Advancing Cross-Cultural Natural Language Processing with a Focus on Sundanese Language and Contextual Nuances

Anggi Muhammad Rifai[1*], Ema Utami[2], Amali[3],
Muhammad Fatchan[4], and Muhamad Ekhsan[5]
[1,3,4]Department of Informatics Engineering, Faculty of Engineering, Universitas Pelita Bangsa
Jawa Barat, Indonesia 17530
[2]Doctoral Program in Informatics, Universitas Amikom Yogyakarta
DI Yogyakarta, Indonesia 55281
[5]Department of Management, Faculty of Economics and Business, Universitas Pelita Bangsa
Jawa Barat, Indonesia 17530
Email: [1]anggimuhammad@pelitabangsa.ac.id, [2]ema.u@amikom.ac.id, [3]amali@pelitabangsa.ac.id,
[4]fatchan@pelitabangsa.ac.id, [5]muhammad.ekhsan@pelitabangsa.ac.id

***Abstract*—The Sundanese language, as one of Indonesia's regional tongues, holds deep cultural value but is still underrepresented in computational linguistics. The research addresses this gap by developing a translation model between Sundanese and Indonesian using a transformer-based sequence-to-sequence (Seq2Seq) architecture. With a parallel dataset of 3,616 sentence pairs, the model is fine-tuned to capture linguistic and contextual subtleties. The evaluation yields strong results: Bilingual Evaluation Understudy (BLEU) score of 44.12, Recall - Oriented Understudy for Gisting Evaluation (ROUGE)-1 F1-Score of 0.72, and ROUGE-L F1-Score of 0.71. Those demonstrate high translation quality despite limited data. Unlike earlier Sundanese translation studies that rely on Recurrent Neural Network (RNN), Long Short-Term Memory (LSTM), or standard transformer models, this research uniquely leverages the multilingual pretrained M2M100 Transformer, enabling transfer learning from high-resource languages to improve low-resource performance. These outcomes highlight the model's potential for real-world applications, such as translation tools for education and cultural exchange. The research emphasizes the importance of improving access to Sundanese texts and promoting its digital presence to aid in language preservation. Overall, the research not only advances Natural Language Processing (NLP) research for low-resource languages but also reinforces the importance of integrating regional languages like Sundanese into modern technology. Building upon prior studies on Indonesian–Sundanese translation, the research novelty lies in fine-tuning a multilingual Seq2Seq Transformer that captures both linguistic and contextual nuances, thereby setting a new benchmark for low-resource language processing.**

Received: July 01, 2025; received in revised form: Sep. 08, 2025; accepted: Sep. 13, 2025; available online: April 09, 2026.
*Corresponding Author

***Index Terms*—Cross-Cultural Natural Language Processing, Seq2Seq Transformer, Sundanese Language Translation, Low-Resource Language Processing**

I. Introduction

THE rapid advancements in Natural Language Processing (NLP) have significantly transformed how languages are processed and understood [1]. These advancements have opened new possibilities for multilingual communication [2]. They have fostered cultural exchange across the globe. By enabling machines to comprehend and generate human language, NLP technologies have bridged linguistic barriers [3]. They have facilitated global interactions. However, despite these advancements, a stark disparity persists in the representation of languages within NLP research [4].

Predominantly, the benefits of extensive datasets and sophisticated models have been concentrated on globally dominant languages [5]. These languages, such as English, Chinese, and Spanish, have a vast digital presence and strong economic influence [6]. As a result, they have seen continuous innovation and refinement in computational processing [7]. In contrast, many regional and minority languages are left underrepresented. This situation creates a profound imbalance. It reflects not only technological inequities but also the potential loss of cultural and linguistic heritage [8].

This imbalance is particularly evident in languages spoken in the Global South. Linguistic diversity is abundant in this region but often neglected in NLP advancements [9]. Sundanese, a regional language of

Indonesia, exemplifies this issue [10]. It is spoken by over 40 million people but remains underrepresented in digital linguistic tools and computational resources [11]. Despite its widespread use and cultural significance, Sundanese faces significant technological gaps [12]. This lack of support hinders its integration into modern systems [13]. For example, translation tools and AI-driven applications often exclude it. As a result, Sundanese remains isolated from the global digital ecosystem [14].

Without adequate NLP research and development, the unique linguistic structures of Sundanese are at risk [15]. Its cultural nuances may be overshadowed by more dominant languages. Addressing this gap is critical, as it can enhance NLP inclusivity and preserve linguistic diversity [16]. Regional languages like Sundanese must be preserved to ensure their continued relevance [17]. In an interconnected, technology-driven world, it is vital for their cultural survival.

Recent research has also explored enhancements to Neural Machine Translation (NMT) models, such as Bidirectional Encoder Representations from Transformers – Joint Audio Text Model (BERT-JAM), which integrates multi-layer representations from BERT to improve translation performance [18]. While this method is not directly applied in this research, it illustrates the broader advances in NMT that motivate continued exploration of alternative architectures. This research focuses on addressing the gap in NLP for the Sundanese language by developing a translation model between Indonesian and Sundanese [19]. Sundanese, as a low-resource language, presents unique challenges [20]. One of the major obstacles is the limited availability of data [21]. The quality of existing data is another concern. Additionally, capturing the linguistic and contextual nuances of Sundanese is particularly difficult [22]. These challenges complicate the development of effective computational models. To tackle these issues, the researchers leverage a dataset of 3,616 parallel sentences. This dataset serves as the foundation for training the translation model. The model employs a sequence-to-sequence (Seq2Seq) architecture [23]. Seq2Seq models are based on transformer technology, which has proven to be highly effective in NLP tasks [24]. The researchers specifically use the multilingual pretrained M2M100 Transformer as the Seq2Seq model, which is fine-tuned for the Indonesian–Sundanese translation task. Transformers excel at capturing contextual dependencies between words and phrases. This ability is crucial for generating translations that are both accurate and contextually appropriate [25].

The primary objective of this research is to explore the capabilities of the model in handling linguistic nuances. In particular, the research focuses on how well the model can capture the unique characteristics of the Sundanese language. These linguistic subtleties are often lost in simpler translation models [26]. Therefore, the research aims to enhance the model's ability to generate high-quality, human-like translations.

Beyond the technical objectives, this research has practical applications. One of the key areas of focus is the development of translation services. These services can be invaluable in educational settings, where accurate translations are essential for cross-cultural communication [27]. Such tools can also be applied in broader cross-cultural contexts, facilitating better understanding and exchange [28]. By focusing on both linguistic and contextual challenges, this research aims to provide a meaningful contribution to NLP for low-resource languages like Sundanese [29].

In addition to developing the translation model, the research also examines the challenges encountered during the process. One of the primary challenges is the quality of the data used to train the model [30]. Data quality issues, such as inconsistencies and errors, can significantly impact the effectiveness of the model. These issues are particularly pronounced when working with low-resource languages like Sundanese. The limited availability of reliable, high-quality data further complicates the development of effective models. Another challenge is the size of the corpus. With only 3,616 parallel sentences, the corpus is relatively small. This limited size can hinder the model's ability to generalize across a wide range of language variations [31]. As a result, the model's performance may not be as robust or accurate as desired. Addressing these data-related challenges is critical for improving the model's performance in real-world applications.

To evaluate the model's effectiveness, the research uses various metrics, including Bilingual Evaluation Understudy (BLEU) and Recall - Oriented Understudy for Gisting Evaluation (ROUGE) scores [32]. These metrics are commonly used in the NLP community to assess the quality of machine-generated translations. By applying these metrics, the researchers aim to establish a benchmark for future work in the field of Sundanese NLP [33]. The evaluation results will provide valuable insights into the model's strengths and areas for improvement.

The research novelty lies in its dual focus. First, it advances technical solutions for Sundanese language translation by building upon existing works such as Bidirectional Long Short-Term Memory (LSTM)-based models, transformer approaches with Sundanese speech level evaluation, and Recurrent Neural Network (RNN)-based neural translators [20, 34]. Second, it contributes to the preservation of a vital regional

Cite this article as: A. M. Rifai, E. Utami, Amali, M. Fatchan, and M. Ekhsan, “Advancing cross-cultural natural language processing with a focus on Sundanese language and contextual nuances”, CommIT Journal 20(1), 155–167, 2026.

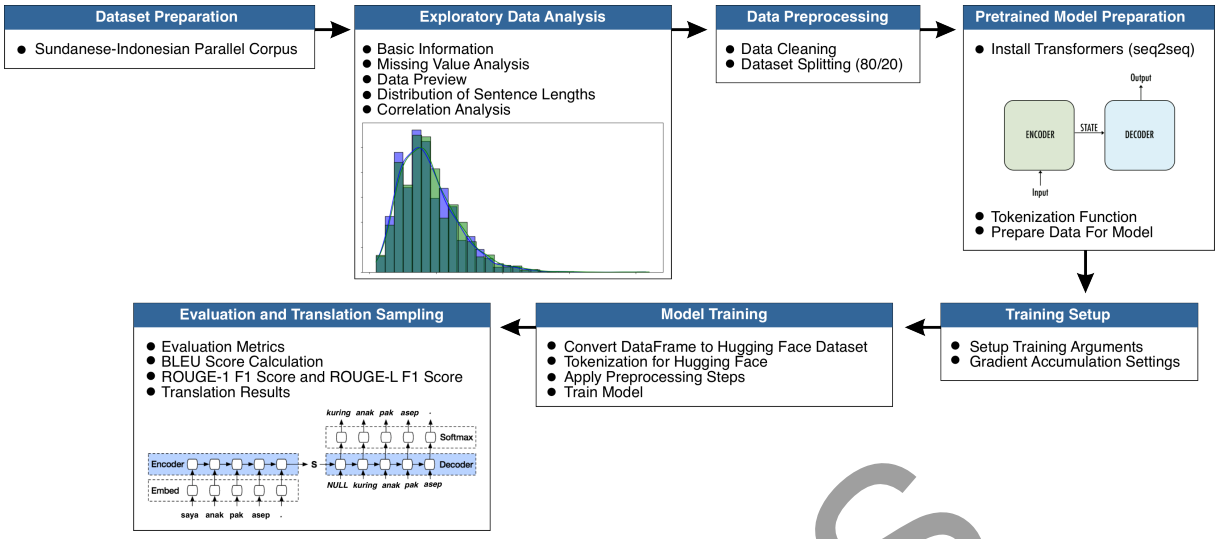

Fig. 1. Research proposal. Note: *saya anak pak Asep* and *kuring anak pak Asep* mean I am Mr. Asep's son. Bilingual Evaluation Understudy (BLEU) and Recall - Oriented Understudy for Gisting Evaluation (ROUGE).

TABLE I
BASIC INFORMATION OF THE DATASET.

Column	Non-Null Count	Data Type
Indonesian	3,616	Object
Sundanese	3,616	Object

language. This dual focus strengthens current research in the field, which often prioritizes technical aspects without considering cultural significance. Through its findings, the research aims to bridge linguistic divides and make Sundanese more accessible in the digital era.

II. RESEARCH METHOD

In the research, a comprehensive methodology is applied. Figure 1 provides an overview of the complete research pipeline for the Indonesian to Sundanese translation task. The pipeline is structured into seven sequential stages: (1) dataset preparation using the Indonesian–Sundanese parallel corpus, (2) Exploratory Data Analysis (EDA) including data preview, distribution analysis, and correlation checks, (3) data preprocessing, such as cleaning and dataset splitting, (4) pretrained model preparation with the installation of transformers and tokenization setup, (5) training setup defining hyperparameters and gradient accumulation, (6) model training using the Seq2Seq architecture with Hugging Face integration, and (7) evaluation and translation sampling with BLEU and ROUGE metrics. Each stage provides transparency and ensure reproducibility for future research.

A. Dataset Preparation

The dataset used is sourced from the Indonesian–Sundanese parallel corpus, a curated collection of bilingual sentence pairs [34]. It contains 3,616 parallel sentences, with each sentence in Sundanese matched with its corresponding Indonesian translation. To ensure the robustness and consistency of this parallel corpus, manual alignment checks are conducted to verify sentence pair accuracy, while duplicate and incomplete entries are removed. This cleaning process, combined with a reproducible 80–20 split, ensures that the dataset maintained a high level of reliability for model training and evaluation.

This dataset forms the foundation for the model's training and evaluation phases, enabling it to learn linguistic patterns and translation rules between the two languages. The use of this parallel corpus is essential in addressing the challenges of low-resource languages like Sundanese, providing a robust and consistent dataset that ensures the quality and reliability of the translation model. By leveraging this dataset, the research aims to improve the accuracy and efficiency of Indonesian–Sundanese machine translation systems.

B. Exploratory Data Analysis

The initial exploratory data analysis of the Indonesian–Sundanese parallel dataset provides a comprehensive understanding of its structure and quality. The dataset consists of 3,616 entries with two columns: indonesian and sundanese. Each column has type object, indicating textual data, and the dataset occupies a memory of 56.6 KB, as shown in Table I. Importantly,

Cite this article as: A. M. Rifai, E. Utami, Amali, M. Fatchan, and M. Ekhsan, "Advancing cross-cultural natural language processing with a focus on Sundanese language and contextual nuances", CommIT Journal 20(1), 155–167, 2026.

TABLE II
SAMPLE DATASET SHOWING ENGLISH, INDONESIAN, AND SUNDANESE TRANSLATIONS.

English	Indonesian	Sundanese
The arrival of Arab and Persian traders...	*Kedatangan pedagang-pedagang Arab dan Persia m...*	*Kadatangan padagang-padagang Arab jeung Pérsia...*
Minangkabau customs and culture follow the line of...	*Adat dan budaya Minangkabau mengikuti garis ib...*	*Adat jeung budaya Minangkabau ngagu-ratkeun ka...*
Based on the influence of Hinduism and Buddhism, be...	*Berdasarkan pengaruh agama Hindu dan Budha, be...*	*Tī pangaruh agama Hindu jeung Buddha, sababara...*
In the western region of Java, in the 4th century BC...	*Di daerah barat Pulau Jawa, pada abad ke-4 sam...*	*Di wewengkon kulon Pulo Jawa, dina abad ka-4 n...*
In the 7th century, there was a Malay kingdom centered...	*Pada abad ke-7 ada kerajaan Melayu yang pusatn...*	*Dina abad ka-7 aya Karajaan Malayu nu puseurna...*

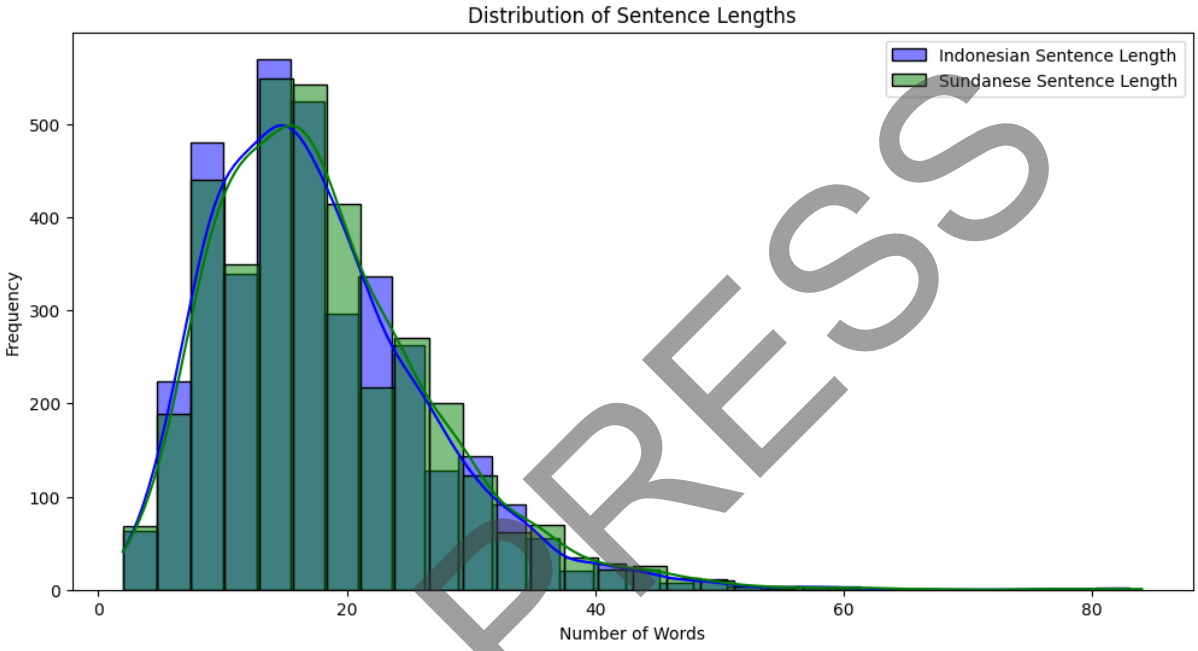

Fig. 2. Distribution analysis of sentence lengths.

a missing value analysis reveals that both columns contain 0 missing values, ensuring that the dataset is complete and ready for further processing.

A preview of the dataset confirms its quality, displaying parallel sentence pairs where each Indonesian sentence is paired with its Sundanese translation. For instance, the first entry highlights a sentence discussing the arrival of Arab and Persian traders (*Kedatangan pedagang-pedagang Arab dan Persia...*), faithfully translated into Sundanese, as shown in Table II. This confirms the dataset's alignment and appropriateness for translation tasks.

To understand the dataset's characteristics, a distribution analysis of sentence lengths is conducted, as illustrated in Fig. 2. The sentence lengths of both languages are primarily concentrated between 10 to 30 words, with a similar distribution pattern observed across both languages. This consistency suggests that the dataset is well-structured and reflects linguistic parallels between Sundanese and Indonesian.

A correlation analysis further supports this observation, showing a strong positive correlation (0.986) between the sentence lengths of the two languages as shown in Fig. 3. This high correlation signifies linguistic consistency, It also indicates that the dataset effectively captures parallelism in sentence structure, an essential feature for training machine translation models.

C. Data Preprocessing

Data preprocessing is a critical step in preparing textual data for NLP tasks. In the research, the preprocessing process begins with the installation of the Natural Language Toolkit (NLTK), a powerful library

Cite this article as: A. M. Rifai, E. Utami, Amali, M. Fatchan, and M. Ekhsan, “Advancing cross-cultural natural language processing with a focus on Sundanese language and contextual nuances”, CommIT Journal 20(1), 155–167, 2026.

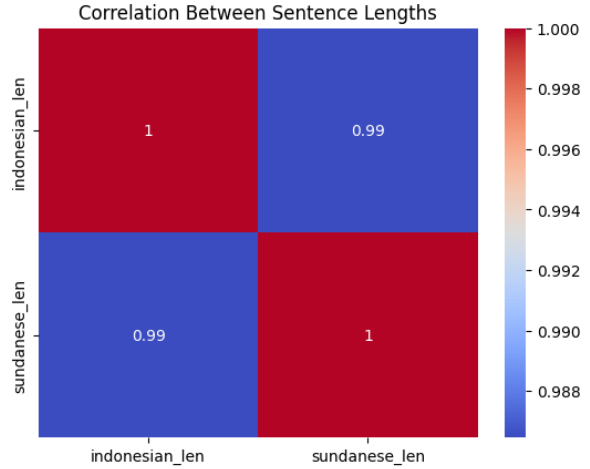

Fig. 3. Correlation between sentence lengths.

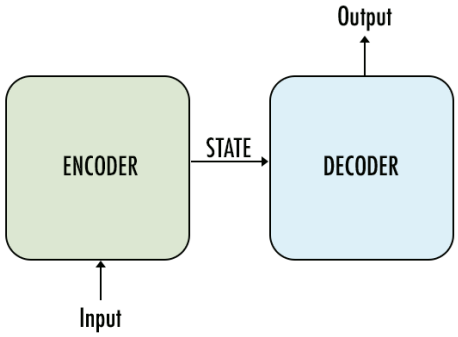

Fig. 4. Pretrained models.

for text analysis [30]. Key resources, including *punkt*, *stopwords*, and *punkt_tab*, are downloaded to enable tokenization, stopword removal, and other essential operations [35]. A preprocessing function is then implemented to clean and standardize the text data [36]. This function converts all text to lowercase for consistency and uses tokenization to split sentences into individual words. To ensure that only relevant information is retained, non-alphabetic tokens such as numbers and special characters are removed. This cleaning process is applied to both the Indonesian and Sundanese text data, resulting in two new columns, *indonesian_cleaned* and *sundanese_cleaned*, which contain the preprocessed text.

To facilitate model training and evaluation, the dataset is then split into training and testing subsets. Using an 80–20 split ratio, 80% of the data are allocated for training, while the remainings (20%) are reserved for testing. A random seed (`random_state=42`) is employed to ensure reproducibility of the split. This systematic preprocessing approach not only cleans and prepares the text data but also preserves its linguistic characteristics, ensuring the dataset is well-suited for further analysis and machine learning applications.

Despite the challenges of working with a low-resource language such as Sundanese, these preprocessing steps play a crucial role in mitigating data quality issues. By systematically removing inconsistencies, ensuring alignment between parallel sentences, and applying a reproducible train-test split, the dataset is prepared in a form that reduces the negative effects of noise and imbalance. Furthermore, by fine-tuning the multilingual pretrained M2M100 model on this cleaned dataset, the researchers leverage transfer learning from high-resource languages, helping to compensate for the limited size and variability of the corpus. This combination of careful preprocessing and fine-tuning ensured that the model can achieve robust performance despite inherent data constraints.

D. Pretrained Model Preparation

In preparing a pretrained model for machine translation tasks, the transformers library and Hugging Face Hub are utilized to access state-of-the-art models and tools. The M2M100 multilingual model, capable of handling translations across numerous languages, is chosen for the research [31]. First, the necessary libraries, including transformers and Hugging Face Hub, are installed to facilitate seamless integration with pretrained models. The tokenizer and model for the selected multilingual translation model are then initialized using the AutoTokenizer and Auto Model for Seq2SeqLM classes Fig. 4.

A tokenization function, `preprocess_batch`, is developed to prepare the text data for input into the model. This function uses the tokenizer to process the cleaned Indonesian and Sundanese text. The tokenization process includes truncating sentences to a maximum length of 128 tokens and padding shorter sentences to this length to maintain uniformity. The function returns tokenized input data (`input_ids`), attention masks (`attention_mask`) to indicate relevant tokens for processing, and tokenized labels for the target language (`labels`). The tokenization process can be formally described as follows in Eq. (1)):

$$X = \{\text{sentence}_1, \ldots, \text{sentence}_n\}. \quad (1)$$

This tokenization function is applied to both the training and testing datasets (`train_data` and `test_data`). The resulting encodings include the processed inputs, (`input_ids` and

attention_mask), and the corresponding labels for supervised training. This systematic preparation ensures the model can effectively learn the mapping between the source language (Indonesian) and the target language (Sundanese) while adhering to the computational constraints of the pretrained architecture. By leveraging a robust multilingual pretrained model and precise tokenization techniques, the research lays the foundation for efficient and accurate machine translation.

E. Training Setup

The training setup for fine-tuning the M2M100 translation model is meticulously designed to balance efficiency and performance while leveraging the capabilities of the Hugging Face transformers library. The training process utilized the `Seq2SeqTrainingArguments` class to define key hyperparameters and strategies. An output directory is specified to store model checkpoints and results, enabling recovery of training in case of interruptions and facilitating evaluation. The training employs a step-based evaluation strategy, performing evaluations every 1,000 steps to monitor the model's progress without incurring excessive computational overhead.

A learning rate of 1×10^{-4} is chosen to ensure gradual and stable optimization, while a weight decay of 0.01 is applied to regularize the model and mitigate overfitting. The batch sizes for both training and evaluation are set to 8, striking a balance between memory limitations and throughput. Given this configuration, gradient accumulation is kept minimal, as the batch size is sufficient for stable gradient updates. Training is conducted for 20 epochs, a duration optimized to avoid underfitting while minimizing the risk of overfitting. Logging intervals are set to every 500 steps, providing regular updates on the training progress without generating excessive output.

Mixed-precision training is enabled through the `fp16=True` setting, leveraging half-precision floating-point arithmetic to accelerate computation and reduce memory usage, particularly beneficial for Graphics Processing Units (GPUs). The `Seq2SeqTrainer` class streamlines the integration of the model, tokenizer, and datasets into the training loop. The training and evaluation datasets are processed into tokenized batches, and the `predict_with_generate` parameter is activated to allow sequence generation during evaluation, ensuring alignment with the machine translation task.

The training objective focuses on minimizing the cross-entropy loss, which quantifies the discrepancy between the predicted token probabilities and the actual target tokens. This loss is computed across all tokens in the sequence and averaged over the training samples. This carefully designed setup ensures efficient and effective fine-tuning, enabling the model to achieve high-quality translations from Indonesian to Sundanese while optimizing computational resources.

F. Model Training

The model training process for Indonesian to Sundanese translation involves converting the preprocessed data into a format compatible with the Hugging Face ecosystem and implementing an optimized training routine. The Hugging Face `Dataset` library is used to transform the training and testing data, initially in the form of Pandas DataFrames, into Hugging Face-compatible datasets. This conversion streamlines the integration of tokenization and preprocessing steps into the training pipeline.

Tokenization is performed using the `preprocess_batch` function, where the text in both Indonesian (`indonesian_cleaned`) and Sundanese (`sundanese_cleaned`) columns is tokenized using the pre-trained M2M100 tokenizer. The tokenization includes truncation to a maximum length of 128 tokens to prevent memory overflow, while padding ensures uniform input lengths within a batch. The resulting tokenized data include input IDs, attention masks, and target labels, encapsulating all components required for sequence-to-sequence learning.

To enhance compatibility with PyTorch, the datasets are formatted into tensors using the `set_format` method, specifying the inclusion of `input_ids`, `attention_mask`, and `labels`. This step enables seamless batching and data loading during the training process. Moreover, preprocessing is applied in a batched manner using the `map` function, which accelerates the process and ensured consistency across the datasets.

The `Seq2SeqTrainer` from Hugging Face's transformers library facilitates the training routine, integrating the model, tokenized datasets, and training arguments. Gradient checkpointing is enabled in the model configuration using `model.config.gradient_checkpointing=True`, a memory optimization technique that reduces GPU memory usage by recomputing activations during backpropagation. It allows the model to handle larger batches or sequences without exceeding memory limits.

The training objective minimized the cross-entropy loss, defined as follows in Eq. (2):

$$L = -\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T} y_{i,t}\log\left(p_{i,t}\right), \tag{2}$$

where N is the number of samples. Then, T is the sequence length, $y_{i,t}$ represents the target token probability, and $p_{i,t}$ denotes the predicted probability for the t-th token in the i-th sample. This loss function guided the model to align its predictions with the actual translation outputs.

The training process is initiated using `trainer.train()`. It leverages the predefined training arguments for controlled optimization. By combining efficient preprocessing, gradient checkpointing, and a carefully designed loss objective, this training setup enables the model to achieve high-quality translations while maintaining computational efficiency.

G. Evaluation

BLEU is selected because it is the most established automatic evaluation method for machine translation, focusing on precision and n-gram overlap, which is particularly useful for assessing word choice accuracy in low-resource settings [34]. ROUGE is included to complement BLEU, as it captures recall-oriented aspects and evaluates sequence-level fluency through the longest common subsequence. The combination of BLEU and ROUGE ensures a balanced assessment of both lexical accuracy and structural coherence, making them suitable for benchmarking against prior Indonesian–Sundanese and other low-resource translation studies.

The BLEU score is a precision-based metric that measures n-gram overlap between the model's output and the reference text. It is defined as Eq. (3):

$$\text{BLEU} = BP \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right), \tag{3}$$

where BP is the brevity penalty used to penalize overly short translations. Then, p_n denotes the precision for n-grams of size n and w_n represents the weight assigned to each n-gram precision (commonly set equally). The BLEU score ranges from 0 to 1, where higher values indicate greater similarity between predicted translations and reference sentences.

In addition to BLEU, the ROUGE metric is employed to further analyze the model's performance. Specifically, ROUGE-1 F1-Score calculates the overlap of unigrams (single words), while ROUGE-L F1-Score considers the Longest Common Subsequence (LCS) between generated translations and references. The ROUGE-L F1-Score is formulated as Eq. (4):

$$\text{ROUGE-L} = \frac{(1+\beta^2)\,\text{Prec}\cdot\text{Prec}_{LCS}\cdot\text{Recall}_{LCS}}{\beta^2\cdot\text{Prec}_{LCS}+\text{Recall}_{LCS}}. \tag{4}$$

It shows Prec_{LCS} and Recall_{LCS} represent the precision and recall computed based on the longest common subsequence, and β is a weighting parameter balancing recall and precision (commonly set to 1).

The BLEU score demonstrates a significant degree of n-gram overlap, reflecting accurate word selection and syntactic structure in the generated sentences [32]. Furthermore, the average ROUGE-1 F1-Score highlights effective word-level matches. Meanwhile, the average ROUGE-L F1-Score confirms the model's ability to generate translations that maintain proper order and coherence.

III. Results and Discussion

The performance of the Indonesian to Sundanese translation model is evaluated using rigorous training and evaluation metrics. The results of the training phase demonstrates the model's convergence and efficiency across the specified hyperparameters. Specifically, the training evaluation yielded a final loss of 0.4297 after 20 epochs, indicating that the model effectively minimized its objective function. The training loss is computed as follows Eq. (5)):

$$L = -\frac{1}{N}\sum_{1}^{N}\log P(y_1 \mid x, 0), \tag{5}$$

where N denotes the total number of tokens in the target sequence, which is the length of the sequence the model aims to predict. The term y_i represents the i-th token in the target sequence, where each token corresponds to a specific word or symbol that the model must predict. The expression $P(y_i \mid x, \theta)$ refers to the conditional probability of predicting the token y_i given the input x and the model parameters θ. The log-likelihood function sums the negative logarithm of these conditional probabilities for each token in the sequence. This function is minimized during training to adjust the model parameters θ so that the model's predictions align more closely with the target sequence, thereby improving its accuracy in generating or translating text. The training runtime was recorded at 11.01 seconds, achieving a throughput of 65.77 samples per second and 8.27 steps per second, reflecting the computational efficiency of the model setup.

The justification for using BLEU and ROUGE lies in their widespread adoption in machine translation

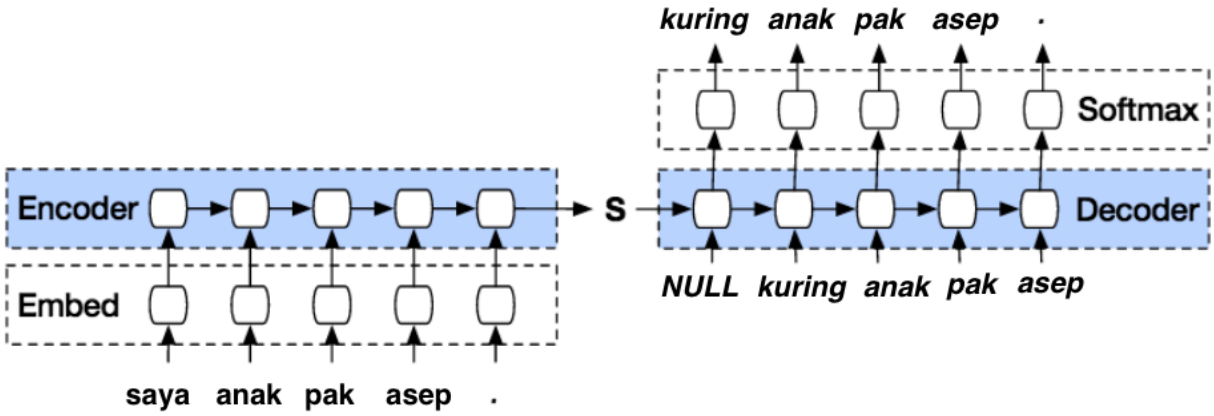

Fig. 5. Predict sequence. *Saya anak pak Asep* and *kuring anak pak Asep* mean I am Mr. Asep's son.

research, allowing direct comparison with previous Indonesian–Sundanese works such as [33]. Moreover, while BLEU captures precise n-gram matches, ROUGE provides insights into fluency and contextual alignment. Together, they offer a more holistic evaluation. Alternative evaluation methods, such as METEOR or human judgment, are not included in the research due to dataset limitations and the primary focus on automated reproducibility, which is consistent with low-resource language evaluation practices.

For the evaluation phase, the model's translation quality is assessed using BLEU and ROUGE metrics. The BLEU score reaches 44.12. It demonstrates a high degree of n-gram overlap between predicted translations and ground-truth references, highlighting the model's ability to reproduce the structure and semantic content of the target language.

Furthermore, the ROUGE-1 F1-Score achieves an average of 0.7165, indicating effective word-level precision and recall. Meanwhile, the ROUGE-L F1-Score yields an average of 0.7092, confirming that the model maintains strong coherence and fluency while preserving structural integrity. The evaluation runtime for ROUGE computation is 9.56 seconds, with a processing rate of 75.73 samples per second and 9.52 steps per second. The results demonstrate both efficiency and consistency.

The results show the effectiveness of the multilingual M2M100 model in capturing the linguistic nuances between Indonesian and Sundanese, as illustrated in Fig. 5. The diagram represents a standard encoder-decoder architecture used in NMT, where the encoder processes the input sentence in Indonesian ("*saya anak pak asep*") by transforming each token into contextualized embeddings. The final encoder state s encapsulates semantic and syntactic information of the entire source sentence and is passed to the decoder to generate the Sundanese translation ("*kuring anak pak asep*").

Each token in the decoder phase is generated sequentially, beginning with a NULL token and conditioned on both the encoder state and previously generated tokens. The softmax layer outputs a probability distribution over the vocabulary at each decoding step, selecting the most probable token to form the output sequence. This mechanism preserves token-level alignment and sentence-level coherence between the source and target languages.

The high BLEU score reflects the model's ability to generate lexically and syntactically consistent translations. Strong ROUGE-1 and ROUGE-L performances further confirm the model's capability to capture unigram overlaps and long-sequence dependencies, indicating preservation of semantic meaning and sentence structure. These results validate the effectiveness of the M2M100 model in handling low-resource language pairs without relying on pivot languages.

The low evaluation loss of 0.4297 confirms that the model achieves an optimal balance between overfitting and generalization. The robust results across both BLEU and ROUGE metrics suggest that the model is highly reliable for practical use in Indonesian to Sundanese translation tasks. However, it is important to note that BLEU and ROUGE have limitations, as they may not fully capture semantic adequacy, idiomatic usage, or speech-level variations in Sundanese. While these metrics provide useful benchmarks, they cannot substitute for human judgment.

These findings provide strong evidence of the multilingual model's ability to generalize across low-

Cite this article as: A. M. Rifai, E. Utami, Amali, M. Fatchan, and M. Ekhsan, "Advancing cross-cultural natural language processing with a focus on Sundanese language and contextual nuances", CommIT Journal 20(1), 155–167, 2026.

TABLE III
RESULTS OF TRANSLATION SAMPLING.

Sentence	Original (Indonesian)	Translated (English)	Translated (Sundanese)	Change
Adverb of time	*Pagi ini*	This morning	*Isuk kéné̊h*	Adverb of time translated into a local idiomatic expression.
Subject	*Saya*	I	*Kuring*	The subject changes according to contextual usage in Sundanese.
Predicate	*Mau makan*	Want to eat	*Hayang dahar*	Expression of desire adapts to local linguistic structure.
Object	*Daging ayam*	Chicken	*Daging ayam*	No lexical change occurred.

TABLE IV
RESULTS OF COMPARISON.

Author	Language	Method and Materials	Result
[37]	Arabic–Italian	Two deep learning models (LSTM and GRU) with attention mechanisms using a Seq2Seq encoder–decoder architecture.	BLEU: LSTM 0.18, GRU 0.17; ROUGE: LSTM 0.17, GRU 0.16
[38]	English–Portuguese	Transformer models with ASR auxiliary loss for spoken language translation.	BLEU improved from 41.21 to 44.69, reaching 46.9 with ensembling
[39]	English–Japanese	Sequence-to-sequence Transformer model for machine translation.	BLEU score: 40.1 (highest with merged datasets)
[40]	Lampung–Indonesian	DMT and SMT	SMT achieves BLEU of 59.85%, outperforming DMT (39.32%)
[34]	Indonesian–Sundanese	Transformer-based Neural Machine Translation.	BLEU 42.72%, average training loss 1.77
Proposed model	Indonesian–Sundanese	Transformer-based Seq2Seq architecture fine-tuned for Sundanese translation.	BLEU 44.12, ROUGE-1 F1 0.72, near-human-level translation quality

Note: Bilingual Evaluation Understudy (BLEU), Recall - Oriented Understudy for Gisting Evaluation (ROUGE), Long Short-Term Memory (LSTM), Gated Recurrent Unit (GRU), Automatic Speech Recognition (ASR), Direct Machine Translation (DMT) and Statistical Machine Translation (SMT).

resource languages, further highlighting the importance of leveraging pretrained models for specific translation applications. Results of translation sampling are presented in Table III, which provides a comparative analysis of sentence components translated from Indonesian into English and Sundanese. This sampling highlights the model's capacity to adapt across lexical, syntactic, and semantic nuances. The adverbial phrase "pagi ini" is rendered as "isuk kéné̊h" in Sundanese, reflecting a culturally and linguistically appropriate idiom rather than a direct word-for-word translation. Similarly, the subject "saya" changes to "kuring", a context-sensitive representation that preserves the speaker's tone and discourse position.

The predicate "mau makan" is translated as "hayang dahar", indicating a meaningful transformation of intent expression. Here, "hayang" and "dahar" reflect preferred local usage, demonstrating the model's effectiveness in preserving contextual accuracy. In contrast, the object "daging ayam" remains unchanged across both languages, suggesting that common nouns without cultural variation are translated through direct lexical mapping.

This analysis confirms that the model preserves semantic integrity while capturing linguistic subtleties inherent in low-resource language pairs. Such transformations are essential for producing fluent and culturally relevant translations rather than rigid literal equivalents. Future work may focus on deeper semantic alignment and expanding idiomatic datasets to further enhance translation performance.

Table IV provides a comparative analysis of machine translation results across various studies. Previous research has translated the Arabic Quran to Italian using LSTM and Gated Recurrent Unit (GRU) models with attention mechanisms, achieving BLEU scores of 0.18 (LSTM) and 0.17 (GRU) with corresponding ROUGE scores of 0.17 and 0.16 [37]. Another research has employed transformer models with Automatic Speech Recognition (ASR) auxiliary loss for English-to-Portuguese translation, improving the BLEU score from 41.21 to 44.69, further increasing to 46.9 with ensembling [38]. Previous study has also applied a Seq2Seq transformer model for English-to-Japanese translation, reaching a BLEU score of 40.1, the highest achieved by merging datasets [39]. Another research has compared Direct Machine Translation (DMT) and Statistical Machine Translation (SMT) for Lampung to Indonesian translation, with SMT outperforming DMT at 59.85% [40]. Next, previous research has experimented to translate Indonesian to Sundanese with transformer model and NMT, achieving a BLEU score of 42.72% and an average training loss of 1.77 [34]. The final entry, "proposed model," details a transformer-based Seq2Seq model fine-tuned for Indonesian to Sundanese translation, achieving a BLEU score of 44.12 and ROUGE-1 F1 score of 0.72, demonstrating near-human-level quality and highlight-

ing its potential for educational and cross-cultural applications.

IV. Conclusion

In the research, the M2M100 multilingual translation model is fine-tuned for Indonesian–Sundanese translation using the Hugging Face transformers library. The model is trained using the cross-entropy loss function to minimize the discrepancy between predicted and target tokens across multiple epochs, achieving a final loss of 0.4297 after 20 epochs. It indicates effective learning and generalization from the training data. Key techniques, such as tokenization, gradient checkpointing, and mixed-precision training, are integrated to ensure high-quality translations while maintaining computational efficiency. The Seq2SeqTrainer framework facilitates the management of training parameters, streamlining the entire process from data preprocessing to evaluation. The model achieves a throughput of 65.77 samples per second, with an evaluation runtime of 11.01 seconds, demonstrating its efficiency during training.

The model's performance is assessed using standard machine translation evaluation metrics, BLEU and ROUGE. A BLEU score of 44.12 is achieved, indicating a significant n-gram overlapping with the reference translations, thus reflecting accurate lexical translation. The ROUGE-1 and ROUGE-L F1-Scores of 0.7165 and 0.7092, respectively, indicate that the model maintains strong word-level precision and recall while preserving sentence structure. These results validate the effectiveness of the pretrained M2M100 model for Indonesian–Sundanese translation, showing its ability to maintain both meaning and syntax in the translated text.

Unlike prior Indonesian–Sundanese translation efforts that primarily rely on LSTMs or standard transformer-based models, the research introduces novelty by fine-tuning the multilingual pretrained M2M100 transformer. Leveraging transfer learning from high-resource languages, the model is able to capture contextual and linguistic nuances more effectively. It improves translation quality compared to earlier approaches. This contribution not only sets a new benchmark for Indonesian–Sundanese low-resource machine translation but also underscores the cultural importance of supporting regional languages in modern NLP systems.

However, a major challenge in the research is the limited availability of a high-quality, diverse dataset for Indonesian-to-Sundanese translation, which affects the model's robustness and generalization. Despite this limitation, the model demonstrates promising results, suggesting its potential for application in low-resource language pairs. For future research, expanding the dataset with more diverse and contextually rich Indonesian–Sundanese parallel corpora will be essential to improve model generalization. Investigating alternative architectures such as encoder-decoder transformers with language-specific adapters may also yield better results in low-resource settings. Additionally, integrating linguistic features or syntactic information can help to preserve grammatical nuances and cultural expressions. Finally, exploring transfer learning techniques from linguistically similar language pairs can further enhance translation quality.

Acknowledgement

The research was funded and supported by a research grant provided by the Department of Research and Community Service, Universitas Pelita Bangsa (Grant Number: 038/KP/7/UPB/2024).

Author Contribution

Developed the conceptual framework, A. M. R.; Designed the research methodology, A. M. R.; Implemented the analytical model, A. M. R.; Conducted data processing and validation, A. M. R.; Interpreted the results, A. M. R.; Prepared visualizations, A. M. R.; Supervised the research process, A. M. R.; Acquired funding support, A. M. R.; Composed the original manuscript draft, A. M. R.; Contributed to methodology refinement, E. U.; Performed data curation and validation, E. U.; Participated in investigation activities, E. U.; Reviewed and edited the manuscript, E. U.; Managed project administration, E. U.; Supported data curation and investigation processes, A.; Contributed to manuscript review and editing, A.; Assisted in project administration, A.; Assisted in investigation activities, M. F.; Contributed to data curation, M. F.; Participated in manuscript review and editing, M. F.; Supported project administration, M. F.; Contributed to methodological development, M. E.; Performed formal analysis, M. E.; and Supported project administration throughout the research process, M. E.

Data Availability

The dataset supporting the findings of this study is publicly available and can be accessed through the following repository: https://doi.org/10.34820/FK2/HDYWXW. The data are provided to ensure transparency and reproducibility of the research results.

Cite this article as: A. M. Rifai, E. Utami, Amali, M. Fatchan, and M. Ekhsan, "Advancing cross-cultural natural language processing with a focus on Sundanese language and contextual nuances", CommIT Journal 20(1), 155–167, 2026.

REFERENCES

[1] 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, 2023.

[2] R. Ruiz-Dolz, J. Taverner, J. Lawrence, and C. Reed, "NLAS-multi: A multilingual corpus of automatically generated natural language argumentation schemes," *Data in Brief*, vol. 57, pp. 1–10, 2024.

[3] N. Ahmed, A. K. Saha, M. A. Al Noman, J. R. Jim, M. F. Mridha, and M. M. Kabir, "Deep learning-based natural language processing in human–agent interaction: Applications, advancements and challenges," *Natural Language Processing Journal*, vol. 9, pp. 1–25, 2024.

[4] Supriyono, A. P. Wibawa, Suyono, and F. Kurniawan, "Advancements in natural language processing: Implications, challenges, and future directions," *Telematics and Informatics Reports*, vol. 16, pp. 1–17, 2024.

[5] H. Wang, J. Li, H. Wu, E. Hovy, and Y. Sun, "Pre-trained language models and their applications," *Engineering*, vol. 25, pp. 51–65, 2023.

[6] J. A. Ruipérez-Valiente, T. Staubitz, M. Jenner, S. Halawa, J. Zhang, I. Despujol, J. Maldonado-Mahauad, G. Montoro, M. Peffer, T. Rohloff, J. Lane, C. Turro, X. Li, M. Pérez-Sanagustín, and J. Reich, "Large scale analytics of global and regional MOOC providers: Differences in learners' demographics, preferences, and perceptions," *Computers & Education*, vol. 180, pp. 1–17, 2022.

[7] Y. Liu and M. Dras, "Using corpora from natural language processing for investigating crosslinguistic influence," *Ampersand*, vol. 12, pp. 1–13, 2024.

[8] W. Khan, A. Daud, K. Khan, S. Muhammad, and R. Haq, "Exploring the frontiers of deep learning and natural language processing: A comprehensive overview of key challenges and emerging trends," *Natural Language Processing Journal*, vol. 4, pp. 1–31, 2023.

[9] N. Nnamoko, T. Karaminis, J. Procter, J. Barrowclough, and I. Korkontzelos, "Automatic language ability assessment method based on natural language processing," *Natural Language Processing Journal*, vol. 8, pp. 1–16, 2024.

[10] S. Sarip, D. Fitriana, A. F. Azhari, A. Absori, E. K. Dewi, H. N. Adiantika, and N. Nurkhaeriyah, "Policy and linguistic considerations in the proposed renaming of West Java Province to Tatar Sunda," *Cepalo*, vol. 8, no. 1, pp. 31–48, 2024.

[11] A. F. Aji, G. I. Winata, F. Koto, S. Cahyawijaya, A. Romadhony, R. Mahendra, K. Kurniawan, D. Moeljadi, R. E. Prasojo, T. Baldwin, J. H. Lau, and S. Ruder, "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: Association for Computational Linguistics, 2022, pp. 7226–7249.

[12] I. Widianingsih, J. J. McIntyre, U. S. Rakasiwi, G. H. Iskandar, and R. Wirawan, "Indigenous sundanese leadership: Eco-systemic lessons on zero emissions: A conversation with Indigenous leaders in Ciptagelar, West Java," *Systemic Practice and Action Research*, vol. 36, no. 2, pp. 321–353, 2023.

[13] M. Javaid, A. Haleem, R. P. Singh, and R. Suman, "Artificial intelligence applications for Industry 4.0: A literature-based study," *Journal of Industrial Integration and Management*, vol. 7, no. 01, pp. 83–111, 2022.

[14] M. Mager, E. Maier, K. von der Wense, and N. T. Vu, "Ethical considerations for machine translation of indigenous languages: Giving a voice to the speakers," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Toronto, Canada: Association for Computational Linguistics, 2023, pp. 4871–4897.

[15] H. Sujaini and A. B. Putra, "Analysis of language identification algorithms for regional Indonesian languages," *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. 13, no. 2, pp. 1741–1752, 2024.

[16] A. Montejo-Ráez, M. D. Molina-González, S. M. Jiménez-Zafra, M. Á. García-Cumbreras, and L. J. García-López, "A survey on detecting mental disorders with natural language processing: Literature review, trends and challenges," *Computer Science Review*, vol. 53, pp. 1–17, 2024.

[17] B. Masua and N. Masasi, "In the heart of Swahili: An exploration of data collection methods and corpus curation for natural language processing," *Data in Brief*, vol. 55, pp. 1–9, 2024.

[18] B. Zhang, P. Williams, I. Titov, and R. Sennrich, "Improving massively multilingual neural machine translation and zero-shot translation," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics,

2020, pp. 1628–1639.

[19] S. Cahyawijaya, H. Lovenia, F. Koto, D. Adhista, E. Dave, S. Oktavianti, S. Akbar, J. Lee, N. Shadieq, T. W. Cenggoro, H. Lunuwih, B. Wilie, G. Muridan, G. Winata, D. Moeljadi, A. F. Aji, A. Purwarianti, and P. Fung, "NusaWrites: Constructing high-quality corpora for underrepresented and extremely low-resource languages," in *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*. Nusa Dua, Bali: Association for Computational Linguistics, 2023, pp. 921–945.

[20] B. D. Wijanarko, Y. Heryadi, D. F. Murad, C. Tho, and K. Hashimoto, "Recurrent neural network-based models as Bahasa Indonesia-Sundanese language neural machine translator," in *2023 International Conference on Computer Science, Information Technology and Engineering (ICCoSITE)*. Jakarta, Indonesia: IEEE, Feb. 16, 2023, pp. 951–956.

[21] E. A. Kusnanti, E. Sierra, G. G. S. Putra, E. S. Cahyadi, A. Haq, and D. Purwitasari, "Indonesian lexical ambiguity in machine translation: A literature review," in *2024 International Conference on Information Technology Research and Innovation (ICITRI)*. Jakarta, Indonesia: IEEE, Sep. 5–6, 2024, pp. 59–64.

[22] A. Tambusai and K. Nasution, "A comparative typology of verbal affixes in Riau-Malay and Sundanese," *Indonesian Journal of Applied Linguistics*, vol. 13, no. 3, pp. 636–647, 2024.

[23] G. L. A. Babu and S. Badugu, "Deep learning based sequence to sequence model for abstractive Telugu text summarization," *Multimedia Tools and Applications*, vol. 82, no. 11, pp. 17 075–17 096, 2023.

[24] A. Rahali and M. A. Akhloufi, "End-to-end transformer-based models in textual-based NLP," *Ai*, vol. 4, no. 1, pp. 54–110, 2023.

[25] N. A. Al-Shameri and H. S. Al-Khalifa, "Arabic paraphrase generation using transformer-based approaches," *IEEE Access*, vol. 12, pp. 121 896–121 914, 2024.

[26] P. Prasada and M. V. P. Rao, "Reinforcement of low-resource language translation with neural machine translation and backtranslation synergies," *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. 13, no. 3, pp. 3478–3488, 2024.

[27] P. U. Ogbogu, L. M. Noroski, K. Arcoleo, B. D. Reese Jr., and A. J. Apter, "Methods for cross-cultural communication in clinic encounters," *The Journal of Allergy and Clinical Immunology: In Practice*, vol. 10, no. 4, pp. 893–900, 2022.

[28] D. Peral-García, J. Cruz-Benito, and F. J. García-Peñalvo, "Comparing natural language processing and quantum natural processing approaches in text classification tasks," *Expert Systems with Applications*, vol. 254, pp. 1–9, 2024.

[29] D. W. Otter, J. R. Medina, and J. K. Kalita, "A survey of the usages of deep learning for natural language processing," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 2, pp. 604–624, 2020.

[30] M. Siino, I. Tinnirello, and M. La Cascia, "Is text preprocessing still worth the time? A comparative survey on the influence of popular preprocessing methods on transformers and traditional classifiers," *Information Systems*, vol. 121, pp. 1–19, 2024.

[31] L. J. Laki and Z. G. Yang, "Neural machine translation for Hungarian," *Acta Linguistica Academica*, vol. 69, no. 4, pp. 501–520, 2022.

[32] D. Roy, S. Fakhoury, and V. Arnaoudova, "Reassessing automatic evaluation metrics for code summarization tasks," in *Proceedings of the 29th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*. New York, United States: Association for Computing Machinery, Aug. 23–28, 2021, pp. 1105–1116.

[33] S. Cahyawijaya, G. I. Winata, B. Wilie, K. Vincentio, X. Li, A. Kuncoro, S. Ruder, Z. Y. Lim, S. Bahar, M. Khodra, A. Purwarianti, and P. Fung, "IndoNLG: Benchmark and resources for evaluating indonesian natural language generation," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, 2021, pp. 8875–8898.

[34] A. A. Suryani, D. H. Widyantoro, A. Purwarianti, and Y. Sudaryat, "Sundanese-Indonesian parallel corpus," 2022. [Online]. Available: https://doi.org/10.34820/FK2/HDYWXW

[35] A. Kathuria, A. Gupta, and R. K. Singla, "A review of tools and techniques for preprocessing of textual data," in *Computational Methods and Data Engineering*. Springer, 2020, pp. 407–422.

[36] R. Egger and E. Gokce, "Natural Language Processing (NLP): An introduction: Making sense of textual data," in *Applied data science in tourism: Interdisciplinary approaches, methodologies, and*

applications. Springer, 2022, pp. 307–334.

[37] H. Hamed, A. M. Helmy, and A. Mohammed, “Deep learning approach for translating Arabic Holy Quran into Italian language,” in *2021 International Mobile, Intelligent, and Ubiquitous Computing Conference (MIUCC)*. Cairo, Egypt: IEEE, May 26–27, 2021, pp. 193–199.

[38] H. K. Vydana, M. Karafiát, K. Zmolikova, L. Burget, and H. Černockỳ, “Jointly trained transformers models for spoken language translation,” in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Toronto, Canada: IEEE, June 6–11, 2021, pp. 7513–7517.

[39] M. Gupta and P. Kumar, “Robust neural language translation model formulation using Seq2Seq approach,” *Fusion: Practice and Applications*, vol. 5, no. 2, pp. 61–67, 2021.

[40] Z. Abidin, P. Permata, and F. Ariyani, “Translation of the Lampung language text dialect of Nyo into the Indonesian language with DMT and SMT approach,” *INTENSIF: Jurnal Ilmiah Penelitian dan Penerapan Teknologi Sistem Informasi*, vol. 5, no. 1, pp. 58–71, 2021.