Building of Informatics, Technology and Science (BITS)
Volume 6, No 2, September 2024 Page: 1240−1249
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i2.5369

Optimasi Kombinasi Hyperparameter dan Augmentasi Korpus dalam Neural Machine Translation Bahasa Indonesia ke Bahasa Melayu Bengkulu

Dewi Soyusiawaty

Prodi Informatika, Universitas Ahmad Dahlan, Yogyakarta, Indonesia
Email: [1,*]dewi.soyusiawaty@tif.uad.ac.id
Email Penulis Korespondensi: dewi.soyusiawaty@tif.uad.ac.id
Submitted: **18/06/2024**; Accepted: **01/08/2024**; Published**: 30/09/2024**

Abstrak−Neural Machine Translation (NMT) dengan mekanisme attention telah menjadi pendekatan yang efektif dalam meningkatkan kualitas terjemahan antarbahasa. Namun, penerapan NMT dengan attention untuk bahasa-bahasa daerah atau minoritas masih menghadapi tantangan, terutama dalam konteks Bahasa Melayu Bengkulu, sebuah varian Bahasa Melayu yang digunakan di Provinsi Bengkulu, Indonesia. Penelitian ini bertujuan untuk meningkatkan akurasi terjemahan dari Bahasa Indonesia ke Bahasa Melayu Bengkulu melalui optimasi kombinasi hyperparameter pada model NMT dengan attention. Metode penelitian ini melibatkan eksperimen dengan berbagai kombinasi hyperparameter, seperti batch size, jumlah dataset, epoch dan dropout rate, yang diterapkan pada model NMT dengan attention. Evaluasi dilakukan menggunakan metrik BLEU untuk mengukur kualitas terjemahan. Augmentasi korpus dilakukan untuk mendapatkan korpus yang lebih besar. Hasil eksperimen menunjukkan bahwa peningkatan akurasi terjemahan dapat dicapai dengan memilih kombinasi hyperparameter yang optimal. Penggunaan data set yang lebih besar menghasilkan kinerja yang lebih baik dibandingkan dengan data set yang lebih kecil. Batch size=16 memberikan hasil yang lebih baik daripada batch size=32 dan 64, terutama ketika digunakan dengan jumlah data set yang lebih besar. Selain itu, dropout rate=0.8 cenderung memberikan kinerja yang lebih baik daripada dropout rate=0.2 dan 0.5 Pada nilai epoch, penelitian menunjukkan bahwa peningkatan epoch hingga suatu titik (kira-kira 30 epoch) meningkatkan kinerja model, tetapi peningkatan lebih lanjut cenderung menyebabkan overfitting pada data training. Penelitian ini memberikan kontribusi penting dalam pengembangan terjemahan mesin untuk Bahasa Melayu Bengkulu dan bahasa-bahasa daerah lainnya. Diharapkan hasil penelitian ini dapat menjadi landasan untuk pengembangan lebih lanjut dalam bidang terjemahan mesin untuk bahasa-bahasa minoritas, serta meningkatkan aksesibilitas informasi dalam masyarakat yang beragam bahasa di Indonesia.

Kata Kunci: Attention; Bahasa Indonesia, Bahasa Melayu Bengkulu, Optimasi Hyperparameter; Neural Machine Translation

Abstract−Neural Machine Translation (NMT) with attention mechanism has become an effective approach in improving the quality of cross-language translation. However, the application of NMT with attention to regional or minority languages still faces challenges, especially in the context of Bengkulu Malay Language, a variant of Malay Language used in the Bengkulu Province, Indonesia. This research aims to enhance the translation accuracy from Indonesian to Bengkulu Malay Language through optimization of hyperparameter combinations in NMT models with attention. The research method involves experiments with various hyperparameter combinations, such as batch size, dataset size, and dropout rate, applied to NMT models with attention. Evaluation is conducted using the BLEU metric to measure translation quality. Corpus augmentation is done to obtain a larger corpus.The experimental results indicate that translation accuracy improvement can be achieved by selecting optimal hyperparameter combinations. The use of a larger dataset yields better performance compared to a smaller dataset. A batch size of 16 yields better results than batch sizes of 32 and 64, especially when used with a larger dataset. Additionally, a dropout rate of 0.8 tends to perform better than dropout rates of 0.2 and 0.5. Regarding epoch values, the research shows that increasing epochs up to a certain point (approximately 30 epochs) enhances model performance, but further increases tend to cause overfitting on the training data. This research provides a significant contribution to the development of machine translation for Bengkulu Malay Language and other regional languages. It is hoped that the findings of this research can serve as a foundation for further development in the field of machine translation for minority languages, as well as improving information accessibility in diverse language communities in Indonesia.

Keywords: Attention; Indonesian Language; Bahasa Melayu Bengkulu; Hyperparameter Optimization; Neural Machine Translation

1. PENDAHULUAN

Dari Sabang hingga Merauke, kekayaan bahasa daerah Indonesia dapat ditemukan. Terdapat sekitar 700 bahasa di Indonesia jika kita melihat sebaran kumulatif bahasa daerah per provinsi. Diperkirakan 145 dari ratusan bahasa daerah yang digunakan oleh kurang dari satu juta orang masih mengalami penurunan. Kematian bahasa, atau keadaan di mana suatu bahasa tidak lagi digunakan, berhubungan dengan kepunahan bahasa. Penurunan populasi penutur yang signifikan merupakan salah satu indikator kepunahan bahasa. Indikator lain bahwa suatu bahasa berada di ambang kepunahan adalah ketidakpedulian para penutur muda terhadap penggunaan bahasa daerah. Berdasarkan asumsi yang tersebar luas, generasi muda saat ini belum mahir dalam bahasa asli daerahnya. Kebanyakan hanya melakukan kontrol pasif. Mereka tidak bisa berbicara dalam bahasa ibu mereka dengan baik, namun mereka memahaminya. Jika hal ini terus berlanjut, maka tidak menutup kemungkinan bahwa dalam beberapa tahun ke depan, semakin banyak bahasa daerah yang secara bertahap akan hilang karena kemerosotan bahasa.

Meskipun upaya untuk melestarikan bahasa tidak bersifat “nyata” dalam arti keuntungan materi atau finansial, namun upaya tersebut mengalami kesulitan untuk memberikan dampak yang berarti dalam melestarikan kekayaan batin negara. Kepunahan bahasa tidak hanya melibatkan hilangnya sintaksis dan kosa kata—kepunahan ini juga melibatkan penghancuran warisan budaya suatu negara yang tak ternilai harganya. Memang, menurut peringatan

Building of Informatics, Technology and Science (BITS)
Volume 6, No 2, September 2024 Page: 1240−1249
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i2.5369

Unesco, hilangnya suatu bahasa berarti dunia kehilangan warisan yang tak ternilai harganya, serta banyak puisi, dongeng, dan informasi yang diwariskan selama berabad-abad. [1]

Dalam era globalisasi yang dipenuhi dengan arus informasi, pentingnya menjembatani komunikasi antarbahasa semakin meningkat. Teknologi Neural Machine Translation (NMT) telah menjadi landasan utama dalam memfasilitasi terjemahan lintas bahasa dengan tingkat akurasi yang tinggi. Namun, penerjemahan dari Bahasa Indonesia ke bahasa daerah lain di Indonesia masih menjadi tantangan yang signifikan. Pada saat ini, masalah utama yang dihadapi dalam terjemahan Bahasa Indonesia ke bahasa daerah lain adalah kurangnya data korpus paralel yang memadai. Korpus paralel, yang merupakan kumpulan teks yang diterjemahkan secara bersamaan ke dalam dua atau lebih bahasa, sangat penting dalam melatih model NMT. Namun, mayoritas bahasa daerah di Indonesia memiliki korpus paralel yang terbatas atau bahkan tidak memiliki sama sekali, menyebabkan kualitas terjemahan yang kurang memuaskan. Data menunjukkan bahwa dari lebih dari 700 bahasa daerah yang ada di Indonesia, hanya sebagian kecil yang memiliki korpus paralel yang cukup untuk melatih model terjemahan mesin dengan baik. Bahkan, beberapa bahasa daerah memiliki kurang dari 100 ribu pasangan kalimat terjemahan, yang merupakan jumlah yang sangat kecil untuk menghasilkan terjemahan yang akurat. Pentingnya membuat terjemahan dari Bahasa Indonesia ke bahasa daerah sangatlah relevan dengan keberagaman budaya dan kekayaan linguistik Indonesia. Terjemahan ini memainkan peran penting dalam meningkatkan aksesibilitas informasi dan memperkuat identitas budaya masyarakat setempat. [1]

Untuk memahami landasan penelitian yang lebih luas, perlu dilihat beberapa kajian terkait yang telah dilakukan dalam domain ini. Penelitian [2], [3], [4], [5], [6], [7] dan masih banyak lainnya mengembangkan NMT Attention untuk bahasa daerah di Indonesia seperti Lampung, Bangka, Pontianak, Kawi, Sunda. Kajian-kajian ini memberikan wawasan yang berharga tentang penggunaan NMT dalam konteks bahasa daerah di Indonesia, tetapi tidak secara langsung terkait dengan Bahasa Melayu Bengkulu. Penelitian tidak secara khusus memperhatikan optimasi kombinasi hyperparameter untuk meningkatkan akurasi terjemahan.

Dari uraian tersebut, dapat disimpulkan bahwa meskipun telah ada beberapa penelitian terkait penggunaan NMT dalam terjemahan antarbahasa di Indonesia, penelitian tentang optimasi kombinasi hyperparameter dalam NMT dengan attention untuk terjemahan Bahasa Indonesia ke Bahasa Melayu Bengkulu masih merupakan hal yang baru. Dengan demikian, penelitian ini bertujuan untuk mengoptimalkan kombinasi hyperparameter dalam model NMT untuk meningkatkan akurasi terjemahan dari Bahasa Indonesia ke Bahasa Melayu Bengkulu. Diharapkan bahwa melalui penelitian ini, masalah rendahnya kualitas terjemahan dapat diatasi, sehingga masyarakat yang menggunakan Bahasa Melayu Bengkulu sebagai bahasa ibu mereka dapat dengan mudah mengakses dan memahami informasi yang tersedia dalam Bahasa Indonesia. Selain itu, peningkatan kualitas terjemahan juga diharapkan dapat meningkatkan kebanggaan akan bahasa daerah dan memperkuat identitas budaya lokal di Indonesia. Oleh karena itu, penelitian ini memiliki nilai penting dalam mengisi kesenjangan pengetahuan tersebut dan diharapkan dapat memberikan kontribusi yang signifikan dalam pengembangan teknologi terjemahan mesin untuk bahasa-bahasa daerah di Indonesia.

2. METODOLOGI PENELITIAN

2.1 Pengumpulan Data

Korpus teks paralel yang terdiri dari pasangan kalimat Bahasa Indonesia dan Bahasa Melayu Bengkulu telah ada pada penelitian sebelumnya. Korpus ini akan digunakan sebagai dataset pelatihan untuk melatih model Neural Machine Translation (NMT) dengan attention. Untuk meningkatkan keberagaman dan jumlah data dilakukan augmentasi data. Augmentasi data dilakukan dengan cara menambahkan variasi sintetis ke dalam korpus, misalnya dengan mengganti sinonim, menambahkan frasa sinonim, atau melakukan penggabungan kalimat.

2.2 Prapemrosesan Data

Prapemrosesan data untuk membersihkan dan menormalkan teks. Langkah-langkah prapemrosesan ini mencakup tokenisasi, penghapusan tanda baca, normalisasi, dan pemisahan kata-kata.

2.3 Pembagian Data

Data dibagi menjadi tiga bagian: data pelatihan, data validasi, dan data pengujian. Data pelatihan digunakan untuk melatih model, data validasi digunakan untuk mengevaluasi kinerja model selama proses pelatihan, dan data pengujian digunakan untuk menguji kinerja model setelah pelatihan selesai.

2.4 Pengembangan Model NMT dengan Attention

Implementasi model NMT dengan attention menggunakan framework atau library deep learning yang tersedia, seperti TensorFlow atau PyTorch. Model ini terdiri dari encoder dan decoder, dengan mekanisme attention untuk memungkinkan fokus pada bagian yang relevan dari kalimat sumber saat melakukan terjemahan.

2.5 Pelatihan Model

Setelah pengembangan model selesai, dilanjutkan dengan melatih model menggunakan data pelatihan yang telah diproses sebelumnya.

Building of Informatics, Technology and Science (BITS)
Volume 6, No 2, September 2024 Page: 1240−1249
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i2.5369

2.6 Validasi dan Evaluasi

Selama proses pelatihan, mengevaluasi kinerja model menggunakan data validasi untuk menghindari overfitting. Setelah pelatihan selesai, evaluasi akhir menggunakan data pengujian untuk mengukur akurasi terjemahan dari Bahasa Indonesia ke Bahasa Melayu Bengkulu. Beberapa metrik dalam pengujian NMT, yaitu : [8] [9] [10]

a. Accuracy (Akurasi)
   Akurasi mengukur seberapa sering model berhasil menerjemahkan kalimat dengan benar dari bahasa sumber ke bahasa target. Akurasi dihitung dengan membagi jumlah prediksi yang benar dengan jumlah total prediksi.
b. Loss (Kehilangan)
   Mengukur seberapa baik atau buruk performa model dalam mempelajari representasi data pada setiap iterasi pembelajaran. Loss umumnya dihitung dengan menggunakan fungsi loss seperti cross-entropy loss. Tujuan dari pelatihan model adalah untuk meminimalkan nilai loss ini, yang berarti meningkatkan kemampuan model dalam melakukan prediksi yang akurat.
c. Val_loss (Kehilangan validasi)
   Val_loss adalah nilai loss yang dihitung pada set data validasi. Data validasi digunakan untuk mengukur kinerja model yang tidak terlihat selama proses pelatihan. Val_loss memberikan informasi tentang seberapa baik model dapat melakukan generalisasi pada data yang belum pernah dilihat sebelumnya. Penurunan val_loss yang konsisten menunjukkan bahwa model mungkin tidak mengalami overfitting dan dapat bekerja dengan baik pada data baru.
d. Val_accuracy (Akurasi validasi)
   Val_accuracy adalah akurasi yang dihitung pada set data validasi. Sama seperti val_loss, val_accuracy memberikan informasi tentang seberapa baik model dapat melakukan generalisasi pada data baru yang tidak terlihat selama proses pelatihan. Tujuan utama adalah untuk mencapai tingkat akurasi yang tinggi baik pada data pelatihan maupun data validasi, menunjukkan bahwa model mampu melakukan terjemahan yang akurat pada berbagai jenis input.

2.7 Optimasi Hyperparameter

Tahap terakhir adalah optimasi kombinasi hyperparameter. Eksperimen dengan berbagai kombinasi hyperparameter, seperti batch size, epoch dan tingkat dropout. Untuk mencari kombinasi yang menghasilkan kinerja terbaik dalam hal akurasi terjemahan. Saat ini, metode pemodelan sering kali melibatkan sejumlah parameter yang diperlukan dalam proses pelatihan data. Berbeda dengan parameter lain yang mungkin berubah seiring waktu atau dalam proses pelatihan data, hyperparameter merupakan parameter yang tetap dan tidak dapat diubah selama proses pelatihan. [11] Hyperparameter ini memiliki pengaruh yang signifikan terhadap keandalan dan performa suatu teknik pemodelan. Hyperparameter merupakan pengaturan yang didefinisikan untuk mengontrol suatu metode pemodelan guna meningkatkan kinerja prediksi. [12] [13]

Tabel 1. Hyperparameter Neural Network yang digunakan

Parameter	Nilai
Batch Size	16, 32, 64
Epoch	10,20,30,40,50
Drop out Rate	0.2, 0.5, 0.8

Kombinasi dari hyperparameter dan ukuran arsitektur Neural Network, antara lain :

a. “Batch Size adalah pembagian data menjadi beberapa bagian sehingga akan memudahkan beban pemrosesan memori. Nilai yang digunakan antara lain adalah 16, 32, 64 dengan mempertimbangkan angka kelipatan 2 yang dapat memudahkan proses pembagian dataset dalam pelatihan model. Batch size yang terlalu kecil mungkin mengakibatkan pelatihan lambat, sementara batch size yang terlalu besar mungkin memerlukan lebih banyak memori GPU.
b. Epoch adalah jumlah iterasi yang diperlukan untuk melatih model. Nilai yang digunakan antara lain adalah 10,20,30,40 dan 50 dengan mempertimbangkan waktu dari proses pelatihan yang lebih cepat. Variasi jumlah epoch untuk menemukan keseimbangan antara underfitting dan overfitting.
c. Dropout Rate adalah regularisasi jaringan syaraf saat neuron dipilih secara acak dan tidak dipakai saat pelatihan. Menggunakan nilai masing-masing adalah 0.2, 0.5, dan 0.8 dengan mempertimbangkan kemerataan nilai dari nol sampai satu. Proporsi unit yang diabaikan selama pelatihan untuk mencegah overfitting. Dropout dapat membantu meningkatkan generalisasi model.” [14]

3. HASIL DAN PEMBAHASAN

3.1 Pengumpulan Data

Berikut adalah beberapa contoh augmentasi data yang dapat diterapkan pada korpus terjemahan Bahasa Indonesia ke Bahasa Melayu Bengkulu:

Building of Informatics, Technology and Science (BITS)
Volume 6, No 2, September 2024 Page: 1240−1249
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i2.5369

a. Penggantian Kata Sinonim
Mengganti beberapa kata dalam kalimat dengan sinonimnya. Misalnya, mengganti kata "makan" dengan "menyantap" atau "memakan". Contoh pada tabel 2 menyatakan kalimat asli dan kalimat hasil augmentasi.

Tabel 2. Penggantian Kata Sinonim

Indonesia		Melayu Bengkulu
Kalimat Asli	Saya makan nasi goreng	Ambo makan nasi guring
Augmentasi	Saya menyantap nasi goreng	Ambo makan nasi guring

b. Penggantian Frasa Sinonim
Mengganti frasa dalam kalimat dengan frasa sinonimnya. Misalnya, mengganti frasa "belajar dengan giat" dengan "belajar dengan rajin". Contoh pada tabel 3 menyatakan frasa dari sinonim.

Tabel 3. Penggantian Frasa Sinonim

Indonesia		Melayu Bengkulu
Kalimat asli	Anak-anak belajar dengan giat di sekolah	Anak-anak tu rajin nian belajar di sekolah.
Augmentasi	Anak-anak belajar dengan rajin di sekolah	Anak-anak tu rajin nian belajar di sekolah.

c. Penambahan Frasa Deskriptif
Menambahkan frasa deskriptif atau keterangan tambahan pada kalimat untuk memberikan informasi lebih lanjut. Contoh pada tabel 4 menyatakan penambahan frasa deskriptif.

Tabel 4. Penambahan Frasa Deskriptif

Indonesia		Melayu Bengkulu
Kalimat asli	Dia pergi ke toko	Die pergi ke toko
Augmentasi	Dia pergi ke toko di dekat rumahnya	Die pergi ke toko

d. Pemisahan Kalimat
Membagi kalimat menjadi dua kalimat terpisah untuk memberikan klarifikasi atau penekanan tambahan. Contoh pada tabel 5 menyatakan pemisahan kalimat.

Tabel 5. Pemisahan Kalimat

Indonesia	Melayu Bengkulu
Kalimat asli	Hari ini cuacanya sangat panas, kami pergi ke pantai
Augmentasi	Hari ini cuacanya sangat panas. Kami pergi ke pantai

e. Penggantian Kata dengan Antonim
Mengganti beberapa kata dalam kalimat dengan antonimnya untuk memberikan kontras. Contoh pada tabel 6 menyatakan penggantian kata dengan antonim.

Tabel 6. Penggantian Kata dengan Antonim

Indonesia		Melayu Bengkulu
Kalimat asli	Dia senang menonton film horror	Die suko nonton film horor
Augmentasi	Dia sedih menonton film horror	Die sedih nonton film horor

3.2 PreProcessing Data

Beberapa tahap dalam preprocessing data, yaitu :

a. Pembersihan Data
Misalkan kalimat dalam Bahasa Indonesia: "Saya suka makan nasi goreng di restoran kecil itu!" Kalimat dibersihkan dari tanda baca yang tidak diperlukan dan karakter khusus, menjadi: "Saya suka makan nasi goreng di restoran kecil itu".

b. Tokenisasi
Kalimat tersebut kemudian dipecah menjadi token atau kata-kata individual: ["Saya", "suka", "makan", "nasi", "goreng", "di", "restoran", "kecil", "itu"].

c. Normalisasi Teks
Kata-kata yang memiliki variasi ejaan atau bentuk yang berbeda disesuaikan menjadi bentuk standar atau normal. Misalnya, kata "goreng" dalam Bahasa Melayu Bengkulu mungkin dieja sebagai "gureng". Normalisasi ini akan memastikan konsistensi antara kedua bahasa.

Building of Informatics, Technology and Science (BITS)
Volume 6, No 2, September 2024 Page: 1240−1249
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i2.5369

d. Pemisahan Kalimat
Proses yang sama akan dilakukan pada kalimat yang sesuai dalam Bahasa Melayu Bengkulu, untuk memisahkan kalimat tersebut menjadi token-token dan membersihkannya.
e. Penyamaan Panjang Kalimat
Setelah itu, panjang kedua kalimat akan disamakan dengan menambahkan token padding jika diperlukan. Misalnya, jika kalimat Bahasa Melayu Bengkulu lebih pendek dari kalimat Bahasa Indonesia, token padding akan ditambahkan ke akhir kalimat untuk menyamakan panjangnya.
f. Penghapusan Kata Rendah Frekuensi
Kata-kata dengan frekuensi yang sangat rendah dalam korpus mungkin dihapus untuk mengurangi kompleksitas model.
g. Tokenisasi Khusus
Setelah itu, token-token khusus, seperti token awal dan akhir kalimat serta token padding, ditambahkan untuk memfasilitasi proses attention pada model NMT.
h. Pembuatan Pasangan Data
Kalimat Bahasa Indonesia dan Bahasa Melayu Bengkulu yang telah diproses akan dibuat menjadi pasangan data input-output yang sesuai untuk pelatihan model NMT dengan attention.

3.3 Pengembangan Model NMT dengan Attention

Adapun alur tahap implementasi :

a. Encoder : tahap awal encoder berfungsi sebagai proses sumber bahasa dalam setiap kata diproses kebentuk angka integer untuk menghasilkan matriks atau gambaran susunan setiap kata. Proses tersebut diperlukan informasi dalam pengaksesan tentang per-kata dalam input urutannya. [15]
b. Attention Mekanisme : tahap penting dalam mekanisme terjemahan dimana attention berperan sebagai pembelajaran input urutan untuk ditentukan nilai skor perhitungannya. Attention dalam Bahdanau biasanya disebut concat atau additive.
c. Decoder : tahap dalam memprediksi setiap kata perlangkah waktu (time_step) berdasarkan proses implementasi dalam mekanisme attention dalam menghasilkan keluaran output. Proses decoder memiliki kesamaan dengan encoder seperti akses layer (2 layer). Untuk menghasilkan keluaran output decoder membutuhkan layer attention. Layer attention merupakan penyedia informasi kepada decoder dalam memfokuskan berdasarkan vektor attention berupa weight attention semua encoder state hidden dalam setiap langkah waktu. Dalam setiap langkah waktu proses decoder berperan untuk input selanjutnya pada vektor konteks.
d. Function Loss : tahap proses pelatihan model dengan teknik optimasi,tujuannya untuk meminimalis nilai loss function. Loss function digunakan untuk mengetahui estimasi parameter atau peristiwa yang dimaksud seperti beberapa fungsi dari perbedaan nilai yang diperkirakan dari sebuah data. Loss function semakin baik pada model apabila semakin rendah nilai loss functionnya (nilai loss function terbaik = 0). Tahap loss function dilakukan pertama dalam pelatihan model pada proses pelengkap model siap latih. Karena dataset model yang dilatih sebelumnya berbentuk bilangan angka integer akan dipresentasikan menjadi tensor data (pemakaian modul sparse categorical cross entropy).
e. Pelatihan Model NMT : tahap pelatihan ini untuk data latih pada model dilatih dengan teknik penyetelan batasan waktu latih tertentu dalam mendapatkan hasil prediksi yang terbaik. Proses setiap langkah waktu pada output encoder yang memiliki isi hidden state akan dikembalikan pada latih tunggal, karena tensor input yang dilewati dalam kalimat masukan (input) milik encoder. Dan dilakukan teknik teacher forcing serta jumlah epoch yang ditentukan. Bagian terakhir dari alur pelatihan adalah perulangan pelatihan dengan jumlah epoch. Bagian ini sebagai hyperparameter yang dilakukan adalah menjalankan beberapa epoch dan pada setiap epoch yang sudah dikalkulasikan dimulai dari epoch 10, 20, 30 sampai hasil akurasi sudah mulai menurun 2 langkah dari nilai akurasi jumlah epoch sebelumnya. Pelatihan jumlah epoch dengan mengulangi dataset dan memanggil fungsi tahap pelatihan train_step pada setiap kumpulan dataset. Beberapa pernyataan seperti if untuk mencatat statistik dan menyimpan data pelatihan. Dalam tahap pelatihan ini akan memakan waktu yang lama jika melatih menggunakan CPU komputer lokal, sedangkan dengan menggunakan google colaboratory notebook akan mendapatkan dukungan GPU untuk mempercepat pelatihan. [9]
f. Prediksi NMT Model : Prediksi model hasil terjemahan didapatkan setelah penyelesaian pada tahap pelatihan model. Penulis dapat memperoleh terjemahan asal yang tidak diketahui tersebut adalah inferensi. Inferensi menggunakan kata-kata yang diprediksi oleh model, sedangkan pada pelatihan yang biasanya memasukan kata-kata target yang benar sebagai input. Encoder output shape pada inference waktu diakses dari kalimat sumber.

3.4 Hyperparameter

Simulasi Hyperparameter Neural Network dilakukan dengan melakukan banyak kombinasi.

Building of Informatics, Technology and Science (BITS)
Volume 6, No 2, September 2024 Page: 1240−1249
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i2.5369

3.4.1 Skenario 1000 Record

Untuk skenario pertama dilakukan dengan dataset berjumlah 1000 record. Tabel 7 menyajikan hasil pengujian kombinasi hyperparameter yang dilakukan sebanyak 9 tahap. Setiap tahap terdiri atas 5 kali pengujian, dimulai dari batch size 16 dan dropout rate 0.2 dengan masing-masing epoch 10, 20, 30, 40 dan 50. Pada penelitian ini, dilakukan simulasi dengan data yang sama seperti sebelumnya dengan menggunakan model terbaik dan terakhir sebagai hasil dari subbab sebelumnya.

Tabel 7. Hasil Pengujian Kombinasi Hyperparameter dengan Dataset 1000 Record

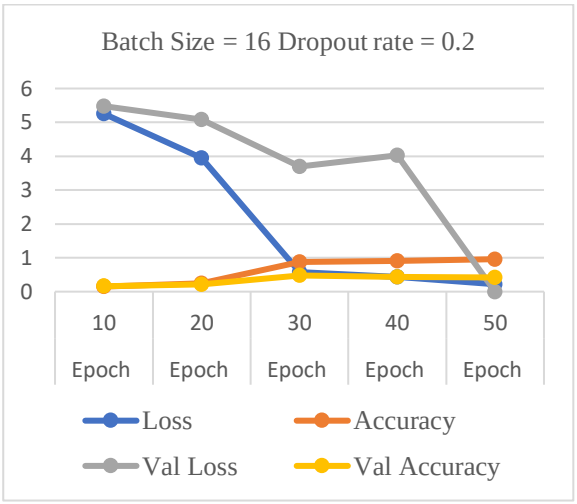
Tahap	Ke-	Batch Size	Dropout rate	Epoch	Loss	Accuracy	Val Loss	Val Accuracy	Ket
	1	16	0.2	10	5.25366	0.14974	5.47790	0.15942	
	2			20	3.95569	0.24978	5.07868	0.21417	
I	3			30	0.58049	0.87939	3.69210	0.47746	
	4			40	0.42724	0.90103	4.02194	0.43317	
	5			50	0.21897	0.95885	4.52492	0.41707	Stop di epoch=33
	6	32	0.2	10	5,54153	0,08462	5,74944	0,08934	
	7			20	4,88267	0,18626	5,30775	0,19163	
II	8			30	4,86345	0,13357	5,30775	0.11987	
	9			40	4.34856	0.13784	5.73345	0.11983	
	10			50	3,48133	0.22763	5.41189	0.17342	
	11	64	0.2	10	5.61360	0.08872	5.76847	0.09018	
	12			20	5.47089	0.09897	5.76531	0.09897	
III	13			30	5.34698	0.10618	5.75632	0.10618	Stop di epoch=26
	14			40	4.95492	0.16453	5.58278	0.16453	Stop di epoch=39
	15			50	2.84891	0.45466	4.12400	0.45466	
	16	16	0.5	10	4.79760	0.16961	5.41108	0.16425	
	17			20	3.92453	0.22261	5.46808	0.20209	
IV	18			30	2.61875	0.30037	5.55827	0.18519	
	19			40	0.13061	0.97978	2.95378	0.59420	Stop di epoch=28
	20			50	0.07252	0.98789	3.05084	0.61353	Stop di epoch=34
	21	32	0.5	10	5.46488	0.11741	5.66491	0.12077	
	22			20	3.50571	0.38821	4.28397	0.36473	
V	23			30	3.81634	0.22403	5.22490	0.18760	Stop di epoch 25
	24			40	1.58173	0.66364	3.58478	0.45572	
	25			50	0.92630	0.76136	4.21416	0.39694	Stop di epoch=39
	26	64	0.5	10	5.63755	0.07162	5.78597	0.07246	
	27			20	5.44450	0.11616	5.67317	0.12882	
VI	28			30	5.04085	0.16007	5.49325	0.15539	Stop di epoch=26
	29			40	3.93495	0.19018	5.53037	0.16989	Stop di epoch=34
	30			50	4.97048	0.12890	5.70004	0.12399	Stop di epoch=29
	31	16	0.8	10	4.94760	0.18662	5.30096	0.18680	
	32			20	1.49576	0.69838	3.23477	0.51208	
VII	33			30	2.13565	0.43337	5.30050	0.23671	
	34			40	0.07542	0.99065	2.87183	0.62319	Stop di epoch=31
	35			50	0.06272	0.99421	3.03551	0.60628	Stop di epoch=35
	36	32	0.8	10	5.49503	0.12168	5.65011	0.13205	
	37			20	5.31667	0.10440	5.77330	0.09823	Stop di epoch=16
VIII	38			30	4.07473	0.22412	5.25582	0.19404	
	39			40	4.06794	0.19535	5.81957	0.16023	Stop di epoch=39
	40			50	0.48553	0.83262	3.81936	0.16023	Stop di epoch=33
	41	64	0.8	10	5.62065	0.08151	5.79672	0.07407	
	42			20	5.39987	0.12827	5.62230	0.13285	
IX	43			30	5.34562	0.10440	5.73922	0.10145	
	44			40	5.01856	0.15642	5.43143	0.55391	Stop di epoch=36
	45			50	4.39022	0.19704	5.33254	0.17472	Stop di epoch=41

Pada tabel 7 dapat dilihat secara umum untuk tiap tahap pengujian yaitu :

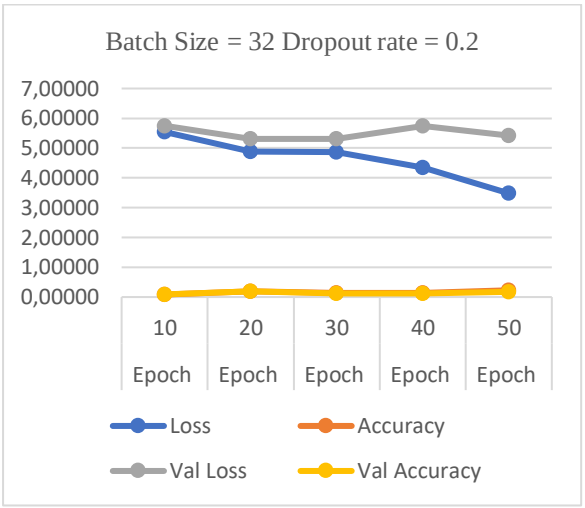
a. Nilai loss selalu turun
b. Nilai akurasi selalu bertambah
c. Nilai validation loss selalu turun
d. Nilai validation accuracy selalu bertambah

Building of Informatics, Technology and Science (BITS)
Volume 6, No 2, September 2024 Page: 1240−1249
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i2.5369

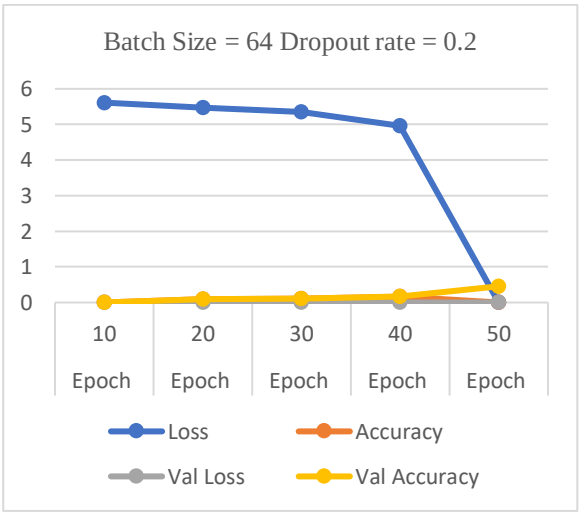
Namun terdapat di beberapa pengujian yang mana proses berhenti sebelum mencapai jumlah yang ditentukan seperti pada tahap I pengujian ke 5 dengan epoch = 50, proses berhenti di epoch = 33 dengan nilai akurasi 95%. Hal ini dapat terjadi untuk menghindari overfitting.



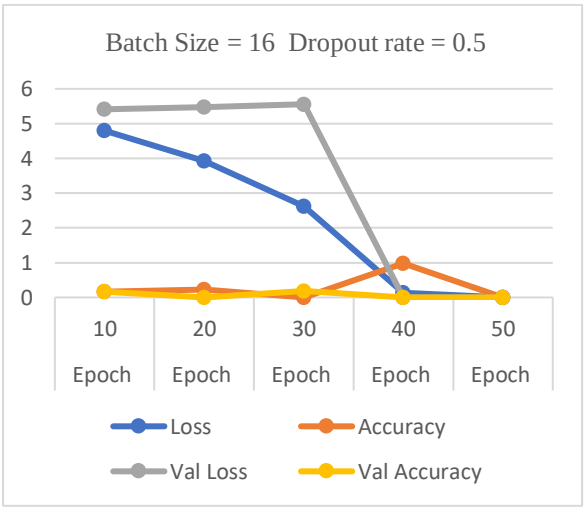
(a)



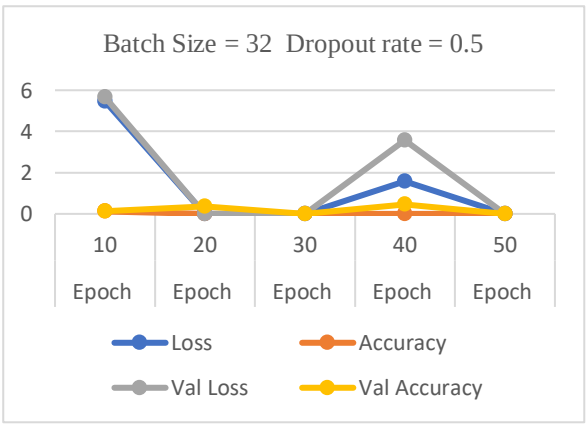
(b)



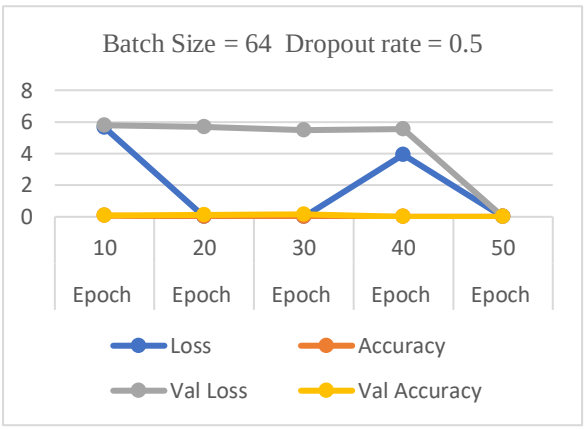
(c)



(d)

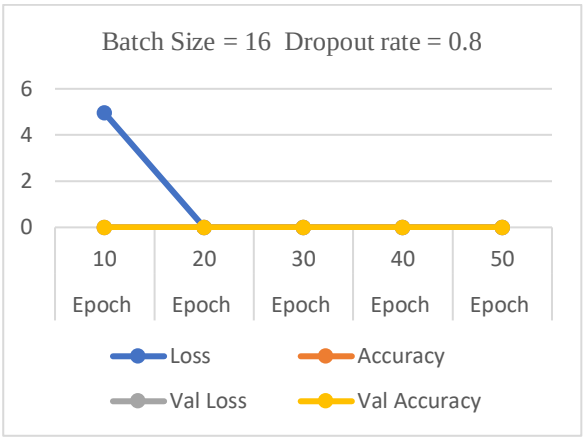


(e)

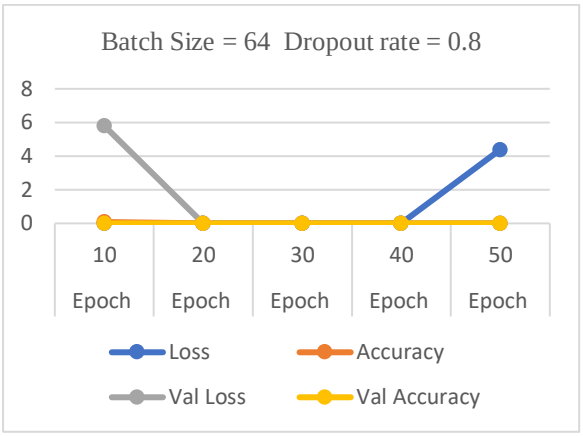


(f)

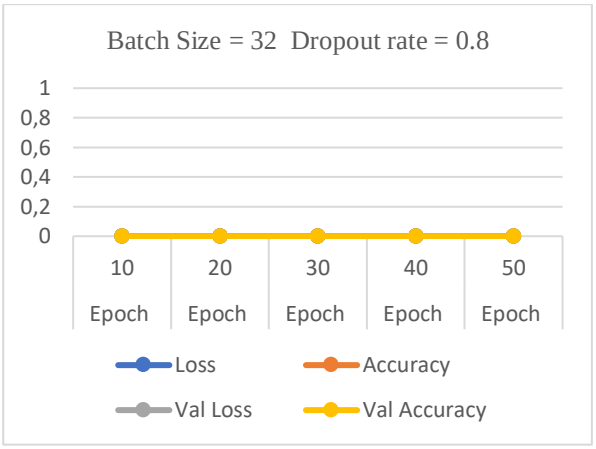
Building of Informatics, Technology and Science (BITS)
Volume 6, No 2, September 2024 Page: 1240−1249
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i2.5369



(g)



(h)



(i)

Gambar 1. Grafik Pengujian Kombinasi Hyperparameter (a) Grafik Batch Size = 16 dan Dropout Rate 0.2 (b) Grafik Batch Size = 32 dan Dropout Rate 0.2 (c) Grafik Batch Size = 64 dan Dropout Rate 0.2 (d) Grafik Batch Size = 16 dan Dropout Rate 0.5 (e) Grafik Batch Size = 32 dan Dropout Rate 0.5 (f) Grafik Batch Size = 64 dan Dropout Rate 0.5 (g) Grafik Batch Size = 16 dan Dropout Rate 0.8 (h) Grafik Batch Size = 32 dan Dropout Rate 0.8 (i) Grafik Batch Size = 64 dan Dropout Rate 0.8

Gambar 1 menyajikan hasil pengujian dalam bentuk grafik untuk dapat melihat trend nilai dari tiap metrik per pengujian. Pada tahap VIII, pengujian ke 40 dengan epoch = 40 dan batch size = 32 dropout rate = 0.8, proses berhenti di epoch = 39 sehingga mengakibatkan nilai akurasi yang turun dari proses sebelumnya yaitu dari 22% ke 19%. Pada pengujian lanjutan dengan epoch = 40, walaupun proses berhenti lagi di epoch =33, namun sistem sudah menghasilkan akurasi yang meningkat jauh dari proses sebelumnya yaitu dari 19% ke 83%.

Tabel 8. Nilai Terbaik dari Setiap Tahapan dengan Dataset 1000 Record

Tahap	Pengujian ke-	Batch Size	Dropout rate	Epoch	Loss	Accuracy	Val Loss	Val Accuracy
I	5	16	0.2	50	0.21897	0.95885	4.52492	0.41707
II	10	32	0.2	50	3.48133	0.22763	5.41189	0.17342
III	13	64	0.2	50	2.84891	0.45466	4.12400	0.45466
IV	20	16	0.5	50	0.07252	0.98789	3.05084	0.60353
V	25	32	0.5	50	0.92630	0.76136	4.21416	0.39694
VI	29	64	0.5	40	3.93495	0.19018	5.53037	0.16989
VII	35	16	0.8	50	0.06272	0.99421	3.03551	0.60628
VIII	40	32	0.8	50	0.48553	0.83262	3.81936	0.16023
IX	45	64	0.8	50	4.39022	0.19704	5.33254	0.17472

Dari tabel 8 dapat diambil beberapa analisis hasil, yaitu :

Building of Informatics, Technology and Science (BITS)
Volume 6, No 2, September 2024 Page: 1240−1249
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i2.5369

a. Pada pengujian tahap I, IV dan VII dengan batch size = 16 dengan berapapun nilai dropout rate didapat nilai akurasi di atas 95%. Akurasi tertinggi dihasilkan pada dropout rate yang terbesar yaitu 0,8. Dapat disimpulkan bahwa kombinasi batch size yang rendah dan dropout rate terbesar akan menghasilkan akurasi terbaik.
b. Pada pengujian tahap VII juga didapatkan nilai loss terkecil, validation loss terkecil dan juga validation accuracy terbesar. Dapat disimpulkan bahwa semakin kecil batch size dan semakin besar dropout rate akan meminimalkan nilai loss dan validation loss serta menaikkan nilai validation accuracy.

3.4.2 Skenario 3000 Record

Pada tabel 9 menyajikan hasil pengujian kombinasi hyperparameter dengan dataset 3000 record yang didapatkan dari augmentasi data. Nilai parameter yang digunakan sama halnya dengan pengujian pada dataset 1000 record, namun yang dibedakan di sini hanya pada nilai epoch yang hanya menggunakan epoch = 30 dan epoch = 50. Hal ini dilakukan untuk efisiensi waktu penelitian, dikarenakan telah dapat disimpulkan pada pengujian dengan dataset 1000 record, nilai terbaik didapat dengan jumlah epoch di atas 30.

Tabel 9. Nilai Terbaik dari Setiap Tahapan dengan Dataset 3000 Record

Tahap	Ke-	Batch Size	Dropout rate	Epoch	Loss	Accuracy	Val Loss	Val Accuracy	Ket
I	1	16	0.2	30	1.01904	0.99675	2.37282	0.6608	Stop di epoch=24
	2			50	0.00671	0.99823	2.28802	0.73259	Stop di epoch=25
II	3	16	0.5	30	0.02556	0.99528	2.28827	0.70992	Stop di epoch=19
	4			50	0.00503	0.99917	2.35766	0.72521	Stop di epoch=26
III	5	16	0.8	30	0.02890	0.9940	2.09385	0.71941	Stop di epoch=20
	6			50	0.00418	0.99941	2.18651	0.73576	Stop di epoch=27
IV	7	32	0.2	30	0.06231	0.99321	2.15251	0.70359	-
	8			50	0.03418	0.99168	2.24935	0.71730	Stop di epoch=20
V	9	32	0.5	30	0.10067	0.98430	2.20266	0.70359	Stop di epoch=29
	10			50	0.01511	0.99835	2.43854	0.70517	Stop di epoch=26
VI	11	32	0.8	30	0.30560	0.94229	3.52985	0.49895	Stop di epoch=24
	12			50	0.01702	0.99717	2.20394	0.73312	Stop di epoch=30
VII	13	64	0.2	30	0.34048	0.94949	2.38291	0.66825	Stop di epoch=29
	14			50	0.07515	0.99091	2.50477	0.66456	Stop di epoch=28
VIII	15	64	0.5	30	0.33663	0.94795	2.38320	0.66350	Stop di epoch=24
	16			50	0.08903	0.98861	2.41640	0.68460	Stop di epoch=23
IX	17	64	0.8	30	5.34562	0.10440	5.73922	0.10145	Stop di epoch=23
	19			50	4.39022	0.19704	5.33254	0.17472	Stop di epoch=23

Akurasi tertinggi didapat pada tahap III dengan batch size = 16 dan dropout rate = 0.8 dan pada epoch = 50 sebesar 99,941%. Loss terkecil, validation loss terkecil dan validation accuracy terkecil juga dihasilkan pada tahap ini.

4. KESIMPULAN

Optimasi kombinasi hyperparameter dan augmentasi korpus memiliki dampak yang signifikan dalam meningkatkan kinerja Neural Machine Translation (NMT) dari Bahasa Indonesia ke Bahasa Melayu Bengkulu. Dengan mengkombinasikan nilai epoch, batch size, dan dropout rate pada jumlah data set yang berbeda (1000 dan 3000), penelitian ini menemukan pola yang menarik. Hasil penelitian menunjukkan bahwa penggunaan data set yang lebih besar (3000) cenderung menghasilkan kinerja yang lebih baik dibandingkan dengan data set yang lebih kecil (1000). Batch size 16 memberikan hasil yang lebih baik daripada batch size 32 dan 64, terutama ketika digunakan dengan jumlah data set yang lebih besar. Selain itu, dropout rate 0.8 cenderung memberikan kinerja yang lebih baik daripada dropout rate 0.2 dan 0.5. Pada nilai epoch, penelitian menunjukkan bahwa peningkatan epoch hingga suatu titik (kira-kira 30 epoch) meningkatkan kinerja model, tetapi peningkatan lebih lanjut cenderung menyebabkan overfitting pada data training. Kesimpulannya, untuk meningkatkan kinerja NMT dari Bahasa Indonesia ke Bahasa Melayu Bengkulu, kombinasi hyperparameter yang optimal adalah dengan menggunakan data set yang lebih besar (3000), batch size 16, dropout rate 0.8, dan jumlah epoch yang moderat (sekitar 30). Penelitian ini belum detail memeriksa kualitas dataset dan juga perlu dibandingkan dengan dataset pada bahasa lain untuk melihat seberapa pengaruh kualitas dataset terhadap output yang dihasilkan.

REFERENCES

[1] H. D. Ismadi, “Kebijakan Pelindungan Bahasa Daerah dalam Perubahan Kebudayaan Indonesia,” 2022.

Building of Informatics, Technology and Science (BITS)
Volume 6, No 2, September 2024 Page: 1240−1249
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i2.5369

[2] Zaenal Abidin, Adi Sucipto, Arief Budiman, "PENERJEMAHAN KALIMAT BAHASA LAMPUNG-INDONESIA dengan NMT Attention," 2018.
[3] Yustiana Fauziyah, Ridwan Ilyas, Fatan Kasyidi, "Mesin Penterjemah Bahasa Indonesia - Bahasa Sunda menggunakan RNN," *Teknoinfo,* 2022.
[4] Lo Bun San, Herry Sujaini, Tursina, "Uji Nilai Akurasi pada NMT Bahasa Indonesia ke Bahasa Tiociu Pontianak dengan Mekanisme Attention Bahdanau," *JEPIN,* 2023.
[5] Fadel Razsiah, Ahmat Josi, Sari Mubaroh, "Aplikasi Penerjemah Bahasa Bangka Ke Bahasa Indonesia dengan NMT Berbasis Web," *Jurnal Inovasi Teknologi Terapan,* 2023.
[6] Nan Peng, Yue Wang, Yingying Wei, Lu Liu, Lei Wang, Yu Wang, "Quality Evaluation Model of Automatic Machine Translation based on Deep Learning Algorithm," dalam *International Conference for Emerging Technology (INCET)*, Belgaum, India, 2023.
[7] Dzulkahfi, Herry Sujaini, Tursina, "Perbandingan Hasil Penerjemahan Neural Machine Translation dengan MarianNMT terhadap sumber korpus Wikimedia dan QED&TED," *JURISTI,* 2023.
[8] Yustiana Fauziyah, Ridwan Ilyas, Fatan Kasyidi, "MESIN PENTERJEMAH BAHASA INDONESIA-BAHASA SUNDA MENGGUNAKAN RECURRENT NEURAL NETWORKS," *TEKNOINFO,* 2022.
[9] Irmawati Carolina, Toto Haryanto, "Modeling Of Hyperparameter Tuned RNN-LSTM and Deep Learning For Garlic Price Forecasting In Indonesia," 2024.
[10] Naufal Ananda, Haryas Subyantara Wicaksana, Yusuf Giri, "HYPERPARAMETER TUNING LSTM SEBAGAI ESTIMATOR SENSOR RELATIVE HUMIDITY PADA AUTOMATIC WEATHER STATION BERBASIS SIMULATED ANNEALING," 2023.
[11] WILSON WONGSO , ANANTO JOYOADIKUSUMO , BRANDON SCOTT BUANA, DERWIN SUHARTONO, "Many-to-Many Multilingual Translation Model for Languages of Indonesia," 2023.
[12] Bert Le Bruyn, Martín Fuchs , Martijn van der Klis , Jianan Liu, Chou Mo, Jos Tellings, Henriëtte de Swart, "Parallel Corpus Research and Target Language Representativeness: The Contrastive, Typological, and Translation Mining Traditions," 2022.
[13] AJAY SHRESTHA AND AUSIF MAHMOOD, "Review of Deep Learning Algorithms and Architectures," *IEEE Access,* 2019.
[14] "Perancangan Mesin Translasi berbasis Neural dari Bahasa Kawi ke dalam Bahasa Indonesia menggunakan Microframework Flask," *JSI,* 2022.
[15] Teguh Ikhlas Ramadhan, Nur Ghaniaviyanto Ramadhan, Agus Supriatman, "Implementation of Neural Machine Translation for English-Sundanese Language using Long Short Term Memory (LSTM)," *BITS,* 2022.
[16] Muhammad Yusuf Aristyanto, Robert Kurniawan, "Pengembangan Metode Neural Machine Translation Berdasarkan Hyperparameter Neural Network (Studi Kasus: Bahasa Jerman – Inggris)," 2021.
[17] Jinyi Zhang and Tadahiro Matsumoto, "Corpus Augmentation for Neural Machine Translation with Chinese-Japanese Parallel Corpora," *MDPI,* 2019.
[18] Sree Harsha Ramesh and Krishna Prasad Sankaranarayanan, "Neural Machine Translation for Low Resource Languages using Bilingual Lexicon Induced from Comparable Corpora," 2018.
[19] Andry Chowandaa, Alan Darmasaputra Chowandab, "Generative Indonesian Conversation Model using Recurrent Neural Network with Attention Mechanism," 2018.
[20] Jiatao Gu, Hany Hassan, Jacob Devlin, Victor O.K. Li†, "Universal Neural Machine Translation for Extremely Low Resource Languages," dalam *Proceedings of NAACL-HLT,* 2018.