BAB I
PENDAHULUAN

1.1 Latar Belakang

Bahasa merupakan alat komunikasi yang sangat penting digunakan oleh manusia dalam menyampaikan informasi, seperti mengenalkan diri, bercerita, dan menyampaikan ide maupun gagasan. Bahasa berperan penting dalam interaksi antar manusia untuk memudahkan memahami apa yang disampaikan dengan baik dan benar. Indonesia memiliki beragam bahasa atas keberagaman suku dan budaya. Menurut artikel yang dimuat pada situs Kemendikbud (2019) jumlah keberagaman bahasa daerah yang dimiliki Indonesia adalah 718 bahasa daerah.

Keberagaman bahasa daerah menjadikan setiap bahasa daerah yang memiliki perbedaan kesulitan tersendiri, tidak terkecuali bahasa Tiochiu Pontianak daerah Kalimantan Barat, dengan susunan kata dalam kalimatnya yang sulit dipahami dan memiliki dialek yang sulit diikuti. Bahasa Tiochiu Pontianak memiliki keunikan sendiri, satu kata yang banyak arti dengan ucapan berbeda atau bahkan sama, dan beberapa susunan katanya berbeda dengan bahasa Indonesia. Contoh kalimat "Saya ke pasar menggunakan motor" dalam bahasa Tiochiunya adalah "Wa eng mopit khe pasat".

Di Indonesia, bahasa Tiochiu berasal perantauan suku Tiochiu ke berbagai belahan dunia, dari daerah Chaoshan dibagian timur provinsi Guangdong, Republik Rakyat Tiongkok. Bahasa Tiochiu banyak mempertahankan pelafalan dan kosakata bahasa Tiongkok kuno yang telah hilang seiring perkembangan jaman. Dengan begitu untuk membantu dalam memahami dan mencegah hilangnya bahasa Tiochiu, akan dibuat mesin penerjemah bahasa Indonesia ke bahasa Tiochiu Pontianak dengan faktor tambahan agar hasil penerjemahan yang dihasilkan memiliki tingkat akurasi yang baik untuk komputasi secara digital.

Mesin penerjemah dapat diartikan sebagai mesin yang bertugas secara otomatis untuk mengkonversikan atau mengalihkan sebuah teks kalimat dari sebuah bahasa ke bahasa yang lain. Mesin penerjemah juga dapat dimanfaatkan sebagai upaya komputasi dalam melestarikan bahasa daerah secara digital. Mesin penerjemah berbasis jaringan saraf tiruan (NMT) adalah sebuah pendekatan baru

dalam teknologi mesin menggunakan arsitektur recurrent neural network (RNN) pada bagian encoder dan decoder-nya, yang telah membuktikan performa NMT lebih baik dibanding SMT (*Statistik Machine Translation*) pada eksperimen 30 jenis penerjemahan. NMT yang digunakan saat ini oleh para peneliti mesin penerjemah adalah NMT berbasis attention (Bahdanau et al., 2015).

Dalam pengembangan mesin penerjemah jaringan saraf tiruan (NMT), beberapa peneliti dengan makalah sebelumnya yaitu arsitektur yang digunakan *recurrent neural network (RNN)* oleh (Kalchbrenner & Blunsom, 2013), *framework* model yang biasa dikenal adalah *sequence to sequence* oleh (Sutskever et al., 2014) dan (Cho et al., 2014), mekanisme attention (*alignment & translate*) oleh (Bahdanau et al., 2015), pendekatan NMT *attention* (Abidin et al., 2018), NMT dengan arsitektur Transformer (Sebagai et al., 2018), *Google Neural Machine Translation* (Untara & Setiawan, 2020), pengembangan metode NMT dengan Neural Network (Aristyanto & Kurniawan, 2021), pendekatan NMT dengan arsitektur Encoder-Decoder (Fauziyah et al., 2022). RNN sama dengan halnya sequence to sequence merupakan jenis arsitektur jaringan saraf tiruan yang pemrosesannya secara bertahap dipanggil berulang-ulang untuk memproses jaringan *encoder* (target input) dan *decoder* (target output) untuk mempelajari urutan informasi dan membuat model prediksi pada mesin penerjemah NMT. Penelitian lainnya yang melakukan penerjemahan satu arah dengan pengujian penggunaan penambahan *epoch* secara konsisten oleh (Gunawan et al., 2021). Penambahan *epoch* merupakan pengaturan parameter perulangan pelatihan dataset(himpunan data) dengan jumlah *epoch* pada mesin NMT. Melakukan penerjemahan dua arah oleh (Luong et al., 2015) dengan mekanisme attention pendekatan global yang menggunakan semua masukan kalimat sumber dan local yang hanya menggunakan subset atau sebagian masukan kalimat sumber dalam satu waktu.

Dalam beberapa tahun terakhir, mesin penerjemah jaringan saraf tiruan (NMT) telah menjadi populer dikalangan penelitian dan mencapai kesuksesan yang luar biasa dan menjadi metode pilihan baru dalam praktik mesin penerjemah bahkan sampai sekarang sudah diterapkan di google. Setelah NMT terkenal dari mesin learning yang ada, pada penelitian terbaru NMT lebih banyak penelitian berfokus

pada arsitektur RNN dan algoritma *end to end*. NMT dibagi menjadi dua jenis yaitu NMT dengan Mekanisme *Attention* atau disebut *classical* NMT dan tanpa *Attention* disebut *Greedy Decoding*. Greedy decoding sangat sedikit dilakukan oleh penelitian karena greedy decoding menggunakan algoritma perkiraan objektif arbitrary (sewenang-wenangnya atau sesuka hati).

Menurut (Gu et al., 2017) dengan judul "Trainable Greedy Decoding for Neural Machine Translation" penelitiannya dilakukan dengan mengikuti jaringan ketergantungan bersyarat (probabilitas) dan simbol dipilih dengan kondisi probabilitas tertinggi sejauh ini pada setiap node. Ini setara dengan memilih simbol terbaik satu per satu dari kiri ke kanan dalam pemodelan bahasa bersyarat. Rumus terjemahan dari greedy decoding dengan sebuah estimasi maksimum-a-posteriori dalam log p(Y | X). Dapat disimpulkan Greedy decoding merupakan sebuah algoritma perkiraan dengan objektif *arbitrary* (sewenang atau sesuka hati), dari kata Greedy yang berarti serakah/tamak dalam mengambil estimasi. Tujuan dari penelitiannya itu sebagai cara untuk mempelajari algoritma decoding untuk mesin terjemahan NMT dengan objektif arbitrary decoding, yang dapat dilatih untuk mengamati dan memanipulasi keadaan hidden state yang terlatih di NMT, dan dilatih dengan Teknik *critic-aware* (kritik-sadar) *actor gradient* dari varian *novel* yang *deterministic* (peristiwa yang nyata terjadi). Secara umum penelitian yang dilakukannya dengan delapan arah pasangan bahasa dan penentuan dua objektif dari validitas dan kegunaannya. Yang diusulkannya dari Greedy decoding adalah ide umum yang dapat diterapkan pada setiap pemodelan bahasa berulang, dan mengantisipasi penelitian di masa mendatang baik tentang dasar-dasar decoding yang dapat dilatih maupun di aplikasi untuk tugas yang lebih beragam seperti gambar pembuatan teks dan pemodelan dialog. Dapat disimpulkan bahwa greedy decoding merupakan encoder-decoder dasar dari NMT.

Menurut penemu mekanisme *attention,* Oleh Bahdanau dkk (2015) bahwa salah satu motivasi dibalik pendekatan yang diusulkan (*attention mechanism*) adalah penggunaan *context vector* yang memiliki panjang tetap dalam pendekatan *encoder-decoder* dasar. Penilitiannya menduga bahwa keterbatasan ini dapat membuat pendekatan *encoder-decoder* dasar berkinerja buruk pada kalimat yang panjang. Dapat disimpulkan mekanisme attention oleh Bahdanau adalah

pendekatan model encoder-decoder dasar dengan nilai skornya berdasarkan panjang kalimat sumber dan kalimat target.

Mesin penerjemah jaringan saraf tiruan adalah perangkat mesin penerjemah dengan penerapan ruang (atau otak) untuk membentuk jaringan saraf-saraf suatu bahasa dengan bahasa terjemahannya. Mesin penerjemah jaringan saraf tiruan dikenal dengan *neural machine translation* (NMT) adalah pendekatan mesin penerjemah yang menggunakan jaringan saraf tiruan atau biasa disebut *neural network* untuk memprediksi kemungkinan urutan masukan kata encoder-decoder dan biasanya melakukan permodelan seluruh kalimat dalam satu model yang terintegrasi. Mesin penerjemah jaringan saraf tiruan (NMT) dibangun dari framework model *sequence to sequence*. *Sequence to sequence* memiliki dua komponen penting yaitu jaringan saraf tiruan berulang (RNN) dengan arsitektur encoder dan decoder. Encoder membaca sumber masukan kata satu persatu untuk mendapatkan representasi vektor dari dimensi tetap, dan decoder mengkondisikan masukan dan mengekstrak kata keluaran satu per satu menggunakan RNN.

Masalah yang terdapat pada *sequence to sequence* adalah bahwa satu-satunya informasi yang diterima decoder dari encoder adalah encoder *hidden state* (representasi vektor yang seperti ringkasan numerik dari urutan masukan) yang mempunyai node panjang tetap dan tidak memungkinkan memiliki panjang yang sama seperti urutan masukan pada encoder. Jadi, untuk teks masukan yang panjang, secara tidak wajar mengharapkan decoder hanya menggunakan satu representasi vektor untuk menghasilkan terjemahan. Dengan seperti itu sistem sulit mengingat urutan dengan masukan yang banyak.

Untuk mengatasi masalah yang ada pada sequence to sequence adalah dengan menggunakan mekanisme attention. Attention adalah antarmuka antara encoder dan decoder yang menyediakan decoder dengan informasi dari setiap hidden state encoder. Dengan pengaturan ini, model dapat secara selektif fokus pada bagian-bagian berguna seperti dari urutan masukan dan mempelajari alignment/penyelarasan di antara keduanya. Dengan masalah tersebut maka mekanisme attention yang bisa membuat satu representasi vektor dari setiap langkah waktu encoder sehingga akan membantu model untuk mengatasi secara

efektif pada kalimat masukan yang panjang dan membuat akurasi terjemahan menjadi lebih baik.

Nilai akurasi merupakan hal yang sangat penting untuk diperhatikan dalam mesin penerjemah jaringan saraf tiruan (NMT). Tingkat kesesuaian penerjemahan bahasa sumber dengan arti sesungguhnya dalam bahasa target dilihat dari nilai akurasi yang dihasilkan. Akurasi dari suatu NMT adalah tingkat kedekatan kuantitas (angka) terhadap nilai yang di uji dalam mesin. Nilai akurasi yang baik pada NMT adalah semakin banyaknya kalimat masukan nilai yang dihasilkan. Semakin tinggi nilai akurasi yang dihasilkan maka semakin bagus kualitas terjemahan yang dihasilkan oleh mesin. Nilai dari kualitas terjemahan merupakan hasil kuantitas akhir yang dihasilkan dari nilai akurasi. Kualitas terjemahan dipengaruhi banyaknya jumlah kalimat masukan yang disediakan. Semakin banyak jumlah kalimat masukan maka kualitas terjemahan semakin baik.

Attention adalah hubungan antar encoder dan decoder yang menyediakan decoder dengan informasi dari setiap hidden state encoder dan memiliki perhitungan nilai skor-nya. Cara kerja attention meniru pada bagaimana ciri-ciri khas kalimat dengan menghubungkan kata-kata menjadi sebuah kalimat, sederhananya hubungan satu kata dengan kata yang lain dalam satu kalimat yang menentukan seberapa kuat dan atau dekat hubungan atau relasi sebuah kata dengan kata yang lain. Dengan mengenali hubungan atau susunan kalimat tertentu, maka NMT dapat menentukan susunan kalimat yang diinginkan. Jadi, mekanisme attention menentukan nilai berdasarkan seberapa kuat sebuah kata yang terhubung dengan kata yang lain. Nilai tersebut kemudian berkontribusi kepada struktur hierarki yang membantu neural network menyimpulkan seberapa penting sebuah kata dalam konteks kalimat tertentu. Contoh kalimat sederhana “Adik makan nasi goreng” dapat dijabarkan hubungan kata antara “makan” dan “nasi” lebih kuat daripada kata “makan” dengan “goreng”. Sehingga secara makna kata “makan” menunjukkan bahwa kalimat tersebut merupakan sejenis makanan seperti “nasi”. Sedangkan kata “goreng” menunjukan makna ciri-ciri suatu makanan tetapi hubungan dengan kata “makan” sangat rendah atau bahkan tidak memiliki hubungan.

Berdasarkan penjelasan yang sudah dipaparkan, maka penelitian ini akan mengimplementasikan mesin penerjemah jaringan saraf tiruan (NMT) dengan mekanisme attention model Bahdanau tersebut untuk mengetahui nilai akurasi dan kualitas terjemahan dengan menggunakan tools framework tensorflow (sebuah kerangka kerja komputasional untuk membuat model *machine learning* yang memiliki modul dan library dari bahasa *Python* dan C++) dalam penerjemahan bahasa Indonesia ke bahasa Tiochiu Pontianak berdasarkan hasil evaluasi otomatis BLEU (*Bilingual Evaluation Understudy*) dan pengujian oleh ahli bahasa.

1.2 Perumusan Masalah

Penelitian terkait NMT masih sedikit dilakukan pada penelitian-penelitian di Indonesia. Khususnya penelitian untuk mengetahui hasil nilai akurasi yang dihasilkan mekanisme Attention Bahdanau dari mesin penerjemah jaringan saraf tiruan untuk bahasa Indonesia ke bahasa daerah yang masih sangat sedikit dilakukan. Dengan demikian pada penelitian ini dapat dirumuskan masalah, bagaimana menguji nilai akurasi pada mesin penerjemah jaringan saraf tiruan (NMT) menggunakan mekanisme Attention Bahdanau pada penerjemahan bahasa Indonesia ke bahasa Tiochiu Pontianak.

1.3 Tujuan Penelitian

Adapun tujuan dari penelitian ini adalah untuk mengetahui nilai akurasi dan kualitas terjemahan berdasarkan evaluasi otomatis BLEU terhadap mesin penerjemah jaringan saraf tiruan (NMT) pada penerjemahan bahasa Indonesia ke bahasa Tiochiu Pontianak dengan mekanisme Attention Bahdanau.

1.4 Pembatasan Masalah

Beberapa hal yang menjadi batasan masalah dalam penelitian ini adalah sebagai berikut:

1. Penerjemahan yang dilakukan adalah penerjemahan satu arah dari bahasa Indonesia ke bahasa Tiochiu Pontianak.
2. Menggunakan arsitektur recurrent neural network RNN encoder-decoder.

3. Implementasi mesin penerjemah jaringan saraf tiruan menggunakan *text generation* dengan sumber terbuka dari *framework deep learning tensorflow*.
4. Menerapkan satu mekanisme yaitu mekanisme Attention Bahdanau dengan *function score alignment additive/concat*.
5. Mengukur akurasi skor dengan BLEU.

1.5 Sistematika Penulisan

Sistematika penulisan penelitian ini disusun untuk memberikan gambaran umum tentang penelitian yang dijalankan. Sistematika laporan tugas akhir ini disusun dalam 5 (lima) bab yang terdiri dari Bab I Pendahuluan, Bab II Tinjauan Pustaka, Bab III Metodologi Penelitian, Bab IV Hasil dan Analisis, serta Bab V Penutup.

Bab I Pendahuluan adalah bab yang berisi latar belakang penelitian, perumusan masalah, tujuan penelitian, pembatasan masalah, dan sistematika penulisan.

Bab II Tinjauan Pustaka adalah bab yang berisi uraian sistematis tentang hasil-hasil penelitian yang didapat dari peneliti-peneliti terdahulu dan landasan teori yang ada hubungannya dengan penelitian yang akan dilakukan.

Bab III Metodologi Penelitian adalah bab yang berisi tentang bahan penelitian, perangkat penelitian yang digunakan, metode yang akan digunakan pada penelitian, diagram alir penelitian serta perancangan pengujian yang akan dilakukan pada penelitian.

Bab IV Hasil dan Analisis adalah bab yang berisi hasil penelitian, penjelasan mengenai implementasi metode yang digunakan, hasil analisis dari setiap pengujian. Bagian yang ditampilkan akan dilakukan analisis terlebih dahulu untuk mengarah kepada suatu kesimpulan.

Bab V Penutup adalah bab yang berisi kesimpulan dari penelitian yang telah dilakukan dan saran atau rekomendasi untuk perbaikan, pengembangan atau kesempurnaan dan kelengkapan penelitian yang telah dilakukan.