Deteksi Plagiat Tesis Berbahasa Indonesia Menggunakan Metode Cosine Similarity

Syukry Ansis[1], Endang Palupi Listyaningsih[2], Prof. Dr. Ir. Hari Soetanto, S.kom, M.Sc.[3]
[123]Universitas Budi Luhur,
Jl. Ciledug Raya, Kec. Pesanggrahan, Kota Jakarta Selatan, Indonesia
E-mail: 211601064@student.budiluhur.ac.id[1], 211160117@student.budiluhur.ac.id[2], hari.soetanto@budiluhur.ac.id[3]

Abstrack - Plagiarism detection plays a crucial role in maintaining academic integrity and intellectual honesty in higher education. One of the ways used to detect plagiarism is by using Cosine Similarity. Cosine Similarity works by comparing a document with other documents based on the number of keywords, bag of words and the frequency of occurrence of certain keyword words. However, the application of Cosine Similar ity is rarely analyzed for accuracy and performance. This research aims to test the performance of the Cosine Similar ity method and make a comparison with the Jaccard Similarity method. Data samples are obtained from Indonesian-language thesis data belonging to Budi Luhur campus students. The test model will be tested by comparing several original theses and documents containing plagiarism. The findings of this study show that Cosine Similarity has high accuracy in classifying documents as plagiarism or not, with an accuracy rate of 96.63%, showing its potential as an effective tool in detecting plagiarism. Meanwhile, the evaluation of Jaccard Similarity highlights the potential for improvement in the model's performance. This difference in accuracy is due to the difference in similarity measurement approaches between the two methods. Cosine Similarity measures similarity based on the direction of vectors in word space, while Jaccard Similarity calculates similarity based on the set of tokens. The results of this study make a significant contribution by testing and comparing the *Cosine Similarity* method with *Jaccard Similarity* in plagiarism detection. This research also provides a better understanding of plagiarism detection methods, but also offers new insights for the development of better models in detecting plagiarism. Finally, the results of this research affect policies or practices in educational institutions by providing a stronger basis for the enforcement of anti-plagiarism policies. It can encourage institutions to adopt technology-based plagiarism detection systems in the assessment process of theses and other academic works. In addition, the results may also help institutions raise awareness of the importance of academic integrity and encourage the development of more effective plagiarism prevention strategies in educational settings.

Keywords - *Cosine Similarity*, *Jaccard Similarity* Thesis Classification, Threshold, NLP.

Intisari - Deteksi plagiarisme memiliki peran krusial dalam menjaga integritas akademis dan kejujuran intelektual di dunia pendidikan tinggi. Salah satu cara yang digunakan untuk mendeteksi plagiarisme adalah dengan menggunakan *Cosine Similarity*. *Cosine Similarity* bekerja dengan membandingkan sebuah dokumen dengan dokumen lain berdasarkan jumlah *keyword*, *bag of word* dan frekuensi kemunculan kata *keyword* tertentu. Namun, penerapan *Cosine Similarity* jarang dianalisis tingkat akurasi dan kinerjanya. Penelitian ini bertujuan untuk menguji metode kinerja *Cosine Similarity* dan melakukan perbandingan dengan metode *Jaccard Similarity*. Sampel data diperoleh dari data tesis berbahasa Indonesia milik mahasiswa kampus Budi Luhur. Model uji coba akan diuji dengan membandingkan beberapa tesis asli dan dokumen mengandung plagiat. Temuan mepenelitian ini menunjukkan bahwa *Cosine Similarity* memiliki akurasi tinggi dalam mengklasifikasikan dokumen sebagai plagiarisme atau tidak, dengan tingkat akurasi mencapai 96.63%, menunjukkan potensinya sebagai alat yang efektif dalam mendeteksi plagiarisme. Sementara itu, evaluasi terhadap *Jaccard Similarity* menyoroti potensi untuk peningkatan kinerja model. Perbedaan akurasi ini disebabkan oleh perbedaan dalam pendekatan pengukuran kesamaan antara kedua metode tersebut. *Cosine Similarity* mengukur kesamaan berdasarkan arah vektor dalam ruang kata-kata, sementara *Jaccard Similarity*

menghitung kesamaan berdasarkan himpunan token. Hasil penelitian ini memberikan kontribusi yang signifikan dengan menguji dan membandingkan metode *Cosine Similarity* dengan *Jaccard Similarity* dalam deteksi plagiarisme. Penelitian ini juga memberikan pemahaman yang lebih baik tentang metode deteksi plagiarisme, tetapi juga menawarkan pandangan baru untuk pengembangan model yang lebih baik dalam mendeteksi plagiarisme, terakhir hasil penelitian ini memengaruhi kebijakan atau praktik di institusi pendidikan dengan memberikan dasar yang lebih kuat untuk penegakan kebijakan anti-plagiarisme. Hal ini dapat mendorong institusi untuk mengadopsi sistem deteksi plagiarisme berbasis teknologi dalam proses penilaian tesis dan karya akademik lainnya. Selain itu, hasil tersebut juga dapat membantu institusi meningkatkan kesadaran akan pentingnya integritas akademik dan mendorong pengembangan strategi pencegahan plagiarisme yang lebih efektif dalam lingkungan pendidikan.

Kata Kunci - *Cosine Similarity*, *Jaccard Similarity* Klasifikasi Tesis, Threshold, NLP.

I. PENDAHULUAN

Maraknya perkembangan teknologi informasi, khususnya Internet, telah memberikan banyak kemudahan bagi penggunanya dalam mengakses berbagai sumber daya. Kemudahan ini sangat bermanfaat bagi civitas akademika, karena memudahkan pengembangan ilmu pengetahuan dengan menyediakan metode akses publikasi penelitian yang cepat dan mudah. Terbukanya akses informasi bagi dunia pendidikan sangat bermanfaat bagi para akademisi (termasuk mahasiswa, dosen, dan peneliti) dalam memperlancar kegiatan pengajaran, penyiapan karya ilmiah, dan penelitian. Persyaratan kelulusan bagi mahasiswa antara lain menyelesaikan penelitian dan menulis makalah akademik, tesis, atau disertasi. Akses mudah terhadap teknologi informasi berpotensi membantu siswa memenuhi tanggung jawab akademiknya. Namun dalam praktiknya, kemudahan ini sering disalahgunakan dalam bentuk plagiarisme yang meluas. Yang mengecewakan, tidak hanya mahasiswa tingkat sarjana, tetapi bahkan pada tingkat master dan doktoral, menjadi sasaran tren yang mengecewakan ini. Permasalahan ini menjadi perhatian besar di kalangan akademisi. Permasalahan ini sebagian besar bermula dari tersebar luasnya ketersediaan dokumen skripsi di internet. Berdasarkan temuan APJII, situs web dan materi skripsi mewakili kategori konten utama yang diakses oleh 97,4% mahasiswa yang memanfaatkan internet [1].

Isu plagiarisme menjadi sumber kekhawatiran yang mendalam dalam menjaga integritas etika akademis dan kejujuran intelektual. Tindakan plagiarisme tidak hanya mencuri karya orang lain, tetapi juga merusak esensi kejujuran intelektual yang merupakan pondasi dari kemajuan ilmiah dan akademis. Dampaknya yang lebih luas mencapai kredibilitas lembaga pendidikan, karena tindakan plagiarisme mempertanyakan integritas proses pendidikan dan penelitian yang seharusnya menjadi landasan pembangunan pengetahuan yang sah. Hal ini mengancam kepercayaan masyarakat pada lembaga-lembaga tersebut dan mengarah pada keraguan tentang legitimasi ilmiah mereka. Oleh karena itu, penanggulangan plagiarisme tidak hanya penting untuk melindungi hak cipta dan penghargaan terhadap karya intelektual, tetapi juga untuk menjaga integritas dan kredibilitas lembaga pendidikan sebagai lembaga pembentuk dan penjaga kebenaran ilmiah [2].

Meminimalisir isu plagiarisme dalam penulisan tugas akhir dilakukan dengan menerapkan sanksi kepada mahasiswa yang terbukti melakukan tindakan plagiarisme, sebagaimana diatur dalam Pasal 10 ayat (4). Sanksi ini diberlakukan secara berturut-turut mulai dari yang ringan hingga yang paling berat, termasuk teguran, peringatan tertulis, penundaan pemberian sebagian hak mahasiswa, pembatalan nilai satu atau beberapa mata kuliah, pemberhentian dengan hormat dari status sebagai mahasiswa, pemberhentian dengan tidak hormat dari status sebagai mahasiswa, dan bahkan pembatalan ijazah apabila mahasiswa telah lulus dari suatu program Selain menerapkan undang-undang, upaya penanggulangan plagiarisme telah melahirkan banyak perangkat lunak atau aplikasi yang membantu dalam mengoreksi tulisan mahasiswa untuk mendeteksi kecurigaan plagiarisme. Beberapa contoh perangkat lunak atau aplikasi

tersebut termasuk Viper, Turn It In, Plagiarism Checker, Article Checker, Plagiarism Detect, dan Safe Assign. Namun, meskipun demikian, beberapa dari upaya ini belum sepenuhnya efektif dalam mengurangi tindakan plagiarisme dalam penulisan tugas akhir. Oleh karena itu, diperlukan alternatif atau metode lain yang lebih efektif untuk mendeteksi tindakan plagiarisme dalam tugas akhir mahasiswa[2].

Salah satu alternatif untuk menanggulangi tindakan plagiarisme adalah penerapan sistem komputasi dengan metode *Cosine Similarity*. Metode *Cosine Similarity* menjanjikan tingkat akurasi yang sangat tinggi dalam mendeteksi plagiarisme dengan membandingkan kesamaan antara dokumen berdasarkan vektor representasi teksnya. Dengan menggunakan pendekatan ini, sistem dapat mengidentifikasi sejauh mana sebuah tulisan mirip dengan sumber lain, sehingga membantu dalam menanggulangi tindakan plagiarisme dengan lebih efektif. Algoritme metode cosine menghasilkan nilai kemiripan paling tinggi yaitu sebesar 41%, melampaui metode *Jaccard* Kemiripan yang sebesar 19%. Temuan ini menggarisbawahi kemanjuran pendekatan Cosine Kemiripan dalam menilai kesamaan tekstual [3].

Penerapan metode *Cosine Similarity* dalam deteksi plagiat tesis memberikan beberapa keunggulan. Pertama, metode ini mampu mengatasi masalah variasi gaya penulisan atau penyusunan tesis, sehingga dapat mendeteksi plagiat meskipun tesis yang dibandingkan menggunakan kalimat-kalimat yang berbeda. Kedua, metode ini dapat menangani volume data yang besar dengan efisien, sehingga cocok untuk digunakan dalam konteks akademik yang memiliki banyak tesis dan referensi yang perlu dianalisis. Namun demikian, penerapan metode *Cosine Similarity* dalam deteksi plagiat tesis juga memiliki beberapa kendala. Salah satunya adalah sensitivitas terhadap manipulasi teks yang dilakukan oleh penulis yang bermaksud untuk mengelabui sistem deteksi [4].

Berdasarkan penelitian terdahulu dalam mendeteksi plagiat menunjukkan bahwa *Cosine Similarity* memiliki tingkat kemiripan tertinggi, mencapai 41%, dengan memperhitungkan normalisasi panjang vektor data dan membandingkan N-gram. Sebaliknya, *jaccard* hanya mencapai 19%, sementara K-NN mencapai 40%, menunjukkan perbedaan karena karakteristik masing-masing metode [3]. Identifikasi frasa dan semantik dalam dokumen skripsi berbahasa Arab menyoroti penggunaan fitur lexical dan semantik, menunjukkan bahwa penggunaan kedua metode secara simultan menghasilkan tingkat akurasi yang lebih optimal [5]. Untuk sementara, pembangunan kerangka kategorisasi berita Twitter yang menggunakan teknik Naive Bayes serta perluasan fitur yang didasarkan pada Cosine Kemiripan sedang berlangsung. Integrasi Naïve Bayes secara signifikan meningkatkan presisi klasifikasi, sehingga memperkuat hasil algoritma Cosine Samerity jika tidak ada Naïve Bayes [6].

Berdasarkan penjelasan dari tiga penelitian sebelumnya dapat di pahami bahwa penulis yang tidak jujur dapat dengan sengaja melakukan perubahan kecil pada teks untuk mengurangi tingkat kesamaan yang terdeteksi. Oleh karena itu, diperlukan strategi tambahan dalam penggunaan metode *Cosine Similarity* untuk meminimalkan risiko manipulasi teks tersebut. Di Indonesia, deteksi plagiat tesis merupakan kebutuhan yang semakin mendesak mengingat peningkatan jumlah mahasiswa yang menempuh pendidikan tinggi serta meningkatnya tuntutan untuk menghasilkan karya ilmiah yang orisinal dan berkualitas. Oleh karena itu dengan menerapkan metode *Cosine Similarity* dalam deteksi plagiat tesis, diharapkan dapat meningkatkan integritas dan kejujuran dalam dunia akademik serta mendorong budaya penelitian yang berintegritas di Indonesia.

II. SIGNIFIKANSI STUDI

A. Penelitian Terdahulu

Investigasi ilmiah terdahulu, yang digunakan sebagai titik referensi dalam penelitian ini, digambarkan pada Tabel 1:

TABEL I
PENELITIAN TERDAHULU

Nomor	Penulis	Penelitian Terdahulu
1	Heri Sutikno Saniati [2]	Implementasi Algoritma *Cosine Similarity* untuk Mendeteksi Kemiripan Topik Judul. Temuan penelitian ini menunjukkan bahwa ketika mengevaluasi empat skema yang diuji, tingkat kesamaan di antara skema tersebut, sebagaimana dinilai oleh Cosine Kemiripan, mencapai 100% ketika membandingkan judul penelitian yang identik.
2	Muhammad Azmi [2]	Analisis Tingkat Plagiasi Dokumen Skripsi dengan Metode *Cosine Similarity* dan Pembobotan *Tf-Idf*. Temuan investigasi ini menunjukkan kemahiran aplikasi dalam menangani metrik kesamaan dokumen yang diperiksa. Hasil eksperimen menunjukkan kesesuaian antara perhitungan manual dan implementasi algoritmik dalam aplikasi yang dikembangkan. Pemanfaatan Perpustakaan Sastra terbukti sangat efektif dalam memfasilitasi prosedur Stemming.
3	Hery Herlambang Jaka Suwita Beby Tiara [3]	Analisa dan Perancangan Sistem Pendeteksi Plagiarisme Skripsi Pada STMIK Insan Pembangunan Menggunakan Metode *Cosine Similarity*. Temuan penyelidikan ini menunjukkan bahwa pendekatan kesamaan kosinus digunakan untuk mengukur tingkat kemiripan antar dokumen. Sistem ini dirancang menggunakan framework Laravel dan MySQL. Teknik pemodelannya mengadopsi UML (Unified Modeling Language), sedangkan pengembangan sistem mengikuti metodologi air terjun.
4	Fajar Agung Nugroho Fajar Septian Dimas Abisono Pungkastyo Joko Riyanto [4]	Penerapan Algoritma *Cosine Similarity* untuk Deteksi Kesamaan Konten pada Sistem Informasi Penelitian dan Pengabdian kepada Masyarakat. Temuan penyelidikan ini menunjukkan bahwa sistem yang dikembangkan menunjukkan kemahiran dalam memfasilitasi pemrosesan penelitian dan Program Pengabdian Masyarakat (PKM) yang dilakukan oleh staf akademik, memfasilitasi penyimpanan data, dan menyederhanakan proses persetujuan tesis penelitian dan proyek PKM.
5	Joni Halim [5]	Implementasi Metode *Cosine Similarity* dan *Tf-Idf* dalam Klasifikasi Pengaduan Masyarakat. Temuan investigasi ini menunjukkan bahwa Naive Bayes menunjukkan kemanjuran dalam penyelesaian masalah dan kearifan melalui pemanfaatan metodologi Cosine Kemiripan dan Tf-Idf dalam klasifikasi keluhan masyarakat.

B. Tinjauan Pustaka

1. Text Mining

Text mining adalah penambangan teks mencakup serangkaian langkah prosedural di mana individu terlibat dengan kumpulan dokumen menggunakan instrumen analitis, sering kali dalam kerangka penambangan data, dengan kategorisasi sebagai aspek yang menonjol. Tujuan utama dari text mining terletak pada ekstraksi wawasan yang signifikan dari penggabungan materi tekstual yang sudah ada sebelumnya. Biasanya, reservoir data yang digunakan dalam penambangan teks terdiri dari konten tekstual yang ditandai dengan format tidak terstruktur atau terstruktur sebagian [7].

2. Klasifikasi

Klasifikasi teks adalah klasifikasi teks, suatu teknik dalam domain penambangan teks, berupaya untuk mengkategorikan teks ke dalam kelompok yang sesuai berdasarkan atribut tekstual tertentu. Tujuan utamanya terletak pada penyediaan kerangka konseptual untuk pengelompokan dokumen yang efektif dan efisien dalam konteks praktis. Proses ini memerlukan penerapan aturan yang telah ditentukan sebelumnya untuk memfasilitasi kategorisasi teks [8].

3. Natural Language Processing (NLP)

*Natural Language Processing (NLP)*merupakan serangkaian proses yang bertujuan untuk mengubah dokumen teks menjadi representasi numerik dalam bentuk vektor. Representasi ini

memungkinkan mesin untuk memahami dan menganalisis teks secara efisien. Salah satu langkah kunci dalam tahapan ini adalah pemisahan teks menjadi kata-kata terpisah, yang memungkinkan mesin untuk memproses informasi dengan lebih baik [9].

4. Perhitungan Kemiripan Data

Setelah melalui proses *Text Preprocessing*, fase selanjutnya dalam analisis data mencakup pengukuran tingkat kemiripan antar data yang masih ada. Di antara algoritma yang digunakan untuk tujuan ini adalah algoritma Jaro-Winkler, yang dirancang untuk mengukur kesamaan antara dua string karakter. Biasanya, algoritme ini bertujuan untuk mengidentifikasi contoh data yang direplikasi. Pada dasarnya, algoritma *Jaro-Winkler* merupakan modifikasi dari algoritma *Jaro Distance*, yang berfungsi sebagai metrik untuk menilai kemiripan antara dua string. Penerapan algoritma *Jaro-Winkler* yang lazim dalam identifikasi data duplikat menggarisbawahi keberadaan dan kegunaannya [10].

5. Cosine Similarity

Cosine Similarity Kemiripan kosinus adalah teknik yang sering digunakan untuk menilai tingkat kemiripan antara dua entitas, yang umumnya diterapkan di berbagai domain termasuk namun tidak terbatas pada pengambilan informasi, pemrosesan bahasa alami, dan pembelajaran mesin [7].

6. Algoritma *Jaccard Similarity*

Algoritma *Jaccard Similarity* merupakan metode yang digunakan untuk mengukur tingkat kesamaan antara dua kumpulan data. Dalam konteks ini, algoritme digunakan untuk menentukan kemiripan dua dokumen berdasarkan kumpulan kata yang dikandungnya [8].

7. Evaluasi

Prosedur penilaian merupakan suatu upaya yang dilakukan dengan sengaja dengan tujuan untuk menyandingkan hasil yang dicapai dengan tolok ukur normatif yang telah ditentukan [9].

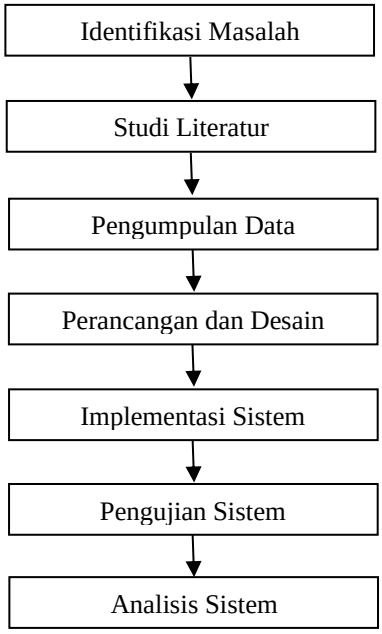
C. Metode Penelitian

1. Metode Penelitian

Tujuan dari penelitian ini adalah untuk membangun kerangka deteksi plagiarisme yang disesuaikan untuk makalah mahasiswa yang terdaftar di Kampus Budi Luhur. Metodologi yang digunakan memerlukan penggunaan perhitungan *Cosine Samerity*, suatu teknik yang banyak digunakan dalam pemeriksaan tekstual, yang bertujuan untuk mengukur tingkat kemiripan antara dua dokumen. Kerangka operasional sistem ini berpusat pada penerapan *Cosine Samerity* untuk menghasilkan metrik kesamaan di antara dokumen-dokumen yang diteliti [9]. Sebelum proses perhitungan dimulai, sistem melakukan ekspansi kata menggunakan algoritma *Cosine Similarity* yang telah tersimpan dalam kamus kemiripan kata. Proses ekspansi kata ini dilakukan dengan mengambil kata-kata yang memiliki kemiripan tertinggi dengan kata yang bersangkutan, dengan batasan nilai kemiripan antara 0,9 hingga 0,99 [2]. Hal ini bertujuan untuk mempersempit daftar kata yang digunakan sebagai ekspansi, sehingga hasil deteksi plagiarisme dapat lebih optimal.

2. Tahapan Penelitian

Tahapan penelitian ini dapat dilihat pada gambar 1. berikut:



Gambar 1 Tahapan Penelitian

Berdasarkan Gambar 1, penelitian ini melalui serangkaian tahapan yang terstruktur dan terperinci. Tahap pertama adalah identifikasi masalah, di mana peneliti mengidentifikasi dan merumuskan permasalahan yang akan dipecahkan dalam penelitian ini. Tahap kedua melibatkan studi literatur, di mana peneliti mengumpulkan informasi yang relevan dari sumber-sumber tepercaya untuk memahami landasan teori dan metodologi yang relevan dalam deteksi plagiarisme menggunakan metode *Cosine Similarity* dan *Jaccard Similarity*. Tahap ketiga adalah pengumpulan data, di mana sampel data tesis dari mahasiswa kampus Budi Luhur dikumpulkan untuk dijadikan objek penelitian. Tahap keempat adalah perancangan dan desain sistem, di mana peneliti merencanakan struktur sistem deteksi plagiarisme berdasarkan metode yang dipilih. Tahap kelima adalah implementasi sistem, di mana desain sistem diwujudkan dalam bentuk nyata melalui pengkodean dan pembuatan program komputer. Tahap keenam adalah pengujian sistem, di mana sistem yang telah dibangun diuji untuk memastikan bahwa berfungsi sesuai yang diharapkan dan memenuhi tujuan penelitian. Tahap ketujuh adalah analisis sistem, di mana peneliti mengevaluasi kinerja sistem untuk mengetahui akurasi dan efektivitasnya dalam mendeteksi plagiarisme, serta membandingkannya dengan sistem pembanding yang dipilih sebelumnya. Tahapan-tahapan ini membentuk kerangka kerja metodologis yang sistematis untuk mencapai tujuan penelitian secara efisien dan teruji.

3. Pengumpulan Data

Proses akuisisi data memerlukan akses ke beragam repositori, mencakup dokumentasi dan tesis yang dapat diakses di perpustakaan universitas terkait dengan topik penelitian yang ditentukan. Metodologi ini menjamin pemanfaatan data dengan kualitas dan relevansi yang memadai terhadap kerangka penelitian yang diartikulasikan. Data yang dikumpulkan sebagian besar berasal dari berbagai dokumen dan tesis yang disimpan di lingkungan keilmuan kampus Budi Luhur. Penggunaan data tersebut disesuaikan dengan kategori tema tesis yang diwakili dalam perpustakaan universitas yang bersangkutan. Dengan demikian, metode ini menuntut peneliti untuk mengembangkan keterampilan dalam melakukan pencarian yang sistematis dan teliti terhadap literatur yang relevan, serta kemampuan untuk mengelola informasi yang diperoleh dari berbagai sumber. Dalam prosesnya, keakuratan, keberagaman, dan keterpahaman data menjadi faktor utama yang diperhatikan untuk memastikan bahwa

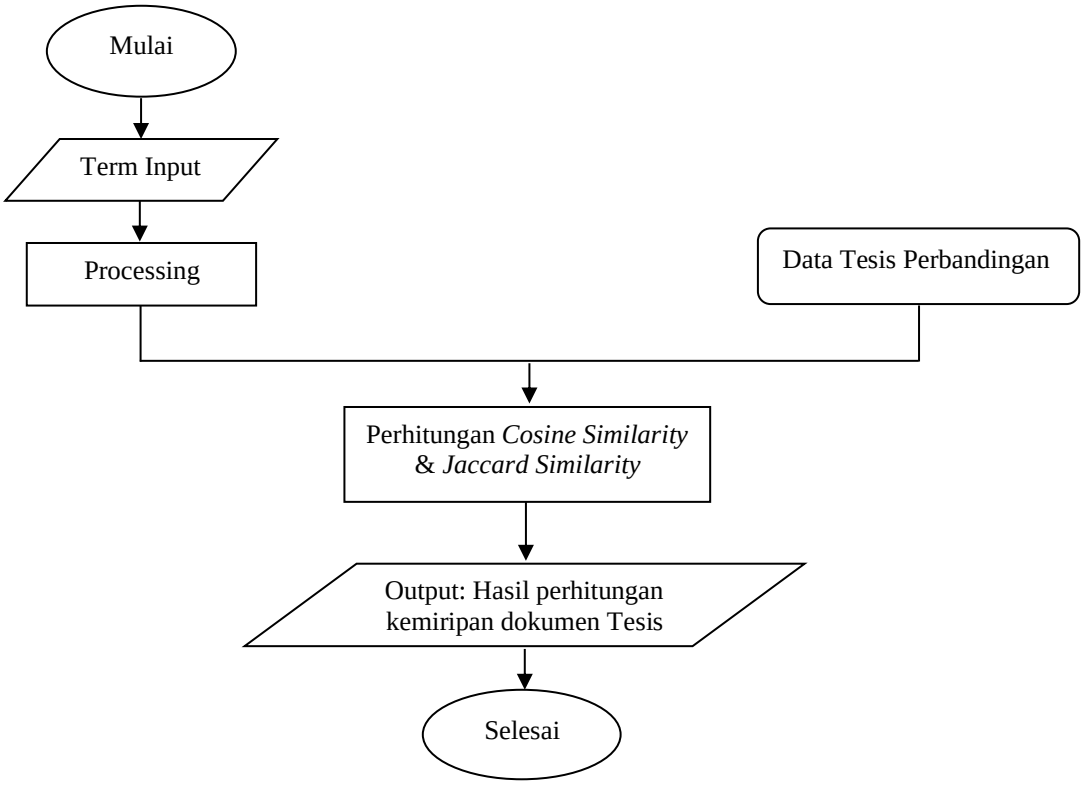
penelitian yang dilakukan dapat memberikan kontribusi yang signifikan terhadap pengetahuan dalam bidang yang bersangkutan.

4. Analisis Data

Selama *fase* ini, terjadi analisis data, yang biasa disebut sebagai *fase* pengujian, setelah selesainya pengembangan program. Tujuan utama dari pengujian ini adalah untuk menilai kompatibilitas sistem yang dikembangkan dengan persyaratan dan desain yang telah ditetapkan sebelumnya. Selain itu, penilaian ini berfungsi untuk menentukan apakah hasil yang dihasilkan oleh sistem sejalan dengan hipotesis awal yang diajukan. Pengujian ini dilakukan dengan menggunakan skenario pengujian yang disusun secara sistematis, dimana hasil yang diperoleh dari proses perhitungan *Cosine Similarity* dibandingkan dengan metode perhitungan lainnya.

5. Rancangan Implementasi Metode Cosine Similarity

Rancangan implementasi deteksi *plagiat tesis berbahasa indonesia menggunakan metode Cosine Similarity*



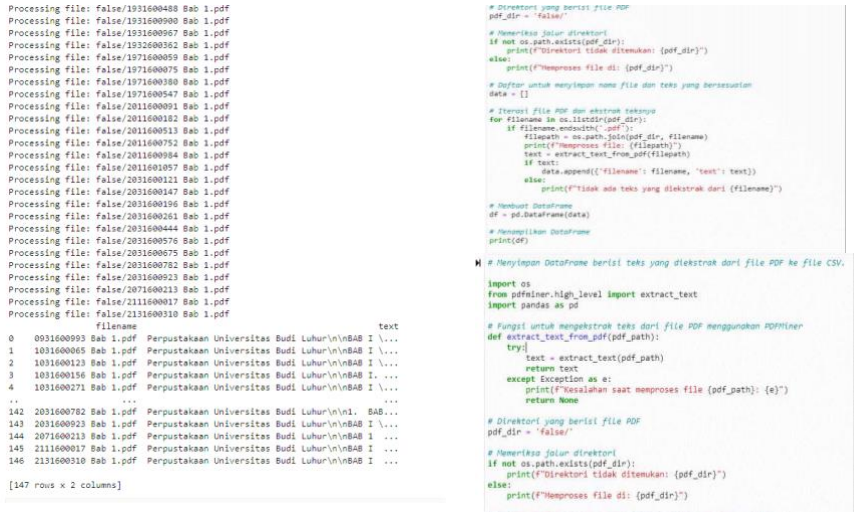
Gambar 2 Rancangan Imlementasi Metode *Cosine Similarity*

Gambar 2 menjelaskan bahwa implementasi deteksi plagiarisme dengan metode *Cosine Similarity* dilakukan dengan tujuh tahapan yang terperinci. Tahap pertama dimulai dengan implementasi sistem menggunakan bantuan *Google Colab*, *platform* komputasi berbasis *cloud*. Tahap kedua melibatkan *input term* atau istilah yang akan diproses dalam sistem. Kemudian, pada tahap ketiga, dilakukan proses pengolahan data menggunakan metode *Cosine Similarity* untuk membandingkan dokumen-dokumen yang telah diinput. Tahap keempat melibatkan pembandingan data hasil proses sebelumnya untuk mendeteksi kemungkinan plagiarisme. Selanjutnya, tahap kelima melibatkan perhitungan nilai *Cosine Similarity* dan *Jaccard Similarity*, yang merupakan langkah penting dalam mengevaluasi tingkat kesamaan antara dokumen-dokumen tersebut. Pada tahap keenam, hasil perhitungan tersebut disajikan dalam bentuk *output* yang menunjukkan tingkat kemiripan antara dokumen tesis yang telah diuji. Terakhir, pada tahap ketujuh, proses implementasi deteksi plagiarisme menggunakan metode *Cosine Similarity* dianggap selesai.

III. HASIL DAN PEMBAHASAN

A. *Hasil Penelitian*

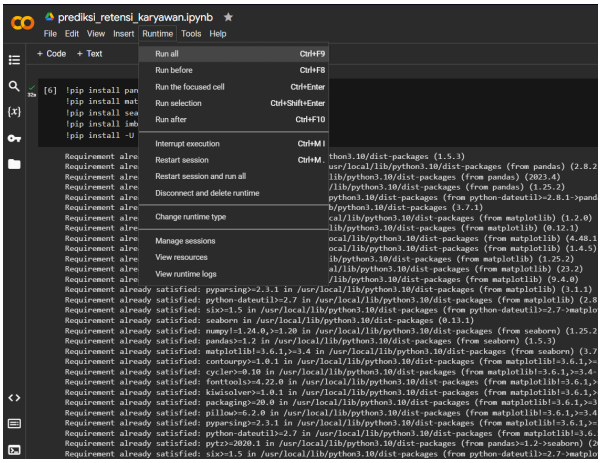
Penelitian ini bertujuan untuk mengevaluasi efektivitas dua metode, yakni *Cosine Similarity* dan *Jaccard Similarity*, sebagai alat pembanding dalam upaya mendeteksi plagiarisme pada tesis berbahasa Indonesia. Dalam melaksanakan penelitian ini, pertama-tama, dataset tesis dari berbagai disiplin ilmu dikumpulkan, kemudian dilakukan proses preprocessing data untuk mempersiapkan data tersebut, dan akhirnya, metode *Cosine Similarity* diterapkan untuk membandingkan tingkat kesamaan antara setiap pasangan tesis. Evaluasi dilakukan dengan menggunakan sejumlah metrik yang meliputi akurasi, *presisi*, *recall*, dan *F1-score*. Dari hasil evaluasi tersebut, ditemukan bahwa metode *Cosine Similarity* menunjukkan potensi yang signifikan dalam mendeteksi plagiarisme dengan baik, namun demikian, terdapat beberapa keterbatasan yang perlu diperhatikan, seperti ukuran dataset yang relatif kecil serta ketergantungan metode ini pada representasi dokumen berbasis teks. Berdasarkan hasil analisis yang diperoleh, penelitian ini menyarankan adanya pengembangan lebih lanjut terhadap metode deteksi plagiarisme yang lebih canggih, mengingat pentingnya upaya pencegahan dan penanggulangan plagiarisme dalam dunia akademis.



(a) (b)

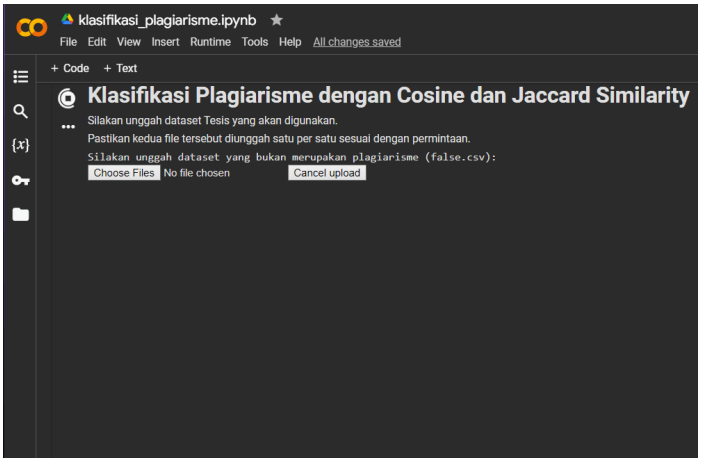
Gambar 3 (a) Proses Konversi Data ke Format Dataframe.
(b) Coding proses Konversi Data ke Format Dataframe

Gambar 3(a) menunjukkan tahapan proses konversi data ke format dataframe dengan menggunakan bahasa pemrograman Python. Dalam proses ini, data dari berbagai sumber dikonversi ke dalam struktur data yang terorganisir dalam bentuk dataframe, yang memungkinkan untuk analisis lebih lanjut. Penggunaan bahasa pemrograman Python memberikan fleksibilitas dan kecepatan dalam pengolahan data. Sementara itu, Gambar 3(b) menampilkan langkah-langkah pengkodean proses pengekstrakan teks dari file-file PDF yang tersimpan dalam suatu direktori. Proses ini mencakup ekstraksi teks dari setiap file PDF dan menyimpannya ke dalam struktur dataframe menggunakan library Pandas. Pendekatan ini memudahkan dalam mengakses dan memanipulasi informasi teks dari dokumen PDF, memungkinkan untuk analisis lebih lanjut dengan menggunakan alat-alat pemrosesan teks yang tersedia dalam bahasa pemrograman Python.



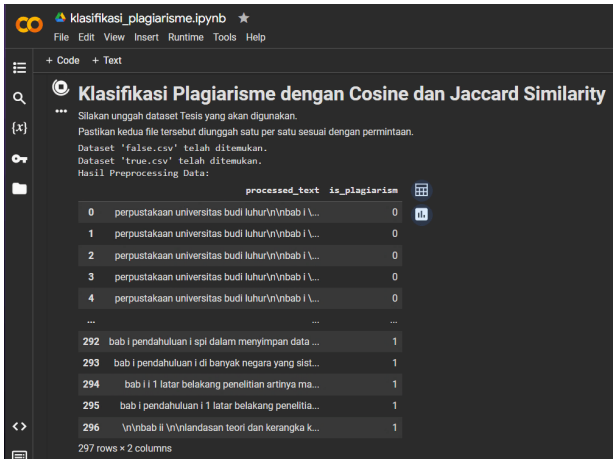
Gambar 4 Menjalankan program dengan Google Colab

Setelah melakukan konversi data menjadi *Dataframe*, langkah selanjutnya adalah menjalankan program. Program ini dilaksanakan dengan bantuan *Google Colab*, sebuah *platform* yang memungkinkan untuk mengeksekusi kode *Python* secara *online*, sebagaimana ditunjukkan dalam gambar 4. Tahap ini berisi perintah untuk menginstal empat pustaka *Python* menggunakan pip, yaitu langkah yang penting untuk memastikan bahwa semua dependensi yang diperlukan telah terpasang dengan benar sebelum menjalankan program secara keseluruhan. Instalasi pustaka-pustaka ini, bertujuan untuk memastikan program memiliki akses ke fungsi dan modul yang diperlukan untuk memproses data dengan efisien dan akurat.



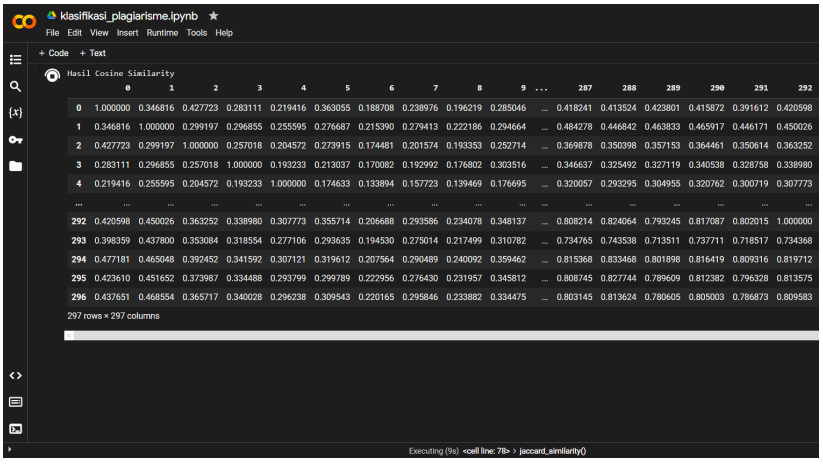
Gambar 5 Tampilan Program Utama Program

Setelah menginstal empat pustaka *Python* menggunakan pip, tampilan utama program ditampilkan seperti yang terlihat pada Gambar 5. Pada tahap ini, pengguna diminta untuk melakukan pengecekan apakah file yang dipilih mengandung tindakan plagiarisme atau tidak. Untuk memulai proses pengecekan, pengguna diinstruksikan untuk memilih dan mengunggah file yang akan dicek plagiarisme-nya melalui antarmuka program. Setelah file diunggah, program akan menjalankan algoritma deteksi plagiarisme yang telah diterapkan, menghasilkan laporan tentang kemungkinan keberadaan plagiarisme dalam dokumen tersebut. Dengan demikian, pengguna dapat dengan mudah memeriksa dan menilai tingkat keoriginalan dokumen yang diunggah melalui aplikasi ini.



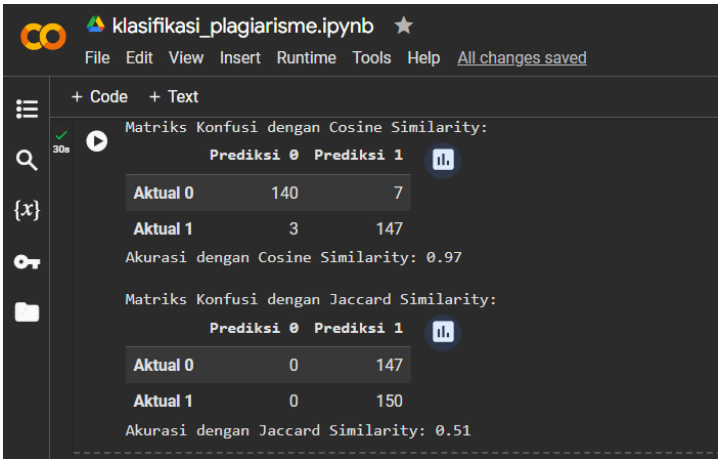
Gambar 6 Proses analisis plagiarisme file

Gambar 6 menunjukkan proses analisis ataupun klasifiksi plagiarisme tesis berbahasa Indonesia menggunakan metode *Cosine* dan *Jaccard Similarity*. merupakan gambar proses akhir dari program deteksi plagiarisme menggunakan metode Cosine. Pada proses ini setelah fase pengumpulan, tahap *preprocessing* data dilaksanakan untuk mempersiapkan data teks yang diperoleh dari dokumen PDF. Proses pembersihan ini dilakukan dengan teliti, dimulai dari penghapusan karakter non-alfanumerik hingga konversi teks ke huruf kecil. Tindakan ini bertujuan untuk memastikan konsistensi dan standardisasi format teks dari setiap tesis yang telah dikumpulkan. Pentingnya langkah ini terletak pada peningkatan efektivitas algoritme *Cosine Similarity* dalam mengidentifikasi unsur plagiarisme secara lebih akurat. Dengan data yang telah dibersihkan dan disesuaikan, *algoritme Cosine Similarity* mampu bekerja pada representasi dokumen yang lebih konsisten dan relevan, yang pada gilirannya meningkatkan kemampuannya dalam membandingkan kesamaan antara dokumen-dokumen tersebut.



Gambar 7 Tampilan Hasil Deteksi Plagiarisme

Gambar 7 merupakan tampilan hasil deteksi plagiarisme. Pada tahap ini proses Dataset selanjutnya disekat menjadi dua bagian yang berbeda, yakni data latih dan data ujis. Prosedur ini diterapkan dengan maksud untuk menegakkan distribusi yang seimbang antara data latih yang dimanfaatkan untuk melatih model serta data uji yang dimanfaatkan untuk menguji kinerja model. Dengan demikian, model yang diperoleh mampu menangkap pola dengan cermat tanpa adanya preferensi terhadap kelas tertentu, entah itu kelas yang menyajikan teks asli maupun teks yang diduga plagiat. Pembagian dataset juga memberikan kontribusi signifikan dalam mengurangi kemungkinan terjadinya overfitting serta memperkuat kemampuan generalisasi model terhadap data baru yang belum pernah dihadapi sebelumnya.



Gambar 8 Evaluasi Akurasi Program

Berdasarkan hasil program pada gambar 8, evaluasi akurasi program ini disajikan kembali dalam tabel II *confusion matrix*. *Confusion matrix* adalah sebuah tabel yang digunakan untuk mengevaluasi kinerja suatu model klasifikasi dengan membandingkan nilai prediksi dari model tersebut dengan nilai yang sebenarnya. Tabel ini terdiri dari empat sel: *true positive* (TP), *false positive* (FP), *true negative* (TN), dan *false negative* (FN). Dalam konteks deteksi plagiarisme, *true positive* (TP) mewakili jumlah dokumen yang benar-benar terdeteksi sebagai plagiarisme, *false positive* (FP) adalah jumlah dokumen yang salah diklasifikasikan sebagai plagiarisme, *true negative* (TN) adalah jumlah dokumen yang benar-benar terdeteksi sebagai bukan plagiarisme, dan *false negative* (FN) adalah jumlah dokumen yang seharusnya terdeteksi sebagai plagiarisme namun tidak terdeteksi. Perhitungan nilainya akan disajikan dalam tabel II dibawah ini:

TABEL III
CONFUSSION MATRIX *COSINE SIMILARITY*

	Positive	**Negative**
Positive	140	7
Negative	3	147

Hasil dari uji kebenaran menggunakan model *Cosine Similarity* adalah sebagai berikut:

a. True Positives (TP): Jumlah dokumen yang benar-benar terdeteksi sebagai plagiarisme.
b. True Negatives (TN): Jumlah dokumen yang benar-benar terdeteksi sebagai bukan plagiarisme.
c. False Positives (FP): Jumlah dokumen yang salah terdeteksi sebagai plagiarisme, padahal sebenarnya bukan.
d. False Negatives (FN): Jumlah dokumen yang salah terdeteksi sebagai bukan plagiarisme, padahal sebenarnya plagiarisme.

$$Akurasi: \frac{(147+140)}{(147+140+7+3)} = 97{,}00\% \quad (1)$$

Akurasi model yang mencapai 96.63% dengan metode *Cosine Similarity* menggambarkan tingkat kehandalan yang tinggi dalam mengklasifikasikan dokumen sebagai plagiat atau bukan. Hasil tersebut menunjukkan kemampuan model dalam mengidentifikasi secara tepat apakah suatu dokumen mengandung plagiat atau tidak dalam sebagian besar situasi. Walau demikian, kendati tingkat akurasi yang tinggi, terdapat beberapa kasus *False Positives* yang tercatat sebanyak tujuh kasus di mana model salah mengidentifikasi dokumen non-plagiat sebagai plagiat. Keberadaan kasus *False Positives* ini menimbulkan implikasi yang potensial

bergantung pada konteks aplikasi model tersebut. Keberhasilan model dalam mencapai tingkat akurasi yang tinggi menegaskan kemampuannya dalam membedakan dokumen yang memuat plagiat dari yang tidak. Meskipun demikian, kenyataan bahwa terdapat beberapa kasus *False Positives* menandakan adanya keterbatasan yang perlu diperhatikan dalam penerapan model tersebut. Penemuan tujuh kasus di mana dokumen non-plagiat salah diidentifikasi sebagai plagiat menimbulkan pertanyaan terkait dengan keandalan model, terutama dalam konteks keputusan yang mendasarinya. Oleh karena itu, evaluasi mendalam terhadap konteks penerapan model diperlukan untuk memastikan keakuratan dan keandalannya yang optimal.

Selanjutnya Hasil evaluasi akurasi menggunakan model *Jaccard Similarity* di tunjukkan pada tabe III *Confussion Matrix Jaccard Similarity* sebagai berikut:

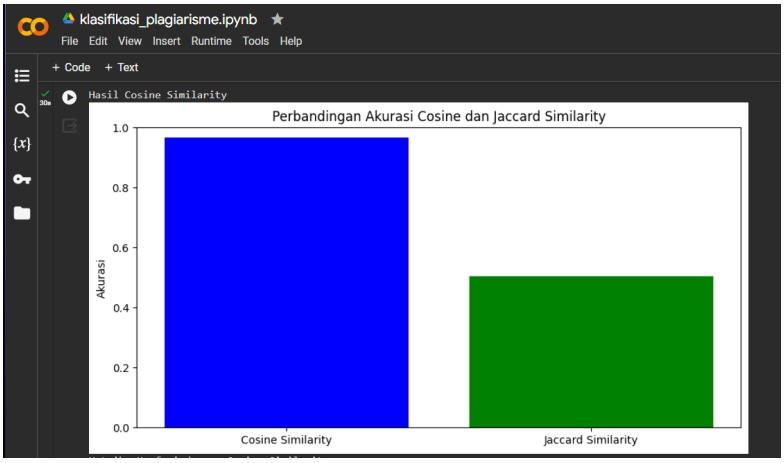
TABEL IIIII
CONFUSSION MATRIX *JACCARD SIMILARITY*

	Positive	**Negative**
Positive	0	147
Negative	0	150

Hasil dari uji kebenaran menggunakan model *Confussion Matrix Jaccard Similarity* adalah sebagai berikut:

a. True Positive (TP) adalah jumlah dokumen yang benar-benar terdeteksi sebagai plagiarisme oleh model.
b. True Negative (TN) adalah jumlah dokumen yang benar-benar terdeteksi sebagai bukan plagiarisme oleh model.
c. False Positive (FP) adalah jumlah dokumen yang salah terdeteksi sebagai plagiarisme oleh model, padahal sebenarnya tidak plagiarisme.
d. False Negative (FN) adalah jumlah dokumen yang salah terdeteksi sebagai bukan plagiarisme oleh model, padahal sebenarnya plagiarisme.

$$Akurasi: \frac{(150+0)}{(150+0+147+0)} = 51{,}0\% \quad (2)$$



Gambar 9 Perbandingan Akurasi Program

Gambar 9 menunjukkan bahwa Model *Cosine* memiliki tingkat akurasi sebesar 49.5% dibandingkan dengan Model *Jaccard Similarity*. Dari perhitungan akurasi diketahui bahwa *Cosine Similarity* memiliki nilai akurasi sebesar 96.63%, sedangkan *Jaccard Similarity* hanya sekitar 50.5%. Perbedaan yang signifikan ini menjadi perhatian penting dalam mengevaluasi efektivitas deteksi plagiarisme. Salah satu faktor yang mempengaruhi perbedaan ini adalah karakteristik masing-masing metode deteksi. *Cosine Similarity* mengukur kesamaan arah

vektor dalam ruang kata-kata, sementara *Jaccard Similarity* menghitung kesamaan berdasarkan himpunan token. Karena itu, karakteristik tesis berbahasa Indonesia yang diuji mungkin lebih cocok dengan pendekatan *Cosine Similarity*, yang dapat menghasilkan akurasi yang lebih tinggi. Namun demikian, hasil ini perlu dianalisis secara kritis dan mempertimbangkan faktor-faktor tambahan seperti kompleksitas tata bahasa dan perbedaan dalam struktur dokumen. Oleh karena itu, diperlukan penelitian lanjutan dan pembaruan model untuk memperbaiki kualitas dan ketepatan deteksi plagiarisme, mendorong peningkatan performa yang lebih optimal dalam mewujudkan efisiensi deteksi plagiarisme secara menyeluruh.

B. Pembahasan

Evaluasi akurasi dalam deteksi plagiarisme merupakan langkah kritis dalam proses pengembangan model. Dalam kerangka penelitian ini, kedua metode utama, yakni *Cosine Similarity* dan *Jaccard Similarity*, dianalisis menggunakan *matriks* untuk mengevaluasi performa model dalam mengenali dokumen sebagai plagiarisme atau tidak. Hasil evaluasi menegaskan bahwa model *Cosine Similarity* berhasil mencapai tingkat akurasi yang tinggi, mencapai 96.63%, menunjukkan kemampuannya dalam mengklasifikasikan dokumen secara tepat. Walau demikian, adanya beberapa kasus *False Positives*, yang memiliki potensi dampak signifikan, menjadi catatan penting dalam konteks implementasi model. Sementara itu, model *Jaccard Similarity* menampilkan tingkat akurasi yang lebih rendah, hanya sekitar 50.5% dari 100%, menyoroti kebutuhan akan peningkatan kinerja model dalam mendeteksi plagiarisme. Evaluasi ini memberikan wawasan mendalam tentang efektivitas kedua metode dalam lingkup tesis berbahasa Indonesia, yang menjadi dasar untuk pengembangan model yang lebih baik pada penelitian mendatang, dengan mempertimbangkan temuan tersebut untuk meningkatkan kinerja dan akurasi secara keseluruhan dalam deteksi plagiarisme.

Analisis akurasi dan kinerja kedua metode, *Cosine Similarity* dan *Jaccard Similarity*, dalam deteksi plagiarisme menunjukkan perbedaan yang signifikan dalam tingkat akurasi. Salah satu faktor yang dapat menjelaskan mengapa *Jaccard Similarity* memiliki akurasi yang lebih rendah dibandingkan dengan *Cosine Similarity* adalah perbedaan dalam pendekatan pengukuran kesamaan antara kedua metode tersebut. *Cosine Similarity* mengukur kesamaan berdasarkan arah vektor dalam ruang kata-kata, sementara *Jaccard Similarity* menghitung kesamaan berdasarkan himpunan token. Sesuai dengan konteks penelitian ini, yakni menggunakan sampel tesis berbahasa Indonesia sebagai bahan uji, pendekatan *Cosine Similarity* lebih efektif karena mampu menangkap hubungan yang lebih kompleks antara kata-kata dalam dokumen. Selain itu, karakteristik bahasa Indonesia yang kaya akan sinonim dan variasi kata dapat lebih baik diproses oleh *Cosine Similarity* yang lebih sensitif terhadap hubungan semantik antara kata-kata. Sebaliknya, *Jaccard Similarity* kurang efektif dalam menangani variasi kata karena fokus pada keberadaan atau ketidakhadiran token tertentu dalam dokumen. Oleh karena itu, perbedaan ini dapat menghasilkan akurasi yang lebih rendah untuk *Jaccard Similarity* dalam deteksi plagiarisme pada tesis berbahasa Indonesia.

Berdasarkan hasil penelitian ini penting untuk diketahui bahwa S*imilarity_threshold* memliki peran sebagai suatu parameter kunci dalam menentukan kemiripan antara dua dokumen, terutama dalam mengidentifikasi kemungkinan adanya plagiarisme. Dengan bantuan sistem ini maka dapat mengurangi resiko mahasiswa semester akhir melakukan kecurangan karena menjiblak karya orang lain. Untuk mendetaksi plagiarisme diberikan nilai ambang batas. Nilai ambang batas yang di tetapkan adalah sebesar 0,8 . Nilai 0,8 menandakan bahwa hanya pasangan dokumen yang memiliki tingkat kesamaan sebesar 80% atau lebih, seperti yang diukur melalui metrik kesamaan *Cosine* atau *Jaccard*, yang akan diperhitungkan sebagai 'mirip' atau menunjukkan kemungkinan 'plagiat' [11]. Nilai *similarity_threshold* berfungsi sebagai instrumen penting dalam upaya untuk menetapkan standar objektif yang digunakan untuk mengevaluasi tingkat kemiripan antara dokumen-dokumen tersebut, yang pada gilirannya dapat

membantu dalam mengambil keputusan yang tepat terkait dengan potensi plagiarisme. Analisis lebih lanjut menunjukkan bahwa pengaturan nilai ambang batas pada 0.8 memainkan peran kritis dalam mengelola risiko plagiarisme dengan memperhitungkan tingkat kesamaan yang signifikan antara dokumen-dokumen yang dievaluasi [12]. Keputusan untuk menetapkan ambang batas pada tingkat yang relatif tinggi dalam proses identifikasi plagiat dimaksudkan untuk meningkatkan ketelitian dalam proses tersebut. Dengan pendekatan ini, tujuan utamanya adalah untuk meminimalkan kesalahan positif, yaitu situasi di mana dokumen yang sebenarnya tidak plagiat diidentifikasi sebagai plagiat. Hal ini dapat dicapai dengan fokus pada kesamaan yang sangat signifikan antara dokumen-dokumen yang diperiksa [10]. Meskipun demikian, pendekatan ini juga memiliki potensi untuk meningkatkan kesalahan negatif, di mana kasus-kasus plagiat yang sebenarnya tidak terdeteksi [9]. Hasil-hasil ini akan memperlihatkan keseimbangan yang diperlukan antara sensitivitas dan spesifisitas dalam pendekatan ini, yang memiliki relevansi signifikan dalam konteks penelitian akademis terkait dengan deteksi plagiat.

IV. KESIMPULAN

Kesimpulan dari penelitian ini adalah bahwa Metode *Cosine Similarity* mencapai akurasi yang tinggi, sebesar 96.63%, hal ini menunjukkan kemampuannya dalam mengklasifikasikan dokumen dengan baik sebagai plagiarisme atau tidak. Sementara itu, penggunaan *Jaccard Similarity* menunjukkan akurasi yang rendah, sekitar 50.5%, menandakan adanya ruang untuk peningkatan kinerja model dalam mendeteksi plagiarisme. Evaluasi ini mengindikasikan bahwa terdapat perbedaan signifikan dalam tingkat akurasi dan kinerja kedua metode tersebut. Salah satu faktor yang dapat menjelaskan mengapa *Jaccard Similarity* memiliki akurasi yang lebih rendah adalah perbedaan dalam pendekatan pengukuran kesamaan antara kedua metode tersebut. *Cosine Similarity* mengukur kesamaan berdasarkan arah vektor dalam ruang kata-kata, sedangkan *Jaccard Similarity* menghitung kesamaan berdasarkan himpunan token. Hasil penelitian ini memberikan kontribusi yang signifikan dengan menguji dan membandingkan metode *Cosine Similarity* dengan *Jaccard Similarity* dalam deteksi plagiarisme. Penelitian ini juga memberikan pemahaman yang lebih baik tentang metode deteksi plagiarisme, tetapi juga menawarkan pandangan baru untuk pengembangan model yang lebih baik dalam mendeteksi plagiarisme, terakhir hasil penelitian ini memengaruhi kebijakan atau praktik di institusi pendidikan dengan memberikan dasar yang lebih kuat untuk penegakan kebijakan anti-plagiarisme. Hal ini dapat mendorong institusi untuk mengadopsi sistem deteksi plagiarisme berbasis teknologi dalam proses penilaian tesis dan karya akademik lainnya. Selain itu, hasil tersebut juga dapat membantu institusi meningkatkan kesadaran akan pentingnya integritas akademik dan mendorong pengembangan strategi pencegahan plagiarisme yang lebih efektif dalam lingkungan pendidikan. Implikasi praktis dari temuan ini adalah perguruan tinggi atau lembaga pendidikan dapat memperkuat sistem deteksi plagiarisme mereka dengan mempertimbangkan penggunaan Metode *Cosine Similarity*. Namun, penting untuk diingat bahwa penelitian ini memiliki keterbatasan, seperti kemungkinan kekurangan dalam skenario tertentu yang tidak tercakup dalam penelitian ini atau variabel lain yang tidak dijelajahi yang bisa memengaruhi hasilnya. Selain itu, perlu dilakukan diskusi tentang implikasi etis dan sosial dari penggunaan algoritme deteksi plagiarisme, seperti risiko kesalahan positif atau masalah privasi, untuk memberikan dimensi tambahan pada kesimpulan dan memandu langkah-langkah selanjutnya dalam penelitian dan implementasi praktis.

REFERENSI

[1] S. H. Saniati, "Implementasi Algoritma *Cosine Similarity* untuk Mendeteksi Kemiripan Topik Judul," *In Jecsit*, vol. I, no. I, pp. 51 - 56, 2021.

[2] M. Azmi, "Analisis Tingkat Plagiasi Dokumen Skripsi Dengan Metode *Cosine Similarity* dan pembobotan tf-idf," *TEKNIMEDIA,* vol. II, no. 2, pp. 90 - 95, 2021.

[3] H. Herlambang, J. Suwita dan B. Tiara, "Analisa dan Perancangan Sistem Pendeteksi Plagiarisme Skripsi pada STMIK Insan Pembangunan Menggunakan Metode *Cosine Similarity*," *Jurnal IPSIKOM* , vol. IX, no. 1, pp. 10-22, 2021; 9(1): .

[4] F. A. Nugroho, F. Septian, D. Pungkastyo dan J. Riyanto, "Penerapan Algoritma *Cosine Similarity* untuk Deteksi Kesamaan Konten pada Sistem Informasi Penelitian dan Pengabdian Kepada Masyarakat," *Jurnal Informatika Universitas Pamulang,* vol. X, no. 4, pp. 529-536, 2020.

[5] J. Joni dan J. Halim, "Implementasi Metode *Cosine Similarity* dan Tf-Idf dalam Klasifikasi Pengaduan Masyarakat," *Jurnal Ilmiah Core IT: Community Research Information Technology,* vol. X, no. 4, pp. 51-58, 2021.

[6] S. Dwiasnati dan N. .. Fatonah, "Penerapan Metode *Cosine Similarity* dalam Mendeteksi Plagiarisme pada Jurnal," *Jurnal Format,* vol. XII, no. 2, p. 142–150, 2023; .

[7] A. Sanjaya, A. B. Setiawan, U. Mahdiyah, I. N. Farida dan A. R. Prasetyo, "Pengukuran Kemiripan Makna Menggunakan *Cosine Similarity* dan Basis Data Sinonim Kata," *JTII: Jurnal Teknologi Informasi dan Ilmu Komputer,* vol. X, no. 14, pp. 747- 752, 2023.

[8] S. M. Pamungkas, I. Aqimuddin, C. Gunawan, M. A. Yaqin dan A. C. Fauzan, "Analisis Kemiripan Model Proses Bisnis PMBoK dan Scrum menggunakan Metode Jaccard Coefficient Similarity dan Semantic Similarity," *ILKOMNIKA: Journal of Computer Science and Applied Informatics,* vol. V, no. 2, pp. 53-64, 2023.

[9] S. Rianti dan R. A. Supono, "Perbandingan Algoritma Edit Distance, Levenshtein Distance, Hamming Distance *Jaccard Similarity* dalam Mendeteksi String Matching," *JSI: Jurnal Sistem Informasi Universitas Suryadarma,* vol. X, no. 1, pp. 305-314, 2023.

[10] S. Rismayani, Nirwana, T. Darwansyah dan I. Mansyur, "Implementasi Algoritma Text Mining dan *Cosine Similarity* untuk Desain Sistem Aspirasi Publik Berbasis Mobile," *Komputika: Jurnal Sistem Komputer,* vol. IX, no. 28, p. 169–176, 2022.

[11] A. Manso, C. G. Marques, V. Alencar dan P. Santos, "Plagiarism Detection in Algorithms - a Case Study Using Algorithmi. Creative Commons License Attribution 4.0 International (CC BY 4.0)," *Journal of Information Technology and Computer Science,* vol. i, no. 1, pp. 1-6, 2020.

[12] M. Davoodifard, "Automatic Detection of Plagiarism in Writing. Studies in Applied Linguistics & TESOL at Teachers College," *Columbia University,* vol. XXI, no. 2, pp. 54-60, 2022.