Bidirectional Multilingual Neural Machine Translation Between Regional Languages and Indonesian with the Transformer Model

Frendy[1], Yohanes Sigit Purnomo W.P.[2], Martinus Maslim[3]
Department of Informatics, Faculty of Industrial Technology, Universitas Atma Jaya Yogyakarta
Jl. Babarsari 43, Sleman 55281, Daerah Istimewa Yogyakarta, Indonesia
Email: [1]220711608@students.uajy.ac.id, [2]sigit.purnomo@uajy.ac.id, [3]martinus.maslim@uajy.ac.id

Abstrak. *Minimnya sumber daya linguistik digital untuk bahasa daerah Indonesia menjadi kendala utama dalam pengembangan sistem penerjemahan otomatis yang berkualitas. Penelitian ini menyajikan perbandingan empiris tiga model Neural Machine Translation (NMT) multilingual berbasis Transformer, yaitu mT5, ByT5, dan NLLB, untuk penerjemahan dua arah antara lima bahasa daerah Indonesia (Jawa, Sunda, Bali, Minangkabau, dan Dayak Maanyan) dan Bahasa Indonesia pada kondisi data terbatas. Menggunakan sekitar 4.000 pasangan kalimat per bahasa, ketiga model dievaluasi dengan BLEU dan ChrF++. Hasil menunjukkan model bilingual baseline secara konsisten mengungguli pendekatan multilingual bidirectional pada ketiga arsitektur, dengan NLLB paling tangguh terhadap penurunan performa dan mencatat hasil terbaik keseluruhan (BLEU 62,16; ChrF++ 79,93); hasil ini berlaku untuk kondisi yang terkontrol dan spesifik-domain. Penelitian ini memberikan bukti empiris awal mengenai pemilihan arsitektur untuk penerjemahan bahasa daerah, dengan keterbatasan pada ukuran dataset dan absennya evaluasi manusia.*
Keywords: *Bahasa Daerah, Multilingual, Penerjemahan Mesin, Penerjemahan Mesin Neural, Transformer*

Abstract. *The limited digital linguistic resources for Indonesian regional languages remain a major obstacle to developing high-quality automatic translation systems. This study presents an empirical comparison of three multilingual Transformer based Neural Machine Translation (NMT) models, mT5, ByT5, and NLLB, for bidirectional translation between five Indonesian regional languages (Javanese, Sundanese, Balinese, Minangkabau, and Dayak Maanyan) and Indonesian under low-resource conditions. Using approximately 4,000 sentence pairs per language, the three models were evaluated with BLEU and ChrF++. Results show the bilingual baseline consistently outperformed the bidirectional multilingual approach across all three architectures, with NLLB proving most resilient to this degradation and achieving the strongest overall result (BLEU 62.16, ChrF++ 79.93) in controlled, domain-specific settings. This study contributes preliminary empirical evidence on architecture selection for low-resource regional language translation, constrained by dataset size and the absence of human evaluation.*
Keywords: *Machine Translation, Multilingual, Neural Machine Translation, Regional Languages, Transformer*

1. Introduction

Indonesia is recognized as one of the most linguistically diverse nations in the world, with more than 700 regional languages still actively used across its communities. However, a UNESCO report noted that 139 of these regional languages in Indonesia are endangered, representing approximately 17% of the total languages in existence [1]. The forces of modernization, migration, and the dominance of Indonesian as the national lingua franca have accelerated the displacement of regional languages from digital spaces, thereby threatening the preservation of local culture and identity [2]. One promising approach to supporting regional language preservation in the digital era is the development of Neural Machine Translation (NMT) systems that automatically translate regional languages into Indonesian. Beyond translation, other technological interventions have also been explored to support regional language preservation, such as rule-based transliteration systems that convert Latin script into traditional regional scripts

like Javanese script, aiming to help younger generations learn and retain their native writing systems [3].

Although NMT has been extensively applied to large-scale international languages, its application to Indonesian regional languages remains confronted with serious challenges. Prior studies have demonstrated that translation quality in Indonesian regional languages remains low, with average BLEU scores ranging from 10 to 30 across language pairs [4]. This is primarily attributed to the scarcity of parallel data for low-resource regional languages. While transfer learning approaches using multilingual models such as NLLB have begun to be adopted, the resulting BLEU scores remain limited, reaching only 30-37 in cases such as Batak Toba→Indonesian translation [5]. This indicates that the low-resource condition remains a significant barrier.

The Multilingual Neural Machine Translation (MNMT) approach presents a more promising solution by allowing a single model to handle multiple language pairs simultaneously through parameter sharing and cross-lingual transfer mechanisms [6]. Research on 45 languages of Indonesia using mT5-based Indo-T5 models showed that intermediate-domain transfer learning can improve translation performance in low-resource settings, although bilingual fine-tuning generally outperformed multilingual fine-tuning on the NusaX benchmark [7]. However, this research is limited to small datasets with specific domains, and does not cover endangered languages such as Dayak Maanyan. Furthermore, no research has systematically compared several multilingual Transformer architectures such as mT5, ByT5, and NLLB in a single study covering all five regional languages simultaneously.

This research used five bilingual datasets: Indonesian-Dayak Ma'anyan [8], Indonesian-Balinese [9], Indonesian-Minangkabau[10], Indonesian-Sundanese [11] and Indonesian-Javanese [12] published by Prabowo et al. By utilizing this dataset, this study aims to fine-tune and empirically compare three existing pretrained multilingual Transformer models (mT5, ByT5, and NLLB) for bidirectional translation between the five regional languages and Indonesian under low-resource conditions. The evaluation is carried out using BLEU and ChrF++ metrics with a comparison against a bilingual model as a baseline, to assess the effectiveness of the multilingual approach and determine the most optimal architecture for the context of Indonesian regional language translation.

2. Literature Review

The development of machine translation has gone through three main phases: Rule-Based Machine Translation (RBMT), Statistical Machine Translation (SMT), and Neural Machine Translation (NMT). RBMT relies on hand-crafted dictionaries and grammar rules, making it effective for simple-structured languages, but inefficient for many language pairs because each pair requires the development of its own rules [13][14]. SMT then came up with a more flexible probabilistic approach but still has limitations in understanding semantic and idiomatic context [15][16]. NMT overcomes the weaknesses of the two previous approaches by utilizing end-to-end artificial neural networks, producing more natural and contextual translations, and has become the industry standard today [17].

The emergence of the Transformer architecture further revolutionized NMT through its self-attention mechanism capable of capturing long-range relations in text in parallel without sequence dependencies like RNN or LSTM [18]. Vaswani et al. demonstrated that this architecture was able to surpass the performance of previous state-of-the-art models, while other studies showed a training process acceleration of up to 11 times faster than sequential-based architectures [18][19]. The Transformer architecture then became the foundation for various large models such as BERT, GPT, mT5, ByT5, and NLLB which are now widely used in various natural language processing tasks.

MNMT is a development of NMT that allows a single model to handle multiple language pairs simultaneously by sharing parameters, so that linguistic similarities between languages can be utilized as a shared source of knowledge through a cross-lingual transfer mechanism [6]. Costa-Jussà et al. showed that a multilingual NMT model was able to improve translation performance

by up to 44% better than previous models [20], while other studies have proven the ability of zero-shot translation on language pairs that have never been explicitly trained [21]. In the context of Indonesian regional languages, Wongso et al. [7] demonstrated the potential of mT5-based transfer learning for low-resource machine translation. However, their results showed that multilingual fine-tuning did not consistently outperform bilingual adaptation. For regional-language-to-Indonesian translation on the NusaX benchmark, the bilingual Indo-T5 NusaX model achieved an average SacreBLEU score of 39.30, compared with 31.72 for its multilingual counterpart. These findings suggest that although multilingual pre-training and transfer can provide useful cross-lingual representations, increasing the number of languages within a shared model does not necessarily improve downstream performance and may introduce a curse-of-multilinguality effect.

However, most NMT research for Indonesian regional languages is still conducted on a limited scale and uses a bilingual approach. Research on Toba Batak, Teochiu, Sundanese, and Sasak languages shows that translation quality is heavily influenced by data availability, with average BLEU scores ranging from only 10 to 30 [5][22][23][24]. Even large-scale multilingual models such as NLLB can only improve BLEU scores to around 30 to 37 for Toba Batak [4], indicating that low-resource conditions remain a major obstacle. Furthermore, no research has systematically compared several multilingual Transformer architectures in a single comprehensive study covering endangered regional languages such as Dayak Maanyan. Prabowo et al. published five bilingual datasets: Indonesian-Dayak Ma'anyan [8], Indonesian-Balinese [9], Indonesian-Minangkabau [10], Indonesian-Sundanese [11], and Indonesian-Javanese [12] without the translation models. This study aims to fill this gap by utilizing an existing corpus to train and compare three multilingual Transformer models, namely mT5, ByT5, and NLLB, in a comparative study under low-resource conditions.

3. Methodology

The research pipeline adopted in this study is illustrated in Figure 1. It comprises seven sequential stages: data collection and annotation, data preprocessing, data exploration, feature engineering, model training, evaluation, and implementation.

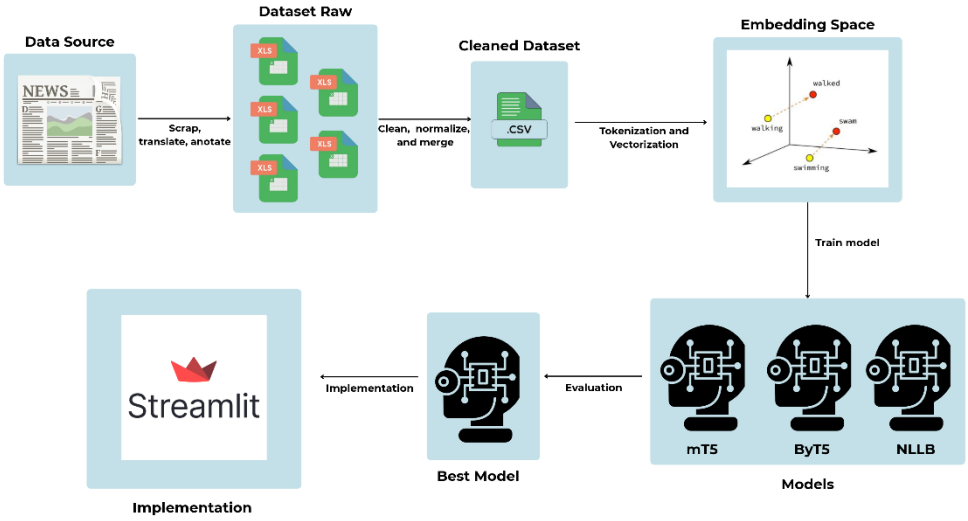


Figure 1. Illustration of the Research Process

3.1. Data Collection

The dataset used in this study is a bilingual corpus published by Prabowo et al., consisting of approximately 4,001 sentence pairs per language across five Indonesian regional languages: Indonesian-Dayak Ma'anyan [8], Indonesian-Balinese [9], Indonesian-Minangkabau[10], Indonesian-Sundanese [11], and Indonesian-Javanese [12]. Source sentences in Indonesian were collected from news articles spanning diverse domains, translated by regional language teachers, and subsequently validated by designated validators. The translation process was carried out using

Microsoft Excel, yielding a total of 20,005 sentence pairs with a balanced distribution of approximately 20% per language.

3.2. Data Preprocessing

Raw data in Excel format was consolidated into a single Pandas dataframe, during which three sentences with missing translations were identified and removed; therefore, the final number of sentence pairs is 3,998, or a total of 19,990 sentence pairs across all languages. The data was restructured from wide to long format to produce source–target sentence pairs compatible with the model's training requirements. Unicode normalization (NFD) was then applied to remove diacritical marks, standardize punctuation, and eliminate excess whitespace. A language tag was prepended to each source sentence to condition the model on translation direction. In the one-directional scenario, only source-language tags were used, in the format <src:{language}> (e.g., <src:jav>), since the target language was always Indonesian. In the bidirectional scenario, two tag types were used to distinguish direction: <src:{language}> was prepended when translating from a regional language into Indonesian (e.g., <src:jav> Alesane, panganan iki bisa nambah... — *"The reason is, this food can increase..."*), while <tgt:{language}> was prepended when translating from Indonesian into a regional language (e.g., <tgt:jav> Pasalnya, makanan tersebut dapat... — *"The reason is, this food can increase..."*). This yields 10 distinct tag values across the five regional languages and both tag types, sufficient to uniquely identify each of the ten direction-language combinations without requiring an explicit output-language token, since each tag value co-occurs with only one target distribution during training. NLLB did not require this scheme as it used its built-in Language ID tokens, such as ind_Latn and jav_Latn.

3.3. Data Exploration

Sentence length distribution was analyzed to determine the optimal max_length value for tokenization. Descriptive statistics revealed an average sentence length of 21 words, with a maximum of 133 words in regional-language sentences and 102 in Indonesian sentences. The cross-language data distribution was confirmed to be balanced, ensuring that the model would not develop bias toward any particular language during training.

3.4. Feature Engineering

The dataset was partitioned into training, validation, and test sets using a group-based splitting strategy to prevent information leakage across regional languages and translation directions. Because the corpus was constructed from a common set of Indonesian source sentences translated into five regional languages, all sentence pairs derived from the same Indonesian source sentence were treated as a single group. Splitting individual sentence pairs independently could otherwise place different regional-language translations, or the reverse translation direction of the same underlying sentence, in different subsets.

After preprocessing, 3,998 unique Indonesian source-sentence groups remained. GroupShuffleSplit from scikit-learn was applied using the source-sentence identifier as the grouping key. The splitting procedure was performed before constructing the reverse translation direction. In the first stage, 80% of the source groups (3,198 groups) were assigned to the training set, while the remaining 800 groups were held out. In the second stage, these 800 groups were divided equally into validation and test sets, yielding 400 groups each. Thus, no Indonesian source sentence, nor any of its associated regional-language translations, could occur in more than one subset.

Each source group contributed five sentence pairs in the one-directional setting, resulting in 15,990 training, 2,000 validation, and 2,000 test pairs. For the bidirectional setting, the reverse Indonesian-to-regional-language pairs were generated only after the group assignment had been completed. Each source group therefore contributed ten sentence pairs, resulting in 31,980 training, 4,000 validation, and 4,000 test pairs. The resulting distribution preserved the intended approximately 80-10-10 partition while maintaining strict group separation between training, validation, and test data.

3.5. Model Configuration

This study compares three multilingual Transformer architectures, described as follows. MT5 (Multilingual Text-to-Text Transfer Transformer) is the multilingual extension of T5, pre-trained on the mC4 corpus covering 101 languages. The variant used in this study is mT5-small, trained with the AdamW optimizer, a learning rate of 3×10^{-4}, and a label smoothing coefficient of 0.1. This model employs a SentencePiece tokenizer with a vocabulary size of 250,000 tokens [25]. The *mt5-small* variant was used with a maximum token length of 128. ByT5 (Byte-level Text-to-Text Transfer Transformer) operates directly at the UTF-8 byte level without subword tokenization, thereby eliminating Out-of-Vocabulary (OOV) issues. This property makes ByT5 particularly suitable for regional languages with complex morphology, such as Dayak Maanyan. The ByT5-small variant was used, trained with a learning rate of 5×10^{-4} and the AdamW optimizer [26]. The *byt5-small* variant was used with a maximum token length of 324, reflecting ByT5's byte-level encoding, which expands sequence length significantly compared to subword models. Meta AI developed NLLB (No Language Left Behind) to support high-quality translation across 200 languages, including low-resource ones. It is the only architecture in this study designed specifically as a machine translation system, and it's pre-training already incorporates several Indonesian regional languages such as Javanese, Sundanese, and Balinese. Fine-tuning was conducted with a learning rate of 2×10^{-5} using the AdamW optimizer [20]. The *nllb-200-distilled-600M* variant was used with a maximum token length of 256.

3.6. Model Training

All training was conducted on Lovelace Platform by Universitas Atma Jaya Yogyakarta using NVIDIA L40S GPU, with a maximum of 20 epochs. All models were implemented using the Hugging Face Transformers library version 4.52.4. Training was configured with a batch size of 16, the AdamW optimizer, a warmup ratio of 0.06, and an early stopping mechanism with a patience of 3 epochs at a threshold of 0.001 for the evaluation loss. Three experimental scenarios were applied. (1) *One-directional*: The model is trained to translate exclusively from regional languages into Indonesian. (2) *Bidirectional*: The model is trained simultaneously on both translation directions (regional language ↔ Indonesian). (3) *Bilingual baseline*: A separate model is trained for each language pair. The multilingual approach trains all language pairs within a single shared-parameter training session to maximize cross-lingual transfer. The best-performing model checkpoint was saved based on the highest ChrF++ validation score.

3.7. Evaluation

Model evaluation was performed using two automatic metrics, BLEU (Bilingual Evaluation Understudy) and ChrF++ (Character n-gram F-score), on a held-out test set that was not used during training or validation. The one-directional test set contained 2,000 sentence pairs, corresponding to 400 source-sentence groups and five regional languages, with exactly 400 pairs contributed by each language. The bidirectional test set contained 4,000 sentence pairs. Because each of the 400 test source groups generated two translation directions for each of the five regional languages, each language contributed 800 test pairs in total, consisting of 400 regional-language-to-Indonesian and 400 Indonesian-to-regional-language pairs. Multilingual model results were compared against the bilingual baseline to quantify the effectiveness of cross-lingual transfer. Translation outputs were generated using beam search decoding for all three models. For mT5 and ByT5, beam search was applied with *num_beams* = 5. For NLLB, the default generation configuration of the *nllb-200-distilled-600M* model was used, with beam search *(num_beams* = 4) and *early_stopping* enabled. Statistical significance testing was additionally conducted using the paired bootstrap resampling method as implemented in the SacreBLEU library, with 1,000 resampling iterations, to confirm that performance differences among the three models were not attributable to random variation.

To assess generalization beyond the training corpus, an additional out-of-corpus evaluation was conducted for the Indonesian-Javanese pair using 100 sentence pairs sampled from the FLORES-200 devtest split, translated in the Indonesian-to-Javanese direction (ind_Latn to

jav_Latn). This language pair was selected because Javanese is the only one of the five regional languages with an available standardized external translation benchmark. The same BLEU and ChrF++ metrics, computed with SacreBLEU, were used to ensure direct comparability with the in-corpus results. This evaluation design carries a construct validity consideration: because training, validation, and test data are drawn from a single, stylistically homogeneous domain, the in-corpus BLEU and ChrF++ scores reported in Section 4 primarily measure each model's ability to adapt to this specific domain and translation style, a narrower construct than the general translation quality these metrics are conventionally used to proxy for. The out-of-corpus evaluation described above is intended to partially address this gap, though it remains limited to a single language pair and direction.

3.8. Implementation

The best-performing model was integrated into a web-based application built with the Streamlit framework in Python, enabling interactive and real-time inference. Users can select the source regional language, input a source sentence, and obtain the Indonesian translation through the application interface. The input sentence is processed through a tokenization stage before being passed to the selected NLLB model for inference. Users may also optionally provide a reference translation to compute on-the-fly BLEU and ChrF++ scores. This implementation serves both as a practical demonstration of the system's usability and as a final validation of the model's generalization capability on out-of-corpus data.

4. Result and Discussion

The results were obtained from the training and testing process of three multilingual Transformer models, namely mT5, ByT5, and NLLB, using a parallel dataset of five Indonesian regional languages that have gone through the preprocessing, tokenization, and feature engineering stages. Performance evaluation was carried out using BLEU and ChrF++ metrics by comparing one-directional, bidirectional, and bilingual baseline scenarios.

4.1. Model Training Results

Table 1 summarizes the training duration and number of epochs completed by each model in the two main scenarios.

Table 1. Training Duration

Scenario	Model	Training Duration	No. of Epoch
One-Directional	mT5	37 minutes 37 seconds	9 epoch
	ByT5	1 hour 5 minutes	6 epoch
	NLLB	4 hour 8 minutes	7 epoch
Bidirectional	mT5	2 hour 17 minutes	20 epoch
	ByT5	1 hour 28 minutes	10 epoch
	NLLB	13 hour 14 minutes	11 epoch

NLLB required the longest training time in both scenarios, taking around 4 hours 8 minutes in the one-directional scenario and 13 hours 14 minutes in the bidirectional scenario, several times longer than mT5 and ByT5. This reflects NLLB's larger model size, which increases the computational cost per epoch even though it relies on standard subword tokenization like mT5. The mT5 model recorded the shortest training time in the one-directional scenario at 37 minutes 37 seconds, while ByT5 was the fastest in the bidirectional scenario at 1 hour 28 minutes. ByT5's longer per-epoch time compared to mT5, despite completing fewer total epochs, is consistent with its byte-level architecture, which produces significantly longer sequences and a heavier computational load per step than subword tokenization-based models. The transition from one-directional to bidirectional scenarios increased training time substantially for mT5 and NLLB, more than tripling their training duration, while ByT5's training time increased more moderately, rising by about 35 percent.

Early stopping occurred in most training configurations. The ByT5 model stopped early at the 6th epoch in the one-directional scenario and the 10th epoch in the bidirectional scenario, suggesting that its byte-level representation reaches validation loss convergence faster despite its slower per-epoch computation. The NLLB model also stopped early, completing 7 epochs in the one-directional scenario and 11 epochs in the bidirectional scenario. In contrast, mT5 stopped early at the 9th epoch in the one-directional scenario but completed all 20 epochs in the bidirectional scenario, indicating that adding translation directions required more extensive training for this architecture to reach its generalization limit.

Analysis of the training and validation loss curves in Figure 2 and Figure 3 reveals consistent differences in convergence behavior among the three models across both scenarios. The mT5 model shows the most stable convergence of the three models, with training and validation loss remaining close together before diverging gradually after around the 4th epoch in the one-directional scenario and the 7th epoch in the bidirectional scenario, resulting in only a mild and comparatively narrow gap between the two curves. ByT5 shows a more pronounced overfitting pattern, with training loss decreasing steadily throughout training while validation loss stagnates early, around the 3rd epoch in the one-directional scenario and the 4th to 5th epoch in the bidirectional scenario, before training is stopped. The NLLB model shows the clearest indication of overfitting among the three models, with validation loss plateauing by around the 3rd to 4th epoch in both scenarios while training loss continues to decrease for the remainder of training, producing the widest gap between the two curves by the end of training, a phenomenon that is common in large pre-trained models that are fine-tuned on limited data [27].

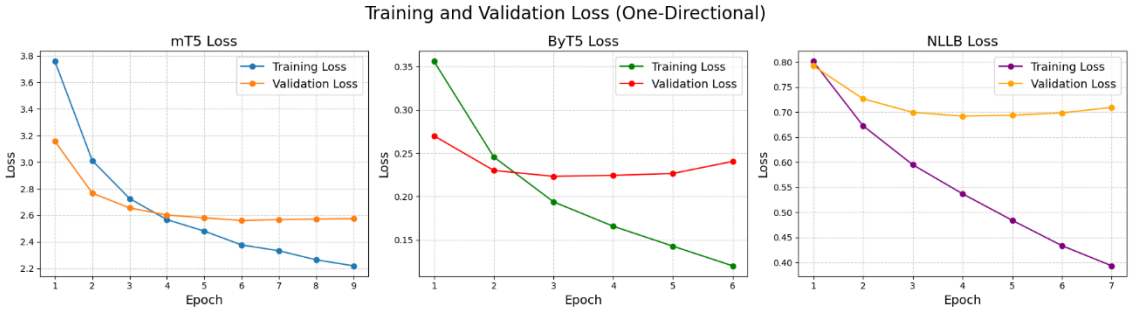


Figure 2. Train and Validation Loss Curve (One-Directional)

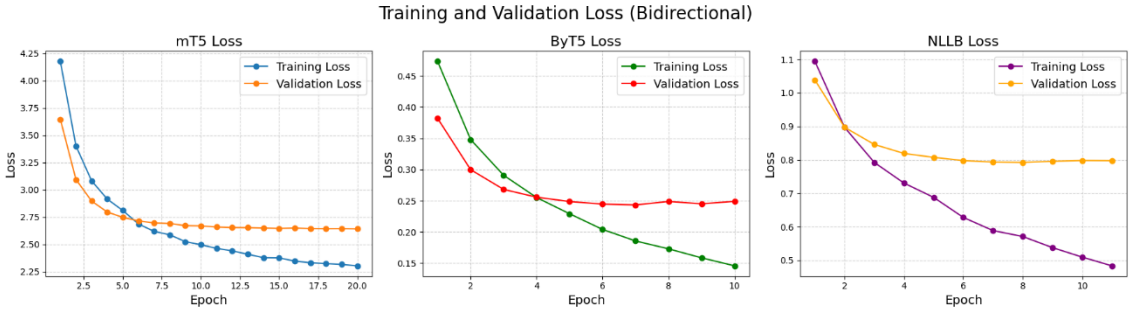


Figure 3. Train and Validation Loss Curve (Bidirectional)

However, the observed indications of overfitting did not significantly degrade translation quality. This is due to three mitigating factors: an early stopping mechanism that saves the model at the highest validation ChrF++ score, cross-language linguistic knowledge stored in the pre-trained weights as a form of implicit regularization, and the ability of large-scale pre-trained models to maintain high performance on test data despite overfitting the loss function during fine-tuning.

4.2. Multilingual Model Performance

Based on the model performance comparison diagram, all three models demonstrate strong performance in the one-directional scenario; these results apply to controlled, domain-specific settings. NLLB recorded the highest score with BLEU 64.35 and ChrF++ 80.48, followed by ByT5 with BLEU 64.18 and ChrF++ 80.18, and mT5 with BLEU 53.74 and ChrF++ 75.17. ByT5's competitive performance, trailing NLLB by only about 0.17 BLEU points and 0.3 ChrF++ points, indicates that byte-based approaches can approach the quality of specialized translation models like NLLB in handling regional languages. In contrast, mT5 lags behind both models, indicating that this architecture is less responsive to the characteristics of regional language datasets without further adjustments (see Figure 4).

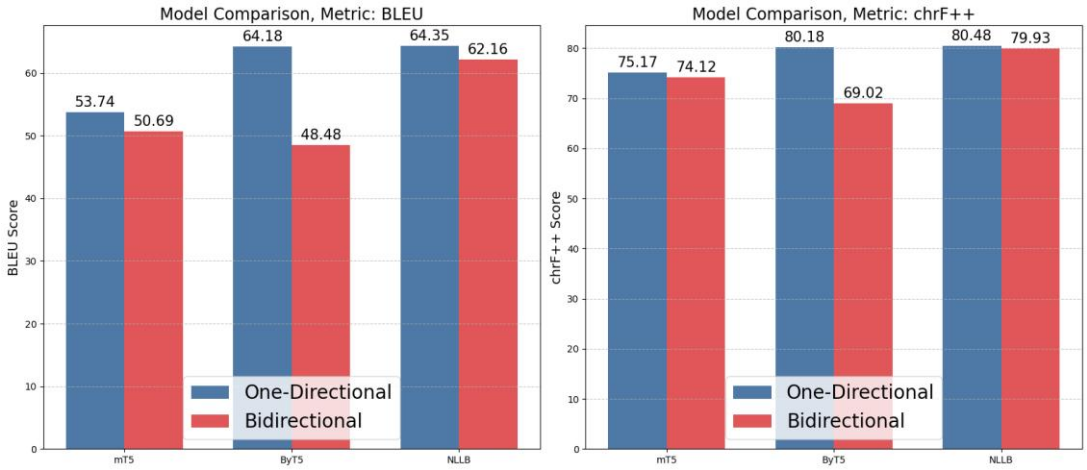


Figure 4. Comparison of Multilingual Model Performance in One-Directional and Bidirectional Scenarios

All three models experienced performance degradation in the bidirectional scenario, though the extent varied considerably across models. ByT5 showed by far the sharpest decline, dropping from BLEU 64.18 to 48.48, a 24.46% decrease, and from ChrF++ 80.18 to 69.02, a 13.92% decrease. mT5 showed a moderate decline, from BLEU 53.74 to 50.69 (5.68%) and ChrF++ 75.17 to 74.12 (1.4%). NLLB was the most robust, declining only from BLEU 64.35 to 62.16 (3.41%) and ChrF++ 80.48 to 79.93 (0.68%). Across the three models, the average degradation was 11.18% in BLEU scores and 5.33% in ChrF++ scores. This degradation occurs because the models must allocate their weight capacity to handle linguistic pattern mapping in two directions simultaneously, leading to feature competition in the hidden state layer. The difference in degradation margins between the two metrics provides an interesting insight: the sharper BLEU degradation indicates the models' difficulty in maintaining word order precision when trained bidirectionally, while the comparatively smaller ChrF++ degradation indicates that the models remain more robust in maintaining character-level similarity and morphological structure, even where BLEU drops sharply, as seen most clearly in ByT5.

To validate the significance of the performance differences among the three models, statistical significance tests were performed using the SacreBLEU library (see Table 2 and Table 3).

Table 2. Significance Test Results in One-Directional Scenario

Model Pair	Chrf++ Score Diff.	p-value	Interpretation
NLLB & ByT5	80.48 & 80.18	0.04495	Significant
NLLB & mT5	80.48 & 75.17	0.00099	Highly Significant
ByT5 & mT5	80.18 & 75.17	0.00099	Highly Significant

Table 3. Significance Test Results in Bidirectional Scenario

Model Pair	Chrf++ Score Diff.	p-value	Interpretation
NLLB & ByT5	79.93 & 69.02	0.00099	Highly Significant

NLLB & mT5	79.93 & 74.12	0.00099	Highly Significant
ByT5 & mT5	69.02 & 74.12	0.00099	Highly Significant

Many comparisons between models showed p-values of 0.00099, which are considered highly significant. NLLB's superiority over ByT5 and mT5 was proven not to be the result of random variation. However, this superiority cannot be attributed to architecture in isolation. NLLB (600M parameters) is roughly twice the size of mT5-small and ByT5-small (approximately 300M parameters each), is the only one of the three models purpose-built as a translation system rather than a general-purpose text-to-text model, and its pre-training already covers three of the five target regional languages (Javanese, Sundanese, and Balinese). NLLB's stronger performance therefore reflects the combined advantages of larger model capacity, task-specific pre-training, and prior exposure to related languages, rather than an inherent superiority of its architecture over mT5 or ByT5. In the one-directional scenario, the comparison of NLLB to ByT5 yielded a p-value of 0.04495, which, while higher, remains below the 0.05 significance threshold.

4.3. Comparison with Bilingual Baseline

To comprehensively test the effectiveness of the multilingual approach, the performance of the multilingual models was compared against a bilingual baseline scenario in which each model was trained separately for each language pair. A summary of the comparison results is presented in Table 4.

Table 4. Comparison Of Multilingual Model with Baseline

	mT5		ByT5		NLLB	
	BLEU	Chrf++	BLEU	Chrf++	BLEU	Chrf++
Baseline (Bidirectional)	**55.6**	**74.27**	**62.66**	**80.26**	**64.05**	**81.20**
Multilingual (Bidirectional)	50.69	74.12	48.48	69.02	62.16	79.93

The results show that the bilingual baseline consistently outperforms the multilingual approach in the bidirectional scenario across all three models, on both BLEU and ChrF++. For mT5, BLEU declined from 55.6 (baseline) to 50.69 (multilingual bidirectional), a decrease of 8.83%, and ChrF++ declined from 74.27 to 74.12, a decrease of 0.20%. ByT5 showed the sharpest decline, with BLEU dropping from 62.66 to 48.48 (22.63%) and ChrF++ dropping from 80.26 to 69.02 (14%). NLLB was the most robust of the three, with BLEU decreasing only from 64.05 to 62.16 (2.95%) and ChrF++ from 81.20 to 79.93 (1.56%). Across the three models, the average degradation was 11.49 % in BLEU scores and 5.25 % in ChrF++ scores. These results suggest that, under the bidirectional training setup used in this study, the added complexity of learning multiple language pairs and both translation directions simultaneously outweighs the benefit of cross lingual transfer, producing feature competition in the shared parameter space that reduces per pair translation quality relative to dedicated bilingual models. NLLB's comparatively small decline indicates that its massively multilingual pretraining and larger model capacity make it more resilient to this competition than mT5 or ByT5, whose smaller, less specialized architectures appear more sensitive to the added training complexity.

4.4. Performance Analysis per Language

Overall, NLLB proved to be the most robust model, achieving the highest BLEU and ChrF++ scores in 9 out of 10 language pairs. NLLB's superiority was most pronounced for Minangkabau and Dayak Maanyan, achieving BLEU scores of 73.15 for Minangkabau-to-Indonesian and 71.59 for Dayak-to-Indonesian, along with ChrF++ scores above 85 for both pairs. This suggests that NLLB's massively multilingual pre-training generalizes well even to regional languages with comparatively limited representation in its pre-training data, within the controlled, domain-specific settings evaluated here. NLLB's only weak spot was Javanese, where ByT5 achieved the highest scores in the Javanese-to-Indonesian direction, with a BLEU score of 53.06 and a ChrF++ score of 72.61, ahead of both mT5 and NLLB (see Table 5).

Table 5. Comparison of Model Performance on Each Language Pair (Bidirectional)

	mT5		ByT5		NLLB	
	BLEU	Chrf++	BLEU	Chrf++	BLEU	Chrf++
Javanese → Ind	51.42	72.18	**53.06**	**72.61**	49.75	71.56
Ind → Javanese	50.36	71.85	53.00	72.18	**54.36**	**73.71**
Bali → Ind	49.63	71.19	48.39	68.4	**64.56**	**80.39**
Ind → Bali	51.69	72.52	49.67	69.43	**55.7**	**75.91**
Dayak → Ind	51.06	72.02	43.02	65.56	**71.59**	**85.12**
Ind → Dayak	50.32	71.71	44.47	66.22	**59.9**	**81.74**
Sunda → Ind	50.83	71.86	48.38	69.54	**61.04**	**78.52**
Ind → Sunda	49.93	71.35	49.71	70.53	**60.44**	**79.19**
Minang → Ind	49.96	71.69	46.56	68.17	**73.15**	**86.52**
Ind → Minang	49.33	71.36	45.28	67	**69.67**	**86.55**

An interesting reversal appears in the Javanese-to-Indonesian pair. ByT5 achieves both the highest BLEU (53.06) and the highest ChrF++ (72.61) here, the only language pair in which NLLB is not the top performer on either metric, and the only pair in which NLLB scores lower than mT5 as well (49.75 versus 51.42 BLEU). This suggests that Javanese, despite being the regional language with the largest speaker population in the dataset, is not necessarily the easiest for NLLB's pre-trained multilingual representations to generalize to. Elsewhere in the table, mT5 outperforms ByT5 on both BLEU and ChrF++ in 8 of the 10 language pairs, including the Minangkabau and Dayak Maanyan pairs, indicating that ByT5's byte-level architecture does not provide a clear advantage over subword-based mT5 for languages with distinctive morphology. Meanwhile, mT5 consistently lags behind NLLB across most language pairs, with the widest gap of more than 23 BLEU points on the Minangkabau-to-Indonesian pair. This indicates a structural constraint in the mT5 architecture, a general text-to-text model with a SentencePiece tokenizer whose vocabulary is dominated by high-resource languages, making it less than optimal for translating regional languages without much more extensive fine-tuning or the addition of large-scale datasets.

4.5. Best Model and Implementation

Based on all evaluation results and significance tests, the NLLB model in the bidirectional scenario was determined to be the best model in this study, with a BLEU score of 62.16 and a ChrF++ score of 79.93. NLLB's superiority is driven by three main factors: its architecture, specifically designed as a machine translation system, pre-training knowledge that already covers several Indonesian regional languages such as Javanese, Sundanese, and Balinese, and a BPE tokenization system with a vocabulary of 256,000 subwords curated using temperature-based sampling to ensure adequate representation of low-resource languages.

Model implementation is carried out to deploy the best model into a system usable directly by users. This system is a web-based application built with the Streamlit framework and Python, enabling interactive, real-time inference. Users enter input by selecting the source and target languages, then entering source-language sentences through the application interface (see Figure 5). The system then processes the input sentences through a tokenization stage. Next, the sentences are fed into the selected NLLB model. The model outputs translated sentences into the target language. These predictions are displayed directly in the application interface (see Figure 6). Users can also input reference sentences to calculate BLEU and ChrF++ scores. This implementation aims to demonstrate the model's feasibility in real-world scenarios and facilitate the evaluation and demonstration of research results.

It should be noted that the high BLEU and ChrF++ scores reported in this study are partly attributable to the controlled nature of the corpus. All sentence pairs originate from a single domain (news articles) and were translated by trained regional language teachers using a consistent methodology. This stylistic homogeneity between the training and test data results in

higher n-gram overlap than in models evaluated on open-domain corpora, a gap that is quantified empirically in Section 4.6. It should be considered when comparing results with prior work. System interface is presented in Figure 5 and System Output Display is presented in Figure 6.

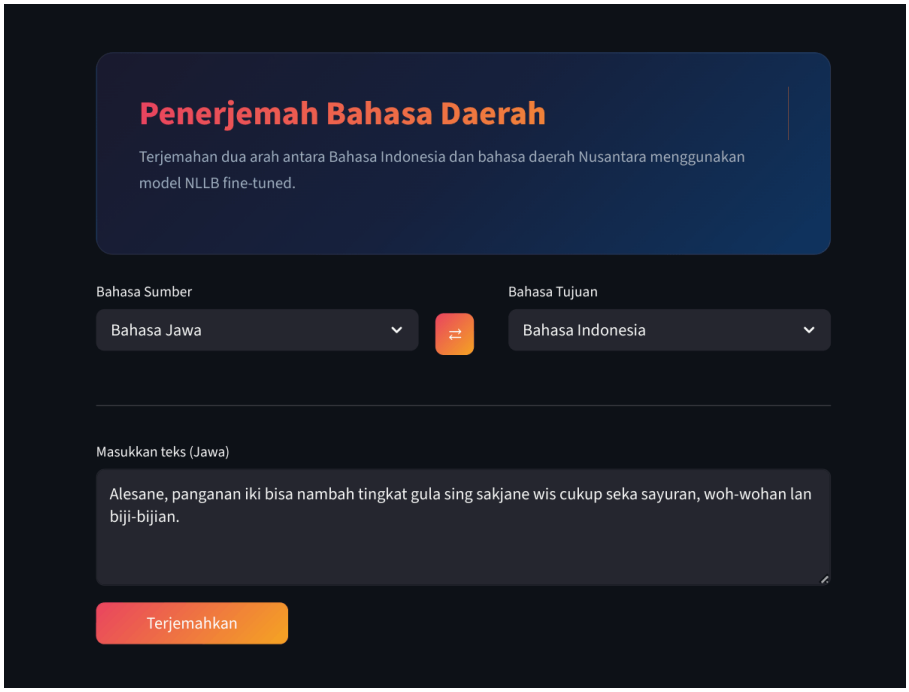


Figure 5. System Interface

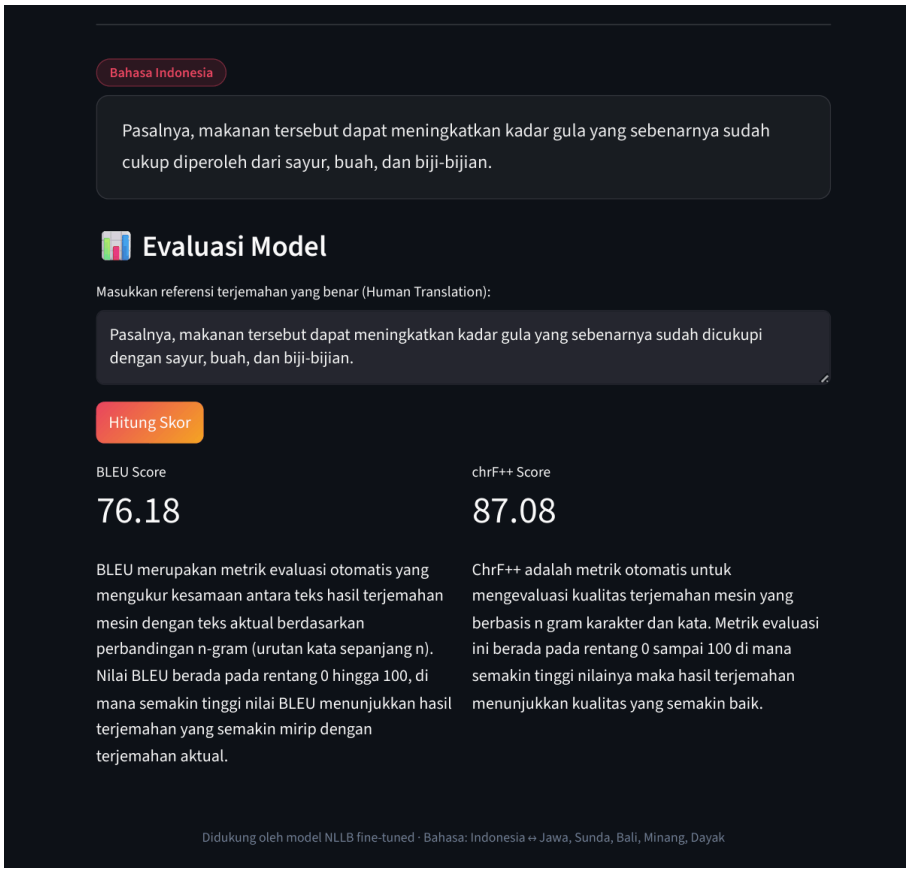


Figure 6. System Output Display

To further support the quantitative evaluation results, the following presents representative translation examples from the implemented NLLB model, along with explanations of the observed outputs (see Table 6).

Table 6. Translation Output Examples

	Successful Case	**Failure Case**
Language Pair	Minangkabau → Indonesian	Dayak Maanyan → Indonesian
Source	Jatim sajauah ko alah mangumpuaan total 210 medali, nan tadiri dari 83 ameh, 67 perak, jo 60 parunggu. *("East Java has so far collected a total of 210 medals, consisting of 83 gold, 67 silver, and 60 bronze.")*	Serangan pangayauan biasa ni terjadi hang ume atawa andri nutung erang kalewu, dan nyamalis katuluh ulun hang wuang lewu hantek here kabur. *("Headhunting raids usually occur in the fields or by burning down a house, and beheading all the house's inhabitants as they flee.")*
Reference	Jatim sejauh ini sudah mengumpulkan total 210 medali, yang terdiri dari 83 emas, 67 perak, dan 60 perunggu. *("East Java has so far collected a total of 210 medals, consisting of 83 gold, 67 silver, and 60 bronze.")*	Serangan pangayauan biasanya terjadi di ladang atau dengan membakar sebuah rumah dan memenggal semua penghuninya ketika mereka melarikan diri. *("Headhunting raids usually occur in the fields or by burning down a house and beheading all its inhabitants as they flee.")*
Model Output	Jatim sejauh ini telah mengumpulkan total 210 medali, yang terdiri dari 83 emas, 67 perak, dan 60 perunggu. *("East Java has so far collected a total of 210 medals, consisting of 83 gold, 67 silver, and 60 bronze.")*	Serangan penghayuan biasanya terjadi di dalam rumah atau dengan memukul di dalam rumah, dan menyambar semua orang di dalam rumah saat mereka melarikan diri. *("Headhunting raids usually occur inside the house or by hitting inside the house, and striking everyone inside the house as they flee.")*
BLEU	87.52	15.98
ChrF++	93.67	54.13

In a successful case from Minangkabau→Indonesian, the model produced a translation with a sentence-level BLEU score of 87.52 and a ChrF++ score of 93.67. The multi-head attention mechanism successfully captured long-distance dependencies, correctly linking the subject and predicate across the full sentence, and the decoder reconstructed grammatically natural Indonesian word order without any loss of meaning.

In a failure case from the Dayak Maanyan to Indonesian direction, the model produced a sentence level BLEU score of only 15.98, despite a relatively higher ChrF++ score of 54.13. Three specific lexical errors were observed: the phrase hang ume (meaning "in the field") was incorrectly rendered as "di dalam rumah" (in the house); nutung (meaning "to burn") was translated as "memukul" (to hit); and nyamalis (meaning "to behead") was translated as "menyambar" (to strike). These errors reflect the limitations of the encoder under very limited training data. One plausible explanation, consistent with the observed lexical substitutions toward more typologically common vocabulary, is that the shared parameter space caused representation borrowing from higher resource languages; confirming this would require attention or attribution analysis, which we leave to future work. While cross lingual transfer generally benefits low resource translation, in this case it led to systematic lexical misalignment. Notably, the decoder still produced syntactically well-formed Indonesian sentences, which explains the comparatively higher ChrF++ score. Sentence structure was preserved, but word level semantics were incorrect. The gap between BLEU (15.98) and ChrF++ (54.13) confirms that the failure is lexical rather than structural, a distinction BLEU captures more sensitively than ChrF++. These examples confirm that while the model performs reliably on higher resource languages such as Javanese,

Sundanese, and Balinese, it remains vulnerable to lexical errors on critically low resource languages such as Dayak Maanyan, consistent with the per language quantitative results reported in Table 5.

4.6. Out-of-Corpus Evaluation

To directly test whether the high in-corpus scores reported in Sections 4.2-4.4 reflect general translation quality or domain-specific adaptation, all three bidirectional models were evaluated on 100 held-out sentence pairs for Indonesian to Javanese, Minang, Sundanese, and Balinese from FLORES-200, a benchmark drawn from Wikipedia-domain text rather than news. Table 7 shows the resulting scores alongside the in-corpus Indonesian-to-Javanese results from Table 5.

Table 7. In-Corpus vs. Out-of-Corpus Performance

Language Pair	Model	In-Corpus BLEU	Out-of-Corpus BLEU	BLEU Gap	In-Corpus Chrf++	Out-of-Corpus Chrf++	Chrf++ Gap
Ind → Javanese	mT5	50.36	15.35	35.01	71.85	42.89	28.96
	ByT5	53.00	20.53	32.47	72.18	48.39	23.79
	NLLB	54.36	27.00	27.36	73.71	53.79	19.92
Ind → Minang	mT5	49.33	17.73	31.6	71.36	45.58	25.78
	ByT5	45.28	25.21	20.07	67	53.65	13.35
	NLLB	69.67	29.19	40.48	86.55	56.24	30.31
Ind → Sunda	mT5	49.93	14.85	35.08	71.35	42.3	29.05
	ByT5	49.71	20.69	29.02	70.53	48.21	22.32
	NLLB	60.44	27.57	32.87	79.19	54.79	24.4
Ind → Bali	mT5	51.69	13.78	37.91	72.52	41.21	31.31
	ByT5	49.67	18.96	30.71	69.43	46.96	22.47
	NLLB	55.7	26.63	29.07	75.91	52.65	23.26

All three models show a substantial drop in performance when moving from the in-corpus test set to FLORES-200. BLEU scores fall by 27 to 35 points, and ChrF++ scores fall by 13 to 31 points, an average BLEU gap of 31.6 points and ChrF++ gap of 24.2 points across the three models. NLLB shows the smallest gap in both metrics (BLEU 27.36, ChrF++ 19.92), consistent with its greater robustness observed elsewhere in this study, while mT5 shows the largest gap (BLEU 35.01, ChrF++ 28.96). This confirms the stylistic homogeneity noted in Section 4.5: because training and test data share a single domain (news articles) and a consistent translation methodology, in-corpus scores overstate how these models would perform on stylistically distinct, open-domain text. The scores reported throughout Section 4 should therefore be read as evidence of strong in-domain adaptation under controlled, domain-specific conditions, not as evidence of general-purpose translation accuracy.

4.7. Limitations

The results of this study should be interpreted with numerous caveats. First, each model configuration was trained with one training seed, although paired bootstrap resampling was employed to compare model outputs. Thus, the statistical analysis captures test sample uncertainty but not stochastic effects like parameter initialization, mini-batch ordering, dropout, and optimization during model fine-tuning. Thus, the reported model differences should be viewed as performance differences under the specific training configuration utilized in this investigation, not as proof of model superiority over multiple training sessions. To better test model stability, future studies should repeat the experiments with several training seeds and report mean performance, standard deviation, and confidence intervals.

Second, beyond the empirical generalization gap quantified in Section 4.6, this evaluation design carries a construct validity limitation. Testing exclusively on a single, highly controlled, homogeneous corpus evaluates how well each model adapts to one specific domain and translation style, not the broader construct of generalizable translation quality that BLEU and

ChrF++ scores are conventionally interpreted as proxying for. Even with a leakage-free, group-based split, training and test data drawn from the same source sentences, translators, and domain cannot, by construction, establish whether a model has learned transferable translation competence or has instead specialized to the idiosyncrasies of this corpus. The results reported in Sections 4.1 to 4.5 should accordingly be read as evidence of narrow domain adaptation rather than generalizable translation quality; the out-of-corpus evaluation in Section 4.6 provides a first, partial empirical test of this distinction for one language pair, but does not resolve it for the study as a whole.

Third, this study's high BLEU and ChrF++ scores are likely in-corpus performance. The controlled corpus of Indonesian sentences from news items and regional-language translations prepared via a consistent translation approach provides training, validation, and test data. Group-based splitting prevented translations from the same Indonesian source sentence from being spread over the training and test sets, but the subsets still share domains, lexical traits, and translation styles. Homogeneity may increase n-gram and character-level overlap between model predictions and reference translations, boosting automatic evaluation scores.

It cannot yet apply the results to open-domain or natural translation scenarios. The present study's out-of-corpus evaluation (Section 4.6) was limited to a single language pair and direction. The other four regional languages, and the reverse (regional-to-Indonesian) direction, remain untested against an external benchmark, so the magnitude of domain-specific inflation observed for Indonesian-to-Javanese may not generalize to the other language pairs in this study. Therefore, future study should compare the models' performance with internal test findings utilizing independently gathered sentences from different sources, genres, and communication settings. Such evaluation would better estimate the models' generalization potential beyond the controlled corpus used in this investigation.

5. Conclusion

This study presented a systematic empirical comparison of three multilingual Transformer based Neural Machine Translation models, mT5, ByT5, and NLLB, for bidirectional translation between five Indonesian regional languages (Javanese, Sundanese, Balinese, Minangkabau, and Dayak Maanyan) and Indonesian under low-resource conditions. Evaluation results show that, contrary to the expected benefit of cross lingual transfer, the multilingual approach in the bidirectional scenario underperformed the bilingual baseline across all three models, with an average degradation of 11.49% in BLEU scores and 5.25% in ChrF++ scores. NLLB was the most resilient to this degradation, while ByT5 was the most affected, suggesting that the benefit of parameter sharing across languages in this study was outweighed by the added burden of jointly learning multiple language pairs and both translation directions within a single model. The NLLB model in the bidirectional scenario proved to be the best-performing model, with a BLEU score of 62.16 and a ChrF++ score of 79.93; these results apply to controlled, domain-specific settings, and statistical significance tests confirmed that its superiority over ByT5 and mT5 was not due to random variation. The ByT5 model demonstrated a competitive advantage specifically for Javanese, confirming that byte-level processing can be robust for languages with non-standard morphology. The Streamlit based implementation of the best-performing model illustrates one potential practical application of these findings for regional language preservation, though further evaluation would be needed to establish real-world deployment feasibility. For future research, it is recommended to expand the size and domain diversity of the training corpus to improve model generalization, explore data imbalance handling techniques such as temperature-based sampling, and supplement automatic evaluation with human evaluation to obtain a more comprehensive assessment of translation quality in terms of semantics and linguistic naturalness.

References

[1] C. Moseley and A. Nicolas, *Atlas of the world's languages in danger*. UNESCO, 2010.

[2] R. Ramli, F. H. Setyawan, and R. Rampeng, "The threat of lingua globalization towards vernaculars in South Sulawesi," *ETNOSIA: Jurnal Etnografi Indonesia*, vol. 6, no. 2, Dec. 2021, doi: 10.31947/etnosia.v6i2.18197.

[3] A. W. Mahastama, "Model Berbasis Aturan untuk Transliterasi Bahasa Jawa dengan Aksara Latin ke Aksara Jawa," *Jurnal Buana Informatika*, vol. 13, no. 02, pp. 146–154, Oct. 2022, doi: 10.24002/jbi.v13i02.6526.

[4] L. Susanto, R. Diandaru, A. Krisnadhi, A. Purwarianti, and D. 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: Association for Computational Linguistics, Nov. 2023, pp. 100–115. doi: 10.18653/v1/2023.sealp-1.8.

[5] C. Samuel and I. T. Ali, "Batak Toba language-Indonesian machine translation with transfer learning using no language left behind," *International Journal of Advances in Applied Sciences*, vol. 13, no. 4, pp. 830–839, Dec. 2024, doi: 10.11591/ijaas.v13.i4.pp830-839.

[6] Q. Wang and J. Zhang, "Parameter Differentiation based Multilingual Neural Machine Translation," *ArXiv*, Dec. 2021, doi: 10.48550/arXiv.2112.13619.

[7] W. Wongso, A. Joyoadikusumo, B. S. Buana, and D. Suhartono, "Many-to-Many Multilingual Translation Model for Languages of Indonesia," *IEEE Access*, vol. 11, pp. 91385–91397, 2023, doi: 10.1109/ACCESS.2023.3308818.

[8] Y. Denny Prabowo, M. Maslim, Y. S. Purnomo W.P., and V. Citrayasa, "Korpus Paralel Dwibahasa Indonesia–Dayak Ma'anyan: Dataset Terjemahan Kalimat Lintas Domain," Sep. 2026, *Zenodo*. doi: 10.5281/zenodo.22499315.

[9] Y. Denny Prabowo, M. Maslim, Y. S. Purnomo W.P., and V. Citrayasa, "Korpus Paralel Dwibahasa Indonesia–Bali: Dataset Terjemahan Kalimat Lintas Domain," Sep. 2026, *Zenodo*. doi: 10.5281/zenodo.22498298.

[10] Y. Denny Prabowo, M. Maslim, Y. S. Purnomo W.P., and V. Citrayasa, "Korpus Paralel Dwibahasa Indonesia–Minang: Dataset Terjemahan Kalimat Lintas Domain," Sep. 2026, *Zenodo*. doi: 10.5281/zenodo.22496100.

[11] Y. Denny Prabowo, M. Maslim, Y. S. Purnomo W.P., and V. Citrayasa, "Korpus Paralel Dwibahasa Indonesia–Sunda: Dataset Terjemahan Kalimat Lintas Domain," Sep. 2026, *Zenodo*. doi: 10.5281/zenodo.22494444.

[12] Y. Denny Prabowo, M. Maslim, Y. S. Purnomo W.P., and V. Citrayasa, "Korpus Paralel Dwibahasa Indonesia–Jawa: Dataset Terjemahan Kalimat Lintas Domain," Sep. 2026, *Zenodo*. doi: 10.5281/zenodo.22490458.

[13] N. Bhadwal, P. Agrawal, and V. Madaan, "A Machine Translation System from Hindi to Sanskrit Language using Rule based Approach," *Scalable Computing: Practice and Experience*, vol. 21, no. 3, pp. 543–554, Aug. 2020, doi: 10.12694/scpe.v21i3.1783.

[14] D. Torregrosa *et al.*, "Aspects of Terminological and Named Entity Knowledge within Rule-Based Machine Translation Models for Under-Resourced Neural Machine Translation Scenarios," *ArXiv*, vol. abs/2009.13398, p., Sep. 2020.

[15] A. Way, "A Critique of Statistical Machine Translation," *Linguistica Antverpiensia, New Series – Themes in Translation Studies*, vol. 8, pp. 17–41, Oct. 2021, doi: 10.52034/lanstts.v8i.243.

[16] L. Macken and E. Lefever, "Translational equivalence in Statistical Machine Translation or meaning as co-occurrence," *Linguistica Antverpiensia, New Series – Themes in Translation Studies*, vol. 7, pp. 193–208, Oct. 2021, doi: 10.52034/lanstts.v7i.215.

[17] J. Dugonik, M. Sepesy Maučec, D. Verber, and J. Brest, "Reduction of Neural Machine Translation Failures by Incorporating Statistical Machine Translation," *Mathematics*, vol. 11, no. 11, p. 2484, May 2023, doi: 10.3390/math11112484.

[18] A. Vaswani *et al.*, "Attention Is All You Need," *ArXiv*, Aug. 2023, doi: https://doi.org/10.48550/arXiv.1706.03762.

[19] C. Fang, A. Zhou, and Z. Wang, "An Algorithm–Hardware Co-Optimized Framework for Accelerating N:M Sparse Transformers," *IEEE Trans. Very Large Scale Integr. VLSI Syst.*, vol. 30, pp. 1573–1586, 2022, doi: 10.1109/tvlsi.2022.3197282.

[20] M. Costa-Jussà *et al.*, "Scaling neural machine translation to 200 languages," *Nature*, vol. 630, pp. 841–846, 2024, doi: 10.1038/s41586-024-07335-x.

[21] C. Mullov, Q. Pham, and A. Waibel, "Decoupled Vocabulary Learning Enables Zero-Shot Translation from Unseen Languages," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 6693–6709. doi: 10.18653/v1/2024.acl-long.362.

[22] L. B. San, H. Sujaini, and T. Tursina, "Uji Nilai Akurasi pada Neural Machine Translation (NMT) Bahasa Indonesia ke Bahasa Tiochiu Pontianak dengan Mekanisme Attention Bahdanau," *Jurnal Edukasi dan Penelitian Informatika (JEPIN)*, vol. 9, no. 3, p. 362, Dec. 2023, doi: 10.26418/jp.v9i3.63346.

[23] R. B. Primandhika, M. N. Munawar, and A. R. Saifullah, "Experiment on a Transformer Model Indonesian-to-Sundanese Neural Machine Translation with Sundanese Speech Level Evaluation," in *Proceedings of the Thirteenth Conference on Applied Linguistics (CONAPLIN 2020)*, 2021. doi: 10.2991/assehr.k.210427.069.

[24] M. Giri, R. Adjie, R. Dwiyansaputra, F. Bimantoro, and A. Aranta, "Algorithm Design for Converting Indonesian Latin to Sasak Latin Using the Sequence-To-Sequence Transformers Method," *Journal of Computer Science and Informatics Engineering*, vol. 9, no. 1, Jun. 2025, doi: 10.29303/jcosine.v9i1.619.

[25] L. Xue *et al.*, "mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Stroudsburg, PA, USA: Association for Computational Linguistics, Mar. 2021, pp. 483–498. doi: 10.18653/v1/2021.naacl-main.41.

[26] L. Xue *et al.*, "ByT5: Towards a Token-Free Future with Pre-trained Byte-to-Byte Models," *Trans. Assoc. Comput. Linguist.*, vol. 10, pp. 291–306, Mar. 2022, doi: 10.1162/tacl_a_00461.

[27] R. K. Mahabadi, Y. Belinkov, and J. Henderson, "Variational Information Bottleneck for Effective Low-Resource Fine-Tuning," Jun. 2021, doi: 10.48550/arXiv.2106.05469.