International Journal Software Engineering and Computer Science (IJSECS)
5 (2), 2025, 519-528
Published Online August 2025 in IJSECS (http://www.journal.lembagakita.org/index.php/ijsecs)
P-ISSN: 2776-4869, E-ISSN: 2776-3242. DOI: https://doi.org/10.35870/ijsecs.v5i2.4105.

RESEARCH ARTICLE Open Access

Palembang to Indonesian Language Translation Machine Using the No Language Left Behind Approach

Muhammad Finaldo *
Intelligent Systems Research Group (ISRG), Faculty of Science Technology, Universitas Bina Darma, Palembang City, South Sumatra Province, Indonesia.
Corresponding Email: muhammadfinaldo7@gmail.com.

Yesi Novaria Kunang
Intelligent Systems Research Group (ISRG), Faculty of Science Technology, Universitas Bina Darma, Palembang City, South Sumatra Province, Indonesia.
Email: yesinovariakunang@binadarma.ac.id.

Received: April 24, 2025; Accepted: June 20, 2025; Published: August 1, 2025.

Abstract: The Palembang language, deeply rooted in the cultural fabric of South Sumatra, continues to serve as a vital means of daily communication for many communities. As globalization accelerates, safeguarding such regional languages has become increasingly urgent, particularly through technological solutions that can bridge communication between local speakers and visitors. This study introduces an automatic translation system designed to convert Palembang text into Indonesian, employing the No Language Left Behind (NLLB) algorithm—a recent development in artificial intelligence for language processing. A dataset containing 7,917 pairs of Palembang and Indonesian sentences was assembled for this purpose. The translation models were trained and assessed using BLEU (Bilingual Evaluation Understudy) and chrF (Character n-gram F-score) metrics. The initial model achieved a BLEU score of 22.55 and a chrF++ score of 43.22. Subsequent improvements raised these scores to 30.72 and 55.39, respectively, reflecting a significant enhancement in translation quality and clarity for Indonesian readers. By focusing on a language with limited digital resources, this research demonstrates the potential of modern translation technologies to support both linguistic preservation and practical communication needs in diverse cultural settings.

Keywords: Palembang Language; No Language Left Behind; Machine Translation.

1. Introduction

Language serves as a primary means of communication, enabling individuals to exchange information and articulate thoughts [1]. Beyond this practical function, language provides a medium through which communities convey ideas and foster social interaction. Its development is deeply rooted in collective experience, evolving through ongoing social engagement. In addition to facilitating self-expression, language underpins social cohesion and adaptation, playing a central role in the advancement of human societies [2]. Indonesia's status as an archipelagic nation has given rise to a remarkable diversity of regional languages. Among these, Palembang language is widely spoken in South Sumatra Province. In Palembang, the provincial capital, local dialects are woven into daily life across various neighborhoods. Sustaining these languages is crucial, as they embody the cultural richness and heritage of their communities [3].

One practical strategy for supporting the continued use of regional languages involves harnessing translation technology. Machine translation systems can help bridge communication gaps between Palembang and Indonesian speakers, making it easier for ideas and information to flow across linguistic boundaries [4]. The preservation of regional languages not only upholds cultural identity but also supports their continued relevance in an era marked by rapid technological change [5]. Translation tools play a meaningful role by broadening access and supporting intergroup communication, which in turn promotes understanding and cooperation among speakers of different backgrounds. Automatic translation from Palembang to Indonesian, and vice versa, can facilitate the exchange of perspectives and foster greater inclusivity in multilingual contexts.

Machine translation systems process natural language with the goal of converting text from one language to another without human intervention [6]. These systems are valued for their ability to streamline communication between speakers of different languages, offering a fast and accessible solution to language barriers. Their efficiency and convenience make them increasingly indispensable in today's interconnected world [7]. To develop a translation model that delivers reliable and accurate results, this study leverages the advanced capabilities of the No Language Left Behind (NLLB) project [8]. NLLB has demonstrated particular promise in supporting languages with limited digital resources, making it a strong candidate for addressing the specific needs of Palembang language translation.

The primary aim of this research is to construct a prototype automatic translation system based on the NLLB model, capable of translating text bidirectionally between Palembang and Indonesian with a high degree of accuracy while preserving intended meaning. Such a model can be integrated into digital applications, enhancing accessibility and supporting the ongoing use of Palembang language in contemporary settings. The adoption of NLLB for regional language translation offers a pathway to greater visibility and continued relevance for Palembang language, helping to ensure that it remains a living part of Indonesia's cultural landscape. This research, entitled "Palembang-Indonesian Machine Translation Using No Language Left Behind," addresses these challenges and opportunities.

2. Related Work

The development of machine translation systems, particularly for translating Palembang language into Indonesian, has gained increasing relevance in the current era of rapid technological advancement. Understanding the multifaceted aspects of effective and efficient translation has thus become a central concern in contemporary research. A foundational element in translation studies is the accurate identification of text types to be translated. Suhardi *et al.* (2019) emphasize the importance of recognizing four categories of proper names—personal names, geographical names, institutional names, and celebration names—in the translation process, as each category requires distinct strategies to preserve meaning and authenticity [17]. Palembang language, sensitivity to local terminology and names is essential to maintain the integrity of the original message. Gunawan (2022) further highlights that a nuanced understanding of context and meaning can significantly enhance user satisfaction and the effectiveness of translation strategies, particularly when navigating cultural differences between Palembang and Indonesian [11]. The choice of translation methodology also plays a pivotal role in determining translation quality. Rakhmyta (2022) demonstrates that collaborative training approaches can improve both comprehension and translation outcomes among translators, fostering the application of contextually appropriate strategies [14]. Such training underscores the necessity for translators to be adaptable, especially when dealing with the cultural and linguistic nuances characteristic of local languages. Technological advancements have increasingly influenced translation practices. Warsidi and Kamal (2022) observe that integrating technological tools with manual translation strategies can yield higher-quality translations, with artificial intelligence serving as a valuable aid to human translators in achieving greater accuracy and efficiency [12]. This synergy between technology and human expertise is particularly pertinent as the demand for rapid and reliable translation grows.

The methodologies employed in translation not only affect the quality but also the success of conveying intended meanings. Nugraha and Prasetya (2023) point out that translating administrative documents from English to Indonesian presents unique challenges, necessitating effective strategies to ensure contextual fidelity [13]. This need for strategic adaptation extends to the psychological dimensions of translation, as discussed by Sobri *et al.* (2024), who explore how psycholinguistic factors shape translators' understanding of the distinct features of different languages [16]. It is also crucial to recognize that translation techniques must be tailored to the specific genre of the source text. For example, Riani (2020) finds that translating religious texts requires heightened sensitivity to both cultural and linguistic contexts, while Akmaliyah *et al.* (2020) stress the importance of preserving the essence and depth of meaning when translating poetry [15][9]. These studies collectively underscore the varying demands of different text types and the necessity of adopting flexible translation approaches.

Adaptation emerges as a central theme in translation research. Fitriani and Ifianti (2021) illustrate how adaptation techniques are employed in translating children's storybooks, allowing translators to render content in ways that resonate with the target audience while remaining faithful to the source [10]. This adaptability is vital for ensuring that translations are both accessible and meaningful across cultural boundaries. Collectively, these studies provide valuable insights for the development of machine translation systems for Palembang-Indonesian language pairs. The integration of advanced technology, effective translation strategies, and deep cultural awareness is essential for producing translations that are both accurate and contextually relevant.

In the broader context of educational and translation policy, the concept of "No Language Left Behind" (NLLB) has gained prominence, especially in discussions about language education for students from minority language backgrounds. The NLLB initiative, rooted in the No Child Left Behind (NCLB) Act in the United States, was designed to ensure equitable access to quality education regardless of students' linguistic backgrounds. Menken (2009) notes that NCLB placed significant emphasis on accountability, compelling schools to meet specific standards in instruction and assessment, including language education [20]. However, Marszalek *et al*. (2022) report that English Language Learners (ELLs) often face challenges in meeting these standards due to limited English proficiency [19]. Research indicates that the implementation of NCLB has, in some cases, led to reduced support and funding for bilingual programs, despite evidence suggesting the superior effectiveness of bilingual education in promoting academic achievement compared to English-only instruction [22]. Rolstad *et al*. (2005), through meta-analysis, further confirm that bilingual education programs have a significantly greater impact on students' academic performance than monolingual approaches [23]. These findings advocate for a more comprehensive approach to language policy that values students' native languages as integral to their educational journey.

Moreover, Menken and Solorza (2012) argue for policy reforms that better accommodate bilingual students, emphasizing the importance of supporting native language instruction rather than focusing exclusively on English acquisition [21]. This perspective raises critical questions about the inclusivity of educational environments shaped by NCLB, particularly for students from diverse linguistic backgrounds. In practice, the NCLB Act, in effect from 2001 to 2015, heavily influenced school policies by prioritizing standardized test results. Menken (2009) critiques this focus, suggesting that an overemphasis on testing can marginalize essential aspects of language teaching [20]. Consequently, ongoing dialogue among researchers and educators is necessary to develop policies that holistically support both teachers and bilingual learners. Hornberger and Link (2012) advocate for educational policies that reflect the complex ethnic and linguistic realities of multilingual classrooms, moving beyond rigid bilingual frameworks to embrace more inclusive and dynamic approaches [18]. The NLLB policy, therefore, should strive for broader and more flexible implementation, enabling active participation and academic success for all students, irrespective of their linguistic heritage.

3. Research Method

This study focuses on the development of a machine translation model for translating Palembang language into Indonesian using the No Language Left Behind (NLLB) algorithm. The initial stage involves a sequence of processes, beginning with data collection from various sources, including Palembang–Indonesian dictionaries. The next steps include language preparation, normalization, and tokenization, which are carried out prior to the model training phase. Once the model has been trained, its performance is evaluated, and the resulting model is uploaded to the HuggingFace platform. The trained model is then utilized to automatically and efficiently translate text between Palembang and Indonesian in both directions. The methodological stages employed in this research are illustrated in Figure 1.

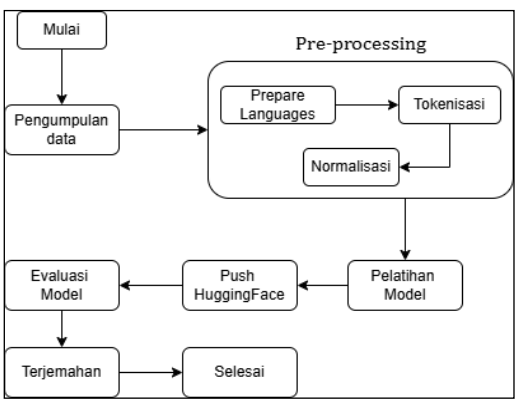
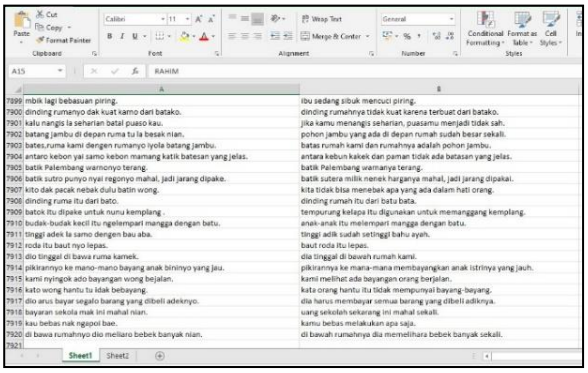


Figure 1. Research Methodology Flowchart.

4. Result and Discussion

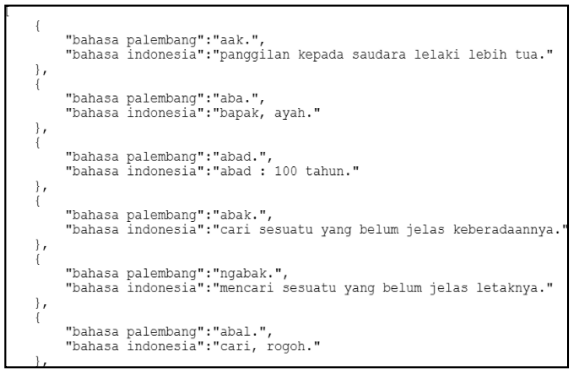
4.1 Results

In this study, the data collected and used originated from scanned versions of the Kamus Sederhana Bahasa Palembang and the Kamus Palembang-Indonesia. The scanning process was conducted meticulously to ensure the accuracy and completeness of the data obtained [9][2][3][5][4][6][7][8]. As a result of this process, a total of 7,917 entries were acquired, consisting of 5,053 words and 2,864 sentences. All collected data were stored in an Excel file. Subsequently, the data were separated from the Excel file and saved in CSV (Comma-Separated Values) format, then converted to JSON. This separation process was intended to facilitate data processing in both Palembang and Indonesian languages, which would be used for translation tasks. This approach is expected to make the translation process more efficient and accurate.



7899	mbik lagi bebasuan piring.	ibu sedang sibuk mencuci piring.
7900	dinding rumanyo dak kuat karno dari batako.	dinding rumahnya tidak kuat karena terbuat dari batako.
7901	kalu nangis la seharian batal puaso kau.	jika kamu menangis seharian, puasamu menjadi tidak sah.
7902	batang jambu di depan ruma tu la besak nian.	pohon jambu yang ada di depan rumah sudah besar sekali.
7903	bates ruma kami dengen rumanyo iyola batang jambu.	batas rumah kami dan rumahnya adalah pohon jambu.
7904	antaro kebon yai samo kebon mamang katik batesan yang jelas.	antara kebun kakek dan paman tidak ada batasan yang jelas.
7905	batik Palembang warnonyo terang.	batik Palembang warnanya terang.
7906	batik sutro punyo nyai ngonyo mahal, jadi jarang dipake.	batik sutera milik nenek harganya mahal, jadi jarang dipakai.
7907	kito dak pacak nebak dulu batin wong.	kita tidak bisa menebak apa yang ada dalam hati orang.
7908	dinding ruma itu dari bato.	dinding rumah itu dari batu bata.
7909	batok itu dipake untuk nunu kemplang .	tempurung kelapa itu digunakan untuk memanggang kemplang.
7910	budak-budak kecil itu ngelempari mangga dengan batu.	anak-anak itu melempari mangga dengan batu.
7911	tinggi adek la samo dengen bau aba.	tinggi adik sudah setinggi bahu ayah.
7912	roda itu baut nyo lepas.	baut roda itu lepas.
7913	dio tinggal di bawa ruma kamek.	dia tinggal di bawah rumah kami.
7914	pikirannyo ke mano-mano bayang anak bininyo yang jau.	pikirannya ke mana-mana membayangkan anak istrinya yang jauh.
7915	kami nyingok ado bayangan wong bejalan.	kami melihat ada bayangan orang berjalan.
7916	kato wong hantu tu idak bebayang.	kata orang hantu itu tidak mempunyai bayang-bayang.
7917	dio arus bayar segalo barang yang dibeli adeknyo.	dia harus membayar semua barang yang dibeli adiknya.
7918	bayaran sekola mak ini mahal nian.	uang sekolah sekarang ini mahal sekali.
7919	kau bebas nak ngapoi bae.	kamu bebas melakukan apa saja.
7920	di bawa rumahnyo dio meliaro bebek banyak nian.	di bawah rumahnya dia memelihara bebek banyak sekali.

Figure 2. Dataset Display

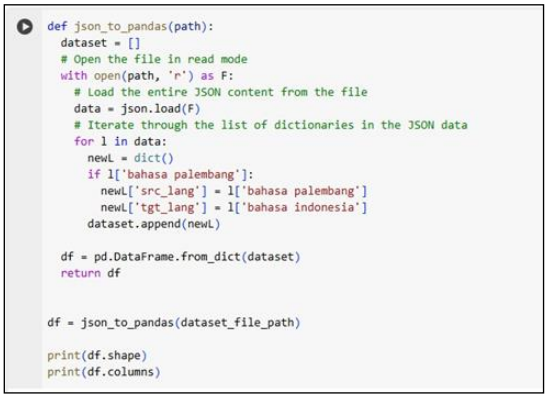


```
{
    "bahasa palembang":"aak.",
    "bahasa indonesia":"panggilan kepada saudara lelaki lebih tua."
},
{
    "bahasa palembang":"aba.",
    "bahasa indonesia":"bapak, ayah."
},
{
    "bahasa palembang":"abad.",
    "bahasa indonesia":"abad : 100 tahun."
},
{
    "bahasa palembang":"abak.",
    "bahasa indonesia":"cari sesuatu yang belum jelas keberadaannya."
},
{
    "bahasa palembang":"ngabak.",
    "bahasa indonesia":"mencari sesuatu yang belum jelas letaknya."
},
{
    "bahasa palembang":"abal.",
    "bahasa indonesia":"cari, rogoh."
},
```

Figure 3. JSON File Display

4.1.1 Processing

1) PrepareLanguages



```
def json_to_pandas(path):
  dataset = []
  # Open the file in read mode
  with open(path, 'r') as F:
    # Load the entire JSON content from the file
    data = json.load(F)
    # Iterate through the list of dictionaries in the JSON data
    for l in data:
      newL = dict()
      if l['bahasa palembang']:
        newL['src_lang'] = l['bahasa palembang']
        newL['tgt_lang'] = l['bahasa indonesia']
      dataset.append(newL)

  df = pd.DataFrame.from_dict(dataset)
  return df


df = json_to_pandas(dataset_file_path)

print(df.shape)
print(df.columns)
```

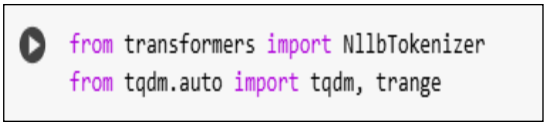
Figure 4. PrepareLanguages Code

This code is designed to read a JSON file containing Palembang-Indonesian language pairs, perform data cleaning and filtering, and convert it into a Pandas DataFrame with a standardized format: the src_lang column as the source language (Palembang) and the tgt_lang column as the target language (Indonesian). This structure simplifies data analysis or subsequent use in language processing.

2) Tokenization

Tokenization is an essential process in natural language processing that aims to break text into smaller units, such as words or subwords, so that it can be processed efficiently and accurately by machine learning models. The following is a step-by-step explanation of the tokenization code functions.

a) At this stage, the code is intended to import the modules and libraries required for the text tokenization process and to display a progress bar for various iterative operations.



```
from transformers import NllbTokenizer
from tqdm.auto import tqdm, trange
```

Figure 5. Tokenization Import Code

b) At this stage, the code is used to initialize the tokenizer from the pre-trained NLLB (No Language Left Behind) model, specifically the nllb-200-distilled-600M version. This tokenizer is responsible for processing raw text into a format that can be understood by the NLLB model for tasks such as machine translation or multilingual language processing.

Figure 6. Tokenizer Initialization Code

c) This code defines a simple tokenizer function using regular expressions to split text into tokens in the form of words, numbers, or punctuation marks. This function is suitable for basic text processing in languages that use the Latin alphabet.

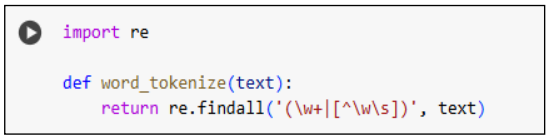


Figure 7. Tokenizer Definition Code

d) This code randomly samples 1,000 rows from the training dataset with replacement, then performs tokenization on the Palembang text column (src_lang) and the Indonesian text column (tgt_lang) using two methods: tokenizer.tokenize for model vocabulary-based tokenization, and a regex-based word_tokenizer for splitting words and punctuation. The tokenization results are stored in new columns to facilitate further data analysis and processing.

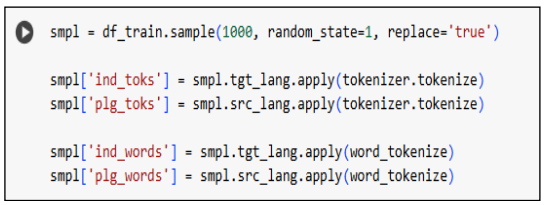


Figure 8. Dataset Sampling Code

e) This code aims to identify and count texts in the Palembang dataset (src_lang) that contain the <unk> token (used for unrecognized words or symbols).

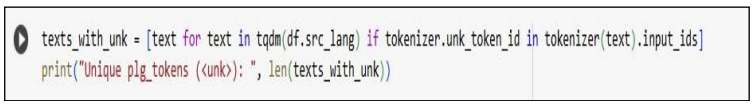


Figure 9. Token Identification Code

f) This code is intended to update or modify the NLLB-200 tokenizer by adding a new language token (in this case, Palembang with Latin script, plg_Latn) to the tokenizer's vocabulary.

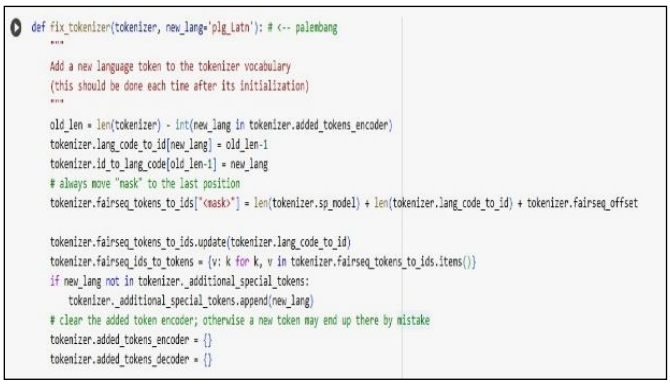


Figure 10. Adding a New Language Code

g) This code prepares the model to support the new language, Palembang Latin (plg_Latn), by identifying the token IDs for the new language and a similar language, Indonesian Latin (ind_Latn). The similar language is used to provide an initial reference for the new language embedding in the model. Since the new language plg_Latn does not yet have a trained embedding representation, ind_Latn is used as the initial reference.

```
# fixing the new/moved token embeddings in the model
added_token_id = tokenizer.convert_tokens_to_ids('plg_Latn')
similar_lang_id = tokenizer.convert_tokens_to_ids('ind_Latn') # <-- needs research!!!
print(added_token_id, similar_lang_id)
```

Figure 11. Supporting a New Language Code

h) This code is used to load the pre-trained NLLB-200 (No Language Left Behind) model, which is a transformer-based translation model trained for 200 languages, and to adjust the model's embedding size to match the modified tokenizer.

```
model = AutoModelForSeq2SeqLM.from_pretrained('facebook/nllb-200-distilled-600M')
model.resize_token_embeddings(len(tokenizer))
```

Figure 12. Adjusting Embedding Size Code

i) This code aims to move the special token embedding (<mask>) to a new position to ensure it continues to function properly within the model, and to initialize the new language token embedding (plg_Latn) with the embedding vector from the similar language (ind_Latn). This provides a semantically reasonable initial representation for the new language token, helping to leverage cross-language relationships and reduce the model's training burden.

```
# moving the embedding for "mask" to its new position
model.model.shared.weight.data[added_token_id+1] = model.model.shared.weight.data[added_token_id]
# initializing new language token with a token of a similar language
model.model.shared.weight.data[added_token_id] = model.model.shared.weight.data[similar_lang_id]
```

Figure 13. Adjusting and Initializing New Token Embeddings Code

3) Normalization

Figure 14. Normalization Code

This code is intended to thoroughly preprocess text by removing unnecessary characters, normalizing punctuation, and converting text into a standard Unicode form that is easy to process. The code focuses on cleaning and preparing the text data to be more consistent and compatible with NLP (Natural Language Processing) tasks.

4) Model Training

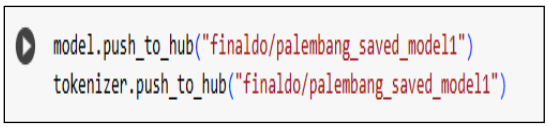
At this stage, the NLLB (No Language Left Behind) model is trained. After the model is created, training is conducted so that the model can perform translation between Palembang and Indonesian. Figure 15 explains the model training code:

```
x, y, loss = None, None, None
cleanup()

tq = trange(len(losses), training_steps)
for i in tq:
    xx, yy, langs, lang2 = get_batch_pairs(batch_size)
    try:
        tokenizer.src_lang = lang1
        x = tokenizer(xx, return_tensors='pt', padding=True, truncation=True, max_length=max_length).to(model.device)
        tokenizer.src_lang = lang2
        y = tokenizer(yy, return_tensors='pt', padding=True, truncation=True, max_length=max_length).to(model.device)
        y.input_ids[y.input_ids == tokenizer.pad_token_id] = -100

        loss = model(**x, labels=y.input_ids).loss
        loss.backward()
        losses.append(loss.item())

        optimizer.step()
        optimizer.zero_grad(set_to_none=True)
        scheduler.step()

    except RuntimeError as e:
        optimizer.zero_grad(set_to_none=True)
        x, y, loss = None, None, None
        cleanup()
        print('error', max(len(s) for s in xx + yy), e)
        continue

    if i % 1000 == 0:
        print(i, np.mean(losses[-1000:]))

    if i % 1000 == 0 and i > 0:
        model.save_pretrained(MODEL_SAVE_PATH)
        tokenizer.save_pretrained(MODEL_SAVE_PATH)
```

Figure 15. Model Training Code

This code is used to train a sequence-to-sequence model by processing source and target data pairs from the training dataset. The process includes tokenizing the data based on the source and target languages, calculating the loss, updating the model parameters through optimization, and handling errors such as GPU memory exhaustion. Every 1,000 iterations, the model's performance is monitored by printing the latest average loss, and the training progress is saved by storing the model and tokenizer. The ultimate goal is to adapt the model to understand the relationships between the trained language pairs.

5) Pushing to Hugging Face

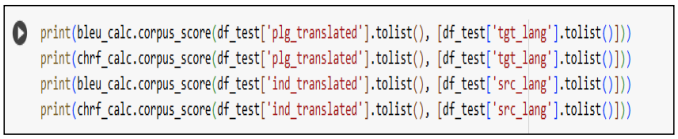


```
model.push_to_hub("finaldo/palembang_saved_model1")
tokenizer.push_to_hub("finaldo/palembang_saved_model1")
```

Figure 16. Hugging Face Push Code

In this stage, the code is used to upload the trained model and tokenizer to the Hugging Face Hub under the repository name "finaldo/palembang_saved_model1", enabling users to store and access the model.

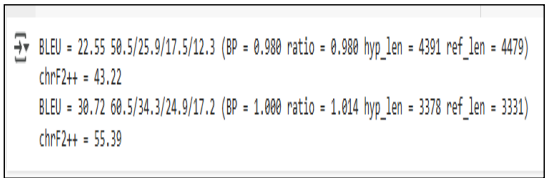
6) Model Evaluation



```
print(bleu_calc.corpus_score(df_test['plg_translated'].tolist(), [df_test['tgt_lang'].tolist()]))
print(chrf_calc.corpus_score(df_test['plg_translated'].tolist(), [df_test['tgt_lang'].tolist()]))
print(bleu_calc.corpus_score(df_test['ind_translated'].tolist(), [df_test['src_lang'].tolist()]))
print(chrf_calc.corpus_score(df_test['ind_translated'].tolist(), [df_test['src_lang'].tolist()]))
```

Figure 17. Model Evaluation Code

This process evaluates the previously trained model using two main evaluation metrics: BLEU (Bilingual Evaluation Understudy) and CHRF (Character F-score). The evaluation is performed by generating translation predictions from the model. These predictions provide an overview of how well the model produces results that match the correct translations.



```
BLEU = 22.55 50.5/25.9/17.5/12.3 (BP = 0.980 ratio = 0.980 hyp_len = 4391 ref_len = 4479)
chrF2++ = 43.22
BLEU = 30.72 60.5/34.3/24.9/17.2 (BP = 1.000 ratio = 1.014 hyp_len = 3378 ref_len = 3331)
chrF2++ = 55.39
```

Figure 18. Model Evaluation Output

The evaluation results using BLEU and chrF++ metrics show satisfactory performance. The first model achieved a BLEU score of 22.55 and a chrF++ score of 43.22, while the second model showed improvement with a BLEU score of 30.72 and a chrF++ score of 55.39. These values indicate that the second model has higher translation accuracy and alignment with the reference, as also supported by the hypothesis-to-reference length ratio approaching the ideal (BLEU BP = 1.000).

7) Translation

The translation process aims to convert text from Palembang to Indonesian, and vice versa. The translation is performed while maintaining the main message and meaning so that it can be well understood by readers in the target language. Below is an example of translation results from Palembang to Indonesian.

Table 1. Translation

Palembang Language > Indonesian	Indonesian > Palembang Language

`t = "Mak mano kabar kau"`
`print(translate(t, 'plg_Latn', 'ind_Latn'))`
`['bagaimana kabarmu']`
	`t = "saya mau pergi ke pasar membeli beras"`
	`print(translate(t, 'ind_Latn', 'plg_Latn'))`
	`['aku nak pegi ke pasar meli beras']`

`t = "pacak kau mang"`
`print(translate(t, 'plg_Latn', 'ind_Latn'))`
`['bisa kamu paman']`
	`t = "Nanti kita makan pempek di pinggir Sungai Musi"`
	`print(translate(t, 'ind_Latn', 'plg_Latn'))`
	`['kagek kito makan pempek di pinggir Sungai Musi']`

`t = "ado bisul rainyo"`
`print(translate(t, 'plg_Latn', 'ind_Latn'))`
`['ada bisul wajahnya']`
	`t = "Jangan suka seperti itu"`
	`print(translate(t, 'ind_Latn', 'plg_Latn'))`
	`['Jangan galak cak itu']`

`t = "Kalu pacak, antar aku balek ke rumah mamang aku"`
`print(translate(t, 'plg_Latn', 'ind_Latn'))`
`['kalau bisa, antar saya pulang ke rumah paman saya']`
	`t = "dia bisa saja minta tolong ke saya"`
	`print(translate(t, 'ind_Latn', 'plg_Latn'))`
	`['dio pacak bae mintak tulung ke aku']`

`t = "Ngapo kau masih tidok"`
`print(translate(t, 'plg_Latn', 'ind_Latn'))`
`['mengapa kamu masih tidur']`
	`t = "ayo kita ke benteng kuto besak sore ini"`
	`print(translate(t, 'ind_Latn', 'plg_Latn'))`
	`['payo kito ke benteng kuto besak sore ni']`

4.2 Discussion

This study highlights the significance of preserving and developing local languages as a vital component of Indonesian cultural identity amid globalization [3][5]. The meticulous collection of data from the Kamus Sederhana Bahasa Palembang and Kamus Palembang-Indonesia ensured the accuracy and completeness of the dataset, reflecting best practices found in the development of web-based dictionary applications [4]. The conversion of data from Excel to CSV and then to JSON facilitated efficient data processing, aligning with established methods in machine translation applications using Natural Language Processing (NLP) [8].

The preparation stage, which organized language pairs into standardized source and target columns, was crucial for supporting subsequent data analysis and processing. The tokenization process, employing both model-based and regular expression-based approaches, enhanced text segmentation and enabled the model to better capture the structure and patterns of the Palembang language. The adaptation of the NLLB-200 model by adding a new language token (plg_Latn) and leveraging the embedding of a related language (ind_Latn) exemplifies an effective transfer learning strategy for low-resource languages [6][7]. This approach has proven successful in accelerating the model's adaptation to new languages and improving translation quality.

Data normalization was also critical to ensure consistency and cleanliness, which are essential prerequisites in NLP model development [8]. The model training process was conducted in stages, with performance monitored regularly using BLEU and chrF++ metrics. The evaluation results showed significant improvements in BLEU and chrF++ scores after further training, indicating that the model produced increasingly accurate translations that closely matched the references [9][10]. The BLEU BP value approaching 1 also demonstrates that the model's translations are proportionate in length to the reference sentences.

The deployment of this model on the Hugging Face Hub not only broadens access for other researchers but also encourages further collaboration and development in local language translation technology. This study demonstrates that modern NLP and transfer learning approaches can be effectively implemented for low-resource languages, making a tangible contribution to the preservation and development of Indonesia's regional languages [2][5]. Therefore, these research outcomes can serve as a foundation for developing automatic translation systems for other local languages in Indonesia, ultimately strengthening the nation's cultural identity and diversity.

5. Conclusion and Recommendations

The conclusion provides a clear summary of the research and the key features of the proposed system. This section effectively highlights the system's ability to accurately translate between Palembang and Indonesian languages while preserving contextual meaning, despite differences in sentence structure and vocabulary between the two languages. The success of the model is greatly influenced by the quality of the training data used and the preprocessing steps performed prior to training. With the right approach, the model can understand and translate various forms of the Palembang language into easily comprehensible Indonesian. These findings are significant not only in the field of technology but also in efforts to preserve regional languages. The presence of this translation engine can help introduce and maintain the use of the Palembang language in the digital era, preventing it from being eroded by the passage of time. In addition, this approach can also be applied to develop automatic translators for other regional languages in Indonesia that face similar challenges. In the future, further development is needed by increasing the amount and variety of training data, as well as conducting more in-depth evaluations, so that translation quality can be further improved and widely used by the community.

The suggestions in this study highlight that the translation engine still has limitations, mainly because it can only translate one sentence at a time. Therefore, it is important to develop more advanced translation engines in the future that can handle more complex texts, such as paragraphs or entire documents. Furthermore, performance improvements should also focus on context processing, so that the translation results become more accurate, natural, and relevant.

References

[1] Mailani, O., Nuraeni, I., Syakila, S. A., & Lazuardi, J. (2022). Bahasa sebagai alat komunikasi dalam kehidupan manusia. *Kampret Journal*, *1*(2), 1-10. https://doi.org/10.35335/kampret.v1i1.8

[2] Sirait, Z., Amalia, A., & Marpaung, D. (2022). Penerapan bahasa Indonesia yang baik dalam memudahkan penjualan barang melalui e-commerce. *Community Development Journal: Jurnal Pengabdian Masyarakat*, 3(1), 26–29. https://doi.org/10.31004/cdj.v3i1.3391

[3] Swarna, M. F., Royani, A., Lestari, S. I., & Rahmawati, C. A. (2024). Peranan Gen Z Dalam Mempertahankan Budaya Lokal Indonesia Di Era Global. *Karimah Tauhid*, *3*(5), 5947-5953. https://doi.org/10.30997/karimahtauhid.v3i5.13298

[4] Andri, A. (2019). Penerapan algoritma pencarian binary search dan quicksort pada aplikasi kamus bahasa Palembang berbasis web. *Jurnal Informatika: Jurnal Pengembangan IT*, 4(1), 70–74. https://doi.org/10.30591/jpit.v4i1.1104

[5] Putri, B. T., Ayu, C. S., Ginting, M. A. B., Saidah, S., & Nasution, S. (2025). Budaya dan bahasa: Refleksi dinamis identitas masyarakat. *Semantik: Jurnal Riset Ilmu Pendidikan, Bahasa dan Budaya*, *3*(1), 20-32. https://doi.org/10.61132/semantik.v3i1.1321

[6] Mokhtar, I. P., Wardani, A., Sajidah, H., Faqih, M., Hati, B. K., Satria, A., ... & Lailani, A. (2025, February). Penerapan Model T5 untuk Penerjemahan Mesin Aceh-Indonesia. In *Prosiding Seminar Nasional Sains dan Teknologi" SainTek"* (Vol. 2, No. 1, pp. 379-393).

[7] Pranaja, M. A., & Nurhidayat, A. I. (2023). Pembuatan Aplikasi Mesin Penerjemah Menggunakan Metode No Language Left Behind Dari Bahasa Indonesia Ke Bahasa Banjar. *Jurnal Manajemen Informatika*, *15*(01).

[8] Ghirrid, A. A., Sari, R. T. K., & Aldisa, R. T. (2024). Algoritma Natural Language Processing Untuk Aplikasi Penerjemah (Indonesia – Jawa) Menggunakan Metode Speech Processing. *Jurnal JTIK (Jurnal Teknologi Informasi Dan Komunikasi)*, *8*(3), 746-759. https://doi.org/10.35870/jtik.v8i3.2244

[9] Akmaliyah, A., Maulidiyah, L., & Supianudin, A. (2020). Seni menerjemahkan puisi: Studi kasus terjemahan Arab atas dua sajak karya Sapardi Djoko Damono oleh Usman Arrumy. *Al-Tsaqafa: Jurnal Ilmiah Peradaban Islam*, 17(2), 140–146. https://doi.org/10.15575/al-tsaqafa.v17i2.6398

[10] Fitriani, E., & Ifianti, T. (2021). Onomatope dalam buku cerita anak dwibahasa Little Abid seri pengetahuan dasar (analisis metode dan prosedur penerjemahan). *Briliant: Jurnal Riset dan Konseptual*, 6(1), 66. https://doi.org/10.28926/briliant.v6i1.584

[11] Gunawan, F. (2022). Strategi penerjemahan kata zina dan rafas: Sebuah reinterpresentasi. *Ranah: Jurnal Kajian Bahasa*, 11(2), 339. https://doi.org/10.26499/rnh.v11i2.4904

[12] Kamal, A. (2022). Pelatihan penerjemahan Indonesian-English dengan menggunakan kombinasi Google Translate dan menerapkan manual translation strategies. *Ash-Shahabah: Jurnal Pengabdian Masyarakat*, 1(2), 9–15. https://doi.org/10.59638/ashabdimas.v1i2.534

[13] Nugraha, N., & Prasetya, R. (2023). Strategi penerjemahan dokumen administrasi berbahasa Inggris dalam lingkungan bisnis startup di Indonesia (studi kasus pada perusahaan XYZ). *Jurnal Serasi*, 20(2), 126. https://doi.org/10.36080/js.v20i2.2243

[14] Rakhmyta, Y. (2022). Pelatihan penerjemahan teks resep makanan dalam pasangan bahasa Inggris dan bahasa Indonesia pada mata kuliah translation. *Catimore: Jurnal Pengabdian Kepada Masyarakat*, 1(2), 6–12. https://doi.org/10.56921/cpkm.v1i2.14

[15] Riani, R. (2020). Teknik penerjemahan People's Religion: Penerjemahan teks religius berbahasa Sunda ke dalam bahasa Inggris. *Genta Bahtera: Jurnal Ilmiah Kebahasaan dan Kesastraan*, 6(2). https://doi.org/10.47269/gb.v6i2.114

[16] Sobri, A., Syahvini, S., Rizqa, R., Padilah, S., Athallah, M., & Fadila, N. (2024). Perbedaan penerjemahan gramatikal bahasa Arab dan bahasa Indonesia. *EDU*, 1(3), 316–324. https://doi.org/10.60132/edu.v1i3.184

[17] Suhardi, S., Widodo, P., & Setiawan, T. (2019). Equivalence of proper name in foreign language into the Indonesian language. *Litera*, 18(1), 1–16. https://doi.org/10.21831/ltr.v18i1.23105

[18] Hornberger, N., & Link, H. (2012). Translanguaging and transnational literacies in multilingual classrooms: A biliteracy lens. *International Journal of Bilingual Education and Bilingualism*, 15(3), 261–278. https://doi.org/10.1080/13670050.2012.658016

[19] Marszalek, J., Balagna, D., Kim, A., & Patel, S. (2022). Self-concept and intrinsic motivation in foreign language learning: The connection between flow and the L2 self. *Frontiers in Education*, 7. https://doi.org/10.3389/feduc.2022.975163

[20] Menken, K. (2009). No child left behind and its effects on language policy. *Annual Review of Applied Linguistics*, 29, 103–117. https://doi.org/10.1017/s0267190509090096

[21] Menken, K., & Solorza, C. (2012). No child left bilingual. *Educational Policy*, 28(1), 96–125. https://doi.org/10.1177/0895904812468228

[22] Pufahl, I., & Rhodes, N. (2011). Foreign language instruction in U.S. schools: Results of a national survey of elementary and secondary schools. *Foreign Language Annals*, 44(2), 258–288. https://doi.org/10.1111/j.1944-9720.2011.01130.x

[23] Rolstad, K., Mahoney, K., & Glass, G. (2005). The big picture: A meta-analysis of program effectiveness research on English language learners. *Educational Policy*, 19(4), 572–594. https://doi.org/10.1177/0895904805278067