ISSN 2723-6471

LSTM-Based Machine Translation for Madurese-Indonesian

Danang Arbian Sulistyo [1,*], Aji Prasetya Wibawa [2], Didik Dwi Prasetya [3], Fadhli Almu'iini Ahda [4]

[1,2,3,4] Universitas Negeri Malang, Jl. Semarang 5, Malang 65145 Jawa Timur Indonesia
[1,4] Institut Teknologi dan Bisnis Asia Malang, Jl. Soekarno Hatta, Rembuksari No.1 A, Malang 65113, Jawa Timur, Indonesia
[1] danangarbian@gmail.com*; [2] aji.prasetya.ft@um.ac.id; [3] didikdwi@um.ac.id; [4] adhi32286@gmail.com
* corresponding author

(Received June 26, 2023; Revised August 21, 2023; Accepted August 28, 2023; Available online September 15, 2023)

Abstract

Madurese is one of the regional languages in Indonesia, which dominates East Java and Madura Island in particular. The use of Madurese as a daily language has declined significantly due to a language shift in children and adolescents, some of which are caused by a sense of prestige and difficulty in learning Madurese. The scarcity of research or scientific titles that raises the Madurese language also helps reduce literacy in the language. Our research focuses on creating a translation machine for Madurese to Indonesian to maintain and preserve the existence of the Madurese language so that learning can be done through digital media. This study use the latest dataset for the Madurese-Indonesian language by using a corpus of 30,000 Madura-Indonesian sentence pairs from the online Bible. This study scrapped online Bible pages to organize the corpus based on the Indonesian and Madurese bilingual Bible. Then This study manually process text to match the two languages' scrapping results, normalization, and tokenization to remove non-printable characters and punctuation from the corpus. To perform neural machine translation (NMT), This study connected the RNN encoder with the RNN decoder of the language model, while for training and testing, This study used a sequential model with LSTM, while the BLEU measure was used to assess the accuracy of the translation results. This study used the SoftMax optimization function with Adam Optimizer and added some settings, including using 128 layers in the training process and adding a Dropout layer so that This study got the average evaluation result for BLEU-1 is 0.798068, BLEU-2 is 0.680932, BLEU-3 is 0.623489, and for BLEU-4 is 0.523546 from five tests conducted. Given the language differences between Madurese and Indonesian, this can be the best approach for machine translation of Indonesian to Madurese.

Keywords: Machine Translation, Indonesian, Madurese, NLP, LSTM

1. Introduction

Ethnic Madurese use the regional language of Madurese for daily communication both inside and outside the island of Madura. Most Madura Island speaks Madurese, which is also spoken in a few other places, including Jember, Pasuruan, and Probolinggo. Some studies have found that children are shifting from using Madurese to Indonesian, the Indonesian official language [1]; this is due to a sense of prestige and the difficulty of learning Madurese, which has a variety of dialects and language levels. The future of Madurese as one of Indonesia's regional languages is seriously threatened by its declining use. As a result, as the originality of Indonesian dialects slowly erodes, Madurese's future is in peril.

Additionally, Madurese academic literature and scientific publications are hard to find in public and academic libraries. Research on reliable Madurese-Indonesian translation systems must therefore be advanced. Through this program, the Madurese language will be preserved as an essential indigenous regional language in Indonesia and made easier to learn in the future through digital media.

The development of a Madurese to Indonesian translation tool is the main topic of this study as one of the solutions to the requirement for producing translation tools for regional languages in Indonesia. Some previous studies have focused more on translating Indonesian into foreign languages, such as [2], which translates English-Indonesian. Several studies that raise the theme of regional language research in Indonesia include Lampung-Indonesian translators [3], Muna-Indonesian [4], Javanese-Indonesian [5][6], Minang-Indonesian [7], Balinese-Indonesian [8] and Madurese-Indonesian [9].

This study recommend employing neural machine translation to create a novel machine that can translate Madurese to Indonesian. This study employ it because it is a better iteration of the translation method intended to displace earlier ones [10]. In contrast to earlier statistical and rule-based machine translation models [11], the neural engine replicates the entire machine translation procedure using a single artificial neural network [12] . In our system, This study use the RNN encoder and decoder [13]. Long Short-Term Memory (LSTM) is employed because it can forecast phrases instead of individual words [14], offering a more pragmatic and practical alternative to conventional Neural Machine Translation (NMT) methods. [15]. The evaluation of machine translation uses the Bilingual Evaluation Understudy (BLEU), often used to evaluate machine translators [16]. A BLEU matrix is created to measure how well the machine translator output matches the translation reference in terms of the output phrase length variable [17].

2. Material and Methodology

The employment of artificial neural networks in this research is followed by training and testing the LSTM sequential model [18] [19], and the outcomes are assessed using the BLEU matrix.

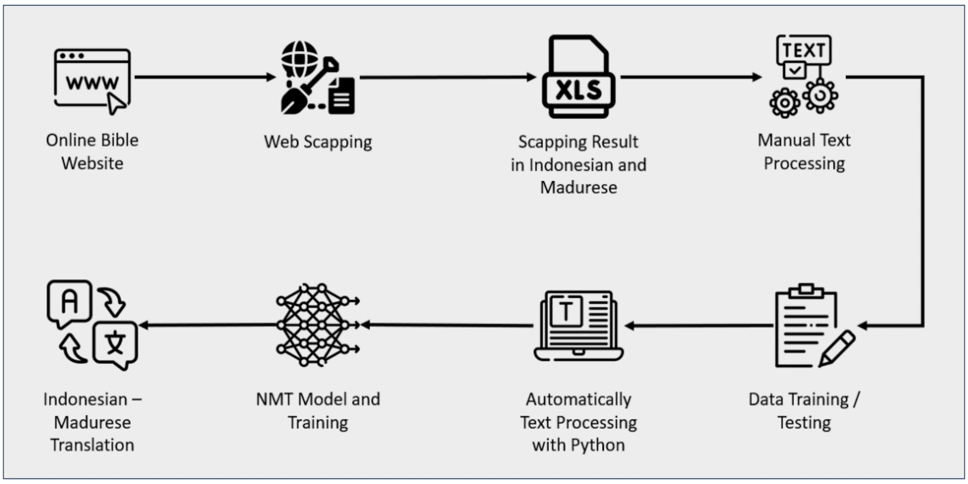

WWW
Online Bible Website
Web Scapping
XLS
Scapping Result in Indonesian and Madurese
TEXT
Manual Text Processing
Data Training / Testing
T
Automatically Text Processing with Python
NMT Model and Training
A
文
Indonesian – Madurese Translation

Figure 1. Design of the Madurese-Indonesian Machine Translation System

Figure 1 shows the processes for developing a machine translation of Madurese to Indonesian. The process begins with scraping the dataset from the website. Then this study do text processing manually for each dataset, combining the two datasets into one .txt format and then doing text processing automatically using Python to remove non-printable chars, punctuation, and numerical tokens and convert them to lowercase. After the dataset cleaning process, the training and testing process uses the NMT model [20] for the best accuracy for translating Madurese to Indonesian.

2.1.Dataset

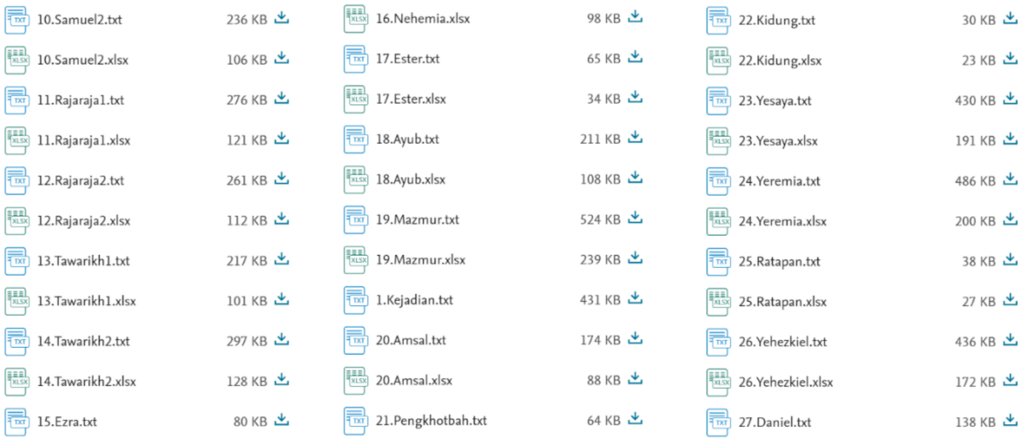
The dataset this study use is from the Madurese Corpus Dataset, based on Madurese and Indonesian online bibles. The dataset contains more than 30,000 pairs of Madurese and Indonesian sentences. As discussed earlier, this study use a clean dataset, which has been manually checked to match the scrapping results of the Indonesian and Madurese translations.

This study manually check the results of the scrapping process because there are differences in the location of the translation between Indonesian and Madurese. The differences occur in several verses in Madurese. Several verses in Madurese refer to one previous verse due to the length of the sentence in the verse. As seen in Table 1, there is a difference in the scrapping results; in verse 18 in the Indonesian translation of the Bible, it is not combined like in the translation of the Madurese language. Here this study make two ways, the first is to break down the Indonesian translation if there are more than three sentences in one verse, or this study use the second way to combine them into one verse if there are less than three sentences.

Table1. Scrapping Result Comparison

Indonesian	Madurese
18.Inilah keturunan Peres: Peres memperanakkan Hezron,	18.Areya' garis toronanna Daud, molae dhari Peres: Peres, Hezron, Ram, Aminadab, Nahason, Salmon, Bowas, Obed, Isay, Daud.
19.Hezron memperanakkan Ram, Ram memperanakkan Aminadab,	19(4:18)
20.Aminadab memperanakkan Nahason, Nahason memperanakkan Salmon,	20(4:18)
21.Salmon memperanakkan Boas, Boas memperanakkan Obed,	21(4:18)
22.Obed memperanakkan Isai dan Isai memperanakkan Daud.	22(4:18)

After going through the manual checking process, the dataset is ready for the following process, as shown in Figure 2, which is an example of the contents of the dataset this study use. There are two files, namely Excel files (.xlsx) and text (.txt); in the Excel file, this study place the translation results from Indonesian and Madurese in different columns, providing tab separators for the two translation results easier when this study move them to text (.txt).



10.Samuel2.txt	236 KB	16.Nehemia.xlsx	98 KB	22.Kidung.txt	30 KB
10.Samuel2.xlsx	106 KB	17.Ester.txt	65 KB	22.Kidung.xlsx	23 KB
11.Rajaraja1.txt	276 KB	17.Ester.xlsx	34 KB	23.Yesaya.txt	430 KB
11.Rajaraja1.xlsx	121 KB	18.Ayub.txt	211 KB	23.Yesaya.xlsx	191 KB
12.Rajaraja2.txt	261 KB	18.Ayub.xlsx	108 KB	24.Yeremia.txt	486 KB
12.Rajaraja2.xlsx	112 KB	19.Mazmur.txt	524 KB	24.Yeremia.xlsx	200 KB
13.Tawarikh1.txt	217 KB	19.Mazmur.xlsx	239 KB	25.Ratapan.txt	38 KB
13.Tawarikh1.xlsx	101 KB	1.Kejadian.txt	431 KB	25.Ratapan.xlsx	27 KB
14.Tawarikh2.txt	297 KB	20.Amsal.txt	174 KB	26.Yehezkiel.txt	436 KB
14.Tawarikh2.xlsx	128 KB	20.Amsal.xlsx	88 KB	26.Yehezkiel.xlsx	172 KB
15.Ezra.txt	80 KB	21.Pengkhotbah.txt	64 KB	27.Daniel.txt	138 KB

Figure 2. Madurese-Indonesian Dataset

2.2.Neural Machine Translation

To execute neural machine translation, this study need to connect the encoder and decoder of the language model [21]. The encoder analyzes the source language sentence word by word and creates a representation of the input sentence.

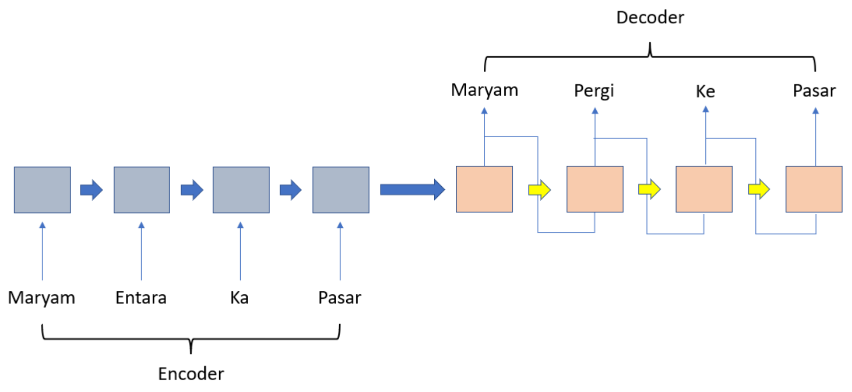


Figure 3. Encoder-Decoder architecture model

The decoder or language model takes the encoder output as input and generates a word-for-word translation of the input words [22]. Figure 3 illustrates how to link the encoder and decoder models together. The model architecture for machine translation is known as the encoder-decoder model.

For each source sentence, the encoder generates a representation. The decoder then creates a sentence-by-sentence translation, with the previously created sentence serving as the global target context. Because the encoder and decoder are iterative, both include loops that process the sequence at different time stages.

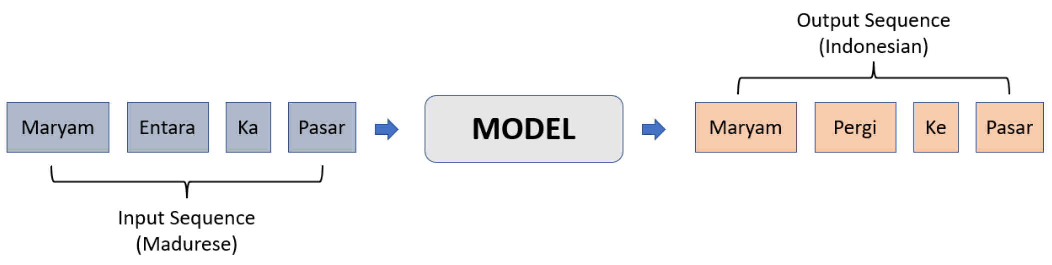


Figure 4. Sequential Model for Madurese-Indonesian Machine Translation

Figure 4 shows that the source sequence is the input language: Madurese, while the target sequence is the expected output language: Indonesian. Because the prediction issue is a multi-class classification problem [23], the model is generated using Adam's effective stochastic gradient descent technique, which minimizes categorical loss. Adam is more efficient than SGD [24], with less memory and training time required. The model was then trained using an epoch of 10000.

2.3.Evaluation

The translation results are evaluated by comparing the translated sentence with the reference sentence using the Bilingual Evaluation Understudy (BLEU). BLEU is an algorithm that evaluates the quality of machine-translated translation results from a source to a destination language. BLEU measures a modified statistical-based precision score between the translated result automatically and the reference translation by using a constant called brevity penalty (BP).

$$BP_{BLUE} = \{1,\ \ if\ c > r\ e^{1\frac{r}{c}},\ \ if\ c \le r \tag{1}$$

$$P_{n} = \frac{\sum C \in corpus\ n-gram\ \in C\ \sum count\ clip^{(n-gram)}}{\sum C \in corpus\ n-gram\ \in C\ \sum count_{(n-gram)}} \tag{2}$$

$$BLEU = BP_{BLEU} . e^{\sum_{n-1}^{N} W_n} \log log\, p_n \qquad (3)$$

The BLEU assessment score ranges from 0 to 1 [16]. A translation will be worth 1 if it exactly matches the reference translation. That is why, even using human translation, it is doubtful to score 1. Note that the higher the score, the more reference translations per sentence. To produce a high BLEU score, the length of the sentence to be translated must also be close to the length of the reference sentence, and more importantly, the sentence to be translated has the same word structure and order as the reference sentence [25]. Eq (1) to (3) show how to write the BLEU formula.

3. Result and Discussion

This study used 30,000 pairs of Madurese and Indonesian sentences based on the online Bible as a dataset. The use of the Bible as the primary reference is due to the lack of reference to Madura books as a dataset. This study collected datasets based on existing Madura books from various sources and only got approximately 1048 pairs of Madura - Indonesian sentences. The number is certainly very less if used as a dataset in NMT [26]. In addition, the Madurese translation of the Bible has gone through an authorized procedure from the Lembaga Alkitab Indonesia (Indonesian Biblical Institute), ensuring that the translation's contents are consistent with the original Bible.

3.1. Dataset Cleaning Process

In this process, this study perform several steps, which include normalization, tokenization, converting all words to lowercase, removing punctuation, and removing numbers in the sentence, each of which can be seen in Table 2. Normalization is accomplished using normalized Unicode letters, which transform string vectors through Normalization Form Canonical Decomposition [27]. Canonical equality is used to deconstruct characters in NFD, and certain combination characters are put in a specific sequence.

Table 2. Dataset Cleaning Process

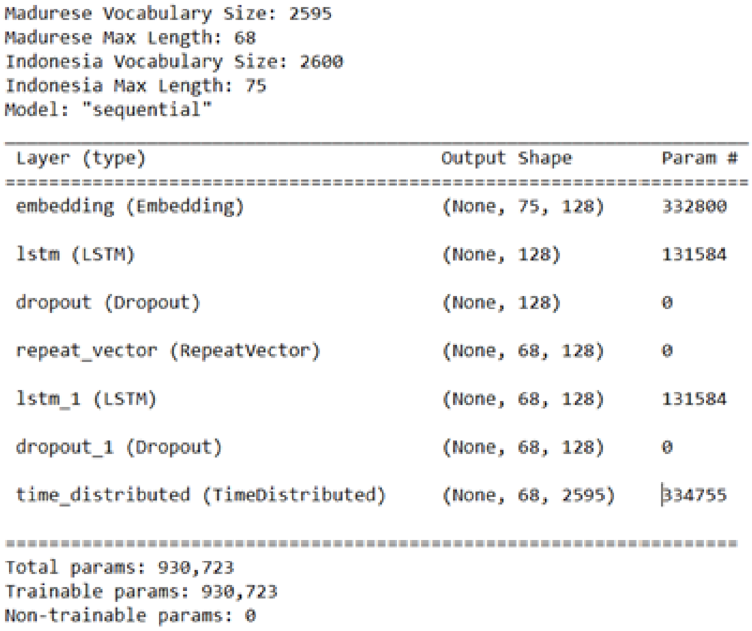
No	Process	Input	Output
1	Normalize Unicode characters	"E bakto Rato Daud buru dhari Rato Saul, laju ngongse ka Ziklag!!". Bannya’ parjurit se esto ban apangalaman dhateng agabung ban Daud.	E bakto Rato Daud buru dhari Rato Saul, laju ngongse ka Ziklag!!. Bannya parjurit se esto ban apangalaman dhateng agabung ban Daud.
2	Tokenize on white space	E bakto Rato Daud buru dhari Rato Saul, laju ngongse ka Ziklag!!. Bannya parjurit se esto ban apangalaman dhateng agabung ban Daud.	E bakto Rato Daud buru dhari Rato Saul, laju ngongse ka Ziklag!!. Bannya parjurit se esto ban apangalaman dhateng agabung ban Daud.
3	Convert to lowercase	E bakto Rato Daud buru dhari Rato Saul, laju ngongse ka Ziklag!!. Bannya parjurit se esto ban apangalaman dhateng agabung ban Daud.	e bakto rato daud buru dhari rato saul, laju ngongse ka ziklag!!. bannya parjurit se esto ban apangalaman dhateng agabung ban daud.
4	Remove punctuation from each token.	e bakto rato daud buru dhari rato saul, laju ngongse ka ziklag!!. bannya parjurit se esto ban apangalaman dhateng agabung ban daud.	e bakto rato daud buru dhari rato saul, laju ngongse ka ziklag bannya parjurit se esto ban apangalaman dhateng agabung ban daud.
5	Remove non-printable chars from each token	e bakto rato daud buru dhari rato saul, laju ngongse ka ziklag bannya parjurit se esto ban apangalaman dhateng agabung ban daud.	e bakto rato daud buru dhari rato saul, laju ngongse ka ziklag bannya parjurit se esto ban apangalaman dhateng agabung ban daud.

6	Remove tokens with numbers in them	e bakto rato daud buru dhari rato saul, laju ngongse ka ziklag bannya parjurit se esto ban apangalaman dhateng agabung ban daud.	e bakto rato daud buru dhari rato saul, laju ngongse ka ziklag bannya parjurit se esto ban apangalaman dhateng agabung ban daud.

Tokenization is a method of aiding comprehension by dividing raw text into little bits (tokens). The tokenization approach used in this work is white space tokenization, which splits phrases into meaningful terms [28]. By switching to lowercase, case folding was accomplished. Lowercase aims to transform all letters in the document from 'a' to 'z', to lowercase [29]. Does noise cleanup eliminate non-printable characters and punctuation marks such as [!,-./:;ó?@_"#$%&'()*.'|] [30].

3.2. Training and Testing

In Figure 5, this study define a function that performs the tokenization process on each text of the Madurese and Indonesian to obtain the vocabulary size as well as the maximum length, so this study perform word embedding by converting words into indices with the encoding sequence (tokenizer, length, lines) function and discover that the maximum length for the Madurese language is 68. The maximum length for the Indonesian language is 75.



```
Madurese Vocabulary Size: 2595
Madurese Max Length: 68
Indonesia Vocabulary Size: 2600
Indonesia Max Length: 75
Model: "sequential"
_________________________________________________________________
 Layer (type)                    Output Shape              Param #
=================================================================
 embedding (Embedding)           (None, 75, 128)           332800

 lstm (LSTM)                     (None, 128)               131584

 dropout (Dropout)               (None, 128)               0

 repeat_vector (RepeatVector)    (None, 68, 128)           0

 lstm_1 (LSTM)                   (None, 68, 128)           131584

 dropout_1 (Dropout)             (None, 68, 128)           0

 time_distributed (TimeDistributed)  (None, 68, 2595)      334755

=================================================================
Total params: 930,723
Trainable params: 930,723
Non-trainable params: 0
```

Figure 5. Sequential Model for Madurese – Indonesian Machine Translation

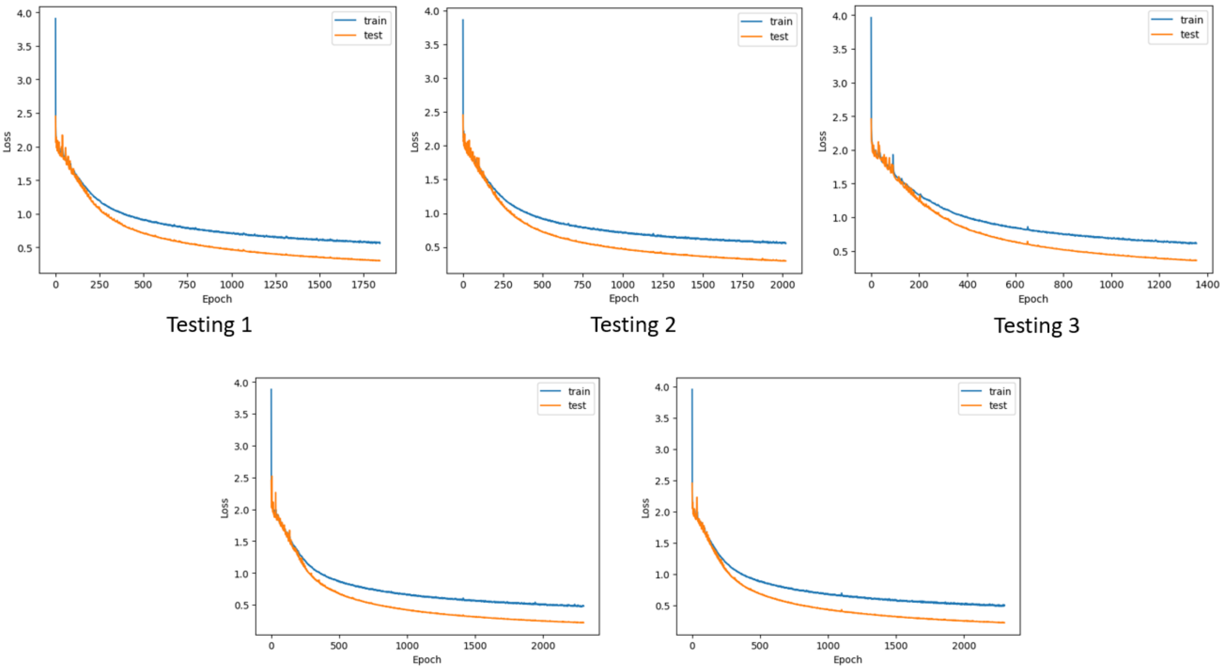
The training data was then separated into trainX (Madurese) and trainY (Indonesian). One-hot encoding is used for predicting each word in the vocabulary as output. This study fine-tuned the settings by incorporating an embedding layer for Madurese sentence length as input and Indonesian vocabulary size as output. Softmax was employed as the activation function, and in addition to 128 layers in the training phase, this study included a Dropout layer with a value of 0.2 to avoid overfitting. This study set the epoch parameter at 10,000, although in five training runs, the most extended epoch was around 2300 because this study believe that batch size and learning speed have a high correlation to achieve good performance. This study summarize the parameters for the architecture model this study used in Table 3.

Table 3. Parameter for the model architecture

Parameter	Value
Optimizer	Adam
Loss	Categorical Crossentropy
Dropout	0.2
Activation	Softmax
Callback	Early Stopping
Epoch	10000

3.3. Evaluation

The BLEU metric is used in the model assessment procedure to evaluate and measure how well the output matches the phrase length. This study tried to experiment 5 times with the results shown in Figure 6 and Table 4 for the BLEU metric score results.



Testing 1

Testing 2

Testing 3

Testing 4

Testing 5

Figure 6. Graphic model from test results

this study obtained the evaluation results from the five tests this study conducted with the BLEU matrix shown in Table 4. This score is calculated by comparing a candidate text translation against one or more reference translations. To double-check our translation, this study reload our datasets and only assess on trainX and testX. The average evaluation result for BLEU-1 is 0.798068, BLEU-2 is 0.680932, BLEU-3 is 0.623489, and BLEU-4 is 0.523546.

Table 4. Evaluating the model with BLEU Score.

Metric	Testing 1 Epoch: 1820	Testing 2 Epoch: 2021	Testing 3 Epoch: 1355	Testing 4 Epoch: 2300	Testing 5 Epoch: 2299	Average
BLEU-1	0.776481	0.787097	0.738298	0.844526	0.843939	0.798068
BLEU-2	0.647631	0.662041	0.591451	0.754327	0.749208	0.680932
BLEU-3	0.583719	0.601285	0.521463	0.711423	0.699556	0.623489
BLEU-4	0.475525	0.496946	0.406568	0.629942	0.608751	0.523546

4. Conclusion

The effectiveness of a Neural Machine Translation (NMT) approach utilizing Long Short-Term Memory (LSTM) for translating Madurese into Indonesian has been demonstrated. The LSTM encoder-decoder model was employed to compute BLEU scores, yielding the highest average of 80% for BLEU-1 and 52% for BLEU-4. Considering the inherent linguistic structure differences between Madurese and Indonesian, the proposed approach emerges as the optimal resolution for machine translation from Indonesian to Madurese.

One potential weakness of this study is the omission of an analysis of the disparities in sentence structure between Indonesian and Madurese languages and the potential variability in interpreting individual words or phrases. The Indonesian translation of the Bible employs active sentence structures, whereas the Madurese translation employs passive sentence structures.

In further investigations, this study suggest examining the disparities in phrase structure and ambiguity variables to ascertain their potential impact on the precision of translation outcomes. In addition, it is vital to contemplate the augmentation of the dataset and the exploration of alternative neural machine translation methodologies to enhance the translation's computational processes and precision. As an innovative methodology, this study propose incorporating Named Entity Recognition (NER) within the Neural Machine Translation (NMT) framework in this forthcoming study.

References

[1] A. F. Aji *et al.*, ‘One Country, 700+ languages: NLP challenges for underrepresented languages and dialects in Indonesia’, *ArXiv Prepr. ArXiv220313357*, 2022.

[2] A. Hermanto, T. B. Adji, and N. A. Setiawan, ‘Recurrent neural network language model for English-Indonesian Machine Translation: Experimental study’, in *2015 International conference on science in information technology (ICSITech)*, IEEE, 2015, pp. 132–136.

[3] Z. Abidin, A. Sucipto, and A. Budiman, ‘Penerjemahan Kalimat Bahasa Lampung-Indonesia Dengan Pendekatan Neural Machine Translation Berbasis Attention Translation of Sentence Lampung-Indonesian Languages With Neural Machine Translation Attention Based’, *J Kelitbangan*, vol. 6, no. 02, pp. 191–206, 2018.

[4] Q. A. Agigi, ‘Language Statistical Machine Translation Muna to Indonesia Language’, *JATISI J. Tek. Inform. Dan Sist. Inf.*, vol. 8, no. 4, pp. 2173–2186, 2021.

[5] A. E. P. Lesatari, A. Ardiyanti, and I. Asror, ‘Phrase Based Statistical Machine Translation Javanese-Indonesian’, *J. MEDIA Inform. BUDIDARMA*, vol. 5, no. 2, pp. 378–386, 2021.

[6] A. P. Wibawa, A. Nafalski, and W. F. Mahmudy, ‘Javanese speech levels machine translation: improved parallel text alignment based on impossible pair limitation’, in *2013 IEEE International Conference on Computational Intelligence and Cybernetics (CYBERNETICSCOM)*, IEEE, 2013, pp. 16–20.

[7] M. S. Alam and A. A. Suryani, ‘Minang and Indonesian Pharase-Based Statistical Machine Translation’, *J. Inform. TELECOMMUNICATION Eng.*, vol. 5, no. 1, pp. 216–224, 2021.

[8] I. G. B. A. Budaya, M. W. A. Kesiman, and I. M. G. Sunarya, ‘The Influence of Word Vectorization for Kawi Language to Indonesian Language Neural Machine Translation’, *J. Inf. Technol. Comput. Sci.*, vol. 7, no. 1, pp. 81–93, 2022.

[9] F. H. Rachman, N. Ifada, S. Wahyuni, G. D. Ramadani, and A. Pawitra, ‘ModifiedECS (mECS) Algorithm for Madurese-Indonesian Rule-Based Machine Translation’, in *2022 International Conference of Science and Information Technology in Smart Administration (ICSINTESA)*, IEEE, 2022, pp. 51–56.

[10] M. Yang, R. Wang, K. Chen, X. Wang, T. Zhao, and M. Zhang, ‘A Novel Sentence-Level Agreement Architecture for Neural Machine Translation’, *IEEEACM Trans. Audio Speech Lang. Process.*, vol. 28, pp. 2585–2597, 2020, doi: 10.1109/TASLP.2020.3021347.

[11] V. Macketanz, E. Avramidis, A. Burchardt, J. Helcl, and A. Srivastava, ‘Machine Translation: Phrase-Based, Rule-Based and Neural Approaches with Linguistic Evaluation’, *Cybern. Inf. Technol.*, vol. 17, no. 2, pp. 28–43, Jun. 2017, doi: 10.1515/cait-2017-0014.

[12] D. Puspitaningrum, ‘A Study of English-Indonesian Neural Machine Translation with Attention (Seq2Seq, ConvSeq2Seq, RNN, and MHA) A Comparative Study of NMT on English-Indonesian’, in *6th International Conference on Sustainable Information Engineering and Technology 2021*, 2021, pp. 271–280.

[13] I. Sutskever, O. Vinyals, and Q. V. Le, ‘Sequence to Sequence Learning with Neural Networks’. arXiv, Dec. 14, 2014. doi: 10.48550/arXiv.1409.3215.

[14] G. Van Houdt, C. Mosquera, and G. Nápoles, ‘A review on the long short-term memory model’, *Artif. Intell. Rev.*, vol. 53, pp. 5929–5955, 2020.

[15] F. Stahlberg, ‘Neural Machine Translation: A Review’, *J. Artif. Intell. Res.*, vol. 69, pp. 343–418, Oct. 2020, doi: 10.1613/jair.1.12007.

[16] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, ‘Bleu: a method for automatic evaluation of machine translation’, in *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 2002, pp. 311–318.

[17] Y. Fan, F. Tian, Y. Xia, T. Qin, X.-Y. Li, and T.-Y. Liu, ‘Searching Better Architectures for Neural Machine Translation’, *IEEEACM Trans. Audio Speech Lang. Process.*, vol. 28, pp. 1574–1585, 2020, doi: 10.1109/TASLP.2020.2995270.

[18] A. Pranolo, Y. Mao, A. P. Wibawa, A. B. P. Utama, and F. A. Dwiyanto, ‘Robust LSTM With tuned-PSO and bifold-attention mechanism for analyzing multivariate time-series’, *IEEE Access*, vol. 10, pp. 78423–78434, 2022.

[19] A. P. Wibawa, I. T. Saputra, A. B. P. Utama, W. Lestari, and Z. N. Izdihar, ‘Long Short-Term Memory to Predict Unique Visitors of an Electronic Journal’, in *2020 6th International Conference on Science in Information Technology (ICSITech)*, IEEE, 2020, pp. 176–179.

[20] K. Dedes, A. B. P. Utama, A. P. Wibawa, A. N. Afandi, A. N. Handayani, and L. Hernandez, ‘Neural Machine Translation of Spanish-English Food Recipes Using LSTM’, *JOIV Int. J. Inform. Vis.*, vol. 6, no. 2, pp. 290–297, 2022.

[21] S. Edunov, A. Baevski, and M. Auli, ‘Pre-trained language model representations for language generation’, *ArXiv Prepr. ArXiv190309722*, 2019.

[22] K. B. Prakash, Y. V. R. Nagapawan, N. L. Kalyani, and V. P. Kumar, ‘Chatterbot implementation using transfer learning and LSTM encoder-decoder architecture’, *Int. J.*, vol. 8, no. 5, 2020.

[23] D. E. Cahyani, A. P. Wibawa, D. D. Prasetya, L. Gumilar, F. Akhbar, and E. R. Triyulinar, ‘Text-Based Emotion Detection using CNN-BiLSTM’, in *2022 4th International Conference on Cybernetics and Intelligent System (ICORIS)*, IEEE, 2022, pp. 1–5.

[24] D. P. Kingma and J. Ba, ‘Adam: A method for stochastic optimization’, *ArXiv Prepr. ArXiv14126980*, 2014.

[25] R. Bawden, B. Zhang, L. Yankovskaya, A. Tättar, and M. Post, ‘A study in improving BLEU reference coverage with diverse automatic paraphrasing’, *ArXiv Prepr. ArXiv200414989*, 2020.

[26] P. Zaremoodi and G. Haffari, ‘Neural machine translation for bilingually scarce scenarios: a deep multi-task learning approach’, *ArXiv Prepr. ArXiv180504237*, 2018.

[27] K. Kim, ‘New canonical decomposition and composition processes for Hangeul’, *Comput. Stand. Interfaces*, vol. 24, no. 1, pp. 69–82, 2002.

[28] A. Rai and S. Borah, ‘Study of various methods for tokenization’, in *Applications of Internet of Things: Proceedings of ICCCIOT 2020*, Springer, 2021, pp. 193–200.

[29] G. V. A. Gutiérrez, ‘A Comparative Study of NLP and Machine Learning Techniques for Sentiment Analysis and Topic Modeling on Amazon’, *Int J Comput Sci Eng*, vol. 9, no. 2, pp. 159–170, 2020.

[30] M. IŞIK and H. Dağ, ‘The impact of text preprocessing on the prediction of review ratings’, *Turk. J. Electr. Eng. Comput. Sci.*, vol. 28, no. 3, pp. 1405–1421, 2020.