

SISTEM DETEKSI PLAGIASI SEDERHANA BERBASIS TEXT SIMILARITY: PERBANDINGAN COSINE SIMILARITY, JACCARD SIMILARITY, DAN RABIN-KARP

Febrianta Immanuel Purba^{1*}, Khairida Octavia Ramadhani², M.Ridho Ananda Aditya³, Hafizh Ariiq⁴, Syairal Fahmy Dalimunthe⁵

^{1,2,3,4} Prodi Ilmu Komputer, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Negeri Medan; Jl. William Iskandar Ps. V, Kenangan Baru, Kec. Percut Sei Tuan, Kabupaten Deli Serdang, Sumatera Utara 20221; 061.6613365

Keywords:

Cosine Similarity; Jaccard Similarity; Rabin-Karp; Deteksi Plagiasi; Text Preprocessing; Kemiripan Teks.

Correspondent Email:

antapurba455@gmail.com

Abstrak. Plagiarisme merupakan tantangan serius dalam dunia akademik yang mengancam integritas ilmiah. Sistem deteksi komersial seperti Turnitin memiliki keterbatasan akses akibat biaya tinggi, sedangkan alat gratis sering membatasi jumlah kata dan menimbulkan masalah privasi. Penelitian ini merancang sistem deteksi plagiasi sederhana berbasis web yang mengimplementasikan tiga algoritma secara bersamaan, yaitu Cosine Similarity, Jaccard Similarity, dan Rabin-Karp, dengan dukungan tahapan text preprocessing berbahasa Indonesia menggunakan pustaka Sastrawi. Pengujian pada 11 pasang dokumen menunjukkan bahwa Cosine Similarity menghasilkan rata-rata skor tertinggi (63,5%), diikuti Jaccard Similarity (23,1%) dan Rabin-Karp (1,7%). Sistem ini berjalan secara offline, gratis, dan dapat digunakan sebagai alternatif deteksi kemiripan teks bagi mahasiswa maupun peneliti pemula.



Copyright © 2026 The Author(s), Published by Jurnal Informatika dan Teknik Elektro Terapan. This is an open-access article distributed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0).

Abstract. Plagiarism is a serious challenge in academia that threatens scientific integrity. Commercial detection systems such as Turnitin have limited access due to high costs, while free tools often limit word counts and raise privacy concerns. This study designs a simple web-based plagiarism detection system that implements three algorithms simultaneously: Cosine Similarity, Jaccard Similarity, and Rabin-Karp, supported by Indonesian text preprocessing using the Sastrawi library. Testing on 11 document pairs showed that Cosine Similarity produced the highest average score (63.5%), followed by Jaccard Similarity (23.1%) and Rabin-Karp (1.7%). This system runs offline, is free, and can be used as an alternative text similarity detection for students and novice researchers.

1. PENDAHULUAN

Plagiarisme merupakan tindakan mengambil ide, hasil penelitian, atau tulisan orang lain tanpa menyebutkan sumbernya [1]. Fenomena ini menjadi tantangan serius dalam dunia akademik karena mengancam integritas ilmiah dan kepercayaan terhadap karya ilmiah. Dengan pesatnya perkembangan teknologi informasi, akses terhadap dokumen digital dan jurnal

online semakin mudah sehingga potensi terjadinya plagiarisme pun meningkat secara signifikan [2].

Sistem deteksi plagiarisme komersial seperti Turnitin memiliki keterbatasan akses karena biaya yang tinggi, sementara alat gratis sering kali membatasi jumlah kata yang dapat dianalisis dan menimbulkan masalah privasi data pengguna [3]. Kondisi ini mendorong

kebutuhan akan sistem deteksi kemiripan teks yang gratis, mudah diakses, dan dapat berjalan secara offline untuk mendukung integritas akademik, terutama bagi mahasiswa dan peneliti pemula.

Salah satu pendekatan yang efektif untuk mendeteksi kemiripan teks adalah dengan menggunakan metode pengukuran kemiripan (similarity measurement) dalam pemrosesan bahasa alami (Natural Language Processing/NLP). Tiga metode yang umum digunakan dalam deteksi kemiripan teks adalah Cosine Similarity, Jaccard Similarity, dan Rabin-Karp. Cosine Similarity mengukur kemiripan berdasarkan sudut antara dua vektor yang merepresentasikan frekuensi kata dalam dokumen, sedangkan Jaccard Similarity bekerja berbasis himpunan dengan menghitung rasio irisan terhadap gabungan kata unik dari dua teks. Adapun Rabin-Karp menggunakan pendekatan n-gram untuk mendeteksi kecocokan substring antar dokumen.

Beberapa penelitian sebelumnya telah mengkaji penerapan kedua metode tersebut pada berbagai domain teks [7][8][9]. Baihaqi & Jananto [15] membandingkan Cosine dan Jaccard Similarity pada dokumen RKPD dan menghasilkan skor Cosine sebesar 58,81% dan Jaccard sebesar 28,57%. Ristias dkk. [13] menemukan bahwa Cosine Similarity dengan pembobotan TF-PDF memiliki kinerja terbaik dalam sistem rekomendasi berbasis konten. Arkan dkk. [12] merancang aplikasi deteksi plagiarisme menggunakan TF-IDF dan Cosine Similarity dengan hasil yang menjanjikan. Namun, masih sedikit penelitian yang mengimplementasikan kedua metode secara bersamaan dalam satu sistem deteksi plagiasi sederhana yang dapat langsung digunakan dan dibandingkan hasilnya secara real-time.

Berdasarkan uraian tersebut, Penelitian ini bertujuan untuk membangun sistem deteksi plagiasi sederhana yang membandingkan kemiripan dua teks menggunakan algoritma Cosine Similarity, Jaccard Similarity, dan Rabin-Karp dengan dukungan tahapan text preprocessing berbahasa Indonesia. Dengan mengintegrasikan ketiga algoritma dalam satu sistem, pengguna dapat melihat perbandingan langsung antara ketiga metode sehingga dapat menentukan algoritma yang paling sesuai dan efektif untuk kebutuhan deteksi plagiasi.

Rumusan masalah dalam penelitian ini adalah: (1) Bagaimana mengimplementasikan algoritma Cosine Similarity, Jaccard Similarity, dan Rabin-Karp dalam sistem deteksi kemiripan dua teks berbahasa Indonesia? (2) Bagaimana perbandingan skor kemiripan antara ketiga algoritma pada berbagai jenis teks berbahasa Indonesia? (3) Algoritma manakah yang paling efektif digunakan sebagai dasar deteksi plagiasi pada teks berbahasa Indonesia?

2. TINJAUAN PUSTAKA

2.1 Plagiarisme dan Deteksi Kemiripan Teks

Plagiarisme adalah tindakan mengambil karya orang lain dan mengakuinya sebagai karya sendiri, baik sebagian maupun seluruhnya, tanpa menyebutkan sumber aslinya [5]. Deteksi plagiarisme secara manual memerlukan waktu yang lama dan tidak efisien, terutama untuk jumlah dokumen yang besar, sehingga diperlukan sistem otomatis berbasis komputer [2]. Sistem deteksi kemiripan teks umumnya mengadopsi pendekatan NLP dengan tahapan preprocessing teks diikuti dengan perhitungan kemiripan antar dokumen [3].

Penelitian mengenai deteksi kemiripan teks telah berkembang pesat. Salman [2] mengembangkan sistem pendeteksi kemiripan judul skripsi berbasis web menggunakan NLP dan word embeddings. Putra dkk. [3] membandingkan metode Cosine, Jaccard, dan Rabin-Karp untuk pengukuran kemiripan judul tugas akhir. Hasibuan & Simangunsong [5] menerapkan Cosine Similarity untuk deteksi plagiarisme artikel ilmiah dan memperoleh nilai recall 13% serta precision 8%.

2.2 Text Preprocessing

Text preprocessing merupakan tahap penting dalam pengolahan teks yang bertujuan mengubah teks mentah menjadi format terstruktur yang dapat diproses oleh algoritma [4]. Hartono dkk. [6] menyatakan bahwa preprocessing yang baik secara signifikan meningkatkan akurasi pengukuran kemiripan. Tahapan preprocessing yang digunakan dalam penelitian ini meliputi:

a) Case Folding: Mengubah seluruh huruf dalam teks menjadi huruf kecil untuk menghindari perbedaan pengenalan akibat

perbedaan kapital [5]. Contoh: "Data SCIENCE" menjadi "data science".

b) Tokenizing: Memecah kalimat menjadi unit kata individual (token) [3]. Contoh: "saya belajar" menjadi ["saya", "belajar"].

c) Stopword Removal: Menghapus kata-kata umum yang tidak memiliki makna signifikan seperti "dan", "di", "yang", "ke", dan "dari" [8].

d) Stemming: Mengubah kata berimbuhan menjadi kata dasar menggunakan algoritma Sastrawi yang dirancang khusus untuk Bahasa Indonesia [7]. Contoh: "memerintah" menjadi "perintah".

2.3 Cosine Similarity

Cosine Similarity adalah metode pengukuran kemiripan antara dua vektor dengan menghitung kosinus sudut di antara keduanya dalam ruang vektor multidimensi [9]. Nilai Cosine Similarity berkisar antara 0 (tidak mirip sama sekali) hingga 1 (identik). Metode ini umum digunakan dalam information retrieval dan pemrosesan teks karena insensitif terhadap panjang dokumen. Rumus Cosine Similarity adalah:

$\text{Cosine Similarity}(A,B) = \frac{A \cdot B}{\|A\| \times \|B\|}$
Tanuwijaya dkk.[11] mengimplementasikan TF-IDF dan Cosine Similarity untuk penyaringan dokumen berita dan menunjukkan efektivitas kombinasi kedua metode tersebut. Rismayani dkk. [7] menerapkan Cosine Similarity dalam sistem aspirasi publik berbasis mobile dengan hasil yang memuaskan. Utami & Putra [10] menggunakan Cosine Similarity dalam text mining clustering untuk pengelompokan topik dokumen penelitian.

2.4 Jaccard Similarity

Jaccard Similarity mengukur kemiripan antara dua himpunan berdasarkan rasio antara irisan (intersection) dan gabungan (union) dari kedua himpunan tersebut [4]. Tidak seperti Cosine Similarity, Jaccard Similarity tidak mempertimbangkan frekuensi kemunculan kata sehingga lebih sederhana namun kurang representatif untuk teks yang panjang [12]. Rumus Jaccard Similarity adalah:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

Imanuel dkk. [4] melakukan analisis komparatif antara Jaccard dan Cosine Similarity untuk pencarian ayat Alkitab dan menemukan bahwa kedua metode memiliki kelebihan masing-masing tergantung konteks penggunaan.

Baihaqi & Jananto [15] membandingkan kedua metode pada dokumen RKPD dan menemukan Cosine menghasilkan skor lebih tinggi, yaitu 58,81% dibandingkan Jaccard sebesar 28,57%.

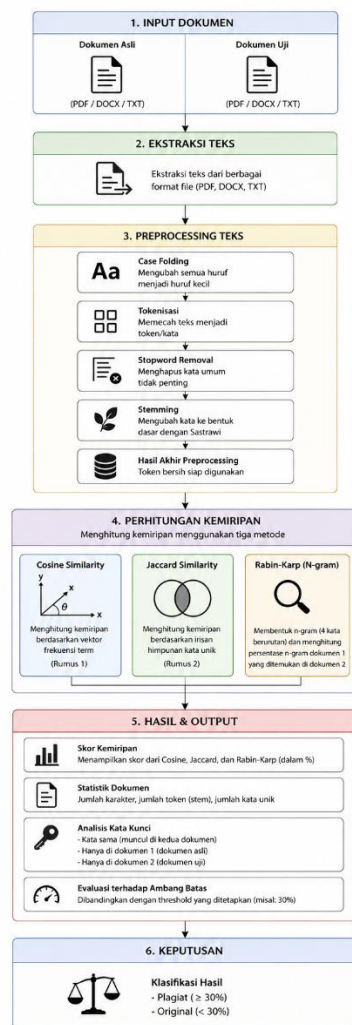
2.5 Penelitian Terdahulu

Supiyanto & Sriyono [9] mendeteksi kemiripan dokumen teks menggunakan Cosine Similarity dan menunjukkan bahwa metode ini efektif untuk berbagai jenis dokumen. Arkan dkk. [12] merancang aplikasi deteksi plagiarisme menggunakan TF-IDF dan Cosine Similarity dengan antarmuka yang mudah digunakan. Mustai'in dkk.[1] menganalisis optimasi tahap pra-pemrosesan pada Cosine Similarity dan menemukan bahwa stemming meningkatkan akurasi kemiripan secara signifikan. Ristias dkk. [13] membandingkan kinerja Cosine dan Jaccard Similarity dalam content-based recommendation system dan menemukan bahwa Cosine Similarity dengan pembobotan TF-PDF memberikan kinerja terbaik. Pratiwi [14] memanfaatkan NLP untuk analisis sentimen yang menunjukkan pentingnya text preprocessing dalam meningkatkan kualitas analisis teks. Penelitian ini berbeda dari penelitian sebelumnya karena mengintegrasikan kedua metode secara bersamaan dalam satu sistem sederhana yang dapat berjalan offline untuk teks berbahasa Indonesia.

3. METODE PENELITIAN

Penelitian ini menerapkan metode eksperimental dengan membangun sebuah sistem deteksi plagiasi sederhana berbasis web yang membandingkan kemiripan dua teks menggunakan tiga algoritma: *Cosine Similarity*, *Jaccard Similarity*, dan *Rabin-Karp*. Sistem dikembangkan dengan pendekatan *Natural Language Processing* (NLP) melalui tahap *preprocessing* teks dan ekstraksi fitur [1][2].

Gambar 1 memperlihatkan arsitektur sistem secara keseluruhan.



Gambar 1. Arsitektur Sistem secara Keseluruhan

3.1 Alur Proses Sistem

Sistem dibangun menggunakan *framework* Flask (Python) dengan antarmuka web (HTML/CSS/JS) [3][5]. Pengguna memasukkan dua dokumen melalui unggahan file (PDF, DOCX, TXT) atau teks manual. Tahapan proses:

1. Ekstraksi teks dari file menggunakan *library* PyPDF, python-docx, dan *built-in* file handler.
2. Preprocessing teks meliputi:
 - o *Case folding* (mengubah ke huruf kecil)
 - o Penghapusan karakter non-alfabet
 - o Tokenisasi (memecah kalimat menjadi kata)
 - o Penghapusan *stopword* (kata umum bahasa Indonesia)
 - o *Stemming* dengan Sastrawi

(mengubah kata berimbuhan menjadi bentuk dasar) [4][6][7].

3. Pembentukan fitur:
 - o Cosine Similarity menggunakan vektor frekuensi kata (TF sederhana).
 - o Jaccard Similarity menggunakan himpunan kata unik.
 - o Rabin-Karp menggunakan *n-gram* ($n=4$) dari token hasil *stemming*.
4. Perhitungan kemiripan dengan ketiga algoritma (Persamaan 1–3).
5. Tampilan hasil: skor kemiripan (dalam persen), statistik dokumen, dan analisis kata kunci.

3.2 Implementasi Algoritma

3.2.1 Cosine Similarity

Dua dokumen direpresentasikan sebagai vektor frekuensi term:

$A = (a_1, a_2, \dots, a_n)$ dan $B = (b_1, b_2, \dots, b_n)$

Kemiripan antar dokumen dihitung menggunakan rumus Cosine Similarity [9][15]:

$$\text{Cosine}(A, B) = \frac{\sum_{i=1}^n A_i B_i}{\sqrt{\sum_{i=1}^n A_i^2} \times \sqrt{\sum_{i=1}^n B_i^2}} \times 100\%$$

3.2.2 Jaccard Similarity

Misalkan S_1 dan S_2 adalah himpunan token unik dari dua dokumen. Nilai kemiripan Jaccard Similarity dihitung dengan rumus berikut [4][8]:

$$J(A, B) = \frac{|S_1 \cap S_2|}{|S_1 \cup S_2|} \times 100\%$$

3.2.3 Rabin-Karp

Algoritma Rabin-Karp membentuk *n-gram* berupa 4 kata berurutan dari setiap dokumen, kemudian menghitung persentase *n-gram* pada dokumen pertama yang juga ditemukan pada dokumen kedua [3][12].

3.3 Skenario Pengujian

Pengujian dilakukan terhadap 11 pasang dokumen yang terdiri atas dokumen asli dan dokumen uji. Dokumen diambil dari artikel jurnal ilmiah berbahasa Indonesia dengan topik yang berdekatan, sehingga memiliki kemiripan leksikal dan tematik yang bervariasi. Setiap pasang dokumen dikategorikan ke dalam tiga tingkat modifikasi, yaitu: (1) modifikasi ringan, berupa perubahan kecil pada susunan kalimat tanpa mengubah isi; (2) modifikasi sedang,

berupa parafrase pada sebagian besar kalimat; dan (3) modifikasi berat, berupa penulisan ulang dengan perubahan struktur dan kosakata yang signifikan.

Penentuan nilai ambang batas (threshold) kemiripan sebesar 25% didasarkan pada beberapa pertimbangan. Pertama, hasil eksperimen awal menunjukkan bahwa dokumen yang tidak memiliki hubungan topik menghasilkan rata-rata skor Cosine Similarity di bawah 20%, sehingga nilai 25% dipilih sebagai batas aman untuk meminimalkan false positive. Kedua, merujuk pada penelitian Baihaqi & Jananto [15] yang menggunakan threshold serupa pada dokumen formal berbahasa Indonesia. Ketiga, penggunaan rata-rata dari ketiga algoritma sebagai dasar keputusan akhir bertujuan untuk mengurangi bias yang muncul dari karakteristik masing-masing metode. Dokumen dengan rata-rata skor ketiga algoritma melebihi 25% dikategorikan sebagai terindikasi plagiat, sedangkan di bawah nilai tersebut dikategorikan sebagai dokumen orisinal.

Setiap pasangan dokumen diproses menggunakan ketiga algoritma secara bersamaan. Hasil pengujian yang dikumpulkan meliputi skor kemiripan dari masing-masing algoritma, statistik dokumen (jumlah karakter, jumlah token hasil stemming, serta jumlah kata unik), dan daftar kata kunci yang diklasifikasikan menjadi kata yang sama, kata yang hanya muncul pada dokumen pertama, dan kata yang hanya muncul pada dokumen kedua. Klasifikasi hasil akhir menggunakan rata-rata skor ketiga algoritma dibandingkan terhadap threshold 25% yang telah ditetapkan.

4. HASIL DAN PEMBAHASAN

4.1 Hasil Pengukuran Kemiripan

Dari 11 pasang dokumen yang diuji, diperoleh ringkasan skor kemiripan sebagaimana ditunjukkan pada Tabel 1.

Tabel 1. Rata-rata Skor Kemiripan dari 11 Pasang Dokumen

Metode	Rata-rata Skor (%)	Skor Tertinggi (%)	Skor Terendah (%)
Cosine Similarity	63,5	87,9	34,1
Jaccard Similarity	23,1	32,2	11,8

Rabin-Karp	1,7	7,2	0,0
-------------------	-----	-----	-----

Sumber: Hasil uji coba sistem (2025)

Cosine Similarity secara konsisten menghasilkan skor kemiripan tertinggi. Hal ini karena metode ini mempertimbangkan frekuensi kemunculan kata, sehingga lebih mampu menangkap kemiripan leksikal dan tematik meskipun terdapat variasi redaksi [10][15]. Sebaliknya, Jaccard Similarity hanya menghitung irisan himpunan kata unik tanpa bobot, sehingga kurang sensitif terhadap panjang dokumen dan pengulangan istilah penting [4][8]. Rabin-Karp memberikan nilai sangat rendah karena dirancang untuk pencocokan *substring* eksak, bukan untuk kemiripan teks umum [3][12].

4.2 Analisis Statistik Dokumen

Tabel 2 menyajikan contoh perbandingan statistik pada pasangan dokumen dengan skor Cosine Similarity tertinggi (87,9%) dan terendah (34,1%).

Tabel 2. Statistik Pasangan dengan Skor Tertinggi

Parameter	Dokumen Asli	Dokumen Uji
Karakter	38.613	27.565
Token Stem	3.544	2.483
Kata Unik	812	655
Cosine Similarity (%)	87,9	87,9

Tabel 3. Statistik Pasangan dengan Skor Terendah

Parameter	Dokumen Asli	Dokumen Uji
Karakter	14.218	26.010
Token Stem	1.305	2.329
Kata Unik	404	691
Cosine Similarity (%)	34,1	34,1

Pada pasangan dengan tingkat kemiripan tinggi, jumlah token dan kata unik cenderung lebih seimbang serta memiliki banyak irisan kata kunci. Sebaliknya, perbedaan ukuran dokumen yang besar dan rendahnya jumlah kata unik yang beririsan menyebabkan nilai Cosine Similarity menurun secara signifikan [5][11].

4.3 Analisis Kata Kunci

Sistem menampilkan maksimal 15 kata hasil *stemming* yang dikategorikan menjadi kata yang sama, kata yang hanya muncul pada dokumen asli, serta kata yang hanya muncul pada dokumen uji. Contoh hasil dari pasangan dengan skor Cosine Similarity sebesar 87,9% adalah sebagai berikut:

Kata sama:

fleksibilitas, buka, tulis, ilmiah, mampu, fokus, efektif, pribadi, penting, materi

Hanya pada dokumen asli:

eksplisit, sesuai, bukan, berbe, conclusions, dimensi, modern, kohesi

Hanya pada dokumen uji:

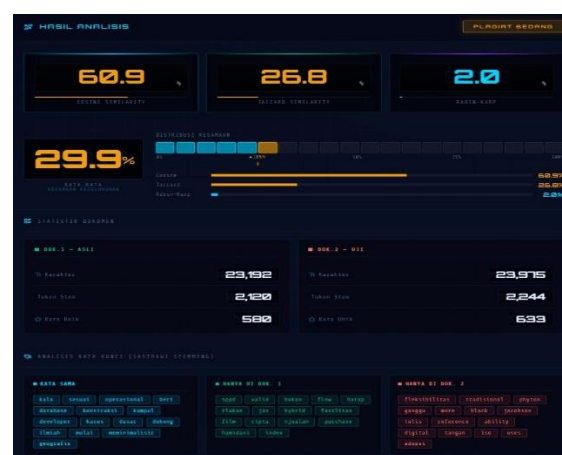
apa, therefore, motivation, experiences, theoretical, ingin, lama

Keberadaan kata-kata yang sama menunjukkan adanya kesamaan topik pembahasan, sedangkan kata unik pada masing-masing dokumen mengindikasikan perbedaan sudut pandang atau penggunaan istilah teknis tambahan. Proses *stemming* menggunakan Sastrawi berhasil menyamakan variasi morfologis kata, misalnya “*memerintah*” menjadi “*perintah*”, sehingga meningkatkan akurasi pengukuran menggunakan Cosine Similarity [6][7].

4.4 Distribusi Kesamaan dan Ambang Batas

Rata-rata kesamaan keseluruhan (rata-rata dari ketiga metode) untuk 11 pasang berkisar antara 11,5% hingga 41,1%, dengan nilai global 28,9%. Sistem menggunakan threshold 25% sebagai batas minimal indikasi plagiat. Karena rata-rata global (28,9%) melebihi 25%, maka secara umum sebagian besar pasangan dokumen dalam uji coba ini dikategorikan sebagai “Plagiat Sedang”, kecuali pasangan dengan rata-rata keseluruhan di bawah 25% (misalnya pasangan 2 yang hanya 11,5%).

Hal ini menunjukkan bahwa threshold 25% lebih sensitif terhadap kemiripan ringan dibandingkan threshold yang lebih tinggi (misalnya 30%), sehingga sistem dengan batas 25% cenderung lebih konservatif dalam mendeteksi potensi plagiarisme sejak dini.



Gambar 2. Contoh Distribusi Kesamaan (Cosine 60,9%; Jaccard 26,8%; Rabin–Karp 2,0%)

Berdasarkan hasil distribusi tersebut, Cosine Similarity merupakan metode yang paling mendekati penilaian manusia karena mempertimbangkan bobot kemunculan kata. Sebaliknya, Jaccard Similarity menghasilkan nilai yang relatif rendah, sedangkan Rabin–Karp kurang memberikan informasi yang memadai untuk kebutuhan deteksi plagiarisi.

4.5 Pembahasan Umum

Hasil penelitian ini sejalan dengan temuan Baihaqi dan Jananto [15] yang menyatakan bahwa Cosine Similarity berbasis TF-IDF lebih representatif (58,81%) dibandingkan Jaccard Similarity (28,57%) pada dokumen formal berdimensi panjang. Temuan ini juga didukung oleh penelitian Ristias dkk. [13] yang menunjukkan keunggulan Cosine Similarity dalam sistem rekomendasi berbasis konten. Adapun kelebihan sistem yang dibangun meliputi:

1. Mampu memproses berbagai format file, seperti PDF, DOCX, dan TXT.
 2. Menyediakan tiga metode analisis secara bersamaan untuk kebutuhan perbandingan langsung.
 3. Memiliki antarmuka web interaktif dengan visualisasi skor digital.
 4. Menggunakan *stemming* bahasa Indonesia berbasis Sastrawi yang efektif.
- Sementara itu, keterbatasan sistem yang ditemukan adalah sebagai berikut:

1. Cosine Similarity masih menggunakan frekuensi kata sederhana dan belum menerapkan TF-IDF, sehingga kata

penting dengan frekuensi rendah belum dapat dibedakan secara optimal.

2. Rabin–Karp menghasilkan skor yang terlalu rendah sehingga kurang efektif untuk deteksi plagiasi berbasis kemiripan dokumen.
3. Sistem saat ini hanya mendukung perbandingan dua dokumen dan belum terintegrasi dengan basis data multi-dokumen.

5. KESIMPULAN

Berdasarkan hasil implementasi dan pengujian sistem deteksi plagiasi sederhana menggunakan algoritma Cosine Similarity, Jaccard Similarity, dan Rabin–Karp, diperoleh beberapa kesimpulan sebagai berikut.

1. Cosine Similarity merupakan metode yang paling efektif dalam mendeteksi kemiripan teks berbahasa Indonesia pada dokumen dengan panjang yang bervariasi. Dari 11 pasang dokumen uji, metode ini menghasilkan rata-rata skor kemiripan sebesar 63,5%, sehingga layak direkomendasikan sebagai algoritma utama pada sistem deteksi plagiasi sederhana.
2. Jaccard Similarity menghasilkan rata-rata skor sebesar 23,1%. Metode ini cukup sesuai untuk pemeriksaan cepat pada teks pendek, tetapi kurang akurat untuk dokumen yang lebih panjang karena hanya mempertimbangkan irisan kata unik tanpa memperhitungkan frekuensi kemunculan kata.
3. Rabin–Karp menunjukkan rata-rata skor yang sangat rendah, yaitu 1,7%, sehingga kurang sesuai untuk pengukuran kemiripan dokumen secara umum. Hal ini disebabkan algoritma tersebut dirancang untuk pencocokan *substring* secara eksak, bukan untuk analisis kemiripan teks berbasis konteks atau distribusi kata.
4. Sistem yang dibangun berhasil mengintegrasikan tahapan *preprocessing* teks, meliputi *case folding*, tokenisasi, *stopword removal*, dan *stemming* menggunakan Sastrawi, serta mendukung ekstraksi teks dari berbagai format file dan penerapan tiga algoritma

similarity dalam satu antarmuka web yang mudah digunakan.

5. Sistem masih memiliki beberapa keterbatasan, antara lain belum menerapkan pembobotan TF-IDF pada Cosine Similarity, belum mendukung perbandingan terhadap basis data multi-dokumen, serta penggunaan metode Rabin–Karp yang menghasilkan informasi kemiripan yang relatif kurang signifikan.

Berdasarkan hasil penelitian dan keterbatasan sistem yang ditemukan, beberapa saran pengembangan untuk penelitian selanjutnya adalah sebagai berikut.

1. Menerapkan metode pembobotan TF-IDF atau pendekatan berbasis *word embedding* seperti Word2Vec dan BERT untuk meningkatkan kemampuan sistem dalam memahami kemiripan semantik antar dokumen [2][13].
2. Menambahkan fitur perbandingan satu dokumen uji terhadap banyak dokumen dalam suatu basis data sehingga sistem dapat digunakan untuk skenario deteksi plagiasi skala yang lebih besar.
3. Mengevaluasi kembali penggunaan algoritma Rabin–Karp, termasuk mempertimbangkan penggantian atau penghapusannya apabila tidak memberikan kontribusi yang signifikan terhadap akurasi sistem.
4. Melakukan pengujian menggunakan korpus dokumen yang lebih besar, lebih beragam, dan berasal dari berbagai domain untuk memperoleh evaluasi performa yang lebih representatif.
5. Penelitian selanjutnya disarankan untuk melakukan validasi sistem menggunakan ground truth dari penilaian pakar guna mengukur precision, recall, dan F1-score sebagai indikator akurasi yang lebih objektif.

UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih kepada seluruh anggota tim peneliti atas kerja sama dan dedikasi dalam menyelesaikan penelitian ini.

DAFTAR PUSTAKA

- [1] A. D. H. Mustai'in, A. Sanjaya, and A. B. Setiawan, "Penerapan Regular Expression dan Cosine Similarity pada Uji Kemiripan Kalimat Bahasa Indonesia," INOTEK, Vol 9, 2025.
- [2] Salman, "Sistem pendeteksi kemiripan judul skripsi berbasis web menggunakan NLP dan word embeddings," Jurnal Pendidikan Indonesia: Teori, Penelitian dan Inovasi, vol. 5, no. 5, Sep 2025.
- [3] R. Putra, S. Sumema, and Z. Mustika, "Similarity judul tugas akhir menggunakan metoda cosine, jaccard, rabin-karp pada jurusan TI," Jurnal Ilmiah Teknologi Sistem Informasi (JITSI), vol. 5, no. 4, 2024.
- [4] Y. Imanuel, S. J. Pinastika, H. Soepandi, and K. Lebai, "Analisis komparatif jaccard dan cosine similarity untuk pencarian ayat alkitab," SINTA Jurnal Sistem Informasi dan Teknologi Komputasi, vol. 2, no. 4, pp. 171–177, Oct 2025.
- [5] F. Hasibuan and J. Simangunsong, "Analisa metode cosine similarity dalam mendeteksi plagiarisme pada artikel ilmiah," JUTISAL (Jurnal Teknik Informatika Komputer Universal), vol. 3, no. 1, 2023.
- [6] R. Hartono, Siswanto, and I. Kanedi, "Implementasi metode vector space modelling untuk pencarian buku pada perpustakaan," Jurnal Komputer, vol. 2, no. 1, Jul-Des 2023.
- [7] Rismayani, Hasyrif SY, Nirwana, T. Darwansyah, and I. Mansury, "Implementasi algoritma text mining dan cosine similarity untuk desain sistem aspirasi publik berbasis mobile," Komputika: Jurnal Sistem Komputer, vol. 11, no. 2, Okt 2022.
- [8] E. Rahmi, E. Yumami, and N. Hidayasari, "Analisis metode pengembangan sistem informasi berbasis website: systematic literature review," Remik: Riset dan E-Jurnal Manajemen Informatika Komputer, vol. 7, no. 1, Jan 2023.
- [9] Supiyanto and Sriyono, "Metode cosine similarity untuk mendeteksi kemiripan pada dokumen teks," SAINS Jurnal MIPA dan Pengajarannya, vol. 1, no. 1, pp. 1–7, Jan 2023.
- [10] N. W. Utami and I. G. J. E. Putra, "Text mining clustering untuk pengelompokan topik dokumen penelitian menggunakan algoritma k-means dengan cosine similarity," JINTEKS (Jurnal Informatika Teknologi dan Sains), vol. 4, no.3 , Agu 2022.
- [11] W. Tanuwijaya, C. E. Setiawan, H. Irsyad, and A. Rahman, "Implementasi TF-IDF dan cosine similarity untuk penyaringan dokumen berita program makan siang gratis pemerintah Indonesia," Device: Journal Of Information System, Computer Science And Information Technology, vol. 6, no. 2, pp. 322–334, Des 2025.
- [12] S. Arkan, H. Sulistiono, and I. Setiadi, "Perancangan aplikasi deteksi plagiarisme dengan TF-IDF dan cosine similarity," Jurnal Riset dan Aplikasi Mahasiswa Informatika (JRAMI), vol. 7, no. 2, 2026.
- [13] A. A. Ristias, E. D. Wahyuni, and S. F. A. Wati, "Komparasi kinerja metode cosine dan jaccard similarity dalam content-based recommendation systems (CBRS) pada aplikasi eventhings," JITET (Jurnal Informatika dan Teknik Elektro Terapan), vol. 12, no. 3, 2024.
- [14] F. Pratiwi, "Pemanfaatan natural language Processing (NLP) untuk analisis sentimen dalam evaluasi risiko kredit," Journal of Accounting And Financial Issue, vol. 6, no. 2, Okt 2025.
- [15] M. H. Al Baihaqi and A. Jananto, "Perbandingan metode cosine dan jaccard similarity dalam pengukuran tingkat kemiripan dokumen rencana kerja pemerintah daerah," (JATI) Jurnal Mahasiswa Teknik Informatika, Vol. 10, no.2, Apr 2026.