

**DETEKSI KEMIRIPAN DOKUMEN EXECUTIVE SUMMARY  
PADA PROSES PENGAJUAN SKRIPSI DI FASILKOM-TI  
DENGAN IMPLEMENTASI *DOC2VEC* DAN *COSINE*  
*SIMILARITY***

**SKRIPSI**

**WARIDA HAFNI HASIBUAN**

**201402018**



**PROGRAM STUDI S1 TEKNOLOGI INFORMASI  
FAKULTAS ILMU KOMPUTER DAN TEKNOLOGI INFORMASI  
UNIVERSITAS SUMATERA UTARA**

**2025**

**DETEKSI KEMIRIPAN DOKUMEN EXECUTIVE SUMMARY PADA  
PROSES PENGAJUAN SKRIPSI DI FASILKOM-TI DENGAN  
IMPLEMENTASI *DOC2VEC* DAN  
*COSINE SIMILARITY***

**SKRIPSI**

Diajukan untuk melengkapi tugas dan memenuhi syarat memperoleh ijazah  
Sarjana Teknologi Informasi

**WARIDA HAFNI HASIBUAN**

**201402018**



**PROGRAM STUDI S1 TEKNOLOGI INFORMASI  
FAKULTAS ILMU KOMPUTER DAN TEKNOLOGI INFORMASI  
UNIVERSITAS SUMATERA UTARA**

**2025**

**PERSETUJUAN**

Judul :DETEKSI KEMIRIPAN DOKUMEN EXECUTIVE SUMMARY  
PADA PROSES PENGAJUAN SKRIPSI DI FASILKOM-TI  
DENGAN IMPLEMENTASI DOC2VEC DAN COSINE  
SIMILARITY

Kategori : SKRIPSI

Nama : WARIDAH HAFNI HASIBUAN

Nomor Induk Mahasiswa : 201402018

Program Studi : SARJANA (S1) TEKNOLOGI INFORMASI

Fakultas : ILMU KOMPUTER DAN TEKNOLOGI INFORMASI  
UNIVERSITAS SUMATERA UTARA

Medan, Juni 2025

Komisi Pembimbing

Pembimbing 2

Pembimbing 1

Ade Sarah Huzaifah, S.Kom., M.Kom.  
NIP.198506302018032001

Dr. Muhammad Anggia Muchtar, S.T., MM.IT.  
NIP.198001102008011010

Diketahui/disetujui oleh  
Program Studi S1 Teknologi Informasi  
Ketua,

Dedy Arisandi, S.T., M.Kom.,  
NIP. 197908312009121002

## **PERNYATAAN**

### **DETEKSI KEMIRIPAN DOKUMEN EXECUTIVE SUMMARY PADA PROSES PENGAJUAN SKRIPSI DI FASILKOM-TI DENGAN IMPLEMENTASI DOC2VEC DAN COSINE SIMILARITY**

#### **SKRIPSI**

Saya mengakui bahwasanya penelitian pada skripsi ini merupakan hasil karya saya sendiri, terkecuali beberapa kutipan dan ringkasan yang masing-masing telah disebutkan sumbernya.

Medan, Juni 2025

Warida Hafni Hasibuan  
201402018

## UCAPAN TERIMA KASIH

Puji syukur penulis ucapkan kepada Tuhan Yang Maha Esa, karena dengan izin dan rahmat-Nya serta kasih karunia-Nya, skripsi dengan judul “DETEKSI KEMIRIPAN DOKUMEN EXECUTIVE SUMMARY PADA PROSES PENGAJUAN SKRIPSI DI FASILKOM-TI DENGAN IMPLEMENTASI DOC2VEC DAN COSINE SIMILARITY” dapat diselesaikan oleh penulis pada waktu yang tepat. Penulisan skripsi ini sebagai salah satu syarat kelulusan dalam meraih gelar Sarjana Komputer pada Program Studi S1 Teknologi Informasi, Fakultas Ilmu Komputer dan Teknologi Informasi Universitas Sumatera Utara. Tentunya penulisan skripsi ini dapat diselesaikan dengan baik berkat dukungan moral dan materi dari berbagai pihak.

Sebagai ungkapan rasa syukur dan terima kasih atas doa, motivasi, serta bantuan yang diberikan oleh berbagai pihak selama proses penyusunan skripsi ini, penulis mengucapkan terima kasih sebesar-besarnya kepada:

1. Tuhan Yang Maha Esa atas berkat-Nya yang membimbing serta memberikan kemampuan kepada penulis sehingga dapat menyelesaikan skripsi ini hingga tahap akhir pada waktu yang tepat.
2. Ibu Dr. Maya Silvi Lydia, B.Sc., M.Sc., selaku Dekan Fasilkom-TI USU.
3. Bapak Dr. Muhammad Anggia Muchtar, S.T., MM.IT. selaku Dosen Pembimbing I saya yang telah meluangkan waktu kepada penulis dan terima kasih atas ilmu, kesabaran, dan bimbingan Bapak dalam penulisan skripsi ini.
4. Ibu Ade Sarah Huzaifah, S.Kom., M.Kom. selaku Dosen Pembimbing II saya yang telah meluangkan waktu kepada penulis dan terima kasih atas ilmu, kesabaran, dan bimbingan Ibu dalam penulisan skripsi ini.
5. Bapak Dedy Arisandi ST., M.Kom selaku Ketua Program Studi S1 Teknologi Informasi Universitas Sumatera Utara.
6. Bapak Ivan Jaya, S.Si., M.Kom., selaku Sekretaris Program Studi S1 Teknologi Informasi Universitas Sumatera Utara
7. Kepada Bapak / Ibu Dosen Penguji yang telah meluangkan waktu serta memberikan saran dan masukan untuk perbaikan skripsi ini agar lebih baik lagi.

8. Seluruh Dosen Program Studi S1 Teknologi Informasi yang telah memberikan ilmu pengetahuan dan pengalaman yang berharga kepada penulis.
9. Seluruh Staf dan Pegawai Fakultas Ilmu Komputer dan Teknologi Informasi yang sudah membantu penulis dalam segala urusan administrasi selama perkuliahan.
10. Seluruh teman-teman Angkatan 2020 yang telah berjuang bersama selama masa perkuliahan.
11. Seluruh pihak yang secara langsung maupun tidak langsung turut membantu penulis dalam menyelesaikan skripsi ini.

Semoga Tuhan Yesus memberkati seluruh pihak yang terlibat dalam penyelesaian skripsi ini. Selain itu, penulis berharap penelitian ini bisa berguna bagi orang-orang yang membacanya.

Medan, Juni 2025

Penulis,

Warida Hafni Hasibuan

## ABSTRAK

*Abstraknya akan aku isi kalau sudah siap semua isinya yah....*

*Jadi boleh isi dulu aja yg kurangnya*

*Ok.*

**Kata Kunci :**

**DETEKSI KEMIRIPAN DOKUMEN EXECUTIVE SUMMARY PADA  
PROSES PENGAJUAN SKRIPSI DI FASILKOM-TI DENGAN  
IMPLEMENTASI DOC2VEC DAN COSINE SIMILARITY**

**ABSTRACT**

*Keywords :*



## DAFTAR ISI

<b>PERSETUJUAN</b>	<b>iii</b>
<b>PERNYATAAN</b>	<b>iv</b>
<b>UCAPAN TERIMA KASIH</b>	<b>v</b>
<b>ABSTRAK</b>	<b>vii</b>
<b>ABSTRACT</b>	<b>viii</b>
<b>DAFTAR ISI</b>	<b>ix</b>
<b>DAFTAR TABEL</b>	<b>x</b>
<b>DAFTAR GAMBAR</b>	<b>xi</b>
<b>DAFTAR PSEUDOCODE</b>	<b>xii</b>
<b>BAB 1</b>	<b>1</b>
<b>PENDAHULUAN</b>	<b>1</b>
1.1 Latar Belakang	1
1.2 Rumusan Masalah	4
1.3 Tujuan Penelitian	4
1.4 Batasan Masalah	5
1.5 Manfaat Penelitian	5
1.6 Metodologi Penelitian	6
1.7 Sistematika Penulisan	7
<b>BAB 2</b>	<b>9</b>
<b>LANDASAN TEORI</b>	<b>9</b>
2.1 Natural Language Processing (NLP)	9
2.2 Doc2Vec	9
2.3 Cosine Similarity	10
2.4 Plagiarisme	11
2.5 Confusion Matrix	11

2.6 Penelitian Terdahulu	13
2.7 Perbedaan Penelitian	19
<b>BAB 3</b>	<b>20</b>
<b>ANALISIS DAN PERANCANGAN SISTEM</b>	<b>20</b>
3.1 Dataset	20
3.2 Analisis Sistem	21
3.3 Diagram Activity	31
3.4 Perancangan Antarmuka Sistem	33
3.4.1 Halaman Login	33
3.4.2 Halaman Register	33
3.4.3 Halaman Dashboard (Admin)	33
3.4.4 Halaman Upload Dokumen (Admin)	33
3.4.5 Halaman Daftar Dokumen (Admin)	34
3.4.6 Halaman Dashboard (User)	34
3.4.7 Halaman Upload File untuk Cek Similarity	34
3.4.8 Halaman Hasil Similaritas	34
3.4.9 Halaman Riwayat Dokumen	34
3.5 Metode Evaluasi	40
<b>BAB 4</b>	<b>41</b>
<b>IMPLEMENTASI DAN PENGUJIAN SISTEM</b>	<b>41</b>
4.1 Implementasi Sistem	41
4.1.1 Spesifikasi Perangkat Keras	41
4.1.2 Spesifikasi Perangkat Lunak	41
4.2 Implementasi Data	42
4.3 Implementasi Model	42
4.4 Evaluasi Testing Model	47
4.4.1 Metrik Evaluasi	48

4.4.2 Hasil Pengujian	48
4.4 Testing model	<b>Error! Bookmark not defined.</b>
4.5 Pengujian Sistem	<b>Error! Bookmark not defined.</b>
4.6 Implementasi Antarmuka Aplikasi	49
4.6.1 Halaman Login	49
4.6.2 Halaman Register	49
4.6.3 Dashboard Admin	50
4.6.4 Halaman Upload Dokumen (Admin)	51
4.6.5 Halaman Daftar Dokumen (Admin)	51
4.6.6 Dashboard User	52
4.6.7 Halaman Upload Dokumen (User)	52
4.6.8 Halaman Similarity exum	53
4.6.9 Halaman Hasil Similarity	53
4.6.10 Halaman Riwayat Dokumen	54
<b>BAB 5</b>	<b>56</b>
<b>KESIMPULAN DAN SARAN</b>	<b>56</b>
4.1 Kesimpulan	56
4.2 Saran	56
<b>DAFTAR PUSTAKA</b>	<b>57</b>

## DAFTAR TABEL

Tabel 2. 1 Penelitian Terdahulu	15
Tabel 3. 1 Jumlah Dataset	21
Tabel 3. 2 Text Cleaning	26
Tabel 3. 3 Case Folding	27
Tabel 3. 4 Tokenization	28
Tabel 3. 5 Lemmitization	29
Tabel 3. 6 Word Removal	30
Tabel 4. 1 Tabel Pengujian Model Terhadap Data Testing	47
Tabel 4. 2 Detail Rincian Hasil	48
Tabel 4. 3 <i>Confusion Matrix</i>	<b>Error! Bookmark not defined.</b>

## DAFTAR GAMBAR

Gambar 2. 1 Perbedaan cara kerja antara PV-DM dan PV-DBOW).	10
Gambar 2. 2 Ilustrasi proses Cosine Similarity	11
Gambar 2. 3 <i>Confusion Matrix</i> dalam matrix 2x2	<b>Error! Bookmark not defined.</b>
Gambar 3. 1 Arsitektur Umum	25
Gambar 3. 2 Activity Diagram	32
Gambar 3. 3 Halaman Login	35
Gambar 3. 4 Halaman Register	35
Gambar 3. 5 Halaman Dashboard Admin	36
Gambar 3. 6 Halaman Upload Dokumen Bagian Admin	36
Gambar 3. 7 Halaman Dokumen Skripsi Bagian Admin	37
Gambar 3. 8 Halaman Dashboard User	37
Gambar 3. 9 Halaman Upload Dokumen Bagian User	38
Gambar 3. 10 Halaman Hasil Simillaritas	38
Gambar 3. 11 Halaman Riwayat Dokumen	39
Gambar 4. 2 Tampilan Halaman Login	49
Gambar 4. 3 Tampilan Halaman Register	50
Gambar 4. 4 Tampilan Halaman Dashboard Admin	50
Gambar 4. 5 Tampilan Halaman Upload Dokumen Bagian Admin	51
Gambar 4. 6 Tampilan Halaman Daftar Dokumen Bagian Admin	51
Gambar 4. 7 Tampilan Halaman Dashboard User	52
Gambar 4. 8 Tampilan Halaman Upload Dokumen Bagian User	53
Gambar 4. 9 Tampilan Halaman Similarity Exum	53
Gambar 4. 10 Tampilan Halaman Dokumen Hasil Similarity	54
Gambar 4. 11 Tampilan Hasil Similarity	54
Gambar 4. 12 Tampilan Halaman Riwayat Dokumen	55

**DAFTAR PSEUDOCODE**

## **BAB 1**

### **PENDAHULUAN**

#### **1.1 Latar Belakang**

Di era digital yang terus berkembang, teknologi telah menjadi elemen fundamental dalam berbagai aspek kehidupan, termasuk dunia pendidikan. Transformasi digital yang terjadi secara masif telah membawa perubahan signifikan dalam cara kita mengakses, menyimpan, dan memproses informasi (Mustariani, 2023). Teknologi tidak hanya merambah dunia bisnis dan komunikasi, tetapi juga memainkan peran penting dalam meningkatkan kualitas pendidikan. Di tingkat global, digitalisasi telah menciptakan sistem yang lebih efisien dan efektif dalam mendukung proses pembelajaran, penelitian, dan evaluasi akademik (Siringoringo & Alfaridzi, 2024). Dalam konteks ini, tantangan yang dihadapi oleh institusi pendidikan tinggi semakin kompleks, terutama dalam mengelola volume dan keragaman dokumen akademik yang terus meningkat.

Di Fakultas Ilmu Komputer dan Teknologi Informasi Universitas Sumatera Utara, proses pengecekan kesamaan dokumen executive summary masih dilakukan secara manual oleh beberapa dosen penguji yang memiliki tanggung jawab di bidang penelitian yang berbeda. Masing-masing dosen melakukan penilaian berdasarkan referensi dan pertimbangan profesional di bidangnya, namun karena keterbatasan akses data serta belum adanya sistem terintegrasi, proses ini dapat menghasilkan variasi dalam evaluasi. Hal tersebut menjadikan proses pengecekan rawan inkonsistensi, terutama ketika jumlah dokumen yang dinilai cukup banyak dan melibatkan banyak pihak dalam waktu yang bersamaan.

Selain itu, proses pengecekan kesamaan dokumen executive summary juga masih menghadapi kendala besar dalam hal ketersediaan data referensi yang memadai.

Salah satu hambatan utama yang dihadapi adalah tidak tersedianya akses penuh terhadap keseluruhan arsip dokumen skripsi alumni secara terpusat dan digital. Dalam praktiknya, dosen penguji hanya dapat mengakses sebagian kecil dokumen terdahulu yang tersedia, sehingga proses verifikasi terhadap kemungkinan adanya kemiripan topik menjadi terbatas cakupannya. Akibatnya, terdapat potensi kemiripan atau duplikasi tema penelitian yang tidak terdeteksi secara dini. Jika hal ini terus berlanjut, maka bukan tidak mungkin akan berdampak pada menurunnya kualitas hasil penelitian mahasiswa dan kredibilitas akademik lembaga. Oleh karena itu, diperlukan pendekatan sistemik berbasis teknologi yang dapat membantu mendukung proses pengecekan dokumen secara lebih luas, konsisten, dan terintegrasi dengan basis data skripsi yang tersedia.

Plagiarisme dan duplikasi penelitian merupakan masalah serius yang dapat mengurangi integritas institusi pendidikan dan menurunkan kepercayaan terhadap kualitas akademik di Indonesia. Selain merusak reputasi pendidikan tinggi, plagiarisme menghambat pengembangan ilmu pengetahuan yang orisinal. Dengan meningkatnya tuntutan terhadap kualitas karya ilmiah, institusi pendidikan perlu memiliki sistem deteksi kesamaan dokumen yang andal untuk memastikan karya mahasiswa memenuhi standar akademik. Sistem ini penting untuk menjaga integritas akademik serta mendorong mahasiswa menghasilkan penelitian yang orisinal dan berkualitas, sehingga dapat mendukung terciptanya penelitian yang bermutu dan berintegritas (Wibowo, 2012).

Dalam konteks ini, teknologi pemrosesan bahasa alami (*Natural Language Processing* atau *NLP*) memainkan peran penting dalam pengembangan sistem deteksi kesamaan dokumen. Salah satu metode yang sangat efektif dalam mendeteksi kesamaan dokumen adalah *Doc2Vec*, sebuah teknik *embedding* dokumen yang mampu merepresentasikan dokumen sebagai vektor dalam ruang multidimensi. Vektor ini kemudian digunakan untuk menghitung kesamaan antar dokumen dengan metode *Cosine Similarity*. *Doc2Vec* dikembangkan untuk mengatasi kekurangan metode berbasis kata (*word-based methods*) seperti *TF-IDF* yang tidak bisa menangkap makna semantik dari sebuah kalimat secara menyeluruh. Sementara *Cosine Similarity* telah terbukti sebagai metode yang sangat andal dalam mengukur kesamaan antara dua vektor.

Berbagai penelitian telah dilakukan untuk mengembangkan sistem deteksi plagiarisme dan kesamaan teks menggunakan beragam metode dan teknologi. Pawestri



& Suyanto (2024) dalam penelitiannya yang berjudul "Analisis Perbandingan Metode Similarity untuk Kemiripan Dokumen Bahasa Indonesia pada Deteksi Kemiripan Teks Bahasa Indonesia" menunjukkan bahwa *Cosine Similarity* memiliki tingkat akurasi yang tinggi, mencapai 98% dalam mendeteksi kemiripan dokumen berbahasa Indonesia, lebih unggul dibandingkan metode lain seperti *Jaccard* dan *Euclidean Distance*. Ansis et al. (2024) dalam penelitian berjudul "Deteksi Plagiat Tesis Berbahasa Indonesia Menggunakan Metode Cosine Similarity" mendukung temuan ini, dengan hasil penelitian yang menunjukkan bahwa *Cosine Similarity* mencapai akurasi 96,63%, jauh lebih tinggi dibandingkan *Jaccard Similarity*.

Wadekar et al. (2021) dalam penelitian berjudul "Plagiarism Detection with Paraphrase Recognizer Using Deep Learning" mengembangkan sistem deteksi plagiarisme menggunakan kombinasi *Doc2Vec*, *Siamese LSTM*, dan *CNN*, yang mencapai akurasi sebesar 97,26% dalam mendeteksi berbagai bentuk plagiarisme, termasuk parafrase. Penelitian ini menunjukkan efektivitas *Doc2Vec* saat dikombinasikan dengan model *deep learning* untuk deteksi kemiripan teks yang lebih mendalam. Sementara itu, Cahyono (2020) dalam penelitiannya yang berjudul "Model Perbandingan Dokumen Karya Ilmiah Dengan Metode Fragmentasi Menggunakan Algoritma Kesamaan Dokumen Doc2Vec" menggabungkan *Doc2Vec* dengan teknik fragmentasi, yang terbukti mampu meningkatkan akurasi dan efisiensi komputasi dalam mendeteksi plagiarisme pada dokumen akademik berbahasa Indonesia.

Berdasarkan penelitian-penelitian yang telah dilakukan, teknologi seperti *Doc2Vec* dan *Cosine Similarity* menunjukkan potensi yang signifikan dalam meningkatkan efisiensi dan akurasi deteksi kesamaan dokumen akademik. *Doc2Vec* memungkinkan pemetaan dokumen ke dalam vektor yang menangkap makna semantik secara lebih mendalam, sedangkan *Cosine Similarity* menyediakan metode yang andal untuk menghitung kesamaan antar dokumen berdasarkan representasi vektor tersebut. Dengan penerapan sistem otomatis yang memanfaatkan teknologi ini, proses pengecekankesamaan dokumen dapat dilakukan dengan lebih cepat dan akurat, tanpa ketergantungan yang berlebihan pada tenaga manusia. Sistem ini tidak hanya mampu melakukan pemeriksaan kesamaan dalam aspek struktur dan konten dokumen, tetapi juga dapat menangkap makna semantik antar dokumen dengan lebih baik, yang sulit dicapai melalui metode manual yang selama ini digunakan. Penggunaan teknologi ini memungkinkan evaluasi yang lebih transparan, objektif, dan adil bagi setiap mahasiswa,

sekaligus membantu menjaga integritas akademik di lingkungan pendidikan tinggi. Dengan meminimalkan risiko kesalahan subjektif yang sering muncul dalam pengecekan manual, sistem otomatis ini diharapkan dapat menciptakan suasana akademik yang lebih kondusif untuk pengembangan penelitian yang orisinal.

Oleh karena itu, penelitian ini bertujuan untuk mengembangkan sistem deteksi kesamaan *executive summary* skripsi di Fakultas Ilmu Komputer dan Teknologi Informasi (Fasilkom-TI) yang lebih efisien dan akurat. Dengan adanya sistem ini, diharapkan proses evaluasi orisinalitas karya ilmiah mahasiswa dapat berlangsung lebih cepat dan objektif, sehingga mendukung tugas kepala bidang dalam memeriksa kesamaan dokumen secara menyeluruh. Selain itu, implementasi sistem otomatis ini diharapkan mampu memberikan kontribusi positif terhadap kualitas penelitian yang dihasilkan oleh

mahasiswa. Dengan pendekatan yang lebih sistematis dan berbasis teknologi, penelitian ini diharapkan dapat memenuhi tuntutan akademik yang semakin tinggi, sekaligus memfasilitasi mahasiswa dalam menghasilkan karya ilmiah yang inovatif dan berkualitas.

## **1.2 Rumusan Masalah**

Proses pengecekan kesamaan *executive summary* di Fakultas Ilmu Komputer dan Teknologi Informasi Universitas Sumatera Utara belum berjalan optimal karena keterbatasan akses terhadap arsip digital skripsi alumni yang terpusat dan lengkap. Kondisi ini membuat verifikasi topik oleh dosen penguji hanya bergantung pada dokumen fisik terbatas atau ingatan personal, sehingga rawan terjadi duplikasi topik yang tak terdeteksi. Ketidaktersediaan data historis yang memadai turut memperlemah proses validasi orisinalitas penelitian mahasiswa. Maka dari itu, dibutuhkan sistem otomatis berbasis teknologi yang dapat mendeteksi kemiripan dokumen secara efisien, akurat, dan konsisten sebagai upaya menjaga integritas serta mutu akademik secara menyeluruh.

## **1.3 Tujuan Penelitian**

Penelitian ini bertujuan mengembangkan sistem deteksi kesamaan dokumen *executive summary* yang efisien dan akurat dengan menggunakan algoritma Doc2Vec untuk representasi dokumen dan Cosine Similarity untuk mengukur kemiripan antar dokumen.

Sistem ini diharapkan dapat membantu dosen penguji dalam verifikasi topik skripsi dan meningkatkan kualitas evaluasi akademik di Fakultas Ilmu Komputer dan Teknologi Informasi Universitas Sumatera Utara.

#### **1.4 Batasan Masalah**

Batasan masalah pada penelitian ini memiliki batasan sebagai berikut:

1. Penelitian ini hanya melibatkan dokumen berbahasa Indonesia dan tidak mencakup jenis dokumen akademik lain seperti proposal penelitian atau laporan ilmiah.
2. Data skripsi alumni yang digunakan sebagai referensi pengecekan kemiripan hanya mencakup sebagian arsip yang tersedia dan dapat diakses oleh peneliti.
3. Dokumen Executive Summary yang menjadi objek deteksi kemiripan merupakan berkas yang diunggah oleh dosen penguji melalui sistem yang dikembangkan.
4. Sistem dikembangkan untuk mendeteksi kemiripan teks antara dokumen Executive Summary yang diunggah dengan dokumen skripsi alumni yang tersedia sebagai referensi.
5. Deteksi kemiripan hanya dilakukan pada konten teks, tidak mencakup elemen non-teks seperti gambar, tabel, atau diagram yang terdapat dalam dokumen.

#### **1.5 Manfaat Penelitian**

Dari penelitian ini diperoleh beberapa manfaat antara lain :

1. Meningkatkan efisiensi dan akurasi dalam pengecekan kesamaan dokumen *executive summary* melalui sistem otomatis.
2. Mendukung orisinalitas dan kualitas akademik mahasiswa melalui sistem deteksikesamaan yang lebih andal dan konsisten.
3. Mengurangi beban kerja manual bagi staf akademik dalam proses penilaianskripsi.

## 1.6 Metodologi Penelitian

Tahapan-Tahapan yang akan dilakukan pada penelitian ini adalah:

### 1. Analisis Permasalahan

Langkah pertama dalam penelitian ini adalah mengidentifikasi dan menganalisis permasalahan yang terjadi dalam proses pengecekan manual terhadap kesamaan dokumen *executive summary* di Fakultas Ilmu Komputer dan Teknologi Informasi Universitas Sumatera Utara. Melalui analisis ini, diidentifikasi sejumlah tantangan utama, seperti waktu pengecekan yang lama, adanya potensi kesalahan subjektif, serta beban kerja yang tinggi pada dosen. Hasil dari analisis ini akan digunakan sebagai dasar untuk merancang solusi yang lebih efisien dalam mendeteksi kesamaan dokumen.

### 2. Studi Literatur

Setelah melakukan analisis permasalahan, dilanjutkan dengan studi literatur untuk mencari solusi teknologi yang sesuai. Pada tahap ini, berbagai referensi yang relevan dikumpulkan dari sumber seperti buku, jurnal ilmiah, dan artikel akademik yang berkaitan dengan *Natural Language Processing* (NLP), khususnya tentang model *Doc2Vec* dan algoritma *Cosine Similarity*. Hasil dari studi ini membantu dalam merumuskan pendekatan yang tepat untuk pengembangan sistem yang lebih efektif.

### 3. Pengumpulan dan Pemrosesan Dataset

Pada tahap pengumpulan dan pemrosesan dataset, peneliti mengumpulkan dokumen skripsi alumni Fakultas Ilmu Komputer dan Teknologi Informasi Universitas Sumatera Utara yang tersedia dan dapat diakses, meskipun terbatas pada sebagian dokumen saja. Setelah data terkumpul, dilakukan proses praproses untuk membersihkan teks dari elemen yang kurang relevan seperti gambar, tabel, dan nomor halaman, yang meliputi tokenisasi, normalisasi, serta transformasi teks menjadi representasi vektor yang akan diolah menggunakan model *Doc2Vec*. Selanjutnya, penyesuaian hyperparameter model seperti ukuran vektor (*vector size*), jendela konteks kata (*window size*), dan jumlah iterasi pelatihan (*epochs*) dilakukan secara seksama untuk mengoptimalkan performa model dalam mendeteksi kemiripan dokumen.

#### 4. Perancangan Sistem dan Model

Berdasarkan hasil analisis dan studi literatur, dirancang sistem deteksi kesamaan dokumen yang menggunakan model *Doc2Vec* untuk menghasilkan representasi vektor dari dokumen. Algoritma *Cosine Similarity* digunakan untuk mengukur kesamaan antara dokumen *executive summary* yang baru dengan skripsi yang telah ada dalam database. Desain sistem juga mencakup arsitektur teknis dan antarmuka pengguna yang memungkinkan pengunggahan *executive summary* secara langsung untuk dianalisis.

#### 5. Implementasi Sistem

Setelah desain sistem selesai, dilakukan tahap implementasi di mana model *Doc2Vec* diintegrasikan untuk memproses teks dokumen menjadi vektor. *Cosine Similarity* digunakan untuk menghitung kesamaan antar dokumen berdasarkan vektor yang dihasilkan. Sistem ini dibangun untuk memberikan hasil kesamaan dokumen dalam bentuk persentase yang dapat digunakan oleh staf akademik dalam mengevaluasi kesamaan antara dokumen yang diajukan.

#### 6. Pengujian Sistem

Pengujian dilakukan setelah implementasi sistem untuk memastikan bahwa sistem bekerja sesuai dengan spesifikasi yang diharapkan. Dataset yang sama digunakan untuk uji coba, di mana *executive summary* yang baru dibandingkan dengan skripsi yang sudah ada dalam database. Evaluasi kinerja sistem dilakukan dengan menggunakan metrik *F1-score* untuk mengukur akurasi sistem dalam mendeteksi kesamaan teks.

#### 7. Dokumentasi dan Penyusunan Laporan

Tahap terakhir adalah penyusunan dokumentasi yang mendetail tentang proses penelitian, mulai dari analisis permasalahan hingga hasil pengujian sistem. Dokumentasi ini digunakan untuk menyusun laporan akhir dalam bentuk skripsi, yang mencakup evaluasi sistem dan rekomendasi pengembangan lebih lanjut.

### 1.7 Sistematika Penulisan

Adapun untuk sistematika penulisan pada penelitian ini yaitu terdiri atas 5 bab, yakni:

Bab 1: Pendahuluan

Bab ini berisi pengantar mengenai penelitian yang dilakukan, meliputi latar belakang

permasalahan, rumusan masalah, tujuan penelitian, batasan masalah, manfaat penelitian, metodologi penelitian, dan sistematika penulisan. Bab ini memberikan gambaran umum tentang pentingnya topik yang diteliti serta alasan penggunaan metode dan pendekatan yang diambil dalam penelitian ini.

## Bab 2: Landasan Teori

Pada bab ini dijelaskan teori-teori yang mendasari penelitian, termasuk konsep-konsep yang terkait dengan *Natural Language Processing* (NLP), model *Doc2Vec*, algoritma *Cosine Similarity*, serta metode-metode pendeteksian kesamaan teks. Kajian literatur dan penelitian terdahulu yang relevan juga dipaparkan sebagai dasar ilmiah bagi pengembangan sistem yang diusulkan.

## Bab 3: Analisis dan Perancangan Sistem

Bab ini menjelaskan hasil analisis terhadap permasalahan yang terjadi dalam proses pengecekan kesamaan dokumen *executive summary* secara manual. Bab ini juga merinci perancangan sistem yang diusulkan, termasuk arsitektur sistem, alur kerja sistem, serta spesifikasi teknis dari sistem deteksi kesamaan dokumen berbasis *Doc2Vec* dan *Cosine Similarity*.

## Bab 4: Implementasi dan Pengujian Sistem

Bab ini menguraikan proses implementasi sistem berdasarkan desain yang telah dirancang pada bab sebelumnya. Sistem diuji untuk memastikan bahwa fungsinya sesuai dengan spesifikasi yang telah ditetapkan. Pengujian sistem dilakukan dengan menggunakan dataset yang telah dipersiapkan dan dievaluasi berdasarkan metrik *F1-score* untuk menilai kinerja sistem dalam mendeteksi kesamaan dokumen.

## Bab 5: Kesimpulan dan Saran

Bab terakhir ini menyajikan kesimpulan dari hasil penelitian, mencakup temuan utama serta pencapaian tujuan penelitian. Selain itu, diberikan pula saran untuk pengembangan sistem lebih lanjut, termasuk potensi pengembangan dan penerapan teknologi yang telah dihasilkan dalam penelitian ini di masa mendatang.

## **BAB 2**

### **LANDASAN TEORI**

#### **2.1 Natural Language Processing (NLP)**

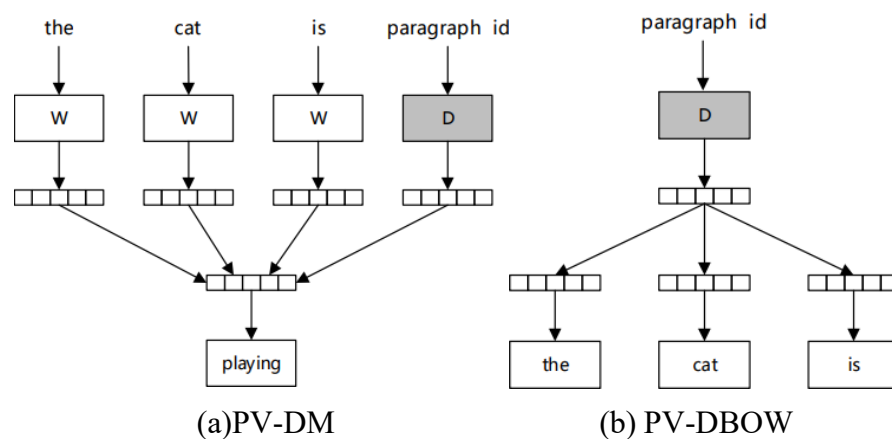
*Natural Language Processing* (NLP) merupakan cabang dari kecerdasan buatan (*Artificial Intelligence*) yang berfokus pada interaksi antara komputer dan bahasa manusia. NLP mencakup berbagai metode dan teknik untuk memungkinkan mesin memahami, menafsirkan, dan memproses bahasa alami manusia, baik dalam bentuk teks maupun suara (Jurafsky & Martin, 2024). Dalam konteks penelitian ini, NLP memainkan peran penting dalam menganalisis teks akademik, terutama untuk mendeteksi kesamaan antara dokumen yang diajukan oleh mahasiswa.

Teknik-teknik dasar dalam NLP meliputi *tokenisasi*, *stemming*, *lemmatization*, dan penghapusan *stop words*, yang membantu dalam praproses teks sebelum digunakan dalam model pembelajaran mesin. Preprocessing ini sangat penting untuk mengurangi kompleksitas data dan meningkatkan akurasi analisis teks. Dalam penelitian ini, NLP akan digunakan sebagai landasan utama untuk mendeteksi kemiripan dokumen akademik dengan memanfaatkan embedding teks yang dihasilkan oleh model *Doc2Vec*.

#### **2.2 Doc2Vec**

*Doc2Vec* adalah teknik embedding dokumen yang dikembangkan dari *Word2Vec*, yang memungkinkan representasi dokumen dalam bentuk vektor berdimensi tinggi. *Doc2Vec* memungkinkan pemodelan dokumen dengan lebih baik karena mampu menangkap konteks dan makna semantik dari teks secara keseluruhan, bukan hanya kata-kata individual seperti *Word2Vec* (Le et al., 2014).

Ada dua pendekatan utama dalam *Doc2Vec*: Distributed Memory Model of Paragraph Vectors (PV-DM) dan Distributed Bag of Words (PV-DBOW). PV-DM mempertahankan urutan kata untuk memahami konteks kalimat, sementara PV-DBOW tidak mempertahankan urutan tetapi tetap dapat menangkap informasi semantik dari dokumen. *Doc2Vec* telah banyak digunakan dalam berbagai aplikasi teks, termasuk klasifikasi dokumen, analisis sentimen, dan tentu saja, deteksi kesamaan dokumen. Dalam penelitian ini, *Doc2Vec* digunakan untuk menghasilkan embedding teks dari skripsi yang diajukan sehingga bisa dibandingkan dengan dokumen lain secara efisien.



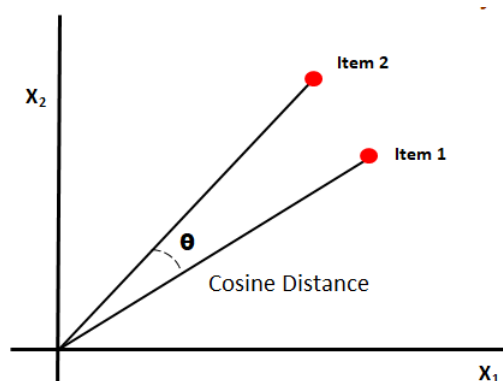
**Gambar 2. 1 Perbedaan cara kerja antara PV-DM dan PV-DBOW).**

### 2.3 Cosine Similarity

*Cosine Similarity* adalah salah satu metode yang paling umum digunakan untuk mengukur kesamaan antar dua vektor dalam ruang vektor. Metrik ini mengukur sudut cosinus antara dua vektor, di mana semakin kecil sudutnya, semakin mirip kedua vektor tersebut (Wahyuni et al., 2017).

Cosine Similarity sering digunakan dalam aplikasi deteksi plagiarisme atau kemiripan dokumen karena mampu mengabaikan panjang dokumen dan fokus pada orientasi vektor dalam ruang multidimensi. Dalam penelitian ini, Cosine Similarity digunakan untuk mengukur kesamaan antara embedding teks yang dihasilkan oleh Doc2Vec. Penerapan Cosine Similarity ini memberikan hasil yang cepat dan akurat, terutama dalam tugas-tugas analisis teks yang memerlukan perbandingan kemiripan antara dokumen yang berbeda seperti pada Gambar 2.2.





**Gambar 2. 2 Ilustrasi proses Cosine Similarity**

## 2.4 Plagiarisme

*Plagiarisme* adalah tindakan mencuri ide, kata-kata, atau karya orang lain dan menyajikannya sebagai milik sendiri tanpa memberikan pengakuan yang semestinya. Dalam dunia akademik, *plagiarisme* adalah pelanggaran serius yang dapat merusak reputasi institusi dan integritas penulis (Sutrisno et al., n.d.).

Plagiarisme dapat muncul dalam berbagai bentuk, dari pengambilan ide hingga penyalinan teks secara verbatim tanpa kutipan yang sesuai. Dengan meningkatnya jumlah dokumen akademik yang dihasilkan, plagiarisme menjadi ancaman yang nyata, dan oleh karena itu, pendeteksian plagiarisme dengan teknologi modern seperti NLP dan machine learning menjadi semakin penting. Dalam konteks penelitian ini, deteksi plagiarisme dilakukan dengan mengidentifikasi kemiripan antara dokumen yang diajukan oleh mahasiswa dan karya-karya yang sudah ada, yang dapat dicegah dengan bantuan teknologi seperti Doc2Vec dan Cosine Similarity.

## 2.5 Hyperparameter Tuning

Tentunya menyesuaikan hyperparameter dengan tepat dapat meningkatkan kinerja model, mempercepat proses pelatihan, dan mencegah masalah seperti *overfitting* atau *underfitting*. Seperti dalam algoritma pohon keputusan, menentukan jumlah pohon yang optimal adalah salah satu aspek penting dalam hyperparameter tuning (Maulid, 2024). Dalam studi ini, Hyperparameter tuning adalah proses menyesuaikan parameter tertentu dalam model pembelajaran mesin yang ditetapkan sebelum pelatihan dimulai. Tidak seperti parameter model yang dipelajari selama pelatihan,

hyperparameter harus diatur secara manual dan dapat mempengaruhi kinerja model secara signifikan (Ivan Belcic, 2024). Berikut ini Pseduocode tahap Hyperparameter Tuning:

```

1. Definisikan parameter grid yang berisi kombinasi nilai:
    - vector_size: [50, 100, 150]
    - window: [3, 5, 7]
    - dm: [0, 1]
    - epochs: [20, 30, 40]

2. Siapkan list kosong untuk menyimpan hasil evaluasi:
    - results = []

3. Lakukan loop untuk setiap kombinasi parameter:
    FOR each vector_size in param_grid["vector_size"]:
        FOR each window in param_grid["window"]:
            FOR each dm in param_grid["dm"]:
                FOR each epochs in param_grid["epochs"]:

                    Tampilkan kombinasi parameter yang sedang
                    diuji

                    Inisialisasi model Doc2Vec dengan
                    parameter:
                        - vector_size
                        - window
                        - dm
                        - learning rate awal dan akhir
                        - min_count = 2

                    Bangun vocabulary dari data train_tagged

                    TRAINING:
                    FOR epoch in jumlah_epochs:
                        latih model 1 epoch
                        kurangi learning rate
                        set ulang min_alpha

```

```

    Evaluasi model:
        - Lakukan infer_vector untuk semua
dokumen validasi/test
        - Hitung similarity antar dokumen
        - Ambil rata-rata skor similarity

    Simpan hasil ke dalam list results:
        - parameter kombinasi
        - nilai similarity rata-rata

4. Setelah semua kombinasi dicoba:
    - Ambil kombinasi parameter dengan skor similarity terbaik
    - Simpan sebagai parameter terbaik

```

#### Pseudocode 2.1 Hyperparameter Tuning

## 2.6 Penelitian Terdahulu

Berbagai penelitian telah dilakukan untuk menilai keefektifan metode *similarity* dalam deteksi kesamaan dokumen akademik, khususnya untuk mendeteksi *plagiarisme* dan kemiripan teks. Salah satunya adalah penelitian oleh Pawestri dan Suyanto (2024) yang berjudul *Analisis Perbandingan Metode Similarity untuk Kemiripan Dokumen Bahasa Indonesia pada Deteksi Kemiripan Teks Bahasa Indonesia*. Dalam studi ini, mereka membandingkan performa beberapa metode seperti *Doc2Vec*, *Jaccard Coefficient*, *Cosine Similarity*, dan *Euclidean Distance*. Hasil penelitian menunjukkan bahwa *Cosine Similarity* unggul dengan akurasi mencapai 98%, presisi 84%, recall 95%, dan *F1-score* 89%, yang menandakan kemampuannya yang lebih tinggi dalam mendeteksi kemiripan dokumen berbahasa Indonesia dibandingkan metode lainnya (Pawestri & Suyanto, 2024).

Penelitian lain oleh Ansis, Listyaningsih, dan Soetanto (2024) dalam studi berjudul *Deteksi Plagiat Tesis Berbahasa Indonesia Menggunakan Metode Cosine Similarity* mengevaluasi efektivitas *Cosine Similarity* dan *Jaccard Similarity* dalam mendeteksi kemiripan antar dokumen tesis. Hasilnya menunjukkan bahwa *Cosine Similarity* memiliki tingkat akurasi sebesar 96,63%, jauh lebih efektif dibandingkan *Jaccard Similarity* yang hanya mencapai 50,5%.

Selanjutnya, Wadekar et al. (2021) dalam penelitian mereka berjudul *Plagiarism Detection with Paraphrase Recognizer Using Deep Learning* mengusulkan kerangka deteksi *plagiarisme* dengan kombinasi *Doc2Vec*, *Siamese LSTM*, dan *CNN*. Metode ini terdiri dari tiga lapisan utama: lapisan pra-pemrosesan yang menghasilkan representasi kata menggunakan *Doc2Vec*, lapisan pembelajaran yang memanfaatkan *Siamese LSTM* untuk pola kemiripan, serta *CNN* untuk deteksi kesamaan dokumen. Dengan akurasi mencapai 97,26%, penelitian ini menunjukkan potensi besar dalam mendeteksi berbagai jenis *plagiarisme*, termasuk parafrase.

Dalam penelitian Resta, Aditya, dan Purwiantono (2021) yang berjudul *Plagiarism Detection in Students' Theses Using The Cosine Similarity Method*, *Cosine Similarity* dan *TF-IDF* diterapkan untuk mendeteksi kesamaan judul dan abstrak dalam tesis mahasiswa. Penelitian ini menunjukkan bahwa metode *Cosine Similarity* efektif dalam mendeteksi kesamaan dokumen, menegaskan potensinya dalam evaluasi teks akademik dengan bantuan teknik pra-pemrosesan.

Selain itu, Cahyono (2020) dalam penelitian *Model Perbandingan Dokumen Karya Ilmiah Dengan Metode Fragmentasi Menggunakan Algoritma Kesamaan Dokumen Doc2Vec* mengusulkan penggabungan teknik fragmentasi dengan *Doc2Vec* untuk meningkatkan akurasi dan efisiensi deteksi *plagiarisme* dalam dokumen akademik berbahasa Indonesia. Teknik fragmentasi memungkinkan analisis bagian-bagian kecil dari dokumen, sehingga memberikan hasil yang lebih optimal dibandingkan dengan metode *Doc2Vec* sekuensial.

Penelitian oleh Setha dan Aliane (2022) dalam studi *Enhancing Automatic Plagiarism Detection Using Doc2Vec* menerapkan *Doc2Vec* dalam mendeteksi kemiripan semantik pada teks berbahasa Inggris dan Arab menggunakan korpus PAN dan AraPlagDet. Hasil penelitian menunjukkan bahwa model ini memiliki akurasi tinggi dalam mendeteksi *plagiarisme* dalam kedua bahasa tersebut, yang menunjukkan potensi penerapannya dalam berbagai bahasa.

Terakhir, Pratama et al. (2019) dalam penelitian *Deteksi Plagiarisme pada Artikel Jurnal Menggunakan Metode Cosine Similarity* meneliti penggunaan *Cosine Similarity* bersama dengan teknik *Grabbing Data* dan *PDF Extractor*. Dari hasil uji coba, tingkat kemiripan mencapai 13% menggunakan *Cosine Similarity*, dengan nilai recall sebesar 8%, memberikan gambaran tentang efektivitas metode ini dalam konteks kemiripan dokumen jurnal.

**Tabel 2. 1 Penelitian Terdahulu**

No	Peneliti	Tahun	Judul	Keterangan
1	Sheraton Pawestri, Yohanes Suyanto	2024	“Analisis Perbandingan Metode Similarity untuk Kemiripan Dokumen Bahasa Indonesia pada Deteksi Kemiripan Teks Bahasa Indonesia”	Penelitian ini bertujuan untuk menilai performa algoritma <i>Doc2Vec</i> dibandingkan dengan metode <i>similarity</i> lain seperti <i>Jaccard Coefficient</i> , <i>Cosine Similarity</i> , dan <i>Euclidean Distance</i> dalam mendeteksi kemiripan antar dokumen berbahasa Indonesia. Hasilnya adalah <i>Cosine Similarity</i> lebih unggul dibandingkan metode lainnya, dengan akurasi 98%, presisi 84%, recall 95%, dan <i>F1-score</i> 89%.
2	Syukry Ansis, Endang Palupi Listyaningsih, Prof. Dr. Ir. Hari Soetanto, S.kom,	2024	“Deteksi Plagiat Tesis Berbahasa Indonesia Menggunakan Metode Cosine Similarity”	Penelitian ini bertujuan untuk mengevaluasi efektivitas dua metode dengan mendeteksi kemiripan antara dua dokumen tesis dengan membandingkan metode <i>Cosine Similarity</i> dan <i>Jaccard Similarity</i> . Hasilnya

				menunjukkan bahwa <i>Cosine Similarity</i> lebih efektif, dengan akurasi hasil mencapai 96,63% dan <i>Jaccard Similarity</i> 50,5%.
--	--	--	--	---

Tabel 3.1 Penelitian Terdahulu (Lanjutan)

No	Peneliti	Tahun	Judul	Keterangan
3	Yogesh Wadekar, Tushar Shendge, Manali Dhokale, Vaishnavi Ohol, Prof. Sagar Dhanake	2021	“Plagiarism Detection with Paraphrase Recognizer Using Deep Learning”	Penelitian ini mengusulkan kerangka kerja deteksi plagiarisme yang memanfaatkan model pembelajaran mendalam untuk meningkatkan efektivitas deteksi plagiarisme. Sistem ini terdiri dari tiga lapisan: prapemrosesan dengan embedding kata, pembelajaran, dan deteksi. Dengan pendekatan ini, sistem berhasil mencapai akurasi 97,26% dalam mengidentifikasi berbagai jenis plagiarisme, termasuk yang melibatkan penggantian kata sederhana dan modifikasi struktur frasa.
4	Oppi Anda Resta, Addin Aditya, Febry Eka Purwiantono	2021	“Plagiarism Detection in Students' Theses Using The Cosine	Penelitian ini bertujuan untuk mengukur kesamaan judul dan abstrak dengan menghitung kesamaan dokumen

			Similarity Method”	menggunakan <i>TF</i> dan <i>IDF</i> untuk mendapatkan skor <i>Cosine Similarity</i> . Hasilnya menunjukkan bahwa penggunaan metode <i>Cosine Similarity</i> , bersama dengan teknik <i>preprocessing</i>
--	--	--	--------------------	---

Tabel 3.1 Penelitian Terdahulu (Lanjutan)

No	Peneliti	Tahun	Judul	Keterangan
				dan <i>TF-IDF</i> , terbukti efektif dalam mendeteksi kesamaan antara tesis yang diajukan dan karya-karya yang sudah ada.
5	Stefanus Christian Cahyono	2020	“Model Perbandingan Dokumen Karya Ilmiah Dengan Metode Fragmentasi Menggunakan Algoritma Kesamaan Dokumen Doc2vec”	Penelitian ini merancang sistem untuk mendeteksi <i>plagiarisme</i> dalam karya ilmiah berbahasa Indonesia dengan menggunakan metode fragmentasi dan algoritma <i>Doc2Vec</i> . Hasil penelitian menunjukkan bahwa penggabungan model fragmentasi dengan metode <i>Doc2Vec</i> memberikan akurasi dan kecepatan komputasi yang lebih optimal dibandingkan metode <i>Doc2Vec</i> sekuensial. Model ini mampu memberikan rekomendasi dokumen literatur yang serupa dengan dokumen terkait.

6	Imene Seta, Hassina Aliane	2022	“Enhancing Automatic Plagiarism Detection Using Doc2Vec”	Penelitian ini menggunakan model <i>Doc2Vec</i> untuk mendeteksi kemiripan semantik pada teks dalam bahasa Inggris dan Arab menggunakan korpus PAN
---	-------------------------------	------	--	--

Tabel 3.1 Penelitian Terdahulu (Lanjutan)

No	Peneliti	Tahun	Judul	Keterangan
				dan AraPlagDet, dengan hasil yang menunjukkan akurasi tinggi dalam mendeteksi <i>plagiarisme</i> dalam kedua bahasa tersebut.
7	Rito Putriwana Pratama, Muhammad Faisal Ajib Hanani	2019	“Deteksi Plagiarisme pada Artikel Jurnal Menggunakan Metode Cosine Similarity”	Penelitian ini mengimplementasikan sistem deteksi plagiarisme dengan menggunakan algoritma Cosine Similarity untuk membandingkan artikel jurnal yang diunggah dengan dokumen yang ada di repositori. Hasilnya adalah berdasarkan uji coba mencari nilai kemiripan artikel jurnal, diperoleh nilai sebesar 13% menggunakan metode <i>Cosine Similarity</i> , sedangkan untuk nilai recall sebesar 8%.



## 2.7 Perbedaan Penelitian

Penelitian ini menerapkan algoritma Doc2Vec sebagai teknik embedding dokumen dan metode Cosine Similarity untuk mengukur tingkat kemiripan antar dokumen executive summary skripsi. Berbeda dengan penelitian Pawestri dan Suyanto (2024) yang membandingkan berbagai metode similarity pada dokumen teks secara umum, penelitian ini fokus pada deteksi kemiripan dokumen akademik berbahasa Indonesia dengan pendekatan embedding dokumen secara menyeluruh. Ansis et al. (2024) dan Resta et al. (2021) menggunakan Cosine Similarity dan TF-IDF untuk mendeteksi plagiarisme pada tesis dan abstrak, sedangkan penelitian ini menggabungkan Doc2Vec dan Cosine Similarity untuk menangkap kemiripan semantik secara lebih mendalam. Selain itu, penelitian ini berbeda dari Wadekar et al. (2021) yang menggunakan model deep learning seperti Siamese LSTM dan CNN, dengan pendekatan yang lebih sederhana namun efisien menggunakan Doc2Vec. Penelitian Cahyono (2020) mengkombinasikan teknik fragmentasi dengan Doc2Vec, sedangkan penelitian ini menggunakan representasi embedding penuh dokumen untuk deteksi kemiripan secara komprehensif. Dengan demikian, penelitian ini menawarkan solusi khusus yang menitikberatkan pada peningkatan efisiensi dan akurasi dalam mendeteksi kemiripan executive summary di Fakultas Ilmu Komputer dan Teknologi Informasi Universitas Sumatera Utara.

## **BAB 3**

### **ANALISIS DAN PERANCANGAN SISTEM**

#### **3.1 Dataset**

Dataset yang digunakan dalam penelitian ini berasal dari dokumen-dokumen skripsi yang dikumpulkan dari Fakultas Ilmu Komputer dan Teknologi Informasi, Universitas Sumatera Utara (USU). Jumlah keseluruhan data skripsi yang berhasil dihimpun sebanyak 300 dokumen, yang mencakup berbagai angkatan mulai dari tahun 2017 hingga 2020. Dokumen-dokumen ini digunakan sebagai sumber pembandingan utama dalam proses deteksi kemiripan terhadap dokumen executive summary yang diunggah oleh mahasiswa melalui sistem. Selain itu, sebanyak 50 dokumen executive summary juga disiapkan sebagai data uji (testing data) untuk mengevaluasi performa sistem.

Seluruh dokumen yang digunakan tersedia dalam format digital, yaitu Portable Document Format (PDF), yang kemudian diarsipkan dan dikelola dalam basis data sistem. Pemanfaatan dokumen digital memberikan kemudahan dalam proses pemrosesan otomatis, sekaligus memungkinkan sistem untuk melakukan pencocokan dan analisis secara efisien. Data digital ini menjadi landasan penting dalam tahap ekstraksi informasi dan pengolahan teks secara komputasional menggunakan pendekatan Natural Language Processing (NLP).

Sebelum digunakan dalam tahap perbandingan, setiap dokumen skripsi terlebih dahulu melalui serangkaian tahapan pra-pemrosesan (preprocessing) untuk menyiapkan data teks agar dapat diolah oleh mesin. Proses pra-pemrosesan ini meliputi text cleaning (penghapusan karakter tidak relevan seperti angka, simbol, dan tanda baca), case folding (konversi seluruh huruf menjadi huruf kecil), tokenization (pemecahan teks menjadi unit kata), stopword removal (penghapusan kata umum yang tidak memiliki makna signifikan), dan lemmatization (mengubah kata ke bentuk dasar). Tahapan ini bertujuan untuk menyederhanakan struktur teks tanpa menghilangkan makna yang terkandung di dalamnya.

Setelah proses pra-pemrosesan selesai, teks hasil pemrosesan diubah menjadi representasi vektor menggunakan algoritma Doc2Vec. Algoritma ini memungkinkan setiap dokumen diwakili dalam bentuk vektor berdimensi tetap yang mengandung informasi semantik dan konteks dari teks tersebut. Representasi vektor inilah yang menjadi dasar dalam proses perbandingan kesamaan antar dokumen, khususnya antara dokumen skripsi dan executive summary yang diunggah pengguna.

Data yang digunakan dalam penelitian ini tidak hanya berfungsi sebagai bahan pengujian sistem, tetapi juga menjadi fondasi pembentukan model deteksi kemiripan dokumen. Keragaman dokumen skripsi dari berbagai tahun memungkinkan sistem untuk mempelajari pola bahasa, gaya penulisan ilmiah, serta struktur naratif yang khas dalam penulisan akademik. Dengan demikian, sistem yang dibangun diharapkan mampu memberikan hasil deteksi kemiripan yang lebih akurat, objektif, dan dapat diandalkan dalam mendeteksi indikasi plagiarisme atau kesamaan konten secara menyeluruh.

**Tabel 3. 1 Jumlah Dataset**

<b>Jenis Dataset</b>	<b>Jumlah Dokumen</b>
Training Set	300 Skripsi
Testing Set	60 <i>Executive Summary</i>

### 3.2 Analisis Sistem

Analisis sistem dilakukan untuk merancang arsitektur sistem secara menyeluruh, meliputi identifikasi kebutuhan perangkat keras, perangkat lunak, serta mekanisme kerja sistem deteksi kemiripan dokumen skripsi. Sistem ini dirancang berbasis web dengan menggunakan framework *Flask* sebagai *backend* dan PostgreSQL sebagai sistem manajemen basis data. Seluruh data yang berkaitan dengan dokumen skripsi, akun pengguna, dan hasil perhitungan similarity disimpan dan dikelola melalui basis data tersebut.

Inti dari sistem ini adalah model pembelajaran mesin *Doc2Vec* yang telah dilatih sebelumnya menggunakan kumpulan dokumen skripsi. Model ini digunakan untuk merepresentasikan setiap dokumen ke dalam bentuk vektor numerik. Proses perhitungan kemiripan antar dokumen dilakukan dengan metode *Cosine Similarity*, yang menghitung tingkat kesamaan antar vektor berdasarkan sudut kemiringan antar dokumen tersebut.

Sebelum proses vektorisasi dilakukan, sistem terlebih dahulu menjalankan tahapan *preprocessing* terhadap teks dokumen, yang meliputi normalisasi huruf, pembersihan karakter non-alfabet, tokenisasi, penghapusan *stopword* bahasa Indonesia, serta stemming menggunakan pustaka Sastrawi. Dengan tahapan ini, sistem dapat menyiapkan data teks yang bersih dan terstruktur untuk selanjutnya direpresentasikan ke dalam model vektor *Doc2Vec*.

Hasil perhitungan similarity ditampilkan melalui antarmuka web yang dirancang secara interaktif dan responsif. Pengguna dapat mengunggah dokumen dalam format PDF, dan sistem akan menampilkan lima dokumen skripsi yang memiliki tingkat kemiripan tertinggi berdasarkan skor similarity. Selain itu, sistem juga menghasilkan file PDF dengan penanda (*highlight*) pada kata-kata yang dianggap mirip antara dokumen yang diunggah dengan dokumen pembanding, guna memudahkan pengguna dalam melakukan analisis visual terhadap kemiripan isi dokumen.

Sistem juga dilengkapi dengan berbagai fitur pendukung, antara lain: riwayat unggahan pengguna, visualisasi statistik penggunaan oleh admin, sistem otentikasi pengguna berbasis peran (admin dan user), serta fitur pengelolaan dokumen skripsi oleh admin secara terpusat. Dengan rancangan tersebut, sistem ini diharapkan mampu memberikan kontribusi nyata dalam mempermudah proses evaluasi kemiripan karya ilmiah secara efektif, akurat, dan terstruktur. Gambaran umum arsitektur sistem ini dapat dilihat pada Gambar 3.1. Penjelasan terkait arsitektur umum pada Gambar 3.1 adalah sebagai berikut:

1. Pengumpulan Data (*Data Acquisition*)

Data skripsi dikumpulkan dari Fakultas Ilmu Komputer dan Teknologi Informasi, Universitas Sumatera Utara (USU). Dokumen-dokumen skripsi ini digunakan sebagai data utama untuk dibandingkan dengan *executive summary* yang diunggah oleh mahasiswa. Tahap ini penting karena skripsi menjadi dasar untuk perbandingan kesamaan dokumen.

2. Pra-Pemrosesan Data (*Data Preprocessing – Document Skripsi*)

Setelah data skripsi terkumpul, langkah berikutnya adalah mempersiapkan data untuk diproses. Tahapan-tahapan preprocessing meliputi:

- a. ***Text Cleaning:*** Menghapus elemen-elemen yang tidak relevan seperti simbol dan tanda baca yang tidak diperlukan. ***Tokenization:*** Memecah

teks menjadi unit-unit kecil (*token*), seperti kata atau frasa seperti pada Tabel 3.2.

- b. Case Folding:** Mengubah seluruh teks menjadi huruf kecil untuk memastikan bahwa kata yang sama dengan kapitalisasi berbeda dianggap serupa seperti pada Tabel 3.4.
- c. Tokenization:** Memecah teks menjadi unit-unit kecil (*token*), seperti kata atau frasa seperti pada Tabel 3.5.
- d. Lemmatization:** Mengubah kata-kata menjadi bentuk dasarnya seperti pada Tabel 3.7.
- e. Stop Words Removal:** Menghapus kata-kata umum yang tidak memberikan banyak makna dalam analisis teks seperti pada Tabel 3.6.
- f. Embedding Text (Text Embedding)** Setelah dokumen skripsi melewati tahap preprocessing, dokumen tersebut diubah menjadi representasi vektor menggunakan model *Doc2Vec*. *Doc2Vec* menghasilkan embedding teks, yaitu representasi numerik dari dokumen, yang mampu menangkap informasi semantik dan konteks dari dokumen tersebut. Embedding ini kemudian digunakan dalam proses perbandingan dengan dokumen *executive summary*.
- g. Menyimpan Vektor ke database (Store Vectors in Database)** Hasil embedding dari dokumen skripsi disimpan dalam database. Database ini akan digunakan sebagai referensi dalam proses perbandingan dengan dokumen *executive summary* yang diunggah oleh pengguna.
- h. Unggah Dokumen (Document Upload)** Mahasiswa mengunggah *executive summary* mereka melalui sistem. Setelah diunggah, dokumen ini akan diproses dan dibandingkan dengan dokumen skripsi yang sudah ada di database.
- i. Pra-Pemrosesan Executive Summary (Data Preprocessing - Executive Summary)**

- j.** Sama seperti dokumen skripsi, *executive summary* juga melalui tahapan
- k.** preprocessing yang serupa:
- a. *Text Cleaning*: Menghapus elemen-elemen yang tidak relevan seperti simbol dan tanda baca yang tidak diperlukan.
  - b. *Case Folding*: Mengubah seluruh teks menjadi huruf kecil untuk memastikan bahwa kata yang sama dengan kapitalisasi berbeda dianggap serupa.
  - c. *Tokenization*: Memecah teks menjadi unit-unit kecil (*token*), seperti kata atau frasa.
  - d. *Lemmatization*: Mengubah kata-kata menjadi bentuk dasarnya.
  - e. *Stop Words Removal*: Menghapus kata-kata umum yang tidak memberikan banyak makna dalam analisis teks.
- l.** Text Embedding (*Executive Summary*) Setelah dokumen *executive summary* diproses, dokumen ini juga diubah menjadi representasi vektor menggunakan model *Doc2Vec*. Proses embedding ini menghasilkan vektor yang menyimpan makna semantik dari dokumen *executive summary* tersebut. Representasi vektor ini nantinya digunakan untuk dibandingkan dengan vektor skripsi yang sudah disimpan dalam basis data.
- m.** Perbandingan Vektor (*Compare Vectors*) Vektor dari *executive summary* yang diunggah dibandingkan dengan vektor dokumen skripsi yang telah disimpan di database menggunakan metode *Cosine Similarity*. *Cosine Similarity* menghitung sudut antara dua vektor di ruang multidimensi untuk menentukan tingkat kemiripan antar dokumen. Semakin kecil sudutnya, semakin besar kesamaan antara dua dokumen tersebut
- n.** Deteksi Kesamaan dan Perhitungan Skor (*Similarity Detection and Score Calculation*) Setelah mendapatkan embedding dari kedua dokumen (skripsi dan *executive summary*), sistem akan melakukan deteksi kesamaan menggunakan metrik *Cosine Similarity* untuk menghitung kedekatan antara dua vektor. Proses ini menghasilkan skor kesamaan,

yang menunjukkan tingkat kemiripan antara executive summary yang diunggah dan dokumen-dokumen skripsi yang ada di database

- o. Tabel Kesamaan (*Similarity Table*) Hasil deteksi kesamaan akan disajikan dalam bentuk tabel yang menampilkan:
  - a. Judul Skripsi: Judul skripsi yang dibandingkan.
  - b. Skor Kesamaan: Tingkat kesamaan antara executive summary dan skripsi.
- c. Aksi (*Actions*): Pengguna dapat melihat dokumen yang lebih mirip, dengan bagian-bagian relevan yang di-*highlight* untuk memudahkan identifikasi kesamaan.



Gambar 3. 1 Arsitektur Umum

**Tabel 3. 2 Text Cleaning**

	Sebelum <i>Text Cleaning</i>	Sesudah <i>Text Cleaning</i>
<b>Skripsi</b>	Jenis pelabelan didasarkan pada kategori yang paling sering ditanyakan pada situs PetCoach. Dengan bantuan pakar dan penyesuaian terhadap kebutuhan klinik, dipilih 3 jenis pelabelan utama yaitu: 'Kesehatan' sebanyak 158 baris data, 'Klinik' sebanyak 61 baris data, dan 'Perilaku' sebanyak 95 baris data. Pertanyaan yang berkaitan dengan perawatan hewan peliharaan, gejala penyakit, dan sejenisnya diberi label 'Kesehatan'. Pertanyaan yang berkaitan dengan layanan dan informasi mengenai UPTD Klinik Hewan Sumatera Utara diberi label 'Klinik'.	Jenis pelabelan didasarkan pada kategori yang paling sering ditanyakan pada situs PetCoach Dengan bantuan pakar dan penyesuaian terhadap kebutuhan klinik dipilih 3 jenis pelabelan utama yaitu Kesehatan sebanyak 158 baris data Klinik sebanyak 61 baris data dan Perilaku sebanyak 95 baris data Pertanyaan yang berkaitan dengan perawatan hewan peliharaan gejala penyakit dan sejenisnya diberi label Kesehatan Pertanyaan yang berkaitan dengan layanan dan informasi mengenai UPTD Klinik Hewan Sumatera Utara diberi label Klinik
<i>Executive Summary</i>	Seiring dengan perkembangan teknologi, chatbot telah menjadi salah satu solusi yang menjanjikan untuk menyediakan layanan informasi yang lebih interaktif dan mudah diakses.Chatbot adalah program komputer yang menggunakan Natural Language Processing (NLP) sebagai bagian dari kecerdasan buatan untuk menanggapi pesan pengguna melalui teks atau suara.	seiring dengan perkembangan teknologi chatbot telah menjadi salah satu solusi yang menjanjikan untuk menyediakan layanan informasi yang lebih interaktif dan mudah diakses chatbot adalah program komputer yang menggunakan natural language processing nlp sebagai bagian dari kecerdasan buatan untuk menanggapi pesan pengguna melalui teks atau suara



**Tabel 3. 3 Case Folding**

	Sebelum <i>Case Folding</i>	Sesudah <i>Case Folding</i>
<b>Skripsi</b>	Jenis pelabelan didasarkan pada kategori yang paling sering ditanyakan pada situs PetCoach Dengan bantuan pakar dan penyesuaian terhadap kebutuhan klinik dipilih 3 jenis pelabelan utama yaitu Kesehatan sebanyak 158 baris data Klinik sebanyak 61 baris data dan Perilaku sebanyak 95 baris data Pertanyaan yang berkaitan dengan perawatan hewan peliharaan gejala penyakit dan sejenisnya diberi label Kesehatan Pertanyaan yang berkaitan dengan layanan dan informasi mengenai UPTD Klinik Hewan Sumatera Utara diberi label Klinik	jenis pelabelan didasarkan pada kategori yang paling sering ditanyakan pada situs petcoach dengan bantuan pakar dan penyesuaian terhadap kebutuhan klinik dipilih 3 jenis pelabelan utama yaitu kesehatan sebanyak 158 baris data klinik sebanyak 61 baris data dan perilaku sebanyak 95 baris data pertanyaan yang berkaitan dengan perawatan hewan peliharaan gejala penyakit dan sejenisnya diberi label kesehatan pertanyaan yang berkaitan dengan layanan dan informasi mengenai uptd klinik hewan sumatera utara diberi label klinik
<i>Executive Summary</i>	Seiring dengan perkembangan teknologi chatbot telah menjadi salah satu solusi yang menjanjikan untuk menyediakan layanan informasi yang lebih interaktif dan mudah diakses Chatbot adalah program komputer yang menggunakan Natural Language Processing NLP sebagai bagian dari kecerdasan buatan untuk menanggapi pesan pengguna melalui teks atau suara	seiring dengan perkembangan teknologi chatbot telah menjadi salah satu solusi yang menjanjikan untuk menyediakan layanan informasi yang lebih interaktif dan mudah diakses chatbot adalah program komputer yang menggunakan natural language processing nlp sebagai bagian dari kecerdasan buatan untuk menanggapi pesan pengguna melalui teks atau suara

**Tabel 3. 4 Tokenization**

	Sebelum <i>Tokenization</i>	Sesudah <i>Tokenization</i>
<b>Skripsi</b>	jenis pelabelan didasarkan pada kategori yang paling sering ditanyakan pada situs petcoach dengan bantuan pakar dan penyesuaian terhadap kebutuhan klinik dipilih 3 jenis pelabelan utama yaitu kesehatan sebanyak 158 baris data klinik sebanyak 61 baris data dan perilaku sebanyak 95 baris data pertanyaan yang berkaitan dengan perawatan hewan peliharaan gejala penyakit dan sejenisnya diberi label kesehatan pertanyaan yang berkaitan dengan layanan dan informasi mengenai updt klinik hewan sumatera utara diberi label klinik	["jenis", "pelabelan", "didasarkan", "pada", "kategori", "yang", "paling", "sering", "ditanyakan", "pada", "situs", "petcoach", "dengan", "bantuan", "pakar", "dan", "penyesuaian", "terhadap", "kebutuhan", "klinik", "dipilih", "3", "jenis", "pelabelan", "utama", "yaitu", "kesehatan", "sebanyak", "158", "baris", "data", "klinik", "sebanyak", "61", "baris", "data", "dan", "perilaku", "sebanyak", "95", "baris", "data", "pertanyaan", "yang", "berkaitan", "dengan", "perawatan", "hewan", "peliharaan", "gejala", "penyakit", "dan", "sejenisnya", "diberi", "label", "kesehatan", "pertanyaan", "yang", "berkaitan", "dengan", "layanan", "dan", "informasi", "mengenai", "uptd", "klinik", "hewan", "sumatera", "utara", "diberi", "label", "klinik"]
<i>Executive Summary</i>	seiring dengan perkembangan teknologi chatbot telah menjadi salah satu solusi yang menjanjikan untuk menyediakan layanan informasi yang lebih interaktif dan mudah diakses chatbot adalah program komputer yang menggunakan natural language processing nlp sebagai bagian dari kecerdasan buatan untuk menanggapi pesan pengguna melalui teks atau suara	["seiring", "dengan", "perkembangan", "teknologi", "chatbot", "telah", "menjadi", "salah", "satu", "solusi", "yang", "menjanjikan", "untuk", "menyediakan", "layanan", "informasi", "yang", "lebih", "interaktif", "dan", "mudah", "diakses", "chatbot", "adalah", "program", "komputer", "yang", "menggunakan", "natural", "language", "processing", "nlp", "sebagai", "bagian", "dari", "kecerdasan", "buatan", "untuk", "menanggapi", "pesan", "pengguna", "melalui", "teks", "atau", "suara"]

**Tabel 3. 5 Lemmatization**

	Sebelum <i>Lemmatization</i>	Sesudah <i>Lemmatization</i>
Skripsi	["jenis", "pelabelan", "didasarkan", "pada", "kategori", "yang", "paling", "sering", "ditanyakan", "pada", "situs", "petcoach", "dengan", "bantuan", "pakar", "dan", "penyesuaian", "terhadap", "kebutuhan", "klinik", "dipilih", "3", "jenis", "pelabelan", "utama", "yaitu", "kesehatan", "sebanyak", "158", "baris", "data", "klinik", "sebanyak", "61", "baris", "data", "dan", "perilaku", "sebanyak", "95", "baris", "data", "pertanyaan", "yang", "berkaitan", "dengan", "perawatan", "hewan", "peliharaan", "gejala", "penyakit", "dan", "sejenisnya", "diberi", "label", "kesehatan", "pertanyaan", "yang", "berkaitan", "dengan", "layanan", "dan", "informasi", "mengenai", "uptd", "klinik", "hewan", "sumatera", "utara", "diberi", "label", "klinik"]	["jenis", "label", "dasar", "pada", "kategori", "sering", "tanya", "pada", "situs", "petcoach", "bantu", "pakar", "dan", "sesuaikan", "butuh", "klinik", "pilih", "3", "jenis", "label", "utama", "yaitu", "sehat", "banyak", "158", "baris", "data", "klinik", "banyak", "61", "baris", "data", "perilaku", "banyak", "95", "baris", "data", "tanya", "kait", "rawat", "hewan", "pelihara", "gejala", "penyakit", "sejenis", "beri", "label", "sehat", "tanya", "kait", "layanan", "dan", "informasi", "mengenai", "uptd", "klinik", "hewan", "sumatera", "utara", "beri", "label", "klinik"]
<i>Executive Summary</i>	["seiring", "dengan", "perkembangan", "teknologi", "chatbot", "telah", "menjadi", "salah", "satu", "solusi", "yang", "menjanjikan", "untuk", "menyediakan", "layanan", "informasi", "yang", "lebih", "interaktif", "dan", "mudah", "diakses", "chatbot", "adalah", "program", "komputer", "yang", "menggunakan", "natural", "language", "processing", "nlp", "sebagai", "bagian", "dari", "kecerdasan", "buatan", "untuk", "menanggapi", "pesan", "pengguna", "melalui", "teks", "atau", "suara"]	["iring", "kait", "kembang", "teknologi", "chatbot", "jadi", "satu", "solusi", "janjikan", "untuk", "sedia", "layanan", "informasi", "interaktif", "mudah", "akses", "chatbot", "program", "komputer", "guna", "natural", "language", "processing", "nlp", "bagi", "cerdas", "buatan", "tanggap", "pesan", "guna", "melalui", "teks", "atau", "suara"]

**Tabel 3. 6 Word Removal**

	Sebelum <i>Stop Words Removal</i>	Sesudah <i>Stop Words Removal</i>
Stop Words Removal	["jenis", "label", "dasar", "pada", "kategori", "sering", "tanya", "pada", "situs", "petcoach", "bantu", "pakar", "dan", "sesuaikan", "butuh", "klinik", "pilih", "3", "jenis", "label", "utama", "yaitu", "sehat", "banyak", "158", "baris", "data", "klinik", "banyak", "61", "baris", "data", "perilaku", "banyak", "95", "baris", "data", "tanya", "kait", "rawat", "hewan", "peliharaan", "gejala", "penyakit", "sejenis", "beri", "label", "sehat", "tanya", "kait", "layanan", "dan", "informasi", "mengenai", "uptd", "klinik", "hewan", "sumatera", "utara", "beri", "label", "klinik"]	["jenis", "label", "dasar", "kategori", "sering", "tanya", "situs", "petcoach", "bantu", "pakar", "sesuaikan", "butuh", "klinik", "pilih", "3", "jenis", "label", "utama", "yaitu", "sehat", "banyak", "158", "baris", "data", "klinik", "banyak", "61", "baris", "data", "perilaku", "banyak", "95", "baris", "data", "tanya", "kait", "rawat", "hewan", "peliharaan", "gejala", "penyakit", "sejenis", "beri", "label", "sehat", "tanya", "kait", "layanan", "informasi", "mengenai", "uptd", "klinik", "hewan", "sumatera", "utara", "beri", "label", "klinik"]
Stop Words Removal	["iring", "kait", "kembang", "teknologi", "chatbot", "jadi", "satu", "solusi", "janjikan", "untuk", "sedia", "layanan", "informasi", "interaktif", "mudah", "akses", "chatbot", "program", "komputer", "guna", "natural", "language", "processing", "nlp", "bagi", "cerdas", "buatan", "tanggap", "pesan", "guna", "melalui", "teks", "atau", "suara"]	["iring", "kait", "kembang", "teknologi", "chatbot", "jadi", "satu", "solusi", "janjikan", "sedia", "layanan", "informasi", "interaktif", "mudah", "akses", "chatbot", "program", "komputer", "guna", "natural", "language", "processing", "nlp", "bagi", "cerdas", "buatan", "tanggap", "pesan", "guna", "melalui", "teks", "suara"]

### 3.3 Diagram Activity

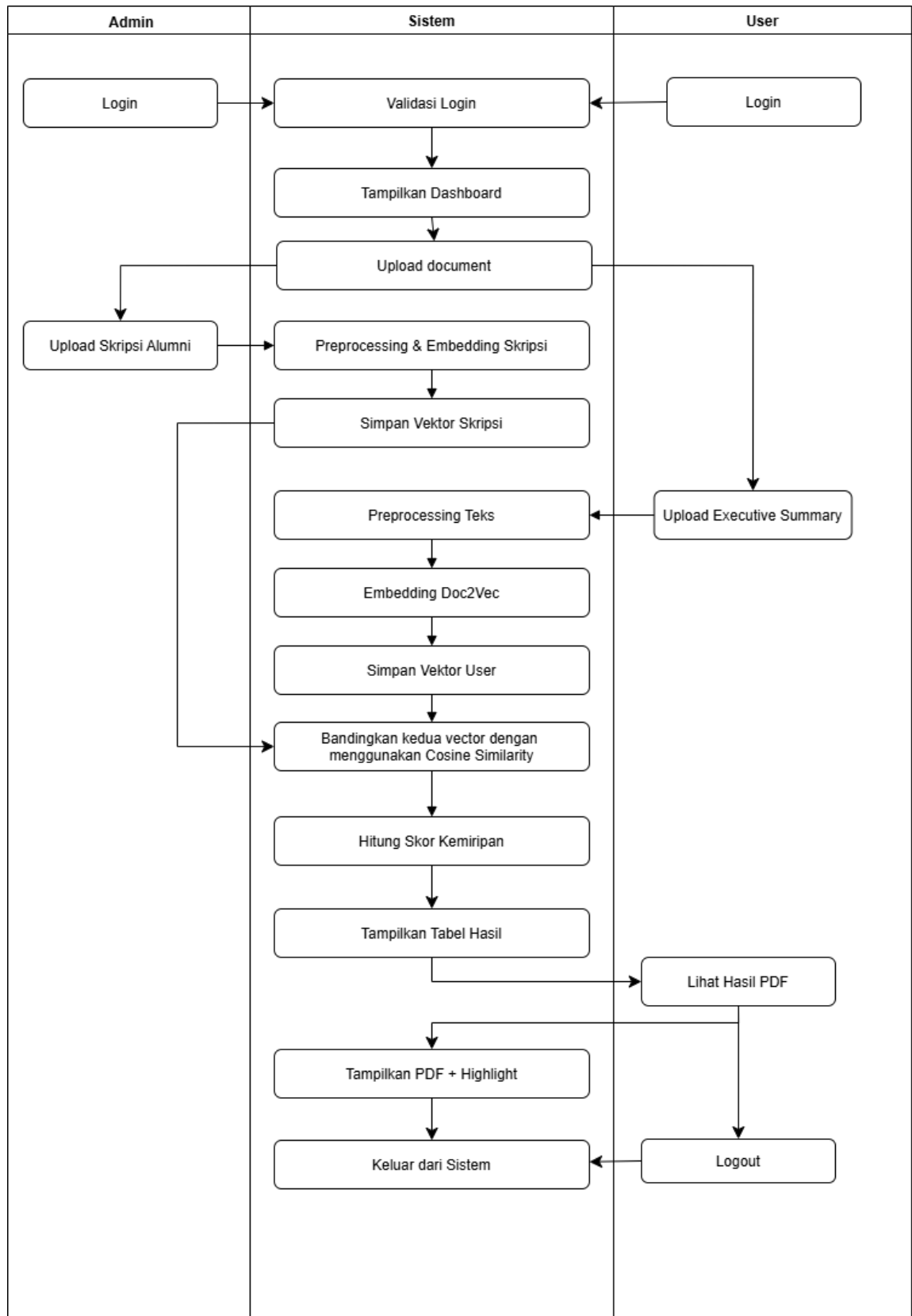
Diagram activity berikut menggambarkan alur sistem deteksi kemiripan dokumen skripsi berbasis web yang melibatkan tiga entitas utama, yaitu Admin, Sistem, dan User. Diagram ini dimulai ketika Admin dan User melakukan proses login ke dalam sistem. Setelah login berhasil, Sistem akan melakukan validasi dan mengarahkan pengguna ke dashboard sesuai perannya.

Pada sisi Admin, proses dimulai dari login lalu dilanjutkan dengan aktivitas mengunggah kumpulan dokumen skripsi alumni ke dalam sistem. Setelah file diunggah, sistem akan menjalankan proses "Upload Document", kemudian dilanjutkan dengan tahapan prapemrosesan teks, seperti cleaning, tokenisasi, stopword removal, dan lemmatization. Selanjutnya, sistem melakukan embedding teks menggunakan algoritma Doc2Vec. Hasil dari embedding berupa vektor numerik dari setiap dokumen akan disimpan dalam basis data sebagai referensi pembandingan.

Sementara itu, User yang telah berhasil login akan diarahkan ke dashboard dan dapat mengunggah dokumen executive summary dalam format PDF. Dokumen ini juga akan melalui tahapan prapemrosesan dan embedding Doc2Vec, sama seperti yang dilakukan pada dokumen skripsi alumni. Vektor dari dokumen user akan disimpan sementara oleh sistem. Setelah kedua vektor tersedia (dokumen user dan dokumen dari database), sistem akan membandingkan keduanya menggunakan metode Cosine Similarity. Proses ini menghitung skor kemiripan antara dua dokumen dalam bentuk nilai numerik. Skor kemiripan tersebut selanjutnya diproses untuk menentukan urutan relevansi antar dokumen.

Sistem kemudian menampilkan hasil perbandingan dalam bentuk tabel kepada User, yang memuat informasi judul skripsi pembandingan dan skor kemiripan. User juga dapat memilih untuk melihat dokumen pembandingan secara lebih detail melalui fitur "Lihat Hasil PDF", yang menampilkan file PDF dengan bagian-bagian yang mirip telah diberi highlight. Setelah semua proses selesai, User dapat menekan tombol Logout untuk keluar dari sistem. Proses logout ini juga tersedia untuk Admin.

Diagram activity ini memberikan gambaran menyeluruh terhadap interaksi antara Admin, User, dan Sistem dalam sistem deteksi kemiripan dokumen, mulai dari login, pengunggahan dokumen, pemrosesan data, pembandingan vektor, hingga penyajian hasil kepada pengguna. Dengan diagram ini, alur sistem dapat dipahami secara runtut dan jelas sesuai implementasi aktual. Diagram Activity disajikan dalam Gambar 3.2.



**Gambar 3. 2 Activity Diagram**

### **3.4 Perancangan Antarmuka Sistem**

Pada bagian ini dijelaskan rancangan antarmuka sistem yang dikembangkan untuk mendeteksi kemiripan dokumen skripsi berbasis web. Rancangan ini mencerminkan struktur tampilan yang akan diimplementasikan dalam sistem, baik untuk pengguna umum (user) maupun admin. Tujuannya adalah memberikan gambaran yang jelas tentang alur interaksi antara pengguna dan sistem serta menjadi acuan dalam proses pengembangan antarmuka.

#### *3.4.1 Halaman Login*

Halaman login merupakan gerbang awal bagi pengguna untuk masuk ke dalam sistem. Pengguna perlu memasukkan username dan password untuk mendapatkan akses ke fitur yang tersedia. Jika kredensial yang dimasukkan tidak valid, sistem akan menampilkan notifikasi kesalahan.

Komponen pada halaman ini mencakup logo institusi di bagian atas, kolom input untuk username dan password, tombol login, serta tautan menuju halaman register bagi pengguna yang belum memiliki akun. Tampak seperti pada Gambar 3.4

#### *3.4.2 Halaman Register*

Halaman register memungkinkan pengguna baru untuk membuat akun. Admin tidak dapat melakukan registrasi melalui halaman ini karena akun admin hanya dapat dibuat oleh sistem secara langsung. Halaman ini terdiri dari logo sistem, kolom input untuk username dan password, pilihan role User, tombol register, serta tautan kembali ke halaman login seperti pada Gambar 3.5.

#### *3.4.3 Halaman Dashboard (Admin)*

Setelah berhasil login, admin diarahkan ke halaman dashboard yang berisi informasi sambutan dan navigasi sistem. Menu navigasi terletak di sebelah kiri dan terdiri atas Dashboard, Upload Dokumen, dan Riwayat Dokumen. Di bagian kanan layar ditampilkan identitas pengguna yang sedang login beserta tombol logout terlihat seperti pada Gambar 3.6.

#### *3.4.4 Halaman Upload Dokumen (Admin)*

Halaman ini memungkinkan admin untuk mengunggah dokumen skripsi ke dalam database sistem. Form upload terdiri atas kolom input untuk judul dokumen, tombol

untuk memilih file berformat PDF, serta tombol untuk mengunggah dokumen tersebut ke sistem seperti pada Gambar 3.7.

#### *3.4.5 Halaman Daftar Dokumen (Admin)*

Halaman daftar dokumen menampilkan seluruh dokumen skripsi yang telah diunggah oleh admin. Informasi yang ditampilkan meliputi nomor urut, judul dokumen, nama file, isi dokumen, dan waktu unggah. Data ini ditampilkan dalam bentuk tabel yang dapat diakses melalui menu navigasi. Tampak seperti pada Gambar 3.8.

#### *3.4.6 Halaman Dashboard (User)*

Pengguna umum (user) setelah login akan diarahkan ke dashboard yang menampilkan ucapan selamat datang dan identitas pengguna. Menu navigasi tersedia di sisi kiri halaman, yang terdiri dari menu Dashboard, Upload Dokumen, dan Riwayat Dokumen. Halaman ini juga menyediakan tombol logout seperti pada Gambar 3.9.

#### *3.4.7 Halaman Upload File untuk Cek Similarity*

Halaman ini digunakan oleh user untuk mengunggah file yang ingin dicek kemiripannya dengan dokumen-dokumen skripsi yang tersedia di database. Halaman ini terdiri dari tombol pemilihan file dan tombol untuk memulai proses pengecekan similarity. File yang dapat diunggah adalah file berformat PDF. Tampak seperti pada Gambar 3.10.

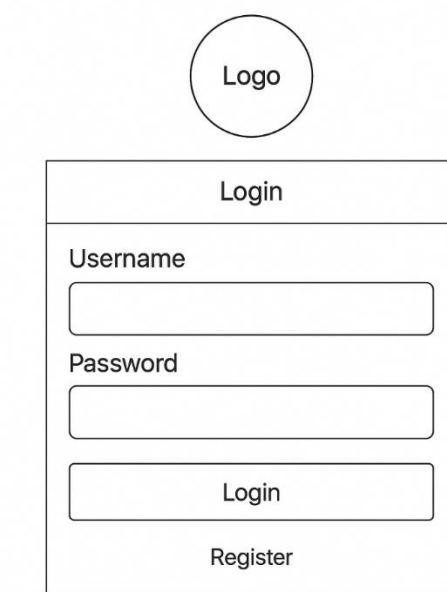
#### *3.4.8 Halaman Hasil Similaritas*

Setelah sistem selesai melakukan perbandingan dokumen, hasil similarity ditampilkan dalam bentuk tabel. Informasi yang disajikan mencakup nomor, judul skripsi yang memiliki kemiripan, skor kesamaan dalam bentuk persentase, serta tombol untuk melihat dokumen hasil similarity dalam format PDF seperti pada Gambar 3.11.

#### *3.4.9 Halaman Riwayat Dokumen*

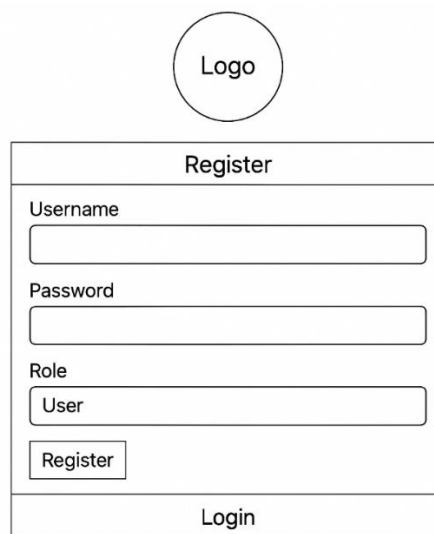
Halaman riwayat menampilkan seluruh dokumen yang pernah diunggah oleh pengguna beserta hasil perbandingan similarity-nya. Informasi yang ditampilkan mencakup nomor urut, judul dokumen, skor similarity, kata yang relevan, waktu unggah, serta tombol untuk melihat dokumen hasil similarity. Tampak seperti pada Gambar 3.12





The login form is centered on the page. Above it is a circular logo containing the word "Logo". The form itself is a rectangular box with a header section labeled "Login". Below the header, there are two input fields: the first is labeled "Username" and the second is labeled "Password". Below these fields is a button labeled "Login". At the bottom of the form box is a link labeled "Register".

**Gambar 3. 3 Halaman Login**

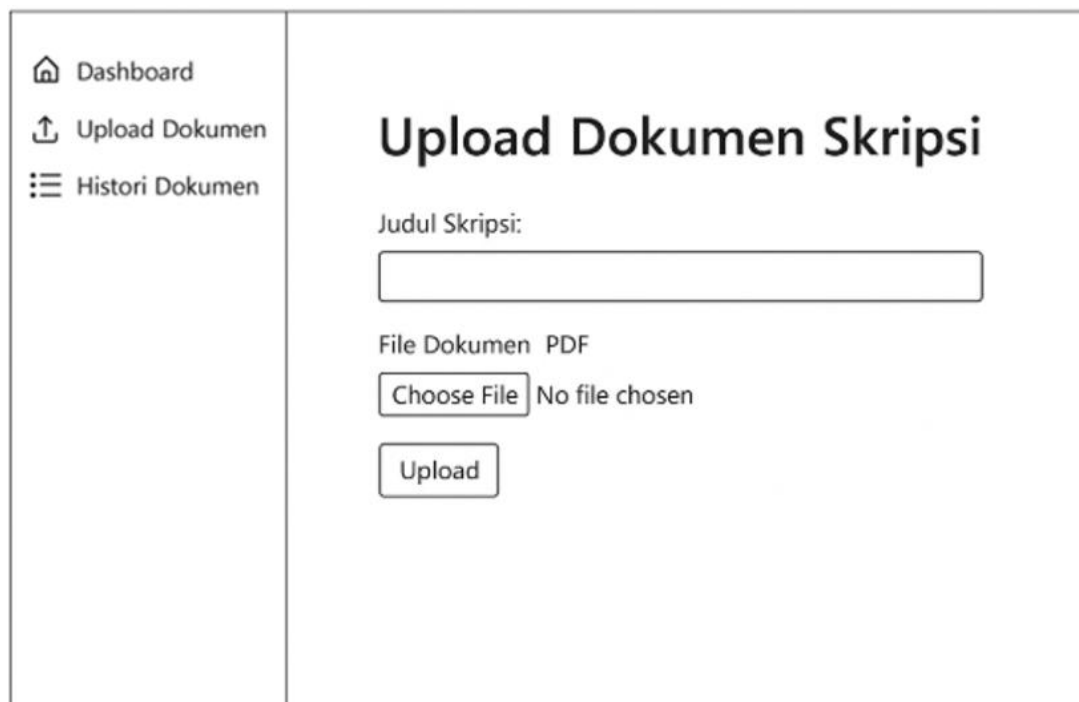


The register form is centered on the page. Above it is a circular logo containing the word "Logo". The form is a rectangular box with a header section labeled "Register". Below the header, there are three input fields: the first is labeled "Username", the second is labeled "Password", and the third is labeled "Role" with the text "User" inside it. Below these fields is a button labeled "Register". At the bottom of the form box is a link labeled "Login".

**Gambar 3. 4 Halaman Register**



**Gambar 3. 5 Halaman Dashboard Admin**



**Gambar 3. 6 Halaman Upload Dokumen Bagian Admin**

Dashboard  
Upload Dokumen  
Histori Dokumen

### Dokumen Skripsi yang Di-upload

No	Judul	Nama File	Isi teks	Waktu Upload
1.	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>
	<input type="text"/>	<input type="text"/>		
	<input type="text"/>			

**Gambar 3. 7 Halaman Dokumen Skripsi Bagian Admin**

Dashboard  
Upload Dokumen  
Histori Dokumen

Dashboard Logout

Selamat Datang, user

**Gambar 3. 8 Halaman Dashboard User**

Dashboard
Upload Dokumen
Histori Dokumen

## Upload File untuk Cek Similarity

Choose File
No file chosen

Upload dan Cek Similarity

Gambar 3. 9 Halaman Upload Dokumen Bagian User




Dashboard
Upload Dokumen
Histori Dokumen

## Hasil Similaritas

### Dokumen yang Tersedia di Database

No	Judul Skripsi	Skor Kesamaan	Aksi
1.			Lihat Hasil PDF
2.			
3.			
4.			
4.			

Gambar 3. 10 Halaman Hasil Simillaritas

<div><div> Dashboard</div><div> Upload Dokumen</div><div> Histori Dokumen</div></div>	<h2>Riwayat Dokumen yang Di-upload</h2>			
	<b>No</b>	<b>Judul File yang DiUpload</b>	<b>Skor Similarity</b>	<b>Kata yang Relevan</b>
	1.	<input type="text"/>		<div>Lihat Hasil PDF</div>
	2.	<input type="text"/>		<div>Lihat Hasil PDF</div>
	3.	<input type="text"/>		<input type="text"/>
	4.	<input type="text"/>		<input type="text"/>

**Gambar 3. 11 Halaman Riwayat Dokumen**

### 3.5 Metode Evaluasi

Metode evaluasi yang digunakan dalam penelitian ini adalah *Cosine Similarity*. Evaluasi ini bertujuan untuk menilai sejauh mana sistem yang dikembangkan mampu mengenali kemiripan antara dokumen Executive Summary (Exume) dengan dokumen skripsi yang telah ada dalam basis data.

*Cosine Similarity* merupakan ukuran sejauh mana dua dokumen memiliki arah yang sama dalam ruang vektor. Nilai cosine similarity berkisar antara 0 hingga 1, di mana:

- a. Nilai mendekati 1 menunjukkan tingkat kemiripan yang tinggi
- b. Nilai mendekati 0 menunjukkan kemiripan yang rendah

Dalam penelitian ini, sistem akan mengukur kemiripan antara satu dokumen exume dengan seluruh dokumen skripsi menggunakan model Doc2Vec yang telah dilatih. Selanjutnya, sistem akan menampilkan Top-N dokumen yang memiliki skor kemiripan tertinggi. Visualisasi hasil evaluasi dilakukan dengan grafik perbandingan skor similarity per epoch untuk melihat performa model selama pelatihan. Selain itu, hasil similarity akhir antara exume dan skripsi juga ditampilkan dalam bentuk tabel atau daftar skor kemiripan tertinggi.

.

## **BAB 4**

### **IMPLEMENTASI DAN PENGUJIAN SISTEM**

#### **4.1 Implementasi Sistem**

Implementasi sistem deteksi kemiripan dokumen skripsi berbasis web ini dilakukan dengan menggunakan perangkat keras dan perangkat lunak yang mendukung proses pengembangan dan pengujian sistem secara optimal. Sistem dibangun dengan menggunakan bahasa pemrograman Python dan framework Flask untuk sisi backend, serta HTML, CSS, dan JavaScript untuk sisi frontend. Proses pengolahan dokumen dan perhitungan kemiripan dilakukan dengan bantuan pustaka NLP seperti Sastrawi dan Gensim. Untuk manajemen basis data digunakan PostgreSQL.

##### *4.1.1 Spesifikasi Perangkat Keras*

Perangkat keras yang digunakan dalam pengembangan dan implementasi sistem adalah sebagai berikut:

1. Perangkat: Laptop HP 245 G8 Notebook PC
2. Prosesor: AMD Ryzen 3 5300U with Radeon Graphics, 2.60 GHz
3. RAM: 8 GB
4. Sistem Operasi: Windows 11 Home Single Language 64-bit

##### *4.1.2 Spesifikasi Perangkat Lunak*

Perangkat lunak yang digunakan dalam pengembangan sistem ini terdiri dari beberapa tools dan library pendukung, yaitu:

1. Bahasa Pemrograman: Python 3.10.9
2. Framework Backend: Flask 3.1
3. Library NLP: Sastrawi, Gensim (Doc2Vec), re, spacy
4. Library Ekstraksi Teks: PyPDF2, pdfplumber, python-docx
5. Library Tambahan: psycpg2, werkzeug.security, functools
6. Frontend: HTML5, CSS3, JavaScript, Jinja2

7. Tools Pengembangan: Visual Studio Code
8. Database: PostgreSQL

## 4.2 Implementasi Data

Data yang digunakan pada penelitian ini terdiri dari kumpulan dokumen Executive Summary (Exum) dan dokumen Skripsi yang berjumlah total sebanyak 70 dokumen. Data tersebut digunakan dalam proses pelatihan model serta sebagai data pembanding untuk mengukur tingkat kemiripan terhadap dokumen yang diunggah oleh pengguna. Data yang digunakan dalam sistem telah melewati proses pre-processing untuk meningkatkan kualitas teks. Tahapan ini bertujuan agar representasi data menjadi lebih bersih dan relevan, sehingga performa dan akurasi model Doc2Vec dalam mendeteksi kemiripan dokumen menjadi lebih optimal.

## 4.3 Pelatihan Model

Implementasi model dilakukan untuk menghasilkan representasi numerik (*embedding*) dari setiap dokumen, sehingga proses perbandingan kemiripan dapat dilakukan secara efisien menggunakan Cosine Similarity.

Model dibangun menggunakan algoritma Doc2Vec dari pustaka Gensim, yang diimplementasikan secara lokal menggunakan Visual Studio Code (VS Code) sebagai lingkungan pengembangan terintegrasi (IDE). Sistem ini dikembangkan dengan pendekatan berbasis Python dan framework Flask, serta terhubung ke basis data PostgreSQL.

Sebelum proses pelatihan dimulai, seluruh dokumen skripsi melalui serangkaian tahapan pra-pemrosesan (*preprocessing*) yang mencakup:

1. Pembersihan teks (*text cleaning*)
2. *Case folding*
3. *Tokenisasi*
4. Penghapusan *stopword*
5. *Lemmatization*

Tahapan ini bertujuan menyederhanakan struktur teks agar dapat diproses secara optimal oleh model. Proses pra-pemrosesan dilakukan menggunakan library seperti Sastrawi, re, dan spaCy.



Dokumen yang telah diproses kemudian diubah menjadi vektor berdimensi tetap dengan Doc2Vec. Setelah proses pelatihan selesai, model disimpan dalam file berformat `.model`, yang kemudian digunakan dalam proses inferensi untuk membandingkan dokumen executive summary dengan dokumen skripsi di basis data menggunakan Cosine Similarity.

Model hanya perlu dilatih satu kali di awal pengembangan sistem. Jika terdapat penambahan dokumen skripsi baru, pelatihan dapat dijalankan ulang untuk memperbarui representasi vektor yang tersimpan di basis data.

Penggunaan Doc2Vec dipilih karena algoritma ini mampu menangkap makna semantik dan konteks dari seluruh dokumen, berbeda dengan metode tradisional seperti TF-IDF yang hanya memperhatikan frekuensi kata. Dengan demikian, sistem dapat mendeteksi kesamaan makna antar dokumen meskipun tidak memiliki kata-kata yang sama secara eksplisit.

Secara keseluruhan, proses implementasi model ini menjadi tulang punggung utama dalam sistem deteksi kemiripan dokumen berbasis teks yang dikembangkan. Selain menggunakan konfigurasi default pada Doc2Vec, dilakukan pula beberapa skenario pelatihan untuk mengeksplorasi pengaruh hyperparameter terhadap performa model. Tujuannya adalah untuk memperoleh akurasi dan nilai loss terbaik sebelum model digunakan dalam sistem. Pada proses pencarian kombinasi parameter pada penelitian ini telah dicoba dengan kombinasi nilai parameter yang beragam seperti berikut:

Vector\_size : [50, 100, 150]

Windows : [3, 5, 7]

DM : [0, 1]

Epcoh : [20, 30, 40]

Random\_seed : 42

Hasil percobaan *hyperparameter tuning* ditunjukkan pada tabel :

**Tabel 4.1 Kombinasi parameter Tuning Kombinasi 1**

Vector_Size	Windows	DM	Epoch	Hasil Val Sum
50	3	0	20	0.9587
50	3	0	30	0.9519
50	3	0	40	0.9454

**Tabel 4.2 Kombinasi parameter Tuning Kombinasi 2**

Vector_Size	Windows	DM	Epoch	Hasil Val Sum
50	3	1	20	0.9136
50	3	1	30	0.9899
50	3	1	40	0.9203

**Tabel 4.3 Kombinasi parameter Tuning Kombinasi 3**

Vector_Size	Windows	DM	Epoch	Hasil Val Sum
50	5	0	20	0.9587
50	5	0	30	0.9520
50	5	0	40	0.9454

**Tabel 4.4 Kombinasi parameter Tuning Kombinasi 4**

Vector_Size	Windows	DM	Epoch	Hasil Val Sum
50	5	1	20	0.8581
50	5	1	30	0.8731
50	5	1	40	0.8712

**Tabel 4.5 Kombinasi parameter Tuning Kombinasi 5**

Vector_Size	Windows	DM	Epoch	Hasil Val Sum
50	7	0	20	0.9581
50	7	0	30	0.9518
50	7	0	40	0.9455

**Tabel 4.6 Kombinasi parameter Tuning Kombinasi 6**

Vector_Size	Windows	DM	Epoch	Hasil Val Sum
50	7	1	20	0.7950
50	7	1	30	0.8185
50	7	1	40	0.8112

**Tabel 4.7 Kombinasi parameter Tuning Kombinasi 7**

Vector_Size	Windows	DM	Epoch	Hasil Val Sum
100	3	0	20	0.9587
100	3	0	30	0.9517
100	3	0	40	0.9450

**Tabel 4.8 Kombinasi parameter Tuning Kombinasi 8**

Vector_Size	Windows	DM	Epoch	Hasil Val Sum
100	3	1	20	0.9162
100	3	1	30	0.9887
100	3	1	40	0.9121

**Tabel 4.9 Kombinasi parameter Tuning Kombinasi 9**

Vector_Size	Windows	DM	Epoch	Hasil Val Sum
100	5	0	20	0.9585
100	5	0	30	0.9517
100	5	0	40	0.9452

**Tabel 4.10 Kombinasi parameter Tuning Kombinasi 10**

Vector_Size	Windows	DM	Epoch	Hasil Val Sum
100	5	1	20	0.8592
100	5	1	30	0.8655
100	5	1	40	8565

**Tabel 4.11 Kombinasi parameter Tuning Kombinasi 11**

Vector_Size	Windows	DM	Epoch	Hasil Val Sum
100	7	0	20	0.9584
100	7	0	30	0.9516
100	7	0	40	0.9451

**Tabel 4.12 Kombinasi parameter Tuning Kombinasi 12**

Vector_Size	Windows	DM	Epoch	Hasil Val Sum
100	7	1	20	0.8150
100	7	1	30	0.8215
100	7	1	40	0.888

**Tabel 4.13 Kombinasi parameter Tuning Kombinasi 13**

Vector_Size	Windows	DM	Epoch	Hasil Val Sum
150	3	0	20	0.9582
150	3	0	30	0.9513
150	3	0	40	0.9449

**Tabel 4.14 Kombinasi parameter Tuning Kombinasi 14**

Vector_Size	Windows	DM	Epoch	Hasil Val Sum
150	3	1	20	0.9070
150	3	1	30	0.9021
150	3	1	40	0.8940

**Tabel 4.15 Kombinasi parameter Tuning Kombinasi 15**

Vector_Size	Windows	DM	Epoch	Hasil Val Sum
150	5	0	20	0.9581
150	5	0	30	0.9512
150	5	0	40	0.9448

**Tabel 4.16 Kombinasi parameter Tuning Kombinasi 16**

Vector_Size	Windows	DM	Epoch	Hasil Val Sum
150	5	1	20	0.8445
150	5	1	30	0.8481
150	5	1	40	0.8358

**Tabel 4.17 Kombinasi parameter Tuning Kombinasi 17**

Vector_Size	Windows	DM	Epoch	Hasil Val Sum
150	7	0	20	0.9581
150	7	0	30	0.9512
150	7	0	40	0.9449

**Tabel 4.18 Kombinasi parameter Tuning Kombinasi 18**

Vector_Size	Windows	DM	Epoch	Hasil Val Sum
150	7	1	20	0.7818
150	7	1	30	0.7789
150	7	1	40	0.7875

Berdasarkan hasil eksperimen tersebut, kombinasi parameter terbaik diperoleh pada epoch = 20, vector\_size = 50, dan windows= 3 dan DM = 0. dengan *Valuasi Similarity* sebesar 95.87%. Konfigurasi ini digunakan sebagai konfigurasi akhir dalam pelatihan model yang diterapkan pada sistem. Peningkatan jumlah epoch dan pemilihan vector\_size yang sesuai terbukti memberikan dampak signifikan terhadap akurasi akhir.

#### 4.4 Implementasi dan Pengujian Model

Pada tahap pelatihan model sebelumnya sudah diperoleh *Best\_Parameter* yang digunakan pada model. Proses pencarian *Best\_Parameter* ini disebut sebagai *Hyperparameter tuning*. Berdasarkan keseluruhan proses pelatihan yang dilakukan sebagaimana sudah dijelaskan pada Tabel 4.1 diperoleh nilai parameter terbaik dan model terbaik yaitu epoch = 20, vector\_size = 50, dan windows= 3 dan DM = 0. dengan *Valuasi Similarity* sebesar 95.87%.

Parameter—parameter yang dihasilkan akan diterapkan pada model. Model yang telah di latih selama tahap pelatihan akan disimpan dan selanjutnya diuji untuk mengukur kinerja model berdasarkan algoritma yang telah digunakan. Dalam pengujian model ini penulis menggunakan data baru yang sebelumnya tidak pernah dipelajari oleh model. Pengujian model dilakukan untuk mengukur performa sistem dalam mendeteksi kemiripan dokumen *executive summary* terhadap dokumen skripsi yang ada di database. Pengujian ini menggunakan 50 dokumen uji yang tidak termasuk dalam data pelatihan (*unseen data*). Hasil pengujian model dapat dilihat pada Tabel 4.1.

**Tabel 4. 1 Tabel Pengujian Model Terhadap Data Exume**

No	Nama File	File Skripsi	Skor Kemiripan
1	Exum_201402011.pdf	KRIPSI_201402011.pdf	0.9282
		SKRIPSI_201402146.pdf	0.9075
		SKRIPSI_201402097.pdf	0.8902
		SKRIPSI_201402115.pdf	0.8828
		SKRIPSI_201401010.pdf	0.8758
2	EXUME_201402023	KRIPSI_171401147.pdf	0.8153
		SKRIPSI_201401139.pdf	0.8064
		SKRIPSI_191402036.pdf	0.7910
		SKRIPSI_211401059.pdf	0.7845
		SKRIPSI_171401049.pdf	0.7801

## 4.5 Evaluasi Model

### 4.4.1 Metrik Evaluasi

Berdasarkan Tabel 4.1 hasil di atas menunjukkan bahwa dokumen dengan nama file Skripsi\_201402011.pdf memiliki skor similarity tertinggi sebesar **0.9282** terhadap dokumen exume yang diuji. Setelah dilakukan pemeriksaan manual terhadap isi dokumen tersebut, ditemukan bahwa topik, tujuan, dan struktur kalimat dalam skripsi tersebut memang memiliki kemiripan dengan dokumen exume.

Model Doc2Vec yang dibangun dapat mengenali kemiripan semantik antar dokumen dengan cukup baik, terlihat dari skor similarity yang tinggi pada beberapa dokumen yang memiliki topik sejenis. Hal ini menunjukkan bahwa sistem yang dibangun memiliki kemampuan untuk membantu proses deteksi kemiripan konten pada proses pengajuan skripsi.

### 4.4.2 Hasil Pengujian

Seperti halnya dalam data latih, data uji juga akan melewati enam tahapan preprocessing yaitu cleaning, case folding, tokenization, lemmitization, dan stopword removal. Setelah melalui tahap preprocessing, setiap kata dalam data uji akan diubah menjadi indeks berupa urutan integer. Selanjutnya, integer ini akan dilakukan prediksi menggunakan model yang sebelumnya telah dilatih. Hasil pengujian model dapat dilihat pada Tabel 4.2

**Tabel 4. 2 Hasil Pengujian Model**

No	File Exume	File Skripsi	Skore
	Exum_211402151	SKRIPSI_201402081.pdf	0.2745
		SKRIPSI_201401025.pdf	0.2516
		SKRIPSI_191401108.pdf	0.2445
		SKRIPSI_181402128.pdf	0.2392
		SKRIPSI_201402128.pdf	0.2381

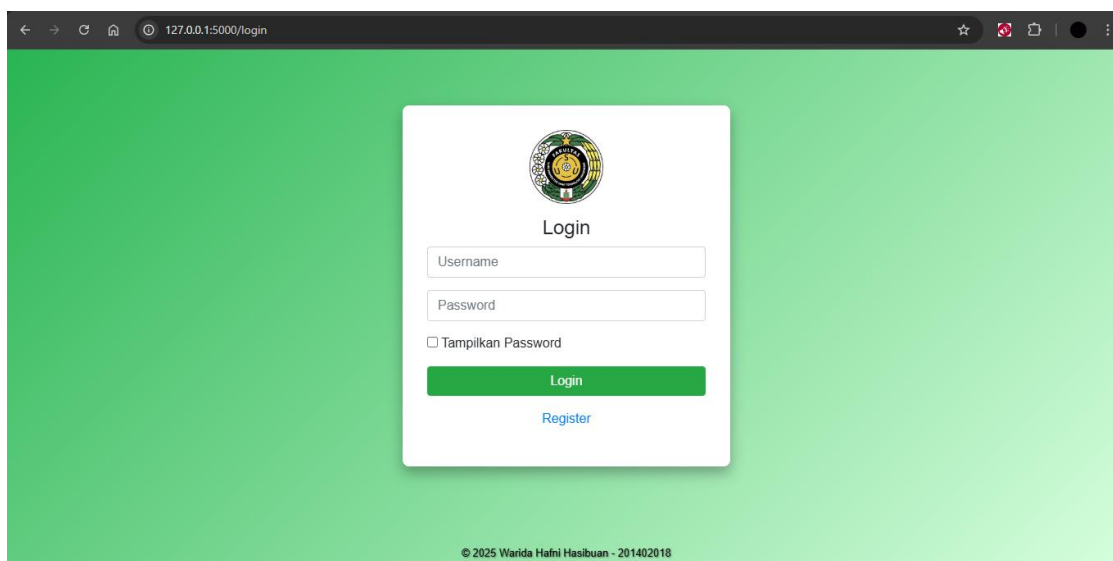
## 4.6 Implementasi Perancangan Antarmuka Aplikasi

Implementasi antarmuka aplikasi merupakan tahap realisasi dari desain UI/UX yang telah dirancang sebelumnya menjadi tampilan fungsional yang dapat digunakan secara langsung oleh pengguna. Proses implementasi dilakukan menggunakan kombinasi teknologi web seperti HTML5, CSS3, JavaScript, serta templating engine Jinja2 dari framework Flask. Seluruh tampilan antarmuka dirancang responsif agar dapat digunakan dengan baik pada perangkat desktop maupun mobile.

Berikut adalah hasil implementasi dari setiap halaman utama dalam aplikasi sistem deteksi kemiripan dokumen:

### 4.6.1 Halaman Login

Halaman login berfungsi sebagai gerbang awal bagi pengguna untuk mengakses sistem. Pengguna diminta memasukkan username dan password yang sesuai. Apabila kredensial tidak valid, sistem akan menampilkan notifikasi kesalahan. Tampilan halaman ini didesain sederhana dan profesional, dengan menyertakan logo institusi di bagian atas sebagai identitas sistem.

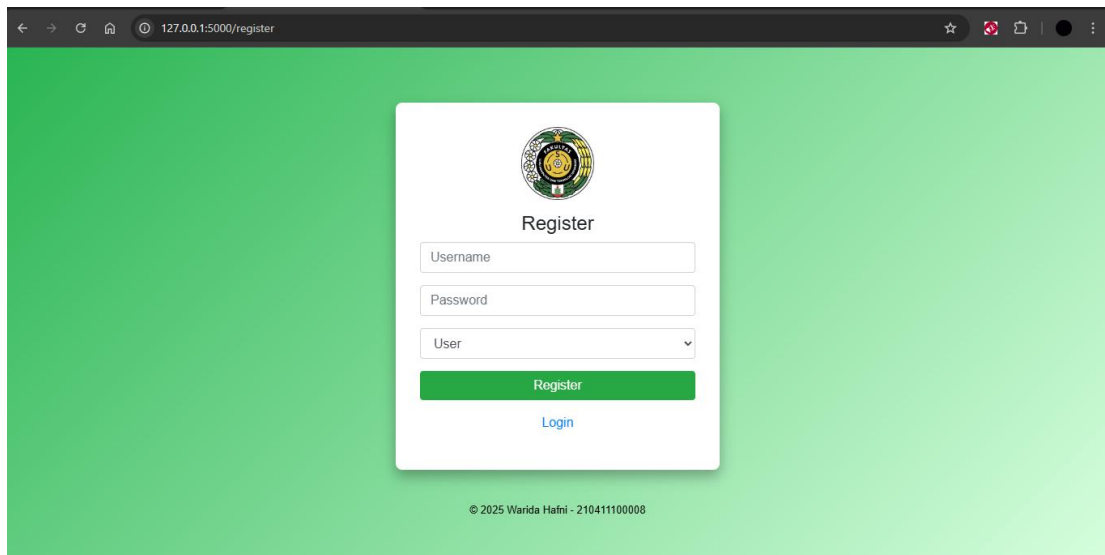


**Gambar 4. 1 Tampilan Halaman Login**

### 4.6.2 Halaman Register

Halaman ini digunakan oleh pengguna baru untuk membuat akun. Formulir registrasi mencakup kolom username dan password, serta pilihan role yang secara default diatur

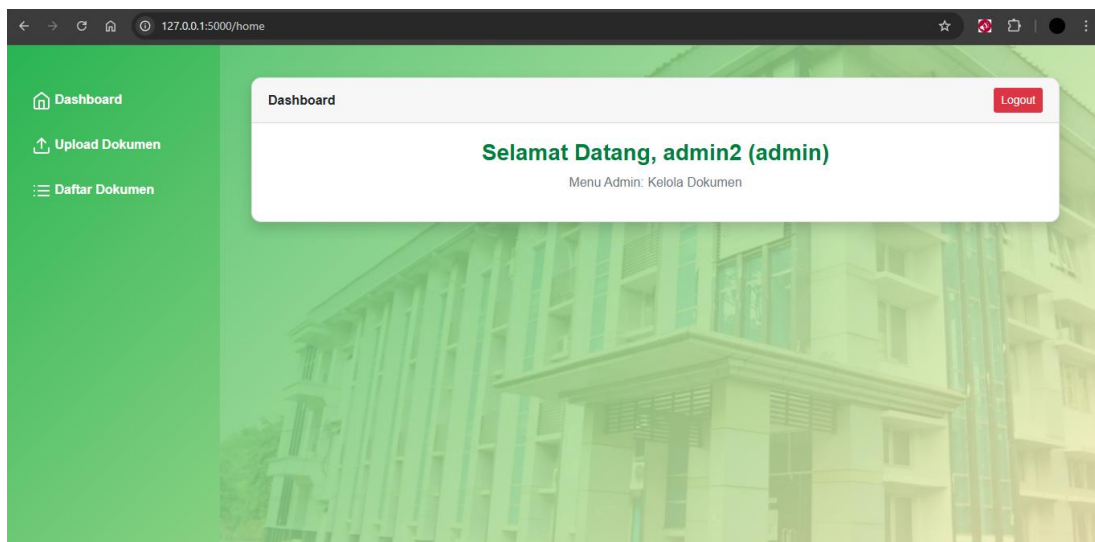
sebagai "User". Validasi input diterapkan untuk memastikan data yang dimasukkan benar sebelum dikirim ke server.

A screenshot of a web browser showing a registration form. The browser's address bar displays '127.0.0.1:5000/register'. The form is centered on a green background and features a circular logo at the top. Below the logo, the word 'Register' is displayed. The form includes three input fields: 'Username', 'Password', and a dropdown menu labeled 'User'. A green 'Register' button is positioned below these fields, with a blue 'Login' link underneath it. At the bottom of the form, a copyright notice reads '© 2025 Warida Hafni - 210411100008'.

**Gambar 4. 2 Tampilan Halaman Register**

#### *4.6.3 Dashboard Admin*

Setelah berhasil login, admin diarahkan ke halaman dashboard yang berfungsi sebagai pusat navigasi. Di sisi kiri halaman terdapat menu navigasi seperti Dashboard, Upload Dokumen, dan Riwayat Dokumen. Di sisi kanan ditampilkan informasi sambutan dan identitas admin yang sedang login.

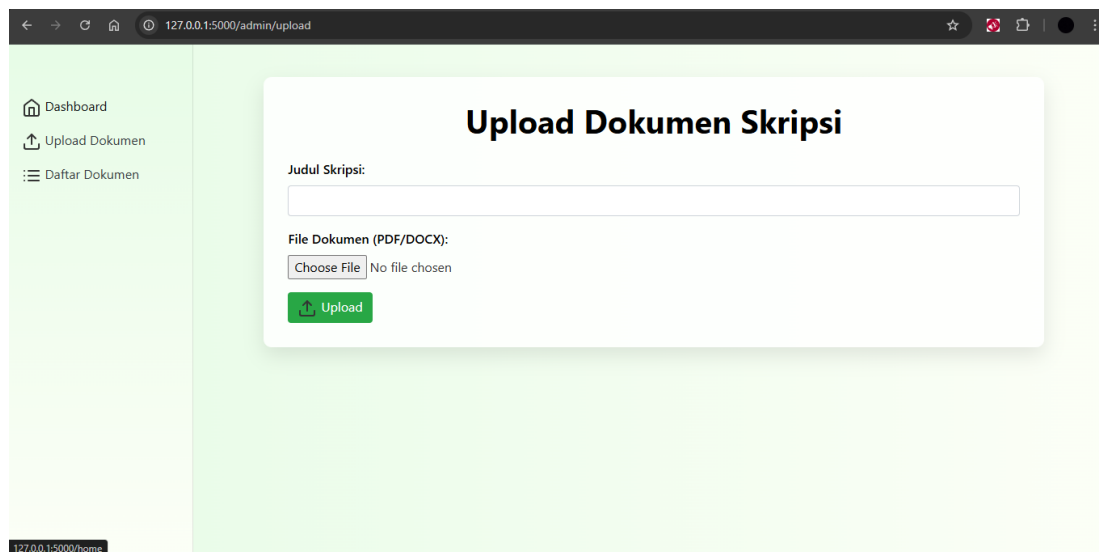


**Gambar 4. 3 Tampilan Halaman Dashboard Admin**



#### 4.6.4 Halaman Upload Dokumen (Admin)

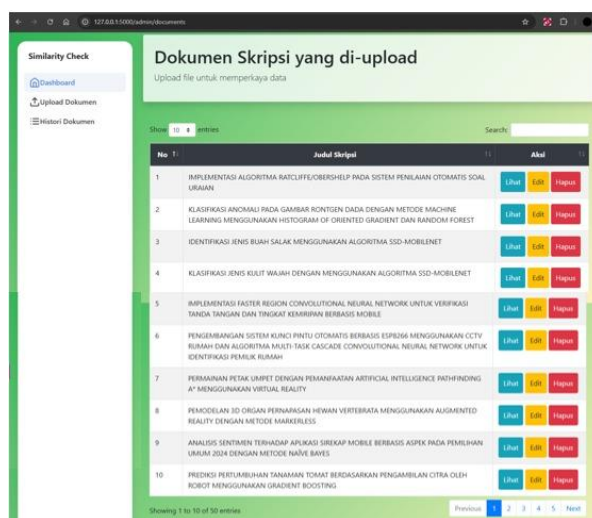
Halaman ini memungkinkan admin untuk mengunggah dokumen skripsi ke dalam sistem. Admin dapat mengisi judul dokumen, memilih file dalam format PDF, lalu mengunggah dokumen tersebut. Setelah berhasil diunggah, sistem akan secara otomatis menyimpan dokumen ke dalam basis data dan menampilkannya pada daftar dokumen.



**Gambar 4. 4 Tampilan Halaman Upload Dokumen Bagian Admin**

#### 4.6.5 Halaman Daftar Dokumen (Admin)

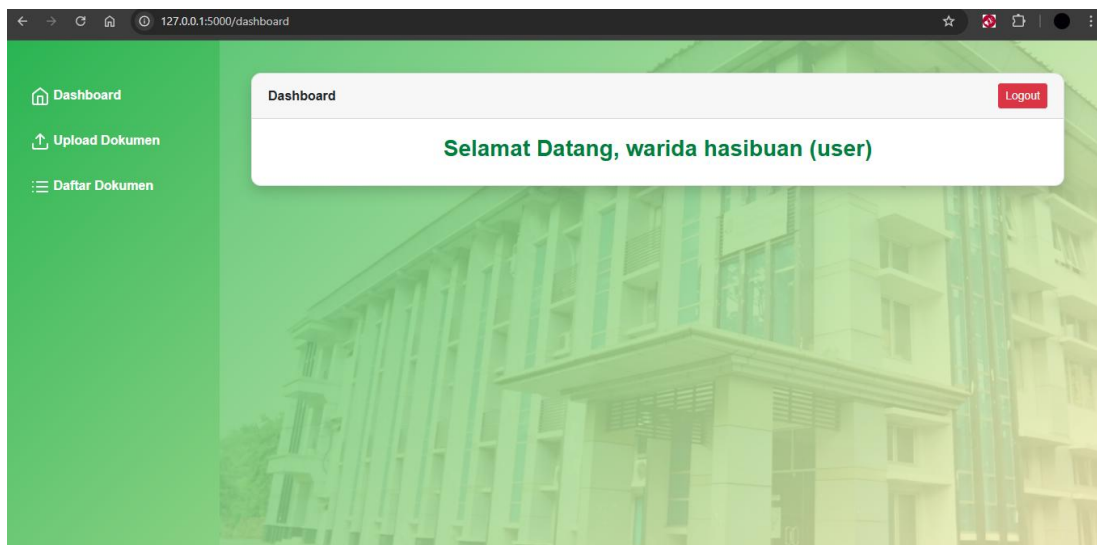
Halaman ini menampilkan seluruh dokumen skripsi yang telah diunggah ke dalam sistem. Informasi yang ditampilkan berupa nomor urut, judul dokumen, nama file, isi ringkas dokumen, dan waktu unggah. Tampilan data disajikan dalam bentuk tabel agar mudah dibaca dan dikelola.



**Gambar 4. 5 Tampilan Halaman Daftar Dokumen Bagian Admin**

#### 4.6.6 Dashboard User

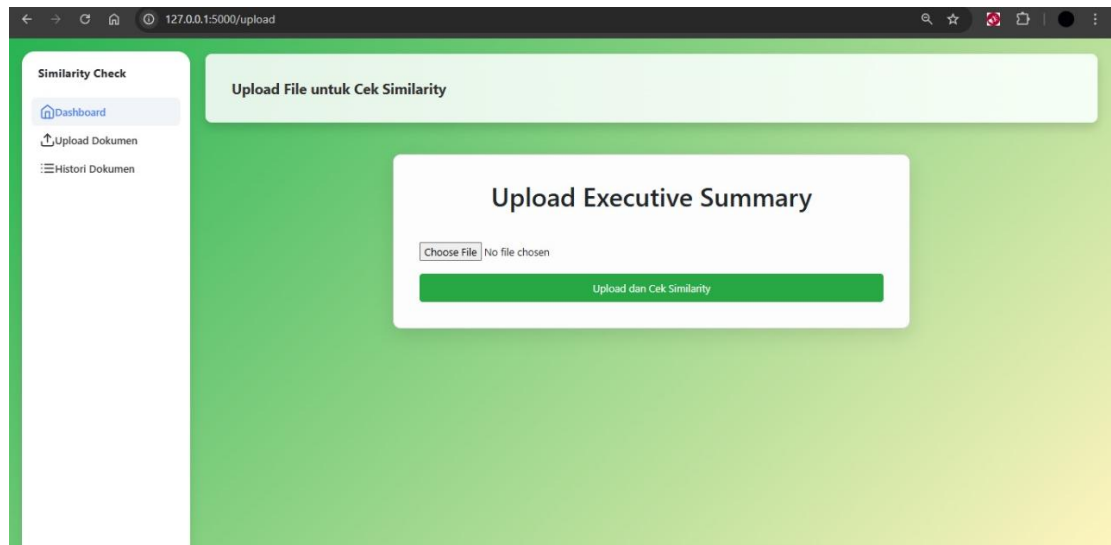
Pengguna umum (user) juga diarahkan ke dashboard setelah login. Tampilan halaman ini menampilkan informasi sambutan serta menu navigasi yang terdiri dari Dashboard, Upload Dokumen, dan Riwayat Dokumen. User juga dapat melihat identitas akunnya dan keluar dari sistem melalui tombol logout.



**Gambar 4. 6 Tampilan Halaman Dashboard User**

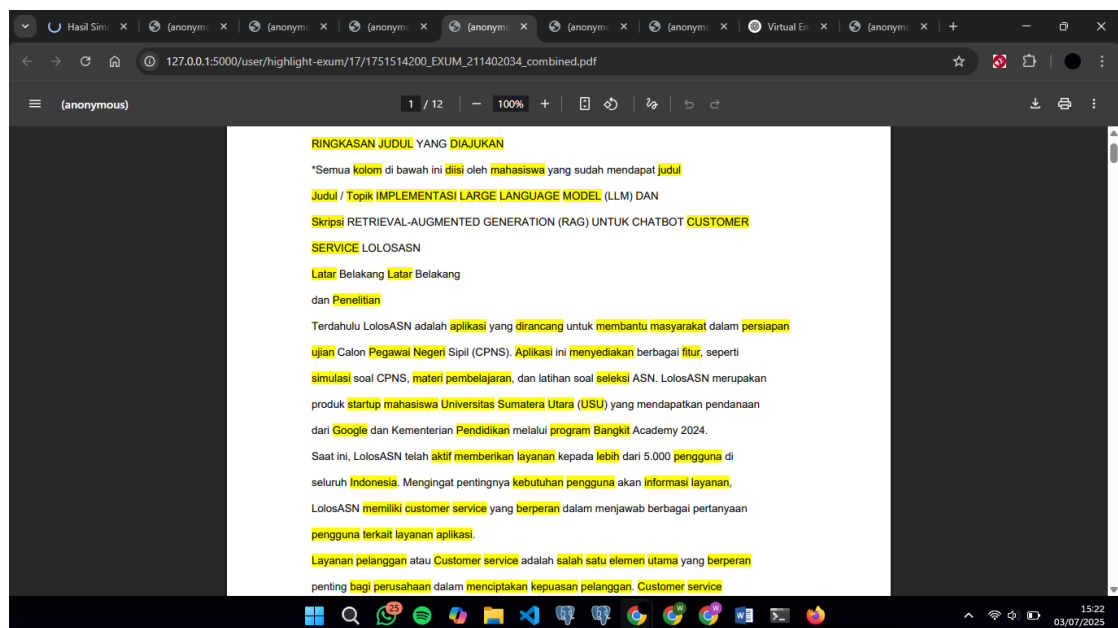
#### 4.6.7 Halaman Upload Dokumen (User)

Halaman ini memungkinkan user untuk mengunggah dokumen executive summary (Exum) yang akan diperiksa tingkat kemiripannya dengan dokumen-dokumen skripsi yang ada di basis data. Setelah file PDF diunggah, sistem akan secara otomatis memproses dokumen tersebut dan menampilkan hasil perbandingan similarity.



**Gambar 4. 7 Tampilan Halaman Upload Dokumen Bagian User**

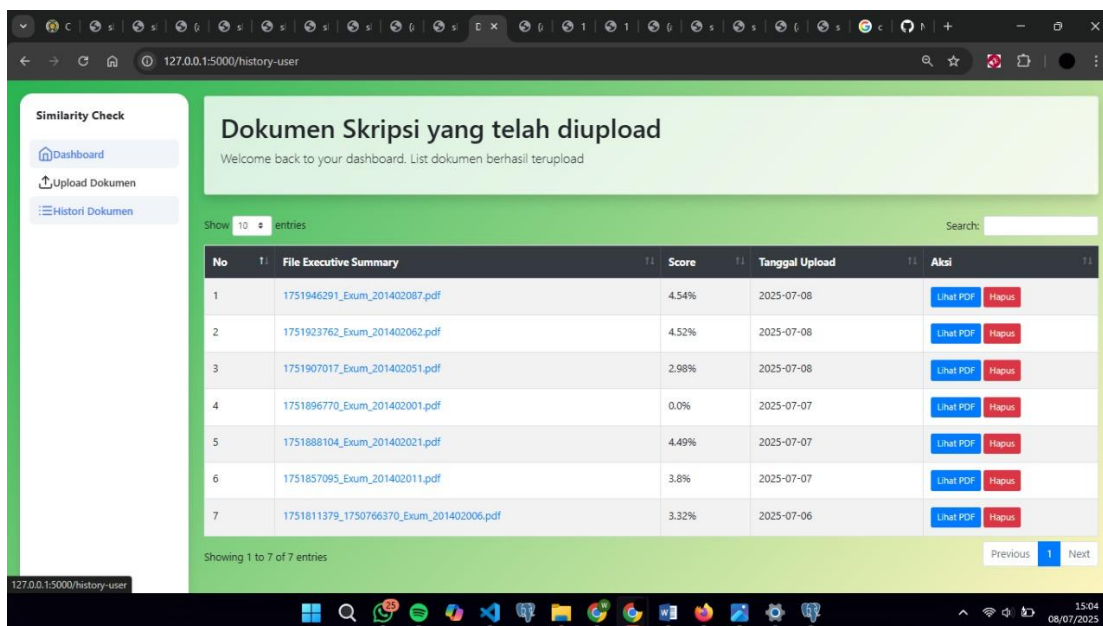
#### 4.6.8 Halaman Similarity exum



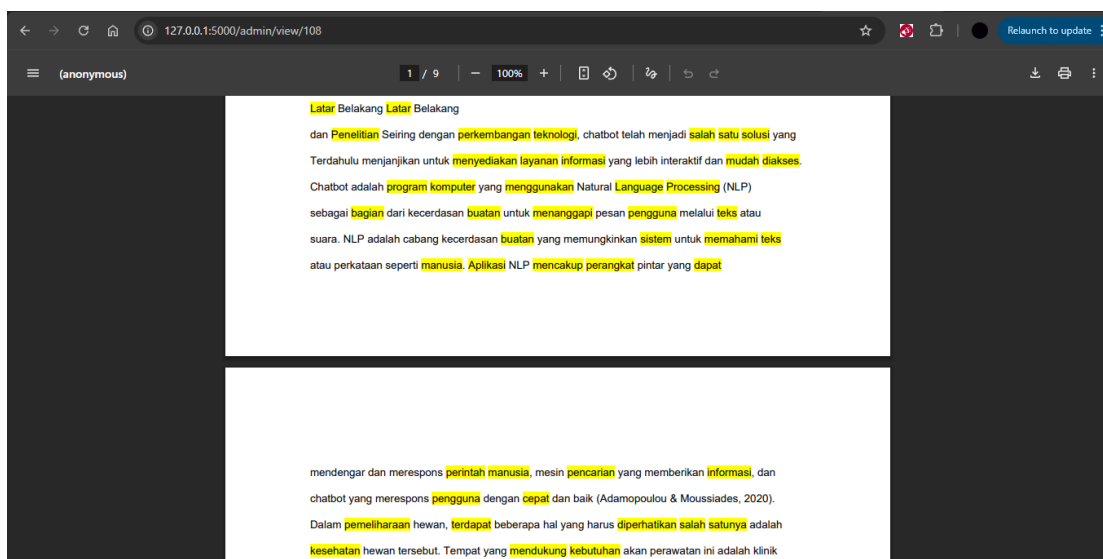
**Gambar 4. 8 Tampilan Halaman Similarity Exum**

#### 4.6.9 Halaman Hasil Similarity

Setelah proses analisis selesai, hasil similarity ditampilkan dalam bentuk tabel. Tabel tersebut memuat informasi berupa nomor, judul dokumen skripsi yang memiliki kemiripan tertinggi, skor kemiripan dalam bentuk persentase, serta tombol aksi untuk melihat dokumen yang relevan secara lebih rinci.



**Gambar 4. 9 Tampilan Halaman Dokumen Hasil Similarity**

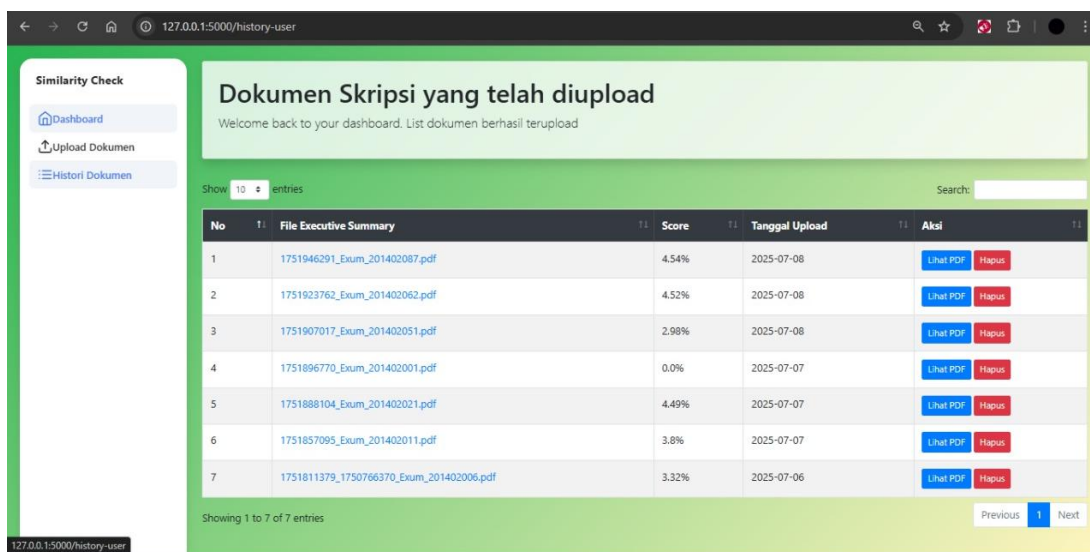


**Gambar 4. 10 Tampilan Hasil Similarity**

#### 4.6.10 Halaman Riwayat Dokumen

Halaman ini berisi daftar dokumen yang pernah diunggah oleh user, lengkap dengan hasil perbandingan similarity. Informasi yang ditampilkan meliputi nomor, judul

dokumen, skor similarity, kata-kata yang dianggap relevan, tanggal unggah, dan opsi untuk melihat detail hasil analisis. Dimana pada bagian ini , user dapat melihat *Executive Summary* yang sudah di cek *Similarity* nya serta akan diberitahukan score kemiripannya terhadap refensi skripsi lain yang dimasukkan pada *database*.



**Dokumen Skripsi yang telah diupload**  
Welcome back to your dashboard. List dokumen berhasil terupload

Show 10 entries Search:

No	File Executive Summary	Score	Tanggal Upload	Aksi
1	1751946291_Exum_201402087.pdf	4.54%	2025-07-08	Lihat PDF Hapus
2	1751923762_Exum_201402062.pdf	4.52%	2025-07-08	Lihat PDF Hapus
3	1751907017_Exum_201402051.pdf	2.98%	2025-07-08	Lihat PDF Hapus
4	1751896770_Exum_201402001.pdf	0.0%	2025-07-07	Lihat PDF Hapus
5	1751888104_Exum_201402021.pdf	4.49%	2025-07-07	Lihat PDF Hapus
6	1751857095_Exum_201402011.pdf	3.8%	2025-07-07	Lihat PDF Hapus
7	1751811379_1750766370_Exum_201402006.pdf	3.32%	2025-07-06	Lihat PDF Hapus

Showing 1 to 7 of 7 entries Previous 1 Next

**Gambar 4. 11 Tampilan Halaman Riwayat Dokumen**

## **BAB 5**

### **KESIMPULAN DAN SARAN**

#### **4.1 Kesimpulan**

1. Dari hasil penelitian ini diperoleh nilai akurasi sebesar 78%.
2. Pada penelitian ini diperoleh nilai dari metrik evaluasi *precision* kelas tinggi lebih tinggi, yaitu mencapai 95%, sedangkan pada kelas sedang dan kelas rendah mempunyai nilai *precision* yang sama, yaitu sebesar 89%. Model yang dibuat juga telah sesuai, yang dimana hal tersebut dapat dibuktikan melalui nilai *recall*, dimana nilai *recall* untuk masing-masing kelas tinggi, kelas sedang dan kelas rendah sebesar 90%, 80% dan 92%. Nilai *f1-score* pada kelas tinggi sebesar 92%, kelas sedang sebesar 84% dan kelas rendah sebesar 90%.
3. Dari hasil kombinasi *hyperparameter* yang telah dilakukan diperoleh parameter terbaik pada epoch sebesar 50, batch size sebesar 16, dan learning rate berniali 0.00001 dengan akurasi sebesar 95.56% dan nilai loss sebesar 0.2571.

#### **4. TAMBAHI YAH WAR**

#### **4.2 Saran**

1. **masukin ya war karena aku gak tahu dimana kelemahan sistem kamu**

## DAFTAR PUSTAKA

- Cahyono, S. C. (2019). *Model Perbandingan Dokumen Karya Menggunakan Algoritma Kesamaan Dokumen Doc2vec*. 75117018.
- Dalianis, H. (2018). *Evaluation Metrics and Evaluation*. 1967, 45–53.
- Jurafsky, D., & Martin, J. H. (2024). *Speech and Language Processing An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models Third Edition draft Summary of Contents*.
- Le, Q., Mikolov, T., & Com, T. G. (2014). *Distributed Representations of Sentences and Documents*. 32.
- Malte Ostendroff, Terry Ruas, Till Blume, Bela Gipp, G. R. (2020). Aspect-based Document Similarity for Research Papers. *COLING 2020 - 28th International Conference on Computational Linguistics, Proceedings of the Conference, October*, 6194–6206. <https://doi.org/10.18653/v1/2020.coling-main.545>
- Mustariani. (2023). *Pengaruh Teknologi Terhadap Perkembangan Pendidikan Siswa di Era Digital*.
- Oppi Anda Resta, Addin Aditya, F. E. P. (2021). Plagiarism Detection in Students' Theses Using The Cosine Similarity Method. *Sinkron*, 5(2), 305–313. <https://doi.org/10.33395/sinkron.v5i2.10909>
- Pawestri, S., & Suyanto, Y. (2024). Analisis Perbandingan Metode Similarity untuk Kemiripan Dokumen Bahasa Indonesia pada Deteksi Kemiripan Teks Bahasa Indonesia. *Jurnal Media Informatika Budidarma*, 8, 1440–1450. <https://doi.org/10.30865/mib.v8i3.7648>
- Pratama, R. P., Faisal, M., & Hanani, A. (2019). Deteksi Plagiarisme pada Dokumen Jurnal Menggunakan Metode Cosine Similarity. *SMARTICS Journal*, 5(1), 22–26. <https://doi.org/10.21067/smartics.v5i1.2848>
- Ramadhanti, N. R. S. M. (2019). *Document Similarity Detection Using Indonesian Language Word2vec Model*. 1–6.
- Setha Imene, & Hassina Aliane. (2022). Enhancing automatic plagiarism detection: Using Doc2vec model. *ICAASE 2022 - 5th Edition of the International Conference on Advanced Aspects of Software Engineering, Proceedings*, 2024. <https://doi.org/10.1109/ICAASE56196.2022.9931542>
- Sutrisno, E., Rochmatika, E., Mahyuni, E. T., Soetijono, I. K., Mayasari, E., Widodo, M. L., & Yuniarti, E. (n.d.). *Plagiarisme Dan Integritas Akademik*.
- Syukry Ansis, Endang Palupi Listyaningsih, Prof. Dr. Ir. Hari Soetanto,

- S.kom, M. S. (2024). Deteksi Plagiat Tesis Berbahasa Indonesia Menggunakan Metode Cosine Similarity. *INOVTEK Polbeng - Seri Informatika*, 9(1), 153–167. <https://doi.org/10.35314/isi.v9i1.4003>
- Wahyuni, R. T., Prastiyanto, D., & Suprpto, E. (2017). Penerapan Algoritma Cosine Similarity dan Pembobotan TF-IDF pada Sistem Klasifikasi Dokumen Skripsi. *Jurnal Teknik Elektro Universitas Negeri Semarang*, 9(1), 18–23. <https://journal.unnes.ac.id/nju/index.php/jte/article/download/10955/6659>
- Wibowo, A. (2012). Preventing and Solving Plagiarism in Educational Institutions. *Jurnal Kesehatan Masyarakat Nasional*, 6(5), 195–200.
- Yogesh Wadekar, Tushar Shendge, Manali Dhokale, Vaishnavi Ohol, P. S. D. (2021). Plagiarism Detection with Paraphrase Recognizer Using Deep Learning. *International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)*, 6(1), 1046–1053. <https://doi.org/10.48175/568>