PENGGUNAAN TEKNIK DATA AUGMENTATION UNTUK MENINGKATKAN KINERJA NEURAL MACHINE TRANSLATION PADA TERJEMAHAN BAHASA INDONESIA KE BAHASA DAYAK BANJUR

SKRIPSI

Program Studi Sarjana Informatika

Jurusan Informatika

Oleh:

NATALIA ARTIKA

NIM D1041211008

FAKULTAS TEKNIK

UNIVERSITAS TANJUNGPURA

PONTIANAK

2025

HALAMAN PERNYATAAN

Yang bertanda tangan di bawah ini:

Nama : Natalia Artika

NIM : D1041211008

menyatakan bahwa dalam skripsi yang berjudul "Penggunaan Teknik Data Augmentation Untuk Meningkatkan Kinerja Neural Machine Translation Pada Terjemahan Bahasa Indonesia Ke Bahasa Dayak Banjur" tidak terdapat karya yang pernah diajukan untuk memperoleh gelar sarjana di suatu perguruan tinggi manapun. Sepanjang pengetahuan Saya, tidak terdapat karya atau pendapat yang pernah ditulis atau diterbitkan oleh orang lain, kecuali yang secara tertulis diacu dalam naskah ini dan disebutkan dalam Daftar Pustaka.

Demikian pernyataan ini dibuat dengan sebenar-benarnya. Saya sanggup menerima konsekuensi akademis dan hukum di kemudian hari apabila pernyataan yang dibuat ini tidak benar.

Pontianak, 14 Juli 2025



Natalia Artika

NIM D1041211008

KEMENTERIAN PENDIDIKAN TINGGI, SAINS, DAN TEKNOLOGI
UNIVERSITAS TANJUNGPURA
FAKULTAS TEKNIK
Jalan Prof. Dr. H. Hadari Nawawi Pontianak 78124
Telepon (0561) 740186, WA: +6282152280907
Email : ft@untan.ac.id Website : http://teknik.untan.ac.id

HALAMAN PENGESAHAN

PENGGUNAAN TEKNIK DATA AUGMENTATION UNTUK MENINGKATKAN KINERJA NEURAL MACHINE TRANSLATION PADA TERJEMAHAN BAHASA INDONESIA KE BAHASA DAYAK BANJUR

Program Studi Sarjana Informatika
Jurusan Informatika

Oleh:

Natalia Artika
NIM D1041211008

Telah dipertahankan di depan Penguji Skripsi pada tanggal 14 Juli 2025
dan diterima sebagai salah satu persyaratan untuk memperoleh gelar sarjana.

Susunan Penguji Skripsi:

Ketua,

Prof. Dr. Herry Sujaini, S.T., M.T.
NIP. 196806291997021001

Penguji Utama,

Dr. Arif Bijaksana Putra Negara, S.T., M.T.
NIP. 197208081998021002

Sekretaris,

Rina Septiriana, S.T., M.Cs.
NIP. 198709232020122001

Penguji Pendamping,

M. Tsana'uddin Farid, S. Kom., M.Eng.
NIP. 199609162024061001

Pontianak, 14 Juli 2025
Dekan,

Dr. Ing. Ir. Slamet Widodo, M.T., IPM
NIP. 196712231992031002

HALAMAN PERSEMBAHAN

Skripsi ini didedikasikan kepada Bapak, Mamak, Kakak, Abang dan sahabat-sahabatku.

PRAKATA

Puji dan syukur penulis panjatkan ke hadirat Tuhan Yang Maha Esa karena atas rahmat dan karunia-Nya penulis dapat menyelesaikan penyusunan skripsi yang berjudul "Penggunaan Teknik Data Augmentation untuk Meningkatkan Kinerja Neural Machine Translation pada Terjemahan Bahasa Indonesia ke Bahasa Dayak Banjur" dengan baik. Skripsi ini disusun sebagai salah satu syarat untuk memperoleh gelar Sarjana Komputer pada Program Studi Sarjana Informatika, Fakultas Teknik, Universitas Tanjungpura.

Penelitian ini dilatarbelakangi oleh pentingnya pelestarian bahasa daerah, khususnya bahasa Dayak Banjur, melalui pemanfaatan teknologi penerjemahan berbasis neural. Dalam prosesnya, penulis menerapkan beberapa teknik *data augmentation* untuk meningkatkan kinerja model *Neural Machine Translation* yang digunakan. Harapannya, hasil penelitian ini dapat memberikan kontribusi dalam bidang pengolahan bahasa alami serta upaya pelestarian bahasa daerah yang semakin langka digunakan.

Penulis mengucapkan terima kasih yang sebesar-besarnya kepada Bapak Prof. Dr. Herry Sujaini, S.T., M.T., selaku dosen pembimbing utama dan Ibu Rina Septiriana, ST, M. Cs selaku dosen pembimbing kedua, yang telah memberikan bimbingan, arahan, dan dukungan selama proses penyusunan skripsi ini.

Penulis menyadari bahwa skripsi ini masih jauh dari sempurna. Oleh karena itu, kritik dan saran yang membangun sangat penulis harapkan untuk perbaikan di masa mendatang. Semoga skripsi ini dapat memberikan manfaat bagi pembaca dan pihak-pihak yang berkepentingan.

Pontianak, 14 Juli 2025

Penulis,



Natalia Artika

ABSTRAK

Penelitian ini dilatarbelakangi oleh kurangnya ketersediaan data korpus paralel bahasa daerah dan kajian penerjemahan otomatis untuk bahasa daerah, khususnya Bahasa Dayak Banjur. Untuk menjawab permasalahan tersebut, penelitian ini bertujuan untuk membangun model *Neural Machine Translation* (NMT) berbasis *attention* dengan menerapkan berbagai teknik augmentasi data teks guna meningkatkan kinerja penerjemahan dari Bahasa Indonesia ke Bahasa Dayak Banjur. Sebanyak lima teknik augmentasi diterapkan, yaitu *synonym replacement, stemming + synonym replacement, random insertion, random swap*, dan *random deletion*. Data awal terdiri atas 5000 pasangan kalimat paralel, yang kemudian di augmentasi menjadi masing-masing 5000 pasangan kalimat baru. 5000 pasangan kalimat baru ini kemudian ditambahkan dengan data awal sehingga menjadi 10000 pasang kalimat untuk setiap data teknik augmentasi. Model dibangun menggunakan arsitektur *Encoder-Decoder* berbasis *Recurrent Neural Network* (RNN) dengan mekanisme Bahdanau *Attention*. Proses pelatihan dilakukan secara terpisah untuk setiap dataset hasil augmentasi, dan evaluasi dilakukan menggunakan metrik BLEU untuk mengukur akurasi hasil terjemahan. Hasil evaluasi menunjukkan bahwa teknik *synonym replacement* menghasilkan skor BLEU tertinggi sebesar 48,19%, diikuti oleh *stemming + synonym replacement* dengan skor 46%, dan *baseline* tanpa augmentasi sebesar 42,17%. Sebaliknya, tiga teknik lainnya justru menunjukkan penurunan performa, di mana *random deletion* memperoleh skor terendah sebesar 33,27%. Temuan ini menunjukkan bahwa augmentasi berbasis semantik lebih efektif dibandingkan teknik yang mengganggu struktur kalimat. Penelitian ini memberikan kontribusi terhadap pengembangan NMT untuk bahasa daerah dengan sumber daya terbatas dan menegaskan pentingnya pemilihan teknik augmentasi yang sesuai untuk meningkatkan performa sistem penerjemahan otomatis.

Kata kunci: Bahasa Dayak Banjur, Bahdanau *Attention*, BLEU, Data *Augmentation*, *Neural Machine Translation*.

ABSTRACT

This study is motivated by the limited availability of parallel corpora and automatic translation research for regional languages, particularly the Dayak Banjur language. To address this gap, the study aims to develop an attention-based Neural Machine Translation (NMT) model by applying various text data augmentation techniques to enhance translation performance from Indonesian to Dayak Banjur. Five augmentation techniques are implemented: synonym replacement, stemming + synonym replacement, random insertion, random swap, and random deletion. The initial dataset consists of 5,000 parallel sentence pairs, which are then augmented to produce 5,000 new sentence pairs for each technique. These augmented pairs are combined with the original data, resulting in 10,000 sentence pairs per augmented dataset. The model is built using an encoder-decoder architecture based on a Recurrent Neural Network (RNN) with Bahdanau Attention mechanism. Each augmented dataset is trained separately, and evaluation is conducted using the BLEU metric to measure translation accuracy. The evaluation results show that the synonym replacement technique achieves the highest BLEU score of 48.19%, followed by stemming + synonym replacement with 46%, and the baseline without augmentation at 42.17%. On the other hand, the remaining three techniques demonstrate a decline in performance, with random deletion obtaining the lowest score of 33.27%. These findings indicate that semantically-based augmentation is more effective than techniques that disrupt sentence structure. This study contributes to the development of NMT for low-resource regional languages and highlights the importance of selecting appropriate augmentation strategies to enhance the performance of automatic translation systems.

Keywords: Bahdanau Attention, BLEU, Data Augmentation, Dayak Banjur Language, Neural Machine Translation.

DAFTAR ISI

HALAMAN PERNYATAAN ii
HALAMAN PENGESAHAN iii
HALAMAN PERSEMBAHAN iv
PRAKATA v
ABSTRAK vi
ABSTRACT vii
DAFTAR ISI viii
DAFTAR TABEL xi
DAFTAR GAMBAR xii
DAFTAR KODE PROGRAM xiv
DAFTAR LAMPIRAN xvi
BAB I PENDAHULUAN 1
1.1 Latar Belakang 1
1.2 Perumusan Masalah 2
1.3 Tujuan Penelitian 3
1.4 Pembatasan Masalah 3
1.5 Sistematika Penulisan 4
BAB II TINJAUAN PUSTAKA 5
2.1 Penelitian Terkait 5
2.2 Korpus 6
2.3 Bahasa Dayak Banjur 7
2.4 Augmentasi Teks 7
2.5 NMT 9
2.6 *Sequence to Sequence Model* 9
2.7 *RNN encoder-decoder* 11
2.8 Tipe Arsitektur RNN 16
2.8.1 *LSTM (Long Short Term Memory)* 16
2.8.2 *GRU (Gated Recurrent Unit)* 16
2.8.3 *Bi-RNN (Bidirectional RNN)* 17
2.9 Mekasnisme *Attention* 17
2.9.1 *Attention* Bahdanau 18
2.10 *Framework deep learning NMT* 19
2.11 Python 20
2.12 BLEU 20
2.13 *Flowchart* 21
BAB III METODOLOGI PENELITIAN 23
3.1 Data Penelitian 23
3.2 Analisis Kebutuhan 23

3.2.1 Perangkat Keras (*Hardware*) 23
3.2.2 Perangkat Lunak (*Software*) 23
3.3 Metodologi Penelitian 24
3.4 Metode Penelitian 25
3.4.1 Pengumpulan Data 26
3.4.2 Pembuatan Korpus Teks Paralel 26
3.4.3 *Preprocessing* 26
3.4.4 Penerapan Augmentasi 27
3.4.5 Membangun Model NMT 31
3.4.6 Latih Model Dengan Data Awal 33
3.4.7 Latih Model Dengan Data Hasil Augmentasi 33
3.4.8 Evaluasi BLEU 33
3.4.9 Perbandingan Hasil Evaluasi BLEU 33
3.4.10 Analisis Hasil Pengujian 34
3.4.11 Penarikan Kesimpulan 34

BAB IV HASIL DAN ANALISIS 35

4.1 Hasil Penelitian 35
4.2 Hasil Pengumpulan Data 35
4.3 Pembuatan Korpus Paralel Data Latih dan Data Uji 35
4.4 *Preprocessing* Data 38
4.5 Augmentasi Data 40
4.5.1 Persiapan Kamus Sinonim 41
4.5.2 *Synonym Replacement* 42
4.5.3 *Stemming + Synonym Replacement* 45
4.5.4 *Random Insertion* 47
4.5.5 *Random Swap* 50
4.5.6 *Random Deletion* 53
4.5.7 Perbandingan Statistik Data Awal dan Data Augmentasi 55
4.6 Pembangunan Model NMT (*Neural Machine Translation*) 57
4.6.1 Menyiapkan *Notebook* Kaggle 58
4.6.2 *Import library* 59
4.6.3 Pembacaan Data 60
4.6.4 Fungsi *Preprocessing* Data 61
4.6.5 Implementasi Fungsi *Preprocessing* Data 62
4.6.6 Menghitung Jumlah Kata Setiap Kalimat 63
4.6.7 Tokenisasi dan *Padding* 63
4.6.8 *Tuning* dan Pengaturan *Hyperparameter* Model 65
4.6.9 *Custom Loss Function* 67
4.6.10 Bahdanau *Attention* 68
4.6.11 *Encoder* 69
4.6.12 *Decoder* 70
4.6.13 Teknik *Teacher Forcing* 71
4.6.14 *Custom Training Loop* 72
4.7 *Training* Model 74
4.7.1 Model dengan Data Tanpa Augmentasi 74
4.7.2 Model dengan Data *Synonym Replacement* 77
4.7.3 Model Dengan Data *Stemming + Synonym Replacement* 79

4.7.4 Model dengan Data *Random Insertion* 82
4.7.5 Model dengan Data *Random Swap* 84
4.7.6 Model dengan Data *Random Deletion* 87
4.7.7 Perbandingan *Training* Model..... 90
4.8 Evaluasi Model 91
4.8.1 Rangkaian Evaluasi Model 91
4.8.2 Evaluasi Model dengan Data Tanpa Augmentasi 95
4.8.3 Model dengan Data *Synonym Replacement* 97
4.8.4 Model dengan Data *Stemming* + *Synonym Replacement* 99
4.8.5 Model dengan Data *Random Insertion* 101
4.8.6 Model dengan Data *Random Swap* 103
4.8.7 Model dengan Data *Random Deletion* 105
4.9 Analisis Hasil Pengujian 107
BAB V KESIMPULAN DAN SARAN 109
5.1 Kesimpulan 109
5.2 Saran 110
DAFTAR RUJUKAN 111

DAFTAR TABEL

Tabel 2.1 Teknik Augmentasi Teks 8
Tabel 2.2 Simbol-simbol pada *flowchart* 22
Tabel 4.1 Contoh Pasangan Kalimat Pararel 37
Tabel 4.2 Korpus Pararel Data Uji.....38
Tabel 4.3 Contoh Data Sesudah *Preprocessing* 40
Tabel 4.4 Perbandingan Kalimat Asli dan Hasil *Synonim Replacement* 45
Tabel 4.5 Perbandingan Kalimat Asli dan Hasil *Stemming* + *Synonym Replacement* 47
Tabel 4.6 Perbandingan Kalimat Asli dan Hasil *Random Insertion* 51
Tabel 4.7 Perbandingan Kalimat Asli dan Hasil *Random Swap* 53
Tabel 4.8 Perbandingan Kalimat Asli dan Hasil *Random Deletion* 55
Tabel 4.9 Perbandingan Parameter Model 91
Tabel 4.10 Hasil Prediksi Model Tanpa Augmentasi 96
Tabel 4.11 Hasil Prediksi Model *Synonym Replacement* 98
Tabel 4.12 Hasil Prediksi Model *Stemming* + *Synonym Replacement* 100
Tabel 4.13 Hasil Prediksi Model *Random Insertion* 102
Tabel 4.14 Hasil Prediksi Model *Random Swap* 104
Tabel 4.15 Hasil Prediksi Model *Random Deletion* 106
Tabel 4.16 Hasil Evaluasi Semua Model NMT 108

DAFTAR GAMBAR

Gambar 2.1 *Encoder-decoder* dalam S*equence to Sequence Model* 9
Gambar 2.2 Jenis S*equence to Sequence Model* 10
Gambar 2.3 Struktur dasar RNN 11
Gambar 2.4 Proses menerjemahkan bahasa 12
Gambar 2.5 Representasi RNN 13
Gambar 2.6 Representasi RNN *loop* 14
Gambar 2.7 Penjabaran RNN 14
Gambar 2.8 *Encoder-decoder* dan *Sequence to Sequence Model* 15
Gambar 2.9 *Attention based encoder-decoder architectures* 17
Gambar 2.10 Proses *attention* bahdanau 18
Gambar 2.11 Rumus evaluasi BLEU 21
Gambar 2.12 Rumus *brevity penalty* 21
Gambar 3.1 Rancangan arsitektur sistem 31
Gambar 3.2 Metodologi penelitian 24
Gambar 3.3 Metode penelitian 25
Gambar 3.4 Alur augmentasi *synonym replacement* 27
Gambar 3.5 Alur augmentasi *stemming + synonym replacement* 28
Gambar 3.6 Alur augmentasi *random insertion* 29
Gambar 3.7 Alur augmentasi *random swap* 30
Gambar 3.8 Alur augmentasi *random deletion* 31
Gambar 4.1 Tampilan *repository github* tesaurus 41
Gambar 4.2 *Install library* sastrawi 46
Gambar 4.3 Statistik jumlah kata 56
Gambar 4.4 Statistik kata unik 56
Gambar 4.5 Statistik rata-rata panjang kalimat 57
Gambar 4.6 Arsitektur model NMT 58
Gambar 4.7 Tampilan *notebook* Kaggle 58
Gambar 4.8 Hasil *Tuning* Model Tanpa Augmentasi 75
Gambar 4.9 Pengaturan *Hyperparameter* Model Tanpa Augmentasi 75
Gambar 4.10 *Training* Model tanpa Augmentasi 76
Gambar 4.11 Grafik *Training* dan Evalusasi *Loss* Tanpa Augmentasi 77
Gambar 4.12 Hasil *tuning* model *synonym replacement* 77

Gambar 4.13 Pengaturan *hyperparameter* model *synonym replacement* 78
Gambar 4.14 Training model *synonym replacement* 78
Gambar 4.15 Grafik *training* dan evaluasi *loss synonym replacement*..... 79
Gambar 4.16 Hasil *tuning* model *stemming + synonym replacement* 80
Gambar 4.17 Pengaturan *hyperparameter* model *stemming + synonym replacement*..... 80
Gambar 4.18 *Training* model *stemming + synonym replacement*..... 81
Gambar 4.20 Hasil tuning model *random insertion* 82
Gambar 4.21 Pengaturan *hyperparameter* model *random insertion* 83
Gambar 4.22 Training model *random insertion* 83
Gambar 4.23 Grafik *training* dan evaluasi *loss random insertion*..... 84
Gambar 4.24 Hasil *tuning* model *random swap* 85
Gambar 4.25 Pengaturan *hyperparameter* model *random swap* 85
Gambar 4.26 Training model *random swap* 86
Gambar 4.27 Grafik *training* dan evaluasi *loss random swap*..... 87
Gambar 4.28 Hasil *tuning* model *random deletion*..... 88
Gambar 4.29 Pengaturan *hyperparameter* model *random deletion*..... 88
Gambar 4.30 *Training* model *random deletion* 89
Gambar 4.31 Grafik *Training* dan Evaluasi *Loss Random Deletion* 90
Gambar 4.32 *Install* dan *import* nltk..... 94
Gambar 4.33 Hasil evaluasi BLEU model tanpa augmentasi..... 97
Gambar 4.34 Hasil evaluasi BLEU model *synonym replacement*..... 99
Gambar 4.35 Hasil evaluasi BLEU model *stemming + synonym replacement*.. 101
Gambar 4.36 Hasil evaluasi BLEU model *random insertion* 103
Gambar 4.37 Hasil evaluasi BLEU model *random swap* 105
Gambar 4.38 Hasil evaluasi BLEU model *random deletion*..... 107

DAFTAR KODE PROGRAM

Kode Program 4.1 Fungsi *preprocessing* 39
Kode Program 4.2 Kamus sinonim bahasa Indonesia dalam file JSON 42
Kode Program 4.3 *Library* pendukung proses augmentasi *synonym replacement* 42
Kode Program 4.4 Kode program untuk membaca kamus sinonim dari JSON.... 43
Kode Program 4.5 Kode program untuk mendapatkan sinonim..... 43
Kode Program 4.6 Kode program untuk mengganti kata dengan sinonimnya 44
Kode Program 4.7 Kode program untuk membaca data dari CSV 44
Kode Program 4.8 Kode program *synonym replacement* pada kalimat..... 44
Kode Program 4.9 Kode program untuk menyimpan hasil augmentasi 44
Kode Program 4.10 Kode program untuk melakukan *stemming* 46
Kode Program 4.11 *Library* pendukung proses augmentasi *random insertion* 47
Kode Program 4.12 Kode program untuk membaca kamus sinonim dari JSON.. 48
Kode Program 4.13 Kode program untuk mendapatkan sinonim..... 48
Kode Program 4.14 Kode program untuk menyisipkan sinonim kata secara acak49
Kode Program 4.15 Kode program untuk membaca data dari CSV 49
Kode Program 4.16 Implementasi *random insertion* pada kalimat 49
Kode Program 4.17 Kode program untuk menyimpan hasil augmentasi 49
Kode Program 4.18 *Library* pendukung untuk proses augmentasi r*andom swap* 50
Kode Program 4.19 Kode program untuk menukar posisi kata 51
Kode Program 4.20 Kode program untuk membaca data dari CSV 51
Kode Program 4.21 Implementasi *random swap* pada kalimat 51
Kode Program 4.22 Kode program untuk menyimpan hasil augmentasi 52
Kode Program 4.23 *Library* pendukung proses augmentasi r*andom deletion*..... 53
Kode Program 4.24 Kode program untuk menghapus secara acak 53
Kode Program 4.25 Kode program untuk membaca data dari CSV 54
Kode Program 4.26 Implementasi *random deletion* pada kalimat..... 54
Kode Program 4.27 Kode program untuk menyimpan hasil augmentasi 54
Kode Program 4.28 Kode program untuk *import library* 60
Kode Program 4.29 Kode program pembacaan dataset 61
Kode Program 4.30 Kode program *preprocessing* data latih dan data validasi.... 62
Kode Program 4.31 Kode program *preprocessing* data uji..... 62
Kode Program 4.32 Implementasi *preprocessing* pada data latih dan data uji..... 63

Kode Program 4.33 Kode program menghitung jumlah kata setiap kalimat 63
Kode Program 4.34 Kode program pembacaan dataset untuk tokenisasi 64
Kode Program 4.35 Kode program tokenisasi 65
Kode Program 4.36 Kode program untuk melakukan *tuning* 66
Kode Program 4.37 Kode program pengaturan *hyperparameter* 67
Kode Program 4.38 Kode program *custom loss function* 68
Kode Program 4.39 Kode program Bahdanau *attention* 69
Kode Program 4.40 Kode program *encoder* 70
Kode Program 4.41 Kode program *decoder* 71
Kode Program 4.42 Kode program teknik *teacher forcing* 72
Kode Program 4.43 Kode program inisialisasi dan *early stopping* 73
Kode Program 4.44 Kode program *epoch* untuk *training* model 74
Kode Program 4.45 Memuat model yang sudah dilatih 92
Kode Program 4.46 Fungsi untuk melakukan penerjemahan kalimat uji 93
Kode Program 4.47 Implemantasi fungsi penerjemahan dan visualisasi 94
Kode Program 4.48 Evaluasi BLEU 95

DAFTAR LAMPIRAN

Lampiran A Data Tanpa Augmentasi A-1
Lampiran B Data Hasil Augmentasi *Synonym Replacement* B-4
Lampiran C Data Hasil Augmentasi *Stemming* + *Synonym Replacement*..... C-7
Lampiran D Hasil Augmentasi *Random Insert*..... D-10
Lampiran E Hasil Augmentasi *Random Swap*.....E-13
Lampiran F Hasil Augmentasi *Random Deletion* F-16
Lampiran G Data Uji..... G-19
Lampiran H Hasil Evaluasi Model Tanpa Augmentasi..... H-22
Lampiran I Hasil Evaluasi Model *Synonym Replacement* I-27
Lampiran J Hasil Evaluasi *Stemming* + *Synonym Replacement*.....J-32
Lampiran K Hasil Evaluasi *Random Insertion* K-35
Lampiran L Hasil Evaluasi Model *Random Swap* L-40
Lampiran M Hasil Evaluasi Model *Random Delete* M-45

BAB I
PENDAHULUAN

1.1 Latar Belakang

Bahasa adalah sistem yang terdiri dari kata-kata dan aturan tata bahasa yang digunakan untuk membentuk kalimat yang bermakna, baik dalam bentuk lisan maupun tulisan. Bahasa digunakan sebagai alat komunikasi yang memungkinkan manusia menyampaikan informasi, mengekspresikan emosi, dan membangun hubungan sosial (Mailani O, et al., 2022).

Indonesia memiliki bahasa daerah yang beragam, bahasa Dayak Banjur merupakan salah satu dari ratusan bahasa daerah yang ada di Indonesia. Bahasa Dayak Banjur dalam pengucapan atau penuturannya relatif tidak terlalu sulit, serta terdapat beberapa bahasa Dayak lain yang hampir mirip seperti Dayak Sebruang. Misalnya dalam kalimat bahasa Indonesianya “Bagaimana cara mengerjakan ini?”, dalam bahasa Dayak Banjur “Kati cara ngerja e tuk?” dan dalam bahasa Dayak Sebruang “Tipa cara ngerjakan e tuk?”.

Bahasa Dayak Banjur adalah bahasa yang dipakai sebagian kecil masyarakat di Kalimantan Barat, secara khusus di beberapa daerah di Kabupaten Sekadau dan Kabupaten Sintang. Seiring perkembangan zaman jumlah penutur bahasa Banjur ini semakin sedikit, bahkan banyak generasi muda yang tidak lagi menggunakan bahasa Banjur. Sebagai salah satu upaya untuk tetap mempertahankan kelestarian Bahasa Dayak Banjur, dilakukan penelitian terkait model mesin penerjemah jaringan saraf tiruan untuk menganalisis penerjemahan dari bahasa Indonesia ke bahasa Dayak Banjur.

Mesin penerjemah jaringan saraf tiruan merupakan mesin terjemahan yang terinspirasi dari cara kerja otak manusia untuk memproses dan menerjemahkan teks. *Neural Machine Translation* (NMT) adalah pendekatan baru dalam teknologi penerjemahan yang menggabungkan *encoder* dan *decoder*. *Encoder* berfungsi sebagai jaringan saraf berulang yang mengonversi bahasa sumber menjadi vektor-vektor dengan panjang tetap, sedangkan *decoder* adalah jaringan saraf berulang lain yang menghasilkan terjemahan secara menyeluruh (Bahdanau, D. 2014)

Untuk membangun model NMT diperlukan data dalam jumlah yang besar dan bervariasi agar model tidak terlalu hafal data latih yang membuat model jadi kurang bagus untuk data baru atau yang disebut dengan *overfitting*, sehingga dalam hal ini teknik *data augmentatio*n diperlukan. Augmentasi data adalah sebuah teknik yang efektif untuk mencegah *overfitting* dalam model pembelajaran mesin. Dengan meningkatkan variasi pada data latih, augmentasi data membantu model untuk belajar pola yang lebih umum dan menjadi lebih *robust*. (Shorten, C., Khoshgoftaar, T. M., & Furht, B. 2021). Teknik augmentasi yang digunakan dalam penelitian ini, yang pertama *synonym replacement* yaitu mengganti kata-kata dalam kalimat dengan sinonimnya. Yang kedua, *stemming + synonym replacement* yaitu teknik mengganti kata dengan sinonimnya juga, hanya saja sebelum proses *synonym repalcement* kata berimbuhan dalam data kalimat Bahasa Indonesia diubah menjadi bentuk dasarnya yang mana proses ini disebut *stemming*. Teknik augmentasi yang ketiga, *random deletion* yaitu menghapus kata-kata secara acak dari kalimat. Yang keempat ada *random swap* yaitu menukar posisi dua kata acak dalam kalimat. Kemudian yang terakhir adalah *random insertion* yaitu memasukkan kata baru secara acak ke dalam kalimat.

Berdasarkan uraian penjelasan yang telah dipaparkan, maka dalam penelitian ini dilakukan lima teknik augmentasi data teks dan membangun sebuah model *Neural Machine Translation* (NMT) bahasa Indonesia ke bahasa Dayak Banjur dengan penerapan teknik *data augmentation* untuk menambah variasi data latih di dalamnya.

1.2 Perumusan Masalah

Berdasarkan latar belakang yang telah diuraikan, karena dalam membangun model NMT diperlukan data latih dalam jumlah besar serta bervariasi, dengan demikian pada penelitian ini dapat dirumuskan masalah, bahwa bagaimana melakukan augmentasi data teks dan bagaimana menerapkan teknik *data augmentation* dapat meningkatkan kinerja mesin penerjemah jaringan saraf tiruan pada penerjemahan bahasa Indonesia ke bahasa Dayak Banjur.

1.3 Tujuan Penelitian

Adapun tujuan dari penelitian ini adalah untuk menerapkan teknik *data augmentation* pada korpus bahasa Indonesia - Dayak Banjur dan membangun model *Neural Machine Translation* (NMT) berbasis *attention*, dengan menganalisis sejauh mana penggunaan *data augmentation* dapat meningkatkan akurasi hasil terjemahan dalam skenario penerjemahan mesin untuk bahasa dengan sumber daya terbatas.

1.4 Pembatasan Masalah

Untuk memperjelas ruang lingkup masalah dalam penelitian ini, peneliti memiliki beberapa batasan sebagai berikut:

a. Penerjemahan yang dilakukan adalah penerjemahan satu arah, yaitu dari bahasa Indonesia ke bahasa Dayak Banjur.
b. Menggunakan mekanisme *Attention* Bahdanau.
c. Menggunakan arsitektur *Recurrent Neural Network* (RNN).
d. Korpus teks paralel data awal yang digunakan sebanyak 5000 kalimat.
e. Menggunakan Teknik augmentasi, *synonym replacement, stemming + synonym replacement, random swap, random deletion, random insertion*.
f. *Synonym Replacement, stemming + synonym replacement* dan *Random Insertion* dilakukan dengan menggunakan kamus sinonim dari Tesaurus yang tersedia di *repository* GitHub.
g. Dalam pengumpulan data bahasa Indonesia, peneliti menyadari adanya beberapa kata serapan atau istilah dalam bahasa Inggris yang tercampur dalam kalimat karena sudah lazim digunakan dalam percakapan sehari-hari. Penelitian ini tetap memasukkan kata-kata tersebut ke dalam data tanpa dilakukan pembersihan, penyesuaian, atau normalisasi khusus, sehingga kata-kata tersebut diterjemahkan apa adanya oleh model NMT dan tidak menjadi fokus evaluasi kualitas terjemahan.

1.5 Sistematika Penulisan

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.

a. BAB I PENDAHULUAN, adalah bab yang berisi latar belakang penelitian, perumusan masalah, tujuan penelitian, pembatasan masalah, dan sistematika penulisan.
b. BAB II TINJAUAN PUSTAKA, adalah bab yang berisi uraian sistematis tentang hasil-hasil penelitian yang didapat oleh peneliti terdahulu dan landasan teori yang ada hubungannya dengan penelitian yang akan dilakukan.
c. BAB III METODOLOGI PENELITIAN, adalah bab yang berisi tentang data penelitian, perangkat penelitian yang digunakan, langkah-langkah penelitian, pengumpulan data, pembuatan korpus teks paralel, membangun mesin penerjemah jaringan saraf tiruan, implementasi mesin penerjemah jaringan saraf tiruan, pengujian dan evaluasi.
d. BAB IV HASIL DAN ANALISIS, adalah bab yang berisi penjelasan mengenai hasil penelitian mulai dari hasil pembuatan korpus paralel, membangun mesin penerjemah jaringan saraf tiruan, implementasi mesin penerjemah jaringan saraf tiruan, hasil pengujian dan analisis hasil pengujian.
e. BAB V PENUTUP, adalah bab yang berisi kesimpulan dari penelitian yang telah dilakukan dan saran atau rekomendasi untuk perbaikan, pengembangan atau kelengkapan penelitian yang telah dilakukan.