Journal of Applied Informatics and Computing (JAIC)
Vol.9, No.6, December 2025, pp. 3588~3593
e-ISSN: 2548-6861

A Comparative Analysis of Character and Word-Based Tokenization for Kawi-Indonesian Neural Machine Translation

I Gede Bintang Arya Budaya[1*], I Gede Putra Mas Yusadara[2]**
* Information Technology, Institute of Technology and Business STIKOM Bali
**Information System, Institute of Technology and Business STIKOM Bali
bintang@stikom-bali.ac.id[1], putramas@stikom-bali.ac.id[2]

Article Info

Article history:

Received 2025-09-21
Revised 2025-10-31
Accepted 2025-11-08

Keyword:

Transformer,
Low Resources Language,
Computational Linguistic,
FLAN-T5.

ABSTRACT

Preserving regional languages is a strategic step in preserving cultural heritage while expanding access to knowledge across generations. One approach that can support this effort is the application of automatic translation technology to digitize and learn local language texts. This study compares two tokenization strategies, word-based and character-based on a Kawi–Indonesian translation model using the FLAN-T5-Small Transformer architecture. The dataset used consists of 4,987 preprocessed sentence pairs, trained for 10 epochs with a batch size of 8. Statistical analysis shows that Kawi texts have an average length of 39.6 characters (5.4 words) per sentence, while Indonesian texts have an average length of 54.9 characters (7.5 words). These findings suggest that Kawi sentences tend to be lexically dense, with low word repetition and high morphological variation, which can increase the learning complexity of the model. Evaluation using BLEU and METEOR metrics shows that the model with word-based tokenization achieved a BLEU score of 0.45 and a METEOR score of 0.05, while the character-based model achieved a BLEU score of 0.24 and a METEOR score of 0.04. Although the dataset size has increased compared to previous studies, these results indicate that the additional data is not sufficient to overcome the limitations of the semantic representation of the Kawi language. Therefore, this study serves as an initial baseline that can be further developed through subword tokenization approaches, dataset expansion, and training strategy optimization to improve the quality of local language translations in the future.

I. INTRODUCTION

Indonesia's linguistic diversity represents a priceless cultural heritage that now faces serious preservation challenges. Many local languages, which represent history and local wisdom, are experiencing a decline in active speakers and a decline in the interest of the younger generation in learning them formally. This challenge is not caused by a lack of preservation efforts [1], but rather by the method of presenting learning resources that has not fully adapted to the digital era. Today, modern society relies heavily on information technology, such as machine translation, as the primary medium for language learning and interaction. [2]. However, efforts to create this kind of technology for local languages are hampered by a fundamental problem, local languages are classified as low-resource languages due to the minimal availability of structured data sources that can be used to develop machine translation [3].

In recent years, much research has been conducted to explore various approaches to building Neural Machine Translation (NMT) systems for low-resource languages. The main challenge faced is the limited availability of adequate parallel data, leading to the emergence of various strategies such as transfer learning, data augmentation, and model and architecture tuning [4], [5]. One widely adopted approach is the use of the Transformer architecture, introduced by Vaswani et al [6]. This architecture has become the modern

standard in natural language processing due to its superior ability to capture complex contextual relationships in text.

Kawi is one of the local languages in Indonesia. It holds high academic and cultural significance as the primary medium for various texts on customary law, philosophy, traditional medicine, and classical literature in Java and Bali [7]. The Kawi language was chosen as the research object due to its significant historical and linguistic significance within the Javanese-Balinese cultural heritage. As an ancient language used in inscriptions and classical manuscripts from the 9th to 16th centuries AD, Kawi serves as a bridge between the development of Sanskrit, Old Javanese, and Balinese [8]. Although many scholars have attempted to translate Kawi into modern languages such as Indonesian for the purpose of disseminating knowledge to the general public, these approaches remain manual and fragmentary. A more fundamental challenge is the lack of structured representations that allow Kawi to be systematically processed by machines. Within the framework of natural language research, this requires the development of computational models that serve not only as translation tools but also as digital preservation tools.

Linguistically, Kawi has a complex morphological system that cause one root word to have many variations in form. The sentence structure is also free and does not always follow pattern like Indonesian. This morphological complexity makes the process of segmenting and tokenizing Kawi texts more challenging because the boundaries between words and morphemes are not always explicit. Research [9] demonstrated this through the development of the E-Translator Kawi to Bahasa based on the Nazief & Adriani algorithm, which emphasized the importance of morphological rules in the process of tokenizing and stemming Kawi. This finding confirms that Kawi requires a careful processing approach to word structure and context. However, this approach did not utilize deep learning to dynamically capture semantic context.

In line with the development of NMT, particularly with the emergence of the Transformer architecture, which has proven superior in modelling long-term dependencies, research opportunities for low-resource languages are increasingly open [10], [11]. Previous research on NMT has generally focused on RNN, LSTM, or GRU architectures for Kawi [12], [13], but the results are still limited both in terms of performance and the availability of parallel corpus. To date, there has been no study that systematically evaluates tokenization strategies for Kawi using the advantages of Transformer. Therefore, this study offers a contribution by conducting a comparative analysis of two tokenization strategies based on characters and words in Kawi-Indonesian translation using the Transformer model. The findings of this study are expected to provide a baseline for the development of NMT in Kawi and support the agenda of preserving the Indonesian literary heritage through more effective digital representation.

II. METHOD

In this research, the research method used is as shown in Figure 1

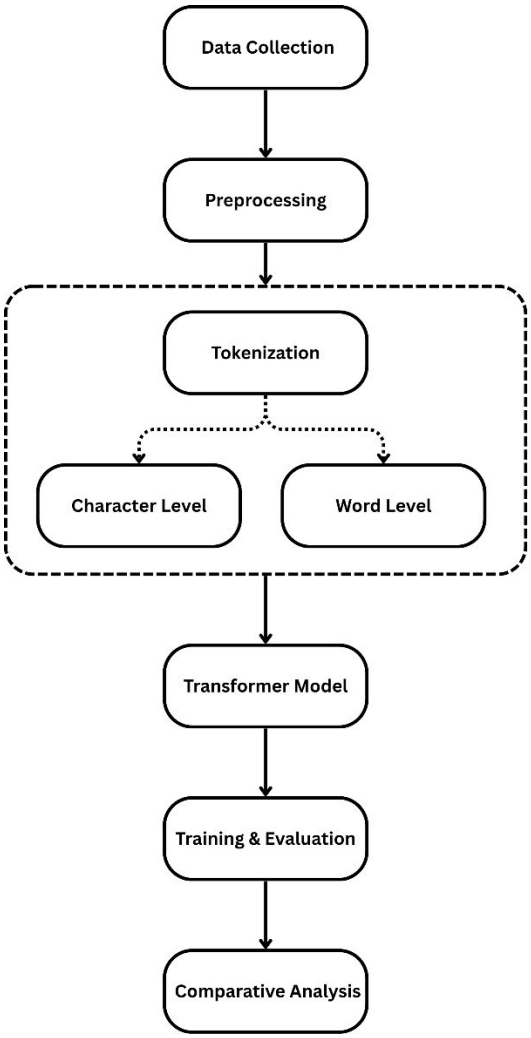

Figure 1. Research method

A. Data Collection

At the data collection stage, the researcher utilized a dataset from a previous study [12], [13] consisting of 1,088 rows and subsequently increased it with newly collected data from *Saramuscaya scripture* in the Kawi language, translated into Indonesian by the Ministry of Religious Affairs of the Republic of Indonesia [14]. Figure 2 illustrates the procedure used to separate the data sources in order to establish the ground truth from Kawi to Indonesian, while Table 1 presents the final number of data instances successfully compiled.

TABEL I
NUMBER OF DATASET

No	Sources	Number of Rows
1	Budaya et al [12], [13]	1,088 rows
2	*Sarasamuccaya scripture* [14]	3,899 rows
Total		4,987 rows

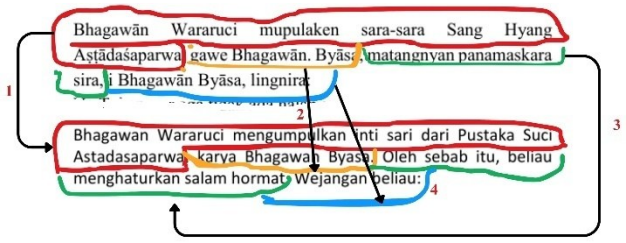

Figure 2. Separating the data sources into four Kawi–Indonesian sentence pairs as ground truth.

B. Preprocessing

In the preprocessing stage, several steps were carried out for both Kawi and Indonesian sentences. First, punctuation marks, numbers, and excessive spaces were removed. Second, all characters were converted to lowercase. Third, special characters were standardized, for example, converting the letter "à" to "a" This third step specifically aims to ensure that characters or symbols with the same reading meaning are not treated as different entities. Table 2 and 3 presents the results of the preprocessing applied to dataset.

TABEL II
PREPROCESSING RESULTS ON KAWI LANGUAGE

No	Before	After
1	*Om Awighnamàstu.*	*om awighnamastu*
2	*syapa ta sira bràhmana sumàpa Sang Wrêsnyàndha...*	*syapa ta sira brahmana sumapa sang wresnyandha*
…	...	...
4897	*Sumahur Bhagawàn Wèùampayana:*	*sumahur bhagawan weuampayana*

TABEL III
PREPROCESSING RESULTS ON INNDONESIAN LANGUAGE

No	Before	After
1	*Oh Hyang Widhi semoga tidak terhalang.*	*oh hyang widhi semoga tidak terhalang*
2	*siapakah itu pendeta agung yang mengutuk Sang Wrêsnyàndha...*	*siapakah itu pendeta agung yang mengutuk sang wresnyandha*
…	...	...
4897	*Lalu dijawab oleh Bhagawàn Wèsampayana.*	*lalu dijawab oleh Bhagawan wesampayana*

C. Tokenization

Following data preparation, two tokenization strategies were employed, tokenization was performed using the built-in tokenizer of the *SentencePiece*-based Flan-T5 model, without modifications to the underlying algorithm, two input variations were used word-level and character-level. Word-level tokenization segments the text based on whitespace, whereas character-level tokenization decomposes the text into individual characters. Table 4 shows the sample result of tokenization

TABEL IV
SAMPLE OF TOKENIZATION RESULTS

No	Category	Sentence	Result
1	*Word-Level*	*oh hyang widhi semoga tidak terhalang*	*[“oh”, “hyang”, “widhi”, “semoga”, “tidak”, “terhalang”]*
2	*Character Level*	*oh hyang widhi semoga tidak terhalang*	*[“o”, “h”, “ ”, “h”, “y”, “a”, “n”, “g”, “ ”, “w”, “i”, “d”, “h”, “i”, “ ”, “s”, “e”, “m”, “o”, “g”, “a”, “ ”, “t”, “i”, “d”, “a”, “k”, “ ”, “t”, “e”, “r”, “h”, “a”, “l”, “a”, “n”, “g”]*

D. Transformer Model

The next step is to configure the Transformer-based model. This study used Flan-T5-small as the primary model because it is relatively lightweight, easy to train, and suitable for experiments in low-resource language scenarios [15] .

E. Training and Evaluation

The model was trained using the Flan-T5-Small built-in tokenizer on two tokenization scenarios (word-level and character-level). Training was performed on a Google Collaboratory GPU with an 80:20 data split, using the Adam optimizer (learning rate 5e-5), 10 epochs, and a batch size of 8 as the experimental baseline. Model performance was evaluated using two main metrics, the BLEU score to measure translation accuracy [16] and the METEOR score to assess the semantic correspondence of the translation to the reference text [17]. Model evaluation was performed using the Hugging Face Evaluate library, namely SacreBLEU and METEOR.

F. Comparative Analysis

The evaluation results were comparatively analyzed by examining the BLEU and METEOR scores for each model and tokenization strategy, as both metrics are widely used for assessing translation performance [18], [19]. Performance differences were visualized in the form of bar charts to clearly demonstrate the strengths and weaknesses of each approach. In this study, statistical analysis of the dataset was also carried out with the aim of understanding the characteristics of the dataset, especially regarding the word density and meaning of

the source and target languages. Through this analysis, the research is expected to provide a deeper understanding of the effectiveness of word- and character-based tokenization in Kawi–Indonesian translation using the Transformer-based NMT approach.

III. Result and Discussion

This section presents the experimental results of implementing the FLAN-T5-Small model for developing a machine translation model from Kawi to Indonesian. In this study, the model was initially trained for 10 epochs with a batch size of 8, and its performance was observed in terms of training loss and validation loss. The model's performance was further evaluated using two metrics, BLEU score and Meteor.

First, the model was trained using word-level tokenization. Based on the training loss and validation loss graph, it can be observed that from epoch 1 to epoch 2, training loss decreased significantly from 4.726 to 0.752, indicating that the model was learning effectively. From epoch 2 to epoch 6, the training and validation loss gradually decreased to 0.565, and from epoch 6 to epoch 10, the loss values decreased and stabilized with no significant changes, reaching 0.526. For word-level validation loss, the trend was consistent with the word-level training loss, decreasing from 0.607 at epoch 1 to 0.271 at epoch 10. Figure 3 shows the visualization of the training process for word-level tokenization.

Second, the model was trained using character-level tokenization. Overall, the trend was similar to that of the word-level model. From epoch 1 to epoch 2, the training loss decreased from 3.990 to 0.424. Between epoch 2 and epoch 10, it continued to decline gradually until reaching 0.279. The validation loss for the character-level model exhibited the same trend, decreasing from 0.607 at epoch 1 to 0.279 at epoch 10. Figure 4 shows the visualization of the training process for character-level tokenization.

Based on the evaluation results, the model trained with word-based tokenization only achieved a BLEU score of 0.45, while the character-based model produced an even lower score. This very low BLEU score indicates that the resulting translation quality is still far from the reference sentence and tends to be close to random predictions. The METEOR score obtained is also very low, indicating that the lexical and word order match between the translated results and the reference is still limited. These results illustrate that, with the current configuration, a relatively small dataset size, word/character-based tokenization without subwords, and a limited number of training epochs the model is not yet able to achieve adequate translation performance. Nevertheless, these results can be used as an initial baseline that can be replicated and developed in further research with a more optimal approach. Figure 5 shows the visualization of both metrics and Table 5 shows the sample of translation.

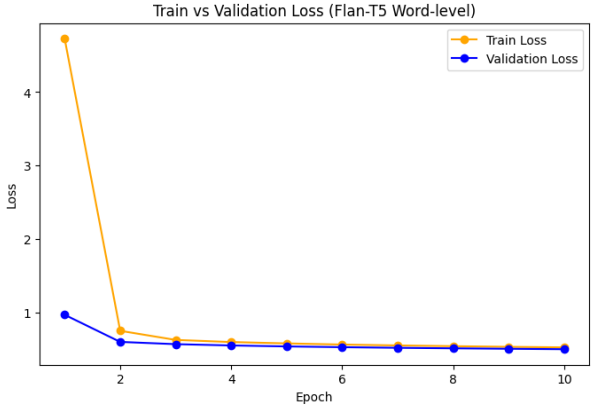

Figure 3. Word level training loss graph

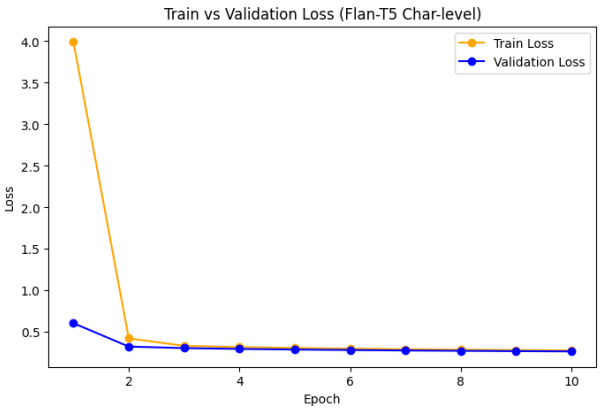

Figure 4. Character level training loss graph

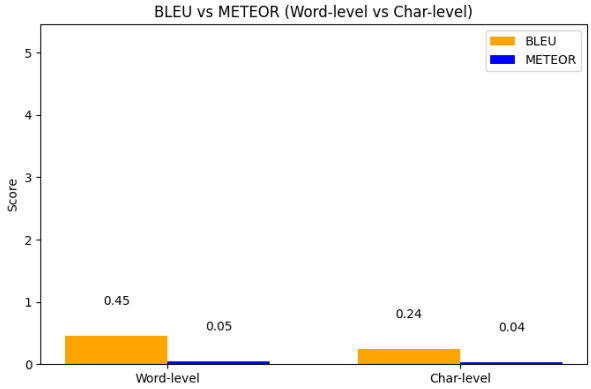

Figure 5. Result of BLEU score and METEOR metrics

TABEL V
TRANSLATION SAMPLE

No	Ground Truth Kawi	Ground Truth Indonesian	Word-Level Result	Char-Level Result
1	*an ikang pehan juga kapangan denya*	*hanya susunya saja diminum olehnya*	*demikian adalah orang yang tidak adalah adalah yang tidak adalah*	*denya denya denya*
2	*magawayang yajna*	*membuat upacara*	*demikian yajna*	*magada yajna*
3	*tar lenok.*	*ia tidak bohong*	*tar lenok*	*tar lenok*

Figures 6 and 7 show the distribution of characters and sentences in the dataset, while Table 6 presents the statistical analysis of the dataset. Statistical analysis shows that sentences in the Kawi corpus have an average length of 39.64 characters, while the equivalent sentences in Indonesian have an average length of 54.85 characters, with a larger standard deviation (31.47) in Indonesian compared to Kawi (19.40). This difference indicates that Indonesian sentences tend to be longer and more varied in character count. This is understandable, as the process of translating from Kawi to Indonesian often requires the addition of auxiliary words or particles to clarify the meaning implicit in a single Kawi word. In other words, Indonesian is more expansive, while Kawi is more morphologically dense.

This morphological density can also be explained linguistically, a single Kawi morpheme can include a combination of a root and an affix that carry both semantic meaning and grammatical function. This makes Kawi text representations appear shorter in character, yet retain a high degree of syntactic complexity. Furthermore, the maximum character range (up to 161 characters for Kawi and 256 characters for Indonesian) indicates extreme length variation. This variation can potentially cause difficulties in padding and sequence alignment during model training, especially for character-based models.

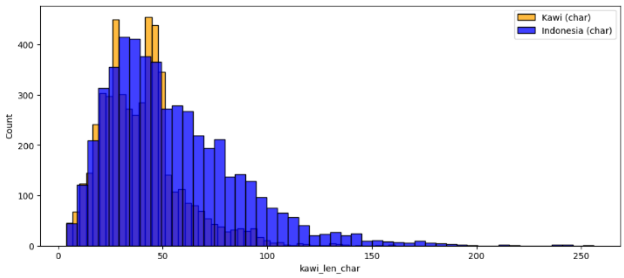

Figure 6. Distribution of Sentence Length Based on the Number of Characters in Kawi and Indonesian

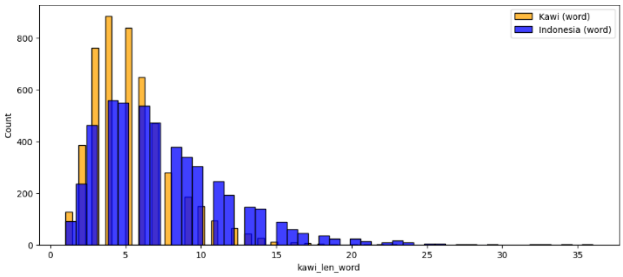

Figure 7. Distribution of Sentence Length Based on Number of Words in Kawi and Indonesian

TABEL VI
SENTENCE LENGTH STATISTICS OF THE KAWI–INDONESIAN CORPUS

Language	Mean Sentence Length (Characters)	Standard Deviation	Mean Sentence Length (Words)	Standard Deviation
Kawi	39.64	19.40	5.38	2.74
Indonesian	**54.85**	**31.47**	**7.53**	**4.37**

In terms of word count, Kawi sentences average around 5.38 words per sentence, while their Indonesian equivalents reach 7.53 words per sentence. The higher standard deviation value in Indonesian indicates a wider variety of sentence structures. This difference indicates that the translation process from Kawi to Indonesian tends to involve lexical expansion, namely the addition of explanatory words or phrases to clarify meanings that are often implicit in Kawi texts. Thus, Kawi sentences are relatively denser morphologically, while Indonesian sentences are more explicit and descriptive. This phenomenon is common in language pairs with different levels of morphological complexity, such as classical and modern languages.

Differences in sentence structure between Kawi and Indonesian influence the effectiveness of tokenization strategies. Word-based tokenization is more stable because each token represents a complete unit of meaning, making it easier for the model to recognize lexical matches and yield higher BLEU and METEOR scores. In contrast, character-based tokenization produces much longer sequences and is prone to losing semantic context, especially in small datasets. Each letter is processed as a separate token, increasing the computational burden and making contextual patterns more difficult to learn. Therefore, under limited datasets, word-based approaches prove more efficient and contextual than character-based approaches.

Although the dataset has increased compared to the previous study [12], [13], this expansion does not necessarily guarantee improved model performance. The Kawi–Indonesian dataset exhibits low word repetition and high lexical diversity, indicating that nearly every token carries a unique meaning. This enriches semantic representation, but also makes it difficult for models to find consistent patterns for generalization. The high morphological complexity of Kawi exacerbates this challenge, as a single word can represent multiple meanings depending on the context. In this situation, models with word-based tokenization demonstrate more stable performance than character-based models, especially

when the data size is limited. Overall, the linguistic and statistical characteristics of this dataset are important factors influencing the effectiveness of translation models, highlighting the need for more adaptive tokenization and data augmentation strategies in future research.

IV. CONCLUSION

This study compares the performance of the FLAN-T5-Small model using two tokenization strategies, word-based and character-based for Kawi to Indonesian translation. Results show that the word-based model yields a higher BLEU score (0.45) compared to the character-based model (0.24), although both still yield low overall accuracy. The low METEOR scores for both approaches indicate that the models are not yet able to optimally capture lexical equivalence and sentence structure. Analysis of dataset characteristics indicates that differences in lexical density and sentence length between languages are factors affecting model performance. Although the amount of data has increased compared to previous studies, this addition does not necessarily result in improved performance, given the high morphological complexity and lexical diversity of Kawi. Therefore, the results of this study serve as an initial baseline that can be further developed through dataset expansion, the application of subword-based tokenization, and optimization of model architecture and training strategies to achieve better translation quality.

ACKNOWLEDGEMENT

We would like to express our gratitude to the Ministry of Higher Education, Science, and Technology of the Republic of Indonesia for the financial support provided through the *Penelitian Dosen Pemula (PDP)* / Beginner Lecturer Research Grant Scheme in 2025.

REFERENCES

[1] A. Hidayat, "Revitalization of ancient Indonesian characters and the maintenance efforts in past 10 years," *LADU: Journal of Languages and Education*, vol. 1, no. 4, pp. 179–186, May 2021, doi: 10.56724/ladu.v1i4.69.

[2] Y. Wang, "Cognitive and sociocultural dynamics of self-regulated use of machine translation and generative AI tools in academic EFL writing," *System*, vol. 126, p. 103505, Nov. 2024, doi: 10.1016/j.system.2024.103505.

[3] P. Koehn and R. Knowles, "Six Challenges for Neural Machine Translation," in *Proceedings of the First Workshop on Neural Machine Translation*, T. Luong, A. Birch, G. Neubig, and A. Finch, Eds., Vancouver: Association for Computational Linguistics, Aug. 2017, pp. 28–39. doi: 10.18653/v1/W17-3204.

[4] J. Wang, L. Yang, J. Wang, Y. Guan, L. Bai, and H. Luo, "A data-guided curriculum towards low-resource neural machine translation," *Expert Systems with Applications*, vol. 283, p. 127673, July 2025, doi: 10.1016/j.eswa.2025.127673.

[5] M. E. Pacheco Martínez, M. Carrillo Ruiz, and M. de L. Sandoval Solís, "Harmony search for hyperparameters optimization of a low resource language transformer model trained with a novel parallel corpus Ocelotl Nahuatl – Spanish," *Systems and Soft Computing*, vol. 6, p. 200152, Dec. 2024, doi: 10.1016/j.sasc.2024.200152.

[6] A. Vaswani *et al.*, "Attention is All you Need," in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2017. Accessed: Sept. 15, 2025. [Online]. Available: https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

[7] P. Susongko and I. Rosdiana, "Educational Theory Based on Ancient Javanese Philosophy".

[8] M. S. Zurbuchen, "Kawi--an Introduction," in *Introduction to Old Javanese Language and Literature*, in A Kawi Prose Anthology. , University of Michigan Press, 1976, pp. 1–12. Accessed: Oct. 30, 2025. [Online]. Available: https://www.jstor.org/stable/10.3998/mpub.11902952.7

[9] O. Sudana, D. Putra, M. Sudarma, R. S. Hartati, R. P. Prastika, and A. Wirdiani, "E-Translator Kawi to Bahasa," *MATEC Web of Conferences*, vol. 159, p. 01047, 2018, doi: 10.1051/matecconf/201815901047.

[10] A. Qorbani, R. Ramezani, A. Baraani, and A. Kazemi, "Multilingual neural machine translation for low-resource languages by twinning important nodes," *Neurocomputing*, vol. 634, p. 129890, June 2025, doi: 10.1016/j.neucom.2025.129890.

[11] B. Li, Y. Weng, F. Xia, and H. Deng, "Towards better Chinese-centric neural machine translation for low-resource languages," *Computer Speech & Language*, vol. 84, p. 101566, Mar. 2024, doi: 10.1016/j.csl.2023.101566.

[12] 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," *Journal of Information Technology and Computer Science*, vol. 7, no. 1, pp. 81–93, Sept. 2022, doi: 10.25126/jitecs.202271387.

[13] I. G. B. A. Budaya, M. W. A. Kesiman, and I. M. G. Sunarya, "Perancangan Mesin Translasi berbasis Neural dari Bahasa Kawi ke dalam Bahasa Indonesia menggunakan Microframework Flask," *Jurnal Sistem dan Informatika (JSI)*, vol. 16, no. 2, pp. 94–103, June 2022, doi: 10.30864/jsi.v16i2.440.

[14] T. P. D. Penerjemah, P. S. Veda, and V. Samiti, "Direktorat Jenderal Bimbingan Masyarakat Hindu Kementerian Agama Republik Indonesia 202".

[15] S. Longpre *et al.*, "The Flan Collection: Designing Data and Methods for Effective Instruction Tuning," in *Proceedings of the 40th International Conference on Machine Learning*, PMLR, July 2023, pp. 22631–22648. Accessed: Sept. 19, 2025. [Online]. Available: https://proceedings.mlr.press/v202/longpre23a.html

[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 on Association for Computational Linguistics - ACL '02*, Philadelphia, Pennsylvania: Association for Computational Linguistics, 2001, p. 311. doi: 10.3115/1073083.1073135.

[17] S. Banerjee and A. Lavie, "METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments," in *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, J. Goldstein, A. Lavie, C.-Y. Lin, and C. Voss, Eds., Ann Arbor, Michigan: Association for Computational Linguistics, June 2005, pp. 65–72. Accessed: Sept. 19, 2025. [Online]. Available: https://aclanthology.org/W05-0909/

[18] "The BLEU Score for Automatic Evaluation of English to Bangla NMT | SpringerLink." Accessed: Sept. 19, 2025. [Online]. Available: https://link.springer.com/chapter/10.1007/978-981-33-4087-9_34

[19] H. Saadany and C. Orasan, "BLEU, METEOR, BERTScore: Evaluation of Metrics Performance in Assessing Critical Translation Errors in Sentiment-oriented Text," in *Proceedings of the Translation and Interpreting Technology Online Conference TRITON 2021*, 2021, pp. 48–56. doi: 10.26615/978-954-452-071-7_006.