**PROPOSAL TUGAS AKHIR**

**IMPLEMENTASI METODE DOC2VEC DAN HDBSCAN PADA** **SISTEM DETEKSI PLAGIARISME BERBASIS WEBSITE**



**Oleh:**

|  |  |
| --- | --- |
| **Bondan Tiur Mahendra** | **(202211420067)** |
|  |  |

**PROGRAM STUDI TEKNIK INFORMATIKA**

**FAKULTAS TEKNIK**

**UNIVERSITAS DR. SOETOMO**

**SURABAYA**

**2024**

**LEMBAR PENGESAHAN  
PROPOSAL TUGAS AKHIR  
<<judul>>**

<<Nama>>  
<<NIM. >>

Disetujui Oleh:

|  |  |  |
| --- | --- | --- |
| Pembimbing  <<Nama Dosen>>  <<NPP>> | | |
|  | Mengetahui: |  |
| Dekan Fakultas Teknik  <<Nama Dekan>>  <<NPP. >> |  | Ketua Jurusan Teknik Informatika  <<Nama KaProdi>>  <<NPP. >> |

**LEMBAR PENGESAHAN  
PROPOSAL TUGAS AKHIR  
<<Judul>>**

<<Nama MHS>>  
<<NIM. >>

|  |  |
| --- | --- |
| Penguji | Tanda Tangan |
| NPP. | ……………………………………….. |
| NPP. | ………………………………………... |
| NPP. | ………………………………………... |

Mengetahui

|  |  |
| --- | --- |
| Dekan Fakultas Teknik  <<Nama Dekan>>  <<NPP. >> | Ketua Jurusan Teknik Informatika  <<Nama KaProdi>>  <<NPP. >> |

# DAFTAR ISI

[DAFTAR ISI 4](#_Toc180333988)

[DAFTAR GAMBAR 5](#_Toc180333989)

[DAFTAR TABEL 6](#_Toc180333990)

[BAB 1 PENDAHULUAN 1](#_Toc180333991)

[1.1. Latar Belakang Masalah 1](#_Toc180333992)

[1.2. Rumusan Masalah 3](#_Toc180333993)

[1.3. Batasan Masalah 3](#_Toc180333994)

[1.4. Tujuan 3](#_Toc180333995)

[1.5. Manfaat 4](#_Toc180333996)

[1.6. Metodologi Penelitian 4](#_Toc180333997)

[1.7. Lokasi Penelitian 6](#_Toc180333998)

[BAB 2 LANDASAN TEORI 7](#_Toc180333999)

[2.1. Landasan Teori 1 7](#_Toc180334000)

[2.1.1. Sub Sub Bab 1 Landasan Teori 1 8](#_Toc180334001)

[2.1.2. Sub Sub Bab 2 Landasan Teori 1 8](#_Toc180334002)

[2.2. Landasan Teori 2 8](#_Toc180334003)

[2.3. Landasan Teori 3 8](#_Toc180334004)

[BAB 3 ANALISIS DAN DESAIN 10](#_Toc180334005)

[3.1. Deskripsi Sistem 10](#_Toc180334006)

[3.2. Pelaku Sistem 10](#_Toc180334007)

[3.3. Kebutuhan Data 10](#_Toc180334008)

[3.4. Arsitektur Sistem 10](#_Toc180334009)

[3.5. Blok Diagram 11](#_Toc180334010)

[3.6. Formulasi Kecerdasan Buatan 11](#_Toc180334011)

[3.7. Model Proses 12](#_Toc180334012)

[3.7.1. Use Case Diagram 12](#_Toc180334013)

[3.7.2. Activity Diagram 13](#_Toc180334014)

[3.7.3. Sequence Diagram 14](#_Toc180334015)

[3.7.4. Class Diagram 14](#_Toc180334016)

[3.8. Desain Proses 14](#_Toc180334017)

[3.9. Desain GUI 14](#_Toc180334018)

# DAFTAR GAMBAR

[Gambar 2.1 Contoh Gambar Landasan Teori 2 4](#_3j2qqm3)

[Gambar 3.1 Contoh Arsitektur Sistem 7](#_49x2ik5)

[Gambar 3.2 Blok Diagram Identifikasi Citra 7](#_147n2zr)

[Gambar 3.3 Flowchart Pemrosesan Citra 8](#_23ckvvd)

[Gambar 3.4 Contoh Context Diagram 9](#_1hmsyys)

[Gambar 3.5 Contoh Use Case Diagram 9](#_2grqrue)

[Gambar 3.6 Contoh Activity Diagram 10](#_vx1227)

[Gambar 3.7 Contoh Sequence Diagram 11](#_3fwokq0)

[Gambar 3.8 Contoh Entity Relationship Diagram 11](#_4f1mdlm)

# DAFTAR TABEL

[Tabel 2.1 Isikan Judul Tabel 1 3](#_35nkun2)

[Tabel 2.2 Isikan Judul Tabel 2 5](#_4i7ojhp)

# PENDAHULUAN

## Latar Belakang Masalah

Perkembangan teknologi di era digital telah menghadirkan aplikasi pembelajaran daring sebagai metode populer yang memudahkan akses materi pembelajaran dan interaksi antara dosen dengan mahasiswa tanpa batasan geografis. Perkembangan teknologi yang pesat ini memberikan fleksibilitas kepada mahasiswa dalam mengakses berbagai sumber pembelajaran dan menciptakan lingkungan belajar yang lebih inklusif. Keuntungan dari aplikasi pembelajaran daring tersebut membawa tantangan baru bagi dosen dalam mengawasi integritas akademik dan menjaga kualitas karya ilmiah mahasiswa. Para dosen memiliki peran strategis dalam pengawasan dan pencegahan plagiarism di lingkungan pembelajaran daring. Minimnya pengawasan terhadap kejujuran akademik ini dapat mengancam integritas karya ilmiah dan menghambat perkembangan keilmuan di perguruan tinggi [1].

Permasalahan dari penelitian ini adalah maraknya tindakan plagiarism berupa pengambilan dan penggunaan karya orang lain tanpa mencantumkan sumber yang tepat sebagai pelanggaran serius terhadap etika akademik. Pembelajaran daring melalui aplikasi digital membuka peluang lebih besar bagi mahasiswa untuk melakukan tindakan plagiarisme karena kemudahan dalam mengakses dan berbagi informasi dari berbagai sumber [1]. Ketersediaan konten digital ini menyebabkan mahasiswa cenderung mengabaikan prinsip amati, tiru, dan modifikasi dalam mengembangkan karya ilmiah. Kondisi tersebut memunculkan urgensi di Universitas Dr. Soetomo untuk mengembangkan mekanisme yang efektif dan handal dalam mendeteksi kemiripan teks pada dokumen-dokumen yang diunggah mahasiswa guna mencegah dan mengurangi tindakan plagiarisme yang dapat merugikan proses pembelajaran dan penilaian.

Berdasarkan studi literature pada penelitian sebelumnya yang membahas tentang pendeteksian plagiarisme. Peneliti [2] dengan dataset berupa jurnal di Website atau Blog menggunakan metode Cosine Similarity memperoleh hasil *recall* sebesar 13% dan *precision* sebesar 8%. Hasil penelitian berikutnya [3], dengan dataset berupa tesis di repositori Universitas Budi Luhur menghasilkan nilai akurasi metode Cosine Similarity sebesar 96.63% dan metode Jaccard Similarity sebesar 50.5%. Sedangkan pada penelitian lain [4] dengan dataset *Plagiarism Analysis, Authorship Identification, and Near-Duplicate Detection* (PAN) menggunakan metode Yake dan Sentence Transformer memperoleh nilai *F-score* sebesar 32%. Akan tetapi, belum ada peneliti yang menggunakan metode Doc2vec untuk pembentukan embedding dan metode *Hierarchical Density-Based Spatial Clustering of Applications with Noise* (HDBSCAN) untuk pengelompokan dokumen mirip.

Tujuan dari penelitian ini adalah untuk membangun aplikasi yang mampu mendeteksi kemiripan teks pada dokumen-dokumen yang diunggah, untuk meningkatkan integritas akademik. Dataset yang digunakan dalam penelitian ini berupa dokumen akademik di repositori Universitas Dr Soetomo. Metode Doc2vec digunakan dalam penelitian ini untuk menghasilkan representasi vektor numerik dari suatu dokumen dengan memperhitungkan konteks kata-kata yang terdapat dalam dokumen tersebut [5], sementara metode HDBSCAN digunakan untuk dapat mengelompokkan data berdasarkan kepadatan dan bersifat *robust* yang artinya tahan terhadap pencilan (*outliers*)[6]. Model evaluasi metode dalam penelitian ini menggunakan *f1-score* [4].

Berdasarkan perumusan masalah diatas, maka penelitian ini diberi judul **“Implementasi Metode Doc2vec Dan HDBSCAN Pada Sistem Deteksi Plagiarisme Berbasis Website”**. Manfaat dari hasil penelitian ini dapat dimanfaatkan oleh dosen atau pengajar untuk memberikan hasil pengecekan kemiripan teks dengan waktu yang lebih efisien dibandingkan dengan metode manual.

## Rumusan Masalah

Berdasarkan permasalahan diatas, maka dapat diambil rumusan masalah sebagai berikut :

1. Belum adanya aplikasi dapat mendeteksi kemiripan dokumen di Universitas Dr. Soetomo.
2. Bagaimana mengetahui nilai *f1-score* metode Doc2Vec dan HDBSCAN untuk membangun sistem yang dapat mendeteksi plagiarisme?

## Batasan Masalah

Batasan masalah pada penelitian ini adalah :

1. Dataset yang digunakan dalam penelitian ini dokumen-dokumen teks dalam format (.docx), dan tidak mencakup tipe file lain seperti gambar atau video. Selain itu, dokumen hanya dalam bahasa Indonesia untuk menjaga konsistensi hasil embedding.
2. Dataset berupa tugas atau esai didapat dari repositori Universitas Dr Soetomo.
3. Penelitian ini menggunakan metode Doc2Vec untuk menghasilkan embedding teks dokumen.
4. Metode HDBSCAN untuk pengelompokan dokumen miripdan *Cosine Similarity* sebagai pengukur kemiripan dokumen.
5. Evaluasi kinerja model dilakukan menggunakan metrik *f1-score*.

## Tujuan

Tujuan dari penelitian ini adalah membuat aplikasi dengan memiliki kemampuan sebagai berikut :

1. Mampu mendeteksi kemiripan dokumen sebagai sistem deteksi plagiarisme.
2. Mampu menampilkan nilai *f1-score* dari perhitungan metode Doc2Vec dan HDBSCAN dalam sistem deteksi plagiarisme.

## Manfaat

Berdasarkan tujuan yang telah dijabarkan, manfaat dari penelitian ini adalah:

1. Membantu dosen atau pengajar dapat memberikan hasil pengecekan kemiripan teks dengan waktu yang lebih efisien dibandingkan dengan metode manual.
2. Mengetahui performa metode Doc2vec dan HDBSCAN berdasarkan hasil nilai *f1-score*.

## Metodologi Penelitian

Metode pengembangan perangkat lunak yang digunakan oleh peneliti dalam penelitian ini menggunakan *System Development Life Cycle* (SDLC), SDLC adalah tahapan-tahapan yang dilakukan oleh analisis dalam membangun sistem informasi. Tahapan yang dilakukan agar penelitian bisa terarah dan sesuai perencanaan, fase yang dilakukan di penelitian ialah:

1. Perencanaan (Planning)

Tahap perencanaan adalah langkah awal mengidentifikasi dalam pengembangan sistem, seperti fitur deteksi plagiarisme, antarmuka website, serta rencana evaluasi menggunakan F1-Score. Kelayakan sumber daya dievaluasi untuk memastikan bahwa proyek dapat dilaksanakan sesuai jadwal dan anggaran.

2. Analisis Kebutuhan (Requirements Analysis)

Di tahap analisis kebutuhan, kami memetakan kebutuhan sistem secara detail, termasuk kebutuhan fungsional dan non-fungsional. Analisis dilakukan terhadap alur proses deteksi plagiarisme, kebutuhan pengguna, dan keamanan data, yang akan menjadi dasar pengembangan lebih lanjut.

3. Desain (Design)

Pada tahap desain, kami merancang arsitektur sistem berbasis client-server, database, antarmuka pengguna, dan algoritma. Integrasi Doc2Vec untuk embedding dan HDBSCAN untuk clustering dirancang untuk mendeteksi kemiripan dokumen. Desain antarmuka dilakukan agar pengguna dapat dengan mudah mengunggah dokumen dan melihat hasil deteksi.

4. Pengembangan (Development)

Pada tahap ini, kami membangun aplikasi secara teknis, termasuk frontend dan backend, serta mengintegrasikan algoritma Doc2Vec dan HDBSCAN untuk pemrosesan data. Setiap komponen diuji secara internal untuk memastikan fungsionalitas sebelum pengujian keseluruhan.

5. Pengujian (Testing)

Lakukan pengujian fungsional, performa, dan pengujian bersama pengguna untuk menilai efektivitas, akurasi, dan kemudahan penggunaan aplikasi. Evaluasi F1-Score digunakan untuk memastikan sistem mendeteksi plagiarisme secara akurat.

6. Implementasi (Deployment)

Tahap implementasi meliputi pengembangan frontend untuk antarmuka website menggunakan HTML, CSS, dan JavaScript agar tampilan interaktif, serta pengembangan backend dengan bahasa pemrograman seperti Python untuk menjalankan proses Doc2Vec dan HDBSCAN serta menyimpan data ke basis data.

7. Pemeliharaan (Maintenance)

Setelah implementasi dan evaluasi, tahap pemeliharaan dan perbaikan dilakukan untuk menjaga dan meningkatkan kinerja aplikasi. Lakukan pembaruan rutin serta perbaikan jika ada kesalahan atau masalah yang ditemukan dari umpan balik pengguna.

## Lokasi Penelitian

Penelitian ini akan diterapkan pada lingkungan kampus Universitas Dr. Soetomo. Data dokumen yang akan digunakan untuk uji coba diambil dari dokumen akademik mahasiswa, seperti tugas atau esai, dengan fokus pada pengujian dan penerapan sistem deteksi plagiarisme yang dapat diakses oleh dosen dan mahasiswa di lingkungan universitas.

# LANDASAN TEORI

Pada Bab ini, silahkan menguraikan semua landasan teori yang terkait dengan kata-kata kunci dari judul proposal yang anda tuliskan di halaman cover depan. Sama halnya dengan latar belakang masalah, pada landasan teori **wajib** disertai sitasi.

Uraian landasan teori bisa didukung dengan tambahan Tabel dan Gambar. Jika ada Tabel atau Gambar yang dimuat untuk mendukung penjelasan narasi, wajib disebutkan pada kalimat uraian serta harus dijelaskan setiap Gambar dan table yang dimasukkan tersebut.

Untuk penulisan sub bab, jangan sampai ada judul sub bab tanpa ada penulisan apapun setelahnya. Selalu mengingat bahwa tidak boleh ada sub sub bab 1 apabila tidak ada sub sub bab 2. Usahakan agar masing-masing sub bab maupun sub sub bab memiliki minimal 2 judul.

## Landasan Teori 1

Uraikan penjelasan dasar teori pertama. Sertakan Gambar dan table sesuai kebutuhan. Jika terdapat tambahan pendukung penjelasan dalam bentuk Tabel, bisa disebutkan Tabel 2.1.

**Tabel 2.1 Isikan Judul Tabel 1**

| **Judul Kolom 1** | **Judul Kolom 2** | **Judul Kolom 3** |
| --- | --- | --- |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

Berikan penjelasan dari Tabel tersebut. Bisa diletakkan setelah table maupun sebelumnya, asal ada nomor dari table yang dijelaskan. Hal serupa dilaksanakan untuk Gambar. Tambahkan Repeat Header Row agar apabila table memiliki panjang melebihi halaman, akan selalu memiliki baris pertama yang biasanya berisi judul kolom dari Tabel.

### Sub Sub Bab 1 Landasan Teori 1

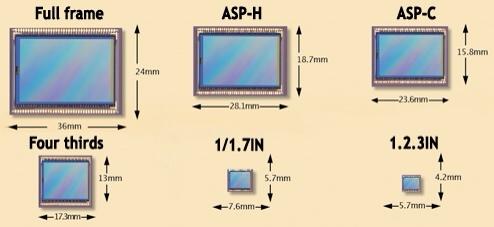
Apabila diperlukan, dapat menambahkan sub bab untuk menjelaskan bagian-bagian dari ladasan teori 1.

### Sub Sub Bab 2 Landasan Teori 1

Apabila diperlukan, dapat menambahkan sub bab untuk menjelaskan bagian-bagian dari ladasan teori 1.

## Landasan Teori 2

Uraikan penjelasan dasar teori kedua. Sertakan Gambar dan table sesuai kebutuhan. Cara penulisan Gambar bisa dilihat pada Gambar 2.1. Berikan pula penjelasan mengenai Gambar tersebut.



**Gambar 2.1 Contoh Gambar Landasan Teori 2**

## Landasan Teori 3

Uraikan penjelasan dasar teori kedua. Sertakan Gambar dan table sesuai kebutuhan. Jika terdapat tambahan pendukung penjelasan dalam bentuk Tabel, bisa disebutkan Tabel 2.1. Penulisan table diletakkan pada bagian atas table bagian tengah. Berikan pula penjelasan dari Tabel tersebut.

**Tabel 2.2 Isikan Judul Tabel 2**

| **Judul Kolom 1** | **Judul Kolom 2** | **Judul Kolom 3** |
| --- | --- | --- |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

Jika narasi pada suatu bab berakhir pada halaman ganjil, maka di halaman genap selanjutnya, tuliskan dengan kalimat: “**Halaman Ini Sengaja Dikosongkan**”. Letakkan pada posisi atas tengah halaman dengan besar tulisan 12 pt, Bold, dan jenis font Times New Roman.

# ANALISIS DAN DESAIN

Bagian ini berisikan segala komponen yang diperlukan dalam menjelaskan penelitian yang akan dilaksanakan. Penjelasan tersebut berupa analisis dan desain sesuai penelitian dan kebutuhan.

## Deskripsi Sistem

Berisi penjelasan mengenai penelitian yang akan dilaksanakan. Pada bagian ini harus terdapat deskripsi dari system yang akan dibangun, yang terdiri atas analisis awal system yang akan dikembangkan, permasalahan yang terjadi beserta rencana solusi pengembangannya.

## Pelaku Sistem

Berisi penjabaran dari pelaku system yang kemungkinan akan terlibat di dalam penggunaan system pada dalam penelitian yang akan dilaksanakan.

## Kebutuhan Data

Kebutuhan data berisi penjelasan mengenai berbagai macam jenis data yang diperlukan dalam penelitian. Kebutuhan data juga mencakup bagaimana proses pengambilan data yang akan digunakan dan bagaimana proses pengolahannya.

## Arsitektur Sistem

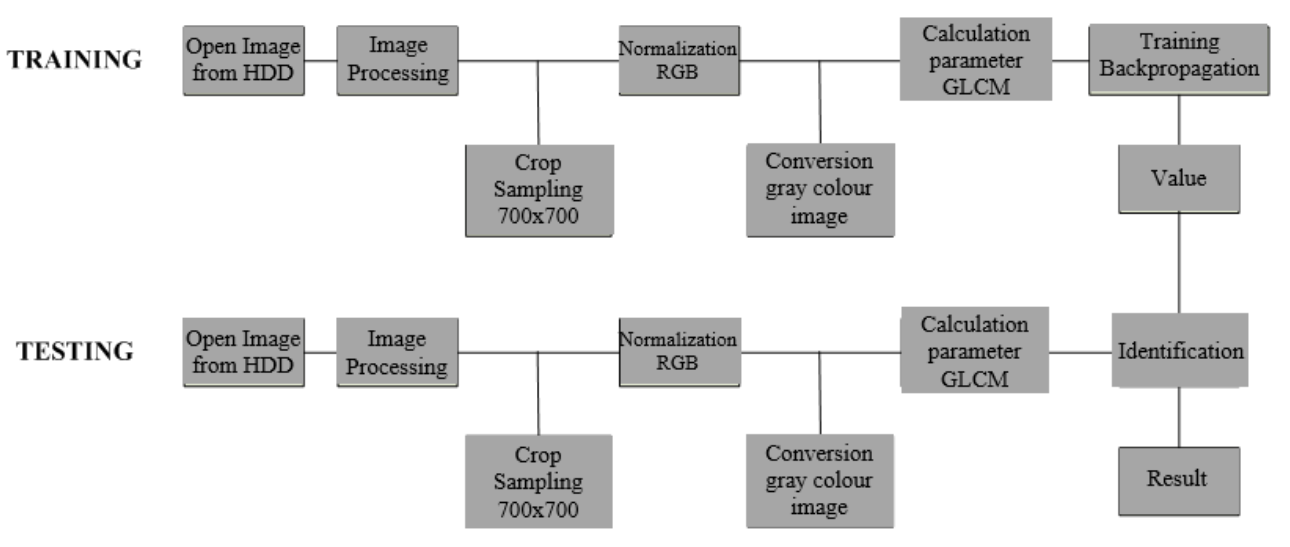
Pada bagian arsitektur sistem digunakan untuk menjelaskan mengenai jenis-jenis komponen yang akan digunakan dan diaplikasikan dalam penelitian. Komponen dapat berupa hardware maupun software yang berhubungan dengan pelaksanaan penelitian. Contoh dari Arsitektur Sistem ditunjukkan oleh Gambar 3.1. Setiap Gambar yang ditampilkan, harus diberi penjelasan.

www.luxoft.com
In-Vehicle Architecture
Computer Vision Box/Head Unit
DSP Camera
HD Camera
Sensors/CAN Navigation System/Ma...

**Gambar 3.1 Contoh Arsitektur Sistem**

## Blok Diagram

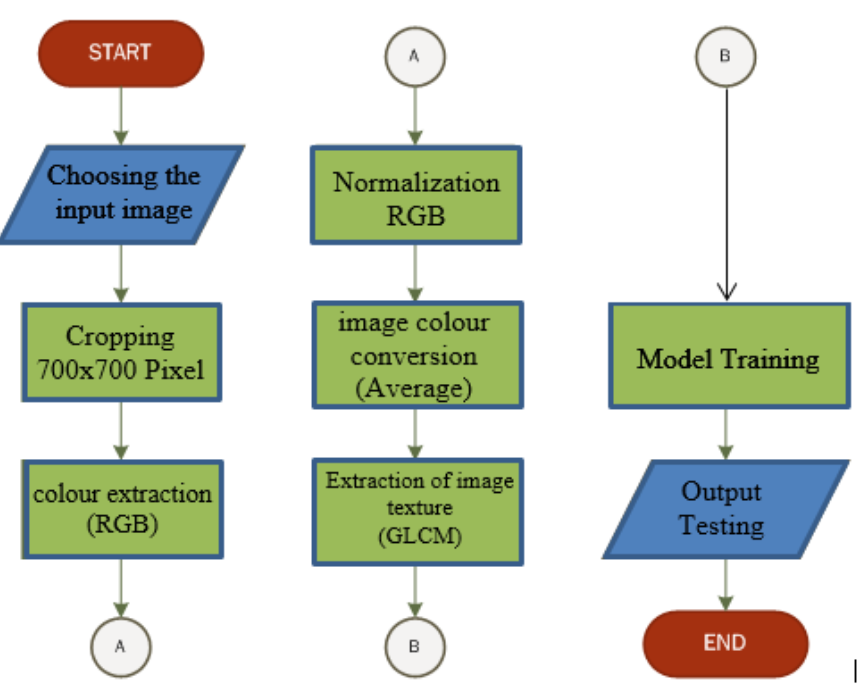
Blok Diagram digunakan untuk menjelaskan mengenai konsep dari penelitian yang akan dibuat. Berikut contoh dari Blok Diagram yang ditunjukkan oleh Gambar 3.1. Setiap Gambar yang ditampilkan, harus diberi penjelasan.



**Gambar 3.2 Blok Diagram Identifikasi Citra**

## Formulasi Kecerdasan Buatan

Flowchart disini digunakan untuk menguraikan bagaimana algoritma penelitian berdasarkan model proses dan formulasi kecerdasan buatan. Contoh dari Flowchart bisa dilihat pada Gambar 3.2. Setiap Gambar yang ditampilkan, harus diberi penjelasan.



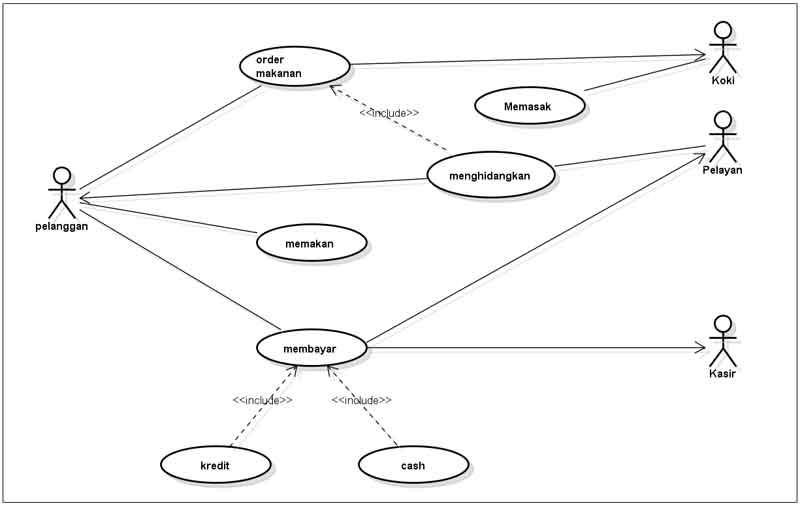
**Gambar 3.3 Flowchart Pemrosesan Citra**

## Model Proses

Pada bagian ini berisikan desain sistem sesuai kebutuhan. Desain yang dimasukkan bergantung kepada penelitian yang dilaksanakan. Desain sistem terbagi atas Data Flow Diagram (DFD) atau Unified Modeling Language (UML) tergantung dari system yang hendak dibuat dan Entity Relationship Diagram (ERD) apabila menggunakan database. Unified Modeling Language (UML) adalah suatu metode permodelan secara visual untuk sarana perancangan sistem berorientasi objek.

### Use Case Diagram

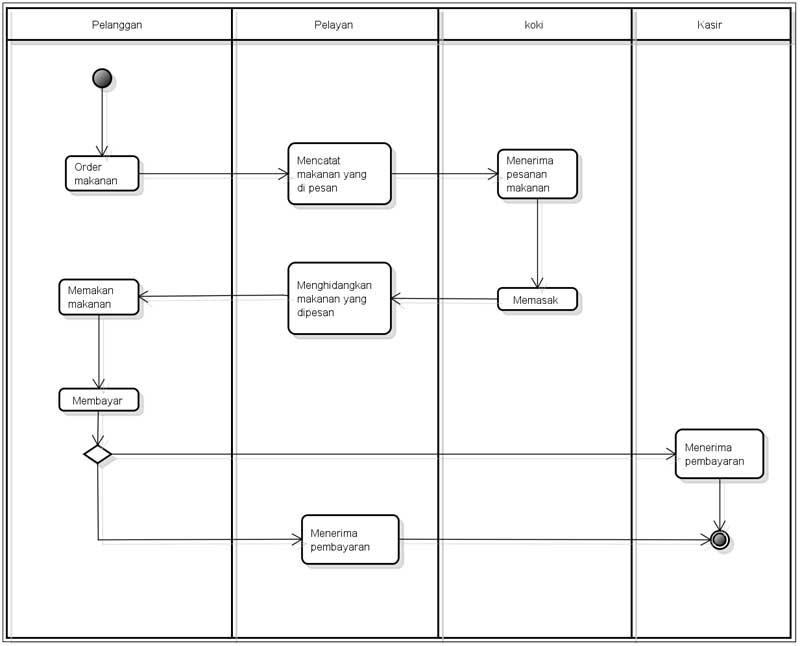
Minimal berisikan use case diagram seperti pada Gambar 3.5. Setiap Gambar yang ditampilkan, harus diberi penjelasan.



**Gambar 3.5 Contoh Use Case Diagram**

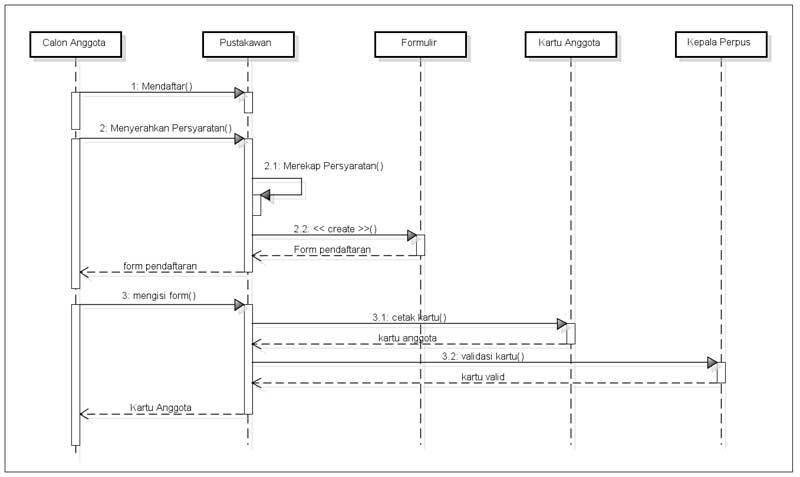
### Activity Diagram

Dapat pula ditambah Activity Diagram, seperti contoh pada Gambar 3.6. Activity diagram atau diagram aktivitas yaitu salah satu jenis diagram pada UML yang dapat memodelkan proses-proses apa saja yang terjadi pada sistem. Setiap Gambar yang ditampilkan, harus diberi penjelasan.



**Gambar 3.6 Contoh Activity Diagram**

Dapat pula ditambah Sequence Diagram, seperti contoh pada Gambar 3.7 apabila memungkinan. Sequence diagram yaitu salah satu jenis diagram pada UML yang menjelaskan interaksi objek yang berdasarkan urutan waktu, sequence diagram juga dapat mengGambarkan urutan atau tahapan yang harus dilakukan untuk dapat menghasilkan sesuatu seperti pada use case diagram. Setiap Gambar yang ditampilkan, harus diberi penjelasan.



**Gambar 3.7 Contoh Sequence Diagram**

### Sequence Diagram

Jika narasi pada suatu bab berakhir pada halaman ganjil, maka di halaman genap selanjutnya, tuliskan dengan kalimat: “**Halaman Ini Sengaja Dikosongkan**”. Letakkan pada posisi atas tengah halaman dengan besar tulisan 12 pt, Bold, dan jenis font Times New Roman.

### Class Diagram

## Desain Proses

## Desain GUI

**DAFTAR PUSTAKA**

[1] Tazkia Aulia Prismadana, “Aplikasi Ruang Tugas Dengan Deteksi Kemiripan Teks Pada Dokumen Tugas Menggunakan Cosine Similarity,” *JURNAL INFORMATIKA & MULTIMEDIA*, vol. 15, no. 1, pp. 31–37, 2023, doi: https://doi.org/10.33795/jtim.v15i1.4405.

[2] Fatimah Zuhra Hasibuan and Juanto Simangunsong, “ANALISA METODE COSINE SIMILARITY DALAM MENDETEKSI PLAGIARISME PADA ARTIKEL ILMIAH,” vol. 3, pp. 1–6, Jun. 2023, [Online]. Available: https://jurnal.universal.ac.id/index.php/teknikinformatika/article/view/9

[3] Syukry Ansis, Endang Palupi Listyaningsih, and S. kom, M. Sc. Prof. Dr. Ir. Hari Soetanto, “Deteksi Plagiat Tesis Berbahasa Indonesia Menggunakan Metode Cosine Similarity,” vol. 9, no. 1, pp. 153–167, 2024, doi: https://doi.org/10.35314/isi.v9i1.4003.

[4] Salsabila Laily Rahma and Umar Taufiq, “Analisis Tingkat Akurasi Metode Pendeteksian Plagiarisme Ide dengan Menggunakan Yake dan Sentence Transformer,” *Journal of Internet and Software Engineering*, vol. 5, no. 1, pp. 15–22, May 2024, doi: https://doi.org/10.22146/jise.v5i1.9073.

[5] Sheraton Pawestri and Yohanes Suyanto, “Analisis Perbandingan Metode Similarity untuk Kemiripan Dokumen Bahasa Indonesia pada Deteksi Kemiripan Teks Bahasa Indonesia,” *JURNAL MEDIA INFORMATIKA BUDIDARMA*, vol. 8, no. 3, p. 1440, Jul. 2024, doi: http://dx.doi.org/10.30865/mib.v8i3.7648\.

[6] GHARDAPATY GHALY GHIFFARY, KEVIN ALIFVIANSYAH, ANWAR FITRIANTO, ERFIANI, and L.M. RISMAN DWI JUMANSYAH, “PERBANDINGAN ALGORITMA HDBSCAN DAN AGGLOMERATIVE HIERARCHICAL CLUSTERING DALAM KLASTERISASI PADA DATA YANG MENGANDUNG PENCILAN,” *Jurnal Riset dan Aplikasi Matematika*, vol. 08, no. 02, pp. 122–135, Oct. 2024, doi: https://doi.org/10.26740/jram.v8n2.p122-135.