah seluruh teks menjadi huruf kecil, *stopword removal* untuk menghapus kata-kata umum yang tidak bermakna, penerapan *regular expression* (*regex*) untuk membersihkan data dari karakter atau simbol yang tidak relevan, *tokenization* untuk memecah teks menjadi unit-unit kata yang lebih kecil, *stemming* untuk mengubah kata Bahasa Indonesia ke bentuk dasarnya, dan *lemmatization* untuk mengubah kata bahasa Inggris ke bentuk datasnya. Hasil dari tahap preprocessing ini disajikan dalam Tabel 2 di bawah ini.

TABEL 2
HASIL PREPROCESSING

Judul	Deskripsi	Gabungan	gabungan_clean
Integrasi Teknologi Informasi Dengan Strategi	Model "the five competitive forces" dan "the three generic strategies"…	Integrasi Teknologi Informasi Dengan Strategi Model "the five competitive forces" dan "the three generic strategies"dari Porter menjadi framework...	integrasi teknologi informasi strategi model the five competitive forces the three generic strategies dari porter framework…
…	…	…	…
Aplikasi Android Game Pembelajaran AUGMENTED REALITY BERBASIS UNITY	Buku berjudul A to Z Pembuatan dengan Mudah Aplikasi Android Game Pembelajaran Augmented Reality Berbasis Unity...	Aplikasi Android Game Pembelajaran AUGMENTED REALITY BERBASIS UNITY Buku berjudul A to Z Pembuatan dengan Mudah Aplikasi Android Game Pembelajaran Augmented Reality Berbasis Unity...	aplikasi android game ajar augmented reality bas unity buku judul a to z buat mudah aplikasi android game ajar augmented reality…

Setelah melalui tahap preprocessing, data diproses lebih lanjut dengan ektraksi fitur menggunakan pembobotan TF-IDF (*Term Frequency-Inverse Document Frequency*). Proses ini melibatkan pemecahan data menjadi daftar kata yang selanjutnya akan disajikan dalam bentuk matriks pembobotan. Jumlah kata dalam matriks ini ditentukan oleh kata yang terdapat dalam dataset, di mana setiap kata diberikan bobot berdasarkan frekuensi kemunculannya dalam dokumen dan seberapa jarang kata tersebut muncul di seluruh dokumen. Semakin besar bobot sebuah kata dalam matriks, semakin spesifik kata tersebut dalam suatu dokumen tertentu [26]. Pada penelitian ini, hasil pembobotan TF-IDF pada kolom deskripsi_clean menghasilkan matriks berukuran 701 x 2973.

C. Perhitungan Similaritas

Setelah dokumen direpresentasikan dalam bentuk vektor dengan menggunakan metode TF-IDF, langkah selanjutnya adalah menghitung tingkat kemiripan antara *query* dengan dokumen yang ada di dalam koleksi. Perhitungan ini bertujuan untuk mengukur relevansi sebuah dokumen terhadap *query* yang diberikan dengan menggunakan metrik kemiripan tertentu. Pada penelitian ini, dua pendekatan yang digunakan untuk menghitung kemiripan adalah *Cosine Similarity* dan *Weighted Jaccard Similarity*.

1) Cosine Similarity

Query yang diberikan oleh pengguna juga akan dilakukan pembobotan menggunakan metode TF-IDF untuk melihat seberapa penting setiap kata dalam konteks pencarian, sehingga dapat meningkatkan relevansi hasil yang dihasilkan. Selanjutnya, menghitung *cosine similarity* antara *query* dan dokumen buku yang ada pada dataset. *Cosine similarity* mengukur derajat kesamaan antara *vector* representasi TF-IDF dari *query* dan dataset buku perpustakaan. Hasil perhitungan ini kemudian digunakan untuk menampilkan 10 buku paling relevan dengan *query* pengguna yang direpresentasikan pada Tabel 3 berikut.

TABEL 3
PERHITUNGAN JARAK *QUERY* DAN DOKUMEN MENGGUNAKAN *COSINE SIMILARITY*

***Query*: buku tentang program dan website**	
Judul	*Cosine Similarity*
Mudah Membuat Website Menggunakan CodeIgniter	0.277024
Cara Menguasai Pemrograman Website Secara Otod...	0.268735
Cara Instan Menguasai Pemrograman Website seca...	0.264945
Website No.1 Cara Mudah Bikin Website dan Prom...	0.260457
Website No.1	0.254199
Membuat Website Sendiri dengan PHP-MySQL	0.243146
Pengembangan Web dengan Jquery	0.241381
Jago Membuat Website & SEO	0.226326
Membuat Website Gratis Tanpa Guru	0.195689
Panduan Praktis Membuat Website Gratis secara ...	0.181827

Berdasarkan hasil pencarian menggunakan *cosine similarity* untuk *query* "buku tentang program dan website", buku dengan nilai cosine similarity tertinggi, yaitu "Mudah Membuat Website Menggunakan CodeIgniter" (0.277024), menunjukkan kemiripan yang paling tinggi dengan *query*. Hal ini mengindikasikan bahwa buku tersebut memiliki relevansi yang lebih besar terhadap topik pencarian, khususnya terkait pembuatan website menggunakan *framework CodeIgniter*. Sementara itu, buku dengan nilai *cosine similarity* lebih rendah, seperti "Panduan Praktis Membuat Website Gratis secara ..." (0.181827), memiliki tingkat relevansi yang lebih rendah, meskipun masih berkaitan dengan topik pembuatan *website*. Secara keseluruhan, semakin tinggi nilai *cosine similarity*, semakin besar kemungkinan buku tersebut relevan dengan kebutuhan pengguna yang mencari informasi tentang program dan pembuatan *website*.

2) Weighted Jaccard Similarity

Pada penelitian ini juga menggunakan perhitungan *weighted jaccard similarity* sebagai metode perbandingan untuk mengukur kesamaan antara *query* pengguna dan dokumen buku yang dimiliki. *Weighted jaccard similarity* mempertimbangkan bobot setiap *term* berdasarkan frekuensi kemunculan atau pembobotan TF-IDF dengan menjumlahkan bobot minimum dari setiap term yang ada pada kedua dokumen sebagai irisan. Hasil implementasi perhitungan *weighted jaccard similarity* direpresentasikan pada Tabel 4 berikut.

TABEL 4
PERHITUNGAN JARAK QUERY DAN DOKUMEN MENGGUNAKAN WEIGHTED JACCARD SIMILARITY

***Query*: buku tentang program dan website**	
Judul	*Weighted Jaccard Similarity*
Membuat Website Sendiri dengan PHP-MySQL	0.099089
Mudah Membuat Website Menggunakan CodeIgniter	0.093380
Membuat Website Gratis Tanpa Guru	0.090698
Panduan Praktis Membuat Website Gratis secara ...	0.086311
Jago Membuat Website & SEO	0.083501
HTML, PHP, dan Mysql untuk Pemula	0.066709
Website No.1	0.066558
Mudah & Cepat Membuat Program Skripsi & TA den...	0.065394
Otodidak Desain dan Pemrograman Website	0.063597
Website No.1 Cara Mudah Bikin Website dan Prom...	0.062558

Hasil pencarian menggunakan *Weighted Jaccard similarity* untuk kueri "buku tentang program dan website" menunjukkan bahwa buku "Membuat Website Sendiri dengan PHP-MySQL" (0.099089) memiliki nilai kemiripan tertinggi, Nilai ini diikuti oleh "Mudah Membuat Website Menggunakan CodeIgniter" (0.093380) dan "Membuat Website Gratis Tanpa Guru" (0.090698) yang juga relevan dengan konteks kueri. Temuan ini menunjukkan bahwa *Weighted Jaccard* dapat menghasilkan hasil pencarian yang lebih sesuai dengan mempertimbangkan bobot *term* karena dia melihat keberadaan kata yang sama dan tingkat kepentingannya dalam dokumen. Meskipun nilai kemiripannya relatif kecil, buku-buku yang berada di peringkat atas tabel memiliki hubungan langsung dengan kueri. Metode ini terbukti lebih efisien dalam mengukur relevansi daripada metode *Jaccard* yang bergantung pada keberadaan kata.

D. Evaluasi

Meskipun kedua metode, yaitu *cosine similarity* dan *weighted jaccard similarity*, digunakan untuk mengukur kesamaan antara *query* pengguna dan dataset buku, tidak dapat dikatakan bahwa salah satu metode lebih baik dari yang lainnnya. Oleh karena itu, untuk mengevaluasi hasil relevansi, sistem diuji menggunakan *precision@K*, yang akan mengukur proporsi dataset yang relevan dalam k dataset teratas, serta memberikan gambaran objektif kinerja kedua metode dalam menghasilkan hasil pencarian yang sesuai kebutuhan pengguna.

Dalam tahap evaluasi ini dilakukan uji coba menggunakan 10 *query* berbeda untuk mengevaluasi kinerja dari kedua metode *similarity*, yaitu *cosine* dan *jaccard*. Hasil pengujian pada Tabel 5 ini dapat dibandingkan

untuk mengetahui metode *similarity* mana yang lebih unggul dalam memberikan hasil pencarian yang relevan dan sesuai dengan kebutuhan pengguna.

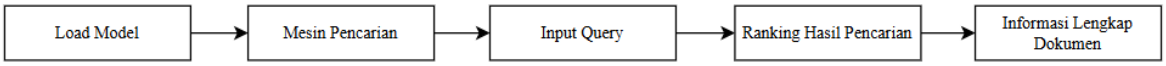
TABEL 5
PERBANDINGAN COSINE SIMILARITY DAN WEIGHTED JACCARD SIMILARITY BERDASARKAN PRECISION@K

No	*Query*	Nilai *Precision@K*			
		Cosine Similarity	*Rata-rata Cosine*	*Weighted Jaccard Similarity*	*Rata-rata Weighted Jaccard*
1	"buku tentang photoshop"	1	0.33971	1	0.1133
2	"buku tentang metodologi penelitian"	1	0.175771	1	0.076609
3	"buku tentang matematika"	0.80	0.270905	0.80	0.12694
4	"buku tentang *augmented reality*"	1	0.137692	1	0.0724
5	"buku tentang *machine learning*"	0.80	0.404968	0.80	0.1145
6	"buku tentang *Microsoft Office*"	0.80	0.401684	o.80	0.14724
7	"buku tentang MySQL"	0.60	0.236572	0.80	0.0736
8	"buku tentang PHP"	1	0.322046	1	0.0926
9	"buku tentang linux"	1	0.225033	1	0.07308
10	"buku tentang arduino"	1	0.322947	1	0.09316
	Rata-rata	0.9	0.283733	0.933	0.098343

Berdasarkan Tabel 5, pengukuran *similarity* menggunakan metode *cosine* menghasilkan nilai *precision@K* dengan rata-rata sebesar 0.9, sedangkan metode *weighted jaccard* menghasilkan nilai *precision@K* sebesar 0.933. Hasil ini menunjukkan bahwa secara akurasi pencarian berdasarkan precision@K, metode *weighted jaccard* sedikit lebih unggul dibandingkan *cosine similarity*. Namun, *cosine similarity* memberikan hasil yang lebih tinggi sebesar 0.2837 dibandingkan dengan nilai *weighted jaccard* 0.0983. Hal ini menunjukkan bahwa *cosine similarity* lebih mampu menunjukkan tingkat kemiripan konsisten antar dokumen, meskipun tidak selalu berdampak langsung pada nilai precision@K. Uji *paired t* dilakukan untuk memastikan bahwa perbedaan performa antara *cosine similarity* dan *weighted jaccard similarity* signifikan secara statistik. Hasilnya menunjukkan nilai t-statistic sebesar 8.1473 dan nilai p-value sebesar 0.0003. Penelitian sebelumnya juga mendukung temuan ini, dengan menunjukkan bahwa *cosine similarity* lebih efektif dalam menangani data teks dan menghasilkan pencocokan yang lebih akurat, sementara *jaccard similarity* cenderung kurang optimal untuk teks yang kompleks [27],[28],[29].

E. Implementasi Sistem

Berdasarkan hasil evaluasi, sistem pencarian perpustakaan mengadopsi pembobotan TF-IDF yang dikombinasikan dengan metode *cosine similarity*. Metode ini dipilih karena terbukti memberikan hasil pencarian yang lebih relevan dan akurat berdasarkan pengujian yang telah dilakukan sebelumnya. Diagram arsitektur sistem pencarian perpustakaan JTIK dengan menggunakan metode ini direpresentasikan pada Gambar 2 berikut.



Gambar 2. Diagram Arsitektur Sistem Pencarian Perpustakaan JTIK

Arsitektur sistem ini dimulai dari proses *load model* yang telah dibangun sebelumnya, dan dijalankan oleh mesin pencarian untuk memproses masukan pengguna. *Query* yang diberikan pengguna diolah dan menghasilkan peringkat dokumen sesuai tingkat kemiripan. Dari hasil tersebut, sistem menampilkan informasi lengkap dokumen yang relevan. Pada penelitian ini digunakan *streamlit* sebagai *framework* antarmuka dalam mengimplementasikan arsitektur sistem yang telah dirancang. Gambar 3 di bawah menunjukkan potongan kode program dalam mengimplementasi sistem.

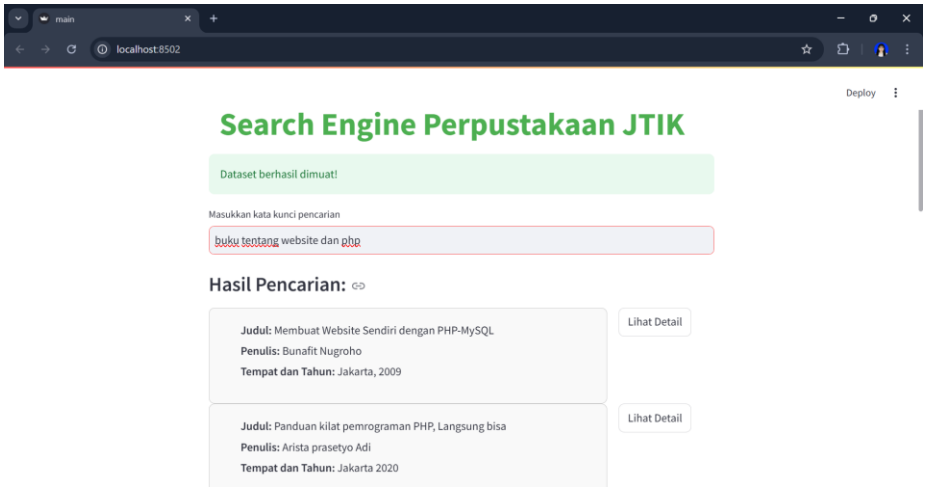
```python
import streamlit as st
import pandas as pd
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.metrics.pairwise import cosine_similarity

st.markdown("<h1 style='text-align: center; color: #4CAF50;'>Search Engine Perpustakaan JTIK</h1>",
unsafe_allow_html=True)

file_path = "Update Dataset STK Preprocessing.xlsx"
try:
    data = pd.read_excel(file_path)
    st.success("Dataset berhasil dimuat!")
except Exception as e:
    st.error(f"Dataset gagal dimuat: {e}")
```

Gambar 3. Potongan Kode Program dalam Mengimplementasi Sistem

Setelah tahap implementasi ditunjukkan melalui potongan kode program Gambar 3 di atas, sistem divisualisasikan dalam antarmuka pengguna. *User Interface* (UI) berfungsi sebagai media utama interaksi pengguna dengan sistem dan memungkinkan pengguna memasukkan pertanyaan pencarian dan secara langsung mendapatkan hasil dokumen yang relevan. Tampilan UI direpresentasikan pada gambar 4 di bawah.

Gambar 4. Implementasi Sistem Pencarian Perpustakaan JTIK

Ketika pengguna memasukkan *query*, sistem secara otomatis memproses permintaan tersebut dan menampilkan 10 buku teratas berdasarkan nilai kemiripan tertinggi. Hasil pencarian ini disusun menggunakan metode *cosine similarity*, sehingga buku yang ditampilkan merupakan yang paling relevan dengan *query* pengguna. Implementasi hasil pencarian ini direpresentasikan pada Gambar 5 di bawah ini.



Gambar 5. Hasil Pencarian Berdasarkan *Query* yang Diberikan

Gambar 6 di bawah menampilkan detail lebih lengkap dari buku yang dipilih oleh pengguna. Informasi yang ditampilkan mencakup judul buku, penulis, tempat dan tahun terbit, serta deskripsi yang relevan untuk membantu pengguna memahami isi buku secara lebih mendalam.

Gambar 6. Tampilan Informasi Detail dari Suatu Buku

Gambar 7 dan Tabel 6 di bawah ini menunjukkan hasil pengujian sistem dengan *query* yang mengandung *typo*, yaitu "artifucial inteligent." Pengujian ini dilakukan untuk mensimulasikan kebiasaan pengguna yang sering melakukan kesalahan ketik saat memasukkan *query* [30]. Berdasarkan hasil pencarian, sistem memberikan rekomendasi buku yang sebagian besar kurang relevan dengan *query* yang dimaksud. Namun demikian, buku yang direkomendasikan pada urutan pertama masih memiliki konteks yang relevan dengan topik *artificial intelligence*, meskipun tingkat akurasi keseluruhan menurun akibat kesalahan pada input *query*.

Gambar 7. Pengujian Sistem Jika *Query* Mengandung *Typo*

Tabel 6 menyajikan rekomendasi buku yang dihasilkan oleh sistem sebagai tanggapan terhadap kueri yang salah eja “artifucial inteligent.” Tabel 6 tersebut menguraikan buku-buku yang diambil beserta skor relevansinya, memberikan wawasan tentang kemampuan sistem untuk memproses kueri yang mengandung kesalahan pengejaan. Hasilnya menunjukkan bahwa meskipun buku dengan peringkat tertinggi tetap relevan dengan kecerdasan buatan, rekomendasi berikutnya menunjukkan penurunan akurasi, dengan beberapa entri hanya terkait sebagian atau sama sekali tidak terkait dengan topik yang dimaksud.

TABEL 6
HASIL PENCARIAN JIKA QUERY MENGANDUNG TYPO

Hasil Pencarian
Aplikasi Android Game Pembelajaran AUGMENTED REALITY BERBASIS UNITY
Integrasi Teknologi Informasi Dengan Strategi
Perancangan Tata Kelola Teknologi Informasi
Pengantar Teknologi Informasi Untuk Bisnis
Teknologi Informasi Pendidikan
Pengantar Teknologi Informasi Edisi Revisi
Membuat Website Gratis Tanpa Guru
Yuk Berbisnis dengan Laravel dan Android
Membuat Website Sendiri dengan PHP-MySQL
Trik Mudah Membuat POS dengan Excel

Secara keseluruhan, sistem ini telah berhasil diimplementasikan untuk mendukung pencarian dan rekomendasi buku. Namun, hasil evaluasi menunjukkan bahwa kesalahan tipografi dalam kueri dapat mengurangi akurasi pencarian. Perbaikan di masa depan harus berfokus pada peningkatan teknik *preprocessing* kueri dan pengoptimalan algoritma pemeringkatan untuk memastikan hasil pencarian yang lebih akurat dan relevan.

F. Pengujian Sistem

Setelah sistem pencarian perpustakaan diimplementasikan, tahap pengujian sistem dilakukan dengan melibatkan 36 responden. Pengujian ini bertujuan untuk mengevaluasi kinerja sistem dalam hal relevansi hasil pencarian, kecepatan respons, dan detail informasi yang disediakan. Responden diminta untuk memberikan penilaian pada berbagai aspek sistem menggunakan skala 1 hingga 5. Hasil user testing direpresentasikan pada Tabel 7 di bawah ini.

TABEL 7
HASIL USER TESTING

No	Pernyataan	SS	S	N	TS	STS
1	Hasil pencarian yang ditampilkan relevan dengan kata kunci yang dimasukkan	14	17	5	-	-
2	*Search engine* ini mampu menampilkan hasil pencarian yang sesuai dengan kebutuhan saya dalam waktu yang cepat	15	15	4	-	-
3	Saya merasa puas dengan urutan hasil pencarian berdasarkan peringkat relevansi	14	15	7	-	-
4	Deskripsi singkat yang muncul pada setiap hasil pencarian cukup informatif	14	16	5	1	-
5	Saya secara keseluruhan merasa puas dengan performa *search engine* ini	20	12	3	1	-

Keterangan:
SS = Sangat Setuju
S = Setuju
N = Netral
TS = Tidak Setuju
STS = Sangat Tidak Setuju

Berdasarkan hasil *user testing* yang melibatkan 36 responden, evaluasi terhadap performa sistem pencarian menunjukkan hasil yang memuaskan. Sistem mendapatkan skor rata-rata sebagai berikut: 4,25 untuk relevansi hasil pencarian, 4,36 untuk kecepatan respons, 4,19 untuk kepuasan terhadap urutan relevansi buku, 4,19 untuk detail informasi buku, dan 4,41 untuk penilaian keseluruhan sistem. Selain evaluasi melalui *user testing*, dilakukan pula *black box testing* untuk memastikan bahwa setiap fungsi utama pada sistem berjalan sesuai dengan yang diharapkan dengan melibatkan 50 responden. Hasil pengujian dapat dilihat pada Tabel 8 berikut.

TABEL 8
HASIL BLACK BOX TESTING

No	Pernyataan	SS	S	N	TS	STS
1	Sistem mampu menerima input dari pengguna berupa kata kunci/topik buku yang sedang dicari.	37	13	-	-	-
2	Sistem menampilkan hasil pencarian/*retrieve* informasi yang relevan dengan input pengguna.	38	12	-	-	-
3	Sistem mampu menampilkan detail informasi buku (judul, penulis, tempat, tahun, dan deskripsi) ketika pengguna memilih opsi Lihat Detail.	41	8	1	-	-
4	Sistem mampu kembali ke halaman hasil pencarian ketika pengguna memilih opsi Kembali ke Hasil Pencarian.	40	9	1	-	-

Hasil pengujian *black box testing* menunjukkan bahwa system berfungsi dengan baik sesuai dengan kebutuhan pengguna. Terlihat pada Tabel 8 di atas, sebagian besar responden memberikan penilaian Sangat Setuju (SS) pada setiap pernyataan yang diuji, dengan nilai 4,74 untuk kemampuan menerima input kata kunci/topik buku, 4,76 untuk relevansi hasil pencarian, 4,80 untuk detail informasi buku, dan 4,78 untuk navigasi kembali ke hasil pencarian. Secara keseluruhan, sistem memperoleh skor rata-rata 4,77. Nilai rata-rata ini diperoleh dengan menghitung total jawaban untuk setiap pernyataan, kemudian membaginya dengan jumlah responden yang terlibat. Hasil evaluasi ini menunjukkan bahwa sistem mampu memberikan pengalaman pengguna yang baik serta memenuhi kebutuhan pencarian informasi secara efektif dan efisien.

IV. SIMPULAN

Berdasarkan penelitian yang dilakukan, dapat disimpulkan bahwa sistem pencarian perpustakaan yang mengimplementasikan metode pembobotan TF-IDF yang dikombinasikan dengan pengukuran *similarity* berhasil meningkatkan relevansi hasil pencarian. Berdasarkan hasil evaluasi menunjukkan bahwa, meskipun *Cosine Similarity* menunjukkan nilai kemiripan rata-rata yang lebih tinggi, *Weighted Jaccard Similarity* menunjukkan nilai Precision@K rata-rata yang lebih tinggi. Terdapaat perbedaan kinerja yang signifikan di antara kedua metode dikonfirmasi oleh uji statistik paired t-test. Selain itu, temuan bahwa metode ini mampu meningkatkan relevansi hasil pencarian dan efisiensi akses informasi diperkuat oleh uji kepuasan pengguna dengan rata-rata secara keseluruhan 4.28 dari 36 responden dan performa sistem dengan rata-rata 4.77 dari 50 responden. Dengan demikian, penelitian ini menegaskan bahwa penggunaan metode pengukuran kemiripan dan pembobotan istilah dapat meningkatkan kualitas sistem pencarian untuk memenuhi kebutuhan akademik mahasiswa. Dengan demikian, sistem yang dikembangkan dapat memberikan pengalaman pencarian yang lebih efisien dan efektif bagi pengguna perpustakaan JTIK. Untuk pengembangan di masa depan, sistem ini dapat diintegrasikan dengan algoritma *Natural Language Processing* (NLP) untuk memahami konteks kueri dengan lebih baik dan menerapkan metode evaluasi tambahan seperti MAP, nDCG, dan MRR untuk memperluas analisis performa sistem. Agar hasil penelitian lebih dapat digeneralisasikan, pengujian harus dilakukan dengan dataset yang lebih besar dan beragam.

DAFTAR PUSTAKA

[1] Indonesia, "Undang-Undang Republik Indonesia Nomor 43 Tahun 2007 tentang Perpustakaan," 43, 2007

[2] R. Aditomo Mahardika Putra, D. Pratiwi, G. Pramita, and F. Dewantoro, "Implementasi Perpustakaan Digital Di SMK Negeri 1 Trimurjo, Kabupaten Lampung Tengah," *JEIT-CS*, vol. 1, no. 3, pp. 180–186, 2023, doi: 10.33365/jeit-cs.v1i3.230.

[3] A. Suhaimah, A. Triayudi, and E. T. E. Handayani, "Cyber Library: Pengembangan Perpustakaan Online Berbasis Web Menggunakan Metode Prototyping (Studi Kasus Universitas Nasional)," *Jurnal JTIK*, vol. 5, 2021.

[4] A. P. Arum and Y. Marfianti, "Pengembangan Perpustakaan Digital untuk Mempermudah Akses Informasi," *SIJALU*, vol. 2, no. 2, pp. 92–100, 2021, doi: 10.26623/jisl.

[5] B. Pratala, "Peningkatan Layanan Perpustakaan IPDN Kampus Jakarta Melalui Sistem Perpustakaan Digital," *CENDEKIA : Jurnal Ilmu Pengetahuan*, vol. 2, no. 1, 2022.

[6] K. P. Sari, A. Masruri, and D. R. Rosalia, "Optimalisasi Temu Kembali Informasi Dengan Teknologi Kecerdasan Buatan di Perpustakaan," *JIPI (Jurnal Ilmu Perpustakaan dan Informasi)*, vol. 8, no. 2, p. 349, Nov. 2023, doi: 10.30829/jipi.v8i2.17775.

[7] F. Cao, J. Zhang, X. Zha, K. Liu, and H. Yang, "A comparative analysis on digital libraries and academic search engines from the dual-route perspective," *The Electronic Library*, vol. 39, pp. 354–372, 2021.

[8] C. Intelligence and Neuroscience, "Retracted: Application of Digital Information Technology in Book Classification and Quick Search in University Libraries," *Comput Intell Neurosci*, vol. 2023, no. 1, Jan. 2023, doi: 10.1155/2023/9892352.

[9] T. K. Wulandari, E. D. Oktaviani, and A. Lestari, "Penerapan Metode Binary Search dan Hamming Distance pada E-library SMAN 2 Katingan Hilir," *KONSTELASI: Konvergensi Teknologi dan Sistem Informasi*, vol. 2, no. 1, 2022.

[10] S. A. Savittri, A. Amalia, and M. A. Budiman, "A relevant document search system model using word2vec approaches," in *Journal of Physics: Conference Series*, IOP Publishing Ltd, Jun. 2021. doi: 10.1088/1742-6596/1898/1/012008.

[11] R. Ramadhan, "Pengelolaan Perpustakaan Digital di Badan Perpustakaan dan Kearsipan Daerah Provinsi Jawa Barat," *Jurnal Pustaka Budaya*, vol. 10, no. 1, pp. 21–31, 2023, [Online]. Available: https://journal.unilak.ac.id/index.php/pb/

[12] A. Aliwy, A. Abbas, and A. Alkhayyat, "NERWS: Towards Improving Information Retrieval of Digital Library Management System Using Named Entity Recognition and Word Sense," *Big Data and Cognitive Computing*, vol. 5, no. 4, Dec. 2021, doi: 10.3390/bdcc5040059.

[13] F. Galatolo, G. Martino, M. Cimino, and C. Tommasi, "Dense Information Retrieval on a Latin Digital Library via LaBSE and LatinBERT Embeddings," INSTICC, Jul. 2023, pp. 518–523. doi: 10.5220/0012134700003541.

[14] K. Ali, "Digital Information Literacy Skills among Library and Information Science Professionals in University Libraries of Sindh Pakistan," *JIMP*, vol. 2, pp. 41–61, 2022.

[15] D. Soyusiawaty, D. Hilmawan, and R. Wolley, "Hybrid Spelling Correction and Query Expansion for Relevance Document Searching," *IJACSA) International Journal of Advanced Computer Science and Applications*, vol. 12, no. 8, p. 2021, 2021, [Online]. Available: www.ijacsa.thesai.org

[16] B. Tang and B. Hu, "Design of Digital Library Data Search Engine Based on Cloud Computing in Big Data Era," in *Journal of Physics: Conference Series*, IOP Publishing Ltd, Oct. 2021. doi: 10.1088/1742-6596/2037/1/012137.

[17] P. Meesad, "Thai Fake News Detection Based on Information Retrieval, Natural Language Processing and Machine Learning," *SN Comput Sci*, vol. 2, no. 6, Nov. 2021, doi: 10.1007/s42979-021-00775-6.

[18] A. Esteva *et al.*, "COVID-19 Information Retrieval with Deep-Learning Based Semantic Search, Question Answering, and Abstractive Summarization," *NPJ Digit Med*, vol. 4, no. 1, Dec. 2021, doi: 10.1038/s41746-021-00437-0.

[19] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. Cambridge: Cambridge University Press, 2008.

[20] O. I. Gifari, Muh. Adha, F. Freddy, and F. F. S. Durrand, "Analisis Sentimen Review Film Menggunakan TF-IDF dan Support Vector Machine," *Journal of Information Technology*, vol. 2, no. 1, pp. 36–40, Mar. 2022, doi: 10.46229/jifotech.v2i1.330.

[21] L. Xiang, "Application of an Improved TF-IDF Method in Literary Text Classification," *Advances in Multimedia*, vol. 2022, pp. 1–10, May 2022, doi: 10.1155/2022/9285324.

[22] M. T. Mohammed and O. F. Rashid, "Document Retrieval Using Term Term Frequency Inverse Sentence Frequency Weighting Scheme," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 31, no. 3, p. 1478, Sep. 2023, doi: 10.11591/ijeecs.v31.i3.pp1478-1485.

[23] Nuzul Hikmah, Dyah Ariyanti, and Ferry Agus Pratama, "Implementasi Chatbot Sebagai Virtual Assistant di Universitas Panca Marga Probolinggo menggunakan Metode TF-IDF," *JTIM : Jurnal Teknologi Informasi dan Multimedia*, vol. 4, no. 2, pp. 133–148, Aug. 2022, doi: 10.35746/jtim.v4i2.225.
[24] X. Li and P. Li, "Rejection Sampling for Weighted Jaccard Similarity Revisited," 2021. [Online]. Available: www.aaai.org
[25] J. Zhu, B. G. Patra, H. Wu, and A. Yaseen, "a Novel NIH Research Grant Recommender Using BERT," *PLoS One*, vol. 18, no. 1, p. e0278636, Jan. 2023, doi: 10.1371/journal.pone.0278636.
[26] R. Wati, S. Ernawati, and H. Rachmi, "Pembobotan TF-IDF Menggunakan Naïve Bayes pada Sentimen Masyarakat Mengenai Isu Kenaikan BIPIH," *Jurnal Manajemen Informatika (JAMIKA)*, vol. 13, no. 1, pp. 84–93, Apr. 2023, doi: 10.34010/jamika.v13i1.9424.
[27] A. Islam, E. Rahman, A. A. Chowdhury, and Md. A. N. Mojumder, "A Deep Learning Approach to Detect Plagiarism in Bengali Textual Content using Similarity Algorithms," in *2023 IEEE International Conference on Contemporary Computing and Communications (InC4)*, IEEE, Apr. 2023, pp. 1–5. doi: 10.1109/InC457730.2023.10262998.
[28] M. Alobed, A. M. M. Altrad, and Z. B. A. Bakar, "a Comparative Analysis of Euclidean, Jaccard and Cosine Similarity Measure and Arabic Wordnet for Automated Arabic Essay Scoring," in *2021 Fifth International Conference on Information Retrieval and Knowledge Management (CAMP)*, IEEE, Jun. 2021, pp. 70–74. doi: 10.1109/CAMP51653.2021.9498119.
[29] Z. Mundher, W. Khater, and L. Ganeem, "Adopting Text Similarity Methods and Cloud Computing to Build a College Chatbot Model," *JOURNAL OF EDUCATION AND SCIENCE*, vol. 30, no. 1, pp. 117–125, Mar. 2021, doi: 10.33899/edusj.2020.127244.1079.
[30] D. Soyusiawaty and D. H. R. Wolley, "Hybrid Spelling Correction and Query Expansion for Relevance Document Searching," *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 8, 2021, doi: 10.14569/IJACSA.2021.0120838.