NusaWrites: Constructing High-Quality Corpora for Underrepresented and Extremely Low-Resource Languages

Samuel Cahyawijaya[1,4*], **Holy Lovenia**[1,4*], **Fajri Koto**[3,4*], **Dea Adhista**[2†], **Emmanuel Dave**[4†],
Sarah Oktavianti[2†], **Salsabil Maulana Akbar**[4†], **Jhonson Lee**[4†], **Nuur Shadieq**[4†],
Tjeng Wawan Cenggoro[8,4†], **Hanung Wahyuning Linuwih**[2†], **Bryan Wilie**[1,4†],
Galih Pradipta Muridan[2†], **Genta Indra Winata**[5,4†],
David Moeljadi[7†], **Alham Fikri Aji**[3,4†], **Ayu Purwarianti**[6,4], **Pascale Fung**[1]

[1]HKUST   [2]Prosa.ai   [3]MBZUAI   [4]IndoNLP   [5]Bloomberg
[6]Institut Teknologi Bandung [7]Kanda University of International Studies   [8]Bina Nusantara University
[*]Equal Contribution   [†] Equal Contribution

Abstract

Democratizing access to natural language processing (NLP) technology is crucial, especially for underrepresented and extremely low-resource languages. Previous research has focused on developing labeled and unlabeled corpora for these languages through online scraping and document translation. While these methods have proven effective and cost-efficient, we have identified limitations in the resulting corpora, including a lack of lexical diversity and cultural relevance to local communities. To address this gap, we conduct a case study on Indonesian local languages. We compare the effectiveness of online scraping, human translation, and paragraph writing by native speakers in constructing datasets. Our findings demonstrate that datasets generated through paragraph writing by native speakers exhibit superior quality in terms of lexical diversity and cultural content. In addition, we present the NusaWrites benchmark, encompassing 12 underrepresented and extremely low-resource languages spoken by millions of individuals in Indonesia. Our empirical experiment results using existing multilingual large language models emphasize the need to extend these models to more underrepresented languages. We release the NusaWrites dataset[1] and code involved in our experiment at https://github.com/IndoNLP/nusa-writes.

1   Introduction

Most of the research works in today's NLP technology are culturally Anglocentric with English as the main language (Søgaard, 2022; Talat et al., 2022). While it is critical to democratize NLP to underrepresented languages, previous works (Cahyawijaya et al., 2022; Kakwani et al., 2020; Koto et al.,
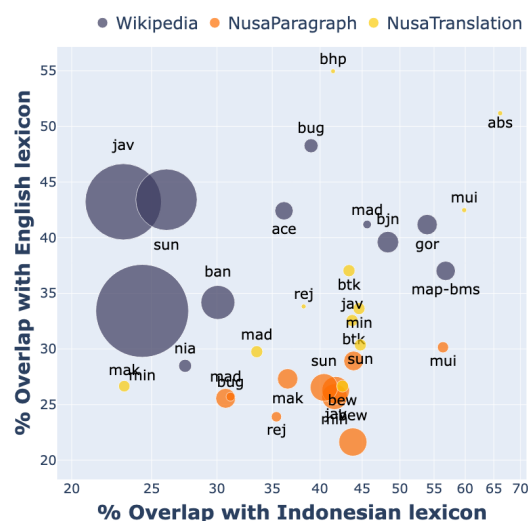

Figure 1: Unique lexicon overlaps of various language corpora with Indonesian and English languages. The Indonesian and English lexicons are from Panlex.[2]

2020; Koto and Koto, 2020; Wilie et al., 2020; Adelani et al., 2021a,b; Cahyawijaya et al., 2021; Ebrahimi et al., 2022; Park et al., 2021; Kumar et al., 2022; Winata et al., 2023; Adilazuarda et al., 2022; Ogundepo et al., 2023; Kabra et al., 2023; Song et al., 2023) have developed labeled and unlabeled corpora in the languages mainly through document translation (Winata et al., 2023) and online scraping (Koto et al., 2021, 2022b). Although such data collection methods could be effective in high-resource languages, applying the methods in underrepresented languages still needs further investigation.

In this work, we compare three corpus collection methods for 12 underrepresented languages in Indonesia, namely Ambon (abs), Batak (btk), Betawi (bew), Bima (bhp), Buginese (bug), Javanese (jav), Madurese (mad), Makassarese (mak), Minangkabau (min), Palembang / Musi (mui), Rejang (rej),

---

[1]Accessible through nusacrowd and HuggingFace datasets packages. Kindly check the README on GitHub for more information.

[2]https://panlex.org/

and Sundanese (sun). We chose Indonesian local languages as our case study because of the language diversity in Indonesia, with more than 700 languages spoken but most of them are underrepresented and extremely low-resource (Cohn and Ravindranath, 2014; Aji et al., 2022b). Bang et al. (2023) categorize Javanese (jav) and Sundanese (sun) as low-resource languages, while the others as extremely low-resource languages. For Ambon (abs), Bima (bhp), Makassarese (mak), Musi (mui), and Rejang (rej), they have no publicly available labeled and unlabeled corpora despite there being millions of speakers. We provide information on 12 low-resource languages under study in Table 10. We conduct two manual data construction efforts for the 12 languages: topic-focused paragraph writing (NusaParagraph) and human translation by native speakers (NusaTranslation),[3] and benchmark them with online scraping. For online scraping, we utilize Wikipedia[4] as the main source as it covers some of the Indonesian local languages under study. Figure 1 summarizes the corpora constructed by each approach: Wikipedia, NusaParagraph, and NusaTranslation for online scraping, paragraph writing, and human translation, respectively. NusaParagraph tends to have fewer English and Indonesian lexicons, indicating they are more relevant to the local cultures than the others.

We build a new benchmark for the 12 Indonesian local languages, namely **NusaWrites**[5], using the texts produced in topic-focused paragraph writing and human translation. NusaWrites covers 5 natural language understanding tasks (e.g., emotion, sentiment classification) and one natural language generation task (i.e., machine translation), and complements NusaX (Winata et al., 2023)—a contemporaneous work on 10 Indonesian local languages for sentiment analysis and machine translation. We also demonstrate the inability of (1) fine-tuned Indonesian and multilingual language models (LMs) and (2) zero-shot prompting via large LMs (LLMs) to adapt to these languages, indicating that these languages are distinct from the existing models.

Our contributions to this work are four-fold:

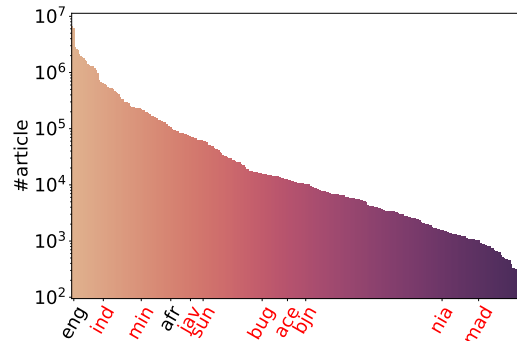- We compare various corpus collection methods for underrepresented and extremely low-

Figure 2: Distribution of Indonesian languages in Wikipedia, compared against all existing languages.

resource languages. We show that paragraph writing is the most promising strategy for building high-quality and culturally-relevant corpora.

- We extend the NLP resource coverage of Indonesian local languages with 5 new languages: Ambon (abs), Bima (bhp), Makassarese (mak), Musi (mui), and Rejang (rej).

- We propose NusaWrites, a benchmark covering new high-quality human annotated corpora consisting of 12 underrepresented languages in Indonesia with 5 downstream tasks.

- We conduct extensive analysis to showcase the similarity between languages under study with Indonesian and the inability of existing LLMs to process these languages.

2 Indonesian Local Languages in Wikipedia

Figure 2 describes Indonesian local languages which are covered in Wikipedia, compared against other existing languages. In total, there are only 11 local languages (out of 700+ (Aji et al., 2022b)), with Minangkabau (min), Javanese (jav), and Sundanese (sun) having a quite large amount of documents around ~100,000 articles, while the remaining languages have less than ~10,000 articles. Despite its relatively large scale in Wikipedia, the text quality is not consistently as good as reported in the WikiMatrix dataset (Schwenk et al., 2021). Kreutzer et al. (2022) further find that ~30% of the correct translation data in English-Javanese are either boilerplates or low-quality texts.

To further verify the quality of Indonesian local languages in Wikipedia, we conduct an analysis to measure lexical diversity in two approaches: 1) calculating the cumulative token distribution per language and 2) measuring the length-agnostic lexical

---

[3]Note that most Indonesians are at least bilingual since they speak Indonesian and their local language (Aji et al., 2022b; Koto and Koto, 2020).

[4]https://www.wikipedia.org/

[5]The "Nusa" of NusaWrites is abbreviated from "Nusantara", which refers to the Indonesian archipelago.

diversity metrics, i.e., *moving average type-token ratio* (MATTR) (Covington and McFall, 2010), *measure of textual lexical diversity* (MTLD) (Mc-Carthy, 2005), and *mean segmental type-token ratio* (MSTTR) (Johnson, 1944). We use Lexical-Richness (Shen, 2021, 2022) v0.5.0[6] to calculate these metrics. Based on our analysis in Table 1, we show that some Indonesian local languages in Wikipedia have much less lexical diversity despite having quite a number of articles in Wikipedia, especially for Buginese (bug), Acehnese (ace), Gorontalo (gor), and Nias (nia). Through a further inspection of the Wikipedia corpus presented in §4.1 and Appendix G, Wikipedia articles for these languages tend to comprise many boilerplate texts, especially for Buginese (bug) Wikipedia.

Indonesian Local Languages in Other Sources Other than Wikipedia, there are other large multilingual corpora such as CommonCrawl,[7] mC4 (Xue et al., 2021), OSCAR (Suárez et al., 2019), FLORES-200 (Guzmán et al., 2019; Costa-jussà et al., 2022), and Bible corpus.[8] Nevertheless, most sources, except the Bible corpus, only support some widely-spoken languages spoken in Indonesia: Indonesian (ind), Javanese (jav), and Sundanese (sun), rendering them ineffective for studying hundreds of local languages spoken in Indonesia. The Bible corpus, on the other hand, consists of 14 Indonesian local languages.[9] Interestingly, these languages have an extremely low number of speakers with an average population of 40k people. On the contrary, Wikipedia covers Indonesian local languages with a larger number of speakers, with Nias (nia) being the smallest (nearly 770k speakers). In this work, we particularly focus on Indonesian local languages with larger population size (∼500k or above), and leave the exploration of the smaller-scale languages for future work.

3 Corpus Construction for Indonesian Local Languages

We conduct corpus construction through human annotation by expert workers in two ways: (1) sentence translation and (2) paragraph writing. Sentence translation is a widely used parallel data collection method (Conneau et al., 2018; Hu et al.,

---

[6] https://pypi.org/project/lexicalrichness/
[7] http://commoncrawl.org
[8] https://huggingface.co/datasets/bible-nlp/biblenlp-corpus
[9] Details of the Indonesian local languages in the Bible corpus are in Appendix A.

lang	category	MATTR	MTLD	MSTTR	Avg.
gor	X-LRL	69.40	37.23	71.36	**47.04**
ace	X-LRL	77.87	30.65	75.91	**51.36**
bug	X-LRL	79.81	28.61	80.12	**53.41**
nia	X-LRL	84.75	68.85	86.33	**57.25**
ban	X-LRL	85.15	53.83	86.57	**57.42**
map-bms	X-LRL	86.62	70.76	87.89	**58.41**
bjn	X-LRL	87.27	83.57	88.20	**58.77**
jav	LRL	89.18	58.94	88.19	**59.32**
ind	MRL	89.88	83.82	90.11	**60.27**
mad	X-LRL	89.88	67.21	90.53	**60.36**
sun	LRL	94.47	70.12	88.92	**61.37**
min	X-LRL	94.23	80.86	92.12	**62.39**

Table 1: Lexical diversity of various Indonesian local languages corpora in Wikipedia. **X-LRL** = Extremely low-resource language, **LRL** = low-resource language, and **MRL** = medium-resource language.

2020; Winata et al., 2023), while paragraph writing (Koto et al., 2022a) is explored to capture a more culturally relevant aspect which is often left out in translation (Kirkpatrick and van Teijlingen, 2009). The details of our expert annotator recruitment are shown in Appendix B. In the following section, we describe how the data construction is done for both methods.

3.1 Sentence Translation

Data Selection We sample data from two sources, i.e., IndoLEM sentiment (Koto and Rahmaningtyas, 2017; Koto et al., 2020), an Indonesian sentiment analysis dataset collected from Twitter and hotel review, and EmoT (Saputri et al., 2018; Wilie et al., 2020), an Indonesian emotion classification dataset collected from Twitter. We take the whole samples from both IndoLEM sentiment (5048 samples) and EmoT (4401 samples) as our source language data, resulting in a total data of 9,449 sentences for translation.

Translation Procedure We translate the source language data into 11 languages: Ambon (abs), Batak (btk), Betawi (bew), Bima (bhp), Javanese (jav), Madurese (mad), Makassarese (mak), Minangkabau (min), Musi (mui), Rejang (rej), and Sundanese (sun). Our expert annotators are instructed to translate while maintaining: (1) the sentence's sentiment/emotion polarity; (2) the named entities; and (3) the completeness of the text. The translation procedure is detailed in Appendix C.

3.2 Paragraph Writing

We conduct paragraph writing by instructing the annotators to write a 100-word paragraph given

a certain topic. The topic for paragraph writing is manually designed to cover a wide coverage of domains. We conduct paragraph writing in 10 languages, i.e., Batak (btk), Betawi (bew), Buginese (bug), Javanese (jav), Madurese (mad), Makassarese (mak), Minangkabau (min), Musi (mui), Rejang (rej), and Sundanese (sun). Note that, unlike sentence translation, there is no Ambon (abs) and Bima (bhp), but instead there is an additional language, Buginese (bug). This happens because of the difficulty of obtaining a large pool of annotators for the Ambon (abs) and Bima (bhp) languages.

Topic Selection We provide a list of topics before instructing the annotators to write paragraphs. The selection of various topics is expected to enrich vocabulary in the corpus as different topics will obviously bring up the use of different terms. The topics provided vary widely, ranging from food and beverages, entertainment/leisure, sports, science, history, politics, and religion. In addition, we also have other topics for describing emotional states such as sadness, happiness, anger, etc. We also provide more specific subtopics for each of the major topics provided. In total, we have 20 main topics with 20 subtopics of each main topic. The list of all topics is given in Appendix D.

Paragraph Writing Procedure The paragraph writing is done with the following criteria: (a) the paragraph consists of a minimum of 100 words, (b) using only the targeted local language except for named entities, (c) the content of the paragraph should be about the provided topics and subtopics, and (d) for each paragraph, the annotator should fill the rhetoric type of the paragraph, which is either narration, description, argumentation, persuasion, or exposition. More details about the paragraph writing procedure are in Appendix E.

3.3 Quality Control

Quality control is conducted to ensure the data are correct through manual and automatic validation. If the data does not meet the desired criteria, it is to be revised. Specifically, through a series of manual and automatic validations, we ensure that all sentences that need to be translated are translated to the target language, with minimal overlap with the source language sentence. For paragraph writing, we ensure that there is no plagiarism from external sources by conducting validation through search engines and we also ensure that there is a minimum 30% distinction between two paragraphs

(measured by using edit distance). The detail of our quality control process is described in Appendix F. The quality control is conducted in several iterations, by asking annotators to rewrite unqualified instances until all quality control passes.

3.4 Resulting Corpora

Through sentence translation, we achieve a total of 72,444 sentences, with 1,579 for Bima; 1,574 each for Musi, Rejang, and Ambon; and 9,449 each for Madurese, Minangkabau, Batak, Betawi, Javanese, Sundanese, and Makassarese. As for paragraph writing, we achieve a total of 57,409 paragraphs, specifically: 5,211 for Maduranese, 8,608 for Minangkabau, 10,188 for Javanese, 9,594 for Sundanese, 9,755 for Betawi, 4,908 for Batak, 5,471 for Makassar, 1,200 for Rejang, 1,474 for Musi, and 1,000 for Buginese. We develop two corpora grouped according to how the data is constructed: **NusaTranslation** and **NusaParagraph**. We use the NusaTranslation and NusaParagraph corpora to build NusaWrites, an underrepresented language benchmark for 12 Indonesian local languages.

4 NusaWrites over Wikipedia

In this section, we compare the quality of various corpus collection methods, i.e., online scraping from Wikipedia, sentence translation (NusaTranslation), and paragraph writing (NusaParagraph), through 5 Indonesian local languages: Buginese (bug), Javanese (jav), Madurese (mad), Minangkabau (min), and Sundanese (sun).

The statistics of each corpus collection method as shown in Appendix G. In general, Wikipedia has a larger token count and unique token coverage for Javanese (jav), Sundanese (sun), and Minangkabau (min). While for Madurese (mad), the corpus size in Wikipedia is very small with only 110k tokens, in this case, the sentence translation and paragraph writing methods provide a huge advantage over collecting through Wikipedia. Interestingly, while the #tokens of the Buginese (bug) in Wikipedia are rather large, the #unique tokens are very small even compared to the smaller data from Madurese (mad). Additionally, the #tokens/document is pretty small, indicating a short document per Wikipedia article. These facts show that the data for Buginese (bug) in Wikipedia comprises many short boilerplate texts, which are not useful for learning the language.

As the data size for each corpus collection method is different, to further compare the cor-
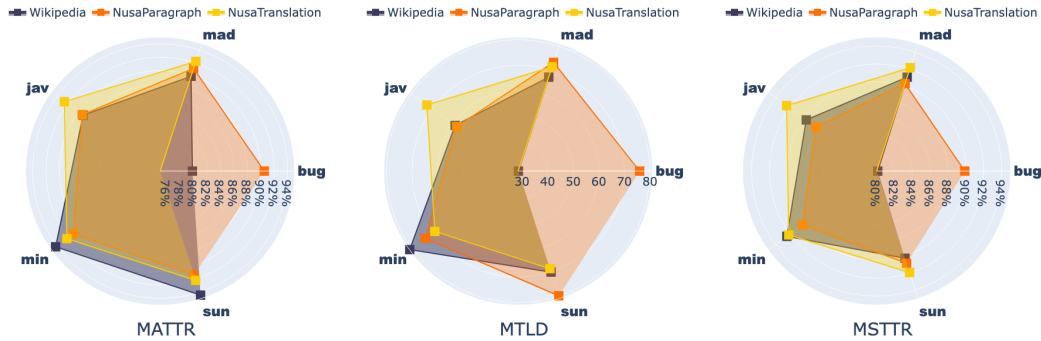
Figure 3: The **(left)** MATTR, **(center)** MTLD, and **(right)** MSTTR scores of different corpus collection methods. Paragraph writing and translation achieve higher diversity on the extremely low-resource languages, i.e., Madurese (mad) and Buginese (bug), compared to scraping from Wikipedia.

pus quality generated by each corpus collection method, we compare three criteria that are less prone to the size of the data, i.e., length-agnostic lexical diversity metrics; the empirical language modeling quality from LMs trained on the generated corpus on hold-out text data, NusaX (Winata et al., 2023), a human-translated Indonesian social media posts and online reviews; and the ratio of borrowed words of the generated corpus.

4.1 Lexical Diversity

To measure the lexical diversity, we measure the length-agnostic lexical diversity metrics, i.e., MATTR (Covington and McFall, 2010), MTLD (McCarthy, 2005), and MSTTR (Johnson, 1944), for each corpus collection method in Figure 3.[10] For MTLD we use a threshold of 0.72, while for MATTR and MSTTR, we use a window size of 20.[11] For low-resource languages, i.e., Javanese (jav) and Sundanese (sun), all three methods produce an almost equally diverse corpus, with a slightly higher diversity for sentence translation. For extremely low-resource languages, compared to other methods, Wikipedia achieves slightly higher diversity scores in Minangkabau (min), and NusaTranslation achieves slightly higher scores in Madurese (mad). Nonetheless, by utilizing permutation test ($n = 1,000$) (Koplenig, 2019), we conclude that the difference between corpora in all metrics and languages are statistically significant ($p < 0.05$), except for Madurese (mad) between NusaTranslation and NusaParagraph on the MATTR and MTLD metrics, and for Sundanese

(sun) between NusaTranslation and NusaParagraph on the MATTR and MSTTR metrics. Interestingly, for Buginese (bug), Wikipedia achieves very low score diversity scores, while NusaParagraph achieves high diversity scores, which shows that there are a large number of sentences in the Wikipedia data for Buginese (bug) that have a repeating pattern like boilerplate. In addition to the diversity metrics, we also measure the lexical overlapping with Indonesian and English lexicons obtained from Panlex (Kamholz et al., 2014). As shown in Figure 1, Wikipedia has a higher overlap with the English lexicon, indicating that it covers many shared foreign terms (e.g., scientific terms) and foreign entities (e.g., the name of cities, tourist attractions, etc.), which are not common in the actual day-to-day use of Indonesian local languages where the languages are commonly used for daily conversation, instead of a formal occasion, such as in the academic setting (Cohn and Ravindranath, 2014; Soeparno, 2015; Nurjanah et al., 2018; Nur, 2018; Sutrisno and Ariesta, 2019).

4.2 Language Modeling Quality

To evaluate the quality of the generated corpus, we evaluate the LM trained on each of the corpus generated by each method. Specifically, we build a small two-layer decoder-only transformer with 128 hidden dimensions and total parameters of ~5.5M, which is a comparable size to a BERT-Tiny model (Devlin et al., 2019) using two different settings: 1) using the same amount of tokens for each corpus by sampling the larger-sized corpora (**balanced**) and 2) using the original corpus size for each collection method (**full**).[12] The first one shows the expected quality of the sentences in the

---

[10]We zero out the Buginese (bug) statistics for the sentence translation as we do not collect Buginese (bug) data in the sentence translation.

[11]According to (Lembersky et al., 2013), lexical diversity through TTR might be inherently different for translated texts which might affect the result from NusaTranslation.

[12]All models are trained using the IndoGPT tokenizer (https://huggingface.co/indobenchmark/indogpt).
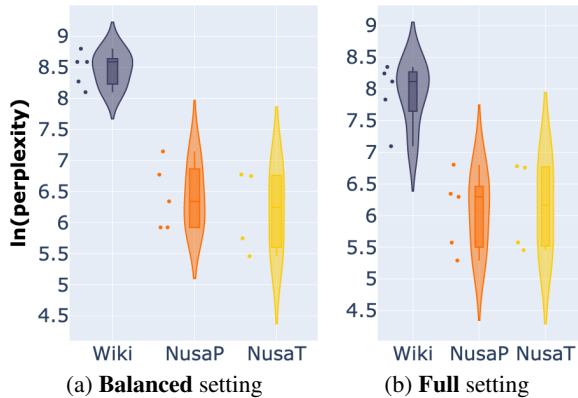
(a) **Balanced** setting (b) **Full** setting

Figure 4: LMs perplexity evaluation of different corpus collection methods. Lower is better. Wiki: Wikipedia, NusaP: NusaParagraph, NusaT: NusaTranslation.
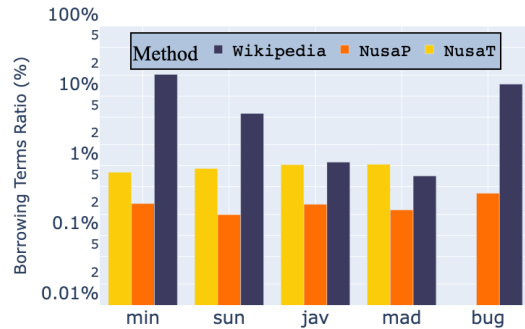

Figure 5: Ratio of loan words per language of different corpus collection methods. Wiki: Wikipedia, NusaP: NusaParagraph, NusaT: NusaTranslation. The ratio is presented in $\log_{10}$ basis.

corpora, while the second one shows the expected empirical performance when utilizing the corpus.

The LM perplexity of the three corpus collection methods is shown in Figure 4a. In general, the performance of LMs from NusaTranslation and NusaParagraph is much lower than the one from Wikipedia, showing that the corpora are more aligned with the colloquial writing of Indonesian local languages which is the common use case of using these languages (Cohn and Ravindranath, 2014; Farisiyah and Zamzani, 2018; Soeparno, 2015; Nurjanah et al., 2018; Nur, 2018; Sutrisno and Ariesta, 2019; Aji et al., 2022b). For the **balanced** setting, we observe that LMs from NusaTranslation produce slightly better results than the LMs from NusaParagraph. This is expected as the source domain of NusaTranslation is more similar to NusaX (Winata et al., 2023), which also covers social media content and online reviews. Nevertheless, as shown in the results from the **full** setting (see Figure 4b), this problem can be alleviated by increasing the coverage of the corpus.

4.3 Loan Words Ratio

To assess the cultural relevance of the generated corpora, we evaluate the ratio of loan words present within each corpus. The loan words are manually curated from the top 200 words that overlap with the English lexicon and an additional list of English loan words[13] in each corpus.[14] The complete list of

loan words is in Appendix H. The ratio is calculated by dividing the number of loan words by the total number of tokens in each corpus, and the results are presented in Figure 5. The findings indicate that NusaParagraph and NusaTranslation exhibit a minimal ratio of loan words, with approximately ∼0.1% and ∼1% respectively. However, some languages in Wikipedia, such as Minangkabau (min), Sundanese (sun), and Buginese (bug), demonstrate significantly higher ratios of loan words, ranging from approximately 5% to 15%. Additionally, in Appendix I, we demonstrate that NusaParagraph and NusaTranslation possess a notably higher ratio of common local words, including terms like `indomie` and `angkot`, in comparison to Wikipedia. These results emphasize the superiority of manually curated methods, particularly paragraph writing, in generating culturally relevant corpora.

5 NusaWrites Benchmark

From our resulting corpora in §3.4, we build the NusaWrites benchmark, which consists of 12 Indonesian local languages: Ambon (abs), Batak (btk), Betawi (bew), Bima (bhp), Buginese (bug), Javanese (jav), Madurese (mad), Makassarese (mak), Minangkabau (min), Palembang / Musi (mui), Rejang (rej), and Sundanese (sun). More details of each language are in Appendix J. 4 languages under study, i.e., Ambon (abs), Bima (bhp), Musi (mui), and Rejang (rej), have a population of <1M speakers, while others have a population of >2M speakers, but are underrepresented in NLP research (van Esch et al., 2022; Aji et al., 2022b). The languages

---

[13]The English loan words for local languages are commonly shared with the English loan words in Indonesian. The list of English loan words is collected from https://id.wiktionary.org/wiki/Wikikamus:ProyekWiki_bahasa_Indonesia/Daftar_kata/Serapan/Inggris.

[14]We use the English lexicon instead of Indonesian because decoupling the word borrowing from Indonesian is impractical

due to the relatively high terms overlapping coming from the shared geopolitical landscape and cultural values.

Models	NusaParagraph			NusaTranslation	
	Emot	Rhetorical Mode	Topic	Emot	Senti
Classical					
Logistic Regression	**78.23**	45.21	**87.67**	**56.18**	74.89
Naive Bayes	75.51	37.73	85.06	52.70	74.89
SVM	76.36	**45.44**	85.86	55.08	**76.04**
Fine-tuning					
IndoLEM IndoBERT$_{BASE}$	66.94	**51.93**	84.87	52.59	69.08
IndoNLU IndoBERT$_{BASE}$	67.12	47.92	**85.87**	54.50	75.24
IndoNLU IndoBERT$_{LARGE}$	62.65	31.75	85.41	**57.80**	77.40
mBERT	63.15	50.01	73.82	44.13	68.72
XLM-R$_{BASE}$	59.15	49.17	71.68	47.02	68.62
XLM-R$_{LARGE}$	**67.42**	51.57	83.05	54.84	**79.06**
Zero-shot					
BLOOMZ-560M	6.57	11.60	4.98	14.27	46.22
BLOOMZ-1.1B	11.72	**12.76**	5.28	13.64	60.64
BLOOMZ-1.7B	8.56	9.92	12.73	11.77	**65.10**
BLOOMZ-3B	8.54	12.35	13.55	16.03	62.23
BLOOMZ-7.1B	13.87	10.04	11.16	11.05	57.84
mT0$_{SMALL}$	9.35	8.61	32.04	15.33	31.92
mT0$_{BASE}$	13.76	7.70	35.35	23.56	27.70
mT0$_{LARGE}$	12.18	7.50	31.00	21.91	35.25
mT0$_{XL}$	**21.97**	8.70	31.76	**30.36**	40.11
mT0$_{XXL}$	19.08	8.48	**40.22**	23.49	35.44

(a) NLU evaluation results of NusaWrites.

Models	SacreBLEU	ChrF++
Classical		
Copy	23.49	41.90
Word Substitution	**23.80**	42.68
PBSMT	25.00	**56.60**
Fine-tuning		
IndoBART	**30.88**	**51.09**
IndoGPT	27.36	49.25
mBART-50	23.40	40.32
mT5	26.16	46.84
Zero-shot		
BLOOMZ-560M	3.14	18.90
BLOOMZ-1.1B	2.12	16.36
BLOOMZ-1.7B	4.68	21.70
BLOOMZ-3B	5.70	**24.34**
BLOOMZ-7.1B	3.65	12.42
mT0$_{SMALL}$	2.35	11.82
mT0$_{BASE}$	3.14	13.28
mT0$_{LARGE}$	2.39	11.29
mT0$_{XL}$	4.22	16.13
mT0$_{XXL}$	**6.33**	16.15

(b) NLG evaluation results of NusaWrites.

Table 2: Overall performance on all tasks in the NusaWrites benchmark. We report the macro-F1 (%) for NLU, and SacreBLEU and ChrF++ for NLG, averaged over all of the languages within the tasks. The best performances in each section are **bolded**, while the best overall performance in each column is underlined.

belong to the Austronesian language family under the Malayo-Polynesian subgroup. While some of the languages are written in multiple scripts, we use the Latin script in NusaWrites, which has become predominant for all covered languages.

5.1 NusaTranslation

We develop three parallel downstream tasks—sentiment analysis, emotion recognition, and machine translation—covering 11 local languages spoken in Indonesia. We generate a new split for each downstream task and keep a reasonable amount of test samples for languages with smaller sample sizes. The labels of the downstream tasks follow the original label from the original dataset. The statistics of each downstream task are shown in Table 11. A detailed description of each downstream task is provided in Appendix K.

5.2 NusaParagraph

We develop three downstream tasks from NusaParagraph—topic modeling, emotion recognition, and rhetoric mode classification—based on the datasets covering 10 local languages spoken in Indonesia. For the topic modeling task, we cover 8 topics: food & beverages, sports, leisure, religion, culture & heritage, a slice of life, technology,

and business. For the emotion recognition task, we cover the 6 basic emotions (Ekman, 1992): fear, disgusted, sad, happy, angry, and surprise, and an additional emotion label: shame (Poulson and of Tasmania. School of Management, 2000). For the rhetoric mode classification, we cover 5 rhetoric modes: narrative, persuasive, argumentative, descriptive, and expository. The statistics of the corpus and the detailed description of each task are shown in Table 12 and Appendix L.

5.3 Baselines

Classical Machine Learning In extremely low-resource settings, the classical approaches can outperform the neural approach, especially if there is no pre-trained model supporting that particular language (Winata et al., 2023). Moreover, with the limited computational access in many regions such as Indonesia, classical machine learning remains a popular choice for researchers and industry (Nityasya et al., 2020; Aji et al., 2022a). Hence, we utilize this approach for NusaWrites.

For NLU tasks, we employ three classical machine learning methods as our baselines, namely (1) Naive Bayes (Zhang, 2004), (2) Logistic Regression (Cramer, 2003), and (3) SVM (Scholkopf et al., 1995). For NLG tasks, we harness three meth-

	abs	bew	bhp	btk	bug	jav	mad	mak	min	mui	rej	sun
XR-L	61.1	80.5	37	71.8	47.5	84.9	72.6	67	82.1	72.7	44.7	81.4
XR-B	51.5	77.9	27.2	68.2	42.4	82.2	70.8	63.2	78.4	36.2	32	77.8
mB	46.8	75.4	27.3	67.9	46.7	80.4	70	64	77.6	57.8	28.1	74
IB-B2	47.2	79	35	69.4	54.7	82.4	70.6	66	78.8	64.7	52.6	74.1
IB-L	60.3	73.8	44.2	67.6	33.8	83	69.9	67.3	79.5	61.5	45.1	76.6
IB-B1	55.7	78.2	47.5	69.7	51	82	71.6	67.7	79.6	66.5	53.9	73.3
SVM	58.2	75.9	57.8	72.8	62	79.7	74	69.9	78.7	64.6	58	75.7
NB	58.9	72.7	55.5	69.6	58.2	76.6	70.8	69.3	73.8	63.3	58.4	72.2
LR	59.2	76	51.7	73.9	63.7	80.4	74.8	70.6	79.5	65.5	59.8	76.5

Figure 6: Per language scores of the classical and fine-tuned baselines. From top to bottom: XLM-R$_{LARGE}$, XLM-R$_{BASE}$, mBERT, IndoLEM IndoBERT$_{BASE}$, IndoNLU IndoBERT$_{LARGE}$, IndoNLU IndoBERT$_{BASE}$, SVM, Naive Bayes, Logistic Regression.

ods to do benchmarking on machine translation tasks; (1) direct copy from the source language, in this case, Indonesian; (2) word lexical substitution via bilingual Panlex lexicons; and (3) phrase-based statistical machine translation (PBSMT) (Koehn et al., 2003). We employ the PBSMT method from Moses toolkit (Koehn et al., 2007).

Massively Multilingual LMs Fine-tuning LMs for downstream tasks has become a popular method in NLP. It enables LMs to learn with a limited dataset and perform better compared to training neural models from scratch (Devlin et al., 2019; Wilie et al., 2020; Gehrmann et al., 2022).

Moreover, recent work has shown that a fine-tuned model for a specific task can outperform general-purpose, larger language models (Bang et al., 2023; Asai et al., 2023; Zhang et al., 2023). We investigate the performances of both large pretrained multilingual and Indonesian monolingual baseline models on low-resource languages used in this work. We follow the hyperparameter settings in (Winata et al., 2023). Details are in Appendix M.

For NLU tasks, we experiment with emotion recognition, sentiment analysis, topic modeling, and rhetoric mode classification. The models used are: (1) mBERT (Devlin et al., 2019); (2) IndoNLU (Wilie et al., 2020); (3) IndoLEM (Koto et al., 2020); and (4) XLM-R (Conneau et al., 2020). For NLG tasks, we experiment on machine translation using the following baselines: (1) IndoGPT (Cahyawijaya et al., 2021); (2) IndoBART (Cahyawijaya et al., 2021); (3) mBART (Liu et al., 2020); and (4) mT5 (Xue et al., 2021).

Zero-Shot LLMs LLMs fine-tuned through diverse instructions show capabilities to generalize across unseen instructions (Wei et al., 2021; Sanh et al., 2021; Ouyang et al., 2022; Yong et al., 2023). Moreover, these models are shown to be able to generalize across different languages, assuming the base model is multilingual (Muennighoff et al., 2022; Adilazuarda et al., 2023; Zhang et al., 2023). Therefore, to assess the zero-shot capabilities of LLMs over our datasets, we benchmark BLOOMZ and mT0 (Muennighoff et al., 2022), both of which are multilingual LLMs that have been fine-tuned with downstream task instructions. We explore the model from 300M up to ∼13B parameters. For NLU, the class output is determined by selecting the most probable label generated after the prompt. For NLG, we generate the translation by using prompts. The prompts used in this experiment can be found in Appendix N.

6 Benchmark Results and Discussion

We present the results of our NLU and NLG experiments in Table 2a and Table 2b, respectively. While the classical baselines have never learned any prior language representations, they perform competitively to the fine-tuning baselines—the fine-tuned Indonesian monolingual models (i.e., IndoBERT, IndoBART, and IndoGPT) and the fine-tuned multilingual models (i.e., mBERT, XLM-R, mBART-50, and mT5)—on both NLU and NLG benchmarks. Furthermore, based on the per language breakdown shown in Figure 6, except for the languages observed during the pre-training, i.e., Javanese (jav) and Sundanese (sun), both Indonesian and multilingual LMs fail to outperform the classical machine learning approaches on most languages and only able to outperform on languages that are closely related to Indonesian (see Appendix J), i.e., Betawi (bew) and Minangkabau (min). These facts demonstrate that most extremely low-resource languages in NusaParagraph and NusaTranslation are beyond the scope of the knowledge transfer from Indonesian and multilingual pre-training due to their distinct linguistic characteristics.

Secondly, the LLMs used in this study: BLOOMZ and mT0, consistently and significantly underperform the fine-tuned and classical baselines, e.g., up to ∼56% gap on emotion recognition and ∼47% on topic modeling in NusaParagraph, as well as ∼17.5 SacreBLEU on machine translation. Despite their ability to generalize to unseen

tasks (Muennighoff et al., 2022), LLMs do not generalize well to unseen languages, which indicates a challenge on knowledge transferability between languages, especially for underrepresented and extremely low-resource language, and underlines the need for more language-inclusive LLMs.

7 Conclusion

In this work, we compare the effectiveness of corpus collection methods for underrepresented and extremely low-resource languages. From our thorough study, we conclude that, although online scraping is effective for high-resource languages, it is not ideal for many extremely low-resource languages. Other approaches such as sentence translation and paragraph writing can be a better alternative for collecting data in extremely low-resource languages because they produce a better corpus with higher lexical diversity and cultural relevance. Furthermore, to measure the capability of existing LLMs to process underrepresented and extremely low-resource languages, we propose the NusaWrites benchmark, which covers 12 Indonesian local languages. Based on the benchmarking results, we demonstrate that both existing zero-shot prompting LLMs and fine-tuned pre-trained LMs fail to outperform the classical baselines, suggesting that LMs cannot generalize to these extremely low-resource languages as most of the extremely low-resource languages under study are distinct from other previously learned languages. Our empirical experiments emphasize the need to extend the language coverage of the models.

Limitations

7.1 Languages for Comparison of Corpus Collection Methods

We explore only 5 Indonesian local languages to compare the effectiveness of different corpus collection methods due to the difficulty of finding eligible annotator candidates for the other languages. We hope future work can explore the generalization of our analysis in broader languages, especially for other underrepresented and extremely low-resource languages in different language families.

7.2 Buginese Data for NusaTranslation

We do not have Buginese data in our NusaTranslation corpus, this is due to the difficulty of finding eligible annotator candidates for Buginese. In fact, during our course of dataset construction, we only found one eligible annotator candidate who would like to participate in our study.

7.3 Few-Shot LLM Prompting

Few-shot in-context learning has been shown to be able to improve the performance of zero-shot prompting (Brown et al., 2020; Sanh et al., 2022; Wei et al., 2022; Chung et al., 2022). However, few-shot in-context learning incurs a high computational cost and, due to a limited computational budget, we only explore zero-shot LLM prompting and we leave the exploration on few-shot in-context learning for future works.

Ethical Consideration

Our work highlights the importance of democratizing access to Natural Language Processing (NLP) technology for underrepresented and extremely low-resource languages. During our study, we are well aware of the ethical responsibility associated with language research and the potential impact it can have on communities. Our study prioritizes inclusivity, cultural relevance, and fairness. Within this work, the annotators are properly rewarded above the national average minimum wage in Indonesia. We have obtained informed consent from all annotators and adhered to data protection and privacy regulations for releasing the corpus and benchmark. Throughout our research process, we have made conscious efforts to engage with the language communities, involve local experts, and respect their linguistic and cultural nuances. Our ultimate goal is to promote linguistic diversity and contribute to a more inclusive NLP landscape providing social good through our work to society, especially in the field of NLP. We encourage further collaboration and engagement with underrepresented language communities to ensure that their voices are heard and their needs are addressed in future language technology development.

References

David Ifeoluwa Adelani, Jade Abbott, Graham Neubig, Daniel D'souza, Julia Kreutzer, Constantine Lignos, Chester Palen-Michel, Happy Buzaaba, Shruti Rijhwani, Sebastian Ruder, Stephen Mayhew, Israel Abebe Azime, Shamsuddeen H. Muhammad, Chris Chinenye Emezue, Joyce Nakatumba-Nabende, Perez Ogayo, Aremu Anuoluwapo, Catherine Gitau, Derguene Mbaye, Jesujoba Alabi, Seid Muhie Yimam, Tajuddeen Rabiu Gwadabe, Ignatius Ezeani, Rubungo Andre Niyongabo, Jonathan Mukiibi, Verrah Otiende, Iroro Orife, Davis David, Samba Ngom, Tosin Adewumi, Paul Rayson, Mofetoluwa Adeyemi, Gerald Muriuki, Emmanuel Anebi, Chiamaka Chukwuneke, Nkiruka Odu, Eric Peter Wairagala, Samuel Oyerinde, Clemencia Siro, Tobius Saul Bateesa, Temilola Oloyede, Yvonne Wambui, Victor Akinode, Deborah Nabagereka, Maurice Katusiime, Ayodele Awokoya, Mouhamadane MBOUP, Dibora Gebreyohannes, Henok Tilaye, Kelechi Nwaike, Degaga Wolde, Abdoulaye Faye, Blessing Sibanda, Orevaoghene Ahia, Bonaventure F. P. Dossou, Kelechi Ogueji, Thierno Ibrahima DIOP, Abdoulaye Diallo, Adewale Akinfaderin, Tendai Marengereke, and Salomey Osei. 2021a. MasakhaNER: Named Entity Recognition for African Languages. *Transactions of the Association for Computational Linguistics*, 9:1116–1131.

David Ifeoluwa Adelani, Jade Abbott, Graham Neubig, Daniel D'souza, Julia Kreutzer, Constantine Lignos, Chester Palen-Michel, Happy Buzaaba, Shruti Rijhwani, Sebastian Ruder, Stephen Mayhew, Israel Abebe Azime, Shamsuddeen H. Muhammad, Chris Chinenye Emezue, Joyce Nakatumba-Nabende, Perez Ogayo, Aremu Anuoluwapo, Catherine Gitau, Derguene Mbaye, Jesujoba Alabi, Seid Muhie Yimam, Tajuddeen Rabiu Gwadabe, Ignatius Ezeani, Rubungo Andre Niyongabo, Jonathan Mukiibi, Verrah Otiende, Iroro Orife, Davis David, Samba Ngom, Tosin Adewumi, Paul Rayson, Mofetoluwa Adeyemi, Gerald Muriuki, Emmanuel Anebi, Chiamaka Chukwuneke, Nkiruka Odu, Eric Peter Wairagala, Samuel Oyerinde, Clemencia Siro, Tobius Saul Bateesa, Temilola Oloyede, Yvonne Wambui, Victor Akinode, Deborah Nabagereka, Maurice Katusiime, Ayodele Awokoya, Mouhamadane MBOUP, Dibora Gebreyohannes, Henok Tilaye, Kelechi Nwaike, Degaga Wolde, Abdoulaye Faye, Blessing Sibanda, Orevaoghene Ahia, Bonaventure F. P. Dossou, Kelechi Ogueji, Thierno Ibrahima DIOP, Abdoulaye Diallo, Adewale Akinfaderin, Tendai Marengereke, and Salomey Osei. 2021b. MasakhaNER: Named Entity Recognition for African Languages. *Transactions of the Association for Computational Linguistics*, 9:1116–1131.

Muhammad Farid Adilazuarda, Samuel Cahyawijaya, and Ayu Purwarianti. 2023. The obscure limitation of modular multilingual language models.

Muhammad Farid Adilazuarda, Samuel Cahyawijaya, Genta Indra Winata, Pascale Fung, and Ayu Purwarianti. 2022. IndoRobusta: Towards robustness against diverse code-mixed Indonesian local languages. In *Proceedings of the First Workshop on Scaling Up Multilingual Evaluation*, pages 25–34, Online. Association for Computational Linguistics.

Alham Aji, Genta Indra Winata, Fajri Koto, Samuel Cahyawijaya, Ade Romadhony, Rahmad Mahendra, Kemal Kurniawan, David Moeljadi, Radityo Eko Prasojo, Timothy Baldwin, et al. 2022a. One country, 700+ languages: Nlp challenges for underrepresented languages and dialects in indonesia. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7226–7249.

Alham Fikri Aji, Genta Indra Winata, Fajri Koto, Samuel Cahyawijaya, Ade Romadhony, Rahmad Mahendra, Kemal Kurniawan, David Moeljadi, Radityo Eko Prasojo, Timothy Baldwin, Jey Han Lau, and Sebastian Ruder. 2022b. One country, 700+ languages: NLP challenges for underrepresented languages and dialects in Indonesia. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7226–7249, Dublin, Ireland. Association for Computational Linguistics.

Akari Asai, Sneha Kudugunta, Xinyan Velocity Yu, Terra Blevins, Hila Gonen, Machel Reid, Yulia Tsvetkov, Sebastian Ruder, and Hannaneh Hajishirzi. 2023. Buffet: Benchmarking large language models for few-shot cross-lingual transfer.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity.

Robert Blust et al. 2013. *The Austronesian Languages*. The Australian National University.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Samuel Cahyawijaya, Holy Lovenia, Alham Fikri Aji, Genta Indra Winata, Bryan Wilie, Rahmad Mahendra, Christian Wibisono, Ade Romadhony, Karissa Vincentio, Fajri Koto, et al. 2022. Nusacrowd: Open source initiative for indonesian nlp resources. *arXiv preprint arXiv:2212.09648*.

Samuel Cahyawijaya, Genta Indra Winata, Bryan Wilie, Karissa Vincentio, Xiaohong Li, Adhiguna Kuncoro, Sebastian Ruder, Zhi Yuan Lim, Syafri Bahar, Masayu Khodra, et al. 2021. Indonlg: Benchmark and resources for evaluating indonesian natural language generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8875–8898.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models.

Abigail C. Cohn and Maya Ravindranath. 2014. LOCAL LANGUAGES IN INDONESIA: LANGUAGE MAINTENANCE OR LANGUAGE SHIFT? *Linguistik Indonesia*, 32(2):131–148.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Alexis Conneau, Guillaume Lample, Ruty Rinott, Adina Williams, Samuel R Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. Xnli: Evaluating cross-lingual sentence representations. *arXiv preprint arXiv:1809.05053*.

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Michael A. Covington and Joe D. McFall. 2010. Cutting the gordian knot: The moving-average type–token ratio (MATTR). *Journal of Quantitative Linguistics*, 17(2):94–100.

J.S. Cramer. 2003. The origins of logistic regression. *SSRN Electronic Journal*.

Sophie Elizabeth Crouch. 2009. *Voice and verb morphology in Minangkabau, a language of West Sumatra, Indonesia*. Ph.D. thesis, The University of Western Australia.

William D Davies. 2010. *A grammar of Madurese*. Mouton De Gruyter.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

David M. Eberhard, Gary F. Simons, and Charles D. Fennig. 2021. *Ethnologue: Languages of the World. Twenty-fourth edition*. Dallas, Texas: SIL International.

Abteen Ebrahimi, Manuel Mager, Arturo Oncevay, Vishrav Chaudhary, Luis Chiruzzo, Angela Fan, John Ortega, Ricardo Ramos, Annette Rios Gonzales, Ivan Meza-Ruiz, et al. 2022. Americasnli: Evaluating zero-shot natural language understanding of pretrained multilingual models in truly low-resource languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6279–6299.

Paul Ekman. 1992. Facial expressions of emotion: New findings, new questions. *Psychological Science*, 3(1):34–38.

Umi Farisiyah and Zamzani Zamzani. 2018. Languange shift and language maintenance of local languages toward indonesian. In *Proceedings of the International Conference of Communication Science Research (ICCSR 2018)*. Atlantis Press.

Sebastian Gehrmann, Abhik Bhattacharjee, Abinaya Mahendiran, Alex Wang, Alexandros Papangelis, Aman Madaan, Angelina Mcmillan-major, Anna Shvets, Ashish Upadhyay, Bernd Bohnet, Bingsheng Yao, Bryan Wilie, Chandra Bhagavatula, Chaobin You, Craig Thomson, Cristina Garbacea, Dakuo Wang, Daniel Deutsch, Deyi Xiong, Di Jin, Dimitra Gkatzia, Dragomir Radev, Elizabeth Clark, Esin Durmus, Faisal Ladhak, Filip Ginter, Genta Indra Winata, Hendrik Strobelt, Hiroaki Hayashi, Jekaterina Novikova, Jenna Kanerva, Jenny Chim, Jiawei Zhou, Jordan Clive, Joshua Maynez, João Sedoc, Juraj Juraska, Kaustubh Dhole, Khyathi Raghavi Chandu, Laura Perez Beltrachini, Leonardo F . R. Ribeiro, Lewis Tunstall, Li Zhang, Mahim Pushkarna, Mathias Creutz, Michael White, Mihir Sanjay Kale, Moussa Kamal Eddine, Nico Daheim, Nishant Subramani, Ondrej Dusek, Paul Pu Liang, Pawan Sasanka Ammanamanchi, Qi Zhu, Ratish Puduppully, Reno Kriz, Rifat Shahriyar, Ronald Cardenas, Saad Mahamood, Salomey Osei, Samuel Cahyawijaya, Sanja Štajner, Sebastien Montella, Shailza Jolly, Simon Mille, Tahmid Hasan, Tianhao Shen, Tosin Adewumi, Vikas Raunak, Vipul Raheja, Vitaly Nikolaev, Vivian Tsai, Yacine Jernite, Ying Xu, Yisi Sang, Yixin Liu, and Yufang Hou. 2022. GEMv2: Multilingual NLG benchmarking in a single line of code. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 266–281, Abu Dhabi, UAE. Association for Computational Linguistics.

Francisco Guzmán, Peng-Jen Chen, Myle Ott, Juan Pino, Guillaume Lample, Philipp Koehn, Vishrav Chaudhary, and Marc'Aurelio Ranzato. 2019. The FLORES evaluation datasets for low-resource machine translation: Nepali–English and Sinhala–English. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6098–6111, Hong Kong, China. Association for Computational Linguistics.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. XTREME: A Massively Multilingual Multi-task Benchmark for Evaluating Cross-lingual Generalization. In *Proceedings of ICML 2020*.

Wendell Johnson. 1944. I. a program of research. *Psychological Monographs*, 56(2):1–15.

Anubha Kabra, Emmy Liu, Simran Khanuja, Alham Fikri Aji, Genta Winata, Samuel Cahyawijaya, Anuoluwapo Aremu, Perez Ogayo, and Graham Neubig. 2023. Multi-lingual and multi-cultural figurative language understanding. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8269–8284, Toronto, Canada. Association for Computational Linguistics.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul N.C., Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. 2020. IndicNLPSuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for Indian languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4948–4961, Online. Association for Computational Linguistics.

David Kamholz, Jonathan Pool, and Susan Colowick. 2014. PanLex: Building a resource for panlingual lexical translation. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3145–3150, Reykjavik, Iceland. European Language Resources Association (ELRA).

Pamela Kirkpatrick and Edwin van Teijlingen. 2009. Lost in translation: Reflecting on a model to reduce translation and interpretation bias. *The Open Nursing Journal*, 3:25–32.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the association for computational linguistics companion volume proceedings of the demo and poster sessions*, pages 177–180.

Philipp Koehn, Franz J. Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 127–133.

Alexander Koplenig. 2019. A non-parametric significance test to compare corpora. *PLOS ONE*, 14(9):e0222703.

Fajri Koto, Timothy Baldwin, and Jey Han Lau. 2022a. Cloze evaluation for deeper understanding of commonsense stories in Indonesian. In *Proceedings of the First Workshop on Commonsense Representation and Reasoning (CSRR 2022)*, pages 8–16, Dublin, Ireland. Association for Computational Linguistics.

Fajri Koto, Timothy Baldwin, and Jey Han Lau. 2022b. LipKey: A large-scale news dataset for absent keyphrases generation and abstractive summarization. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3427–3437, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Fajri Koto and Ikhwan Koto. 2020. Towards computational linguistics in Minangkabau language: Studies on sentiment analysis and machine translation. In *Proceedings of the 34th Pacific Asia Conference on Language, Information and Computation*, pages 138–148, Hanoi, Vietnam. Association for Computational Linguistics.

Fajri Koto, Jey Han Lau, and Timothy Baldwin. 2021. IndoBERTweet: A pretrained language model for Indonesian Twitter with effective domain-specific vocabulary initialization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10660–10668, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Fajri Koto, Afshin Rahimi, Jey Han Lau, and Timothy Baldwin. 2020. Indolem and indobert: A benchmark dataset and pre-trained language model for indonesian nlp. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 757–770.

Fajri Koto and Gemala Y Rahmaningtyas. 2017. Inset lexicon: Evaluation of a word list for Indonesian sentiment analysis in microblogs. In *2017 International Conference on Asian Language Processing (IALP)*, pages 391–394. IEEE.

Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, Monang Setyawan, Supheakmungkol Sarin, Sokhar Samb, Benoît Sagot, Clara Rivera, Annette Rios, Isabel Papadimitriou, Salomey Osei, Pedro Ortiz Suarez, Iroro Orife, Kelechi Ogueji, Andre Niyongabo Rubungo, Toan Q. Nguyen, Mathias Müller, André Müller, Shamsuddeen Hassan Muhammad, Nanda Muhammad, Ayanda Mnyakeni, Jamshidbek Mirzakhalov, Tapiwanashe Matangira, Colin Leong, Nze Lawson, Sneha Kudugunta,

Yacine Jernite, Mathias Jenny, Orhan Firat, Bonaventure F. P. Dossou, Sakhile Dlamini, Nisansa de Silva, Sakine Çabuk Ballı, Stella Biderman, Alessia Battisti, Ahmed Baruwa, Ankur Bapna, Pallavi Baljekar, Israel Abebe Azime, Ayodele Awokoya, Duygu Ataman, Orevaoghene Ahia, Oghenefego Ahia, Sweta Agrawal, and Mofetoluwa Adeyemi. 2022. Quality at a Glance: An Audit of Web-Crawled Multilingual Datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72.

Aman Kumar, Himani Shrotriya, Prachi Sahu, Amogh Mishra, Raj Dabre, Ratish Puduppully, Anoop Kunchukuttan, Mitesh M. Khapra, and Pratyush Kumar. 2022. IndicNLG benchmark: Multilingual datasets for diverse NLG tasks in Indic languages. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5363–5394, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Eri Kurniawan. 2013. *Sundanese complementation*. Ph.D. thesis, The University of Iowa.

Gennadi Lembersky, Noam Ordan, and Shuly Wintner. 2013. Improving statistical machine translation by adapting translation models to translationese. *Computational Linguistics*, 39(4):999–1023.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Philip M. McCarthy. 2005. *An assessment of the range and usefulness of lexical diversity measures and the potential of the measure of textual, lexical diversity (MTLD)*. Ph.D. thesis, The University of Memphis.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, et al. 2022. Crosslingual generalization through multitask finetuning. *arXiv preprint arXiv:2211.01786*.

Made Nindyatama Nityasya, Haryo Akbarianto Wibowo, Radityo Eko Prasojo, and Alham Fikri Aji. 2020. Costs to consider in adopting nlp for your business. *arXiv preprint arXiv:2012.08958*.

Apriani Nur. 2018. Realisasi uu kebahasaan dalam bidang pendidikan terwujudkah itu? *Kongres Bahasa Indonesia XI*.

Fajrin Nurjanah et al. 2018. Pengembangan kemampuan berbahasa indonesia siswa sekolah dasar desa terpencil melalui metode karyawisata berbasis potensi lokal. *FKIP e-PROCEEDING*, pages 167–176.

Odunayo Ogundepo, Tajuddeen R Gwadabe, Clara E Rivera, Jonathan H Clark, Sebastian Ruder, David Ifeoluwa Adelani, Bonaventure FP Dossou, Abdou Aziz

DIOP, Claytone Sikasote, Gilles Hacheme, et al. 2023. Afriqa: Cross-lingual open-retrieval question answering for african languages. *arXiv preprint arXiv:2305.06897*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Sungjoon Park, Jihyung Moon, Sungdong Kim, Won Ik Cho, Jiyoon Han, Jangwon Park, Chisung Song, Junseong Kim, Yongsook Song, Taehwan Oh, et al. 2021. Klue: Korean language understanding evaluation. *arXiv preprint arXiv:2105.09680*.

C. Poulson and University of Tasmania. School of Management. 2000. *Shame: The Master Emotion?* Working paper series (University of Tasmania. School of Management). School of Management, University of Tasmania.

Victor Sanh, Albert Webson, Colin Raffel, Stephen Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Fevry, Jason Alan Fries, Ryan Teehan, Teven Le Scao, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M Rush. 2022. Multitask prompted training enables zero-shot task generalization. In *International Conference on Learning Representations*.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H. Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Fevry, Jason Alan Fries, Ryan Teehan, Stella Biderman, Leo Gao, Tali Bers, Thomas Wolf, and Alexander M. Rush. 2021. Multitask prompted training enables zero-shot task generalization.

Mei Silviana Saputri, Rahmad Mahendra, and Mirna Adriani. 2018. Emotion classification on Indonesian Twitter dataset. In *2018 International Conference on Asian Language Processing (IALP)*, pages 90–95. IEEE.

Bernhard Scholkopf, Chris Burges, and Vladimir Vapnik. 1995. Extracting support data for a given task.

Association for the Advancement of Artificial Intelligence.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2021. Wiki-Matrix: Mining 135M parallel sentences in 1620 language pairs from Wikipedia. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1351–1361, Online. Association for Computational Linguistics.

Lucas Shen. 2021. Measuring political media slant using text data.

Lucas Shen. 2022. LexicalRichness: A small module to compute textual lexical richness.

Soeparno. 2015. Kerancuan fono-ortografis dan orto-fonologis bahasa Indonesia ragam lisan dan tulis. *Diksi*, 12(2).

Anders Søgaard. 2022. Should we ban English NLP for a year? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5254–5260, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yueqi Song, Catherine Cui, Simran Khanuja, Pengfei Liu, Fahim Faisal, Alissa Ostapenko, Genta Indra Winata, Alham Fikri Aji, Samuel Cahyawijaya, Yulia Tsvetkov, Antonios Anastasopoulos, and Graham Neubig. 2023. Globalbench: A benchmark for global progress in natural language processing.

Pedro Javier Ortiz Suárez, Benoît Sagot, and Laurent Romary. 2019. Asynchronous pipeline for processing huge corpora on medium to low resource infrastructures. In *7th Workshop on the Challenges in the Management of Large Corpora (CMLC-7)*. Leibniz-Institut für Deutsche Sprache.

Bejo Sutrisno and Yessika Ariesta. 2019. Beyond the use of code mixing by social media influencers in Instagram. *Advances in Language and Literary Studies*, 10(6):143–151.

Zeerak Talat, Aurélie Névéol, Stella Biderman, Miruna Clinciu, Manan Dey, Shayne Longpre, Sasha Luccioni, Maraim Masoud, Margaret Mitchell, Dragomir Radev, Shanya Sharma, Arjun Subramonian, Jaesung Tae, Samson Tan, Deepak Tunuguntla, and Oskar Van Der Wal. 2022. You reap what you sow: On the challenges of bias evaluation under multilingual settings. In *Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models*, pages 26–41, virtual+Dublin. Association for Computational Linguistics.

Daan van Esch, Tamar Lucassen, Sebastian Ruder, Isaac Caswell, and Clara E Rivera. 2022. Writing system and speaker metadata for 2,800+ language varieties. In *Proceedings of the 13th Language Resources and Evaluation Conference*, Marseille, France.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022. Finetuned language models are zero-shot learners.

Bryan Wilie, Karissa Vincentio, Genta Indra Winata, Samuel Cahyawijaya, Xiaohong Li, Zhi Yuan Lim, Sidik Soleman, Rahmad Mahendra, Pascale Fung, Syafri Bahar, and Ayu Purwarianti. 2020. IndoNLU: Benchmark and resources for evaluating Indonesian natural language understanding. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 843–857, Suzhou, China. Association for Computational Linguistics.

Genta Indra Winata, Alham Fikri Aji, Samuel Cahyawijaya, Rahmad Mahendra, Fajri Koto, Ade Romadhony, Kemal Kurniawan, David Moeljadi, Radityo Eko Prasojo, Pascale Fung, Timothy Baldwin, Jey Han Lau, Rico Sennrich, and Sebastian Ruder. 2023. NusaX: Multilingual parallel sentiment dataset for 10 Indonesian local languages. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, Dubrovnik, Croatia.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Zheng-Xin Yong, Ruochen Zhang, Jessica Zosa Forde, Skyler Wang, Arjun Subramonian, Holy Lovenia, Samuel Cahyawijaya, Genta Indra Winata, Lintang Sutawika, Jan Christian Blaise Cruz, Yin Lin Tan, Long Phan, Rowena Garcia, Thamar Solorio, and Alham Fikri Aji. 2023. Prompting multilingual large language models to generate code-mixed texts: The case of south east asian languages.

Harry Zhang. 2004. The optimality of naive bayes. In *Proceedings of the the 17th International FLAIRS conference (FLAIRS2004)*, pages 562–567.

Ruochen Zhang, Samuel Cahyawijaya, Jan Christian Blaise Cruz, and Alham Fikri Aji. 2023. Multilingual large language models are not (yet) code-switchers.

934

A Indonesian Local Languages in the Bible Corpus

We list out all the Indonesian local languages that are covered in the Bible corpus in Table 3.

Language	ISO code	#Speakers
Balantak	blz	33,000
Nggem	nbq	4,400
Alune	alp	21,300
Bambam	ptu	40,000
Yawa	yva	10,000
Dhao	nfa	5,000
Helong	heg	14,000
Kupang Malay	mkn	200,000
Mamasa	mqj	100,000
Luang	lex	18,000
Ambai	amk	10,000
Sabu	hvn	8,000
Amarasi	aaz	70,000
Kisar	kje	20,000

Table 3: Description of Indonesian local languages covered in the Bible corpus.

B Pre-Annotation Procedure

B.1 Annotator Recruitment

Our recruitment process involves multiple steps. Firstly, we conduct a strict selection process to filter out applicants. Subsequently, we proceed with knowledge transfer sessions for the selected annotators. The primary objective of our recruitment process is to identify and engage proficient annotators with expertise in relevant local languages.

Qualification In developing data for a local language, a competent and experienced team in the required local language is certainly needed. Annotators play a crucial role in compiling high-quality local language data. Therefore, strict qualifications are required for the candidate annotators who will be recruited. The qualifications include educational background and experience related to language. Annotator candidates must have good knowledge of the language and the sentence structure of the local language they are proficient in.

Recruitment Process The recruitment process starts with an assessment test comprising three questions for each task. This test is designed to provide an overview of the candidate's abilities in sentence translation and paragraph creation in

the relevant local language for future tasks. During this stage, the priority in candidate selection is based on the assessment test results, followed by employment status and educational background. From this process, we gathered a total of 892 annotator candidates from different languages. There are 29 candidates for Madurese, 141 for Minangkabau, 319 for Javanese, 217 for Sundanese, 28 for Betawi, 52 for Batak, 45 for Makassarese (including Bugisnese), and 65 for other languages (including Acehnese, Ambonese, Rejang Lebong, Sumbawanese, Papuan, Balinese, Bimanese, Cirebonese, Dayak, Leti, Lombokese, Pontianak Malay, Palembangese, and Tolaki).

Out of a total of 892 applicants, only 127 candidates ($\sim$14%) were eligible to participate in the annotation process, and among which, only 83 ($\sim$65%) candidates expressed their willingness to proceed. Some of the annotators withdraw during the course of the annotations which further increases the complexity of the recruitment process. With this obstacle, the recruitment process faces complex challenges. Finding speakers of certain local languages can be difficult, making the recruitment process long and ongoing throughout the annotation process.

Knowledge Transfer All the selected annotators will join groups and receive explanations regarding this project through knowledge transfer and overview meetings before starting their work. The information provided covers various aspects related to project management and the annotation process in detail. Annotators will gain a clear understanding of the methods and guidelines to be followed in performing these annotations. With this explanation, it is expected that the annotator will have a comprehensive understanding of their responsibilities in this work and a detailed understanding of the task. This will assist them in carrying out their task effectively and producing high-quality output.

C Sentence Translation Procedure

Human translation is carried out by determining the boundaries of the rules in the translation process. We instructed the annotators to retain the meaning of the text and to keep entities, such as persons, organizations, locations, and time with no target language translation the same. Specifically, we instructed them to: (1) maintain the sentence's sentiment polarity; (2) preserve entities; and (3) maintain the complete information content of the

original text.

Besides, we asked the annotators to maintain the typography. Most sentences from the original dataset are written in an informal tone, with non-standard spelling, e.g., elongated vowels and punctuation. When the sentence is translated into the target language, direct translation can sound unnatural. For example, translating the Indonesian word *kangeeeen* (originally *kangen*; en: *miss*) to *taragaaaak* (originally *taragak*) in Minangkabau may sound unnatural. Similarly, the original sentence may also contain typos. Due to the difficulty of accurately assessing the typographical consistency of translations, we removed this as a criterion.

The translation annotation phase is planned to last for approximately 2–6 weeks depending on the number of annotators involved in one language group. Each annotator gets around 1,000–3,000 sentences (with the same reasons as the previous explanation). Each annotator is required to complete a translation of 500 sentences per week. However, there were issues of commitment to achieving weekly targets and availability of annotators, extending the annotation process to 9 weeks.

This translation method achieved a total of 72,444 sentences. The details are: 1,579 sentences for Bima language, 1,574 sentences each for Palembang, Rejang Lebong, and Ambon languages, 9,449 sentences each for Madura language, Minangkabau language, Batak language, Betawi language, Javanese, Sundanese, and Makassar language.

D List of Topics for Paragraph Writing

Here we provide the list of topics and subtopics for the paragraph writing data collection. Each topic consists of 10 subtopics unless stated otherwise.

1. **Food & Beverages.** News about food; Restaurant reviews and fast food; Recipes; Favorite food and disliked food; Favorite drink and disliked drink; Favorite snacks and disliked snacks; Professions related to cuisine; Cooking utensils and kitchens; Vegetable and fruit; Seafood.

2. **Leisure.** Nature tourist attractions; Hidden tourist attractions (hidden gems); Popular tourist spots; Hotel and lodging; Transportation for tourism; The most memorable vacation experience; Vacation activities with friends/family; Activity in leisure time; Watching movies; Hobby.

3. **Religion.** Routine religious activities; Religious figures; The story in the scriptures; Daily stories related to religion; The last religious sermon that you heard; Religious Holidays; Religious scriptures; The most memorable religious teachings; Religious ceremony; House of worship.

4. **Culture & Heritage.** Traditional event; Traditional houses; Folk songs; Folklores from local regions; Traditional weapon; Special souvenirs from local regions; Traditional musical instruments; Regional/traditional dance; Traditional figure; Regional specialty cuisine.

5. **Sports.** Favorite exercise as a child; Easy light exercise; Favorite athlete(s); Sports equipment; Sports at school (extracurricular activity); Sports match; Extreme sports; Unexpected events during sports; Benefits of sports; Watching a sports match.

6. **Technology.** Favorite video game; Handphone; Laptop; Television and radio; Washing machine; Camera; Other electronic appliances; Robot; News about technology; Latest technology.

7. **Business (5 topics).** Businessman; Work at office; Ideas to sell stuff from home; Tips while losing money; Online selling from the internet.

8. **Science.** Animal; Plant; Energy sources; Discoveries; Known figures or researchers; Environmental problems; Diseases and other disorders; Planet and the solar system; Favorite subjects at school; Scientific experiments conducted at school.

9. **History.** The history of the house that is inhabited; The history of public facilities in the neighborhood; The history of the city where you grew up; History of Indonesian independence; National and local heroes.

10. **Politics.** Favorite and disliked political parties; Favorite and disliked political figures; Known political teachings; Election stories; Rules, laws, known regulations.

11. **Emotion: Happy.** Happy to be accepted at college; Happy to pass the exam; Happy to buy goods after saving; Happy because of winning the lottery; Happy to find the item you are looking for; The most fun experience in life; Happy for winning an award; Happy to meet favorite idol(s); Marriage; Happy because of

childbirth.

12. **Emotion: Sad.** Sad due to layoff; Sad due to the death of family members; Failure in life; Childhood trauma; Sad due to failing university entrance exam; Disappointing beloved people; The saddest experience in life; Deepest regret for life; Sad because getting separated from friends/parents' divorce; Sad because of loneliness.

13. **Everyday Life.** Living with a partner/significant other; Preparing an emergency fund; Living with neighbors; How to survive on the mountain; Friendship; Preparing for natural disasters; How to get rid of stress; Farming; Lifestyle in the village; How to ride public transportation.

14. **Emotion: Surprise.** Shocked to see a ghost; Surprised to get a present; Shocked to hear an illness diagnosis; Shocked to because of a prank; Surprised to miss the plane/bus; Surprised to accidentally meet old friends; Surprised to win the lottery; Shocked by positive pregnancy test; Surprised to get a birthday surprise; Surprised to find treasure.

15. **Emotion: Angry.** Angry because someone skipped the queue; Angry because of an insult about physique; Angry to get cheated on; Angry to get hit; Angry because of a traffic jam; Angry without reason; Angry because of unfair treatment; Angry because of discrimination; Angry because other people take our work; Angry at an officer.

16. **Emotion: Fear (20 subtopics).** Scared because of stage fright; Afraid to be at a gunpoint/knifepoint; Fear that evolves to trauma (phobia); Afraid and at the death's door; Scared of seeing ghosts; Afraid of being bullied; Afraid due to being threatened; Fear of drowning; Fear of surgery; Fear of being scolded; Fear of needles; Fear when committing sin/mistakes; Afraid to be a witness to an accident; Afraid when natural disasters occur; Afraid when being chased by animals; Afraid when being stalked; Fear of being alone; Panic attack; Afraid of being in a strange place; Scared when the brake fails.

17. **Emotion: Disgusted (5 subtopics).** Disgusted at vomit; Disgusted at a dirty toilet; Disgusted with animal waste; Disgusting fishy smell; Disgusted at a house full of insects;

Disgusted by other people's saliva; Disgusted at a pile of trash; Disgusted at rotten food; Disgusted at dirty food; Other disgusting experiences/stories.

18. **Emotion: Shame (5 subtopics).** Embarrassed because of torn pants; Embarrassed of farting in public places; Embarrassed because of accidentally getting into the wrong toilet; Embarrassed to be laughed at by classmate(s); Embarrassed to misidentify someone; Embarrassed because of accidentally wearing the same clothes as strangers; Other embarrassing experiences; Embarrassed to get caught red-handed; Embarrassed to be stinky; Embarrassed to be in debt.

19. **Stance: Support/Neutral/Contradict (20 subtopics).** Abortion; Atheism; 1 week = 4 working days + 3 holidays compared to 1 week = 5 working days + 2 holidays; The elimination of national examination; Celibacy; Liberalism; Socialism; Communism; Body positivity; LGBTQ+ in social life; Cloning; The existence of spirits (ghosts/demons/etc.); Reincarnation; The need for college; Culture preservation; Panda conservation; The need for shaving leg hair; Death penalty; Friendship between men and women without being more than friends; The legalization of assisted suicide.

E Paragraph Writing Procedure

For paragraph writing, we initially provide a list of topics before instructing the annotators to write paragraphs. The topics provided vary widely, ranging from simple topics such as food and beverages, entertainment/leisure, and sports, to quite heavy topics such as science, history, politics, and religion. We also provided more specific subtopics from each of the major topics provided. In total, we have 20 main topics with 20 subtopics of each main topic.

The provision of topics (especially, subtopics) aims to facilitate the annotators in the process of writing paragraphs. That way, the annotators only need to write paragraphs by developing ideas from the topics and subtopics that have been given without the need to think about which topic to choose. In addition, the selection of various topics is also expected to enrich vocabulary in the corpus; different topics will obviously bring up the use of different dictionaries.
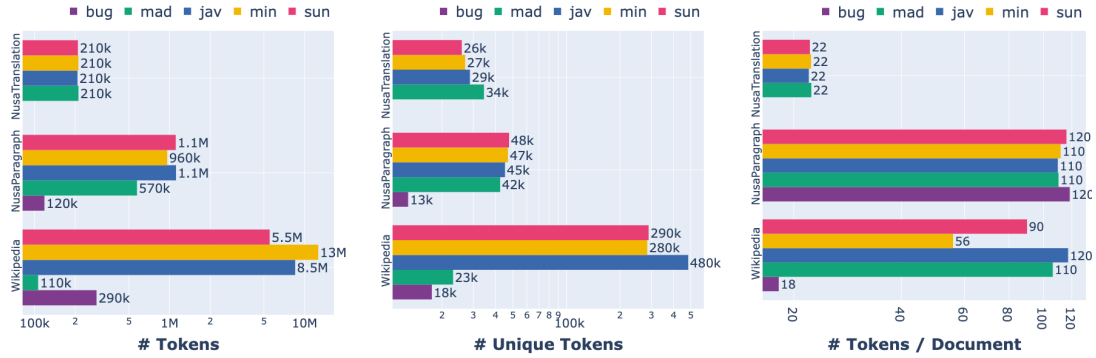
Figure 7: **(left)** #Tokens, **(center)** #Unique Tokens, and **(right)** #Tokens/Document statistics of the corpus collection methods under study: NusaTranslation, NusaParagraph, and Wikipedia.

For conducting the paragraph writing, the annotators are instructed to write short paragraphs with the following criteria: (a) the paragraph consists of a minimum of 100 words, (b) using the targeted local language, (c) the topic is according to the provided topics and subtopics, (d) the type of the paragraphs are narration, description, argumentation, persuasion, and exposition, (e) The content must not defame the name of public entities or contain sensitive and personal information of specific individuals.

The paragraph writing procedure is started (1) after a transfer of learning given to the annotators about the general procedure and knowledge about writing and paragraph types. (2) After that, every annotator is given access to their own worksheet in Google Spreadsheet that already contains all topics and subtopics that they can develop according to the procedure. Every annotator had time around 15 weeks to finish 100–160 paragraphs every week. (3) While annotators already start their paragraph writing process, QC annotators check and validate their work every week (around 2–3 times a week) to the paragraphs that are already done. (4) Lastly, every two weeks, every annotator is gathered in an online meeting to discuss the evaluation of any errors found in their data to prevent mistakes in the future.

Through paragraph writing, we achieved a total of 56395 paragraphs. The details are: 5017 paragraphs for Madura language, 8538 paragraphs for Minangkabau, 10189 paragraphs for Javanese, 9729 paragraphs for Sundanese, 9756 paragraphs for Betawi, 4711 paragraphs for Batak, 5338 paragraphs for Makassar, 1200 paragraphs for Rejang Lebong, 1473 paragraphs for Palembang, 1059 paragraphs for Bugis language, and 44 paragraphs

for Ambon language.

F   Post Annotation Procedure

F.1   Manual Validation

For sentence translation, QC annotators check manually through the data to ensure that all words are translated to the target language and not a single word is skipped by the translator. For paragraph writing, QC annotators check the data by skimming through the paragraphs one by one, checking for any apparent typos, and making sure that the annotators are using the local language and not Indonesian. There are some cases where local languages still use Indonesian words, but it should only be below 30%, while most of the paragraphs must be in the desired local language. To ensure there is no plagiarism, QC annotators also check by sampling some paragraphs from the data and check whether a similar paragraph is found through a search engine.

F.2   Automatic Validation

To further ensure the diversity of the samples, we run an automatic validation to ensure there are no similar paragraphs written by any annotators. Our automatic validation matches two paragraphs by first removing all punctuation marks and then performing string matching using Levensthein distance, and normalizing the distance by dividing with the average length of the two paragraphs. We conduct the process for all the paragraph pairs and we ask the corresponding annotators to revise when the normalized distance of two paragraphs is less than 30%.

Javanese (jav)			Madurese (mad)			Minangkabau (min)			Sundanese (sun)		
word	freq.	prop. (%)	word	freq.	prop. (%)	word	freq.	prop. (%)	word	freq.	prop. (%)
ac	380	0.18	ac	482	0.23	ac	473	0.22	ac	460	0.22
wifi	278	0.13	wifi	294	0.14	wifi	311	0.15	wifi	290	0.14
airy	273	0.13	airy	270	0.13	airy	278	0.13	airy	277	0.13
hotel	243	0.12	tv	255	0.12	tv	259	0.12	tv	259	0.12
tv	227	0.11	hotel	242	0.11	hotel	236	0.11	hotel	249	0.12
hp	201	0.10	hp	202	0.10	hp	225	0.11	hp	181	0.09
video	76	0.04	wc	99	0.05	mode	122	0.06	via	87	0.04
paste	94	0.04	via	116	0.05	motor	61	0.03			
via	89	0.04	shower	89	0.04	twitter	52	0.02			
video	77	0.04	wc	77	0.04						
room	58	0.03									
mode	57	0.03									

Table 4: Common loan words in NusaTranslation from top-200 overlap with the English lexicon and loan word list.

Buginese (bug)			Javanese (jav)			Madurese (mad)			Minangkabau (min)			Sundanese (sun)		
word	freq.	prop. (%)	word	freq.	prop. (%)	word	freq.	prop. (%)	word	freq.	prop. (%)	word	freq.	prop. (%)
media	683	0.58	online	448	0.04	wedding	165	0.03	film	333	0.03	film	548	0.05
hp	535	0.45	media	432	0.04	laptop	83	0.01	motor	253	0.03	duo	543	0.05
tv	396	0.33	laptop	424	0.04	online	82	0.01	tv	239	0.02	hp	410	0.04
laptop	377	0.32	instagram	337	0.04	wc	69	0.01	media	182	0.02	media	407	0.04
hotel	304	0.26	tv	291	0.04	bully	42	0.01	laptop	172	0.02	tv	333	0.03
video	298	0.25	robot	275	0.04	media	35	0.01	robot	155	0.02	laptop	276	0.02
online	270	0.23										video	232	0.02
internet	258	0.22												

Table 5: Common loan words in NusaParagraph from top-200 overlap with the English lexicon and loan word list.

G Token Statistics of the Corpora Under Study

We provide the token statistics of Wikipedia, NusaTranslation, and NusaParagraph in Figure 7. Especially in Buginese (bug), the document length and the number of unique tokens in Wikipedia are rather low, indicating that there is a lot of boilerplate text in the Buginese Wikipedia data.

H List of Loan Words in Indonesian Local Languages

We present the list of manually curated loan words with their frequency and proportion in the corresponding corpus for each language in Table 4, Table 5, and Table 6, for NusaTranslation, NusaParagraph, and Wikipedia, respectively.

I List of Common Local Words In Indonesian Local Languages

We present the list of manually curated common local words with their frequency and proportion in the corresponding corpus for each language in Table 7, Table 8, and Table 9 for NusaTranslation, and NusaParagraph, Wikipedia, respectively.
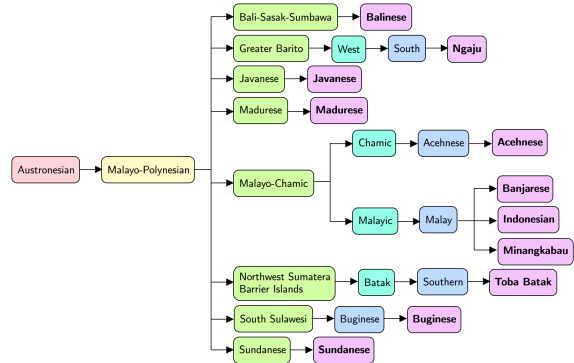
Figure 8: Taxonomy of the languages under study. We show all of the 12 Indonesian local languages under study and the national language of Indonesia, i.e., Indonesian (ind).

J Languages Under Study

In this study, we explore 12 low-resource languages in Indonesian, i.e., Ambon (abs), Batak (btk), Betawi (bew), Bima (bhp), Buginese (bug), Javanese (jav), Madurese (mad), Makassarese (mak), Minangkabau (min), Palembang / Musi (mui), Rejang (rej), and Sundanese (sun). We show the list of all the languages under study in Table 10 and their family tree along with Indonesian (ind) in Fig-

Buginese (bug)			Javanese (jav)			Madurese (mad)			Minangkabau (min)			Sundanese (sun)		
word	freq.	prop. (%)	word	freq.	prop. (%)	word	freq.	prop. (%)	word	freq.	prop. (%)	word	freq.	prop. (%)
kaisne	2450	0.85	the	22125	0.26	of	554	0.52	asteroid	251354	1.99	the	51608	0.94
somme	2323	0.81	of	18454	0.22	planet	301	0.28	ordo	159089	1.26	apollo	23442	0.43
eure	2027	0.70	and	8649	0.10	and	257	0.24	filum	151064	1.20	of	23181	0.42
manche	1805	0.63	a	7256	0.09	gregorian	225	0.21	animalia	148731	1.18	amor	19682	0.36
hautegaronne	1770	0.61	data	3809	0.04	wikipedia	207	0.19	kingdom	148008	1.17	international	15804	0.29
dordogne	1673	0.58	web	3629	0.04	influenza	203	0.19	arthropoda	147996	1.17	planet	15321	0.28
hautesaône	1637	0.57	new	3482	0.04				insecta	146516	1.16	center	15260	0.28
gironde	1628	0.57	film	6100	0.07				cerambycidae	101158	0.80	union	15247	0.28
vosges	1547	0.54	for	2731	0.03				diptera	94388	0.75	orbit	15242	0.28
hautespyrénées	1424	0.49	to	2651	0.03				nebula	88656	0.70	minor	15054	0.27
gers	1391	0.48	tv	2421	0.03				planet	66110	0.52	asteroid	15015	0.27
ardennes	1391	0.48	no	2267	0.03				database	58923	0.47	astronomical	14970	0.27
yonne	1363	0.47	as	2262	0.03				larva	52188	0.41	primordial	14931	0.27
hautemarne	1302	0.45	family	2260	0.03				coleoptera	51860	0.41	and	14887	0.27
ain	1259	0.44	john	2063	0.02				planetesimal	50875	0.40	for	4663	0.08
eureetloir	1211	0.42	isbn	2058	0.02				primordial	46516	0.37	by	3536	0.06
hautrhin	1133	0.39	world	2046	0.02				titan	33741	0.27	that	3303	0.06
drôme	1109	0.39	university	2019	0.02				the	33386	0.26	are	3235	0.06
gard	1063	0.37							orbit	23121	0.18	with	2859	0.05
ardèche	1020	0.35							limoniidae	21274	0.17	be	2636	0.05
ariège	998	0.35							tachinidae	20520	0.16	or	2483	0.05
allier	963	0.33							linear	19213	0.15	as	5071	0.09
deuxsèvres	917	0.32							socorro	19196	0.15	this	2027	0.04
aube	868	0.30							antena	19017	0.15	at	2002	0.04
corrèze	861	0.30							jpl	18739	0.15	gregorian	1728	0.03
finistère	850	0.30							semi	18407	0.15	which	1722	0.03
picardy	816	0.28							minor	18392	0.15	new	1721	0.03
yvelines	787	0.27							porifera	18208	0.14	was	1550	0.03
hauteloire	782	0.27							browser	17753	0.14	ubar	1523	0.03
maineetloire	603	0.21							genus	17721	0.14	not	1279	0.02
alpesdehauteprovence	602	0.21							international	17509	0.14			
vienne	566	0.20							center	17313	0.14			
hautesalpes	533	0.19							union	17232	0.14			
alpesmaritimes	491	0.17							astronomical	17184	0.14			
hautevienne	403	0.14							of	16642	0.13			
essonne	395	0.14							muscidae	15810	0.13			
pyrénéesorientales	185	0.06							ado	15738	0.12			
communauté	163	0.06							asilidae	14571	0.12			
loiretcher	142	0.05							apollo	14350	0.11			
guadeloupe	99	0.03							demospongiae	13249	0.10			
communes	55	0.02							history	13053	0.10			
the	33	0.01							cecidomyiidae	13050	0.10			
of	32	0.01							american	12950	0.10			
community	15	0.01							spider	12833	0.10			
singapore	14	0.00							catalog	12832	0.10			
from	12	0.00							araneae	12814	0.10			
language	12	0.00							ceratopogonidae	11324	0.09			
wedding	11	0.00							bombyliidae	10934	0.09			
proton	10	0.00							salticidae	10431	0.08			
canton	10	0.00							parasit	10294	0.08			

Table 6: Common loan words in NusaParagraph from top-200 overlap with the English lexicon and loan word list. We only show the top 50 words for Buginese (bug) and Minangkabau (min). In total, the top 200 overlapping Buginese (bug) and Minangkabau (min) data with the English lexicon contain 54 and 90 loan words, respectively.

ure 8. We provide a more detailed overview of each language in the following paragraphs.

Ambonese Malay (abs) is spoken in various parts of Maluku province. It was developed on the island of Ambon in the 16th century, firstly used as (spice) trade language, and now it is used as lingua franca for interethnic communication in the market domain and some media. Being a Malay-based creole language, it has around 81% of lexical similarity with Indonesian. The speakers have marginal intelligibility with Indonesian and difficult intelligibility with North Moluccan Malay (Eberhard et al., 2021). It is written in Latin script.

Batak Toba (bbc) is spoken in North Sumatra province. It is slowly being replaced by Indonesian in urban and migrant areas. It used to be written in the Batak script but is mainly written in Latin script now.

Betawi (bew) is spoken in Tangerang, Ban-ten province, Jakarta, and some cities in West Java province such as Depok, Bekasi, Bogor, and Karawang. It is a Malay-based creole distinct from both Indonesian and other Malay-based pidgins and creoles. It was evolved around mid-19th century. It functions as a Low variety in a diglossic situation, but has covert prestige when used by the upper class. It has unique phonological, morphological, and lexical traits. It was influenced by Peranakan Indonesian language and Balinese.

Bima (bhp) is spoken in Komodo island area in East Nusa Tenggara province and some islands in West Nusa Tenggara province such as Sumbawa island and Banta and Sangeang islands. It has five dialects: Kolo, Sangar, Toloweri, Bima, and Mbojo (Eberhard et al., 2021). It is written in Latin script.

Batak languages (btk) are a subgroup of the languages of Northwest Sumatra-Barrier Islands

Topic	Word	Javanese (jav)		Madurese (mad)		Minangkabau (min)		Sundanese (sun)	
		freq	prop (%)	freq	prop (%)	freq	prop (%)	freq	prop (%)
food	indomie	9	0.0076	7	0.0006	11	0.0019	11	0.0011
	rendang	3	0.0025	3	0.0003	0	0.0000	3	0.0003
	tempe	3	0.0025	2	0.0002	2	0.0003	3	0.0003
	gule	0	0.0000	0	0.0000	0	0.0000	0	0.0000
	sate	3	0.0025	3	0.0003	3	0.0005	3	0.0003
transportation	angkot	6	0.0051	4	0.0004	6	0.0010	5	0.0005
	ojol	5	0.0042	4	0.0004	6	0.0010	5	0.0005
	gojek	15	0.0127	13	0.0012	15	0.0026	14	0.0015
religion	doa	1	0.0008	12	0.0011	47	0.0082	36	0.0037
	gaib	3	0.0025	2	0.0002	1	0.0002	2	0.0002
	alhamdulilah	2	0.0017	4	0.0004	2	0.0003	14	0.0015
	insyaallah	4	0.0034	4	0.0004	2	0.0003	1	0.0001
adjective	bule	9	0.0076	8	0.0007	16	0.0028	15	0.0016
	santun	2	0.0017	0	0.0000	2	0.0003	3	0.0003
	alay	9	0.0076	9	0.0008	11	0.0019	7	0.0007
Total		74	**0.0625**	75	**0.0059**	124	**0.0197**	122	**0.0120**

Table 7: Common local words with their frequency and proportion in the NusaTranslation corpus.

Topic	Word	Buginese (bug)		Javanese (jav)		Madurese (mad)		Minangkabau (min)		Sundanese (sun)	
		freq	prop (%)	freq	prop (%)	freq	prop (%)	freq	prop (%)	freq	prop (%)
food	indomie	0	0.0000	24	0.0021	9	0.0016	12	0.0012	15	0.0013
	rendang	7	0.0059	27	0.0024	9	0.0016	4	0.0004	32	0.0029
	tempe	2	0.0017	101	0.0090	45	0.0078	29	0.0030	31	0.0028
	gule	0	0.0000	8	0.0007	32	0.0056	0	0.0000	2	0.0002
	sate	7	0.0059	69	0.0062	122	0.0213	112	0.0117	91	0.0082
transportation	angkot	8	0.0068	52	0.0047	3	0.0005	57	0.0059	254	0.0228
	ojol	0	0.0000	6	0.0005	0	0.0000	0	0.0000	16	0.0014
	gojek	12	0.0101	15	0.0013	15	0.0026	4	0.0004	17	0.0015
religion	doa	11	0.0093	5	0.0004	3	0.0005	102	0.0106	82	0.0074
	gaib	0	0.0000	46	0.0041	1	0.0002	13	0.0014	18	0.0016
	alhamdulilah	0	0.0000	1	0.0001	6	0.0010	2	0.0002	24	0.0022
	insyaallah	0	0.0000	1	0.0001	12	0.0021	17	0.0018	8	0.0007
adjective	bule	1	0.0008	5	0.0004	8	0.0014	1	0.0001	9	0.0008
	santun	0	0.0000	9	0.0008	5	0.0009	11	0.0011	14	0.0013
	alay	0	0.0000	2	0.0002	3	0.0005	9	0.0009	1	0.0001
Total		48	**0.0405**	371	**0.0332**	273	**0.0471**	373	**0.0379**	614	**0.0551**

Table 8: Common local words with their frequency and proportion in the NusaParagraph corpus.

spoken by the Batak people in the North Sumatra province and surrounding areas. Batak languages can be divided into three groups: Northern, Simalungan, and Southern. The Northern group consists of three languages: Batak Alas-Kluet (btz), Batak Dairi (btd), and Batak Karo (btx). The Simalungan group has one language only, i.e., Batak Simalungun (bts). The Southern group consists of three languages: Batak Angkola (akb), Batak Mandailing (btm), and Batak Toba (bbc) (Eberhard et al., 2021). Batak languages are predicate-initial, and have verb systems reminiscent of Philippine languages, although they differ from them in many details (Blust et al., 2013). They were written using the Batak script, but the Latin script is now used for most writing. Our annotators are originating from Batak Toba and Batak Mandailing which are part of the Southern group.

Batak Mandailing (btm) is spoken in North Sumatra (south interior from Padang Sidempuan into Riau) and West Sumatra provinces. The speakers are shifting to Indonesian in urban and migrant areas (Eberhard et al., 2021). It is written in Batak script.

Batak Toba (bbc) is a language spoken in the North Sumatra province. Similarly to Acehnese,

Topic	Word	Buginese (bug)		Javanese (jav)		Madurese (mad)		Minangkabau (min)		Sundanese (sun)	
		freq	prop (%)	freq	prop (%)	freq	prop (%)	freq	prop (%)	freq	prop (%)
food	indomie	0	0.0000	36	0.0032	1	0.0002	1	0.0001	51	0.0046
	rendang	0	0.0000	36	0.0032	0	0.0000	14	0.0015	24	0.0022
	tempe	0	0.0000	123	0.0110	0	0.0000	29	0.0030	11	0.0010
	gule	0	0.0000	26	0.0023	0	0.0000	2	0.0002	4	0.0004
	sate	0	0.0000	223	0.0200	2	0.0003	169	0.0176	21	0.0019
transportation	angkot	0	0.0000	14	0.0013	0	0.0000	8	0.0008	56	0.0050
	ojol	0	0.0000	0	0.0000	0	0.0000	0	0.0000	0	0.0000
	gojek	0	0.0000	18	0.0016	1	0.0002	20	0.0021	0	0.0000
religion	doa	0	0.0000	88	0.0079	2	0.0003	35	0.0036	65	0.0058
	gaib	0	0.0000	90	0.0081	0	0.0000	17	0.0018	70	0.0063
	alhamdulilah	0	0.0000	2	0.0002	0	0.0000	0	0.0000	0	0.0000
	insyaallah	0	0.0000	2	0.0002	0	0.0000	0	0.0000	0	0.0000
adjective	bule	0	0.0000	19	0.0017	0	0.0000	5	0.0005	7	0.0006
	santun	0	0.0000	39	0.0035	0	0.0000	7	0.0007	14	0.0013
	alay	0	0.0000	7	0.0006	0	0.0000	0	0.0000	0	0.0000
Total		**0**	**0.0000**	**723**	**0.0647**	**6**	**0.0010**	**307**	**0.0319**	**323**	**0.0291**

Table 9: Common local words with their frequency and proportion in the Wikipedia corpus.

Language	ISO code	Status
Ambonese Malay	abs	wider communication
Mandailing	btm	threatened
Batak Toba	bbc	threatened
Betawi	bew	threatened
Bima	bhp	vigorous
Buginese	bug	wider communication
Javanese	jav	educational
Madurese	mad	developing
Makassarese	mak	threatened
Minangkabau	min	developing
Palembang / Musi	mui	wider communication
Rejang	rej	vigorous
Sundanese	sun	developing

Table 10: List of all languages under study in the NusaWrites benchmark along with their status of language development versus language endangerment.

it is slowly being replaced by Indonesian in urban and migrant areas. It used to be written in the Batak script but is mainly written in Latin script now. The Batak languages are verb-initial, and have verb systems reminiscent of Philippine languages, although they differ from them in many details (Blust et al., 2013).

Javanese (jav) is a language spoken mainly in Java island. It is the de facto language of provincial identity in central and eastern Java. The word order is SVO. It has 21 consonants and 8 vowels. It used to be written in Javanese script but since the 20th century, it was mostly written in Latin script. Javanese differs from most other languages of western Indonesia in contrasting dental and retroflex stops and in the feature of breathy voice or mur-

mur as a phonetic property of its voiced obstruents. Javanese also differs from most languages of the Philippines and western Indonesia in allowing a number of word-initial consonant clusters. It has an elaborate system of speech levels (Blust et al., 2013).

Madurese (mad) is a language spoken in the East Java province, mainly on Madura Island, south and west of Surabaya city, Bawean, Kangean, and Sapudi islands. It has vowel harmony, gemination, rich affixation, three types of reduplication, and SVO basic word order (Davies, 2010).

Makassarese (mak) is mainly spoken in South Sulawesi province. It has three dialects that form a chain. Those dialects are Lakiung (Gowa), Turatea (Jeneponto), and Bantaeng (Maros-Pangkep). The Gowa dialect is prestigious. It has 17 consonants and 5 vowels. The stress is on the penultimate syllable. Similar to other Western Malayo-Polynesian languages, it has inclusive and exclusive pronouns, noun head initials, prepositions, definite markers, classifiers, passive markers, and aspect markers (Eberhard et al., 2021). The speakers, especially young people in the cities, are shifting to Indonesian and Makassar Indonesian. It is taught as a subject in primary schools and written in Latin script. The Makassar script is no longer used.

Minangkabau (min) is a language spoken mainly in West Sumatra and other provinces on Sumatra Island such as Bengkulu and Riau. Although it is classified as Malay, it is not intelligible with Indonesian. The word order is SVO written in Latin script. Standard Minangkabau

voice can be characterized as an Indonesian-type system whereas colloquial Minangkabau voice is more effectively characterized as a Sundic-type system (Crouch, 2009).

Musi (mui) is mainly spoken in South Sumatra province, widespread in the northern two-thirds of the province from the Musi River upstream to Bukit Barisan Mountains and downstream to the coastal swamplands. It is also spoken in the northeast region of Lampung province and around the border areas in Jambi and Bengkulu provinces. It has twelve dialects. A mutually intelligible dialect chain stretches along the Musi River with two subgroups: Musi and Palembang. The speakers use it for cultural stories and songs, but they prefer Indonesian for educational and religious materials (Eberhard et al., 2021).

Rejang (rej) is a language spoken in some parts of Bengkulu and South Sumatra provinces. It has five dialects: Lebong, Kepahiang (Kebanagung), Pasisir, Musi (Curup), and Rawas. Lebong is recognized as a central dialect (Eberhard et al., 2021). It is written in Latin script and Kaganga script. About 85% of the speakers live in remote rural areas. Most of them are Muslims.

Sundanese (sun) is a language spoken mainly in the Banten and West Java provinces. It is the de facto language of provincial identity in western Java. The main dialects are Bogor (Krawang), Pringan, and Cirebon. It is non-tonal and has 18 consonant and 7 vowel phonemes. The stress is on the penultimate syllable. It has elaborate coding of respect levels. It has been written in Latin script since the middle of the 19th century but was previously written in Arabic, Javanese, and Sundanese scripts. Sundanese is a predominantly SVO language. It has voice marking and incorporates some (optional) actor-verb agreement, i.e., number and person (Kurniawan, 2013).

K Task Description and Statistics of NusaTranslation

We developed three tasks from NusaTranslation, i.e., emotion recognition, sentiment analysis, and machine translation. The statistics of the dataset are shown in Table 11.

K.1 Emotion Recognition

From the EmoT (Saputri et al., 2018; Wilie et al., 2020) part of the translation, develop a parallel emotion recognition dataset of 11 low-resource lan-

guages. For most of the languages, we provide new 3,000/401/1,000 train-validation-test splits for all datasets and attach the ids of the original split. In some languages with fewer data, we randomly sample all the original splits and ensure that the test split has a reasonable amount of test samples.

K.2 Sentiment Analysis

We develop a parallel sentiment analysis from the IndoLEM Sentiment (Koto and Rahmaningtyas, 2017; Koto et al., 2020) part of the translation. We provide new 3400/448/1200 train-validation-test splits for most datasets and keep the ids of the original split attached. Similar to the emotion recognition dataset, for the languages with less number of annotators we randomly sample the original splits maintaining a reasonable amount of samples on the test split.

K.3 Machine Translation

Using the whole constructed translation data, we develop a ind↔xxx machine translation task for 11 languages. The scale of our machine translation task is close to one magnitude higher compared to NusaX (Winata et al., 2023) which also develops a parallel corpus for 10 local languages spoken in Indonesia.

L Task Description and Statistics of NusaParagraph

We developed three tasks from NusaParagraph, i.e., rhetoric mode classification, emotion recognition, and topic modeling. The statistics of the dataset is shown in Table 12.

L.1 Rhetoric Mode Classification

We develop a new rhetoric mode multi-class classification task ranging across 10 low-resource languages. The train-validation-test splits vary between languages as shown in Table 12. Paragraphs in the dataset are labeled into one of 5 categories: (1) narrative; (2) argumentative; (3) expository; (4) descriptive; and (5) persuasive.

L.2 Emotion Recognition

Using the whole NusaParagraph dataset, we compose an emotion recognition multi-class classification task for 10 low-resource languages. The train-validation-test splits vary between languages as shown in Table 12. The emotions expressed in each dataset are labeled into one of 7 emotions: (1)

Language	#MT	#Emotion	#Sentiment	Split MT	Split Emot	Split Sentiment
btk	9,449	4,401	5,048	(6,600, 849, 2,000)	(3,000, 401, 1,000)	(3,400, 448, 1,200)
bew	9,449	4,401	5,048	(6,600, 849, 2,000)	(3,000, 401, 1,000)	(3,400, 448, 1,200)
sun	9,449	4,401	5,048	(6,600, 849, 2,000)	(3,000, 401, 1,000)	(3,400, 448, 1,200)
jav	9,449	4,401	5,048	(6,600, 849, 2,000)	(3,000, 401, 1,000)	(3,400, 448, 1,200)
mad	9,449	4,401	5,048	(6,600, 849, 2,000)	(3,000, 401, 1,000)	(3,400, 448, 1,200)
mak	9,449	4,401	5,048	(6,600, 849, 2,000)	(3,000, 401, 1,000)	(3,400, 448, 1,200)
min	9,449	4,401	5,048	(6,600, 849, 2,000)	(3,000, 401, 1,000)	(3,400, 448, 1,200)
mui	1,574	733	841	(1,000, 174, 400)	(200, 83, 450)	(250, 91, 500)
rej	1,574	746	828	(1,000, 174, 400)	(200, 96, 450)	(250, 78, 500)
abs	1,574	726	848	(1,000, 174, 400)	(200, 76, 450)	(250, 98, 500)
bhp	1,579	719	860	(1,000, 179, 400)	(200, 69, 450)	(260, 100, 500)

Table 11: Statistics of the NusaTranslation corpus. Split means (training, development, test) respectively.

Language	#Rhetorical	#Emotion	#Topic	Split Rhetorical	Split Emot	Split Topic
btk	4,908	1941	2,125	(3,200, 508, 1,200)	(1,150, 292, 500)	(1,350, 275, 500)
bew	9,755	3,928	3,885	(6,900, 855, 2,000)	(2,700, 430, 800)	(2,650, 435, 800)
bug	1,000	437	443	(500, 100, 400)	(87, 50, 300)	(93, 50, 300)
jav	10,188	4,040	3898	(7,300, 888, 2,000)	(2,800, 440, 800)	(2,650, 448, 800)
mad	5,211	1,762	2,867	(3,500, 511, 1,200)	(1,000, 263, 500)	(1,800, 367, 700)
mak	5,471	2,303	2,576	(3,700, 571, 1,200)	(1,500, 304, 500)	(1,500, 376, 700)
min	8,608	3,153	3,599	(6,000, 608, 2,000)	(2,000, 357, 800)	(2,400, 399, 800)
mui	1,474	676	648	(900, 174, 400)	(200, 75, 400)	(168, 80, 400)
rej	1,200	486	505	(650, 150, 400)	(136, 50, 300)	(105, 50, 350)
sun	9,594	3,598	4,168	(6,700, 894, 2,000)	(2,400, 400, 800)	(2,800, 468, 900)

Table 12: Statistics of the NusaParagraph corpus. Split means (training, development, test) respectively.

Hyperparams	Values
batch size	8
num epochs	100
early stop	3
max norm	10
optimizer	Adam
Adam β	(0.9, 0.999)
Adam ϵ	1e-8

Table 13: Hyperparameters of pre-trained LMs on machine translation tasks.

Hyperparams	Values
learning rate	1e-5
batch size	32
num epochs	100
early stop	3
max norm	10
optimizer	Adam
Adam β	(0.9, 0.999)
Adam ϵ	1e-8

Table 14: Hyperparameters of pre-trained LMs on classification tasks.

fear; (2) disgusted; (3) sad; (4) happy; (5) shame; (6) angry; and (7) surprise.

L.3 Topic Modeling

Each paragraph proposed in the NusaParagraph dataset is annotated with a topic to form a topic modeling task for 10 low-resource languages. The train-validation-test splits vary between languages depending on the dataset size. The details are provided in Table 12. Each paragraph is labeled into one of 8 topics: (1) food & beverages; (2) sports; (3) leisures; (4) religion; (5) culture & heritage; (6) slice of life; (7) technology; and (8) business.

M Experiment Hyperparameters

M.1 Statistical Machine Translation

For PBSMT, we set the n-gram value of the Moses toolkit to 3. Other parameters were kept to their default values.

No	Zero-Shot Prompts
	Emotion Recognition
1	[INPUT] => Emotion: [LABELS_CHOICE]
2	Text: [INPUT] => Emotion: [LABELS_CHOICE]
3	[INPUT]\nWhat would be the emotion of the text above? [LABELS_CHOICE]
4	What is the emotion of this text?\nText: [INPUT]\nAnswer: [LABELS_CHOICE]
5	Text: [INPUT]\nPlease classify the emotion of above text. Emotion: [LABELS_CHOICE]
	Rhetoric Mode Classification
6	[INPUT] => Rhetorical mode: [LABELS_CHOICE]
7	Text: [INPUT] => Rhetorical mode: [LABELS_CHOICE]
8	[INPUT]\nWhat would be the rhetorical mode of the text above? [LABELS_CHOICE]
9	What is the rhetorical mode of this text?\nText: [INPUT]\nAnswer: [LABELS_CHOICE]
10	Text: [INPUT]\nPlease classify the rhetorical mode of above text. Rhetorical mode: [LABELS_CHOICE]
	Topic Modeling
11	[INPUT] => Topic: [LABELS_CHOICE]
12	Text: [INPUT] => Topic: [LABELS_CHOICE]
13	[INPUT]\nWhat would be the topic of the text above? [LABELS_CHOICE]
14	What is the topic of this text?\nText: [INPUT]\nAnswer: [LABELS_CHOICE]
15	Text: [INPUT]\nPlease classify the topic of above text. Topic: [LABELS_CHOICE]
	Sentiment Analysis
16	[INPUT] => Sentiment: [LABELS_CHOICE]
17	Text: [INPUT] => Sentiment: [LABELS_CHOICE]
18	[INPUT]\nWhat would be the sentiment of the text above? [LABELS_CHOICE]
19	What is the sentiment of this text?\nText: [INPUT]\nAnswer: [LABELS_CHOICE]
20	Text: [INPUT]\nPlease classify the sentiment of above text. Sentiment: [LABELS_CHOICE]
	Machine Translation
21	Translate the following text from [SOURCE] to [TARGET].\nText: [INPUT]\nTranslation:
22	[INPUT]\nTranslate the text above from [SOURCE] to [TARGET].
23	Text in [SOURCE]: [INPUT]\nHow would you translate that in [TARGET]?
24	Translate the following [SOURCE] text from to [TARGET].\nText: [INPUT]\nTranslation:
25	Text in [SOURCE]: [INPUT]\nText in [TARGET]:

Table 15: List of prompts used in our zero-shot prompting experiments.

M.2 Neural Machine Translation

Table 13 shows the hyperparameters used in deep learning models on machine translation experiments in this work. For the learning rate, it follows the configuration of NusaX (Winata et al., 2023), while the rest are shown in the following table.

M.3 Multi-Class Classification

Table 14 shows the hyperparameters used in deep learning models on classification experiments in this work. Tasks that follow the following parameters include: sentiment analysis, rhetoric mode classification, emotion recognition, and topic modeling. We follow the hyperparameter settings in Winata et al. (2023) that were found to work best.

N List of Zero-Shot Prompts

We provide the full list of prompts used in our zero-shot prompting experiment in Table 15.