BAB III
ANALISIS DAN PERANCANGAN

3.1 Analisis

3.1.1 Identifikasi Masalah

Jumlah bahasa daerah di Indonesia adalah sebanyak 718 bahasa, dengan 90 persen di antaranya tersebar di wilayah Indonesia timur. Terdapat 428 jenis bahasa di Papua, salah satunya adalah bahasa di daerah Kokas, yang kini menghadapi ancaman kepunahan. Hal ini disebabkan oleh kurangnya dokumentasi, pemanfaatan, dan pelestarian bahasa tersebut di tengah arus modernisasi serta dominasi bahasa nasional dan global (Sumatera & Rahima, 2024).

Pelestarian bahasa daerah menjadi semakin penting dalam era globalisasi untuk memastikan keberlanjutan budaya dan identitas lokal (Apriani dkk., 2016). Namun, di daerah seperti Kokas, tidak banyak orang yang masih fasih atau bahkan memahami bahasa daerah tersebut. Kurangnya akses terhadap informasi dan materi pembelajaran dalam Bahasa Kokas mempercepat proses kepunahan.

Dalam konteks di atas, penelitian ini mengidentifikasi kebutuhan untuk mengembangkan alat penerjemah otomatis yang dapat menerjemahkan Bahasa Indonesia ke Bahasa Papua Kokas. Alat ini diharapkan dapat memudahkan komunikasi dan akses informasi bagi masyarakat yang berbicara dalam Bahasa Kokas, sekaligus menjadi upaya untuk melestarikan bahasa yang semakin terpinggirkan ini.

3.1.2 Pemecahan Masalah

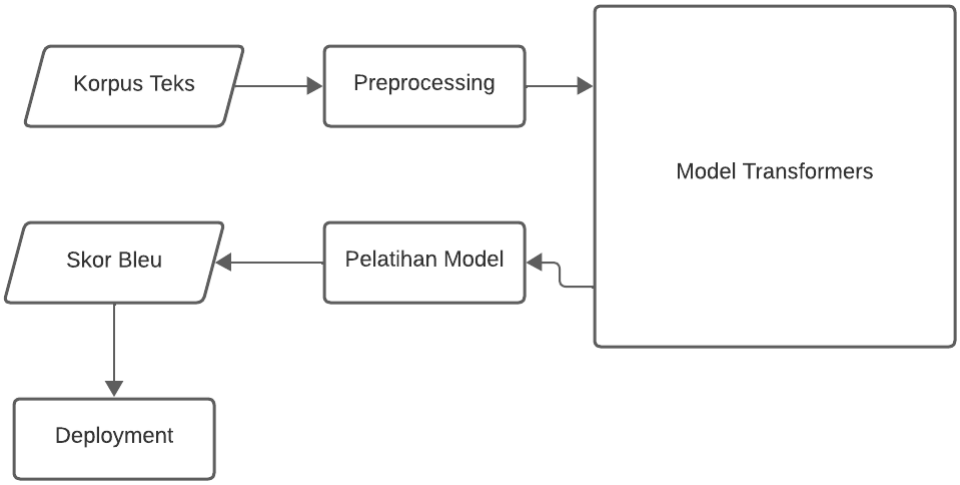
Berkaitan dengan permasalahan yang ditemukan, peneliti kemudian mengusulkan pengembangan sistem penerjemahan otomatis berbasis Transformer. Metode Transformer dipilih karena telah terbukti efektif dalam tugas penerjemahan

mesin dan dapat menangani konteks kalimat secara keseluruhan tanpa memerlukan urutan waktu seperti pada model RNN. Sistem ini dirancang untuk menerjemahkan teks dari Bahasa Indonesia ke Bahasa Papua Kokas dengan akurasi yang tinggi. Teknologi penerjemahan mesin dapat menjadi alat yang efektif untuk mendukung pelestarian bahasa daerah melalui penerjemahan dan pembelajaran yang lebih mudah diakses (Fauziyah dkk., 2022).

3.2 Perancangan

3.2.1 Perancangan Sistem Transformers

Pada penelitian metode Transformers untuk melakukan terjemahan bahasa terdapat beberapa tahapan yaitu:



Gambar 3.1 Perancangan Sistem Transformers

3.2.1.1 Korpus Teks

Korpus teks merujuk pada kumpulan data teks yang digunakan sebagai bahan dasar dalam proses pelatihan model penerjemahan. Dalam penelitian ini, korpus teks yang digunakan terdiri atas pasangan kalimat dalam Bahasa Indonesia dan Bahasa Papua Kokas berjumlah 2908 data paralel. Korpus ini diperoleh melalui wawancara dengan masyarakat asli di daerah Kokas, Papua. Data yang telah terkumpul kemudian disusun dalam format CSV dengan dua kolom: Bahasa Indonesia dan Bahasa Papua Kokas.

Proses pengumpulan data ini sangat penting karena korpus teks menjadi fondasi bagi model penerjemah yang akan dibangun. Kualitas dan representasi data yang digunakan sangat memengaruhi hasil akhir dari model penerjemah. Kuantitas dan kualitas korpus berperan penting dalam meningkatkan akurasi model penerjemahan statistik (Apriani dkk., 2016).

indonesian	papua_kokas
ada apa	amia kusafa
adopsi dapat menciptakan keluarga yang kuat dan beragam	akatik auwa ita karagatuni kuat adi beragam
aktivitas fisik teratur mendukung kesehatan jantung dan sistem kekebalan tubuh	aktivitas fisik teratur adukung kesehatan jantung adi sistem kekebalan tubuh
aku akan mencuci mobil sore ini	emau eburi mobil rera titi ige
aku baik-baik saja, bagaimana denganmu	yai kues weswatan akape adi o
aku baik-baik saja, terima kasih	yai fifian watan o
aku baru saja menonton film dokumenter tentang alam	yai fatak matan efogim film dokumenter tentang alam
aku baru saja menonton film petualangan, sangat seru!	yai matantami efogim film petualangan seru paskali
aku belajar tentang sejarah dunia	yai balajar tentang sejarah dunia
aku berencana untuk mendaki gunung	yai rencana buat esa ami kehi nasin
aku berencana untuk pergi ke jepang	yai rencana buat ati jepang
aku berencana untuk pergi ke jepang dan korea	e rencana buat eti jepang adi korea
aku biasanya membaca buku atau menonton film	yai biasa e baca buku atau efogim film
aku biasanya membuat jadwal harian yang seimbang	yai biasa ena jadwal harian yang seimbang
aku biasanya mengadakan pesta kecil dengan teman dan keluarga	yai biasa ana pesta girif adi e tamang adi keluarga sina

Gambar 3.2 Contoh Data korpus

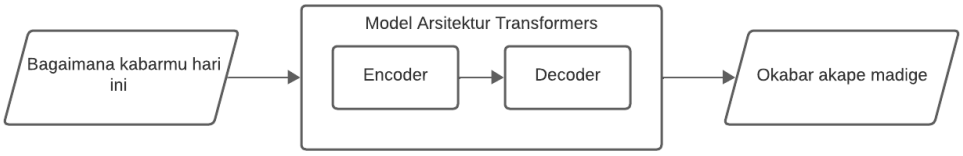
3.2.1.2 Preprocessing

Tahapan *preprocessing* pada penelitian ini dilaksanakan untuk memastikan bahwa data yang digunakan dalam pelatihan model bersih dan siap untuk diproses. *Preprocessing* mencangkup langkah-langkah berikut :

1. Memuat Dataset : pada bagian ini dataset yang dikumpulkan lalu di proses untuk membagi dua data yaitu data pelatihan dan data validasi. Pembagian dataset ini dilakukan untuk memastikan bahwa model dapat dilatih dengan jumlah data yang cukup dan diuji dengan data yang tidak pernah dilihat sebelumnya untuk mengukur kinerjanya secara objektif.
2. *Tokenizing* : Proses ini dilakukan untuk memecah teks dalam bahasa Indonesia dan Papua Kokas menjadi token-token yang dapat diproses oleh model. Tokenisasi dilakukan dengan menggunakan *MarianTokenizer* dari model *Helsinki-NLP/opus-mt-id-en*. Tokenisasi ini secara otomatis mencangkup proses seperti *case folding* dan *removing punctuation*. Tokenisasi merupakan tahap awal yang penting dalam text mining karena membantu mengubah teks yang tidak terstruktur menjadi lebih mudah di proses (Wahyuni dkk., 2017).
3. *Encoding* : Setelah tokenisasi, teks yang telah diubah menjadi token kemudian di-encode menjadi tensor menggunakan *Pytorch*. Tensor ini diperlukan agar data dapat diproses oleh model selama pelatihan. *Encoding* juga mencangkup *Padding*, yaitu penambahan token khusus agar semua input memiliki panjang yang sama, dan *Truncation*, yaitu pemotongan teks yang terlalu panjang. Pentingnya proses dalam mengubah teks menjadi format numerik yang dapat dipahami oleh model pembelajaran mesin (Sulistiyo, 2016).

3.2.1.3 Model Transformers

Model *Transformers*, Metode ini memungkinkan mesin untuk mengatasi pemahaman bahasa alami, terjemahan mesin, pengenalan entitas berbasis teks, dan berbagai tugas lainnya dengan sangat baik. Dalam penelitian ini, model menyertakan lapisan *encoder-decoder*. Lapisan encoder mengekstrak fitur dari kalimat *input* dan lapisan *decoder* mengembalikan nilai yang diekstraksi ke kalimat yang diterjemahkan.



Gambar 3.3 (contoh model transformers)

Model *Transformers* yang digunakan dalam penelitian ini adalah Model *MarianMT* dari *Helsinki-NLP/opus-mt-id-en*.Model ini sudah terlatih sebelumnya (*pre-trained*) dan digunakan sebagai dasar untuk melatih model penerjemah dari bahasa Indonesia ke bahasa Papua Kokas. Arsitektur *Transformers* dipilih karena kemampuannya yang ungguk dalam menangani masalah penerjemahan mesin tanpa memerlukan struktur urutan seperti RNN. Model ini terbukti efektif dalam tugas penerjemahan bahasa karena menggunakan *self-attention* yang mampu menangkap konteks kalimat secara keseluruhan (Vaswani dkk., 2017).

Pada tahap ini, model dimuat bersama dengan *tokenizer* yang sesuai. Model ini kemudian dilatih menggunakan dataset yang telah diproses sebelumnya. Selama pelatihan, model diberi pasangan kalimat dalam bahasa Indonesia dan Papua Kokas,

dan model diharapkan dapat mempelajari pola terjemahan yang tepat di antara kedua bahasa ini.

3.2.1.4 Pelatihan Model

Proses pelatihan model dilakukan dengan menggunakan dataset yang telah diproses. Model dilatih menggunakan *Trainer* dari pustaka *Transformers* dengan konfigurasi yang sudah ditentukan, seperti jumlah *epoch*, ukuran *batch*, dan strategi evaluasi. Selama pelatihan, model terus dievaluasi menggunakan data validasi untuk memantau kinerja dan memastikan bahwa model tidak mengalami *overfitting*. Dalam sebuah penelitian yang menggunakan pendekatan serupa untuk melatih model pembelajaran mesin, dimana proses evaluasi berkala sangat penting untuk mendapatkan model yang optimal (Roihan dkk., 2020).

Pelatihan dilakukan dalam beberapa epoch untuk memastikan model dapat mengkonvergensi ke solusi yang optimal. Selama pelatihan, model terus diperbarui untuk meningkatkan akurasi terjemahan, dengan log pelatihan yang disimpan untuk analisis lebih lanjut.

3.2.1.5 Skor Bleu

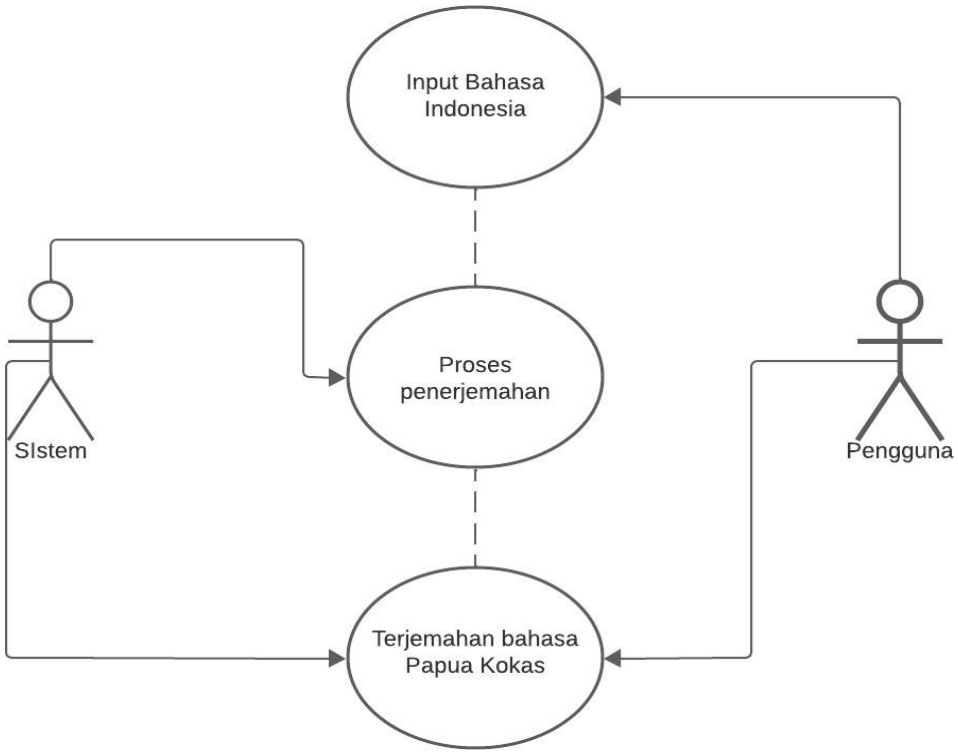
Metrik *BLEU (Bilingual Evaluation Understudy)* digunakan menilai kualitas hasil terjemahan yang dihasilkan pada model. Skor *BLEU* dihitung dengan membandingkan hasil terjemahan model dengan referensi terjemahan yang benar. Dalam penelitian ini, *BLEU* digunakan untuk mengevaluasi sejauh mana model menerjemahkan teks dari bahasa Indonesia ke bahasa Papua Kokas. *BLEU* adalah metrik yang umum digunakan dalam evaluasi penerjemahan mesin, karena mampu

memberikan gambaran tentang akurasi terjemahan berdasarkan n-gram yang cocok dengan referensi. Nilai *BLEU* yang tinggi menunjukkan bahwa hasil terjemahan sangat mendekati terjemahan yang diharapkan, yang sangat penting dalam konteks penerjemahan otomatis (Abidin, 2017).

3.2.1.6 Deployment

Pada tahap *deployment*, model yang telah dilatih diimplementasikan dalam sebuah sistem yang dapat diakses oleh pengguna. Dalam penelitian ini deployment dilakukan dengan mengintegrasikan model penerjemahan otomatis kedalam produk aplikasi website. Aplikasi ini dirancang untuk memungkinkan teks dalam bahasa Indonesia dan mendapatkan terjemahan dalam bahasa Papua Kokas secara real-time.

3.2.2 Perancangan *Use Case*



Gambar 3.4 Use Case

1. Input bahasa Indonesia

   Pengguna cukup memasukkan teks bahasa Indonesia
2. Proses penerjemahan

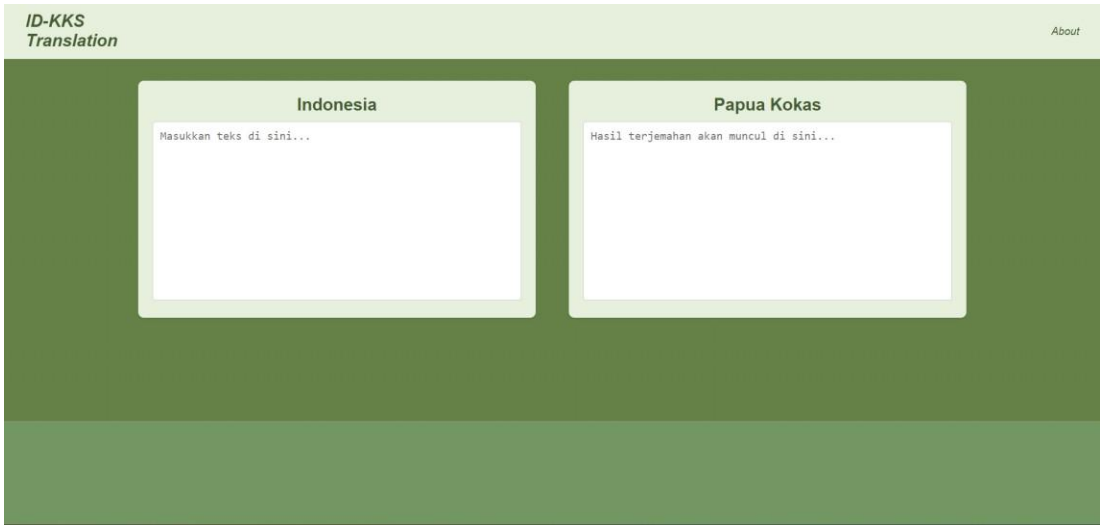
   Pada proses ini sistem akan langsung memproses bahasa Indonesia yang sudah di input oleh pengguna
3. Terjemahan bahasa Papua Kokas

   Pada tahap ini output yang di keluarkan sistem akan terbentuk dalam terjemahan yang di input oleh pengguna pada sebelumnya

3.2.3 Perancangan *User Interface* / *Mock-up* aplikasi

Dibawah ini merupakan rancangan website yang digunakan sebagai pengujian hasil penelitian:

- Halaman Home



Gambar 3.5 Halaman Home

- Halaman About

Gambar 3.6 Halaman About

3.3 Rancangan Pengujian

Pada rancangan pengujian menjelaskan tentang bagaimana rencana pengujian yang akan dilakukan. Metode yang digunakan contohnya *white box, black box, grey box* dan lain-lain.

Dalam Penelitian ini menggunakan dua pengujian yaitu:

3.3.1 Pengujian *Black box*

Pengujian black box pada model penerjemah bertujuan untuk memastikan bahwa model dapat secara konsisten menghasilkan terjemahan yang akurat dan berkualitas tinggi dari bahasa Indonesia ke bahasa Papua Kokas tanpa memerlukan pemahaman mendalam tentang algoritma atau struktur internal model. Pengujian ini penting untuk memastikan bahwa model siap digunakan dalam aplikasi dunia nyata, seperti di website yang kembangkan.

3.3.2 Pengujian *Matrik BLEU*

Pengujian BLEU yang dilakukan secara konsisten di berbagai tahap pelatihan dan dengan dataset uji yang berbeda akan memberikan pemahaman yang lebih mendalam mengenai kemampuan model dalam penerjemahan otomatis. proses pengujian dilakukan beberapa langkah :

1. Pengumpulan data uji yang digunakan dalam pengujian *BLEU* terdiri dari kalimat parallel bahasa Indonesia dan bahasa Papua Kokas. Data ini terbagi menjadi dua bagian utama, yaitu data pelatihan dan data validasi, yang masing-masing berfungsi untuk melatih dan menguji model (Shaw dkk., 2018).
2. Pengukuran Skor *BLEU*: setelah model dilatih menggunakan data pelatihan, model diuji dengan data validasi. Skor *BLEU* dihitung untuk menilai seberapa baik model dapat menghasilkan terjemahan yang mendekati terjemahan referensi. Metrik ini menggunakan pendekatan n-gram untuk mengevaluasi keakuratan terjemahan, dimana semakin tinggi skor *BLEU*, semakin baik kualitas terjemahan yang dihasilkan (Shaw dkk., 2018).

3. Evaluasi berkala: Evaluasi dilakukan pada setiap epoch pelatihan untuk memantau perkembangan skor *BLEU* dan memastikan model tidak mengalami overfitting. Evaluasi ini penting untuk menentukan jumlah *epoch* yang optimal dalam pelatihan model, mengacu pada metode yang telah diusulkan (Abidin, 2017).
4. Analisis hasil: Hasil dari pengukuran skor BLEU dianalisis untuk melihat tren peningkatan atau penurunan performa model. Peningkatan skor BLEU menunjukkan bahwa model semakin akurat dalam menghasilkan terjemahan. Stagnasi atau penurunan skor dapat mengindikasikan perlunya penyesuaian pada parameter pelatihan atau data yang digunakan (van der Wees dkk., 2017).