

BAB 1

PENDAHULUAN

1.1 Latar Belakang

Indonesia adalah negara dengan keanekaragaman budaya dan suku. Setiap kebudayaan menjadi identitas unik dari masing-masing kelompok etnis di tanah air. Hal ini menggambarkan betapa kayanya Indonesia akan budaya dari masyarakatnya di berbagai wilayah. Keragaman multikultural, baik dari individu maupun kelompok, timbul secara alami melalui interaksi, perbedaan sudut pandang, atau latar belakang, yang pada akhirnya membentuk perilaku budaya dan tradisi di tiap daerah [1].

Dalam kamus besar bahasa Indonesia, budaya (*culture*) diartikan sebagai pikiran, adat istiadat, sesuatu yang sudah berkembang, sesuatu yang sudah menjadi kebiasaan yang sukar diubah. Dalam pemakaian sehari-hari, orang biasanya mensinonimkan pengertian budaya dengan tradisi. Dalam hal ini tradisi diartikan sebagai kebiasaan masyarakat yang tampak [2].

Rokan Hulu adalah salah satu Kabupaten yang ada di Provinsi Riau yang juga memiliki beberapa kebudayaan Melayu. Kabupaten ini memiliki beragam objek budaya lokal seperti rumah adat, pakaian tradisional, alat musik daerah, seni ukir, serta tradisi lisan yang sarat makna historis. Namun, seiring waktu, banyak objek budaya tersebut kurang terdokumentasi secara digital dan belum terintegrasi dalam sistem informasi yang modern, sehingga potensi pemanfaatannya sebagai sarana edukasi dan promosi budaya masih belum optimal.

Menurut hasil wawancara di Bidang Budaya Pada Dinas Pariwisata dan Kebudayaan Rokan Hulu, permasalahan yang di lihat saat ini adalah belum adanya sistem cerdas yang mampu mengenali dan mendeskripsikan objek-objek budaya lokal secara otomatis dari gambar atau citra. Sehingga, informasi mengenai budaya masih banyak diakses secara manual. Ini tentu kurang sesuai dengan kebutuhan masyarakat saat ini yang cenderung mengandalkan kecepatan, visualisasi, dan interaktivitas dalam mengakses informasi. Di sisi lain, pemanfaatan teknologi kecerdasan buatan dalam bidang kebudayaan juga masih sangat minim. Padahal, teknologi seperti ini bisa menjadi salah satu solusi untuk mendukung transformasi digital, khususnya dalam upaya pelestarian dan promosi budaya lokal agar lebih dikenal luas, terutama oleh generasi muda.

Sebagai solusi, diperlukan pendekatan teknologi yang mampu menggabungkan kemampuan pengolahan citra (gambar) dan pemahaman bahasa alami (narasi) dalam satu sistem salah satunya dengan menerapkan *Transformer Multimodal* . Salah satunya dengan menggunakan teknik *BLIP* (Bootstrapped Language–Image Pretraining) yang mampu melakukan pengenalan gambar sekaligus menghasilkan deskripsi naratif yang relevan secara otomatis. Dengan memanfaatkan *BLIP*, sistem dapat mendeteksi objek budaya dari gambar dan secara langsung menghasilkan narasi penjelasan berbasis konteks budaya lokal. Penerapan teknologi ini tidak hanya meningkatkan efisiensi dalam dokumentasi budaya, tetapi juga dapat dimanfaatkan sebagai media edukasi digital yang interaktif dan informatif, serta sebagai sarana pelestarian budaya lokal berbasis teknologi kecerdasan buatan.

Beberapa peneliti telah melakukan penelitian tentang *Transformer Multimodal*. Sebelum penelitian dilakukan, penulis telah membaca referensi terhadap beberapa penelitian sebelumnya. Seperti yang dilakukan penelitian oleh Yuchun Lu, dkk tahun 2024 dengan judul “*Application of Multimodal Transformer Model in Intelligent Agricultural Disease Detection and Question-Answering System*”. Dari hasil penelitian dapat disimpulkan dalam penelitian ini, pendekatan komprehensif berdasarkan data multimoda dan model *Transformer* diusulkan untuk mengatasi tantangan utama dalam sistem deteksi penyakit pertanian dan tanya jawab. Pertama, dalam eksperimen deteksi penyakit, berbagai model termasuk *AlexNet*, *GoogLeNet*, *VGG*, *ResNet*, dan metode yang diusulkan dalam makalah ini dibandingkan. Hasilnya menunjukkan bahwa metode yang diusulkan mencapai nilai tertinggi dalam presisi, perolehan kembali, dan akurasi, dengan skor masing-masing 0,95, 0,92, dan 0,94, yang secara signifikan mengungguli model perbandingan lainnya. Hal ini menunjukkan bahwa metode yang diusulkan memiliki keuntungan signifikan dalam mengidentifikasi berbagai penyakit pertanian, terutama dalam memproses data kompleks dan fitur halus. Kedua, dalam eksperimen penerapan teks gambar pertanian, kinerja *BLIP*, *mPLUG-Owl*, *Instruct BLIP*, *CLIP*, *BLIP 2*, dan metode yang diusulkan dalam makalah ini diperiksa [3].

Penelitian lainnya yang telah diteliti oleh Stragel Alandia Sinaga, dkk tahun 2024 dengan judul implementasi *Multimodal* convolutional neural network dan extreme gradient boosting untuk grading buah jambu kristal skala industri. Dari penelitian dapat disimpulkan Arsitektur dan konfigurasi pelatihan model

Multimodal CNN dan *XGBoost* untuk grading buah jambu kristal: Penelitian ini berhasil menentukan kombinasi hyperparameter terbaik untuk masing-masing model yang digunakan. Pada model *Convolutional Neural Network (CNN)*, konfigurasi terbaik yang ditemukan meliputi optimizer Adam, filter size 3×3 , pool size 2×2 , learning rate 0.001, serta konfigurasi layer yang terdiri dari 32, 64, dan 128 filter dengan 128 neuron aktivasi. Sementara itu, *model Extreme Gradient Boosting (XGBoost)* mencapai performa terbaik dengan menggunakan *booster gbtrees*, *learning rate 0.1*, *max depth 3*, dan *min child width 5* [4].

Untuk menjawab permasalahan tersebut, diperlukan sebuah solusi berbasis teknologi cerdas yang mampu mengintegrasikan informasi visual dan teks secara otomatis. Salah satu pendekatan yang dapat digunakan adalah *Transformer Multimodal*, khususnya model *BLIP* (Bootstrapped Language Image Pretraining), yang mampu melakukan pengenalan gambar sekaligus menghasilkan deskripsi naratif yang relevan secara otomatis. Dengan memanfaatkan *BLIP*, sistem dapat mendeteksi objek budaya dari gambar dan secara langsung menghasilkan narasi penjelasan berbasis konteks budaya lokal. Penerapan teknologi ini tidak hanya meningkatkan efisiensi dalam dokumentasi budaya, tetapi juga dapat dimanfaatkan sebagai media edukasi digital yang interaktif dan informatif, serta sebagai sarana pelestarian budaya lokal berbasis teknologi kecerdasan buatan.

1.2 Rumusan Masalah

Berdasarkan latar belakang yang telah dijelaskan, terdapat beberapa permasalahan utama yang menjadi fokus dalam penelitian ini, yaitu : Bagaimana

merancang dan membangun sistem yang dapat mengenali suatu sistem objek budaya lokal Rokan Hulu berbasis gambar secara otomatis?

1.3 Tujuan Penelitian

Adapun tujuan penelitian ini adalah Menerapkan model *Transformer Multimodal* , khususnya *BLIP (Bootstrapped Language Image Pretraining)*, untuk menghasilkan narasi deskriptif otomatis berdasarkan objek budaya yang dikenali.

1.4 Batasan Masalah

Adapun batasan dalam penelitian ini guna menghindari kesalahan persepsi dan ruang lingkup yang terlalu luas, maka batasan masalah ditentukan sebagai berikut :

1. Objek budaya yang dikenali terbatas pada objek visual seperti bangunan bersejarah dan alat musik khas Rokan Hulu yang memiliki dokumentasi gambar. Disini saya menggunakan 6 objek, sebagai berikut : A. Bangunan Prasejarah (Istana Rokan, Benteng Tujuh Lapis, Makam Raja-Raja Rambah), B. Alat Musik (Colempong, Gong (ogong), Gondang) Dari beberapa objek tersebut saya mengambil 50 gambar pada setiap objek untuk dilatih pada model, berarti ada 300 gambar yang akan saya latih untuk pengenalan objek ini.
2. Model yang digunakan untuk pengenalan gambar dan deskripsi otomatis adalah model *BLIP (Bootstrapped Language Image Pretraining)*.
3. Sistem yang dibangun berbasis *Website* agar dapat diakses secara online menggunakan perangkat desktop maupun mobile, namun belum mencakup integrasi dengan media sosial.

1.5 Manfaat Penelitian

Penelitian ini diharapkan dapat memberikan manfaat sebagai berikut :

1. Menerapkan model *Transformer Multimodal* seperti *BLIP* untuk pengolahan data visual dan teks secara terpadu.
2. Menyediakan sebuah prototipe sistem cerdas yang mampu mengenali dan mendeskripsikan objek budaya lokal secara otomatis, sehingga dapat digunakan sebagai media edukasi dan pelestarian budaya Rokan Hulu.
3. Membantu pemerintah daerah, lembaga kebudayaan, maupun pelaku pendidikan dalam mempromosikan budaya lokal secara digital dan interaktif kepada generasi muda maupun masyarakat umum.

1.6 Sistem Penulisan

Sistematika penulisan dari SKRIPSI ini terdiri dari pokok – pokok permasalahan yang dibahas pada masing – masing yang diuraikan menjadi beberapa bagian:

BAB 1 PENDAHULUAN

Bab ini berisi latar belakang, rumusan masalah, tujuan penelitian batasan masalah, manfaat penelitian dan sistematika penulisan

BAB 2 LANDASAN TEORI

Bab ini berisi teori-teori yang digunakan pada penelitian ini, teori-teori yang berhubungan dengan kecerdasan buatan, *Transformer Multimodal* , *BLIP*.

BAB 3 METODOLOGI PENELITIAN

Bab ini berisi tahapan-tahapan dalam pengumpulan data, perancangan sistem perumusan masalah dan analisa

BAB 4 ANALISA DAN PERANCANGAN

Bab ini berisi analisa dan perancangan aplikasi pengenalan dan deskripsi objek budaya lokal rokan hulu berbasis gambar dan narasi otomatis.

BAB 5 IMPLEMENTASI DAN PENGUJIAN

Bab ini berisi implementasi dari analisa dan perancangan dan pengujian pada aplikasi yang berhasil dibangun.

BAB 6 PENUTUP

Bab ini berisi rangkuman dari hasil penelitian yang telah dilakukan dan saran-saran untuk pengembangan aplikasi atau penelitian selanjutnya.

BAB 2

LANDASAN TEORI

2.1 *Deep Learning*

Deep Learning adalah bagian dari pembelajaran mesin yang berkaitan dengan algoritma dimana cara kerja dari algoritma ini meniru struktur dan fungsi otak yang disebut jaringan saraf tiruan. Penelitian terkait *Deep Learning* telah banyak dilakukan dalam berbagai bidang. Seperti misalnya di bidang sosial, ekonomi dan bahkan hingga masalah emosi, musik juga social media. Beberapa penelitian terkait bidang kesehatan adalah deteksi kanker payudara, deteksi covid tumor otak, klasifikasi image medis dan sebagainya [5].

Deep Learning atau sering dikenal dengan istilah *deep structured learning* atau *hierarchical learning* adalah salah satu cabang dari ilmu pembelajaran mesin machine learning yang terdiri algoritma pemodelan abstraksi tingkat tinggi pada data menggunakan sekumpulan fungsi transformasi nonlinear yang ditata secara berlapis - lapis dan mendalam [6].

Deep Learning adalah metode khusus pembelajaran mesin yang menggabungkan jaringan saraf di lapisan berturut-turut untuk belajar dari data secara iterative dengan meniru cara kerja otak manusia sehingga komputer dapat dilatih untuk menangani abstraksi dan masalah yang didefinisikan dengan buruk, Keberadaan *Deep Learning* untuk mengatasi kekurangan dari metode machine learning konvensional biasa. Salah satunya adalah kemampuan feature engineering yang dapat merekayasa fitur secara otomatis. Teknologi *Deep Learning* mampu

memberikan hasil yang lebih baik dan sebanding dengan jumlah data yang di tambahkan [7].

2.2 Kecerdasan Buatan (*AI*)

Kecerdasan buatan (*AI*) adalah bidang ilmu komputer yang menggunakan simbol daripada angka untuk menyampaikan informasi dan menggunakan metode heuristik atau sejumlah aturan untuk memproses data [8]. Salah satu bidang sains dan teknik yang paling baru adalah kecerdasan buatan, atau *AI* penelitian tentang kecerdasan buatan dimulai secara resmi setelah Perang Dunia II, dan nama *AI* pertama kali digunakan pada tahun 1956 [9].

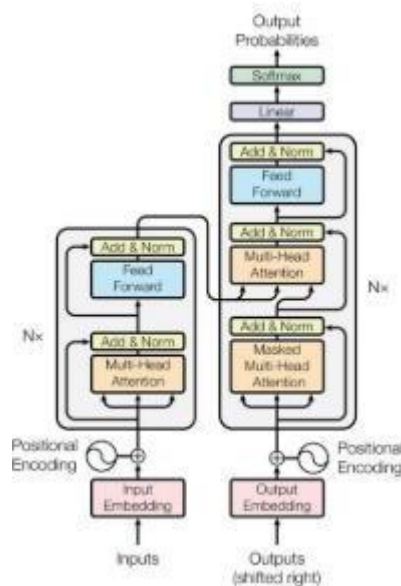
Semakin banyak pengembangan hardware dan *software* yang menggunakan teknik kecerdasan buatan menunjukkan betapa berkembangnya kecerdasan buatan saat ini di bidang ilmu komputer [10]. Kemampuan mesin untuk meniru kecerdasan manusia dalam hal berpikir, belajar, dan membuat keputusan disebut kecerdasan buatan (*AI*). Ini mencakup penggunaan teknik komputasi yang kompleks dan algoritma untuk mengolah data, mengenali pola, dan membuat prediksi atau tindakan cerdas [11].

Kecerdasan buatan (*Artificial Intelligence/AI*) mengacu pada kemampuan mesin untuk meniru kecerdasan manusia dalam mempelajari, berpikir, dan mengambil keputusan, ini melibatkan penggunaan algoritma dan teknik komputasi yang kompleks untuk mengolah data, mengenali pola, dan membuat prediksi atau tindakan yang cerdas, contoh penggunaannya adalah *machine learning*, *neural networks*, *natural language processing*, *computer vision*, dan robotika, kecerdasan

buatan ini bertujuan untuk meningkatkan efisiensi dan kualitas hidup melalui teknologi cerdas [12].

2.3 Model *Transformer*

Model *Transformer* s berisi lapisan *encoder* dan *decoder*, dimana masing-masing terhubung ke lapisan multi-head dan memberi masukan pada layer-layer yang terhubung secara *fully connected*. Model akan mengingat posisi dan urutan kata dengan bantuan fungsi cosine dan membuat pengkodean posisi tersebut. Di layer *encoder* dan *decoder* multi-head menerapkan mekanisme yang disebut *self-attention* yang menjadi masukan dan dibagi menjadi tiga lapisan yang terhubung untuk membuat vektor *Query (Q)*, *Key (K)* dan *Value (V)* [13].



Gambar 2. 1 Model dan Arsitektur *Transformer*

Pada Model *Transformer* terdapat dua pendekatan yaitu *pre-training* dan *fine-tuning*. Dalam *pre-training* dilakukan dengan melatih model dengan data yang tidak berlabel sedangkan *fine-tuning* dilakukan dengan inialisasi dari parameter dari pretrained dan *fine-tuning* dari data yang telah berlabel [14].

Transformer adalah model *Deep Learning* dengan arsitektur *encoder - decoder*. *Encoder* menerima token berita sebagai masukan. Pada saat training, *decoder* menerima token label (referensi ringkasan) yang disisipkan token “” di posisi paling depan, deretan token digeserkan ke sebelah kanan (*shifted right*) agar jumlah token yang masuk ke dalam *decoder* sama dengan jumlah token label. Sedangkan, pada saat *inference*, *decoder* menghasilkan ringkasan dengan melakukan iterasi untuk mengeluarkan prediksi token satu per satu sampai *decoder* menemukan token “” atau sampai dengan batas iterasi yang ditentukan [15].

2.4 Multimodal Learning

Multimodal sangat sering dikaitkan dengan konsep dari multiliterasi. Terdapat empat komponen penting dalam konsep pedagogi multiliterasi yang dikembangkan oleh New London Group (1996) yaitu, Situation Practice, Overt Intructions, Critical Framing, dan Tranformation Practice. Pendekatan pedagogi multiliterasi tersebut dihubungkan dengan tiga elemen desain (Available design, designing, dan redesigned) untuk membangun pola makna dari lingkungan *Multimodal* yang tersedia, *Multimodal* terjadi melalui banyaknya mode yang terlibat (berbagai macam indera), antara satu mode dengan mode lainnya, mereka saling berinteraksi dengan apa yang mereka hasilkan [16].

Multimodal itas adalah penggunaan banyak modalitas komunikasi dalam upaya kolaboratif untuk menciptakan makna, agar pembaca atau pendengar dapat memahami dan bereaksi terhadap keseluruhan makna yang terkandung, beberapa komponen penting perlu diperhitungkan. Istilah "mode pertama" menggambarkan

jenis komunikasi atau media yang digunakan, termasuk teks tertulis, grafik, suara, dan video. Kedua, media yang digunakan untuk berkomunikasi, termasuk buku, televisi, internet, dan bentuk media lainnya. Ketiga, desain adalah susunan komponen seperti model atau bentuk, tata letak, warna, dan ukuran untuk menghasilkan makna. Keempat, konten pesan yang ingin Anda sampaikan dengan menggunakan gaya dan desain tertentu. Kelima keadaan sejarah, sosial, dan budaya ini mempengaruhi bagaimana orang menafsirkan pesan [17].

Multimodal adalah istilah yang digunakan untuk merujuk pada cara manusia berkomunikasi dengan menggunakan berbagai mode sekaligus. Ini melibatkan penggunaan beberapa mode semiotik (seperti bahasa, gambar, suara, gerakan, dll.) secara bersamaan dalam desain produk atau peristiwa semiotik. Dalam konteks analisis teks, *Multimodal* itas berarti menganalisis teks yang menggabungkan mode verbal dan mode visual (gambar) [18].

2.5 Website

Website adalah salah satu bentuk komunikasi melalui media massa dengan jaringan internet yang dapat memberikan informasi tertentu serta dapat diakses oleh orang banyak. *Website* berasal dari kata *World Wide Web*, yakni layanan yang didapati oleh pemakai komputer yang terhubung dengan jaringan internet. *Website* merupakan aplikasi tertentu yang berjalan di atas *platform* atau *operation system browser*[19]

2.6 Budaya Rokan Hulu

Kabupaten Rokan Hulu, yang terletak di Provinsi Riau, merupakan daerah yang kaya akan nilai-nilai budaya Melayu dan warisan sejarah yang masih lestari.

Budaya Rokan Hulu mencerminkan perpaduan antara adat istiadat lokal, pengaruh kerajaan lama, serta kearifan leluhur yang diwariskan dari generasi ke generasi.

Berikut aspek-aspek penting dari budaya Rokan Hulu [20] :

1. Bangunan Prasejarah

a) Istana Rokan

Istana kerajaan Rokan atau Istana Raja Rokan terletak di Kecamatan Rokan IV Koto kurang lebih 48 Kilometer dari pusat pemerintahan Kabupaten Rokan Hulu. Untuk menuju kesana kita bisa menggunakan roda empat maupun roda dua. Istana Kerajaan Rokan juga dikenal dengan rumah tinggi, berusia sekitar 200 tahun dan dibangun pada masa pemerintahan Kesultanan Nagari Tuo. Dibangun dengan gaya arsitektur khas melayu rokan dengan ukiran naga, serta berbagai motif tumbuhan. Bentuk bangunan Istana Rokan merupakan rumah panggung yang terlihat pada keberadaan kandang, tiang-tiang yang tinggi dan juga sandi yang berada di bawah tiang. Bentuk bangunan Istana Rokan mirip dengan bentuk rumah adat minangkabau, namun lengkungan gonjong sangat rendah bila dibandingkan dengan gonjong Rumah Gadang. Disekitar istana juga terdapat perkampungan penduduk dengan suku yang terdiri dari Suku Melayu, Suku Bendang, Suku Patapang, Suku Maeh, Suku Mandailing Dan suku Caniago[21].



Gambar 2. 2 Istana Rokan IV Koto

(Dokumentasi Peneliti, 27 Juli 2025)

b) Benteng Tujuh Lapis

Benteng Tujuh Lapis ini merupakan benteng pertahanan yang terbuat dari gundukan-gundukan tanah yang memiliki ketinggian pada masa dahulu sekitar 3 meter. Di antara lapis gundukan tanah ini dibuat parit yang berguna untuk menghalangi musuh masuk. Benteng ini terletak diantara aliran sungai dan lereng bukit. Benteng ini merupakan benteng artifisial yang dibuat masyarakat untuk kubu pertahanan dengan luas 105.000 meter dengan ukuran panjang 350 meter dan lebar 300 meter. Benteng ini didirikan pada tahun 1815 [21].



Gambar 2. 3 Benteng Tujuh Lapis

(Sumber : Peneliti, 10 Agustus 2025)

c) Makam Raja-Raja Rambah

Pemakaman ini merupakan kompleks pemakaman raja-raja Rambah yang kedua. Lokasi pertama berada di Kampung Rambahan Tanjung Beling. Mulai pindah diperkirakan awal tahun 1800-an. Kompleks pemakaman ini dahulunya berada dalam kompleks istana Kerajaan Rambah yang berada di pinggir sungai Rokan Kanan dengan jarak sekitar 250 meter dari jalan raya Pasir Pengaraian – Dalu-Dalu dengan jalan menuju lokasi pemkaman sudah ditembok. Luas dari komplek pemakaman Raja-Raja Rambah ini sekitar 600m dengan panjang 30 meter dan lebar 20 meter. Luas keseluruhan dari lokasi ini menurut data yang diperoleh seluas 4ha. Dalam areal 600m² tersebut terdapat 27 makambesar dan kecil. Menurut data makam yang kecil merupakan anak-anak dari keluarga kerajaan[21].



Gambar 2. 4 Makam Raja-Raja Rambah

(Sumber : Peneliti, 26 Agustus 2025)

2. Alat Musik

a) Calempong

Sorawan (satu set) atau seperangkat alat musik berbentuk seperti gong berukuran kecil, berjumlah enam buah, berdasarkan enam urutan nada yang khas. Diletakkan pada sebuah tempat dinamakan rurumah còlempong (rumah-rumah còlempong).Còlempong dimainkan dengan iringan dua gendang disebut dengan istilah bógondang bocòlempong (bergendang bocòlempong) atau sering disingkat dengan sebutan gondang còlempong. Còlempong bisa dimainkan tanpa iringan gendang atau gong. Struktur fisik còlempong terdiri dari pusek, bahu, badan, muluik dan bibie (bibir). Terbuat dari bahan tembaga dicampur dengan logam lain semisal perak perunggu, suasa, emas, timah dll. Kadar timah putih dan besi dalam campuran tembaga berpengaruh pada lengkingan dan dentangan suara còlempong[22].



Gambar 2. 5 Calempong

(Sumber : Peneliti, 28 Agustus 2025)

b) Gong (ogong)

Bagi orang Melayu Sungai Rokan, ogong termasuk barang yang dihormati karena fungsinya untuk upacara. Disimpan dengan baik di tempat khusus, diasuh dengan upacara memandikan ogong menggunakan air limau, diasap dengan kemenyan, dan disimpan dalam balutan kain kuning atau kain putih. Ada beberapa sebutan penamaan atas ogong ini, yakni Selembang Karang; Selimang Karang; atau Selumang Karang. Ogong Selembang Karang digelar Ogong sakti, yakni; gong yang mempunyai keistimewaan[22].



Gambar 2. 6 Gong (ogong)

(Dokumentasi Peneliti, 27 Juli 2025)

c) Gondong

Gondang terdiri dari dua, pertama gondang polalu (melodi) dan kedua gondang poningkah (Bass). Kedua bentuk gondang ini pada dasarnya sama, baik ukuran maupun bentuk, hanya setelan ketegang kulit, dan cara pemukulan yang berbeda, sehingga menghasilkan

nada melodi dan bass. Fungsinya dalam musik tradisional Gondang Borogong yaitu sebagai pembawa ritme dan juga tempo.



Gambar 2. 7 Gondong

(Dokumentasi: Peneliti 28 Agustus 2025)

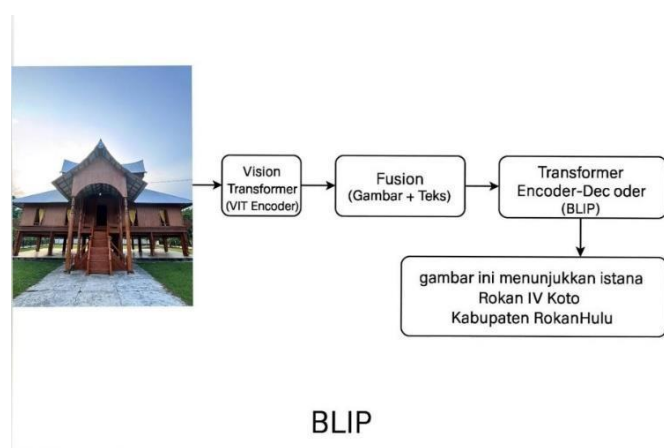
2.7 BLIP (Bootstrapped Language-Image Pretraining)

BLIP atau *Bootstrapped Language-Image Pretraining* merupakan sebuah model *Transformer Multimodal* yang dikembangkan untuk memahami serta menghubungkan informasi visual (gambar) dan tekstual (bahasa). *BLIP* dirancang untuk mendukung berbagai tugas vision-language seperti *image captioning*, *image- text retrieval*, dan *visual question answering*. Salah satu fitur kunci dari *BLIP* adalah pendekatan *bootstrapping* yang digunakan untuk melatih model dengan data gambar dan teks dari web secara efisien dan akurat. Hal ini memungkinkan model belajar dari data yang bersifat noisy tanpa memerlukan kurasi manual secara ekstensif [23].

BLIP memberikan keunggulan kompetitif dibandingkan model-model sebelumnya dalam bidang *Multimodal learning* karena kemampuannya menangani data skala besar dengan efisiensi tinggi. Model ini telah menunjukkan performa

unggul dalam benchmark populer seperti *COCO* dan *Flickr30k*. Aplikasi *BLIP* meluas pada bidang pendidikan budaya, sistem deskripsi otomatis objek, arsip museum digital, hingga pengembangan sistem interaktif berbasis gambar dan teks. Dalam konteks pengenalan objek budaya lokal, *BLIP* dapat dimanfaatkan untuk mendeskripsikan warisan budaya secara otomatis dan memperluas akses masyarakat terhadap informasi berbasis visual dan narasi [23].

Arsitektur *BLIP* menggabungkan *Vision Transformer (ViT)* sebagai *encoder* visual dan *Transformer* language model untuk pemrosesan teks. *BLIP* dapat dioperasikan dalam dua mode: *encoder* -only untuk memahami gambar dan teks secara bersamaan, dan *encoder* -*decoder* untuk menghasilkan teks dari gambar (*captioning*). Salah satu komponen penting dalam proses pretraining *BLIP* adalah *Captioning and Filtering (CapFilt)* yaitu teknik untuk menghasilkan caption dari data gambar secara otomatis dan kemudian menyaring hasil yang berkualitas tinggi. Ini memungkinkan *BLIP* menyempurnakan data latih yang berasal dari web secara efisien [24].



Gambar 2. 8 Contoh Alur *BLIP*

2.8 *Python*

Python diciptakan oleh Guido van Rossum di Belanda pada tahun 1990 dan namanya diambil dari acara televisi kesukaan Guido Monty *Python* 's Flying Circus. Van Rossum mengembangkan *Python* sebagai hobi, kemudian *Python* menjadi bahasa pemrograman yang dipakai secara luas dalam industri dan pendidikan karena sederhana, ringkas, sintak intuitif dan memiliki putaka yang luas, *Python* adalah bahasa pemrograman yang menggunakan interpreter untuk menjalankan kode programnya. Interpreter tersebut dapat menerjemahkan kode secara langsung, dan *Python* dapat dijalankan di berbagai platform seperti Windows, Linux, dan lain- lain [25].

Python adalah sebuah bahasa pemrograman dinamis yang sering digunakan dalam pengembangan aplikasi pada berbagai domain. Hal ini memungkinkan suatu program ditulis dalam beberapa pendekatan sekaligus. Misalnya, antar muka grafis dibuat dalam bentuk orientasi objek, sedangkan pemrosesan dalam bentuk fungsional atau prosedural [26].

Python adalah bahasa pemrograman tingkat tinggi yang sangat populer, dikenal karena sintaksnya yang sederhana dan mudah dipahami. *Python* telah berkembang menjadi salah satu bahasa pemrograman yang paling banyak digunakan di dunia, terutama dalam bidang data science, pengembangan web, dan otomatisasi [27].

2.9 *Google Collab*

Google Collab adalah sebuah IDE untuk pemrograman *Python* dimana pemrosesan akan dilakukan oleh server Google yang memiliki perangkat keras

dengan performa yang tinggi. Dari sisi perangkat lunak, *Google Collab* telah menyediakan hampir sebagian besar pustaka (library) yang dibutuhkan [28].

Google Collab adalah layanan cloud computing yang disediakan oleh Google untuk mendukung pengembangan dan penelitian ilmiah. Collaboratory, atau “Collab” adalah produk dari Google Research. Collab memungkinkan untuk menulis dan mengeksekusi kode *Python* arbitrer melalui browser, dan sangat cocok untuk machine learning, analisis data, dan pendidikan [29].

Google Collab terkadang dikenal sebagai *Google Collaboratory* adalah layanan pembelajaran mesin gratis yang ditawarkan oleh Google. Platform ini memungkinkan pengguna untuk menelusuri dan menjalankan kode *Python* , mengunggah dan menyimpan kode bersama data, dan memungkinkan kolaborasi waktu nyata dengan pengguna lain. Selain itu, *Google Collab* menyediakan akses ke mesin virtual dengan *CPU*, *GPU*, dan *TPU* (Unit Pemrosesan Tensor) yang dapat digunakan untuk menjalankan kode *Python* [31].

2.10 UML (Unified Modeling Language)

UML (Unified Modelling Language) adalah bahasa visual untuk pemodelan dan komunikasi sebuah sistem dengan menggunakan diagram dan teks – teks pendukung [32]. *UML (Unified Modelling Language)* adalah salah satu standar bahasa yang banyak digunakan di dunia industry untuk mendefinisikan *Requirement*, membuat analissi dan design, serta menggambarkan arsitektur dalam pemrograman berorientasi objek. Berdasarkan sifatnya yang statis maupun dinamis, berikut beberapa jenis diagram *UML* yang akan digunakan oleh penulis adalah sebagai berikut [33] :

1. *Use Case Diagram*

Use Case Diagram merupakan pemodelan untuk kelakuan sistem informasi yang akan dibuat. *Use Case* digunakan untuk mengetahui fungsi apa saja yang ada di dalam sistem dan siapa saja yang berhak untuk menggunakan fungsi-fungsi tersebut.

2. *Activity Diagram*

Activity Diagram menggambarkan aliran kerja atau aktivitas dari sebuah sistem.

3. *Sequence Diagram*

Sequence diagram menggambarkan kelakuan objek pada *Use Case* dengan mendeskripsikan waktu hidup objek dan pesan yang dikirimkan dan diterima antar objek.

4. *Class Diagram*

Class Diagram menggambarkan sebuah spesifikasi yang menghasilkan objek dan merupakan inti dari pengembangan dan orientasi objek

2.11 Penelitian Terdahulu

Untuk tabel penelitian dapat dilihat pada tabel berikut :

Tabel 2. 1 Penelitian Terdahulu

No	Nama Peneliti (Tahun)	Judul Penelitian	Metode	Kesimpulan

1	Wang, C., Shen, Y., Ji, L. (2021)	<i>Geometry Attention Transformer with Position-aware LSTMs for Image Captioning</i>	<i>Transformer (GAT) + LSTM</i>	Integrasi informasi geometris meningkatkan kualitas caption pada dataset COCO dan Flickr30K
2	Dubey, S. et al. (2021)	<i>Label-Attention Transformer with Geometrically Coherent Objects for Image Captioning</i>	<i>Transformer (LATGeO)</i>	Menggabungkan label dan posisi objek memperkaya caption secara semantik
3	Fan, Z. et al. (2021)	<i>TCIC: The Image Concepts Learning Cross Language and Vision for Image Captioning</i>	<i>Transformer (TTN)</i>	Representasi tema lintas bahasa dan visi memperbaiki pemahaman semantik tingkat tinggi

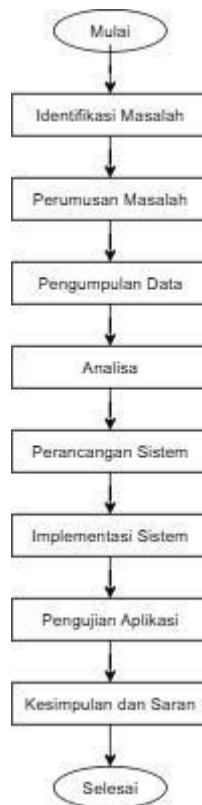
4	Liu, W. et al. (2021)	<i>CPTR: Full Transformer Network for Image Captioning</i>	<i>Fully Transformer (CPTR)</i>	<i>Model Transformer lengkap (tanpa CNN) unggul dibanding metode CNN+Transformer</i>
5	Hu, W., Qiao, L., Kang, W., Shi, X. (2023)	<i>Thangka Image Captioning Based on Semantic Concept Prompt and Multimodal Feature Optimization</i>	<i>Prompt Multimodal + Transformer</i>	Pendekatan prompt-konsep mampu mendeskripsikan Thangka dengan akurasi lebih tinggi.
6	Stefanini, M. et al. (2023)	<i>DCAT: Dynamic Context-Aware Image Captioning in Tamil Language</i>	<i>Transformer kontekstual</i>	Model dinamis menyesuaikan konteks bahasa lokal secara efektif
7	Caron, G., Bellon, O. R.	<i>Computer Vision and</i>	<i>Deep Learning +</i>	Teknologi visi-komputer dan

		<i>Robotics</i> <i>for</i>		
--	--	-------------------------------	--	--

BAB 3

METODOLOGI PENELITIAN

Tahapan penelitian yang akan dilakukan dan penyelesaian masalah terhadap penerapan *Transformer Multimodal* untuk pengenalan dan deskripsi objek budaya lokal Rokan Hulu berbasis gambar dan narasi otomatis. Adapun tahapan metodologi yang dilakukan selama penelitian dapat dilihat pada Gambar 3.1 yang mana merupakan proses yang dimulai dari studi literatur hingga diperoleh Kesimpulan.



Gambar 3. 1 Tahapan Metodologi Penelitian

3.1 Identifikasi Masalah

Dalam pelestarian budaya lokal, khususnya objek budaya yang berasal dari daerah Rokan Hulu, masih ditemukan berbagai tantangan dalam hal dokumentasi

dan pengenalan yang efektif. Berdasarkan latar belakang yang telah dijabarkan, maka dapat diidentifikasi beberapa permasalahan sebagai berikut :

1. Kurangnya media digital yang mampu mengenali dan mendeskripsikan objek budaya lokal secara otomatis, sehingga proses dokumentasi masih bersifat manual dan memerlukan keterlibatan tenaga ahli budaya secara langsung.
2. Informasi budaya cenderung tidak terdokumentasi dalam bentuk narasi yang menarik dan mudah dipahami, khususnya oleh generasi muda atau masyarakat umum yang lebih familier dengan media digital dan otomatisasi.
3. Belum tersedia sistem interaktif yang dapat digunakan masyarakat untuk mengakses informasi budaya secara visual dan deskriptif secara langsung, misalnya dengan mengunggah gambar lalu menerima narasi budaya yang relevan.

3.2 Perumusan Masalah

Berdasarkan hasil dari tahapan pengamatan pendahuluan sebelumnya, maka tahapan selanjutnya adalah tahapan perumusan masalah, pada tahapan perumusan masalah akan dirumuskan masalah yang dianggap sebagai penelitian dalam tugas akhir ini, permasalahan-permasalahan yang menjadi rumusan masalah dalam penelitian ini didapatkan dari penelitian dari penelitian terkait data pengamatan pendahuluan sebelumnya solusi yang didapatkan pada tahapan perumusan masalah ini yang akan menjadi judul penelitian tugas akhir ini “Penerapan

Transformer Multimodal Untuk Pengenalan dan Deskripsi Objek Budaya Lokal Rokan Hulu Berbasis Gambar Dan Narasi Otomatis”. Pengumpulan Data

Pengumpulan data adalah tahapan-tahapan yang bertujuan dalam memperoleh data-data informasi yang berhubungan dengan penelitian tugas akhir ini, pada tahapan pengumpulan data ini juga berguna untuk mengumpulkan semua kebutuhan data yang akan diproses nantinya menggunakan *Transformer Multimodal BLIP*. Pengumpulan data tersebut dilakukan dengan beberapa cara, antara lain:

1. Observasi

Observasi dilakukan secara langsung terhadap objek budaya lokal Rokan Hulu yang menjadi fokus pengenalan, seperti rumah adat, pakaian tradisional, senjata, atau ornamen khas. Kegiatan observasi ini bertujuan untuk mendokumentasikan gambar objek budaya secara visual (foto), mengidentifikasi ciri-ciri fisik, bentuk, warna, dan atribut penting yang relevan secara budaya serta mencatat konteks penggunaannya dalam kehidupan masyarakat, seperti upacara adat atau fungsi keseharian.

2. Wawancara

Wawancara dilakukan secara terstruktur maupun semi-terstruktur dengan narasumber yang memiliki pengetahuan tentang budaya lokal, antara lain tokoh adat, budayawan atau sejarawan lokal dan Petugas dinas kebudayaan. Wawancara ini bertujuan untuk memperoleh informasi tekstual dan naratif tentang makna, fungsi, dan sejarah dari objek budaya.

3. Studi Pustaka

Studi pustaka dilakukan terhadap berbagai referensi tertulis yang berkaitan dengan teknologi kecerdasan buatan berbasis *Multimodal* dan model *BLIP* dan dokumentasi budaya lokal Rokan Hulu yang tersedia dalam buku, jurnal, prosiding, laporan dinas, dan situs web resmi pemerintah daerah.

3.3.1 Spesifikasi Data dan Output Sistem

Penelitian ini menggunakan dataset berupa gambar-gambar objek budaya lokal yang dikumpulkan dari dokumentasi lapangan, sumber terpercaya seperti arsip Dinas Kebudayaan Rokan Hulu, Buku dan Internet. Data gambar memiliki format .jpg dan .png dengan resolusi minimal 224x224 piksel agar kompatibel dengan *Vision Transformer (ViT)*. Gambar yang digunakan mencakup bangunan prasejarah dan alat musik khas Melayu.

Sistem ini menghasilkan output berupa narasi deskriptif otomatis yang menjelaskan isi gambar dalam konteks budaya lokal. Narasi ini dihasilkan oleh model *BLIP* yang menggabungkan representasi visual dan tekstual. Sebagai contoh, input berupa gambar rumah adat Istana Rokan IV Koto akan menghasilkan teks seperti: "Gambar ini menunjukkan Istana Rokan IV Koto, rumah adat tradisional yang dibangun dengan arsitektur khas Melayu."

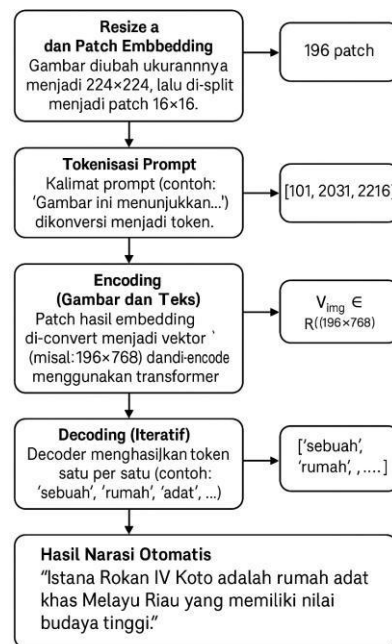
3.3 Analisa

Tahapan selanjutnya adalah melakukan analisa metode sistem dari penelitian skripsi ini, adapun tahapan analisa dalam penelitian tugas akhir ini adalah sebagai berikut :

3.4.1 Analisa Tranformer *Multimodal BLIP*

Model *Transformer Multimodal BLIP (Bootstrapped Language-Image Pretraining)* merupakan arsitektur kecerdasan buatan yang menggabungkan kemampuan pemrosesan gambar dan teks secara simultan untuk menghasilkan deskripsi naratif dari input visual. *BLIP* menggunakan dua komponen utama, yaitu *Vision Encoder* (biasanya berbasis ViT – *Vision Transformer*) untuk memahami fitur visual dari gambar, dan *Text Decoder* (berbasis *Transformer language* model seperti BERT atau GPT) untuk menghasilkan narasi yang relevan secara kontekstual.

Keunggulan *BLIP* terletak pada kemampuannya untuk melakukan image captioning, visual question answering, serta image-text retrieval dengan akurasi tinggi karena telah dilatih menggunakan teknik pretraining yang disesuaikan untuk modalitas ganda. Dalam konteks pengenalan objek budaya, *BLIP* memungkinkan sistem untuk tidak hanya mengenali bentuk visual objek, tetapi juga menyajikan deskripsi yang bersifat informatif dan bermakna budaya, menjadikannya sangat cocok untuk penerapan pelestarian pengetahuan budaya lokal berbasis AI.



Gambar 3. 2 Proses *Transformer Multimodal* Menggunakan *BLIP*

3.4.2 Analisa fungsi sistem aplikasi

Setelah melakukan tahapan analisis terhadap *Transformer Multimodal BLIP* maka selanjutnya adalah analisa fungsional sistem yang akan dibangun, adapun tahapan-tahapan analisis fungsional yaitu dalam pembuatan usecase diagram, activity diagram, sequence diagram dan class diagram.

3.4.3 Analisa Sistem Lama

Analisis sistem lama merupakan langkah penting dalam proses pengembangan atau peningkatan aplikasi, fungsi utama dari analisis sistem lama adalah untuk memahami kondisi saat ini dari sistem yang ada sebelum melakukan perubahan atau pengembangan baru, dengan melakukan analisis sistem lama secara menyeluruh, organisasi dapat membuat keputusan yang lebih baik mengenai bagaimana cara mengembangkan atau memperbarui sistem yang ada, hal ini juga membantu dalam meminimalkan risiko dan memastikan bahwa sistem

baru atau yang ditingkatkan dapat memenuhi kebutuhan bisnis dan pengguna dengan lebih efektif.

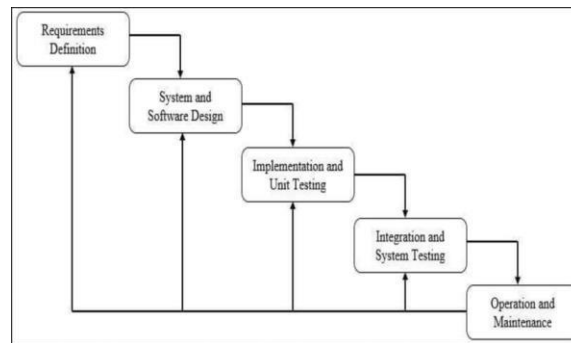
3.4.4 Analisa Sistem Baru

Analisis sistem baru adalah langkah penting dalam pengembangan atau implementasi aplikasi baru. Fungsi utama dari analisis sistem baru adalah untuk memastikan bahwa sistem yang akan dikembangkan atau diimplementasikan sesuai dengan kebutuhan dan tujuan organisasi. Dengan melakukan analisis sistem baru yang komprehensif, organisasi dapat memastikan bahwa sistem yang dikembangkan atau diimplementasikan akan memenuhi kebutuhan bisnis, meningkatkan efisiensi operasional, dan memberikan nilai tambah yang signifikan bagi organisasi dan pengguna.

3.4 Perancangan Sistem Aplikasi

Setelah tahapan analisa selesai dilakukan, maka tahapan selanjutnya adalah perancangan sistem, tahapan perancangan sistem pada penelitian ini menggunakan metode *waterfall* terdiri dari :

Metode *waterfall* adalah model yang paling banyak digunakan untuk tahap pengembangan. Model *waterfall* ini juga dikenal dengan nama model tradisional atau model klasik. Model air terjun ini menyediakan pendekatan alur hidup perangkat lunak secara sekuensial terurut dimulai dari analisis, desain, pengkodean, pengujian dan tahap pendukung (*support*) [34].



Gambar 3. 3 Metode *Waterfall*

Tahapan-tahapan metode pengembangan sistem informasi ini berikut adalah uraiannya [34]:

1. *Requirement*

Pada tahap ini pengembang harus mengetahui seluruh informasi mengenai kebutuhan *software* seperti kegunaan *software* yang diinginkan oleh pengguna dan batasan *software*. Informasi tersebut biasanya diperoleh dari wawancara, survey, ataupun diskusi. Setelah itu informs dianalisis sehingga mendapatkan data-data yang lengkap mengenai kebutuhan pengguna akan *software* yang akan dikembangkan.

2. *Design*

Tahap selanjutnya yaitu Desain. Desain dilakukan sebelum proses coding dimulai. Ini bertujuan untuk memberikan gambaran lengkap tentang apa yang harus dikerjakan dan bagaimana tampilan dari sebuah sistem yang diinginkan. Sehingga membantu menspesifikan kebutuhan hardware dan sistem, juga mendefinisikan arsitektur sistem yang akan dibuat secara keseluruhan.

3. *Development*

Proses penulisan code ada di tahap ini. Pembuatan *software* akan dipecah menjadi modul-modul kecil yang nantinya akan digabungkan dalam tahap selanjutnya. Dalam tahap ini juga akan dilakukan pemeriksaan lebih dalam terhadap modul yang sudah dibuat, apakah sudah memenuhi fungsi yang diinginkan atau belum.

4. *Testing*

Pada tahap keempat ini akan dilakukan penggabungan modul-modul yang sudah dibuat sebelumnya. Setelah itu akan dilakukan pengujian yang bertujuan untuk mengetahui apakah *software* sudah sesuai desain yang diinginkan dan apakah masih ada kesalahan atau tidak.

5. *Maintenance*

Maintenance adalah tahapan terakhir dari metode pengembangan *waterfall*. Di sini *software* yang sudah jadi akan dijalankan atau dioperasikan oleh penggunanya. Disamping itu dilakukan pula pemeliharaan yang termasuk perbaikan kesalahan, perbaikan implementasi unit sistem, peningkatan jasa sistem sesuai kebutuhan baru.

3.5.1 Perancangan Struktur Menu

Rancangan struktur menu ini kita perlu untuk memberikan gambaran terhadap menu-menu atau fitur pada sistem yang akan dibuat.

3.5.2 Perancangan Antar Muka (*Interface*)

Dalam mempermudah komunikasi antara sistem dengan pengguna, maka perlu di rancang antar muka (*interface*), dalam perancangan *interface* hal terpenting yang harus dilakukan ialah bagaimana menciptakan tampilan yang baik dan mudah dimengerti oleh pengguna

3.5 Implementasi Sistem Aplikasi

Beberapa komponen pendukung yang memiliki peran yang sangat penting dalam implementasi sistem diantaranya adalah perangkat keras (*hardware*) dan perangkat lunak (*software*) adapun spesifikasi dari perangkat keras (*hardware*) dan perangkat lunak (*software*) yang di gunakan sebagai berikut :

1. Perangkat keras (*hardware*), antara lain :

Processor : Intel Core i3

Memory : 8 GB RAM

Harddisk : 500 GB

2. Perangkat lunak (*software*), antara lain :

Operating System : Windows 11 Home Language

Tools : Visual Studio Code, Web Browser Google Chrome, Github, Streamlit.

3.6 Pengujian Aplikasi

Pengujian sistem ini dilakukan dengan cara menggunakan Black Box. Pengujian *Blackbox* ini berfokus pada perangkat lunak untuk mendapatkan hasil yang baik, apabila terjadi error atau tidak sesuai dengan tujuan yang diharapkan maka dilakukan penganalisaan sistem kembali hingga tidak ditemukan error.

3.7.1 *Blackbox* testing

Blackbox testing merupakan pengujian kualitas perangkat lunak yang berfokus pada fungsionalitas perangkat lunak, pengujian *Blackbox* testing bertujuan untuk menemukan fungsi yang tidak benar kesalahan antarmuka kesalahan pada struktur data, kesalahan perfomansi, kesalahan inisialisasi dan terminasi, dalam pengujian *Blackbox* testing digunakan alat untuk pengumpulan data yang disebut dengan *user acceptance test*, dokumen ini terdiri deskripsi indikator dari prosedur-prosedur pengujian fungsionalitas dari perangkat lunak

3.7 Kesimpulan dan Saran

Tahapan terakhir adalah menarik kesimpulan dari hasil penelitian yang didapatkan dalam penerapan *Transformer Multimodal* untuk pengenalan dan deskripsi objek budaya lokal Rokan Hulu berbasis gambar dan narasi otomatis, pada tahapan ini juga berisikan saran peneliti bagi pembaca untuk melakukan pengembangan terhadap penelitian ini kedepannya.