Benchmarking Transformer Models for Low-Resource Language Translation: A Case Study on the Tegalan-Indonesian Language Pair

Dwi Intan Af'idah
Informatics Department, Harkat Negeri University, Tegal, Indonesia
dwiintanafidah@harkatnegeri.ac.id (corresponding author)

Sharfina Febbi Handayani
Informatics Department, Harkat Negeri University, Tegal, Indonesia
sharfina.handayani@harkatnegeri.ac.id

Ratri Wikaningtyas
Electronics Engineering Department, Harkat Negeri University, Tegal, Indonesia
ratriwikaningtyas@harkatnegeri.ac.id

Received: 18 November 2025 | Revised: 16 December 2025 | Accepted: 26 December 2025

 DOI: https://doi.org/10.48084/etasr.16348

ABSTRACT

This study investigates how well Transformer models can translate a low-resource local language, Tegalan, to Indonesian. Three Transformer-based models, mBART-50, mT5, and NLLB-200, were tested using a new parallel Tegalan-Indonesian dataset, collected from everyday conversations and online news texts. The translations were manually reviewed by native speakers and cultural experts. The preprocessing steps included case folding, spelling normalization, and subword tokenization to ensure consistency and handle dialect differences. Each model was fine-tuned under controlled conditions, with manual adjustment of hyperparameters. BLEU, METEOR, and TER were used to evaluate the models, offering insight into word-level accuracy, semantic alignment, and the required number of edits. The results show that NLLB-200 achieved the highest performance, with BLEU scores of 85.51, METEOR scores of 76.91, and TER scores of 17.73, clearly exceeding both mBART-50 and mT5. A qualitative review of the output indicated that NLLB-200 generated more natural and contextually appropriate translations. The results suggest that Transformer models can be a practical option for translation in low-resource environments and may also contribute to ongoing efforts to document and maintain regional languages. Further work is planned to enlarge the dataset and examine the extent to which semi-supervised techniques might strengthen both accuracy and overall model robustness.

Keywords-Low-Resource Language; Tegalan–Indonesian; Transformer Models; Neural Machine Translation

I. INTRODUCTION

Recent developments in digital communication have made it easier for people to exchange information and collaborate across regions. This growing connectivity has supported scientific, economic, and social activities at an international level [1]. Nevertheless, differences in language continue to limit cross-language communication, particularly in settings where digital interaction is essential. In an attempt to address this issue, research in Natural Language Processing (NLP) has increasingly focused on tools that can bridge communication gaps across languages [2]. Among the various NLP technologies, Machine Translation (MT) occupies a central position by providing automated methods for converting one language to another. Recent developments, through Neural Machine Translation (NMT), have achieved high accuracy for well-resourced language pairs, such as English-Mandarin, French-German, and English-Indonesian, improving access to knowledge, international trade, and scientific cooperation [3].

However, these technical advances have not been impartially beneficial to all languages. There are more than 7,000 languages spoken around the world, with many of them in danger of going extinct because of changes in society, globalization, and the fact that only a few major languages are used online [4]. Indonesia ranks among the most linguistically diverse countries globally, boasting over 700 regional languages. Some languages are seeing a decrease in active speakers and a deficiency in digital resources [5]. Therefore, inclusive research on regional languages is important to conduct, as it supports the preservation of linguistic heritage for languages with limited resources [6].

Tegalan refers to a distinctive Javanese language spoken in the northern coastal region of Central Java, including Tegal City, the Tegal Regency, the Brebes Regency, and the Pemalang Regency. Tegalan plays an important role in local cultural identity, but its presence is still rare in the digital world [7] and is rarely studied in computational linguistics. This limitation hinders the preservation and revitalization of the language [8]. Meanwhile, the development of NLP technology for the Tegalan language is a significant cultural need for its preservation and advancement in the digital era [9].

NMT can improve bilingual and multilingual resources, allowing translation of languages even with different writing systems [10]. Furthermore, NMT demonstrates the effectiveness of contemporary neural machine translation models, especially the Transformer architecture [11]. Transformers have proven to be highly effective in translating languages. Meanwhile, resources are determined by large and well-organized datasets [12]. However, this architecture requires a large parallel corpus containing millions of aligned sentence pairs to function optimally [13]. This dependency is a barrier for Low-Resource Languages (LRLs) due to the lack of quality digital parallel data [14].

In order to resolve this issue, several studies have developed NMT by applying the Transformer architecture. Several Transformer models that provide solutions to NMT problems with LRLs include mBART-50 (Multilingual Bidirectional and Auto-Regressive Transformers 50) [15], mT5 (Multilingual Text-to-Text Transfer Transformer) [16], and NLLB-200 (No Language Left Behind 200) [17]. In recent years, research on low-resource machine translation has expanded. However, the Tegalan-Indonesian translation has not yet been addressed in previous studies. In addition, no comparative evaluation of state-of-the-art Transformer models has been conducted for this language pair, limiting efforts in technological development and digital preservation.

The mBART-50 model, the mT5 model, and the NLLB-200 have been proven to perform exceptionally well for NMT tasks involving LRLs. Other studies have indicated that mBART50 is capable of providing significant improvements in translation quality for languages with low resources. Examples include research on MT for the English-Manipuri language [18], fine-tuning MT for French to Farsi [19], and research on low-resource MT in Vietnamese-Chinese [20]. Meanwhile, the mT5 model has also demonstrated satisfactory performance for MT with LRLs in previous studies [21, 22]. As for NLLB-200, many previous studies have also shown that it can improve translation quality in LRLs [23, 24].

A literature review reveals a gap, as, currently, there is no publicly accessible and scientifically validated Tegalan-Indonesian parallel corpus, which is crucial for a quality NMT system. Furthermore, no study has systematically compared the performance of Transformer-based multilingual models, including mBART-50, mT5, and NLLB-200, on any Indonesian regional language, such as Tegalan. Therefore, this research focused on creating the first NMT system for Tegalan-Indonesian.

This study focuses on three objectives, taking into account previous research. The primary goal is to determine which of the mBART-50, mT5, and NLLB-200 models is most effective for translation in low-resource settings, such as the Tegalan-Indonesian corpus. Three different but complementary methods are used to evaluate the overall performance of the system: BLEU, METEOR, and TER. In addition, a qualitative analysis was conducted to identify differences in error patterns between models in handling idiomatic expressions and typical linguistic structures.

This research also presents the first parallel Tegalan-Indonesian corpus and a specific NMT baseline model for this language pair. This corpus is used to perform a controlled evaluation of the performance of three high-level multilingual models, mBART-50, mT5, and NLLB-200. Based on the evaluation results, this study analyzed the types of errors in each model, improving the understanding of automated results. The novelty of this study lies in establishing the first controlled benchmarking of multilingual transformer models for the Tegalan-Indonesian language pair using a newly constructed low-resource corpus with varying levels of linguistic complexity.

II. MATERIALS AND METHODS

Figure 1 shows the implementation of a structured method in crafting translation models for languages with limited resources. Data coherence is the purpose for which every phase is planned to yield an optimal model. The very first step was to gather the datasets using a manual collection and translation of Tegalan language data. This way, a parallel corpus of the Tegalan and Indonesian languages was produced.

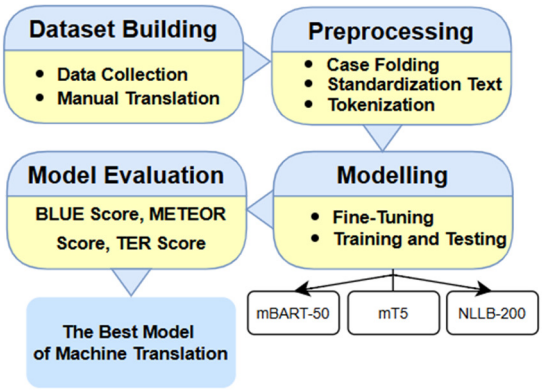

Fig. 1. The systematic procedure of this research.

After the dataset was collected, the next step was preprocessing, which involved case folding, spelling normalization, and tokenization. This step ensures that the data is accurate and consistent for model training. The processed data were then used in the NMT modeling phase, which includes training and evaluation of the transformer models and their enhancement. The mBART-50, NLLB-200, and mT5 were among the MT models employed in this procedure. The models were then evaluated using assessment metrics [25], namely BLEU, METEOR, and TER, to compare their performance.

A. Datasets

The Tegalan-Indonesian dataset was developed in the form of parallel sentence pairs that were manually aligned during the translation process. Each sentence was translated into the other language by four native speakers. At the time, the translations were stored in a separate file while preserving strict line-by-line correspondence. As a result, each line in the Indonesian file directly matched its Tegalan counterpart on the same line in the Tegalan file, forming a one-to-one sentence alignment. To ensure data quality, all sentence pairs were manually reviewed by two cultural experts of Tegalan.

This study employed three district datasets. The first dataset, referred to as Dataset A, consists of daily conversational data from Harsh Jain's study, publicly available on Kaggle [26]. This dataset contains paired conversational sentences in English and Indonesian. In this research, only the Indonesian text was used and subsequently translated into the Tegalan language by native speakers. The translated texts were then checked by cultural experts from Tegal to confirm that they kept the original meaning, were grammatically correct, and included local expressions in an appropriate way.

Original Tegalan language texts were also collected from the online news site www.kumparan.com, which formed the second dataset (Dataset B). The articles dealt with crime, politics, economics, or other local matters. All Tegalan language texts were manually translated by native speakers, who provided Tegalan-Indonesian sentence pairs. The third dataset, created from the combination of the two, incorporates the conversational context of the first dataset (Dataset A) along with the linguistic variation among informal and formal language found in the second dataset (Dataset B), named Dataset C.

Table I shows details on the three main Tegalan-Indonesian language datasets. Dataset A has 7,138 pairs of daily conversation utterances, informal in style and with an average length of five words per sentence. Dataset B contains 12,503 pairs of sentences written in a more formal narrative style with an average length of 14 words per sentence. Dataset C was created by merging the two previous datasets, resulting in 19,641 sentences. This dataset exhibits a high level of lexical diversity, containing over 33,000 unique words in Tegalan and roughly 28,000 unique words in Indonesian.

TABLE I. DATASET STATISTICS

Statistics	Dataset A	Dataset B	Dataset C
Number of Sentence Pairs	7,138	12,503	19,641
Average Sentence Length (Tegalan)	5.29	14.90	11.41
Average Sentence Length (Indonesian)	5.25	14.73	11.29
Vocabulary Size (Tegalan)	7,483	28,336	33,026
Vocabulary Size (Indonesian)	6,383	25,230	28,629

The dataset was not specifically created with in-depth linguistic analysis in mind, as this study focused on benchmarking for Tegalan-Indonesian MT rather than on detailed linguistic analysis. Therefore, idiomatic expressions and local cultural terms are included only when they naturally occur in the data, whereas honorifics, detailed morphological variation, and systematic code-mixed constructions are left for future work. Furthermore, regarding the nature of the writing in the dataset, all non-text elements (emojis, hashtags, and phonetic spellings) were removed. Thus, the short conversation dataset was cleaned up, while the news-based dataset consists of formal written text. The Tegalan-Indonesian parallel corpus created in this study is available upon reasonable request from the corresponding author.

B. Preprocessing and Tokenization

Before developing the Transformer model, preprocessing was performed to standardize the dataset. The goal of this phase was to bring about as much consistency as possible and, therefore, usability across datasets [27]. The first step, case-folding, was performed to transform all letters to lowercase. Case-folding was necessary because the text sources contained both highly formal and highly informal language. The second phase of the process was to standardize the text, maintaining the consistency of the Tegalan language without relying on spelling and dialectal variations. This process changed words to different forms or spellings, but still preserved their original meanings.

After the first step, there was further standardization of the text to normalize the writing and dialect differences among the phrase pairs in Tegalan. This stage aimed to maintain the same vocabulary and align the different spellings and forms of words to ensure that the dataset was treated in the same way at each point. After normalization, tokenization with subword segmentation was applied. This method breaks the sentence into smaller units, making them easier to work with [28]. Subword tokenization was performed using the SentencePiece algorithm, which is natively supported by the mBART-50, mT5, and NLLB-200 models. This is a technique used in language modeling and has been successful even for LRLs. The choice of subword tokenization was made based on its ability to handle language variations and yet let the model pick up the larger language patterns, thus improving its ability to generalize effectively [29].

After the tokenization step, 80% of the preprocessed dataset was allocated for training, with the other 20% was used for testing. This ratio was selected to maximize the number of samples for training the model while still providing a large enough holdout for performance testing. Given the small size of the Tegalan-Indonesian corpus, no separate validation set was created. Early stopping and checkpointing techniques were used in the fine-tuning process to control overfitting and ensure the model was at its best performance.

C. Model Architecture

This study performed a comparative analysis of various transformer models in Tegalan-Indonesian MT. Three models were selected, known for their outstanding capabilities for MT in LRLs: mBART-50, mT5, and NLLB-200. These models employ an encoder-decoder architecture that enables understanding the relationships between phrases within a sentence. However, each model has a different approach to developing NMT, especially in cases of limited datasets.

The first method used was mBART-50, which performs complete pre-training on the encoder and decoder simultaneously. Initially, the mBART model was trained in 25 languages, showing that it is capable of massively improving MT performance, mainly for LRLs [30]. Over time, mBART was developed with training in 50 languages. This was detailed in [15], hence the name mBART-50.

The second model utilized is mT5. Previous research introduced T5 as a model that allows various text analysis tasks to be performed within a single architecture. However, T5 was only trained in English [31]. Subsequently, T5 was developed into mT5 through large-scale pretraining in 101 languages, making it more inclusive of low-resource languages [16].

NLLB-200 was the third model used for machine translation. NLLB was first developed by META AI, and high performance was achieved by training it in hundreds of languages, including low-resource ones [32]. NLLB was expanded in 2022 to support 200 languages, in order to ensure that languages with low and high resources have equivalent translation quality. NLLB-200 successfully improved the performance of NLLB for low-resource languages [17].

D. Model Fine-Tuning Setup

The first stage involved fine-tuning the pre-trained mBART-50, mT5, and NLLB-200 models. This process aimed to assess the influence of hyperparameters on their performance in Tegalan-Indonesian translation. The fine-tuning process adjusted three hyperparameters: learning rate, batch size, and number of epochs. This study used two levels of learning rate, 0.00001 and 0.0001. Three batch sizes were used, namely 4, 8, and 16, while the dimensions 5, 10, and 15 represent variations of epochs used in this research.

The hyperparameter values at the beginning of the experiment were determined based on previous studies on language models with limited resources. Only representative parameter configurations were selected, rather than evaluating all parameter combinations, since this provides efficiency in executing the fine-tuning stages. All experiments were conducted in a GPU-accelerated environment equipped with a NVIDIA Tesla T4, and model fine-tuning was performed using the Adam optimizer.

During each fine-tuning session, the training loss was continuously monitored. Monitoring of loss values enabled a performance assessment for each model iteration. Additionally, the model evaluation score for each epoch was also recorded using three main metrics: BLEU, METEOR, and TER. BLEU assesses the suitability of word structure, whereas METEOR measures the equivalence of meaning between the translated result and the reference text. TER calculates the number of corrections required to bring the translation closer to the reference text.

E. Evaluation Metrics

All models were evaluated under identical data splits, experimental settings, and evaluation metrics to ensure a fair and comparable performance assessment. Three evaluation metrics were used to assess the performance of the fine-tuned translation models: BLEU, METEOR, and TER. Employing more than one metric provides a broader view of translation quality, capturing different aspects of the output. This approach is particularly valuable for low-resource language pairs, such as Tegalan-Indonesian, where translation characteristics can vary widely, and a single metric is often insufficient to represent overall model behavior.

BLEU was used as the primary metric because it enables consistent comparison across different experiments and is widely applied in MT research. The score reflects the amount of n-gram overlap between the system output and the reference text. BLEU is also less sensitive to formatting or tokenization differences, which helps maintain reproducibility. The formulation applied was:

$$BLEU = BP.exp\left(\sum_{n-1}^{N} w_{n \log P_n}\right) \quad (1)$$

where P_n denotes the n-gram precision for $n = 1$ to N (typically up to 4), w_n represents uniform weights ($w_n = 1/N$), and BP is the brevity penalty defined as $BP = 1$ if $> r$, else $BP = exp(1 - r/c)\, c$, with c and r being the lengths of the candidate and reference sentences, respectively. A higher BLEU score indicates n-gram overlap between the model output and the reference, indicating good translation.

METEOR provides a broader view of translation quality by incorporating linguistic features such as stemming, synonym matching, and more flexible word order. With these components, METEOR can align the system output with the reference using both exact and semantically related matches, making it more responsive to meaning preservation. The score is computed from unigram precision and recall through a weighted harmonic mean, with recall given greater weight to reflect completeness. This allows METEOR to capture nuances that may not be fully represented by BLEU.

METEOR calculates its final score using a harmonic mean of unigram precision (P) and recall (R). The specific formulas used in this study are given in (2) and (3). A penalty component is also included to account for fragmented alignments and differences in word order between the system output and the reference text. Taken together, these elements allow METEOR to evaluate translation quality in a more meaning-oriented way, making it sensitive to cases where the intended meaning is preserved even if the wording differs.

$$METEOR = Fmean\ x\ (1 - Penalty) \quad (2)$$

$$Fmean\ x = \frac{10\ x\ P\ x\ R}{R+9P} \quad (3)$$

Translation Edit Rate (TER) serves as a complementary metric by estimating how many edits are needed to adjust a system-generated translation so that it matches the reference. The edits counted include insertions, deletions, substitutions, and shifts. TER is calculated by dividing the minimum number of such edits by the total number of words in the reference text, as shown in (4). In this way, TER reflects the amount of human-level correction required to produce an acceptable translation.

$$TER = \frac{E}{R} \quad (4)$$

where E refers to the total number of edit operations and R represents the word length of the reference translation. A lower TER score means that fewer revisions are needed to obtain a correct translation, which suggests a more fluent and accurate output. This metric is particularly useful in low-resource evaluation settings, where translation output can vary considerably and surface-level lexical matching, such as that used by BLEU, does not always reflect the actual effort required to correct a translation.

III. RESULTS AND DISCUSSION

A. *Quantitative Evaluation Metrics Results*

The three MT models were trained on the three Tegalan-Indonesian datasets marked A, B, and C. The final results were obtained using a learning rate of 0.0001, a batch size of 8, and 10 training epochs, which provided the best balance between translation performance and training stability across models. NLLB-200 achieved the highest accuracy on Dataset C with an excellent BLEU score of 85.51, a METEOR score of 76.91, and a TER of 17.73. However, mBART-50 achieved the lowest METEOR results. The mT5 achieved the worst BLUE score on dataset A, due to its informal structure, but was competent on Dataset C, which gave a boost for both formal and informal translation settings. NLLB-200 achieved the lowest TER score, which means that it can render translations with higher fluency and naturalness. mBART-50 exhibited relatively consistent scores on BLEU and METEOR, despite the limited data available on the Tegalan language. Finally, mT5 is still affected by changes in the size and arrangement of the dataset.

mBART-50 required the longest training time, but its performance was actually below that of the other two models, whereas NLLB-200 achieved the best overall performance for all metrics. These findings demonstrate that the effectiveness of a model is determined not only by its duration of training but also by its architecture and pretraining quality.

Figure 2 presents the differences in translation quality. Consistent BLEU and METEOR values on datasets A to C indicate that more balanced and clean data can enhance model learning. Thus, it can be inferred that NLLB-200 outperforms the other models, exhibiting adaptability and consistency with different linguistic styles. There are inconsistencies in the data that affect the performance of mT5, while mBART-50 still manages to produce almost similar results, although less accurate. However, the low TER scores of NLLB-200 indicate high quality, but it still requires more computational power.

(a)

(b)

(c)

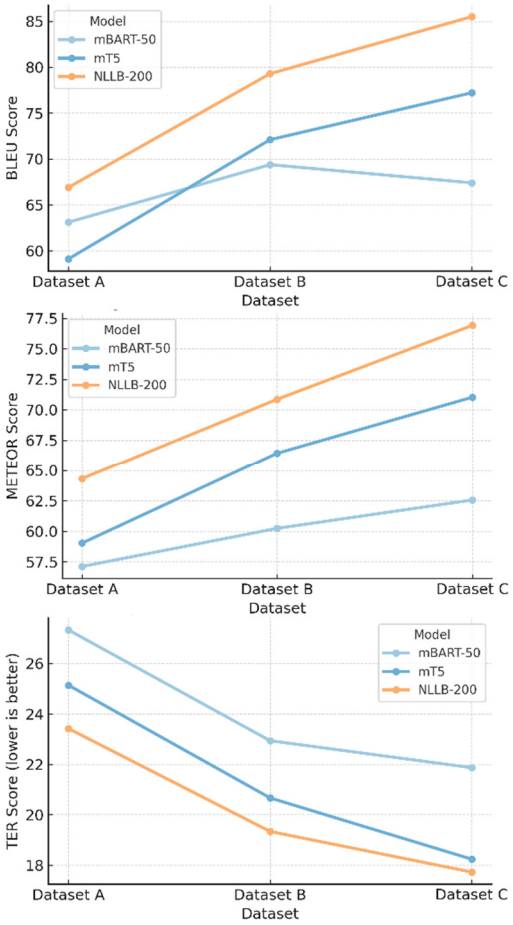

Fig. 2. Performance across datasets: (a) BLEU, (b) METEOR, (c) TER.

TABLE II. MODEL EVALUATION

Dataset	Transformer model	BLEU	METEOR	TER	Training Time (hours)
Dataset A	mBART-50	55.51	75.76	27.81	4.18
	mT5	23.64	46.78	57.57	3.57
	NLLB-200	35.36	70.94	19.19	2.32
Dataset B	mBART-50	69.39	84.01	20.372	4.62
	mT5	72.33	71.17	33.78	2.17
	NLLB-200	79.58	80.31	13.51	2.53
Dataset C	mBART-50	67.41	80.97	21.84	5.18
	mT5	79.44	71.17	17.73	3.93
	NLLB-200	85.51	76.91	17.73	2.88

B. *Qualitative Evaluation Results*

In addition to the quantitative results, a qualitative analysis was also performed to determine how mBART-50, mT5, and NLLB-200 translated Tegalan into Indonesian. The instances given in Table III are from models that were retrained with Dataset C. This evaluation focuses on the ability of each model to deal with dialect variations and sentence structures in the case of insufficient data, in addition to their general understanding, fluency, and semantic accuracy in translation. Table III illustrates a comparison of the translation output from the three Transformer-based models. In general, NLLB-200 produced the most fluent and contextually appropriate Indonesian translations among all models.

TABLE III. SAMPLE QUALITATIVE ASSESSMENT

Source (Tegalan)	Reference (Indonesian)	mBART50 output	mT5 output	NLLB-200 output
dheweke nyoba turu nang lantai.	dia mencoba tidur di lantai.	Dia mencoba turu di lantai.	Dia mencoba turu nang lantai.	Dia mencoba tidur di lantai.
Akhire Danu nggawe keputusan sing bener.	Pada akhirnya Danu membuat keputusan yang tepat.	Akhire Danu membuat keputusan yang benar.	Akhirnya Danu membuat keputusan yang benar.	Akhirnya Danu membuat keputusan yang tepat.
kiye kamera sing paling apik nang toko kuwe	ini adalah kamera terbaik di toko tersebut.	Ini kamera paling bagus nang toko itu.	Ini kamera paling bagus di toko tersebut.	Ini adalah kamera paling bagus di toko tersebut.

Some expressions from Tegalan, such as "turu" and "nang," were kept in the respective outputs of mBART-50 and mT5, which suggests that their translations were not very refined. mBART-50 shows greater syntactic accuracy, while mT5 shows greater sensitivity to semantic nuances. Therefore, mT5 tends to give more weight to the meaning of the sentence, while mBART gives more priority to the grammatical structure. Despite these differences, both models carry out the Tegalan syntax with a good degree of accuracy. Although mT5 and NLLB-200 achieved identical TER scores on Dataset C, NLLB-200 demonstrated superior overall performance due to its substantially higher BLEU score of 85.51 and METEOR score of 76.91.

C. Discussion

According to previous research [21, 22], the improvements in BLEU and METEOR support the ability of NLLB-200 to translate more than 200 languages. In [16], it was stated that the grammatical constructions formed in mT5 translation had more complex semantics, which is in agreement with the findings of this study. Subsequently, based on [19], it appears that the stability demonstrated by mBART-50 is good, although it was not subjected to intensive retraining. Scalability and tokenization seem to be more important than introducing a new architectural design in the context of a language with many morphologies, such as Tegalan. Based on this, some of these results are in line with findings from other Transformer models that support multiple languages.

The more accurate subword segmentation of NLLB-200 and its extensive multilingual pretraining are probably the main reasons for its better performance compared to mBART-50. Better knowledge transfer between languages is made possible by these features, even in situations where parallel data is scarce. However, some shortcomings are suggested by sporadic literal translations or untranslated Tegalan words such as "turu" and "nang" [33]. This restriction is likely a result of the relatively small and homogenous dataset of around 4,000 sentence pairs. The limitation of using only a single GPU hindered the testing of larger batch sizes or the possible tuning of hyperparameters in alternative ways.

The technological implications of this research are numerous and, at the same time, reveal the areas of technological advance and cultural preservation. In addition, it is also inferred that multilingual approaches can be applied to dialects with scarce resources through the use of linguistic normalization, balanced learning rates, and optimized tokenization strategies, among other methods. Future work should build up the Tegalan corpus so as to improve translation quality and cross-linguistic generalization even more.

IV. CONCLUSION

This study assessed the performance of three multilingual Transformer-based models, mBART-50, mT5, and NLLB-200, in Tegalan-Indonesian MT, showing that the most reliable and effective was NLLB-200. The remarkable BLEU and METEOR scores of NLLB-200 demonstrate that it is capable of producing correct and contextually appropriate translations, although it has a relatively low TER score. The results also indicate that there has been considerable improvement in pre-training and resulting translation quality, particularly in low-resource languages.

However, various problems need to be resolved in future research. The model in question faces difficulties with idioms and generalization due to insufficient data. To create a more diverse data pool, future studies could add to the Tegalan corpus by asking the local community for input. The use of semi-supervised methods and back-translation in future studies could increase the variety of data. The accuracy could also be improved by including syntactic and morphological aspects, which might be important variables. This study promotes the digital preservation of regional languages and establishes a framework for AI-driven translation systems in tourism, education, and cultural heritage.

ACKNOWLEDGMENT

This work was funded by the Indonesian Ministry of Higher Education, Science, and Technology (KEMDIKTISAINTEK) for the financial year 2025, grant scheme of Penelitian Fundamental Reguler (PFR), and contract number 202/C3/DT.05.00/PL-BATCH II/2025, 002/LL6/PL-BATCH II/AL.04/2025, 105.16/P3M.PHB/VIII/2025.

REFERENCES

[1] V. Peltokorpi and E. Vaara, "Language Policies and Practices in Wholly Owned Foreign Subsidiaries: A Recontextualization Perspective," in *Language in International Business*, M. Y. Brannen and T. Mughan, Eds. Springer International Publishing, 2017, pp. 93–138.

[2] I. Rivera-Trigueros, "Machine translation systems and quality assessment: a systematic review," *Language Resources and Evaluation*, vol. 56, no. 2, pp. 593–619, June 2022, https://doi.org/10.1007/s10579-021-09537-5.

[3] D. Khurana, A. Koli, K. Khatter, and S. Singh, "Natural language processing: state of the art, current trends and challenges," *Multimedia Tools and Applications*, vol. 82, no. 3, pp. 3713–3744, Jan. 2023, https://doi.org/10.1007/s11042-022-13428-4.

[4] S. Bird, "Decolonising Speech and Language Technology," in *Proceedings of the 28th International Conference on Computational Linguistics*, Barcelona, Spain, 2020, pp. 3504–3519, https://doi.org/10.18653/v1/2020.coling-main.313.

[5] A. F. Aji *et al.*, "One Country, 700+ Languages: NLP Challenges for Underrepresented Languages and Dialects in Indonesia," in *Proceedings of the 60th Annual Meeting of the Association for Computational*

Linguistics (Volume 1: Long Papers), Dublin, Ireland, 2022, pp. 7226–7249, https://doi.org/10.18653/v1/2022.acl-long.500.

[6] F. Li, B. Liu, H. Yan, P. Xie, J. Li, and Z. Zhang, "Incorporating bilingual translation templates into neural machine translation," *Scientific Reports*, vol. 15, no. 1, Feb. 2025, Art. no. 5547, https://doi.org/10.1038/s41598-025-86754-w.

[7] F. Reza, Z. Rohmah, and R. Ismail, "Dialect shift and cultural dynamism among Betawi community in urban Jakarta in *Palang Pintu* and *Rebut Dandang* traditional ceremonies," *Cogent Arts & Humanities*, vol. 11, no. 1, Dec. 2024, Art. no. 2410542, https://doi.org/10.1080/23311983.2024.2410542.

[8] U. Kulsum, A. D. Darnis, and A. Asdari, "Language Shift and Endangerment of Sundanese Banten Dialect in South Tangerang," *Insaniyat : Journal of Islam and Humanities*, vol. 7, no. 2, pp. 99–111, May 2023, https://doi.org/10.15408/insaniyat.v7i2.28905.

[9] D. A. Sulistyo, "LSTM-Based Machine Translation for Madurese-Indonesian," *Journal of Applied Data Sciences*, vol. 4, no. 3, pp. 189–199, Sept. 2023, https://doi.org/10.47738/jads.v4i3.113.

[10] D. E. Messaoudi and D. Nessah, "Enhancing Neural Arabic Machine Translation using Character-Level CNN-BILSTM and Hybrid Attention," *Engineering, Technology & Applied Science Research*, vol. 14, no. 5, pp. 17029–17034, Oct. 2024, https://doi.org/10.48084/etasr.8383.

[11] A. Vaswani *et al.*, "Attention Is All You Need." arXiv, 2017, https://doi.org/10.48550/ARXIV.1706.03762.

[12] Z. Tan *et al.*, "Neural machine translation: A review of methods, resources, and tools," *AI Open*, vol. 1, pp. 5–21, 2020, https://doi.org/10.1016/j.aiopen.2020.11.001.

[13] S. Ranathunga, E. S. A. Lee, M. P. Skenduli, R. Shekhar, M. Alam, and R. Kaur, "Neural Machine Translation for Low-Resource Languages: A Survey." arXiv, 2021, https://doi.org/10.48550/ARXIV.2106.15115.

[14] L. Susanto, R. Diandaru, A. Krisnadhi, A. Purwarianti, and D. T. Wijaya, "Replicable Benchmarking of Neural Machine Translation (NMT) on Low-Resource Local Languages in Indonesia," in *Proceedings of the First Workshop in South East Asian Language Processing*, Nusa Dua, Bali, Indonesia, 2023, pp. 100–115, https://doi.org/10.18653/v1/2023.sealp-1.8.

[15] Y. Tang *et al.*, "Multilingual Translation with Extensible Multilingual Pretraining and Finetuning." arXiv, Aug. 02, 2020, https://doi.org/10.48550/arXiv.2008.00401.

[16] L. Xue *et al.*, "mT5: A massively multilingual pre-trained text-to-text transformer." arXiv, 2020, https://doi.org/10.48550/ARXIV.2010.11934.

[17] NLLB Team *et al.*, "Scaling neural machine translation to 200 languages," *Nature*, vol. 630, no. 8018, pp. 841–846, June 2024, https://doi.org/10.1038/s41586-024-07335-x.

[18] S. M. Singh and T. D. Singh, "Low resource machine translation of english–manipuri: A semi-supervised approach," *Expert Systems with Applications*, vol. 209, Dec. 2022, Art. no. 118187, https://doi.org/10.1016/j.eswa.2022.118187.

[19] B. Namdarzadeh, S. Mohseni, L. Zhu, G. Wisniewski, and N. Ballier, "Fine-tuning MBART-50 with French and Farsi data to improve the translation of Farsi dislocations into English and French," in *Proceedings of Machine Translation Summit XIX, Vol. 2: Users Track*, Macau SAR, China, June 2023, pp. 152–161.

[20] T. N. Son, N. A. Tu, and N. M. Tri, "An Efficient Approach for Machine Translation on Low-resource Languages: A Case Study in Vietnamese-Chinese." arXiv, Jan. 31, 2025, https://doi.org/10.48550/arXiv.2501.19314.

[21] R. Oida-Onesa and M. A. Ballera, "Fine Tuning Language Models: A Tale of Two Low-Resource Languages," *Data Intelligence*, vol. 6, no. 4, pp. 946–967, Dec. 2024, https://doi.org/10.3724/2096-7004.di.2024.0016.

[22] V. N. M. Abadi and F. Ghasemian, "Enhancing Persian text summarization through a three-phase fine-tuning and reinforcement learning approach with the mT5 transformer model," *Scientific Reports*, vol. 15, no. 1, Jan. 2025, Art. no. 80, https://doi.org/10.1038/s41598-024-78235-3.

[23] V. Akerman *et al.*, "The eBible Corpus: Data and Model Benchmarks for Bible Translation for Low-Resource Languages." arXiv, Apr. 19, 2023, https://doi.org/10.48550/arXiv.2304.09919.

[24] D. Degenaro and T. Lupicki, "Experiments in Mamba Sequence Modeling and NLLB-200 Fine-Tuning for Low Resource Multilingual Machine Translation," in *Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP 2024)*, Mexico City, Mexico, 2024, pp. 188–194, https://doi.org/10.18653/v1/2024.americasnlp-1.22.

[25] S. Lee *et al.*, "A Survey on Evaluation Metrics for Machine Translation," *Mathematics*, vol. 11, no. 4, Feb. 2023, Art. no. 1006, https://doi.org/10.3390/math11041006.

[26] "Machine Translation | Seq2Seq | LSTMs." Kaggle, [Online]. Available: https://kaggle.com/code/harshjain123/machine-translation-seq2seq-lstms.

[27] D. N. De Oliveira and L. H. D. C. Merschmann, "Joint evaluation of preprocessing tasks with classifiers for sentiment analysis in Brazilian Portuguese language," *Multimedia Tools and Applications*, vol. 80, no. 10, pp. 15391–15412, Apr. 2021, https://doi.org/10.1007/s11042-020-10323-8.

[28] T. Bergmanis, A. Stafanovičs, and M. Pinnis, "Robust Neural Machine Translation: Modeling Orthographic and Interpunctual Variation," in *Frontiers in Artificial Intelligence and Applications*, A. Utka, J. Vaičenonienė, J. Kovalevskaitė, and D. Kalinauskaitė, Eds. IOS Press, 2020.

[29] K. Imamura and M. Utiyama, "An Empirical Study of Multilingual Vocabulary for Neural Machine Translation Models," in *Proceedings of the Eleventh Workshop on Asian Translation (WAT 2024)*, Miami, FL, USA, Aug. 2024, pp. 22–35, https://doi.org/10.18653/v1/2024.wat-1.2.

[30] Y. Liu *et al.*, "Multilingual Denoising Pre-training for Neural Machine Translation," *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 726–742, Dec. 2020, https://doi.org/10.1162/tacl_a_00343.

[31] C. Raffel *et al.*, "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer," *Journal of Machine Learning Research*, vol. 21, no. 140, pp. 1–67, 2020.

[32] NLLB Team *et al.*, "No Language Left Behind: Scaling Human-Centered Machine Translation." arXiv, 2022, https://doi.org/10.48550/ARXIV.2207.04672.

[33] D. A. Sulistyo, A. P. Wibawa, D. D. Prasetya, F. A. Ahda, I. N. G. A. Astawa, and F. A. Dwiyanto, "Multilingual Parallel Corpus for Indonesian Low-Resource Languages," *JOIV : International Journal on Informatics Visualization*, vol. 9, no. 5, pp. 2176–2182, Sept. 2025, https://doi.org/10.62527/joiv.9.5.3412.