IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Mathematical Modelling of Engineering Problems
Vol. 12, No. 11, November, 2025, pp. 4077-4090
Journal homepage: http://iieta.org/journals/mmep

Optimizing Machine Translation with Sequential Tokenizer Embeddings Using Convolutional Neural Network for Low-Resource Language

Andi Sofyan Anas[1*], Abdul Muhid[2], Pahrul Irfan[3], Kurniadin Abd Latif[1], Muhamad Fikri Arjuna[1]

[1] Faculty of Engineering, Bumigora University, Mataram 83127, Indonesia
[2] Postgraduate Program in English Literature, Faculty of Cultural Science, Bumigora University, Mataram 83127, Indonesia
[3] Faculty of Engineering, Mataram University, Mataram 83115, Indonesia

Corresponding Author Email: andi.sofyan@universitasbumigora.ac.id

https://doi.org/10.18280/mmep.121134

Received: 28 September 2025
Revised: 13 November 2025
Accepted: 21 November 2025
Available online: 30 November 2025

Keywords:
Neural Machine Translation, low-resource language, Seq2Seq, CNN, TF-IDF, IndoBERT

ABSTRACT

Low-resource languages, including regional languages with limited parallel corpora, remain a major challenge in Neural Machine Translation (NMT). The main issue lies in finding an appropriate feature representation that can capture linguistic and contextual meaning under data-scarce conditions. Moreover, model instability and poor adaptation to specific feature representations often lead to suboptimal translation performance. To address this problem, this study develops an optimized NMT framework for the Sasak and Indonesian language pair as a representative case of regional low-resource translation. The proposed model integrates sequential tokenizer embeddings (Seq2Seq) within a Convolutional Neural Network (CNN) based encoder and decoder architecture. In addition, two baseline approaches, namely Term Frequency–Inverse Document Frequency (TF-IDF) and the pre-trained IndoBERT model, are investigated to examine alternative feature representation methods. All models are tested bidirectionally, from Sasak to Indonesian and vice versa, using accuracy and BLEU metrics (BLEU-1 to BLEU-4). The experimental results show that the CNN-based Seq2Seq embedding model achieves the best performance, obtaining 84.47% accuracy with BLEU-1 of 56.63%, BLEU-2 of 42.68%, BLEU-3 of 31.91%, and BLEU-4 of 22.31% for Sasak to Indonesian translation. For Indonesian to Sasak translation, the same model reaches 81.09% accuracy with BLEU-1 of 44.25%, BLEU-2 of 29.99%, BLEU-3 of 19.89%, and BLEU-4 of 12.24%. These findings confirm that sequential embedding offers a more effective feature representation, enhancing translation performance in low-resource regional languages.

1. INTRODUCTION

Regional languages represent an essential part of cultural identity and serve as vital media for transmitting local knowledge. However, many of these languages are categorized as low-resource languages due to the limited availability of parallel corpora that can support the development of Neural Machine Translation (NMT) systems [1, 2]. This scarcity of data limits technological support for regional languages in the digital space, increasing the risk of their marginalization in the era of globalization. Therefore, developing effective NMT solutions for low-resource languages has become an urgent necessity, not only for daily communication but also for preserving linguistic heritage.

In NMT research, sequence-to-sequence (Seq2Seq) models based on Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM), and Transformer architectures have long dominated the field [3, 4]. These models show strong capabilities in capturing semantic dependencies but tend to be unstable and prone to overfitting in low-resource scenarios [5]. Convolutional Neural Networks (CNNs), although less explored, possess advantages in extracting local linguistic patterns efficiently with fewer parameters [6], making them a promising alternative for small-scale datasets. CNNs also provide lightweight architectures, which are beneficial for regional language translation tasks where computational and data resources are limited.

Previous studies have shown the effectiveness of CNNs in various NLP tasks such as text classification, sentiment analysis, and n-gram-based feature extraction [7, 8]. However, their potential in NMT remains underexplored because CNNs are considered less capable of modeling long-range dependencies. This limitation opens an opportunity to combine CNNs with richer representation techniques such as Term Frequency–Inverse Document Frequency (TF-IDF), IndoBERT, or Seq2Seq tokenizer embeddings. Such integration could enhance CNNs' ability to capture both local patterns and broader semantic or sequential relationships within language data.

Research on the Sasak language illustrates these broader challenges in low-resource NMT. Aranta et al. [9] investigated multi-dialect Sasak translation (Ngento-Ngente, Meno-Mene, Ngeno-Ngene, and Meriak-Meriku) using a low-resource approach, but their work was limited to one-way translation

(Indonesian–Sasak) and lacked a structured corpus. Shabrina et al. [10] developed a static dictionary that could not handle contextual variation, while Wardhana et al. [11] implemented an NMT system for a single dialect without integrating cultural aspects. More recent work using mBART achieved a BLEU score of 79.78 [12], yet cultural and linguistic validation remained absent. Meanwhile, studies on other regional languages [13-19] mainly relied on attention-based or Transformer models for Javanese and Sundanese, which benefited from large datasets that do not reflect the real constraints of low-resource languages like Sasak. These findings indicate a persistent research gap in developing compact yet robust NMT architectures that perform effectively under data scarcity.

To address this gap, the present study introduces an optimized Sasak–Indonesian NMT framework that integrates both technical and cultural considerations. It constructs the first structured Sasak–Indonesian parallel corpus of over 10,000 sentence pairs covering five main dialects, enriched with cultural annotations such as bebasan, wariga, and ritual expressions. The proposed model compares three feature extraction techniques—TF-IDF, IndoBERT, and sequential tokenizer embeddings—to evaluate their effectiveness in representing bilingual text under limited data. Methodologically, a CNN-based Seq2Seq architecture integrated with Attention is developed to capture both local and long-range dependencies, while a linguistic postprocessing module incorporating Sasak affixation and reduplication rules ensures culturally accurate translation outputs.

Overall, this study contributes to advancing low-resource machine translation by demonstrating that CNN-based architectures, when combined with sequential embeddings and linguistically validated data, can serve as efficient and culturally grounded alternatives to Transformer-based models for regional languages such as Sasak.

2. METHODOLOGY

This study adopts a NMT approach to develop an automatic translation system between the Sasak and Indonesian languages. The overall design follows an encoder–decoder framework enhanced with an attention mechanism to more effectively capture contextual relationships within each word sequence. The encoder is responsible for representing input sentences from the source language into continuous vector representations that encode semantic information, while the decoder generates sentences in the target language by leveraging the encoded representations along with the attention mechanism that aligns token-to-token correspondence. In general, the system pipeline consists of five main components: data collection, data preprocessing, feature extraction, translation using the attention-based encoder–decoder architecture, and translation evaluation. The proposed NMT system flow is illustrated in Figure 1. The initial stage begins with the collection of parallel data comprising Sasak sentences and their corresponding Indonesian translations. The collected data are then processed through preprocessing steps to ensure text quality and readiness for subsequent processing. The preprocessed data are further transformed into numerical representations that can be utilized as inputs for the NMT model. The translation process is carried out using a CNN architecture that follows the encoder–decoder framework, augmented with an attention mechanism. The model outputs are evaluated both automatically using BLEU metrics and through linguistic validation, allowing translation quality to be assessed comprehensively. The results from both automated evaluation and linguistic validation are then employed for model tuning to achieve optimal performance.

2.1 Data collection

The data collection stage was carried out systematically to ensure the quality and authenticity of the Sasak–Indonesian parallel corpus. The data were obtained directly from native Sasak speakers who are fluent and actively use the language in their daily lives. This process was supervised by a linguist with expertise in both regional linguistics and Indonesian, ensuring that each collected sentence was verified in terms of transcription, semantic accuracy, and conformity to formal linguistic structures. The involvement of the linguist in this study served as a crucial factor to guarantee that the resulting corpus not only reflects natural language usage but also adheres to linguistic principles necessary for automatic translation modeling.

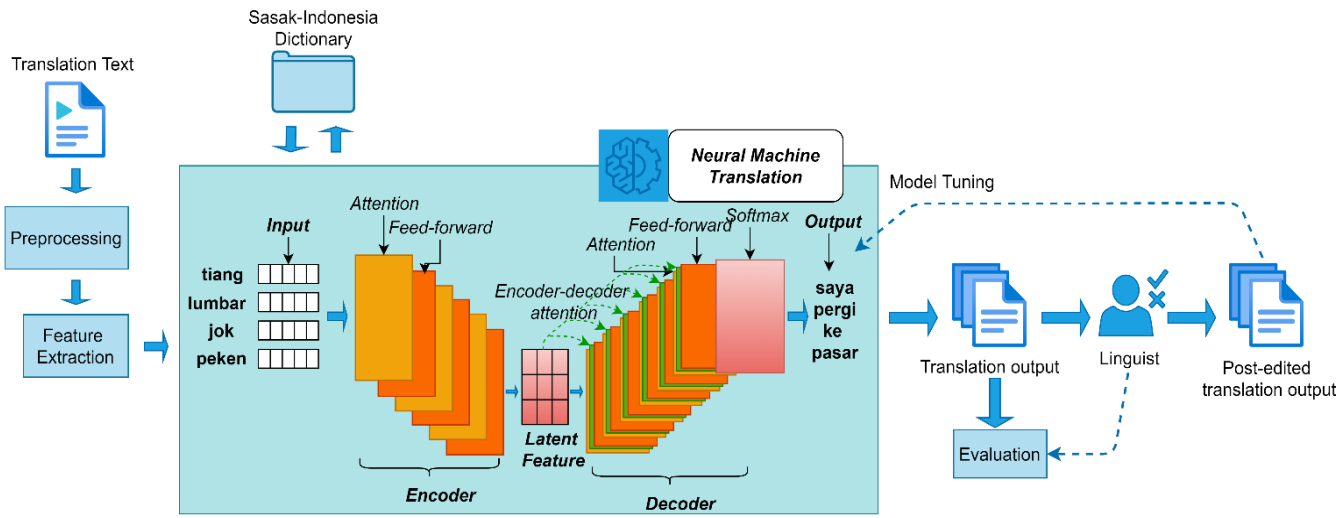


Figure 1. Workflow of the NMT system using CNN

Figure 2. Stages of text preprocessing in the Sasak–Indonesian NMT system

The type of data collected consisted of simple conversational sentences frequently used in everyday social interactions, such as communication at home, in markets, schools, and within community activities. The selection of this type of data was based on the need to develop a translation system that is both practical and relevant, enabling the model to accurately translate real-life conversational contexts. At the same time, this approach reflects the richness of Sasak language expressions in daily life settings. From a research ethics perspective, participation of native speakers was voluntary, with clear explanations provided regarding the objectives of the study and the potential use of the data. Informed consent was obtained prior to interviews and transcription, and all personal identities were kept confidential through anonymity protocols. Accordingly, this data collection procedure not only yielded a valid dataset but also complied with modern research ethics principles, respecting the rights and privacy of participants.

2.2 Data preprocessing

The text preprocessing procedure was conducted systematically to ensure that the input data were in a consistent format and to support the performance of the NMT model. The preprocessing stages are illustrated in Figure 2. The first step is lowercasing, which converts all text into lowercase letters to simplify vocabulary representation so that words differing only in capitalization, such as “Mele” and “mele,” are not treated as distinct entities. This step effectively reduces model complexity while improving generalization capability [20]. Next, punctuation removal was applied by eliminating all punctuation marks that do not carry significant semantic contributions in the Sasak–Indonesian translation process. Retaining punctuation would unnecessarily increase vocabulary size and introduce noise that could hinder model learning. Following this, whitespace normalization was performed to remove redundant spaces as well as unnecessary leading and trailing spaces in sentences. This normalization is essential to prevent empty tokens that may interfere with the tokenization process and generate out-of-vocabulary (OOV) instances [21].

The next stage is sequence delimitation, where special tokens <sos> (start of sequence) are added at the beginning of each sentence and <eos> (end of sequence) at the end. These tokens provide explicit markers for the model to indicate sentence boundaries, enabling the system to determine when prediction should start and stop [22, 23]. Following this, tokenization is applied, in which each sentence is mapped into tokens or words through a tokenizer so that it can be further processed by the model. This process is then continued with text-to-sequence conversion, which transforms the tokenized output into sequences of numerical indices to be projected into the embedding vector space, allowing the semantic relationships among words to be learned more representatively. The final step is sequence padding, where empty tokens are appended at the end of sentences to ensure that all sequences have uniform length. This step is crucial since neural networks require fixed-dimensional input to facilitate batch training. Through this series of preprocessing stages, the data become cleaner, more consistent, and ready to support optimal performance in NMT-based translation modeling.

2.3 Feature extraction

The feature extraction stage is a crucial component of this study, as it determines how the Sasak–Indonesian parallel texts are represented in numerical form prior to further processing. Three main approaches were employed to construct these representations: TF-IDF, contextual embeddings based on IndoBERT, and sequential embeddings built through tokenizers. The first approach, TF-IDF, calculates the relative weight of each word based on its frequency within a sentence and its distribution across the entire corpus. In this way, words that frequently occur but carry little meaning, such as stopwords, are automatically assigned lower weights, while semantically more informative words are emphasized in the representation [7, 8, 24]. The output of this stage is a high-dimensional vector that reflects the importance of each word in a sentence relative to the entire dataset. Word weighting in a document using TF-IDF is computed using Eqs. (1)-(3) [25].

$$\mathrm{TF-IDF(i,j)} = \mathrm{TF(i,j)} \times \mathrm{IDF(i)} \tag{1}$$

where,

$$\mathrm{TF(i,j)} = \frac{\mathrm{n(i,j)}}{\sum_{k=1}^{q} \mathrm{n(k,j)}} \tag{2}$$

$$\mathrm{IDF(i)} = \log \frac{\mathrm{N}}{1 + \mathrm{df(i)}} \tag{3}$$

Here, $\mathrm{n(i,j)}$ denotes the number of occurrences of term t_i in document D_j, $\mathrm{n(k,j)}$ represents the total number of terms in the document, N is the total number of documents, and $\mathrm{df(i)}$ is the number of documents that contain the term t_i.

The second approach in this study employs IndoBERT, a pretrained language model based on the BERT architecture that was specifically developed for the Indonesian language [26]. IndoBERT has been trained on a large-scale Indonesian corpus, enabling it to generate contextual representations. Each text is processed through the IndoBERT tokenizer, producing a sequence of hidden states for every token. The final representation is then obtained by applying mean pooling over all hidden states, resulting in a fixed-dimensional embedding that captures both syntactic and semantic relationships among words in a sentence. Recent studies have confirmed the effectiveness of IndoBERT across various NLP tasks. For instance, Syahputra et al. [27] demonstrated that IndoBERT significantly improves clickbait detection accuracy compared to the Roberta method. Similarly, studies by Marutho and Utomo [28], as well as Kamdan et al. [29], reported that IndoBERT outperformed mBERT and

IndoRoBERTa in text classification tasks. These findings indicate that the contextual representations produced by IndoBERT are more accurate in capturing the semantic nuances of the Indonesian language.

The third approach is tokenizer sequence embedding, where each word is mapped to a numerical index based on a vocabulary constructed from the training data. These indices are then transformed into fixed-dimensional vectors through an embedding layer. Unlike IndoBERT, which relies on prior pretraining, tokenizer-based embeddings are built and fine-tuned directly on the research dataset, making them more adaptive to the unique vocabulary of the Sasak language that frequently appears in everyday conversations. By combining these three approaches, this study aims to examine the extent to which frequency-based representations, pretrained contextual embeddings, and locally trained tokenizer embeddings influence the quality of Sasak–Indonesian bilingual representations.

2.4 Modeling

The modeling in this study was designed to leverage three types of data representations: TF-IDF, IndoBERT contextual embeddings, and tokenizer-based embeddings. Since each representation possesses distinct characteristics, the model architecture was tailored accordingly to optimally extract information from each input form. For TF-IDF and IndoBERT representations, the sequential order of words in a sentence is no longer explicitly preserved. TF-IDF converts text into a static high-dimensional vector that solely represents word weights, while IndoBERT, in this implementation, produces a single embedding obtained through mean pooling of hidden states. Consequently, both representations are more similar to text classification approaches than to pure sequential modeling. To overcome this limitation, the feature-extracted vectors are fed into dense layers to generate compact representations, which are then repeated across the length of the target sequence via a repeat vector operation. These repeated representations are subsequently processed by convolutional blocks in the decoder, enabling the model to learn sequence patterns even though the initial input lacks explicit sequential information. An attention mechanism is then employed to align the global representations with the target tokens being predicted.

In contrast, for tokenizer-based embeddings, the sequential nature of words is preserved from the outset. Each sentence is converted into a sequence of token indices and projected into a fixed-dimensional embedding space through an embedding layer. The architecture employed here is a sequence-to-sequence (Seq2Seq) model with convolutional layers in both the encoder and decoder. The encoder utilizes convolutional filters with varying kernel sizes to capture local patterns at different granularities. The convolution outputs are normalized, re-projected, and enriched with residual connections to ensure stable representations. The decoder employs causal convolution to guarantee that the prediction of the current token depends only on previous tokens. An attention mechanism is also integrated to balance the decoder's focus on specific segments of the input sequence that are relevant to the target token. Thus, the TF-IDF and IndoBERT-based approaches are more closely aligned with text classification paradigms due to the loss of sequential information, while tokenizer-based embeddings better conform to the characteristics of sequence-to-sequence translation. Figure 3 illustrates the sequence-to-sequence translation architecture, comprising the embedding layer, encoder–decoder, and classifier layer. This architectural separation is essential to ensure that each representation is matched with the most appropriate modeling strategy according to the nature of the data it encodes.

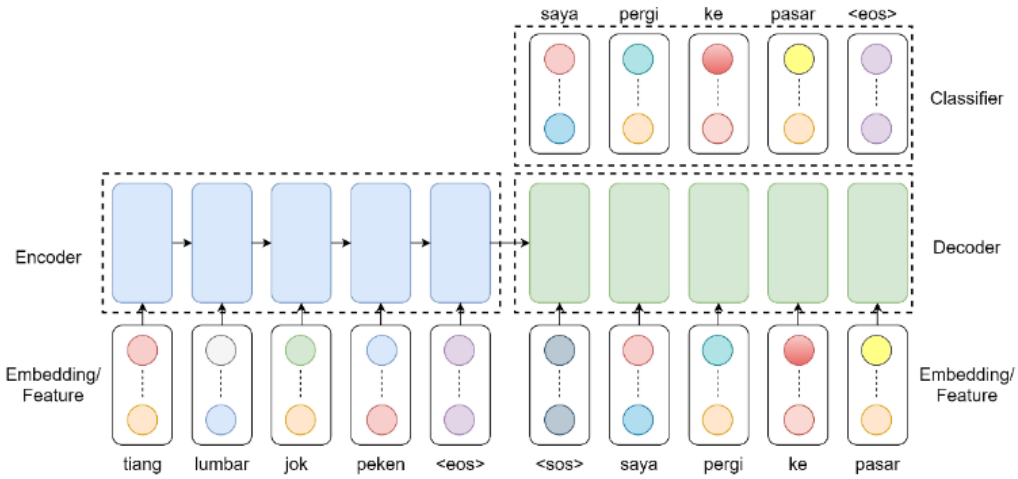


Figure 3. The general architecture of NMT consists of an embedding layer for word representation, an encoder network, a decoder network, and a classification layer [3]

The NMT model is designed to represent the probability distribution $p(y|x)$. In this formulation, $x = (x_1, x_2, x_3, \ldots, x_T)$ represents the sequence of source language tokens, while $y = (y_1, y_2, y_3, \ldots, y_T)$ represents the sequence of target language tokens [4]. The translation process begins with the encoder model, such as a Neural Network or feature representation, which transforms the sequence of source tokens into an encoded representation. This representation is then utilized by the decoder to gradually generate the sequence of target tokens. Accordingly, for a given source sentence x, the NMT system models the probability distribution over the target sentence y as expressed in Eqs. (4) and (5) [30]. Where θ denotes the set of model parameters. Furthermore, the NMT model can be trained by maximizing the log-likelihood, with $D_{x,y}$ representing the set of source–target sentence pairs.

$$P(y \mid x;\theta) = \prod_{t=1}^{T} P(y_t \mid y_{<t}, x;\theta) \quad (4)$$

$$\hat{\theta} = \text{argmax}_\theta \left\{ \sum_{\langle x,y\rangle \in D_{x,y}} \log P(y \mid x;\theta) \right\} \quad (5)$$

In the inference stage, this study employs greedy decoding, a straightforward decoding strategy in which the model selects the token with the highest probability at each step. This approach offers advantages in terms of speed and determinism, although it is prone to producing suboptimal outputs. Nevertheless, several studies have emphasized the importance of greedy decoding as a strong baseline in NMT experiments, particularly in low-resource scenarios [5, 31]. This strategy remains relevant for evaluating basic performance prior to applying more complex decoding methods. In the greedy approach, the selected token $\hat{y}_t$ at step t is determined by choosing the candidate that provides the maximum probability given the preceding context and the input x, as expressed in Eq. (6). The process iterates until the system generates the EOS symbol or reaches the decoding limit T [32].

$$\begin{aligned} y_0 &= SOS \\ y_t &= \text{argmax}_{y\in\bar{v}} \log\, q(y \mid x, y_{<t}) \text{ (for t} > 0) \end{aligned} \quad (6)$$

2.5 Evaluation

During model training, token-level accuracy was used as an initial metric to monitor the proportion of predicted tokens that matched the reference tokens. In the evaluation stage, this study focused on measuring translation quality using the Bilingual Evaluation Understudy (BLEU) metric. The main principle of BLEU is to compare machine-generated translations (candidate translations) with reference translations produced by native speakers or language experts. The evaluation process is carried out by calculating the modified n-gram precision, which considers clipped counts to prevent the model from achieving a high score merely by excessively repeating certain words. In addition, BLEU also applies a brevity penalty (BP) to control the tendency of the system to produce shorter sentences that may appear correct in terms of precision but are semantically incomplete. Furthermore, the BLEU metric evaluates translation quality from two perspectives: adequacy, through word-level precision, and fluency, by computing n-gram precision at levels n = 1, 2, 3, and 4 [33]. Precision is calculated by dividing the number of matched n-grams by the total number of n-grams. The precision for the n-th n-gram (p_n) can be formulated as in Eq. (7).

$$p_n = \frac{\sum_{p\in \text{ hypothesis}} \; n-\text{ gram} \; \sum_p \; \text{Count}_{\text{clip}}\,(n-\text{ gram})}{\sum_{p\in \text{ hypothesis}} \; n-\text{ gram} \; \sum_p \; \text{Count}\,(n-\text{ gram})} \quad (7)$$

In this calculation, BLEU takes into account the maximum frequency of n-gram matches. To prevent repeated counting of the same n-gram, a clipping mechanism is applied, which limits the number of n-gram matches based on the highest frequency found in one of the reference translations. This process produces the clipped count, which is mathematically formulated in Eq. (8).

$$\text{Count}_{\text{clip}} = min\Big(n-\text{ gram count}, \underset{r\in x}{max}(n-\text{ gram count in } r)\Big) \quad (8)$$

Thus, the BLEU score can be calculated using Eqs. (9) and (10).

$$BP = \begin{cases} 1 & \text{if } |h| > |r| \\ e^{(1-|r|/|h|)} & \text{if } |h| \le |r| \end{cases} \quad (9)$$

$$BLEU = BP \cdot \exp\left(\sum_{n=1}^{N} w_n \log\, p_n\right) \quad (10)$$

where, $|r|$ denotes the length of the reference sentence, $|h|$ denotes the length of the hypothesis sentence, p_n represents the precision value of the n-th n-gram, while w_n indicates the weight assigned to the corresponding n-gram. BLEU was chosen because it is a fast, efficient, and language-independent automatic evaluation method, based on the principle of calculating n-gram precision between machine translations and human references [34]. To ensure that the evaluation covers various levels of accuracy, this study adopts four variations of BLEU: BLEU-1, BLEU-2, BLEU-3, and BLEU-4. BLEU-1 assesses word-level (unigram) accuracy, BLEU-2 incorporates bigram context, while BLEU-3 and BLEU-4 extend the scope to trigram and four-gram to evaluate longer phrase coherence. Using BLEU at these different levels is essential to capture translation quality that is not only lexically correct but also syntactically and contextually appropriate. Thus, the evaluation provides a more comprehensive picture of the NMT system's performance in translating Sasak–Indonesian sentence pairs.

3. RESULT AND DISCUSSION

This subsection presents the results and discussion of the study, covering all stages from data collection, preprocessing, feature extraction, modeling, to evaluation. The key findings and observations at each stage are discussed to provide a comprehensive understanding of the overall process and the outcomes obtained.

3.1 Data collection

The data collection stage resulted in a total of 10,000 parallel sentence pairs in Sasak–Indonesian. Each entry consists of a source sentence in Sasak and its corresponding translation in Indonesian. The construction of these text pairs was carried out manually under strict supervision by researchers and linguists to ensure the quality of semantic equivalence. The process involved recording conversations, transcription, editing to remove inconsistencies or spelling errors, and final verification before being included in the research corpus. All data were stored in tabular format with column structures: index number, sentence in Sasak, and the corresponding sentence in Indonesian. Table 1 presents an example of the collected Sasak–Indonesian translation pairs. This format was chosen to facilitate preprocessing, feature extraction, and direct integration with the NMT pipeline. The data distribution covers various conversational domains, such as family interactions, market activities, school environments, and general social communication, providing the corpus with sufficiently representative contextual coverage.

Table 1. Sample of Sasak–Indonesian parallel text pairs

No.	Sasak Language	Indonesian Language	English
1	Sai aran side?	Siapa nama anda?	What is your name?
2	Pire umur side?	Berapa umur anda?	How old are you?
3	Embe taoq de mero mangkin?	Dimana tempat anda tinggal sekarang?	Where do you live now?
4	Jejah ku dengah kamu tempuk.	Khawatir saya dengar kamu dipukul.	I was worried when I heard you were hit.
5	Ape siq me jejahang?	Apa yang kamu takutkan?	What are you afraid of?
6	Ku nujaqang inambi jeje.	Saya menumbukkan ibumu gabah.	I pounded your mother's unhulled rice.
7	Jejeh kepeng leq atas meje.	Deretkan uang di atas meja.	Line up the money on the table.
8	Kanak-kanak no tejejeh isiq gurune.	Anak-anak itu di deretkan oleh gurunya.	The children were lined up by their teacher.
9	Dengan sugih sino jejeh kepeng ojok dengan jeleng.	Orang kaya itu bagikan uang kepada orang miskin.	The rich man distributed money to the poor.
10	Kereng uah bih tejejeh.	Sarung sudah habis dibagikan.	The sarong has been distributed.

Table 2. Preprocessing results of text for Seq2Seq model

No.	Sasak Language	Sasak Tokens	Sasak Text to Sequence & Padding
1	Sai aran side?	[‘<sos>’, ‘sai’, ‘aran’, ‘side’, ‘<eos>’]	[2, 11, 12, 4, 3, 0, 0, 0, 0, 0]
2	Pire umur side?	[‘<sos>’, ‘pire’, ‘umur’, ‘side’, ‘<eos>’]	[2, 13, 14, 4, 3, 0, 0, 0, 0, 0]
3	Embe taoq de mero mangkin?	[‘<sos>’, ‘embe’, ‘taoq’, ‘de’, ‘mero’, ‘mangkin’, ‘<eos>’]	[2, 15, 16, 17, 18, 19, 3, 0, 0, 0]
4	Jejah ku dengah kamu tempuk	[‘<sos>’, ‘jejah’, ‘ku’, ‘dengah’, ‘kamu’, ‘tempuk’, ‘<eos>’]	[2, 20, 5, 21, 22, 23, 3, 0, 0, 0]
5	Ape siq me jejahang?	[‘<sos>’, ‘ape’, ‘siq’, ‘me’, ‘jejahang’, ‘<eos>’]	[2, 24, 25, 26, 27, 3, 0, 0, 0, 0]
6	Ku nujaqang inambi jeje	[‘<sos>’, ‘ku’, ‘nujaqang’, ‘inambi’, ‘jeje’, ‘<eos>’]	[2, 5, 28, 29, 30, 3, 0, 0, 0, 0]
7	Jejeh kepeng leq atas meje	[‘<sos>’, ‘jejeh’, ‘kepeng’, ‘leq’, ‘atas’, ‘meje’, ‘<eos>’]	[2, 6, 7, 31, 32, 33, 3, 0, 0, 0]
8	Kanak-kanak no tejejeh isiq gurune	[‘<sos>’, ‘kanak’, ‘kanak’, ‘no’, ‘tejejeh’, ‘isiq’, ‘gurune’, ‘<eos>’]	[2, 8, 8, 34, 9, 35, 36, 3, 0, 0]
9	Dengan sugih sino jejeh kepeng ojok dengan jeleng	[‘<sos>’, ‘dengan’, ‘sugih’, ‘sino’, ‘jejeh’, ‘kepeng’, ‘ojok’, ‘dengan’, ‘jeleng’, ‘<eos>’]	[2, 10, 37, 38, 6, 7, 39, 10, 40, 3]
10	Kereng uah bih tejejeh	[‘<sos>’, ‘kereng’, ‘uah’, ‘bih’, ‘tejejeh’, ‘<eos>’]	[2, 41, 42, 43, 9, 3, 0, 0, 0, 0]

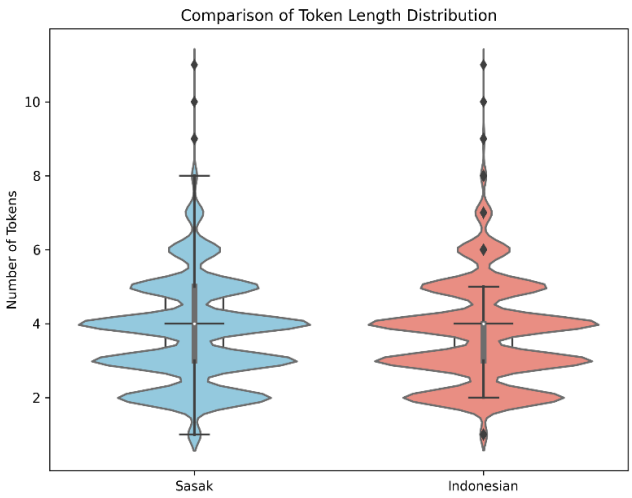


Figure 4. Violin plot of token length distribution per sentence

The parallel corpus consisting of 10,000 sentence pairs can be categorized as a medium-sized dataset for research on low-resource languages. This amount is consistent with prior studies in the development of machine translation systems for minority languages, which typically employ corpora ranging from thousands to tens of thousands of sentence pairs [1, 2, 35]. Based on the analysis of the 10,000 Sasak–Indonesian parallel sentence pairs, the corpus exhibits diverse linguistic characteristics. On the Sasak side, the vocabulary size reaches 5,517, while the Indonesian vocabulary consists of 4,256 entries. This difference indicates that Sasak possesses greater lexical variety, likely influenced by dialectal variations, morphology, and more complex word derivations compared to Indonesian. Furthermore, sentence length analysis shows that the maximum number of words in a single sentence for both languages is relatively balanced (up to 11 words), which facilitates alignment in the translation model. The sentence length distribution for each language is illustrated in Figure 4. The violin plot demonstrates that token lengths in Sasak and Indonesian are relatively balanced, with similar medians and dispersions. This suggests that sentences in both languages tend to be comparable in length when measured by token count. Such a balance is important because it allows the model to learn cross-lingual representations more stably without significant disparities in input–output lengths. Thus, this dataset not only holds practical value for building a Sasak–Indonesian translation system but also contributes to efforts in preserving and digitizing regional languages within both academic and technological domains.

3.2 Preprocessing

Before training the Sasak–Indonesian NMT model, the text data was first processed through a series of preprocessing steps to ensure consistency and compatibility with subsequent stages. This process included text cleaning (lowercasing, punctuation removal, whitespace normalization), tokenization, and the addition of special tokens <sos> and <eos> to mark the beginning and end of each sentence. The results of preprocessing are shown in Table 2. The cleaned text, tokenized sequence, and the addition of special tokens are presented in the *Sasak Tokens* column. Subsequently, these tokens were converted into numerical sequences using a tokenizer, followed by sequence padding to standardize the length, ensuring that all data maintained a consistent size. This preprocessing yielded numerical representations of the Sasak/Indonesian target sentences, which were ready to be used as input and output in the encoder–decoder model. Here, <sos> serves as the start-of-sequence marker with a numerical value of 2, <eos> as the end-of-sequence marker with a numerical value of 3, and padding guarantees efficiency in batch processing.

3.3 Feature extraction

The feature extraction stage revealed three distinct types of representations corresponding to the approaches used. In the TF-IDF-based representation, each sentence was successfully converted into a high-dimensional vector corresponding to the size of the vocabulary. This representation is effective in identifying key words within everyday conversational corpora; however, it is static, cannot capture inter-word contextual relationships, and produces sparse vectors. In contrast, IndoBERT generates fixed-dimensional embeddings rich in contextual information with a dimension of 768. These embeddings not only reflect the presence of words but also account for syntactic and semantic interactions among tokens within a sentence. Meanwhile, the tokenizer sequence embeddings produce a sequence of vectors standardized through padding, with an embedding dimension of 300. These representations are built directly from the research corpus, making them more flexible in capturing Sasak–Indonesian specific vocabulary. The distribution of the generated embeddings shows distinct semantic clustering, particularly for frequently occurring words in local conversations, indicating that data-trained embeddings possess strong adaptive capabilities to the characteristics of the source language. Overall, the three approaches provide complementary results: TF-IDF is effective in highlighting important words, IndoBERT excels at capturing contextual relationships, and tokenizer sequence embeddings are adaptive to local vocabulary.

3.4 Modeling

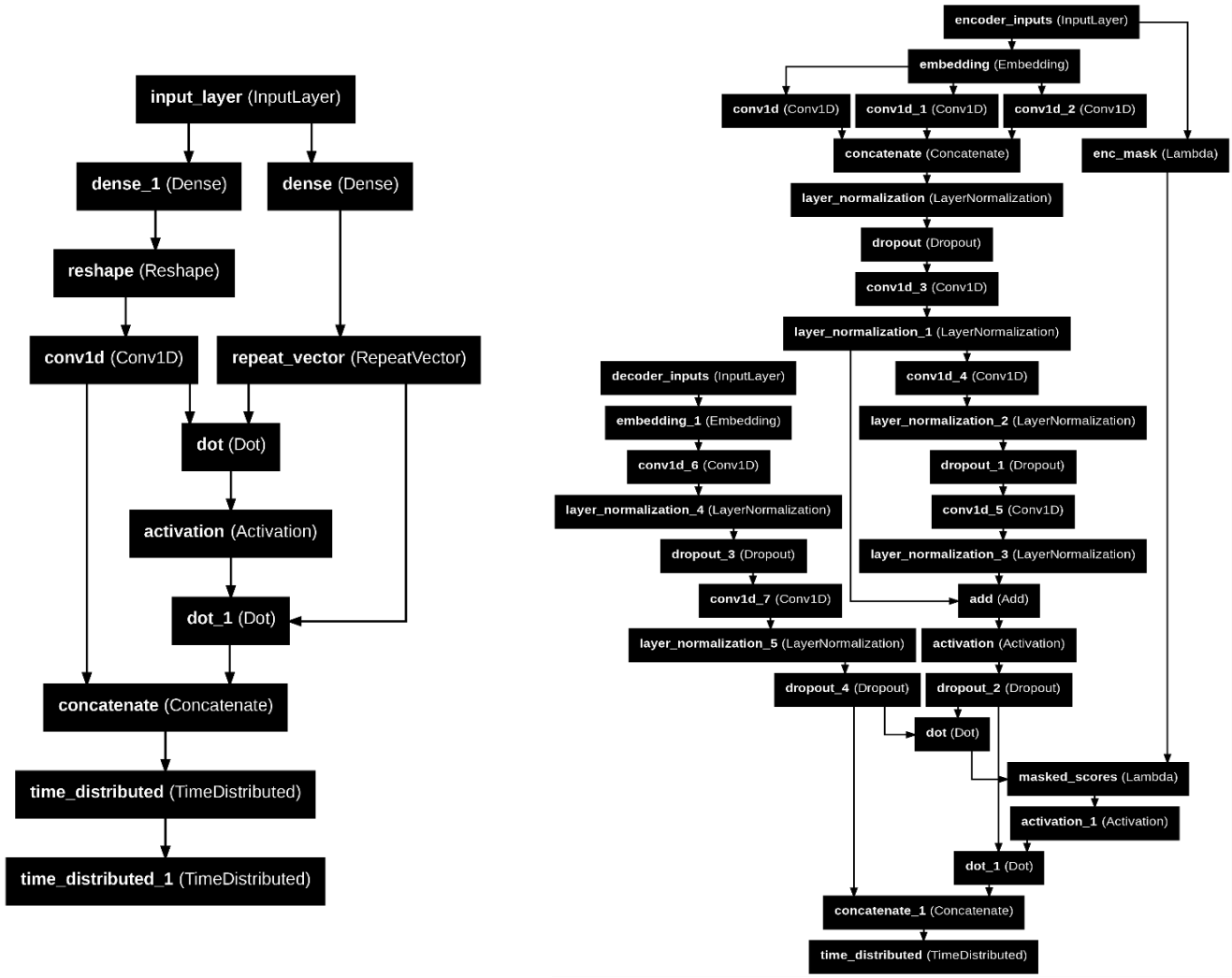
The implementation of the modeling architecture based on the three feature extraction approaches resulted in model structures tailored to the type of input used. In the TF-IDF- and IndoBERT-based models, the compilation results indicate that the initial layers are not sequential embeddings but fixed-dimensional input vectors. These vectors are then projected through dense layers and expanded using a repeat vector to resemble a sequence. Convolutional layers (Conv1D) are applied in the decoder to extract patterns from the repeated representations, while the attention mechanism is constructed using a dot product operation between the convolutional outputs and the context vector. This process enables the model to generate sequential outputs even when the initial input lacks inherent sequential properties. The structure is reflected in the model flow, consisting of Input, Dense, RepeatVector, Conv1D, Attention, TimeDistributed Dense, and Softmax layers. As an illustration, Figure 5 presents the NMT model architectures developed according to the type of input. Figure 5(a) shows the architecture using TF-IDF and IndoBERT feature representations as the encoder to produce contextual embeddings. Meanwhile, Figure 5(b) depicts the Seq2Seq-based tokenizer embedding architecture, which directly captures sequential information from the parallel data.

Meanwhile, in the tokenizer embedding-based model, the architecture is more complex because the input consists of token sequences. Embedding layers are applied to both the encoder and decoder, followed by several convolutional layers with varying kernel sizes (3, 5, and 7) in the encoder to capture local pattern variations. The convolution outputs are combined, normalized, and enhanced with residual connections before being re-projected. The decoder employs causal convolution to ensure that token predictions follow the sequential order. The attention mechanism is then formed via a dot product between the decoder outputs and the encoder representations, which is subsequently combined to generate the probability distribution for each target token. The complete model flow consists of Input, Embedding, Conv1D (multi-kernel), Residual, Decoder Causal Conv, Attention, TimeDistributed Dense, and Softmax layers. Additionally, dropout and layer normalization are incorporated to improve training stability, maintain consistent activation distributions, and prevent overfitting [6, 36]. Model training was conducted over 100 epochs with a batch size of 32, with the data split into 80% for training, 10% for validation, and 10% for testing.

3.5 Evaluation

This study evaluates the performance of several feature representation approaches in NMT models for the Sasak–Indonesian language pair and vice versa. The NMT models utilize three types of feature representations: TF-IDF, IndoBERT, and tokenizer sequence embeddings. The evaluation results for Sasak–Indonesian and Indonesian–Sasak translation are presented in Table 3. Overall, the Seq2Seq + CNN model demonstrates superior performance compared to the TF-IDF + CNN and IndoBERT + CNN approaches in both translation directions. Token-level accuracy reaches 84.47% for Sasak–Indonesian and 81.09% for Indonesian–Sasak, while BLEU-1 scores are 56.63% and 44.25%, respectively, indicating the model's ability to produce more natural translations that are consistent with the source sentence context. The BLEU score is considered capable of representing translation comprehensibility and fluency, where a score above 30% indicates that a sentence is understandable, and a score above 50% reflects good and fluent translations [37]. This advantage underscores the Seq2Seq model's capability to preserve sequential information and leverage the attention mechanism to capture contextual relationships between tokens, resulting in more coherent and fluent translations. In contrast, the TF-IDF-based model performs less effectively due to its sparse, static vector representations, whereas IndoBERT, despite providing contextual embeddings, is less sensitive to word order, leading to lower

BLEU scores. These performance differences highlight the importance of sequential modeling for low-resource languages, where inter-token relationships are critical for producing comprehensible translations that retain semantic meaning. The reported results are derived from single experimental runs using the same random seed (42) across all models to ensure a fair and consistent comparison. While this setup may limit the statistical generalizability of the findings, the consistent improvements observed across all metrics indicate that the proposed sequential CNN-based embeddings demonstrate stable and robust translation performance.



(a) Architecture for TF-IDF and IndoBERT features

(b) Architecture for Seq2Seq tokenizer embeddings

Figure 5. NMT model architectures based on input type

Table 3. Token-level accuracy and BLEU scores of bidirectional NMT models

No.	Model	Accuracy	BLEU-1	BLEU-2	BLEU-3	BLEU-4
1	TFIDF + CNN Sasak-Indo	72.37%	36.78%	22.02%	13.68%	8.48%
2	IndoBERT + CNN Sasak-Indo	76.16%	32.93%	17.23%	8.57%	4.19%
3	Seq2Seq + CNN Sasak-Indo	84.47%	56.63%	42.68%	31.91%	22.31%
4	TFIDF + CNN Indo-Sasak	71.53%	30.19%	16.27%	8.80%	4.33%
5	IndoBERT + CNN Indo-Sasak	76.37%	34.35%	19.05%	10.62%	5.60%
6	Seq2Seq + CNN Indo-Sasak	81.09%	44.25%	29.99%	19.89%	12.24%

The phenomenon of high token-level accuracy accompanied by lower BLEU scores highlights the distinction between individual token prediction accuracy and overall sentence quality. In the context of low-resource languages, the model can correctly predict words in isolation, but the word order or sentence structure in the translated output is not fully optimal, resulting in reduced n-gram alignment with the reference. This trend is observed across all models, where accuracy ranges from 71% to 84%, while BLEU scores remain relatively lower, indicating persistent challenges in capturing complex syntactic structures and phrase coherence. These findings underscore the importance of a sequence-to-sequence approach that preserves contextual information and token order patterns to produce translations that are more coherent and comprehensible. Analysis of BLEU-1 through BLEU-4 trends provides additional insights into the model's ability to maintain both word-level accuracy and sequential coherence. The performance trend based on BLEU metrics is illustrated in Figure 6.

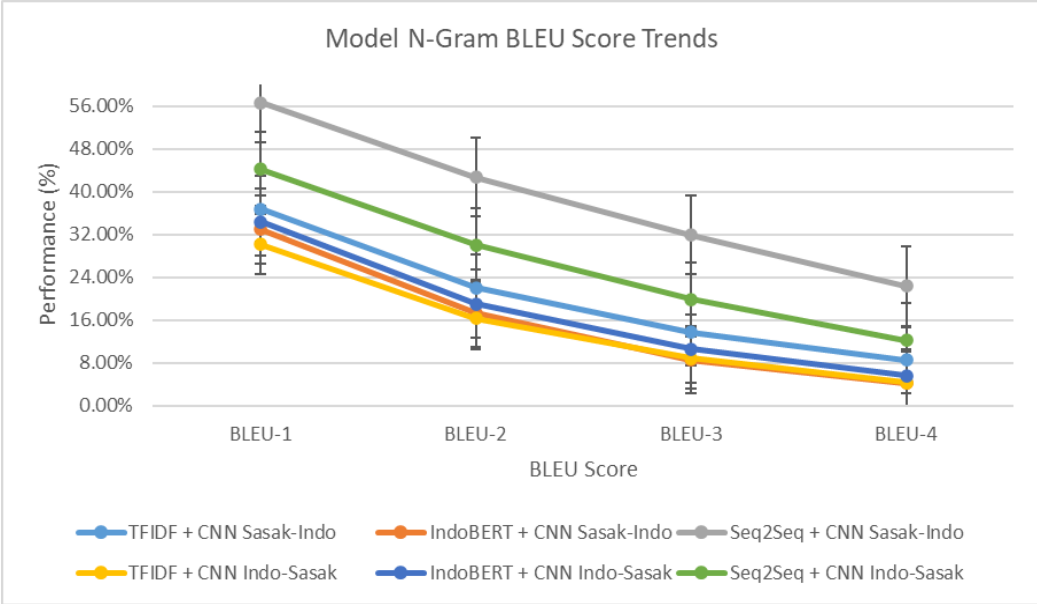


Figure 6. NMT performance trends based on BLEU scores

Table 4. Evaluation of model prediction errors

No.	Model	Sasak	Reference	Prediction	BLEU-1
1	TFIDF + CNN Sasak-Indo	Sai me suruq popoq ye	Siapa kamu suruh cuci itu (Who told you to wash that?)	*Siapa suruh kamu cuci* (Who told you to wash?)	77.88%
2	IndoBERT + CNN Sasak-Indo			*Siapa yang kamu cuci* cuci (Who that you wash wash?)	60%
3	Seq2Seq + CNN Sasak-Indo			*Siapa kamu akan cuci itu* (Who you will wash that?)	80%
4	TFIDF + CNN Indo-Sasak	Jangan main hp terus (Don't keep playing phone.)	Endaq main hp doang	*Endaq behape doang* (Don't keep playing phone.)	47.76%
5	IndoBERT + CNN Indo-Sasak			*Endaq main doang* (Don't keep playing.)	71.65%
6	Seq2Seq + CNN Indo-Sasak			*Endaq main hp doang* (Don't keep playing phone.)	100%

In general, all models exhibit a decline in BLEU scores as the n-gram length increases, from unigram (BLEU-1) to four-gram (BLEU-4). This indicates that the models are relatively more capable of recognizing key words (unigrams) but struggle to maintain longer word sequences and complex phrase coherence (trigrams and four-grams). This behavior is influenced by the dataset characteristics, where most training sentences are only 3–5 words long, whereas some test sentences extend up to 11 words. The difference in sequence length causes the models to have difficulty consistently predicting long word sequences, resulting in lower BLEU scores for higher-order n-grams. Additionally, for Sasak-to-Indonesian translation, the TF-IDF-based model performs better than IndoBERT, whereas for Indonesian-to-Sasak translation, IndoBERT demonstrates superior performance. This can be attributed to IndoBERT's prior training and familiarity with Indonesian sentence structures and sequences, enabling it to capture context and word order more effectively in the source language.

Based on an analysis of a parallel corpus comprising 10,000 Sasak–Indonesian sentence pairs, the comparison of translation performance between the Sasak-to-Indonesian and Indonesian-to-Sasak directions reveals a significant difference. Overall, translations from Sasak to Indonesian tend to perform better than the reverse direction. This can be explained by the linguistic characteristics of the corpus: the Sasak language exhibits a larger vocabulary size compared to Indonesian. Such variation reflects greater lexical diversity in Sasak, likely influenced by dialectal variations, morphology, and more complex derivational forms. This diversity poses challenges for the model in generating natural Sasak text, whereas translation into Indonesian is comparatively easier due to a more consistent vocabulary and standardized sentence structures. To further understand model performance, error analysis was conducted on selected prediction examples with the assistance of linguists, as presented in Table 4.

The main types of errors can be categorized as lexical substitution, omission, insertion, and word order errors. In the Sasak-to-Indonesian translation direction, the sentence "sai me suruq popoq ye" has the reference translation "siapa kamu suruh cuci itu" (Who told you to wash that). The TF-IDF + CNN model produced "siapa suruh kamu cuci," showing an omission error as the word "itu" is missing; however, the word order is relatively correct, resulting in a BLEU-1 score of 77.88%. The IndoBERT + CNN model predicted "siapa yang kamu cuci cuci," which includes an insertion of the word "yang" and repetition of "cuci," along with slight word order disruption, yielding a lower BLEU-1 score of 60%. The Seq2Seq + CNN model generated "siapa kamu akan cuci itu," adding the word "akan" (insertion), but key words were preserved and word order was correct, resulting in a relatively high BLEU-1 of 80%.

In the Indonesian-to-Sasak translation direction, the sentence "jangan main hp terus" (don't keep playing phone)

has the reference “endaq main hp doang.” The TF-IDF + CNN model produced “endaq behape doang,” indicating a lexical substitution error for “hp” to “behape” and a slight word order shift, resulting in a low BLEU-1 score of 47.76%. Semantically, however, “main hp” in Sasak can also be interpreted as “behape,” highlighting that TF-IDF and BLEU evaluation do not fully capture the semantic meaning of translations. The BLEU method essentially evaluates translation quality by matching n-grams between strings. However, one of its main limitations is its inability to capture semantic similarities between words [38]. The IndoBERT + CNN model predicted “endaq main doang,” omitting the word “hp,” but retaining the main meaning of the sentence, yielding a BLEU-1 score of 71.65%. The Seq2Seq + CNN model produced the exact reference translation, “endaq main hp doang,” achieving a BLEU-1 of 100%. This analysis indicates that model errors primarily arise from omissions or lexical substitutions, particularly in translations into Sasak. An overview of prediction errors is presented in Figures 7 and 8, highlighting the performance of the best-performing model.

Figures 7 and 8 depict the probability distributions of the top five predicted tokens at each token position for the Sasak–Indonesian sentence pair “sakit gati tian ke” – “sakit sekali perut ku”. The vertical axis represents token positions, while the horizontal axis shows candidate predicted tokens. Color intensity reflects prediction probability, with darker shades indicating higher probabilities. The visualization shows that both models accurately predicted the start <sos> and end <eos> tokens with a probability of 1.0, indicating reliable initialization and termination mechanisms. However, predicting the tokens forming the actual sentence remains challenging for the Indonesian–Sasak model; several tokens appear with low probability and incorrect order. For example, the Sasak word “tian”, which aligns with “perut” (stomach), is incorrectly predicted as “pinggang” (waist) in Indonesian. Various strategies have been explored to enhance NMT for low-resource languages, including fine-tuning large models with synthetic data, knowledge distillation, preprocessing and subword tokenization, multimodal information, cross-lingual transfer, and adversarial or multilingual training. Table 5 provides a contextual overview of recent advances in NMT for low-resource languages. While the metrics, dataset sizes, and evaluation setups vary considerably across studies, this comparison is not intended as a direct numerical benchmark. Instead, it aims to position the methodological contribution of this study, specifically the use of sequential embeddings integrated within a CNN-based Seq2Seq architecture for the Sasak–Indonesian language pair, within the broader landscape of low-resource NMT research. This contextual framing emphasizes how the proposed approach differs from dominant Transformer or LLM-based methods by focusing on lightweight architectures that can perform effectively under data scarcity.

Table 5. Recent advances in techniques for improving NMT in low-resource languages

Author	Language	Dataset (Size)	Model	Low-Resource Strategy	Metrics	Main Results	Contribution
Tan and Zhu [39]	Balinese, Minangkabau ↔ Indonesian / English	~55.2 k cleaned parallel pairs	Fine-tuned LLaMA2-7B (Komodo-7B base)	Continued pretraining, LLM-based data cleaning, backtranslation/self-learning	spBLEU	For translations *into* local languages, beats SoTA by +6.69 spBLEU	Shows that fine-tuning LLM with rigorous data cleaning + synthetic data can significantly improve low-resource MT
Roy et al. [40]	Indic languages (Hindi, Kannada, Punjabi)	Samanantar corpora; large parallel sets	Seq2Seq with multilingual encoder (XLM-R) + decoder	Complementary Knowledge Distillation (CKD), encoder freezing	BLEU-4, chrF, Human Eval	CKD model outperforms baselines by +1.22 to +5.15 BLEU	Demonstrates benefit of multilingual encoder + knowledge distillation for low-resource Indic NMT
Agrawal et al. [41]	English ↔ Manipuri; English ↔ Assamese	~53 k Eng–Assamese, ~24.3 k Eng–Manipuri	Transformer	Preprocessing (deduplication, noise removal, sentence length filtering, language ID)	BLEU, chrF2, RIBES, TER	Eng→Manipuri BLEU = 15.02; Eng→Assamese BLEU = 5.47	Baseline Transformer performance for low-resource Indic languages
Singh et al. [42]	English ↔ Manipuri	~21,687 training pairs; 1,000 dev/test	Transformer (4-layer encoder & decoder)	Subword tokenization (BPE), hyperparameter tuning	BLEU, chrF, RIBES, TER, COMET	Eng→Manipuri BLEU = 22.75; Man→English = 26.92	A competitive NMT setup for English-Manipuri, benchmark across multiple metrics
Yang et al. [43]	Multimodal (text + images)	Parallel text + image corpora (COCO + bilingual captions)	Transformer with cross-modal alignment module	Use image–text alignment to supplement scarce text parallel data	sacreBLEU	Gains in BLEU when image context is available	Shows multimodal alignment can help translation under data scarcity
Goyal et al. [44]	Indo-Aryan languages ↔	ILCI (~50 k per language),	Transformer (6-layer)	Multilingual transfer learning +	BLEU	~+5 BLEU improvement	Leverages related

	English	IIT-Bombay EN–HI (~1.5M)		transliteration + shared subword vocab		over baseline	languages (transliteration + shared subwords) to boost low-resource translation
Sun et al. [45]	Multiple low-resource languages → English / English → LRL	Several extremely low-resource pairs (e.g., Ind-Eng, Tgl-Eng, Kor-Eng) from Tatoeba dataset (few thousand pairs)	Seq2Seq generator + adversarial discriminator (LAC model)	Adversarial learning + transfer learning, pretrained discriminator	BLEU, ChrF	BLEU improvement over baseline Seq2Seq models (e.g., +4.3 BLEU on Turkish–English, gains also on Ind-Eng, Tgl-Eng)	Proposes LAC model combining adversarial & transfer learning to better generalize in extremely low-resource translation
Rubino et al. [1]	English ↔ Asian languages (Japanese, Lao, Malay, Vietnamese)	ALT corpus: 20,106 parallel sentences + large monolingual corpora	Transformer	Hyper-parameter tuning, multilingual training, back-translation with tagging	BLEU, chrF	Back-translation & multilingual training improved BLEU significantly	Provides best practices for extremely low-resource Transformer MT, emphasizing tagging and multilingual transfer
Our Study (2025)	Sasak ↔ Indonesian	~10 k parallel pairs	Seq2Seq CNN encoder-decoder + embeddings (TF-IDF, IndoBERT, Sequential)	Sequential embeddings with CNN; comparison with TF-IDF and IndoBERT	Accuracy, BLEU-1 to BLEU-4	Sasak→Indo: Acc 84.47%, BLEU-1=56.63, BLEU-4=22.31%; Indo→Sasak: Acc 81.09%, BLEU-1=44.25, BLEU-4=12.24	Shows sequential embeddings + CNN outperform TF-IDF and IndoBERT; potential for preserving regional low-resource languages

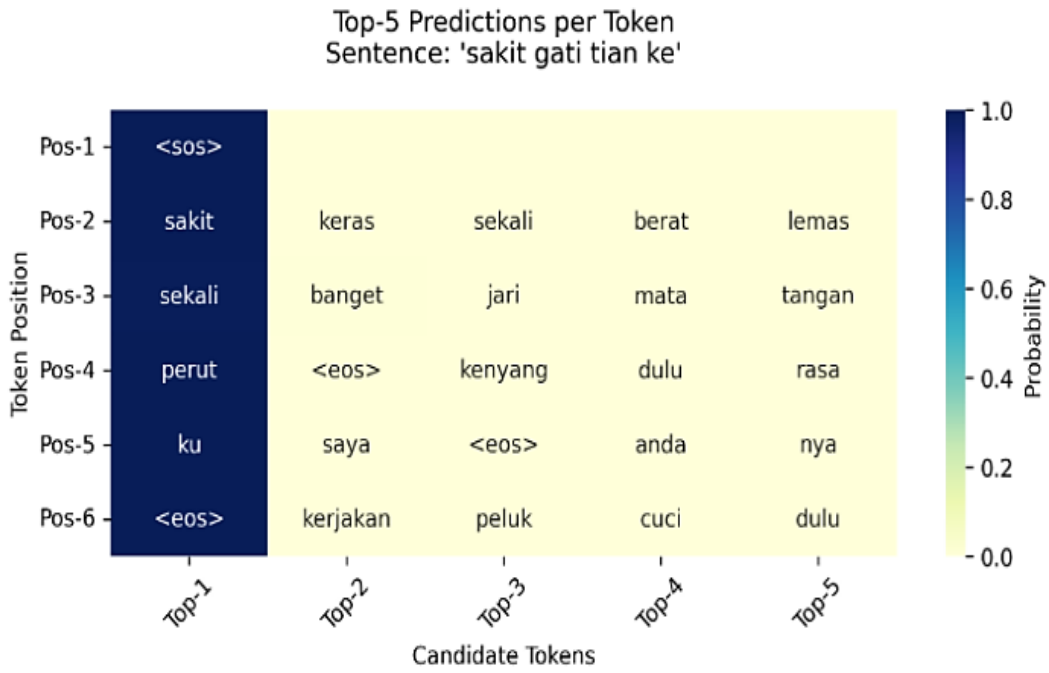


Figure 7. Top-5 token prediction probabilities per position for the Sasak-to-Indonesian NMT model

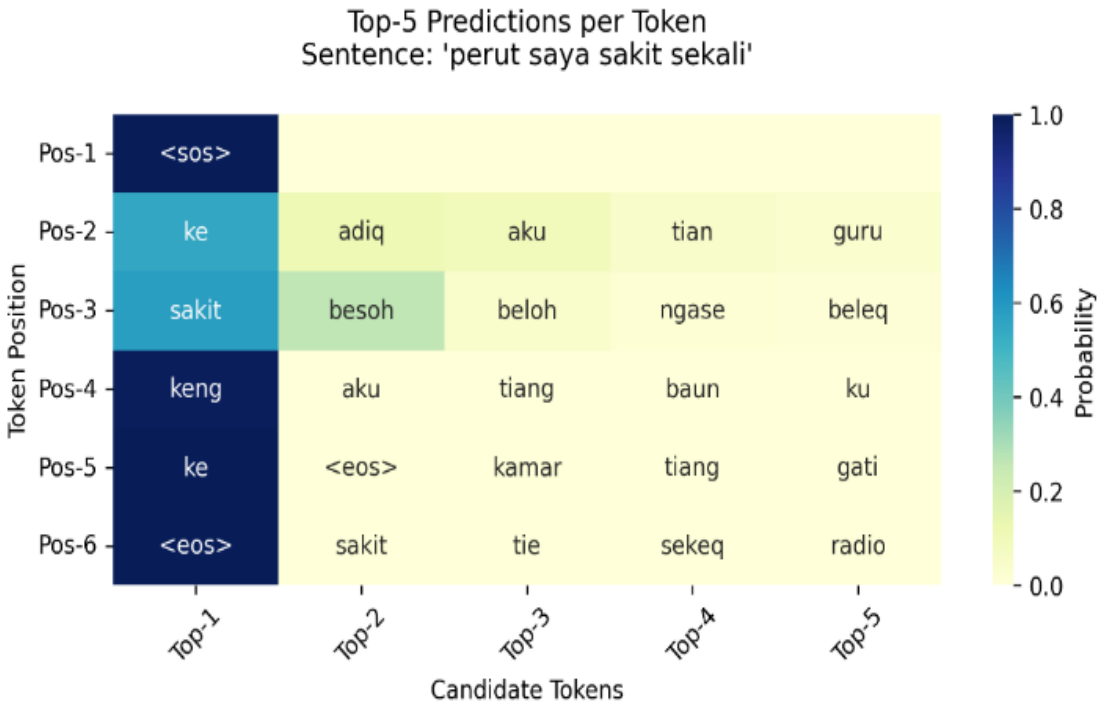


Figure 8. Top-5 token prediction probabilities per position for the Indonesian-to-Sasak NMT model

As summarized in Table 5, most prior works have relied on Transformer-based or multilingual pretraining approaches, often supported by large-scale or synthetic parallel corpora. In contrast, this study investigates a CNN-based Seq2Seq model leveraging sequential embeddings as a compact yet effective alternative for low-resource translation. Despite the relatively small dataset size of approximately 10,000 pairs, the model demonstrates competitive accuracy and BLEU scores within its experimental scope. However, because metrics, preprocessing pipelines, and corpus sizes differ substantially among studies, the comparison presented here is qualitative rather than quantitative. In this study, IndoBERT is used as the embedding model, which is specifically designed for the Indonesian language and has also been applied effectively in another work [46]. Future research should aim to establish standardized benchmarks and shared datasets for Indonesian regional languages to enable more rigorous and reproducible cross-study evaluations.

4. CONCLUSIONS

This study demonstrates that the sequential tokenizer embedding/Seq2Seq approach with a CNN encoder-decoder outperforms TF-IDF and IndoBERT representations in NMT for the Sasak–Indonesian language pair and vice versa. The Seq2Seq + CNN model effectively preserves contextual information and token order, achieving high token-level accuracy (84.47% for Sasak–Indonesian and 81.09% for Indonesian–Sasak) along with improved BLEU scores. The decline in BLEU scores for longer n-grams indicates the model's difficulty in capturing complex word sequences, particularly for test sentences longer than those in the training data. vocabulary size in both languages also affects predictions, especially for models that do not adequately preserve token order and semantic context. Error analysis reveals that the primary sources of errors are omission, insertion, substitution, and word order mistakes, with translations into Sasak being more susceptible to lexical errors. However, several limitations remain, particularly the restricted dataset size, limited vocabulary coverage, and the model's difficulty in generalizing long or syntactically complex sentences. For future research, it is recommended to pre-train richer contextual embeddings and integrate them within a Transformer-based Seq2Seq NMT architecture. Expanding the corpus with more diverse linguistic structures and exploring multilingual pretraining strategies are also suggested. This approach is expected to enhance the model's ability to capture more complex sequential and semantic relationships, thereby improving translation quality, particularly for long sentences and low-resource languages.

ACKNOWLEDGMENT

The authors gratefully acknowledge the Ministry of Higher Education, Science, and Technology of the Republic of Indonesia, particularly through the Directorate of Research, Technology, and Community Service, for the financial support (Grant No.: 0070/C3/AL.04/2025) dated 23 May 2025 under the Fundamental Research Scheme for the Fiscal Year 2025. Special thanks are also addressed to the Institute for Research and Community Service of Bumigora University for providing invaluable support, facilities, and coordination throughout the course of this study.

REFERENCES

[1] Rubino, R., Marie, B., Dabre, R., Fujita, A., Utiyama, M., Sumita, E. (2020). Extremely low-resource neural machine translation for Asian languages. Machine Translation, 34: 347-382. https://doi.org/10.1007/s10590-020-09258-6

[2] Agyei, E., Zhang, X.L., Bannerman, S., Quaye, A.B., Yussi, S.B., Agbesi, V.K. (2024). Low resource Twi-

English parallel corpus for machine translation in multiple domains (Twi-2-ENG). Discover Computing, 27: 17. https://doi.org/10.1007/s10791-024-09451-8

[3] Tan, Z.X., Wang, S., Yang, Z.H., Chen, G., Huang, X.C., Sun, M.S., Li, Q., Liu, Y. (2020). Neural machine translation: A review of methods, resources, and tools. AI Open, 1: 5-21. https://doi.org/10.1016/j.aiopen.2020.11.001

[4] Pang, R.Y., He, H., Cho, K. (2022). Amortized noisy channel neural machine translation. In Proceedings of the 15th International Conference on Natural Language Generation, pp. 131-143. https://doi.org/10.18653/v1/2022.inlg-main.11

[5] Ekle, O.A., Das, B. (2025). Low-resource neural machine translation using recurrent neural networks and transfer learning: A case study on English-to-Igbo. arXiv preprint arXiv:2504.17252. https://doi.org/10.48550/arXiv.2504.17252

[6] Israr, H., Khan, S.A., Tahir, M.A., Shahzad, M.K., Ahmad, M., Zain, J.M. (2023). Neural machine translation models with attention-based dropout layer. Computational Materials and Continua, 75(2): 2981-3009. https://doi.org/10.32604/cmc.2023.035814

[7] Wang, Y. (2024). Research on the TF–IDF algorithm combined with semantics for automatic extraction of keywords from network news texts. Journal of Intelligent Systems, 33(1): 20230300. https://doi.org/10.1515/jisys-2023-0300

[8] Agustina, C.A.N., Novita, R., Mustakim, Rozanda, N.E. (2024). The implementation of TF-IDF and Word2Vec on booster vaccine sentiment analysis using support vector machine algorithm. Procedia Computer Science, 234: 156-163. https://doi.org/10.1016/j.procs.2024.02.162

[9] Aranta, A., Djunaidy, A., Suciat, N. (2024). Preserving Sasak dialectal features in English to Sasak machine translation through locked tokenization with transformer models. In 2024 International Seminar on Intelligent Technology and its Applications (ISITIA), Mataram, Indonesia, pp. 19-24. https://doi.org/10.1109/ISITIA63062.2024.10667682

[10] Shabrina, M.M., Aranta, A., Irmawati, B. (2024). Design and development of Indonesian voice conversion algorithm to Sasak language Latin text using dictionary based method. JTIKA (Jurnal Teknologi Informasi, Komputer dan Aplikasi), 6(1): 364-375. https://doi.org/10.29303/jtika.v6i1.371

[11] Wardhana, H., Dharma, I.M.Y., Marzuki, K., Hidayatullah, I.S. (2024). Implementation of neural machine translation in translating from Indonesian to Sasak language. Matrik: Jurnal Manajemen, Teknik Informatika, dan Rekayasa Komputer, 23(2): 456-476. https://doi.org/10.30812/matrik.v23i2.3465

[12] Aranta, A., Djunaidy, A., Suciati, N. (2024). Developing an English-to-Indonesian speech-to-text as a foundation for Sasak language translation using the mBART algorithm. IET Conference Proceedings, 2024(30): 568-573. https://doi.org/10.1049/icp.2025.0311

[13] Lo, S. (2024). The effects of NMT as a de facto dictionary on vocabulary learning: A comparison of three look-up conditions. Computer Assisted Language Learning, 38(7): 1611-1631. https://doi.org/10.1080/09588221.2024.2310285

[14] Chistova, E.V. (2025). Principles of developing a Chinese-Russian polysemantic dictionary as a means of improving interpretability of neural machine translators. Professional Discourse & Communication, 7(1): 89-107. https://doi.org/10.24833/2687-0126-2025-7-1-89-107

[15] Dwivedi, R.K., Nand, P., Pal, O. (2024). Hybrid NMT model and comparison with existing machine translation approaches. Multidisciplinary Science Journal, 7(4): e2025146. https://doi.org/10.31893/multiscience.2025146

[16] De Silva, D.I., Hansadi, D.G.P. (2024). Enhancing machine translation: Cross-approach evaluation and optimization of RBMT, SMT, and NMT techniques. In 2024 International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems (ICSES), Chennai, India, pp. 1-8. https://doi.org/10.1109/ICSES63760.2024.10910801

[17] Dimakis, A., Markantonatou, S., Anastasopoulos, A. (2024). Dictionary-aided translation for handling multi-word expressions in low-resource languages. In Findings of the Association for Computational Linguistics: ACL 2024, Bangkok, Thailand, pp. 2588-2595. https://doi.org/10.18653/v1/2024.findings-acl.152

[18] Chakrawarti, R.K., Pathik, N., Chauhan, P.S., Jaiswal, L. (2025). NMT Framework for verse translation from Hindi to English with CDEC. In 2025 First International Conference on Advances in Computer Science, Electrical, Electronics, and Communication Technologies (CE2CT), Bhimtal, Nainital, India, pp. 457-464. https://doi.org/10.1109/CE2CT64011.2025.10939390

[19] Mazi, C.C., Anya, O.I., Nwanakwaugwu, A.C., Anichebe, G. (2024). Machine translation and natural language processing. International Journal of Science and Management Studies, 7(1): 65-69. https://doi.org/10.51386/25815946/ijsms-v7i1p110

[20] Siino, M., Tinnirello, I., La Cascia, M. (2024). Is text preprocessing still worth the time? A comparative survey on the influence of popular preprocessing methods on Transformers and traditional classifiers. Information Systems, 121: 102342. https://doi.org/10.1016/j.is.2023.102342

[21] Palomino, M.A., Aider, F. (2022). Evaluating the effectiveness of text pre-processing in sentiment analysis. Applied Sciences, 12(17): 8765. https://doi.org/10.3390/app12178765

[22] Cho, H., Cha, J., Awasthi, P., Bhojanapalli, S., Gupta, A., Yun, C. (2024). Position coupling: Improving length generalization of arithmetic transformers using task structure. arXiv preprint arXiv:2405.20671. https://doi.org/10.48550/arXiv.2405.20671

[23] Zaoad, M.S., Mannan, M.M.R., Mandol, A.B., Rahman, M., Islam, M.A., Rahman, M.M. (2023). An attention-based hybrid deep learning approach for Bengali video captioning. Journal of King Saud University - Computer and Information Sciences, 35(1): 257-269. https://doi.org/10.1016/j.jksuci.2022.11.015

[24] Mohd Sofi, S., Selamat, A. (2023). Aspect based sentiment analysis: Feature extraction using Latent Dirichlet Allocation (LDA) and Term Frequency - Inverse Document Frequency (TF-IDF) in Machine Learning (ML). Malaysian Journal of Information and Communication Technology, 8(2): 158-168. https://doi.org/10.53840/myjict8-2-102

[25] Zhang, L. (2025). Features extraction based on Naive

Bayes algorithm and TF-IDF for news classification. PLoS One, 20(7): e0327347. https://doi.org/10.1371/journal.pone.0327347
[26] Koto, F., Rahimi, A., Lau, J.H., Baldwin, T. (2020). IndoLEM and IndoBERT: A Benchmark dataset and pre-trained language model for Indonesian NLP. In 28th International Conference on Computational Linguistics, Barcelona, Spain, pp. 757-770. https://doi.org/10.18653/v1/2020.coling-main.66
[27] Syahputra, M.E., Kemala, A.P., Ramdhan, D. (2023). Clickbait detection in Indonesia headline news using IndoBERT and RoBERTa. Jurnal Riset Informatika, 5(3): 425-430. https://doi.org/10.34288/jri.v5i4.237
[28] Marutho, D., Utomo, V.G. (2025). Benchmarking IndoBERT and transformer models for sentiment classification on Indonesian e-government service reviews. Jurnal Transformasi, 23(1): 86-95. https://doi.org/10.26623/transformatika.v23i1.12095
[29] Kamdan, K., Anugrah, M.P., Almutaali, M.J., Ramdani, R., Kharisma, I.L. (2025). Performance analysis of IndoBERT for detection of online gambling promotion in YouTube comments. Engineering Proceedings, 107(1): 66. https://doi.org/10.3390/engproc2025107066
[30] Chen, Y., Li, V.O.K., Cho, K., Bowman, S. (2018). A stable and effective learning strategy for trainable greedy decoding. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, pp. 380-390. https://doi.org/10.18653/v1/d18-1035
[31] Park, C., Yang, Y., Park, K., Lim, H. (2020). Decoding strategies for improving low-resource machine translation. Electronics, 9(10): 1562. https://doi.org/10.3390/electronics9101562
[32] Wiher, G., Meister, C., Cotterell, R. (2022). On decoding strategies for neural text generators. Transactions of the Association for Computational Linguistics, 10: 997-1012. https://doi.org/10.1162/tacl_a_00502
[33] Lee, S., Lee, J., Moon, H., Park, C., Seo, J., Eo, S., Koo, S., Lim, H. (2023). A survey on evaluation metrics for machine translation. Mathematics, 11(4): 1006. https://doi.org/10.3390/math11041006
[34] Islam, M.A., Mukta, M.S.H. (2022). A comprehensive understanding of popular machine translation evaluation metrics. International Journal of Computer Science and Engineering, 25(5): 467-478. https://doi.org/10.1504/ijcse.2022.126258
[35] Shi, X.Y., Yue, P., Liu, X.Y., Xu, C., Xu, L. (2022). Obtaining parallel sentences in low-resource language pairs with minimal supervision. Computational Intelligence and Neuroscience, 2022: 5296946. https://doi.org/10.1155/2022/5296946
[36] Mao, Z.Y., Dabre, R., Liu, Q.Y., Song, H.Y., Chu, C.H., Kurohashi, S. (2023). Exploring the impact of layer normalization for zero-shot neural machine translation. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Toronto, Canada, pp. 1300-1316. https://doi.org/10.18653/v1/2023.acl-short.112
[37] Inácio, A.d.S., Lopes, H.S. (2023). Evaluation metrics for video captioning: A survey. Machine Learning with Applications, 13: 100488. https://doi.org/10.1016/j.mlwa.2023.100488
[38] ElNokrashy, M., Kocmi, T. (2023). eBLEU: Unexpectedly good machine translation evaluation using simple word embeddings. In Proceedings of the Eighth Conference on Machine Translation, Singapore, pp. 746-750. https://doi.org/10.18653/v1/2023.wmt-1.61
[39] Tan, W., Zhu, K. (2024). NusaMT-7B: Machine translation for low-resource Indonesian languages with large language models. arXiv preprint arXiv:2410.07830. https://doi.org/10.48550/arXiv.2410.07830
[40] Roy, A., Ray, P., Maheshwari, A., Sarkar, S., Goyal, P. (2024). Enhancing low-resource NMT with a multilingual encoder and knowledge distillation: A case study. In Proceedings of the Seventh Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2024), Bangkok, Thailand, pp. 64-73. https://doi.org/10.18653/v1/2024.loresmt-1.7
[41] Agrawal, G., Das, R., Biswas, A., Thounaojam, D.M. (2023). Neural machine translation for English - Manipuri and English – Assamese. In Proceedings of the Eighth Conference on Machine Translation, Singapore, pp. 931-934. https://doi.org/10.18653/v1/2023.wmt-1.86
[42] Singh, K.B., Singh, N.A., Meetei, L.S., Bandyopadhyay, S., Singh, T.D. (2023). NITS-CNLP low-resource neural machine translation systems of English-Manipuri language pair. In Proceedings of the Eighth Conference on Machine Translation, Singapore, pp. 967-971. https://doi.org/10.18653/v1/2023.wmt-1.92
[43] Yang, Z., Fang, Q.K., Feng, Y. (2022). Low-resource neural machine translation with cross-modal alignment. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, Abu Dhabi, United Arab Emirates, pp. 10134-10146. https://doi.org/10.18653/v1/2022.emnlp-main.689
[44] Goyal, V., Kumar, S., Sharma, D.M. (2020). Efficient neural machine translation for low-resource languages via exploiting related languages. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop, Online, pp. 162-168. https://doi.org/10.18653/v1/2020.acl-srw.22
[45] Sun, M.T., Wang, H., Pasquine, M., Hameed, I.A. (2021). Machine translation in low-resource languages by an adversarial neural network. Applied Sciences, 11(22): 10860. https://doi.org/10.3390/app112210860
[46] Muhajir, M., Rosadi, D. (2024). Mathematical modelling of engineering problems integrating decision tree and BIRCH clustering algorithms of BERTopic for analyzing public sentiment on Dirtyvote movie. Mathematical Modelling of Engineering Problems, 11(12): 3391-3401. https://doi.org/10.18280/mmep.111217