Jurnal Komputer Antartika
Volume 2 Nomor 1 2024 Hal. 26-34
ISSN: 2988-5272

Neural Machine Tranlation Untuk Bahasa Sunda Loma – Sunda Halus Menggunakan Long Short Term Memory

Marsela Arsya Sakinah[1], Teguh Ikhlas Ramadhan[2], Rudi Hartono[3]

[1,2,3]Program Studi Teknik Informatika, Universitas Pejuangan Tasikmalaya, Indonesia

e-mail korespondensi: 2003010086@unper.ac.id

Informasi Artikel: Submit: 01-02-2024 | Revisi : 08-02-2024 | Terima : 12-02-2024

Abstrak - Bahasa Sunda, dengan kompleksitas undak-usuk basa, memiliki peran penting dalam komunikasi di Jawa Barat, Indonesia. Mesin penerjemah Bahasa Sunda Loma ke Bahasa Sunda Halus menjadi tantangan karena penggunaan kata harus tepat sesuai konteksnya. Mesin ini penting sebagai alat pembelajaran Bahasa Sunda, mengingat banyaknya generasi yang kehilangan pemahaman terhadap bahasa daerah. Neural Machine Translation (NMT), terutama dengan model Long Short Term Memory (LSTM), menjadi solusi yang menjanjikan. Penelitian ini bertujuan mengimplementasikan LSTM dalam penerjemah Bahasa Sunda, tetapi penelitian terbaru masih kurang. Hasil penelitian menunjukkan bahwa optimizer ADAM pada dataset duplikat menghasilkan akurasi terbaik, meskipun masih ada evaluasi yang kurang baik. Mesin penerjemah ini diharapkan dapat membantu pelestarian Bahasa Sunda di era digital. Evaluasi BLEU score menunjukkan kualitas terjemahan yang rendah pada dataset asli dan dataset duplikat dengan optimizer RMS, sementara dengan optimizer ADAM menunjukkan peningkatan signifikan, terutama pada dataset duplikat. Meskipun demikian, masih ditemukan evaluasi yang kurang baik pada dataset yang diduplikat.

Kata Kunci : Bahasa Sunda, LSTM, NMT, RMS, ADAM

Abstracts *- Sundanese, with its intricate speech levels, holds a pivotal role in West Java's communication. Translating from Loma to Formal Sundanese poses challenges due to precise contextual word usage. Crucial for Sundanese language preservation, a Neural Machine Translation (NMT) system using Long Short Term Memory (LSTM) models emerges promising, yet current research is limited.This study aims to implement LSTM in Sundanese translation, focusing on the ADAM optimizer's efficacy on duplicate datasets. While ADAM yields the highest accuracy, some evaluations remain suboptimal. The translation engine's role is vital in preserving Sundanese in the digital era. BLEU score evaluations show low translation quality with RMS and significant improvements with ADAM, especially in duplicates. However, deficiencies persist, notably in duplicated datasets. This endeavor addresses the decline in regional language comprehension among younger generations, fostering Sundanese language education.*

Keywords : Sundanese, LSTM, NMT, RMS, ADAM

1. Pendahuluan

Bahasa sunda merupakan salah satu bahasa daerah yang banyak digunakan di Indonesia, tepatnya di wilayah Jawa Barat. Dalam menggunakan bahasa sunda terdapat istilah speech levels ‘tingkat tutur’ atau lebih dikenal dengan sebutan undak-usuk basa.. Undak-usuk basa merupakan bentuk dari kesopanan dan etika dalam bahasa Sunda. Tidak banyak yang mengetahui akan pentingnya berkomunikasi menggunakan bahasa sunda dengan melibatkan undak usuk basa tersebut. Seiring dengan beberapa perubahan sosial, undak-usuk basa tersebut cenderung mengalami banyak perubahan, mulai dari pembagian yang kompleks menuju tingkatan yang lebih sederhana[1].

Membangun penerjemah bahasa Sunda Loma – bahasa Sunda Halus merupakan sebuah tugas yang sangat khusus, karena pemilihan kata yang digunakan seringkali memiliki arti yang sama namun penggunaan kata harus ditempatkan sesuai dengan kondisinya. Adanya undak-usuk basa ini tergantung pada tiga hal, yaitu (a) pemakai bahasa, siapa penuturnya dan siapa yang dibicarakannya; (b) kedudukan umur pemakai bahasa bawahan, setara, atau atasan; (c) bagaimana situasi yang digambarkan pada waktu menggunakannya, hormat, biasa, loma, atau kasar [2]. Saat ini mesin penerjemah untuk bahasa Sunda masih sangat terbatas, terutama untuk penerjemah bahasa Sunda Loma – bahasa Sunda Halus.

Dibuatnya mesin penerjemah bahasa sunda dapat digunakan sebagai media pembelajaran bahasa sunda. Karena semakin bertambahnya generasi baru, banyak pula yang tidak mengetahui bahasa Sunda meskipun tinggal dan tumbuh di daerah Sunda [3]. Bahasa sunda merupakan warisan budaya takbenda yang termasuk ke dalam

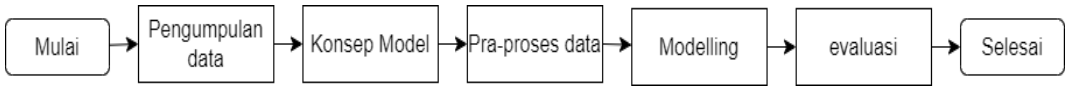
anggota bahasa yang besar dan penting di dunia [4]. Mesin penerjemah yang dapat menerjemahkan bahasa sunda yang tepat dan akurat dapat memberi manfaat yang besar dalam mendukung pelestarian bahasa sunda.

Penerjemah bahasa pada era digital sekarang sangat mudah digunakan untuk kehidupan sehari hari. Mesin penerjemah dapat memudahkan dalam berkomunikasi dan memahami arti dari sebuah bahasa. Mesin penerjemah telah mengalami banyak perkembangan, salah satunya dengan menggunakan Neural Machine Translation (NMT). Model NMT digunakan untuk menerjemahkan dua bahasa yang berbeda [5]. Sebagian besar model penerjemahan yang menggunakan pendekatan berbasis saraf adalah Recurrent Neural Network yang dimodifikasi menjadi model Long Short Term Memory (LSTM) [6]. LSTM merupakan algoritma yang dapat digunakan dalam membangun model prediksi [7]. Dalam hal ini LSTM akan digunakan untuk membuat model penerjemah bahasa. Dengan adanya model LSTM ini memperbesar peluang untuk membangun sebuah mesin penerjemah bahasa untuk bahasa Sunda loma – bahasa Sunda Halus.

Neural Machine Translation adalah sebuah pendekatan pemrosesan bahasa alami yang digunakan untuk menerjemahkan suatu bahasa ke bahasa lain. NMT merupakan mesin penerjemah jaringan syaraf tiruan yang menggunakan pendekatan baru pada teknologi mesin penerjemah yang menggabungkan komponen jaringan saraf tiruan berulang[8]. NMT memanfaatkan konsep deep learning yang kompleks untuk menerjemahkan bahasa. Konsep kerja dalam NMT adalah dengan menerjemahkan suatu bahasa dengan mengumpulkan data dalam bentuk teks, sehingga dapat mengajarkan komputer agar memiliki kemampuan yang dapat menyerupai kemampuan manusia yaitu belajar dari pengalaman. Jika dibandingkan dengan model lain, model NMT memerlukan lebih sedikit pengetahuan linguistic, namun dapat menghasilkan kinerja yang memuaskan[9].

2. Metode Penelitian

Penelitian ini digambarkan dalam bentuk diagram alur untuk menggambarkan alur penelitiaan agar lebih mudah dalam menyampaikan informasi. Tahapan yang dilakukan dalam penelitian ini adalah sebagai berikut.



Gambar 1. Diagram Alur Penelitian

2.1 Pengumpulan data

Penelitian ini dimulai dengan pengumpulan data, termasuk studi literatur dari jurnal, buku, e-book, dan penelitian sebelumnya. Sumber-sumber ini mendukung pemahaman masalah penelitian.

Tabel 1. Referensi kata dalam Bahaasa Sunda loma dan Sunda halus. BL adalah bahasa loma, BHS adalah Bahasa Sunda halus untuk sendiri dan BHB adalah Bahasa Sunda halus untuk orang lain.

No	BL	BHS	BHB
1	anjang	ngadeuheus	natamu
2	balik	wangsul	mulih
3	mawa	ngabantun	nyandak

Wawancara dengan ahli juga dilakukan untuk mendapatkan informasi secara langsung. Selain itu, dilakukan pencarian korpus teks. Korpus teks berisi berbagai teks dalam bahasa Sunda seperti cerita, dongeng, dan cerpen. Korpus ini akan digunakan untuk melatih model bahasa dan mesin penerjemah. Ketidaktersediaan dataset Bahasa Sunda loma – Sunda halus mengharuskan pengumpulan data dari penyusunan dataset.

Tabel 2. Proses pemecahan kalimat dan pengelompokan kalimat menjadi dataset yang akan digunakan. Kalimat yang telah dikumpulkan akan disimpan dalam kolom LOMA atau Halus sesuai dengan referensi kata.

Contoh kalimat	"ceu maneh érék kamana?" Taros Ceu éha Ninggali Ceu Titin mapah. "érek indit ka pasar" waler Ceu Titin. "Rék naon ceu?" Taros Ceu éha
LOMA	HALUS
ceu maneh érék kamana	
	Taros Ceu éha Ninggali Ceu Titin mapah
érek indit ka pasar	
	waler Ceu Titin
Rék naon ceu?	
	Taros Ceu éha

2.2 Konsep model

Pada tahap ini penulis akan membuat rancangan mengenai model yang akan digunakan dalam penelitian. Pemodelan akan digunakan untuk mengenalkan pola bahasa Sunda loma dan Sunda halus. Pada proses ini

perancangan model diharapkan dapat menerjemahkan bahasa dunda dengan mendapatkan nilai akurasi paling tinggi.

2.3 Pra-proses data

Pra proses data adalah langkah langkah yang dilakukan untuk mengelola data mentah sebelum data tersebut dapat digunakan. Dataset yang digunakan adalah hasil penyusunan secara manual, sehingga proses pembersihan data hanya berupa data cleaning yaitu membersihkan data dari karakter yang tidak diperlukan dan menggunakan fungsi lower() dalam python untuk mengganti huruf kapital menjadi huruf kecil.

Tabel 3. Contoh dataset yang telah melaui tahap data cleaning, dengan bahasa sumber (loma) dan bahasa target (halus)

LOMA	HALUS
ceu maneh érék kamana	ceu anjeun bade kamana
tanya ceu éha nempo ceu titin leumpang	taros ceu éha ninggali ceu titin mapah
tanya ceu éha	taros ceu éha

Selanjutnya dilakukan tokenisasi, tokenisasi merupakan proses mengubah data bahasa target dan bahasa sumber menjadi token. Proses tokenisasi mengubah kalimat menjadi kata yang akan digunakan untuk membangun kosa kata. Kemudian token tersebut akan diberi kode dalam bentuk integer.

Tabel 4. Tokenisasi bahasa sumber dan bahasa target

LOMA		HALUS	
Kalimat	Token	Kalimat	Token
ceu maneh érék kamana	ceu, maneh, érék, kamana	ceu anjeun bade kamana	ceu, anjeun, bade, kamana
tanya ceu éha nempo ceu titin leumpang	tanya, ceu, éha, nempo, ceu, titin, leumpang	taros ceu éha ninggali ceu titin mapah	taros, ceu, éha, ninggali, ceu, titin, mapah

Tabel 5. Pengkodean token dengan integer

LOMA		HALUS	
Kalimat	Token	Kalimat	Token
maneh	2	anjeun	22
érék	3	badé	23
kamana	4	kamana	24

Tabel 6. Menyandingkan kalimat menjadi urutan pengkodean bilangan bulat. Ini adalah tahapan akhir untuk mengubah kalimat menjadi vector sehingga model dapat memprosesnya.

LOMA		HALUS	
Kalimat	Token	Kalimat	Token
ceu maneh érék kamana	[1, 2, 3, 4]	ceu anjeun bade kamana	[21, 22, 23, 24]
tanya ceu éha nempo ceu titin leumpang	[5, 1, 6, 7, 1, 8, 9]	taros ceu éha ninggali ceu titin mapah	[25, 21, 26, 27, 21, 28, 29]

2.4 Modelling

Langkah awal dilakukan dengan pembuatan objek model dan penambahan lapisan LSTM. Definisi input dan output diikuti, di mana input berupa urutan kata atau karakter dalam bentuk tensor, sementara outputnya adalah urutan kata dalam bahasa target. Selanjutnya, konfigurasi LSTM dilakukan dengan menentukan parameter untuk setiap lapisan model LSTM, termasuk jumlah unit dan fungsi aktivasi. Proses terakhir adalah pelatihan model, dimana model akan memperbarui parameter internalnya secara iteratif untuk meminimalkan fungsi kerugian selama proses pelatihan.

2.5 Evaluasi

Pada tahap ini, setelah keseluruhan tahap modelling dilakukan, makal model LSTM harus diuji dengan menggunakan data pengujian. Pengujian menggunakan Skor BLEU dan pengujian sampel oleh narasumber ahli dibidang Bahasa Sunda.

Tabel 7. Contoh hasil terjemahan dari teks sumber ke teks target

LOMA (sumber)	Halus (Target)	Hasil terjemahan
urang hayang dahar	Abdi hoyong neda	Aing hoyong dahar

Dari contoh pada Tabel 7 tersebut dilakukan perhitungan menggunakan BLEU. Skor BLEU dihitung dengan rumus berikut.

$$Precision = \frac{jumlah\ \ kata\ yang\ cocok\ \ kalimat\ dalam\ referensi}{Jumlah\ kata\ dalam\ kalimat\ referensi} \quad (1)$$

$$Precision\ = \frac{1}{3} = 0{,}333$$

Kemudian dilakukan perhitungan BP (Brevity Penalty). Dalam contoh tersebut, terjemahan mesin memiliki 3 kata.

BP = min(1, e(1 - (3/3)) = 1 (2)

Langkah terakhir adalah menghitung skor BLEU dengan menggabungkan semua nilai presisi dan faktor BP:

BLEU = BP * exp(1 - (r1 + r2 + r3 +r4)) * (p1 * p2 * p3*p4) (3)
BLEU = 1 * exp(1 - (0 + 0 + 0+0)) * (0,333 * 0 * 0*0) = 0

3. Hasil dan Pembahasan

3.1. Pengumpulan data

Penyusunan dataset dilakukan dengan mengutip naskah biantara, dialog, dan cerpen Bahasa Sunda yang dibagi menjadi beberapa kalimat dan disimpan dalam Google Spreadsheet. Totalnya, terdapat 1600 dataset dalam dataset asli. Dalam penelitian ini, dataset asli digunakan bersama dengan dataset yang diduplikatkan menjadi 8000 dataset. Dataset asli terbagi menjadi 1400 data training dan 160 data testing, sementara dataset yang diduplikatkan memiliki 7200 data training dan 800 data testing. Dengan demikian, data training diambil sebanyak 90%, dan data testing sebanyak 10% dari jumlah kalimat paralel yang ada.

	LOMA	HALUS
410	aya telepon telepon jang maneh.	Aya telepon telepon pikeun anjeun.
1547	maneh nyaho teu?	anjeun uninga teu?
1030	nepi ka huntu garing	dugi ka waos garing
42	urang pikir maneh the salah	Abdi pikir anjeun ngagaduhan kasalahan.
208	aing panggih jeung manehna di stasion.	Abdi pendak sareng anjeunna di stasion.

Gambar 2. Contoh dataset mentah yang belum melalui tahapn pengolahan data

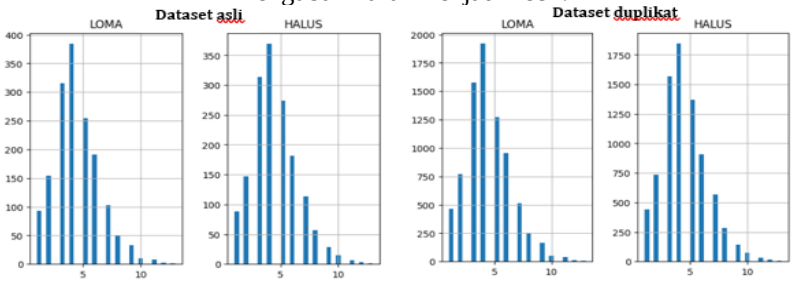
3.2. Konsep Model

Dalam penerjemahan dari Bahasa Sunda loma ke Bahasa Sunda halus, model LSTM sering digunakan dalam neural machine translation (NMT) untuk memahami konteks dan urutan kata. Untuk mengoptimalkan struktur LSTM, penelitian ini menggunakan metode Root Mean Square (RMS) dan Adaptive Moment Estimation (ADAM). RMS mengurangi divergensi nilai gradien yang tinggi, sedangkan ADAM mengadaptasi laju pembelajaran untuk setiap parameter secara adaptif. Kedua metode ini digunakan dalam pelatihan model NMT untuk mempercepat konvergensi dan menghindari masalah seperti vanishing atau exploding gradients.

3.3. Pra proses data

	0	1
0	aya telepon telepon jang maneh	aya telepon telepon pikeun anjeun
1	maneh nyaho teu	anjeun uninga teu
2	nepi ka huntu garing	dugi ka waos garing
3	urang pikir maneh the salah	abdi pikir anjeun ngagaduhan kasalahan
4	aing panggih jeung manehna di stasion	abdi pendak sareng anjeunna di stasion

Gambar 3. Dataset setelah tahap *preprocessing* hanya menghilangkan karakter yang tidak diperlukan dan mengubah huruf menjadi kecil.



Gambar 4. Grafik jumlah dataset dan frekuensi banyaknya panjang kalimat pada dataset.

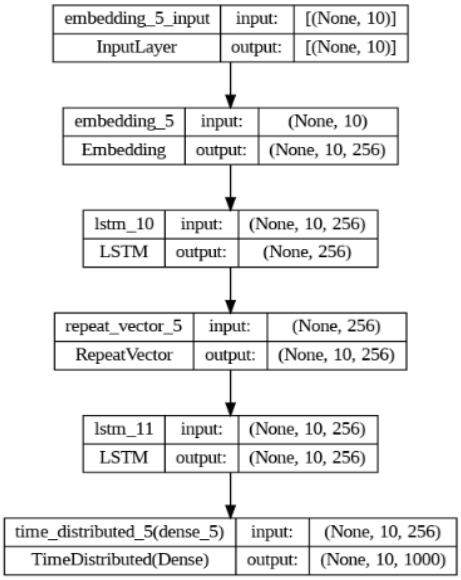
Tabel 8. Jumlah encode setelah tokenisasi (variasi kata dalam dataset).

LOMA	HALUS
1533	1567

Tabel 9. Hasil penggabungan encode dengan kalimat token

LOMA		HALUS	
Kalimat	Encode	Kalimat	Encode
anu kaayaan pasar na keur rarujit	[31, 32, 12, 33, 34, 35]	anu kaayaan pasar na nuju rarujit	[52, 53, 22, 54, 55,56]
sabab geus hujan badag cikeneh	[36, 37, 38, 39, 40]	margi atos hujan ageung nembe	[57, 58, 59, 60, 61]

3.4. Modelling



Gambar 5. Arsitektur modeldari tahapan input sampai tahap output

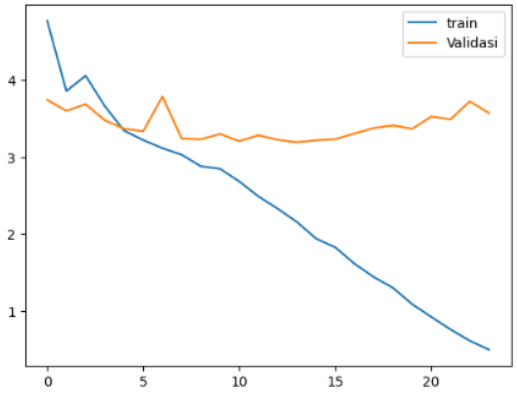
Lapisan pertama berperan sebagai lapisan input untuk membaca vektor input berukuran 10 dari bahasa sumber. Kemudian, vektor input disematkan ke lapisan embedding sebelum diproses lapisan LSTM. Lapisan LSTM pertama bertindak sebagai encoder, dan outputnya akan diteruskan ke lapisan RepeatVector. Lapisan RepeatVector berfungsi sebagai jembatan antara lapisan LSTM pertama dan lapisan LSTM kedua yang juga berperan sebagai encoder. Decoder LSTM akan menerima output dari lapisan LSTM kedua dan menuju ke lapisan dense. Pada lapisan dense, distribusi probabilitas untuk setiap kelas dalam kosakata target dihitung menggunakan aktivasi softmax. Berikut adalah gambaran model pada Gambar 5 dan Gambar 6.

```
Model: "sequential_5"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 embedding_5 (Embedding)     (None, 10, 256)           256000

 lstm_10 (LSTM)              (None, 256)               525312

 repeat_vector_5 (RepeatVec  (None, 10, 256)           0
 tor)

 lstm_11 (LSTM)              (None, 10, 256)           525312

 time_distributed_5 (TimeDi  (None, 10, 1000)          257000
 stributed)

=================================================================
Total params: 1563624 (5.96 MB)
Trainable params: 1563624 (5.96 MB)
Non-trainable params: 0 (0.00 Byte)
_________________________________________________________________
```

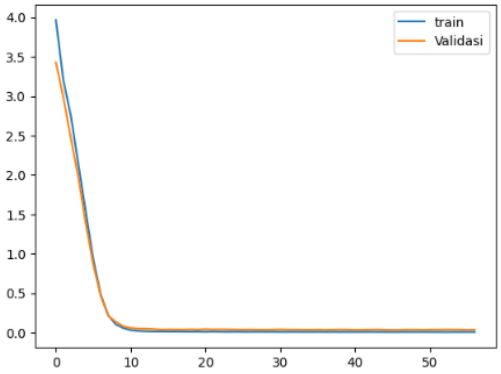
Gambar 6. Lapisan embedding menerima input berukuran 10, yang merupakan vektor maksimum dalam bahasa sumber. Sedangkan lapisan akhir menghasilkan output berukuran 10, yang juga merupakan ukuran vektor maksimum untuk target.

3.5. Evaluasi

Dalam kasus ini, optimizer RMSprop dan fungsi kerugian 'sparse_categorical_crossentropy' dipilih. Proses pelatihan menggunakan fungsi fit dengan jumlah epoch 200 dan batch size 64. Data validasi sebesar 0.1 juga ditentukan untuk memantau proses pelatihan. Grafik loss dapat divisualisasikan untuk mengamati kinerja model selama proses pembelajaran.



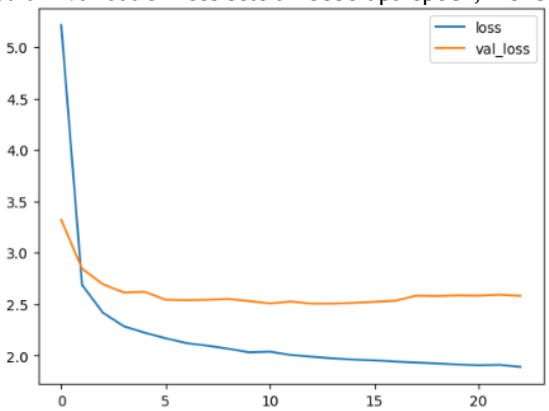
Gambar 7. Grafik loss pelatihan menggunakam optimizer RMS dengan dataset asli



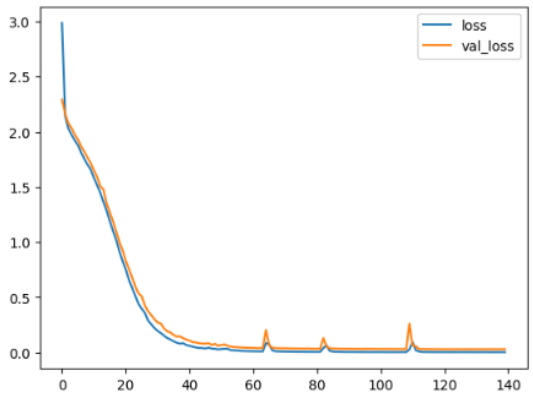
Gambar 8. Garfik loss pelatihan menggunakan optimizer RMS dengan dataset duplikat

Pada Gambar 7 .grafik loss selama pelatihan model menampilkan garis validasi (jingga) menjauhi garis *train* (biru). Garis *training* dan validasi yang saling menjauh menandakan bahwa pelatihan menggunakan model tersebut kurang efektif [10]. Grafik akurasi pada garis validasi cenderung stabil sedangkan garis *train* mengalami mengalami penurunan pada *epoch* pertama sampai akhir. Model berhenti di epoch ke 24 yang menandakan tidak ada peningkatan validasi selama 10 *epoch* berturut turut.Sedangkan pada Gambar 8 menunjukan garis *train* dan garis validasi saling mendekat dan menurun. Kedua garis tersebut cenderung stabil dan berhenti setelah *epoch* ke 90. Kedua garis tersebut membentuk garis yang hampir sejajar, menunjukkan bahwa performa model baik pada data pelatihan maupun data validasi relatif stabil dan konsisten selama periode tersebut.

Proses pelatihan model dilakukan dengan ADAM menggunakan fungsi fit, yang memanfaatkan data latih untuk mengoptimalkan parameter-model. Jumlah epoch 200 dan ukuran batch 64 ditentukan untuk mengontrol iterasi dan proses pembelajaran secara batch. Callback EarlyStopping digunakan untuk menghentikan pelatihan jika tidak terjadi peningkatan dalam validation loss setelah beberapa epoch, mencegah overfitting.

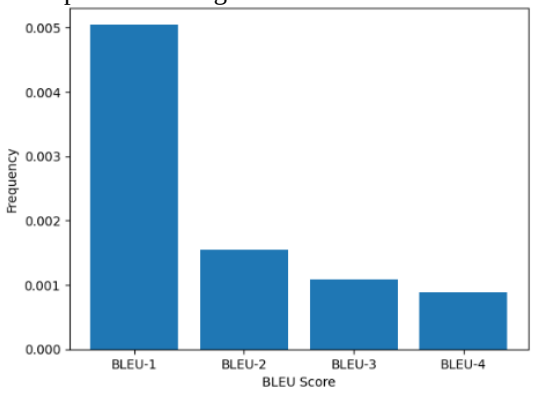


Gambar 9. Grafik loss pelatihan menggunakan optimizer ADAM dengan dataset asli

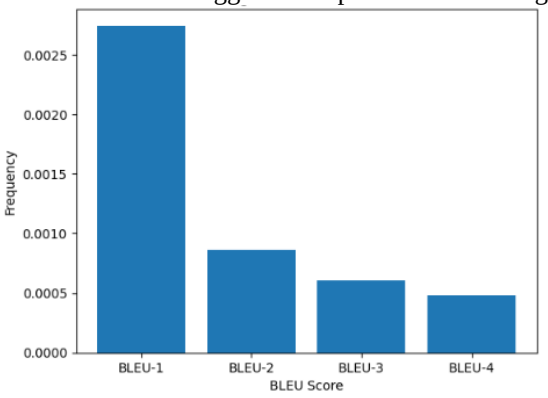


Gambar 10. Grafik losspelatihan menggunakan optimizer ADAM dengan dataset duplikat

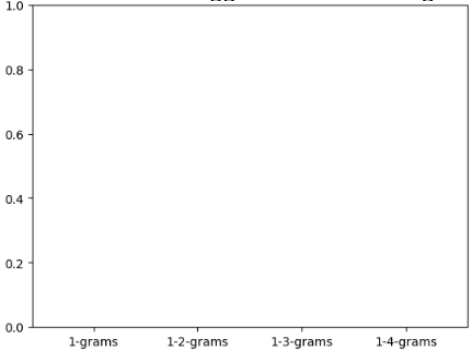
Grafik loss pada Gambar 9 selama pelatihan model menunjukkan bahwa garis validasi (jingga) menjauhi garis train (biru). Kedua garis yang saling menjauh menandakan bahwa model memiliki performa yang konsisten pada kedua set data. Hal ini mengindikasikan bahwa model dapat menggeneralisasi dengan baik, tidak hanya pada data pelatihan tetapi juga pada data yang belum pernah dilihat sebelumnya (data validasi). Pelatihan model berhenti di epoch ke-21. Sedangkan pada Gambar 10. menunjukkan hasil pelatihan menggunakan optimizer ADAM dengan data yang diduplikat. Garis train dan garis validasi mendekat dan menurun, stabil rapat, dan berhenti pada epoch ke 140. Model berhasil dalam proses pelatihan dan generalisasi.



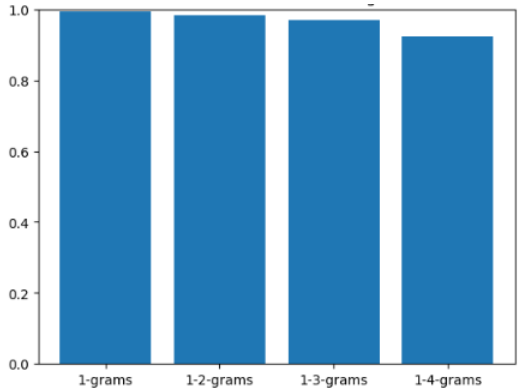
Gambar 11. Hasil BLEU skor menggunakan optimizer RMS dengan dataset asli.



Gambar 12. Hasil BLEU skor menggunakan RMS dengan dataset duplikat



Gambar 13. Hasil BLEU skor menggunakan ADAM dengan dataset asli



Gambar 14. Hasil BLEU skor menggunakan ADAM dengan dataset duplikat

Gambar 11 menampilkan hasil perhitungan BLEU score pada data test menggunakan dataset asli (RMS). Skor BLEU-1, BLEU-2, BLEU-3, dan BLEU-4 secara berturut-turut adalah 0.0049, 0.0016, 0.00125, dan 0.001. Hasil ini menunjukkan kualitas terjemahan mesin yang rendah dibandingkan dengan teks referensi manusia. Hanya sekitar 0.49% dari unigram dalam hasil terjemahan yang sesuai dengan unigram dalam teks referensi. Hasil perhitungan BLEU score pada Gambar 13 dengan menggunakan dataset duplikat(RMS) menunjukkan skor BLEU-1, BLEU-2, BLEU-3, dan BLEU-4 masing-masing adalah 0.00250, 0.000582, 0.000540, dan 0.005. Pada gambar 13, grafik tersebut menujukan evaluasi dari dataset asli yang menganmpilkan hasil perhitungan yang kurang baik. Skor pada 1-grams, 1-2 grams, 1-3 grams dan 1-4 grams masing masing adalah 0.000, 0.000, 0.000 dan 0.000 Meskipun terjadi peningkatan, kualitas terjemahan masih rendah. Pada Gambar 14, hasil perhitungan BLEU score pada model mesin penerjemah LSTM dengan optimizer ADAM menunjukkan skor yang cukup baik, dengan BLEU-1 sebesar 99.2%, BLEU-2 sebesar 98.3%, BLEU-3 sebesar 96.9%, dan BLEU-4 sebesar 92.4%. Meskipun demikian, terdapat evaluasi yang kurang baik pada dataset yang diduplikat.

Tabel 10. Contoh hasil terjemah dan skor penilaian oleh narasumber ahli. Semakin mendekati 0 hasil terjemahn berarti tidak sesuai, mendekati 1 berarti terjemahan sesuaai.

Bahasa Sumber	Hasil Terjemahan Otomatis (RMS data asli)	skor
sabab boga keneh pagawean	manehna maneh urang ka maneh teu ka urang maneh maneh teu urang ka maneh teu urang ka urang teu ka anu maneh teu maneh teu maneh urang maneh maneh urang ka maneh urang maneh ...	0
da butuh jang dahar sapopoe	manehna maneh urang ka maneh ka teu ka urang maneh maneh teu urang ka maneh anu maneh teu urang teu urang teu ka urang teu ka maneh teu maneh teu maneh urang maneh urang ka maneh...	0
Bahasa Sumber	**Hasil Terjemahan Otomatis (RMS data duplikat)**	**skor**
naha maneh teu bisa indit sorangan	manehna maneh anu urang teu teu anu teu urang urang teu teu maneh anu urang teu teu urang teu anu teu ka anu ka teu urang teu urang teu anu teu ka teu anu ka urang anu teu anu t...	0
ih aing ge nempo	manehna urang urang ka urang anu urang ka urang maneh maneh ka aing ka urang aing maneh anu teu urang teu urang maneh teu anu urang teu ka teu ka teu ka urang aing anu teu anu maneh anu k...	0
Bahasa Sumber	**Hasil Terjemahan Otomatis (ADAM data asli)**	**skor**
sabaraha umur maneh	urang maneh	0
maneh baraya	urang	0
Bahasa Sumber	**Hasil Terjemahan Otomatis (ADAM data duplikat)**	**skor**
Tong jadi hinaan	Teu kenging janten hinaan	1
Ari maneh boga kartu pos	Dupi anjeun ngagaduhan kartu pos	0,8

Berdasarkan evaluasi oleh narasumber ahli Sastra Sunda, hasil terjemahan menunjukkan bahwa penggunaan optimizer ADAM pada dataset yang diduplikatkan memberikan hasil terbaik. Meskipun demikian, terdapat beberapa evaluasi yang perlu diperhatikan, seperti kurangnya pengenalan subjek dalam beberapa hasil terjemahan dan satu kesalahan penggunaan kata kerja yang kurang sesuai dari 11 sampel. Meskipun demikian, kesalahan tersebut tidak signifikan dan tidak mempengaruhi makna keseluruhan kalimat. Sebaliknya, penggunaan optimizer RMS dan ADAM dengan dataset asli tidak menghasilkan terjemahan yang sesuai menurut penilaian

narasumber ahli. Meskipun ada satu kalimat yang sama, namun memiliki arti dan makna yang berbeda sehingga dinilai tidak sesuai.

4. Kesimpulan

Penelitian mengenai implementasi Neural Machine Translation (NMT) dengan algoritma Long Short Term Memory (LSTM) dari Bahasa Sunda loma ke Bahasa Sunda halus menghadapi tantangan kompleksitas karena adanya *undak usuk basa*. Percobaan dilakukan untuk menyelesaikan permasalahan ini, dan hasil menunjukkan adanya overfitting yang dipengaruhi oleh jumlah data. Variasi dataset dapat meningkatkan akurasi terjemahan, dengan optimizer ADAM memberikan hasil terbaik, baik pada dataset asli maupun yang diduplikat. Evaluasi oleh narasumber ahli menunjukkan bahwa hasil terjemahan yang paling optimal diperoleh dengan optimizer ADAM pada dataset yang diduplikatkan, sementara prediksi menggunakan optimizer RMS dan ADAM dengan data asli menunjukkan hasil yang jauh dari target yang diharapkan.

Referensi

[1] C. Sobarna, G. Gunardi, and A. S. Afsari, "Toponim daToponim dalam Upaya Pemertahanan Bahasa Sunda di Wilayah Jawa Tengah: Kasus di Kecamatan Dayeuhluhur, Kabupaten Cilacaplam Upaya Pemertahanan Bahasa Sunda di Wilayah Jawa Tengah: Kasus di Kecamatan Dayeuhluhur, Kabupaten Cilacap," *Makna (Jurnal Kaji. Komunikasi, Bahasa, dan Budaya)*, vol. 4, no. 1, pp. 154–173, 2019, doi: 10.33558/makna.v4i1.1678.

[2] M. A. Pangestu and S. Sudjianto, "Analisis Struktur dan Pemakaian Keigo dan Perbandingannya dengan Undak Usuk Basa Sunda," *IDEA J. Stud. Jepang*, vol. 3, no. 1, pp. 1–11, 2021, doi: 10.33751/idea.v3i1.3328.

[3] N. Komalasari, E. W. Hidayat, and A. P. Aldya, "Aplikasi Pengenalan Bahasa Sunda Berbasis Multimedia Dengan Konsep V.I.S.U.a.L.S," *J. Nas. Pendidik. Tek. Inform.*, vol. 9, no. 1, p. 21, 2020, doi: 10.23887/janapati.v9i1.21654.

[4] Y. Hendayana, "TEKS DAN KONTEKS DALAM JEJAK BUDAYA TAKBENDA STUDI KASUS: BABASAN DAN PARIBASA SUNDA," *Pros. Balai Arkeol. Jawa Barat*, vol. 3, no. 1 SE-, pp. 215–223, Dec. 2020, doi: 10.24164/prosiding.v3i1.24.

[5] Z. Munawar, Iswanto, D. Widhiantoro, and N. I. Putri, "Analisis Sentimen Covid-19 Pada Media Sosial Dengan Model Neural Machine Translation," *Tematik*, vol. 9, no. 1, pp. 15–20, 2022, doi: 10.38204/tematik.v9i1.899.

[6] T. I. Ramadhan, N. G. Ramadhan, and A. Supriatman, "Implementation of Neural Machine Translation for English-Sundanese Language using Long Short Term Memory (LSTM)," *Build. Informatics, Technol. Sci.*, vol. 4, no. 3, pp. 1438–1446, 2022, doi: 10.47065/bits.v4i3.2614.

[7] Moch Farryz Rizkilloh and Sri Widiyanesti, "Prediksi Harga Cryptocurrency Menggunakan Algoritma Long Short Term Memory (LSTM)," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 6, no. 1, pp. 25–31, 2022, doi: 10.29207/resti.v6i1.3630.

[8] W. Gunawan, H. Sujaini, and T. Tursina, "Analisis Perbandingan Nilai Akurasi Mekanisme Attention Bahdanau dan Luong pada Neural Machine Translation Bahasa Indonesia ke Bahasa Melayu Ketapang dengan Arsitektur Recurrent Neural Network," *J. Edukasi dan Penelit. Inform.*, vol. 7, no. 3, p. 488, 2021, doi: 10.26418/jp.v7i3.50287.

[9] S. Yang, Y. Wang, and X. Chu, "A Survey of Deep Learning Techniques for Neural Machine Translation," 2020, [Online]. Available: http://arxiv.org/abs/2002.07526

[10] G. P. Natakusumah and E. Ernastuti, "Implementasi Metode CNN Multi-Scale Input dan Multi-Feature Network untuk Dugaan Kanker Payudara," *JOINTECS (Journal Inf. Technol. Comput. Sci.*, vol. 7, no. 2, p. 43, 2022, doi: 10.31328/jointecs.v7i2.3637.