A Multimodal Transformer-LSTM Framework for Cross-Lingual Lexical Alignment of Indonesian Regional Languages

Ema Utami
Faculty of Computer Science, Universitas AMIKOM Yogyakarta, Indonesia
ema.u@amikom.ac.id (corresponding author)

Sri Ngudi Wahyuni
Faculty of Computer Science, Universitas AMIKOM Yogyakarta, Indonesia
yuni@amikom.ac.id

Mulia Sulistiyono
Faculty of Computer Science, Universitas AMIKOM Yogyakarta, Indonesia
muliasulistiyono@amikom.ac.id

Suwanto Raharjo
Faculty of Information Technology and Business, Universitas AKPRIND, Indonesia
wa2n@akprind.ac.id

Anggit Dwi Hartanto
Faculty of Computer Science, Universitas AMIKOM Yogyakarta, Indonesia
anggit@amikom.ac.id

Arif Nur Rohman
Faculty of Computer Science, Universitas AMIKOM Yogyakarta, Indonesia
arif.nurrohman@amikom.ac.id

Bambang Krismono Triwijoyo
Faculty of Engineering, Universitas Bumigora, Indonesia
bkrismono@universitasbumigora.ac.id

Titik Ceriyani Miswaty
Faculty of Humanities, Law and Tourism, Universitas Bumigora, Indonesia
titikceriyani@universitasbumigora.ac.id

Elyakim Nova Supriyedi Patty
Faculty of Education, Universitas Bumigora, Indonesia
elyakim@universitasbumigora.ac.id

Fahry
Faculty of Engineering, Universitas Bumigora, Indonesia
fahry@universitasbumigora.ac.id

Received: 21 December 2025 | Revised: 7 February 2026 and 28 February 2026 | Accepted: 2 March 2026

 DOI: https://doi.org/10.48084/etasr.17086

ABSTRACT

Research on cross-linguistic lexical alignment in Indonesian regional languages is still limited, especially in the use of multimodal data and the systematic evaluation of multilingual Transformer models. Most previous studies have relied on single-text data and semantic similarity-based approaches without providing adequate empirical evidence regarding the performance of literal lexical alignment derived from multimodal transcription. These limitations indicate the urgency for research to examine multimodal approaches in the task of cross-linguistic lexical alignment in Indonesian regional languages. This study contributes through the development and evaluation of a multimodal framework based on a Transformer and Long Short-Term Memory (LSTM) architecture for lexical alignment across languages. The novelty of this research lies in the utilization of primary field data from native speakers curated into a multimodal corpus of Indonesian regional languages, as well as in the comparative evaluation of a multilingual Transformer model for literal lexical alignment based on audio and visual transcription. Audio and visual data were transcribed using Automatic Speech Recognition (ASR) and Optical Character Recognition (OCR) to produce cross-modality text representations. The representations were processed using the mBERT, DistilBERT, XLM-R Large, and LaBSE models, with the support of sequential modeling using LSTM. This dataset consists of 6,000 parallel multimodal utterances. Experimental results showed that the multilingual Transformer model was effective in performing cross-linguistic lexical alignment, with LaBSE producing the most consistent performance in terms of Accuracy, Mean Reciprocal Rank (MRR), and BLEU metrics. These findings make a significant empirical contribution to the study of multimodal-based cross-language lexical alignment for Indonesian regional languages.

Keywords-SaSaMbo; multimodal; transformer; LSTM; ASR; OCR

I. INTRODUCTION

Low-resource languages remain a major challenge in Natural Language Processing (NLP) due to limited parallel corpora, inconsistent linguistic standardization, and high morphological variation, which collectively restrict the performance of multilingual and cross-lingual models in tasks such as lexical alignment and translation [1-4]. Recent studies also highlight the importance of scalable corpus infrastructures and database-driven linguistic resources to support multilingual annotation and AI-ready language datasets for endangered languages [5]. Studies in under-resourced and indigenous languages consistently show that the lack of aligned lexical resources leads to weak cross-lingual mappings and unstable literal correspondences, particularly when regional languages are excluded from large-scale multilingual datasets [6]. Within the Indonesian linguistic landscape, regional languages such as Sasak, Samawa, and Mbojo (SaSaMbo) are actively used in daily communication but remain minimally represented in computational corpora, making them a relevant case for research on low-resource cross-lingual lexical alignment [7]. Previous work has demonstrated that surface-level lexical variation, phonological divergence, and limited orthographic consistency significantly complicate alignment and translation tasks for such languages.

Recent advances in Transformer-based architecture have substantially improved multilingual representation learning by leveraging self-attention mechanisms to construct shared semantic embedding spaces across languages, allowing knowledge transfer even in low-resource scenarios [8, 9]. Survey and empirical studies report that multilingual Transformer models outperform traditional approaches in cross-lingual and multimodal language processing tasks, including literal lexical matching and alignment [10, 11]. Recent studies have also explored hybrid Transformer-based architectures to enhance contextual representation and modeling performance across diverse NLP tasks [12]. In addition, Transformer-based models have shown strong adaptability in various NLP applications due to scalable pretraining and contextual representation learning [13]. However, linguistic data derived from spoken and multimodal sources introduce temporal and sequential dependencies that may not be fully captured by Transformers alone, motivating the use of hybrid architectures that integrate LSTM networks for temporal modelling [14, 15].

In low-resource settings, multimodal data acquisition using Automatic Speech Recognition (ASR) and Optical Character Recognition (OCR) further supports corpus construction by converting audio and visual inputs into aligned textual representations, thereby improving data availability and robustness [16, 17]. Prior studies indicate that multimodal and hybrid learning frameworks can enhance alignment stability and performance in multilingual NLP systems [18]. Based on these considerations, this study evaluates Transformer-based cross-lingual lexical alignment for the SaSaMbo languages using a multimodal corpus and literal evaluation metrics, providing empirical insights into the effectiveness of multilingual Transformer models in low-resource regional language scenarios.

The novelty of this study lies in the empirical and methodological novelty in the context of Indonesian regional language processing with the application and evaluation of a multimodal approach and temporal modeling for lexical alignment across regional languages, which was very limitedly discussed in previous research. Unlike previous studies that generally used single-text data and semantic similarity-based evaluations, this study utilized a primary multimodal corpus collected directly from native speakers and focused on literal lexical alignment. Research contributions include the evaluation of four Transformer models for cross-linguistic lexical alignment, the analysis of the influence of ASR-OCR-based multimodal transcription and LSTM sequential modeling, as well as the provision of literal evaluation benchmarks based on Accuracy, MRR, and BLEU.

II. MATERIALS AND METHODS

This study employed an experimental approach based on multimodal NLP to evaluate cross-linguistic lexical alignment in Indonesian regional languages. The method was designed as a sequential processing flow that includes primary data collection, multimodal transcription, text representation formation, Transformer- and LSTM-based modeling, and quantitative evaluation using lexical alignment metrics. All stages are designed to ensure the consistency of the experiment and the repeatability of the results. Figure 1 illustrates the overall AI-oriented research framework, highlighting the transformation pipeline from multimodal data acquisition to feature extraction, hybrid modelling, and final inference outputs. The overall framework consists of multimodal data acquisition, automatic transcription using ASR and OCR, contextual and temporal representation learning through a hybrid Transformer-LSTM model, multimodal feature fusion, and multi-task classification. This pipeline directly addresses key challenges in low-resource NLP, including data sparsity, cross-modal misalignment, and limited linguistic standardization [19-22].

A. Multimodal Corpus and Data Acquisition

The SaSaMbo corpus was used in this study as an experimental data source to evaluate the performance of the model on cross-lingual lexical alignment tasks. This research does not focus on the construction or development of the corpus, but on the empirical analysis of lexical alignment across languages. The dataset used in this study contains primary data from field collection from native speakers of Sasak, Samawa, and Mbojo, which is then curated, annotated, and compiled into the SaSaMbo corpus. The data collection process involved audio recording, video capture, transcription, and text synchronization, carried out in a controlled manner. From the total field data collected, this study utilized 6,000 parallel multimodal speeches, consisting of 2,000 samples for the Indonesian-Sasak, Indonesian-Samawa, and Indonesian-Mbojo pairs. Each sample represents a single speech synchronized across modalities (audio, visual, and text), allowing for consistent cross-lingual and multimodal analysis.

The SaSaMbo corpus is organized into three language pairs—Indonesian-Sasak, Indonesian-Samawa, and Indonesian-Mbojo—with balanced data for each pair to reduce bias during training and evaluation. Balanced corpus composition is important in multilingual experiments, as uneven data distribution can influence alignment accuracy and model generalization. Table I presents the detailed corpus composition for each language pair. The dataset has been published in [23] and is intended solely for academic research purposes without containing personal information.

TABLE I. MULTIMODAL CORPUS COMPOSITION

Language	Original language	Translated language	Number of samples
Indo-Sasak	1000	1000	2000
Indo-Samawa	1000	1000	2000
Indo-Mbojo	1000	1000	2000
Total	3000	3000	6000

B. Data Validation and Preprocessing

Before modelling, all multimodal data were subjected to validation procedures to ensure data quality and consistency. Audio recordings were inspected for clarity and noise interference, visual data were checked for frame continuity and text visibility, and textual annotations were manually verified to ensure semantic alignment. Ensuring high-quality multimodal inputs is essential, as noisy or misaligned data can severely degrade learning performance [24]. Preprocessing was conducted separately for each modality. Audio signals were normalized and segmented, visual data were converted to representative frames, and textual data underwent normalization processes, such as case folding and noise removal. These preprocessing strategies are commonly adopted in multimodal NLP pipelines to ensure compatibility across heterogeneous representations [25]. This validation and preprocessing stage ensures that the multimodal data used is of adequate quality and consistency before the modeling stage.

C. Automatic Transcription Using ASR and OCR

Automatic transcription was employed to transform raw multimodal inputs into machine-readable linguistic representations that can be processed by neural models. Audio signals were converted to textual form using an ASR system, while textual information appearing in video frames was extracted using an OCR system [26]. Recent studies on low-resource speech processing have shown that Transformer-based ASR architectures can improve transcription quality for regional and dialectal languages, supporting the development of language resources for underrepresented speech communities [27]. These transcription processes enable the conversion of heterogeneous multimodal inputs into a unified textual modality suitable for cross-lingual processing. In this study, the ASR component is applied to audio recordings to automatically generate textual transcriptions of spoken utterances. In parallel, the OCR component is applied to selected video frames to extract visible textual content. The ASR and OCR outputs are treated as complementary textual sources that capture linguistic information from both auditory and visual modalities [28]. When available, manually verified transcriptions are also incorporated to improve the completeness of the textual representation.

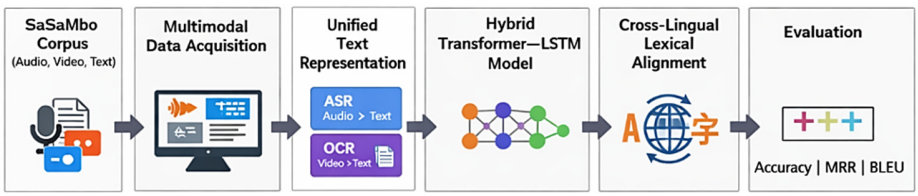

Fig. 1. Research flowchart: Multimodal corpus engine for SaSaMbo languages using a hybrid Transformer-LSTM.

The resulting textual outputs from ASR, OCR, and manual transcription are combined to form a unified textual representation, which serves as the primary linguistic input to the subsequent stages of the proposed framework. Importantly, both ASR and OCR models are used solely as transcription tools and are not retrained or fine-tuned during the lexical alignment experiments. This design choice ensures that the focus of the study remains on evaluating the effectiveness of multilingual Transformer models for cross-lingual lexical alignment, rather than optimizing the transcription components themselves [29].

D. Unified Text Representation

The transcription results from ASR and OCR are combined to form a unified text representation that serves as the main input for cross-language processing models. The unified textual data are transformed into dense contextual embeddings using a Transformer encoder. The Transformer employs self-attention mechanisms to model global dependencies within the input sequence [30, 31]. Given an input embedding matrix $X \in \mathbb{R}^{n\,x\,d}$, the scaled dot-product attention is defined as:

$$Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \qquad (1)$$

where $Q = XW_q$, $K = XW_K$, and $V = XW_V$ denote the query, key, and value matrices, respectively, and d_k represents the dimensionality of the key vectors. This mechanism enables the model to capture semantic, syntactic, and pragmatic linguistic cues relevant to downstream inference tasks.

E. Contextual Representation Using Transformer Models

Unified text representations are mapped into contextual embedding using a multilingual Transformer model to capture lexical and semantic information across languages. To comprehensively evaluate contextual representation learning for resource-constrained regional languages, this study combines four Transformer-based pre-trained language models as textual encoders: multilingual BERT (mBERT), DistilBERT, XLM-R Large, and Language-Agnostic BERT Sentence Embedding (LaBSE) [32, 33]. These models were selected to reflect different architectural scales, multilingual learning strategies, and computational considerations.

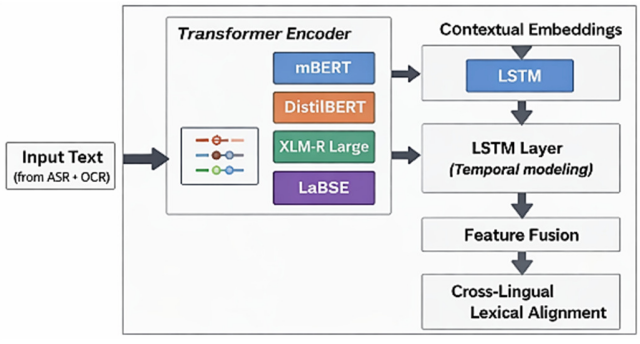

Fig. 2. Integration of Transformer variants (mBERT, DistilBERT, XLM-R Large, and LaBSE) within the hybrid Transformer-LSTM architecture.

All Transformer models are integrated into the same hybrid Transformer-LSTM pipeline to ensure experimental consistency. The contextual embeddings generated by each model serve as the semantic input to the temporal modelling and multimodal fusion stages.

- mBERT is a Transformer-based multilingual model that was used as a baseline encoder to evaluate cross-lingual transfer to low-resource SaSaMbo languages. Although SaSaMbo languages are not explicitly included in its pretraining data, mBERT can capture general contextual semantics across languages. Its main limitation lies in its relatively high computational cost due to the large number of parameters [25].
- DistilBERT is a lightweight Transformer model obtained through knowledge distillation, offering faster inference and lower memory usage. This study employed DistilBERT to assess the feasibility of efficient Transformer encoders for multimodal NLP tasks. However, its reduced model depth may limit its ability to capture fine-grained sociolinguistic and psycholinguistic patterns [34].
- XLM-R Large is a high-capacity multilingual Transformer pretrained on extensive multilingual corpora. It is employed to generate rich semantic and syntactic representations for SaSaMbo languages, making it particularly suitable for low-resource NLP scenarios. This performance advantage is achieved at the expense of higher computational complexity [35].
- LaBSE (Language-Agnostic BERT Sentence Embedding) is designed to generate language-independent sentence embeddings that enable effective cross-lingual semantic alignment. In this study, LaBSE is used to evaluate the effectiveness of sentence-level multilingual embeddings for lexical alignment across Indonesian regional languages. Its architecture is optimized for multilingual representation learning and cross-language retrieval tasks [36].

F. Transformer Models Configuration (mBERT, DistilBERT, XLM-R, LaBSE)

Table II summarizes the Transformer-based language models used as contextual encoders.

TABLE II. TRANSFORMER-BASED LANGUAGE MODELS USED IN THIS STUDY

Model	Main strength	Computational cost	Function in the proposed system
mBERT	Broad multilingual coverage	High	Cross-lingual baseline
DistilBERT	Efficiency and speed	Low	Lightweight alternative
XLM-R Large	Deep semantic representation	Very High	High-capacity encoder
LaBSE	Cross-lingual semantic alignment	Moderate	Best-performing encoder

mBERT is adopted as a multilingual baseline due to its broad language coverage, although it requires relatively high computational resources. DistilBERT provides a more efficient alternative with faster inference and lower memory usage, making it suitable for resource-limited settings [32]. XLM-R Large offers stronger semantic and syntactic representation capabilities for low-resource languages, but at a higher

computational cost. LaBSE focuses on language-agnostic sentence embeddings and cross-lingual semantic alignment, enabling robust generalization across dialectal variations in the SaSaMbo corpus. Overall, this comparison illustrates the trade-off between model capacity and efficiency and justifies the comparative evaluation of Transformer encoders in the proposed hybrid framework. All four Transformer models are used as text encoders in the same configuration to ensure fair performance comparisons on cross-language lexical alignment tasks.

G. Sequential Modeling Using LSTM

In this study, LSTM was used as a component of sequential modeling to capture the temporal dependency of multimodal transcription results. Given an input vector x_t at time step t and the previous hidden state, the LSTM cell is computed as follows.

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \quad (2)$$

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \quad (3)$$

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \quad (4)$$

$$\tilde{c}_t = \tanh(W_c \cdot [h_{t-1}, x_t] + b_c) \quad (5)$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t \quad (6)$$

$$h_t = o_t \odot \tanh(c_t) \quad (7)$$

Here, $\sigma(\cdot)$ denotes the sigmoid activation function, $\tanh(\cdot)$ is the hyperbolic tangent function, and $\odot$ represents element-wise multiplication. This formulation allows the LSTM to preserve long-term dependencies and temporal speech patterns. The mathematical equations presented in this section are used to describe the basic mechanism of self-attention in Transformers and sequential modeling in LSTMs. In its implementation, all operations are realized using a standard deep learning library according to the configuration of the pre-trained model used. Mathematical parameters and operations are not implemented manually, but rather follow the built-in specifications of each architecture.

H. Sequential Integration of Transformer and LSTM

The Transformer model is used as a key component to construct cross-linguistic lexical representations, while LSTM serves to model the temporal dependencies of multimodal transcription outcomes. The process starts with a Transformer encoder that converts unified text from ASR and OCR into contextual embeddings, capturing general semantic information such as word meaning, sentence structure, and discourse context [34]. These embeddings are then sequentially passed to an LSTM layer, which focuses on learning temporal patterns across utterances, including speech rhythm, hesitation, intonation, and discourse flow. By combining the Transformer's strength in modelling long-range semantic dependencies with the LSTM's ability to capture sequential and temporal dynamics, this hybrid architecture overcomes the limitations of single-model approaches and provides a robust representation for complex multimodal NLP tasks.

I. Model Configuration and Reproducibility Details

The ASR model used is a Transformer-based speech recognition system that has been previously trained on large-scale Indonesian data, while the OCR model used is a CNN-based character recognition system that is optimized for Latin text. Both systems are used without retraining or parameter adjustments, so they function purely as transcription tools. The Transformer models are used with the default configurations of their respective pretrained models. The entire model is used as a text encoder without advanced fine-tuning. The LSTM layer is added as a sequential modeler with one hidden layer, a hidden state size of 256, and a tanh activation function. The training process was carried out using the Adam optimizer with a learning rate of 1e-4 and a batch size of 32. The dataset was divided into training, validation, and testing data in a random but controlled 70:15:15 ratio for each language pair. The entire experiment was run on the same hardware configuration to ensure consistency of the results.

J. Lexical Alignment Task Definition

The main task evaluated in this study is cross-linguistic lexical alignment, namely mapping the form of words or phrases in Indonesian regional languages to their lexical equivalents in Indonesian. Evaluation focused on the suitability of literal lexical forms, not on high-level semantic similarities. Semantic features from the Transformer encoder and temporal features from the LSTM layer are combined using feature-level fusion. The fused representation is defined as:

$$F_{fusion} = [F_{semantic}; F_{temporal}] \quad (8)$$

Table III describes the functional roles of each model component in the multimodal feature fusion process. The Transformer encoder is responsible for extracting semantic features that capture contextual linguistic information, including word meaning, syntactic structure, and discourse-level cues. The LSTM layer complements this representation by modelling temporal and sequential patterns, such as speech rhythm, hesitation, and behavioral dynamics across utterances [14]. The fusion layer integrates both semantic and temporal features into a unified representation, enabling the system to jointly exploit linguistic content and temporal behavior for robust multimodal inference in downstream classification tasks. The fused multimodal representation is subsequently evaluated for cross-lingual lexical alignment using standard evaluation metrics.

TABLE III. TRANSFORMER-BASED LANGUAGE MODELS USED IN THIS STUDY

Component	Feature Type
Transformer	Semantic
LSTM	Temporal
Fusion Layer	Combined

K. Experimental Setup and Evaluation Metrics

All experiments were conducted using an NVIDIA L4 GPU with 22.5 GB of GPU memory and 53 GB of system RAM. Model performance was evaluated using Accuracy, Mean Reciprocal Rank (MRR), and BLEU score. These metrics provide complementary perspectives on lexical alignment

accuracy, ranking effectiveness, and n-gram similarity between predicted and reference sequences. Previous studies recommend the use of multiple evaluation metrics to ensure comprehensive performance assessment in cross-lingual and multimodal NLP experiments.

III. EXPERIMENTAL RESULTS

This section presents the experimental results of the proposed multimodal cross-lingual lexical alignment framework applied to the SaSaMbo corpus. Four Transformer-based models—mBERT, DistilBERT, XLM-R Large, and LaBSE—were evaluated and further integrated with an LSTM layer to capture temporal dependencies.

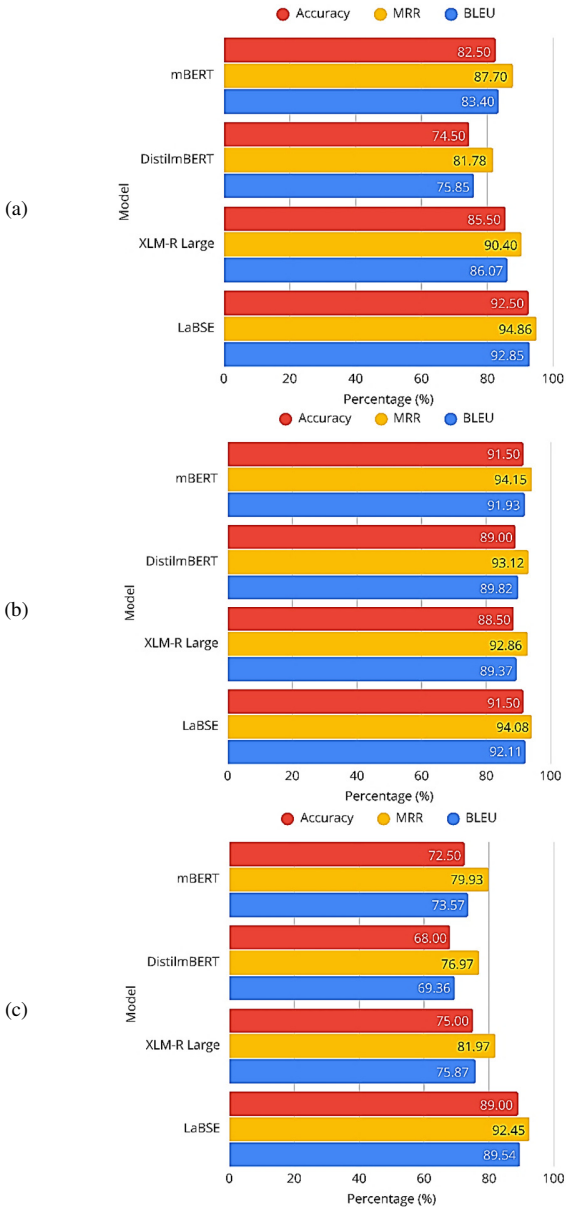

Fig. 3. Model evaluation for: (a) Indo-Sasak dataset, (b) Indo-Samawa dataset, (c) Indo-Mbojo dataset.

Model performance was assessed using Accuracy, Mean Reciprocal Rank (MRR), and BLEU score. Figure 4 illustrates the comparative performance trends of the evaluated models across the three language pairs. To establish a baseline for cross-lingual lexical alignment, the four Transformer-based models were first evaluated without the integration of the LSTM sequential layer. This baseline experiment aims to measure the capability of each Transformer model in capturing cross-lingual lexical representations derived from the multimodal transcription pipeline. Table IV reports the average Accuracy, Mean Reciprocal Rank (MRR), and BLEU scores obtained for the Indonesian-Sasak, Indonesian-Samawa, and Indonesian-Mbojo datasets. Subsequently, these baseline results are compared with the hybrid Transformer-LSTM configuration to analyze the impact of LSTM-based temporal smoothing on lexical alignment performance.

TABLE IV. AVERAGE ACCURACY, MRR, AND BLEU FOR EACH MODEL

Dataset	Model	Accuracy	MRR	BLEU
Indonesian-Mbojo	mBERT	72.50%	79.93%	73.57%
	DistilBERT	68.00%	76.97%	69.63%
	XLM-R Large	75.00%	81.97%	75.87%
	LaBSE	89.00%	92.45%	89.54%
Indonesian-Samawa	mBERT	75.00%	81.97%	75.87%
	DistilBERT	89.00%	92.45%	89.54%
	XLM-R Large	75,00%	87.27%	82.97%
	LaBSE	89.00%	93.12%	89.82%
Indonesian-Sasak	mBERT	82.50%	87.70%	83.40%
	DistilBERT	74.50%	81.78%	75.85%
	XLM-R Large	85.50%	90.40%	86.07%
	LaBSE	92.50%	94.86%	92.85%

LaBSE demonstrated the most consistent performance across the evaluated datasets. In the Indonesian-Mbojo dataset, LaBSE achieved an accuracy of 89.00%, an MRR of 92.45%, and a BLEU score of 89.54%, outperforming mBERT, DistilBERT, and XLM-R Large. A similar trend was observed in the Indonesian-Samawa dataset, where LaBSE obtained an accuracy of 89.00%, an MRR of 93.12%, and a BLEU score of 89.82%. In the Indonesian-Sasak dataset, LaBSE again achieved the highest performance with an accuracy of 92.50%, an MRR of 94.86%, and a BLEU score of 92.85%. This advantage is attributed to LaBSE's ability to generate sentence-level semantic representations, enabling more effective cross-lingual alignment compared to token-based models such as mBERT and XLM-R [37]. These findings are consistent with previous studies on multilingual Transformer models for cross-lingual tasks, which report that models with larger representation capacity, such as XLM-R and LaBSE, generally produce more stable cross-language alignment, particularly in scenarios with limited data availability and high morphological variation [8, 10, 12]. Dual-encoder architectures such as LaBSE are designed to generate language-agnostic sentence embeddings that support consistent alignment across languages and domains. While most prior studies rely on single-text datasets, the results of this study extend these findings to a multimodal setting by incorporating ASR- and OCR-based transcription combined with temporal modelling using LSTM. This indicates that multimodal integration can further strengthen lexical alignment performance in low-resource regional language scenarios.

A smoothing process was carried out in the word alignment process using LSTM, with the results presented in Table V. Models with stronger multilingual representations maintain higher accuracy after the LSTM layer is applied, although the absolute accuracy values are slightly lower than those of the Transformer-only configuration. For example, in the Indonesian-Mbojo dataset, the Top-1 accuracy ranges from 58.5% (mBERT-LSTM) to 84.5% (LaBSE-LSTM). A similar pattern is observed in the Indonesian-Samawa dataset, where accuracy ranges from 81.5% to 90.5%, and in the Indonesian-Sasak dataset from 71.5% to 88.5%. These results indicate that models with stronger cross-lingual sentence embeddings, such as LaBSE, maintain higher alignment accuracy even after introducing temporal modelling.

The LSTM layer offers improved semantic stability, as reflected by consistently high MRR and BERTScore values. However, it does not always improve Top-1 accuracy compared to the Transformer-only configuration (Table IV). This suggests that LSTM primarily acts as a smoothing mechanism that reduces embedding noise but may also attenuate fine-grained lexical distinctions, a behavior commonly associated with the over-smoothing phenomenon in sequential modeling. Overall, the results indicate that cross-linguistic representation capability plays a greater role in lexical alignment performance than additional architectural complexity.

Table VI provides a direct comparison between the Transformer models without LSTM and the Hybrid Transformer-LSTM configuration. The results indicate that although the LSTM layer improves semantic stability, the Top-1 accuracy slightly decreases across all evaluated datasets. In the Indo-Mbojo dataset, the accuracy drops from 89.00% to 84.50%, representing a decrease of 4.50%. In the Indo-Samawa dataset, the decrease is relatively small, from 91.50% to 90.50% (−1.00%), indicating that lexical alignment remains stable when linguistic similarity between languages is high. In the Indo-Sasak dataset, the accuracy decreases from 92.50% to 88.50%, corresponding to a reduction of 4.00%.

TABLE V. AVERAGE ACCURACY, MRR, AND BLEU FOR EACH MODEL AFTER SMOOTING USING LSTM

Dataset	model	n_samples	top1_ correct	topk_ correct	top1_ acc	topk_ acc	mean_ sentence_ bleu	mean_ reciprocal_ rank	mean_ bertscore_ precision	mean_ bertscore_ recall	mean_ bertscore_ f1
Indonesian-Mbojo	mBERT-LSTM	200	117	151	0.585	0.755	0.572	0.692	0.883	0.885	0.884
	DistilmBERT-LSTM	200	112	152	0.560	0.760	0.553	0.680	0.875	0.875	0.875
	XLMR-Base-LSTM	200	128	160	0.640	0.800	0.640	0.739	0.898	0.899	0.899
	XLMR-Large-LSTM	200	136	161	0.680	0.805	0.664	0.758	0.911	0.911	0.911
	LaBSE-LSTM	200	169	188	0.845	0.940	0.827	0.895	0.956	0.957	0.957
Indonesian-Samawa	mBERT-LSTM	200	163	187	0.815	0.935	0.810	0.877	0.944	0.943	0.944
	DistilmBERT-LSTM	200	161	182	0.805	0.910	0.798	0.867	0.940	0.940	0.940
	XLMR-Base-LSTM	200	178	187	0.890	0.935	0.879	0.918	0.968	0.968	0.968
	XLMR-Large-LSTM	200	179	188	0.895	0.940	0.883	0.924	0.968	0.970	0.969
	LaBSE-LSTM	200	181	191	0.905	0.955	0.893	0.932	0.973	0.973	0.973
Indonesian-Sasak	mBERT-LSTM	200	143	169	0.715	0.845	0.699	0.790	0.917	0.915	0.916
	DistilmBERT-LSTM	200	147	168	0.735	0.840	0.723	0.796	0.924	0.921	0.922
	XLMR-Base-LSTM	200	156	176	0.780	0.880	0.764	0.840	0.933	0.935	0.934
	XLMR-Large-LSTM	200	170	183	0.850	0.915	0.831	0.894	0.954	0.954	0.954
	LaBSE-LSTM	200	177	190	0.885	0.950	0.867	0.924	0.967	0.966	0.967

TABLE VI. DIRECT COMPARISON OF RESULTS WITH AND WITHOUT LSTM SMOOTHING

Dataset	Model no LSTM Smoothing				Model with LSTM Smoothing				
	Model	Accuracy	MRR	BLEU	Model	Top-1 Acc	MRR	BLEU	Accuracy
Indo-Mbojo	LaBSE	89.00%	92.45%	89.54%	LaBSE-LSTM	84.50%	89.53%	82.70%	−4.50%
Indo-Samawa	LaBSE	91.50%	94.08%	92.50%	LaBSE-LSTM	90.50%	93.25%	89.25%	−1.00%
Indo-Sasak	LaBSE	92.50%	94.86%	92.85%	LaBSE-LSTM	88.50%	92.35%	86.70%	−4.00%

TABLE VII. COMPARISON OF MULTILINGUAL TRANSFORMER MODELS IN LOW-RESOURCE CROSS-LINGUAL SETTINGS

Study	Language pairs	Resource level	Task	Model	Metric	Reported result
[38]	Czech-English, Czech-French	Medium-Low	Cross-lingual sentiment analysis	XLM-R	Accuracy	96.0% (Czech), 97.6% (French)
[33]	English-Sinhala, Tamil, Bengali	Low	Sentiment classification	XLM-R-base	Macro F1	71.45% (Sinhala), 64.67% (Tamil), 43.26% (Bengali)
[39]	Roman Urdu-English	Low	Code-mixed sentiment prediction	mBART	Accuracy	72%
[40]	English-Swahili, English-Amharic	Low	Machine translation	Transformer (multi-task)	BLEU	32.5 (Swahili), 30.1 (Amharic)
[41]	Swahili-English	Low	Multimodal language generation	RAG (dual-encoder+T5)	BLEU	28.4
[42]	English-Turkish (Medical)	Low	NLI translation & inference	BERTurk	Accuracy	75.94%
This study	**Indo-Mbojo, Indo-Samawa, Indo-Sasak**	**Low**	**Cross-lingual lexical alignment**	**LaBSE**	**Accuracy**	**92.50% (Indonesian-Sasak)**

To contextualize the performance of the proposed framework, Table VII presents a comparative evaluation of multilingual Transformer models applied in low-resource cross-lingual settings, including sentiment analysis, translation, and alignment tasks. Despite the variations in datasets, linguistic distance, and task formulation limiting direct numerical equivalence, this structured comparison offers meaningful insight into relative model behavior and performance trends across comparable multilingual scenarios.

The comparative analysis presented in Table VII situates the proposed framework within the broader landscape of multilingual Transformer-based models applied to low-resource cross-lingual tasks. Previous studies in sentiment classification and translation tasks typically report performance ranging between 43% and 97%, depending on dataset scale, linguistic distance, and task formulation. For example, XLM-R achieves 96.0–97.6% accuracy in Czech–French settings, which involve relatively structured European languages with richer digital corpora [38]. In contrast, performance drops significantly in truly low-resource languages such as Sinhala and Bengali, where macro-F1 scores remain below 72% [33].

In translation-oriented low-resource settings, the BLEU scores reported for Transformer-based systems generally range between 28 and 32, particularly for African language pairs such as Swahili and Amharic. These values reflect the inherent difficulty of low-resource neural machine translation tasks, especially when multimodal augmentation is not extensively applied. Similarly, multimodal retrieval-augmented systems report BLEU scores of around 28.4 [41], indicating moderate generation quality under data-constrained scenarios.

Compared to these studies, this work achieved 92.50% accuracy on the Indo-Sasak lexical alignment task using LaBSE without LSTM smoothing. Although direct numerical comparison should be interpreted cautiously due to differences in task formulation (classification vs. translation vs. lexical alignment), the results demonstrate that sentence-level multilingual embeddings remain highly effective for structurally related regional languages. The high accuracy observed in this study may be partially attributed to linguistic proximity between Indonesian and SaSaMbo languages, as well as balanced corpus construction.

Importantly, unlike most previous studies that rely solely on text-based corpora, the proposed framework integrates multimodal transcription derived from the ASR and OCR pipelines. This design potentially enhances lexical stability in the presence of phonological and orthographic variation. Therefore, the findings not only align with existing literature that shows the strength of multilingual Transformer models but also extend previous results by empirically validating multimodal lexical alignment in underrepresented Indonesian regional languages.

IV. CONCLUSIONS

This study presented a multimodal framework for the cross-lingual lexical alignment of Indonesian regional languages using a hybrid Transformer-LSTM architecture applied to the SaSaMbo corpus. Audio and visual data were transcribed using ASR and OCR and processed using four multilingual Transformer models: mBERT, DistilBERT, XLM-R Large, and LaBSE. Experimental results show that multilingual Transformer models are effective for lexical alignment in low-resource regional languages. Among the evaluated models, LaBSE demonstrated the most consistent performance, achieving an average accuracy of about 90%, with an MRR of approximately 93.5% and a BLEU score around 90.7%. Dataset-level analysis indicates that the Indo-Samawa pair achieved the most stable results due to higher lexical similarity, whereas Indo-Mbojo showed slightly lower performance because of greater morphological variation.

Compared with previous studies that mainly rely on single-text datasets and semantic similarity-based evaluation, this work demonstrates the potential of combining multimodal transcription and multilingual sentence embeddings for lexical alignment in low-resource languages. The results highlight the effectiveness of language-agnostic embeddings such as LaBSE for structurally related regional languages and emphasize the importance of considering linguistic surface characteristics in multilingual NLP systems. Future research may extend this work by incorporating semantic-aware evaluation metrics, expanding the corpus with additional dialectal variations and spontaneous speech, and exploring richer multimodal cues, such as prosody and visual gestures, to further enhance multilingual language resources and preservation efforts.

ACKNOWLEDGMENT

The authors acknowledge the valuable funding provided by Direktorat Penelitian dan Pengabdian kepada Masyarakat, Direktorat Jenderal Riset dan Pengembangan, Kementerian Pendidikan Tinggi, Sains, dan Teknologi in Hibah Riset Konsorsium Unggulan Berdampak (RIKUB) for the fiscal year 2025, with Master Contract Number: 031/C3/DT.05.00/ RIKUB/2025 and Master Contract Date 10 September 2025.

REFERENCES

[1] R. Pramana, M. Jonathan, H. S. Yani, and R. Sutoyo, "A Comparison of BiLSTM, BERT, and Ensemble Method for Emotion Recognition on Indonesian Product Reviews," *Procedia Computer Science*, vol. 245, pp. 399–408, Jan. 2024, https://doi.org/10.1016/j.procs.2024.10.266.

[2] M. Lupaşcu, A. C. Rogoz, M. S. Stupariu, and R. T. Ionescu, "Large multimodal models for low-resource languages: A survey," *Information Fusion*, vol. 131, July 2026, Art. no. 104189, https://doi.org/10.1016/j.inffus.2026.104189.

[3] P. Pakray, A. Gelbukh, and S. Bandyopadhyay, "Natural language processing applications for low-resource languages," *Natural Language Processing*, vol. 31, no. 2, pp. 183–197, Mar. 2025, https://doi.org/10.1017/nlp.2024.33.

[4] D. D. Baishya, D. R. Baruah, D. M. Bora, and B. Sarma, "Processing Low-Resource Languages: A Review Of Challenges And Strategies For Inclusive NLP And Sustainable Environment," *International Journal of Environmental Sciences*, pp. 7730–7739, Sept. 2025, https://doi.org/10.64252/w55rwj24.

[5] S. Raharjo, E. Utami, E. Sutanta, and N. R. Az–Zahra Raharema, "Korpus.id: A Database-Driven Approach to Linguistic Annotation and Analysis," in *2025 Eighth International Women in Data Science Conference at Prince Sultan University (WiDS PSU)*, Apr. 2025, pp. 151–156, https://doi.org/10.1109/WiDS-PSU64963.2025.00040.

[6] T. Dalai, A. Das, T. K. Mishra, and P. K. Sa, "OdNER: NER resource creation and system development for low-resource Odia language," *Natural Language Processing Journal*, vol. 11, June 2025, Art. no. 100139, https://doi.org/10.1016/j.nlp.2025.100139.

[7] A. S. Ekakristi, A. F. Wicaksono, and R. Mahendra, "Intermediate-task transfer learning for Indonesian NLP tasks," *Natural Language Processing Journal*, vol. 12, Sept. 2025, Art. no. 100161, https://doi.org/10.1016/j.nlp.2025.100161.

[8] A. Maesya, Y. Arifin, A. Zahra, and W. Budiharto, "AMSunda: A novel dataset for Sundanese information retrieval," *Data in Brief*, vol. 61, Aug. 2025, Art. no. 111796, https://doi.org/10.1016/j.dib.2025.111796.

[9] Z. Zainuddin, Mudassir, and Z. Tahir, "Entity Extraction in Indonesian Online News Using Named Entity Recognition (NER) with Hybrid Method Transformer, Word2Vec, Attention and Bi-LSTM," *JOIV: International Journal on Informatics Visualization*, vol. 9, no. 3, pp. 964–973, May 2025, https://doi.org/10.62527/joiv.9.3.2902.

[10] A. Banerjee and D. Banik, "A Comprehensive Survey on Transformer-Based Machine Translation: Identifying Research Gaps and Solutions for Large Language Models," *ACM Computing Surveys*, vol. 58, no. 5, Art. no. 124, Sept. 2025, https://doi.org/10.1145/3773076.

[11] L. Qin *et al.*, "A survey of multilingual large language models," *Patterns*, vol. 6, no. 1, Jan. 2025, https://doi.org/10.1016/j.patter.2024.101118.

[12] H. Boutouta, A. Lakhfif, F. Senator, and C. Mediani, "A Transformer-based Hybrid Model for Implicit Emotion Recognition in Arabic Text," *Engineering, Technology & Applied Science Research*, vol. 15, no. 3, pp. 23834–23839, June 2025, https://doi.org/10.48084/etasr.10261.

[13] H. Zhang and M. O. Shafiq, "Survey of transformers and towards ensemble learning using transformers for natural language processing," *Journal of Big Data*, vol. 11, no. 1, Feb. 2024, Art. no. 25, https://doi.org/10.1186/s40537-023-00842-0.

[14] Z. Chen *et al.*, "Evolution and Prospects of Foundation Models: From Large Language Models to Large Multimodal Models," *Computers, Materials & Continua*, vol. 80, no. 2, pp. 1753–1808, 2024, https://doi.org/10.32604/cmc.2024.052618.

[15] I. D. Mienye, T. G. Swart, and G. Obaido, "Recurrent Neural Networks: A Comprehensive Review of Architectures, Variants, and Applications," *Information*, vol. 15, no. 9, Aug. 2024, https://doi.org/10.3390/info15090517.

[16] N. Khan *et al.*, "Systematic Literature Review of Machine Learning Models and Applications for Text Recognition," *IEEE Access*, vol. 13, pp. 177647–177670, 2025, https://doi.org/10.1109/ACCESS.2025.3618109.

[17] S. Bhushan, V. Prakash Mishra, V. Rishiwal, S. Arunkumar, and U. Agarwal, "Advancing Text-to-Speech Systems for Low-Resource Languages: Challenges, Innovations, and Future Directions," *IEEE Access*, vol. 13, pp. 155729–155758, 2025, https://doi.org/10.1109/ACCESS.2025.3605236.

[18] X. Dong, "Computer-Assisted Multimodal Translanguaging Analysis in English Classrooms: A Deep-Learning and NLP Framework," *Informatica*, vol. 49, no. 37, Dec. 2025, https://doi.org/10.31449/inf.v49i37.10365.

[19] T. Seng, "Analysis of multimodal data recorded during videoconferences," Ph.D. dissertation, Université de Toulouse, France, 2025.

[20] T. Baltrusaitis, C. Ahuja, and L.-P. Morency, "Multimodal Machine Learning: A Survey and Taxonomy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 2, pp. 423–443, Feb. 2019, https://doi.org/10.1109/TPAMI.2018.2798607.

[21] Y. Hao and B. Zhou, "Multi-modal smart classroom topic segmentation," in *Proceedings of the 2025 5th International Conference on Applied Mathematics, Modelling and Intelligent Computing*, Mar. 2025, pp. 262–268, https://doi.org/10.1145/3745533.3745577.

[22] E. Marevac, E. Kadušić, N. Živić, N. Buzađija, E. Tabak, and S. Velić, "Multimodal Video Summarization Using Machine Learning: A Comprehensive Benchmark of Feature Selection and Classifier Performance," *Algorithms*, vol. 18, no. 9, Sept. 2025, Art. no. 572, https://doi.org/10.3390/a18090572.

[23] S. Wahyuni and E. Utami, "SaSamBo Corpus Dataset." Mendeley Data, Feb. 10, 2026, https://doi.org/10.17632/F2JP385ZKR.1.

[24] M. I. Ragab, E. H. Mohamed, and W. Medhat, "Multilingual Propaganda Detection: Exploring Transformer-Based Models mBERT, XLM-RoBERTa, and mT5," in *Proceedings of the first International Workshop on Nakba Narratives as Language Resources*, Jan. 2025.

[25] H. Wei *et al.*, "General OCR Theory: Towards OCR-2.0 via a Unified End-to-end Model." arXiv, Sept. 03, 2024, https://doi.org/10.48550/arXiv.2409.01704.

[26] P. T. Krishnan, A. N. Joseph Raj, and V. Rajangam, "Emotion classification from speech signal based on empirical mode decomposition and non-linear features: Speech emotion recognition," *Complex & Intelligent Systems*, vol. 7, no. 4, pp. 1919–1934, Aug. 2021, https://doi.org/10.1007/s40747-021-00295-z.

[27] S. N. Endah, . Suprapto, and Y. Suyanto, "Enhancing Low-Resource Dialectal ASR in Indonesian Using Speech-Transformer Models and Data Augmentation," *Engineering, Technology & Applied Science Research*, vol. 15, no. 5, pp. 28095–28101, Oct. 2025, https://doi.org/10.48084/etasr.12734.

[28] P. T. Krishnan, A. N. Joseph Raj, and V. Rajangam, "Emotion classification from speech signal based on empirical mode decomposition and non-linear features: Speech emotion recognition," *Complex & Intelligent Systems*, vol. 7, no. 4, pp. 1919–1934, Aug. 2021, https://doi.org/10.1007/s40747-021-00295-z.

[29] H. Kashid and P. Bhattacharyya, "RoundTripOCR: A Data Generation Technique for Enhancing Post-OCR Error Correction in Low-Resource Devanagari Languages," in *Proceedings of the 21st International Conference on Natural Language Processing (ICON)*, Sept. 2024, pp. 274–284.

[30] A. P. Bhopale and A. Tiwari, "Transformer based contextual text representation framework for intelligent information retrieval," *Expert Systems with Applications*, vol. 238, Mar. 2024, Art. no. 121629, https://doi.org/10.1016/j.eswa.2023.121629.

[31] S. Islam *et al.*, "A comprehensive survey on applications of transformers for deep learning tasks," *Expert Systems with Applications*, vol. 241, May 2024, Art. no. 122666, https://doi.org/10.1016/j.eswa.2023.122666.

[32] Y. Ma, "Cross-language Text Generation Using mBERT and XLM-R: English-Chinese Translation Task," in *Proceedings of the 2024 International Conference on Machine Intelligence and Digital Applications*, May 2024, pp. 602–608, https://doi.org/10.1145/3662739.3672320.

[33] V. Dhananjaya, S. Ranathunga, and S. Jayasena, "Lexicon-based fine-tuning of multilingual language models for low-resource language sentiment analysis," *CAAI Transactions on Intelligence Technology*, vol. 9, no. 5, pp. 1116–1125, 2024, https://doi.org/10.1049/cit2.12333.

[34] B. Li, "A Study of DistilBERT-Based Answer Extraction Machine Reading Comprehension Algorithm," in *Proceedings of the 2024 3rd International Conference on Cyber Security, Artificial Intelligence and Digital Economy*, Mar. 2024, pp. 261–268, https://doi.org/10.1145/3672919.3672968.

[35] K. R. Mabokela, T. Celik, and M. Raborife, "Multilingual Sentiment Analysis for Under-Resourced Languages: A Systematic Review of the Landscape," *IEEE Access*, vol. 11, pp. 15996–16020, 2023, https://doi.org/10.1109/ACCESS.2022.3224136.

[36] F. Feng, Y. Yang, D. Cer, N. Arivazhagan, and W. Wang, "Language-agnostic BERT Sentence Embedding," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2022, pp. 878–891, https://doi.org/10.18653/v1/2022.acl-long.62.

[37] M. I. Salih, S. M. Mohammed, A. K. Ibrahim, O. M. Ahmed, and L. M. Haji, "Fine-Tuning BERT for Automated News Classification," *Engineering, Technology & Applied Science Research*, vol. 15, no. 3, pp. 22953–22959, June 2025, https://doi.org/10.48084/etasr.10625.

[38] P. Přibáň, J. Šmíd, J. Steinberger, and A. Mištera, "A comparative study of cross-lingual sentiment analysis," *Expert Systems with Applications*, vol. 247, Aug. 2024, Art. no. 123247, https://doi.org/10.1016/j.eswa.2024.123247.

[39] E. Hashmi, S. Y. Yayilgan, and S. Shaikh, "Augmenting sentiment prediction capabilities for code-mixed tweets with multilingual transformers," *Social Network Analysis and Mining*, vol. 14, no. 1, Apr. 2024, Art. no. 86, https://doi.org/10.1007/s13278-024-01245-6.

[40] C. Kaoutar, L. Yasser, and S. Maha, "Machine Translation with Neural Networks Based on a Transformer," *International Journal For Multidisciplinary Research*, vol. 6, no. 5, Sept. 2024, Art. no. 26674, https://doi.org/10.36948/ijfmr.2024.v06i05.26674.

[41] Z. Li and Z. Ke, "Cross-Modal Augmentation for Low-Resource Language Understanding and Generation," in *Proceedings of the 1st Workshop on Multimodal Augmented Generation via Multimodal Retrieval (MAGMaR 2025)*, 2025, pp. 90–99, https://doi.org/10.18653/v1/2025.magmar-1.9.

[42] İ. Ü. Oğul, F. Soygazi, and B. E. Bostanoğlu, "TurkMedNLI: a Turkish medical natural language inference dataset through large language model based translation," *PeerJ Computer Science*, vol. 11, Jan. 2025, Art. no. e2662, https://doi.org/10.7717/peerj-cs.2662.