E-ISSN 2797-9512
P-ISSN 2797-9504

ANALISA METODE *COSINE SIMILARITY* DALAM MENDETEKSI PLAGIARISME PADA ARTIKEL ILMIAH

[1]Fatimah Zuhra Hasibuan, [2]Juanto Simangunsong
[1,2]Akademi Manajemen Informatika dan Komputer Universal
e-mail : juantosmg@gmail.com

ABSTRAK

Plagiarisme merupakan tindakan mengambil ide, hasil penelitian, atau rangkuman dari suatu tulisan tanpa menyebutkan sumbernya. Salah satu metode yang digunakan untuk mengukur tingkat kesamaan antara artikel adalah metode cosine similarity. Sistem melakukan langkah-langkah untuk menghitung nilai kemiripan antara artikel jurnal yang diunggah di repositori, yang diperoleh dari data DOAJ. Setelah dilakukan perhitungan, diperoleh presentase nilai kemiripan antara artikel. Kemudian, dilakukan perhitungan kembali untuk mencari nilai kemiripan antara artikel jurnal dari berbagai penerbit yang ada di repositori. Berdasarkan uji coba, nilai recall pada Aplikasi Deteksi Plagiarisme Menggunakan Metode Cosine Similarity adalah 17%, yang dihitung dengan membagi jumlah artikel relevan yang terambil dengan jumlah keseluruhan artikel dan dikalikan dengan 100%. Sementara itu, untuk mendapatkan nilai precision, skenario pengujian dilakukan dengan membagi jumlah artikel relevan yang terambil dengan jumlah artikel yang relevan dalam pencarian, lalu hasilnya dikalikan dengan 100%, dan diperoleh hasil 6%.

***Kata Kunci*:** *Cosine Similarity*, Artikel Ilmiah, Deteksi, Plagiarisme.

1. PENDAHULUAN

Perkembangan pesat teknologi saat ini telah mengubah gaya hidup manusia menjadi serba digital. Teknologi telah menjadi kebutuhan penting dalam kehidupan sehari-hari manusia di berbagai bidang. Dalam era digital ini, dampak teknologi memiliki sisi positif dan negatif. Salah satu bidang yang terpengaruh oleh perkembangan teknologi adalah pendidikan, termasuk dokumen-dokumen digital.

Dokumen digital menjadi salah satu hasil dari perkembangan teknologi di era digital ini, seperti jurnal online. Jurnal online merupakan dokumen digital yang sangat penting dan dibutuhkan dalam berbagai bidang. Jurnal merupakan bagian dari publikasi yang ada di perpustakaan dan berisi berita serta hasil penelitian tentang berbagai topik. Jurnal dapat hadir dalam dua format, yaitu tercetak dan digital. Jurnal online adalah versi digital dari jurnal cetak yang biasanya ada di perpustakaan. Jurnal online dapat diakses melalui email, situs web, atau akses internet. Sama seperti jurnal cetak, jurnal online juga merupakan terbitan berseri, namun perbedaannya terletak pada bahan baku yang digunakan. Jurnal cetak menggunakan kertas, sementara jurnal online dapat dibaca secara langsung tanpa perlu mencetaknya.

Adanya jurnal online memiliki beberapa keuntungan, seperti kemudahan dalam membacanya di mana saja tanpa perlu membawa kertas. Namun, dampak negatif dari jurnal online adalah munculnya potensi plagiat. Plagiat dapat terjadi dalam berbagai bentuk, dan salah satu yang umum dilakukan adalah menyalin dan memodifikasi artikel

dari jurnal online. Plagiat dapat ditemukan dalam bentuk kutipan yang tidak sesuai pada sebuah dokumen [1].

Mendeteksi plagiat bisa dilakukan secara manual, tetapi tidak efisien karena harus membandingkan satu artikel dengan ribuan artikel lain dan menganalisis gaya penulisannya. Cara yang lebih mudah untuk mendeteksi plagiat adalah menggunakan mesin pencarian dengan memasukkan kata kunci tema artikel, dan mesin pencarian akan menemukan artikel yang mungkin dijiplak [1]. Mesin pencarian adalah program komputer yang membantu pengguna menemukan informasi yang relevan dengan kata kunci yang dimasukkan. Dalam waktu singkat, mesin pencarian akan memberikan hasil berupa artikel yang relevan dengan kata kunci tersebut. Metode ini efektif jika plagiat dilakukan pada seluruh dokumen, tetapi kurang efektif jika plagiat hanya terjadi pada sebagian artikel dan menggabungkan beberapa sumber.

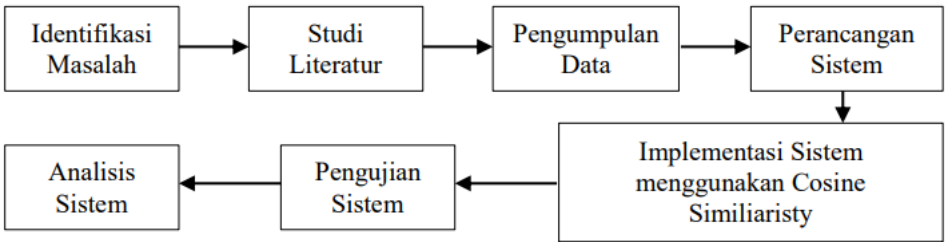
Untuk menjalankan mesin pencarian pendeteksi plagiat, diperlukan cara untuk mengambil halaman web secara otomatis agar diperoleh dokumen jurnal yang terbaru. Web crawler digunakan untuk melakukan hal ini, karena dapat menjelajahi halaman web secara rekursif dan otomatis dengan mengikuti hyperlink yang tersedia dan mengunduh URL untuk mengambil link dari halaman web lain [2].

Sebelum mencari kemiripan, konten jurnal harus difilter menggunakan pdf extractor untuk mengambil semua konten jurnal dalam bentuk PDF, termasuk metadata dan isi artikel. Salah satu metode yang digunakan dalam pencarian dan penilaian tingkat kemiripan adalah metode Cosine Similarity. Metode Cosine Similarity digunakan untuk menghitung tingkat kemiripan antara dua dokumen [3]. Perhitungan metode ini didasarkan pada dua vektor yang mewakili kata-kata dalam dokumen yang dibandingkan. Peneliti memilih metode Cosine Similarity karena memiliki tingkat keakuratan yang lebih tinggi daripada metode Jaccard Similarity, karena metode Cosine Similarity melakukan normalisasi panjang vektor data dan membandingkan N-gram yang sejajar dari dua dokumen yang dibandingkan [4].

2. METODE PENELITIAN

Bagian metodologi penelitian menjelaskan langkah-langkah yang diambil dalam melaksanakan penelitian ini. Penelitian ini berjudul "Aplikasi Deteksi Plagiarisme Menggunakan Metode Cosine Similarity."

a. Prosedur penelitian ini direpresentasikan dalam diagram pada gambar 1.



Gambar 1 Prosedur Penelitian

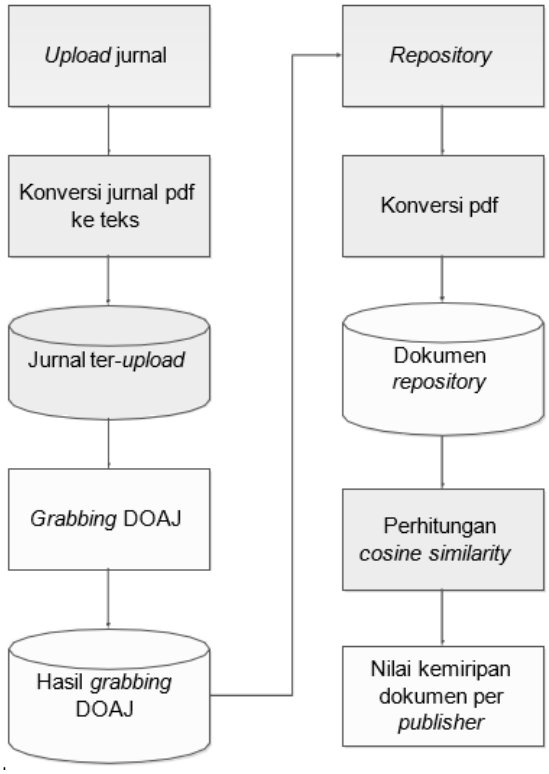
Penelitian ini dimulai dengan mengidentifikasi masalah dan menetapkan pertanyaan penelitian. Tahap selanjutnya adalah studi literatur, di mana teori-teori yang mendukung penelitian dikumpulkan, termasuk proses grabbing dan Cosine Similarity. Langkah berikutnya adalah pengumpulan data, di mana jurnal-jurnal digunakan

sebagai dokumen repository. Tahap perancangan sistem dilakukan untuk memahami alur sistem yang akan dibuat, termasuk implementasi web crawler untuk mengambil konten dari setiap jurnal dan metode cosine similarity untuk menghitung kemiripan teks dalam deteksi plagiarisme. Setelah perhitungan nilai kemiripan dokumen selesai, sistem diuji untuk memastikan keakuratannya dengan metode yang telah digunakan. Setelah semua tahap selesai, dilakukan analisis terhadap hasil perhitungan metode dan sistem yang telah dibangun.

b. Pengumpulan Data
   - Jurnal Online. Objek penelitian menggunakan jurnal online. Tahap awal adalah mengumpulkan atau mencari referensi jurnal online dari berbagai website atau blog.
   - Konversi PDF digunakan untuk mengubah file PDF menjadi teks agar nilai kemiripan dapat dihitung dengan dokumen repository yang ada dalam database.
   - Isi jurnal diambil dari semua isi dokumen jurnal dalam bentuk teks untuk dihitung kemiripannya.
   - Database. Tahap terakhir adalah memasukkan konten yang telah diperoleh ke dalam database yang sudah dibuat. Data dalam database ini akan diproses oleh aplikasi deteksi plagiarisme artikel jurnal.

c. Perancangan Sistem

Pada tahap ini terdapat desain rancangan sistem yang akan memberikan gambaran yang harus dikerjakan serta bagaimana sistem memecahkan masalah deteksi plagiarisme.



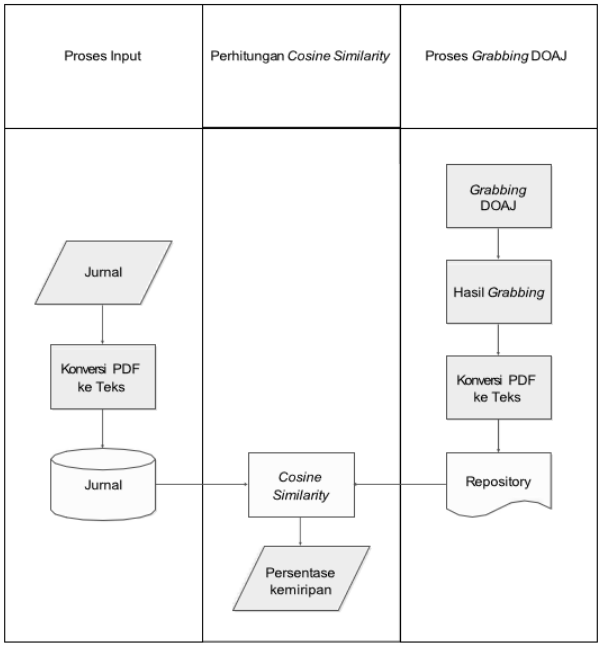
Gambar 2 Perancangan Sistem

Proses pertama adalah mengunggah jurnal. Pada tahap ini, mahasiswa akan mengunggah jurnal dalam format PDF untuk memeriksa kemiripannya dengan

dokumen-dokumen di repositori. Jurnal yang diunggah akan dikonversi menjadi teks dan hasilnya akan dimasukkan ke dalam database. Selanjutnya, untuk mencari dokumen di repositori, proses grabbing akan dilakukan dengan mengambil jurnal-jurnal yang ada di Directory of Open Access Journal (DOAJ). Pengambilan ini dapat disesuaikan dengan tema dan jumlah jurnal yang diinginkan. Seluruh hasil pengambilan dari DOAJ akan dimasukkan ke dalam database dan kemudian dijadikan dokumen repositori setelah dikonversi dari format PDF menjadi teks untuk perhitungan algoritma cosine similarity. Perhitungan algoritma dilakukan untuk membandingkan jurnal yang diunggah dengan jurnal-jurnal yang ada di repositori.

d. Desain Sistem

Desain sistem bertujuan untuk memberikan gambaran yang jelas dan rancang bangun yang lengkap kepada pemakai sistem.



Gambar 3 Desain Sistem

- Proses pertama adalah memasukkan jurnal yang akan dideteksi kemiripannya dengan dokumen-dokumen yang sudah ada di dalam database.
- Selanjutnya, dilakukan konversi dari format PDF ke teks agar nilai kemiripannya dapat dihitung.
- Tahap berikutnya adalah melakukan grabbing pada DOAJ untuk menjelajahi dan mengambil halaman-halaman web yang diperlukan dengan mengikuti hyperlink yang ada di DOAJ.
- Terakhir, dilakukan perhitungan nilai kemiripan untuk mendapatkan persentase kemiripan antara dokumen jurnal dengan dokumen-dokumen di repositori yang telah didapatkan.

3. HASIL DAN PEMBAHASAN

A. Perhitungan

Pada tahap pengujian sistem, beberapa dokumen uji diambil dan kemudian dihitung nilai kemiripannya dengan dokumen pembanding yang telah diperoleh dari hasil

E-ISSN 2797-9512
P-ISSN 2797-9504

grabbing DOAJ. Nilai persentase yang diperoleh menunjukkan tingkat kemiripan antara dokumen-dokumen tersebut yang dihitung menggunakan rumus cosine similarity, yaitu:

$$similarity(d_j, q) = \frac{\sum_{i=l}^{t}(w_{ij} \cdot w_{iq})}{\sqrt{\sum_{i=l}^{t} w_{ij}^2 \cdot \sum_{i=l}^{t} w_{iq}^2}}$$

Perhitungan Recall

Dilakukan untuk menilai akurasi suatu sistem berdasarkan hasil pencarian. Dalam proses pencarian dengan kata kunci "data mining", diperoleh 12 dokumen jurnal yang sesuai dengan nama file yang diinputkan pada proses pencarian dalam tabel dokumen repository. Untuk mengetahui tingkat akurasi sistem, dilakukan perhitungan nilai recall menggunakan rumus berikut ini:

$$\begin{aligned} Recall &= \frac{\sum dokumen\ relevan\ yang\ terambil}{\sum dokumen\ relevan\ dalam\ database} \text{ x } 100\% \\ &= \frac{12}{91} \times 100\% \\ &= 13\% \end{aligned}$$

Perhitungan recall merupakan parameter untuk mengukur tingkat akurasi suatu sistem berdasarkan dokumen relevan yang terdapat dalam databse. Pada sistem ini, diperoleh nilai recall yaitu 13%.

Perhitungan Precision

Dilakukan untuk menilai akurasi suatu sistem berdasarkan jumlah dokumen relevan yang terambil dalam proses pencarian. Pada proses pencarian dengan kata kunci "data mining", diperoleh 12 dokumen jurnal yang relevan. Namun, dari total jurnal dalam pencarian sebanyak 12 file, hanya 1 dokumen jurnal yang sesuai dengan nama file pada proses pencarian "data mining" dalam dokumen repository dengan nilai cosine similarity di atas 0%. Dengan data jumlah dokumen jurnal yang diperoleh dari proses pencarian "data mining" seperti dijelaskan di atas, dilakukan perhitungan nilai precision menggunakan rumus berikut ini:

$$\begin{aligned} Precision &= \frac{\sum dokumen\ relevan\ yang\ terambil}{\sum dokumen\ relevan\ dalam\ pencarian} \text{ x } 100\% \\ &= \frac{1}{12} \times 100\% \\ &= 8\% \end{aligned}$$

Perhitungan precision merupakan parameter untuk mengukur tingkat akurasi sebuah sistem berdasarkan dokumen yang relevan pada pencarian yang dilakukan. Pada sistem ini, diperoleh nilai precision yaitu 8%.

B. Perangkat

Implementasi merupakan tahap penerapan sistem yang akan dilakukan jika sistem disetujui sebagai program yang telah dibuat pada tahap perancangan. Selain itu, implementasi sistem merupakan sebuah proses pembuatan dan penerapan sistem secara utuh baik dari sisi perangkat keras maupun perangkat lunaknya.

C. Implementasi *Interface*

Interface merupakan tampilan yang dapat melakukan interaksi antara pengguna dengan sistem, dimana interface dapat menerima informasi dari pengguna dan memberikan informasi kepada pengguna yang bertujuan untuk menginput pengetahuan baru ke dalam basis pengetahuan sistem pakar, menampilkan penjelasan sistem dan memberikan panduan pemakaian sistem secara menyeluruh sehingga dapat dipahami oleh pengguna.

4. KESIMPULAN

Berdasarkan hasil penelitian dan pembahasan mengenai aplikasi deteksi plagiarisme menggunakan metode cosine similarity, dapat disimpulkan bahwa akurasi sistem dapat diukur melalui perhitungan recall dan precision dari metode cosine similarity dengan membandingkan data yang diambil dengan repository yang ada. Nilai recall dalam kasus ini adalah 13%, yang diperoleh dari jumlah dokumen relevan yang berhasil diambil dibagi dengan jumlah total dokumen yang ada dalam database, kemudian dikalikan dengan 100%. Sedangkan nilai precision adalah 8%, yang diperoleh dari jumlah dokumen relevan yang berhasil diambil dibagi dengan jumlah total dokumen relevan dalam pencarian, kemudian dikalikan dengan 100%.

DAFTAR PUSTAKA

1. Firdaus, Hari Bagus. 2003. “Algoritma Rabin-Karp.” Jurnal Ilmu Komputer dan Teknologi Informasi III No. 2: 1–5.
2. Zuliarso, Eri. 2010. “Aplikasi Web Crawler Berdasarkan Breadth First Search Dan Back-Link.” Fakultas Teknologi Informasi, Universitas Stikubank Semarang XV(1): 52–56.
3. Sugiyamta. 2015. “Sistem Deteksi Kemiripan Dokumen Dengan Algoritma Cosine Similarity Dan Single Pass Clustering.” Dinamika Informatika 7(2): 7.
4. Nurdiana, Ogie, Jumadi, and Dian Nursantika. 2016. “Perbandingan Metode Cosine Similarity Dengan Metode Jaccard Similarity Pada Aplikasi Pencarian Terjemah Al-Qur‟an Dalam Bahasa Indonesia.” Jurnal Online Informatika (JOIN) 1(1): 59–63.
5. Santoso, Hari. 2015. “Pencegahan Dan Penaggulangan Plagiarisme Dalam Penulisan Karya Ilmiah Di Lingkungan Perpustakaan Perguruan Tinggi Oleh : Drs. Hari Santoso, S.Sos. 1.” Perpustakaan UM Malang (1): 1–23.
6. Kumar, Manish, Ankit Bindal, Robin Gautam, and Rajesh Bhatia. 2018. “Keyword Query Based Focused Web Crawler.” Procedia Computer Science 125: 584–90. http://linkinghub.elsevier.com/retrieve/pii/S1877050917328399.
7. Rungsawang, Arnon, and Niran Angkawattanawit. 2005. “Learnable Topic-Specific Web Crawler.” Journal of Network and Computer Applications 28(2): 97–114.
8. Pahlevi, Irfan, Moch Arief Bijaksana, and M Ir Tech. “Perhitungan Kemiripan Dokumen Bahasa Indonesia Menggunakan Metode Cosine Similarity (Studi Kasus : Abstrak Tugas Akhir Fakultas Informatika Universitas Telkom).”