NusaBERT: Teaching IndoBERT to be Multilingual and Multicultural

Wilson Wongso[1,2*], **David Samuel Setiawan**[2*], **Steven Limcorn**[2*], **Ananto Joyoadikusumo**[2*]
[1]University of New South Wales [2]LazarusNLP
[*] Equal Contribution

Abstract

Indonesia's linguistic landscape is remarkably diverse, encompassing over 700 languages and dialects, making it one of the world's most linguistically rich nations. This diversity, coupled with the widespread practice of code-mixing and the presence of low-resource regional languages, presents unique challenges for modern pre-trained language models. In response to these challenges, we developed NusaBERT, building upon IndoBERT by incorporating vocabulary expansion and leveraging a diverse multilingual corpus that includes regional languages. Through rigorous evaluation across a range of benchmarks, NusaBERT demonstrates state-of-the-art performance in tasks involving multiple languages of Indonesia, paving the way for future natural language understanding research for under-represented languages. Our models and code are publicly available.[1]

1 Introduction

Indonesia's exceptional linguistic landscape, encompassing over 700 languages and dialects (Aji et al., 2022), presents a significant challenge for current natural language processing (NLP) techniques, such as pre-trained language models. These techniques often fall short in handling the nation's intricate and multifaceted linguistic tapestry. Furthermore, the bilingual nature of Indonesian colloquial conversations (mixing Indonesian and English) with the majority continuing to also communicate in regional languages as their daily conversational language poses a complex problem to be solved by language models.

Nonetheless, pre-trained language models have shown remarkable progress in recent years showing their ability to solve a wide range of natural language processing tasks, including the Indonesian language. These language models are trained on a large corpus and are fine-tuned to solve specific, downstream tasks. Language models such as BERT (Devlin et al., 2018) and GPT (Radford et al., 2018, 2019) are typically trained on a monolingual corpus and were originally trained on an English corpus. In the studies that followed, language-specific language models like IndoBERT (Wilie et al., 2020) and IndoBART (Cahyawijaya et al., 2021) have been tailored for the Indonesian language and regional languages of Indonesia like Javanese and Sundanese. Despite the large size discrepancy between the English and Indonesian corpus, IndoBERT managed to leverage the contextualized Indonesian language model to attain exceptional results on multiple downstream natural language understanding (NLU) tasks.

Although demonstrating remarkable capabilities across various tasks, these models often perform poorly when applied to languages with unique characteristics like those found in the many regions of Indonesia. For instance, IndoBERT faces limitations when addressing the intricacies of code-mixing (Adilazuarda et al., 2022) and the specific needs of low-resource languages (Cahyawijaya et al., 2023b). Furthermore, while efforts like XLM-R (Conneau et al., 2020) and mBERT (Devlin et al., 2018) have aimed to introduce cross-linguality, their focus on achieving state-of-the-art performance in cross-lingual language understanding tasks may not fully address the unique issues faced by language models operating within Indonesia's complex multilingual and multicultural environment. Cahyawijaya et al. (2023b) showed that even these large multilingual models fail to outperform classical baselines on extremely low-resource languages.

In light of this, we propose NusaBERT, a model that builds upon IndoBERT and targets the linguistic complexities of low-resource regional languages in Indonesia. NusaBERT leverages the vocabulary expansion technique proposed by PhayaThaiBERT (Sriwirote et al., 2023), and aims to achieve state-of-the-art performance on multilingual benchmarks.

[1] https://github.com/LazarusNLP/NusaBERT

Proceedings of the Second Workshop in South East Asian Language Processing, pages 10–26
January 20, 2025.

2 Related Works

Recent years have witnessed significant progress in Indonesian NLP research. Pre-trained language models like IndoBERT (Wilie et al., 2020) and IndoBART (Cahyawijaya et al., 2021) have demonstrated the effectiveness of this approach for various Indonesian language tasks. IndoBERT, based on BERT (Devlin et al., 2018), was specifically trained on a large Indonesian text corpus. It achieved state-of-the-art performance on the IndoNLU benchmark (Wilie et al., 2020), a collection of Indonesian-specific NLU tasks like text classification, question answering, and named entity recognition, demonstrating its competence in understanding the nuances of the Indonesian language. IndoBART, based on the BART architecture (Lewis et al., 2020), focuses on sequence-to-sequence tasks within the Indonesian language. This model has found success in language generation tasks like machine translation and text summarization, highlighting its ability to process and produce natural Indonesian text.

NusaX (Winata et al., 2023), a benchmark for 10 under-resourced Indonesian local languages, shows that when IndoBERT and IndoBART are fine-tuned for these languages, they achieve impressive results in sentiment analysis and machine translation, respectively. Afterward, NusaWrites (Cahyawijaya et al., 2023b) was released and complements NusaX by providing a more lexically diverse and culturally relevant dataset on 12 underrepresented local languages. Upon fine-tuning different models on these new benchmarks, results show that multilingual models like XLM-R (Conneau et al., 2020) and mBERT (Devlin et al., 2018) and monolingual models (IndoBERT, IndoBART, IndoGPT) fail to outperform classical machine learning models on several extremely low-resource languages.

The success of PhayaThaiBERT (Sriwirote et al., 2023), a Thai language model specifically designed to address the challenge of unassimilated loanwords, offers valuable inspiration for tackling challenges faced by NLP models in Indonesia. Similar to Thai, low-resource regional languages in Indonesia are frequently influenced by other languages due to code-mixing. This phenomenon leads to a significant number of unassimilated loanwords, which are words from other languages adopted into the regional language but fully integrated into its grammar. PhayaThaiBERT addresses this challenge by incorporating techniques such as vocabulary expansion. This technique involves augmenting the model's vocabulary with these loanwords and variations, allowing it to better recognize and understand them within the context of the regional language. Similarly, IndoBERTweet (Koto et al., 2021), an extension of IndoLEM's IndoBERT (Koto et al., 2020), tackles the challenge of informal language and social media slang by specifically augmenting its vocabulary with terms commonly found in Indonesian Twitter data. Their vocabulary expansion and subword embedding averaging technique (Cahyawijaya et al., 2024) helped the model better understand and process the nuances of informal communication, which often deviates from standard Indonesian grammar and incorporates slang terms.

3 NusaBERT

This section introduces the vocabulary expansion method applied to IndoBERT (Wilie et al., 2020), the corpus dataset used for training, and the continued pre-training procedure. Subsequently, we will evaluate our resultant models on downstream tasks to measure their natural language understanding, multilinguality, and multicultural capabilities.

3.1 Vocabulary Expansion and Dataset

3.1.1 Pre-training Corpus

Following PhayaThaiBERT (Sriwirote et al., 2023), we expanded IndoBERT's vocabulary to introduce foreign tokens by collecting monolingual texts in various Indonesian languages from open-source corpora. We utilized the dataset catalog from NusaCrowd (Cahyawijaya et al., 2023a), which streamlined the process of locating Indonesian datasets. To ensure quality, we focused on clean, rigorously filtered datasets, particularly CulturaX (Nguyen et al., 2023), which uses mC4 (Raffel et al., 2023) and OSCAR (Suárez et al., 2019). CulturaX, however, only covers Indonesian (`ind`), Javanese (`jav`), and Sundanese (`sun`). We also included Standard Malay (`msa`) due to its use in parts of Sumatra and West Kalimantan (Wahyudi et al.; Corporation, 2007).

To add further linguistic diversity, we used an open-source, deduplicated and filtered Wikipedia dataset[2] for Indonesian languages. This dataset includes Acehnese (`ace`), Balinese (`ban`), Banjarese (`bjn`), Banyumasan (`jav`[3]), Buginese (`bug`),

[2] https://hf.co/datasets/sabilmakbar/indo_wiki

[3] A dialect of Javanese (`jav`), sometimes given the ISO

Gorontalo (gor), Minangkabau (min), Malay (msa), Nias (nia), Sundanese (sun), and Tetum (tet). Tetum (tet) was included as it is still spoken in parts of West Timor.

Since the Wikipedia dataset is smaller than typical web-based corpora, we supplemented it with a filtered Indonesian subset of NLLB corpus (Costa-jussà et al., 2022), called KoPI-NLLB[4]. KoPI-NLLB covers Acehnese (ace), Balinese (ban), Banjarese (bjn), Javanese (jav), Minangkabau (min), and Sundanese (sun) and we deliberately excluded Indonesian (ind) from KoPI-NLLB as it was well represented in CulturaX. Our final pre-training corpus comprises 13 languages, integrating CulturaX, Wikipedia, and KoPI-NLLB with a focus on quality via strict filtering and deduplication, summarized in Appendix B.

3.1.2 Vocabulary Expansion

Unlike PhayaThaiBERT, we did not transfer the non-overlapping vocabulary of XLM-R (Conneau et al., 2020). Instead, we decided to train a new WordPiece tokenizer (Wu et al., 2016) based on the IndoBERT tokenizer on the newly formed corpus. There are several design choices considered when training the new tokenizer, such as the target vocabulary size and the subsets to be included during tokenizer training. For the latter, we decided not to include the Indonesian subset of CulturaX due to its large percentage and that it would diminish the importance of non-Indonesian tokens, which contradicts the goal of NusaBERT. However, the relatively smaller Indonesian Wikipedia is still included as there might be new words that might have not been included in the IndoBERT tokenizer.

On the other hand, for the former, we followed a close estimate to that of Typhoon language models (Pipatanakul et al., 2023) whose design choice is based on another previous study that investigated the most efficient target vocabulary size (Csaki et al., 2023). Both studies suggested a vocabulary size of 5,000, but our preliminary experiments found that a target vocabulary size of 5,000 has very few new tokens to be added to the current tokenizer. Due to this, we increased the target vocabulary size to 10,000 and found 1,511 new, non-overlapping tokens to be added.

While this increase is not as significant as originally proposed in PhayaThaiBERT, we considered the downstream effects of significantly increasing the number of parameters if we decided to exactly follow their approach. Moreover, WangchanBERTa (Lowphansirikul et al., 2021), the base model of PhayaThaiBERT, has a deeper issue of only supporting mainly Thai tokens and struggles with unassimilated loanwords in the Latin alphabet. The IndoBERT tokenizer, on the other hand, has been trained on an Indonesian corpus that uses the Latin alphabet and NusaBERT aims to only introduce regional language tokens. Therefore, we finalized this set of additional tokens which increased IndoBERT's vocabulary size from 30,521 to 32,032.

3.2 Continued Pre-training

3.2.1 Model Configuration and Initialization

Like PhayaThaiBERT, we conducted continued pre-training with IndoBERT's initial model checkpoints. We experimented with two size variants of IndoBERT, namely $\text{IndoBERT}_{\text{BASE}}$ and $\text{IndoBERT}_{\text{LARGE}}$. In both variants, we used phase one checkpoints of IndoBERT. Therefore, the initial parameters of our model are identical to that of IndoBERT with the exception of the new vocabulary's embeddings, which are initialized from the mean of the old word embeddings (Hewitt, 2021). There are no additional architectural changes added to the original BERT architecture and call our new extended models $\text{NusaBERT}_{\text{BASE}}$ (111M) and $\text{NusaBERT}_{\text{LARGE}}$ (337M), respectively.

3.2.2 Data Pre-processing

During the continued pre-training, we decided to keep the same sequence length of 128 as IndoBERT phase one models. Our data pre-processing procedures follow a typical masked language modeling pre-processing setup. Firstly, a random 5% sample of our corpus described in Section §3.1.1 is held out for evaluation purposes. Secondly, all texts are tokenized using the newly extended tokenizer described in §3.1.2. Since our tokenizer follows exactly from the original IndoBERT tokenizer, special [CLS] and [SEP] tokens are added at the start and end of all texts. Finally, batches of tokenized texts are then concatenated into one long sequence and then grouped into sequences of length 128 tokens each. Sequences shorter than 128 are thus discarded. These batches of fixed-length tokenized sequences are thereby ready for training purposes.

3.2.3 Pre-training Objective and Procedures

Instead of using the original BERT (Devlin et al., 2018) objective of both next sentence prediction

code map-bms.

[4] https://hf.co/datasets/acul3/KoPI-NLLB

(NSP) and masked language modeling (MLM), we opted for the more robust RoBERTa (Liu et al., 2019) objective. With this setup, we conducted continued pre-training for 500,000 optimization steps with hyperparameters shown in Appendix C. Unlike PhayaThaiBERT, our continued pre-training procedure doesn't involve sophisticated fine-tuning techniques. Instead, we simply trained our models with 24,000 warmup steps to the peak learning rate and applied a linear learning rate decay to zero, with a batch size of 256 on a single GPU.

3.3 Evaluation Benchmark

Our benchmark concentrates on three aspects: (1) natural language understanding (NLU), (2) multilinguality, and (3) multicultural. Therefore, we decided to utilize the Indonesian NLU benchmark IndoNLU (Wilie et al., 2020), and multilingual NLU benchmarks such as NusaX (Winata et al., 2023), and NusaWrites (Cahyawijaya et al., 2023b) which contain a wide range of regional languages of Indonesia and closely reflect the local cultures. The tasks in these benchmarks can be divided into five major categories: (a) single-sentence classification, (b) single-sentence multi-label classification, (c) sequence-pair classification, (d) token classification, and (e) sequence-pair token classification.

3.3.1 Datasets

The IndoNLU benchmark consists only of Indonesian datasets from various NLU tasks. On the other hand, NusaX (Winata et al., 2023) and NusaWrites (Cahyawijaya et al., 2023b) provide NLU benchmarks for a variety of regional languages of Indonesia. A high-level overview of the benchmarks is shown in Appendix A. The list of all languages and dialects involved in this study and its details are found in Appendix B.

IndoNLU IndoNLU (Wilie et al., 2020) is a comprehensive benchmark corpus designed to facilitate research in Indonesian natural language understanding. It comprises multiple datasets covering a variety of NLU tasks, which can be categorized into two main tasks: text classification and sequence labeling. The benchmark aims to provide a standard for evaluating the performance of models on Indonesian language tasks, addressing the need for more resources in languages other than English. The dataset supports text classification tasks like emotion classification, sentiment analysis, textual entailment, and aspect-based sentiment analysis (ABSA) making it versatile for testing different aspects of language understanding models. Further, the sequence labeling datasets include subtasks such as part-of-speech tagging, span extraction, and named entity recognition.

NusaX NusaX (Winata et al., 2023) is a multilingual benchmark that focuses on assessing the capabilities of NLU performance of language models across 10 low-resource local Indonesian languages, with the addition of Indonesian and English. The dataset was originally the IndoNLU's SmSA sentiment analysis dataset, which was then translated into 11 other languages. Its main task is therefore sentiment analysis, although the dataset is likewise usable for machine translation purposes. For the evaluation of our model, we utilized the sentiment analysis dataset only.

NusaWrites NusaWrites (Cahyawijaya et al., 2023b) is a multilingual benchmark that serves as an extension of NusaX (Winata et al., 2023) and encompasses 12 underrepresented and low-resource languages in Indonesia. By its design, NusaWrites is more locally nuanced than generic corpora like Wikipedia and is lexically more diverse. It contains 2 sub-corpus defined by the way the data is constructed, topic-focused paragraph writing from human annotators (NusaParagraph) and human translation by native speakers (NusaTranslation). NusaParagraph contains three downstream tasks which include topic classification, emotion classification, and rhetoric mode classification. On the other hand, NusaTranslation contains three parallel downstream tasks which are sentiment analysis, emotion classification, and machine translation. Like NusaX, NusaTranslation is a translated version of IndoNLU's EmoT emotion classification dataset and IndoLEM's sentiment analysis dataset (Koto et al., 2020).

3.3.2 Benchmarking Models

We compared the performance of our NusaBERT models against the reported benchmark results without any further fine-tuning of the baseline models. The IndoNLU benchmark results include monolingual Indonesian language models IndoBERT$_{\text{BASE}}$, IndoBERT$_{\text{LARGE}}$, IndoBERT-lite$_{\text{BASE}}$, IndoBERT-lite$_{\text{LARGE}}$, as well as multilingual language models like mBERT (Devlin et al., 2018), XLM-MLM (Conneau and Lample, 2019), and both XLM-R$_{\text{BASE}}$ and XLM-R$_{\text{LARGE}}$ (Conneau et al., 2020). Addition-

ally, NusaX (Winata et al., 2023) and NusaWrites (Cahyawijaya et al., 2023b) report on the same set of models, with the inclusion of the IndoLEM IndoBERT (Koto et al., 2020), and classical machine learning models.

3.3.3 Fine-Tuning Setup

To fairly compare our results with the baselines, we adhere to similar fine-tuning procedures outlined in their respective benchmark codebases. Appendix C details the hyperparameters employed for fine-tuning the models across various tasks, reflecting the benchmarks' recommended settings with minor adjustments to learning rates and batch sizes for certain tasks. For IndoNLU, NusaX, and NusaTranslation benchmarks, we used a sequence length of 128, while for NusaParagraph, we increased the sequence length to 512 due to its much longer input text length. We applied early stopping based on the evaluation metrics and chose the best-scoring model. All fine-tuning processes utilize the Trainer API from Hugging Face's transformers library (Wolf et al., 2020). For other hyperparameters not mentioned in Appendix C, we followed the default hyperparameter from the Trainer API.

3.3.4 Evaluation Metrics

We evaluated the performance of our fine-tuned models using the macro-averaged F1 score for classification tasks, as specified in the IndoNLU, NusaX, and NusaWrites. Likewise, we followed the sequence labeling evaluation procedure used for CoNLL for token classification tasks of IndoNLU.

4 Results and Analysis

4.1 Pre-training Results

Both NusaBERT$_{\text{BASE}}$ and NusaBERT$_{\text{LARGE}}$ converged smoothly during the continued pre-training phase (§3.2). After 500,000 steps, NusaBERT$_{\text{BASE}}$ achieved an evaluation loss of 1.488 (4.427 PPL). Similarly, NusaBERT$_{\text{LARGE}}$ achieved a lower evaluation loss of 1.327 (3.769 PPL).

4.2 Fine-tuning Results

IndoNLU We report the official baseline results as well as the results of NusaBERT in Table 1. As shown, our models' performance on classification tasks of IndoNLU slightly deteriorates from that of the original IndoBERT models. The average score of NusaBERT decreases by about 1-2%, with NusaBERT$_{\text{BASE}}$ decreasing from 85.41% to 84.28% (-1.13%) and NusaBERT$_{\text{LARGE}}$ decreasing from 88.43% to 86.84% (-1.59%). Our models struggle particularly with aspect-based sentiment analysis tasks (CASA and HoASA), and the NusaBERT$_{\text{LARGE}}$ result on SmSA drops by 5%. In contrast, NusaBERT significantly improves the sequence labeling results of IndoBERT, increasing the average score by about 2-3%. NusaBERT$_{\text{BASE}}$ improves the base IndoBERT model score from 77.47% to 79.86% ($+2.39\%$), while NusaBERT$_{\text{LARGE}}$ improves the score from 81.21% to 84.09% ($+2.88\%$). NusaBERT especially improves the results on part-of-speech tagging tasks (POSP, BaPOS) and named entity recognition tasks (NERGrit, NERP).

Further, since the results of IndoBERT are similar to those of multilingual models like XLM-R, we observed a similar trend when comparing NusaBERT with the latter. That is, our models are slightly worse on classification tasks (-0.87% NusaBERT$_{\text{BASE}}$, -1.43% NusaBERT$_{\text{LARGE}}$), yet better on sequence labeling tasks ($+0.1\%$ NusaBERT$_{\text{BASE}}$, $+2.17\%$ NusaBERT$_{\text{LARGE}}$) than XLM-R. These indicate that our models remain competitive on Indonesian NLU tasks, retaining most of its initial knowledge found in the base IndoBERT model. Further experiments are required to fully retain and improve the results of IndoBERT across all tasks while still introducing multilingual capabilities to NusaBERT.

NusaX The official baseline and NusaBERT results on NusaX are shown in Table 2. From the baseline result, the monolingual IndoBERT models outperformed larger multilingual models like mBERT and are on par against XLM-R models despite being trained only on Indonesian texts, suggesting strong transferability from Indonesian to regional languages (Winata et al., 2023). It thus remains whether NusaBERT's introduction to regional languages will benefit the model when fine-tuned on multilingual, regional language tasks. On average, our models improve the results of both size-variants of IndoBERT. The NusaBERT$_{\text{BASE}}$ model increases the average score from 78.5% to 79.8% ($+1.3\%$) while NusaBERT$_{\text{LARGE}}$ increases the average score from 80.0% to 82.6% ($+2.6\%$). In particular, NusaBERT significantly improves the results on most languages that were included during the continued pre-training phase such as Acehnese (ace), Balinese (ban), Banjarese (bjn), Buginese (bug), Javanese (jav), and Sundanese (sun). How-

Model	Classification						Sequence Labeling							
	EmoT	SmSA	CASA	HoASA	WReTE	μ	POSP	BaPOS	TermA	KEPS	NERGrit	NERP	FacQA	μ
mBERT	67.30	84.14	72.23	84.63	84.40	78.54	91.85	83.25	89.51	64.31	75.02	69.27	61.29	76.36
XLM-MLM	65.75	86.33	82.17	88.89	64.35	77.50	95.87	88.40	90.55	65.35	74.75	75.06	62.15	78.88
XLM-R_{BASE}	71.15	91.39	91.71	91.57	79.95	85.15	95.16	84.64	90.99	68.82	79.09	75.03	64.58	79.76
XLM-R_{LARGE}	78.51	92.35	92.40	**94.27**	83.82	88.27	92.73	87.03	91.45	70.88	78.26	78.52	**74.61**	81.93
IndoBERT-lite$_{BASE}$	73.88	90.85	89.68	88.07	82.17	84.93	91.40	75.10	89.29	69.02	66.62	46.58	54.99	70.43
+ phase two	72.27	90.29	87.63	87.62	83.62	84.29	90.05	77.59	89.19	69.13	66.71	50.52	49.18	70.34
IndoBERT-lite$_{LARGE}$	75.19	88.66	90.99	89.53	78.98	84.67	91.56	83.74	90.23	67.89	71.19	74.37	65.50	77.78
+ phase two	70.80	88.61	88.13	91.05	**85.41**	84.80	94.53	84.91	90.72	68.55	73.07	74.89	62.87	78.51
IndoBERT$_{BASE}$	75.48	87.73	93.23	92.07	78.55	85.41	95.26	87.09	90.73	70.36	69.87	75.52	53.45	77.47
+ phase two	76.28	87.66	93.24	92.70	78.68	85.71	95.23	85.72	91.13	69.17	67.42	75.68	57.06	77.34
IndoBERT$_{LARGE}$	77.08	**92.72**	**95.69**	93.75	82.91	**88.43**	95.71	90.35	91.87	71.18	77.60	79.25	62.48	81.21
+ phase two	**79.47**	92.03	94.94	93.38	80.30	88.02	95.34	87.36	**92.14**	71.27	76.63	77.99	68.09	81.26
NusaBERT$_{BASE}$	76.10	87.46	91.26	89.80	76.77	84.28	95.77	96.02	90.54	66.67	72.93	82.29	54.81	79.86
NusaBERT$_{LARGE}$	78.90	87.36	92.13	93.18	82.64	86.84	**96.89**	**96.76**	91.73	**71.53**	**79.86**	**85.12**	66.77	**84.09**

Table 1: Evaluation results of baseline models and NusaBERT on the IndoNLU benchmark, measured in macro-F1 (%). Baseline results are obtained from Wilie et al. (2020). The best performance on each task is **bolded**.

Model	ace	ban	bbc	bjn	bug	eng	ind	jav	mad	min	nij	sun	μ
Logistic Regression	77.4	76.3	76.3	75.0	77.2	75.9	74.7	73.7	74.7	74.8	73.4	75.8	75.4
Naive Bayes	72.5	72.6	73.0	71.9	73.7	76.5	73.1	69.4	66.8	73.2	68.8	71.9	72.0
SVM	75.7	75.3	**76.7**	74.8	**77.2**	75.0	78.7	71.3	73.8	76.7	75.1	74.3	75.4
mBERT	72.2	70.6	69.3	70.4	68.0	84.1	78.0	73.2	67.4	74.9	70.2	74.5	72.7
XLM-R_{BASE}	73.9	72.8	62.3	76.6	66.6	90.8	88.4	78.9	69.7	79.1	75.0	80.1	76.2
XLM-R_{LARGE}	75.9	77.1	65.5	86.3	70.0	**92.6**	91.6	84.2	74.9	83.1	73.3	**86.0**	80.0
IndoLEM IndoBERT$_{BASE}$	72.6	65.4	61.7	71.2	66.9	71.2	87.6	74.5	71.8	68.9	69.3	71.7	71.1
IndoNLU IndoBERT$_{BASE}$	75.4	74.8	70.0	83.1	73.9	79.5	90.0	81.7	77.8	82.5	75.8	77.5	78.5
IndoNLU IndoBERT$_{LARGE}$	76.3	79.5	74.0	83.2	70.9	87.3	90.2	85.6	77.2	82.9	75.8	77.2	80.0
NusaBERT$_{BASE}$	76.5	78.7	74.0	82.4	71.6	84.1	89.7	84.1	75.6	80.8	74.9	85.2	79.8
NusaBERT$_{LARGE}$	**81.8**	**82.8**	74.7	**86.5**	73.4	84.6	**93.3**	**87.2**	**82.5**	**83.5**	**77.7**	82.7	**82.6**

Table 2: Evaluation results of baseline models and NusaBERT on NusaX sentiment analysis, measured in macro-F1 (%). Baseline results are obtained from Winata et al. (2023). The best performance on each task is **bolded**.

ever, this improvement is not consistent across all cases, particularly noting a slight decline in the performance of NusaBERT$_{BASE}$, even for languages included in the continued pre-training phase. Moreover, the results of languages not included in the continued pre-training phase like Madurese (mad) and Ngaju (nij) are still improved especially in NusaBERT$_{LARGE}$.

Overall, NusaBERT$_{LARGE}$ attained state-of-the-art results on most languages of NusaX, except for English (eng) and Sundanese (sun). XLM-R, which was pre-trained on these two languages (Conneau et al., 2020), is unsurprisingly still best. Likewise, classical machine learning algorithms like SVM and Logistic Regression achieved the highest scores on Buginese (bug) and Toba Batak (bbc), two extremely low-resource languages. Our findings align with the suggestion of Winata et al. (2023) whereby these languages are highly distinct from other languages of Indonesia and hence do not exhibit strong cross-lingual transferability. We also note that both languages stem from different language families than most of the other languages, even though they are all grouped into one Malayo-Polynesian subgroup (Eberhard et al., 2022). Buginese (bug) is spoken mostly in the South Sulawesi region, while Toba Batak (bbc) is spoken primarily in the Northwestern Sumatra and Barrier Islands regions. In addition, while Buginese (bug) is included in our pre-training corpus, it is the third smallest subset within our Wikipedia dataset, with only about 9,000 documents. Therefore, it remains our interest to find other ways to improve the results of languages that are not only extremely low-resource but are also highly distinct from other languages of Indonesia.

NusaWrites The official baseline result of NusaWrites aggregates the scores across all languages into a single mean score for each subtask (Cahyawijaya et al., 2023b). Fortunately, the individual raw results for each subtask and each language are available on the official NusaWrites repository[5], enabling us to thoroughly examine and compare per-language results. The aggregated baseline and NusaBERT results are shown in Table 3, while the detailed per-task and per-language results are

[5] https://github.com/IndoNLP/nusa-writes/

Model	NusaP			NusaT	
	Emot.	Rhet.	Topic	Emot.	Senti.
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
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
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
NusaBERT$_{BASE}$	67.18	51.34	83.32	56.54	77.07
NusaBERT$_{LARGE}$	71.82	**53.06**	85.08	**61.40**	**79.54**

Table 3: Evaluation results of baseline models and NusaBERT on the NusaWrites benchmark tasks, measured in macro-F1 (%) and averaged over all of the languages found in each task. Detailed per-task and per-language results are shown in Appendix D. Baseline results are obtained from Cahyawijaya et al. (2023b). The best performance on each task is **bolded** for clarity.

Dataset	Prop. (%)	Relative Improvement	
		IndoBERT$_{BASE}$	IndoBERT$_{LARGE}$
NusaX	8.46	+1.3	+2.6
NusaT Emotion	8.21	+2.04	+3.61
NusaT Sentiment	6.83	+1.83	+2.14
NusaP Topic	11.26	−2.55	−0.33
NusaP Rhetoric	11.63	+3.42	+21.30
NusaP Emotion	11.41	+0.06	+9.18

Table 4: Proportion of new tokens found only in the extended NusaBERT tokenizer compared with the performance gain of NusaBERT over IndoBERT for each dataset.

shown in Appendix D.

Like our results on NusaX, NusaBERT increases the average score on the two tasks of NusaTranslation. Specifically, NusaBERT$_{BASE}$ improves the NusaTranslation emotion classification score of IndoBERT from 52.59% to 57.80% (+5.21%) and NusaBERT$_{LARGE}$ from 54.50% to 61.40% (+6.9%). Further, on the sentiment analysis task, NusaBERT$_{BASE}$ improves the IndoBERT score from 75.24% to 77.07% (+1.83%) and NusaBERT$_{LARGE}$ from 77.40% to 79.54% (+2.14%). Overall, NusaBERT$_{LARGE}$ is state-of-the-art on both NusaTranslation tasks.

Notably, unlike NusaX, most languages of NusaTranslation are not found in the pre-training corpus of NusaBERT and are extremely low-resource. Nonetheless, based on the results alone, it seems that the introduction of additional new regional languages during the continued pre-training phase benefits the robustness of NusaBERT on these new languages as well, suggesting cross-lingual transferability. Similarly, NusaBERT's results on languages that were included in the continued pre-training corpus like Javanese (`jav`) and Minangkabau (`min`) significantly improve that of IndoBERT. However, as noted by Cahyawijaya et al. (2023b), NusaTranslation and NusaX share a similar source domain of social media texts, therefore it is expected that our findings are parallel.

NusaParagraph, on the contrary, presents a more challenging task by consisting of not only languages that are not found in our pre-training corpus but is also lexically more diverse and contains a remarkably higher ratio of local/colloquial words (Cahyawijaya et al., 2023b). Indeed, the gains of NusaBERT over IndoBERT are lacklustre when evaluated on the NusaParagraph topic classification task. For instance, NusaBERT failed to improve the results of IndoBERT, dropping the result of IndoBERT$_{BASE}$ from 85.87% to 83.32% (−2.55%), and for NusaBERT$_{LARGE}$, the result dropped from 85.41% to 85.08% (−0.33%). Nevertheless, it still improved the IndoBERT results on both the rhetorical mode (+3.42% NusaBERT$_{BASE}$, +21.3% NusaBERT$_{LARGE}$) and emotion classification (+0.06% NusaBERT$_{BASE}$, +9.18% NusaBERT$_{LARGE}$) tasks. It is only on NusaParagraph rhetorical mode classification where NusaBERT$_{LARGE}$ is state-of-the-art.

Like the findings of Cahyawijaya et al. (2023b), NusaBERT fails to outperform classical machine learning baselines on languages that are highly distinct from Indonesian (`ind`). We also note that NusaBERT was pre-trained on Wikipedia and common crawl corpora, which explains its effectiveness on and closeness to NusaX and NusaTranslation source domains, but not so for NusaParagraph. Due to the high linguistic and lexical discrepancies found in NusaParagraph (Cahyawijaya et al., 2023b), NusaBERT's capabilities to exploit knowledge and cross-lingual transfer to these extremely low-resource languages remain largely ineffective.

4.3 Impact of New Tokens

We investigated the impact of the new tokens on downstream tasks, especially noting that our extended tokenizer was additionally trained on the regional languages of Indonesia and that the IndoBERT tokenizer might not be suitable for this purpose. We modified the approach conducted by Sriwirote et al. (2023), where they calculated the proportion of unassimilated English words with respect to the number of total words in the downstream task. However, since we are unable to dis-

tinguish the regional languages' words from the Indonesian words programmatically, we defined a new metric as follows:

$$\text{Proportion of New Tokens} = \frac{\#\text{new tokens}}{\#\text{total tokens}} \quad (1)$$

We re-tokenized all downstream tasks' texts using the extended NusaBERT tokenizer and calculated the percentage of new tokens with respect to the total number of tokens. This way, we can closely inspect and compare the relation between the newly introduced tokens and the gains of NusaBERT over IndoBERT. Table 4 shows the aforementioned results. While the trend of the proportion of new tokens with the gains of NusaBERT over IndoBERT isn't always linear, there is generally a correlation between the two – parallel with the findings of Sriwirote et al. (2023). This, however, doesn't apply to NusaParagraph topic classification where NusaBERT performed worse than IndoBERT. Despite these findings, the new tokens might not definitively be the only factor behind the improved results of NusaBERT (e.g. continued pre-training), and further investigation is required. We analyzed tokenizer fertility, comparing NusaBERT's extended tokenizer to IndoBERT's original tokenizer in Appendix E.

4.4 Code-mixing Robustness

Although NusaBERT doesn't directly address the issue of code-mixing, we examined its code-mixing robustness by evaluating our models on IndoRobusta-Blend (Adilazuarda et al., 2022). Following its procedure, we took NusaBERT models which have been fine-tuned on the original Indonesian EmoT and SmSA datasets, and conducted zero-shot inference on code-mixed versions of their respective test sets. To have a fair comparison with the official reported results, we similarly applied a perturbation ratio $R = 0.4$ and mixed English (`eng`), Javanese (`jav`), Malay (`msa`), and Sundanese (`sun`) as target L2 languages. We report the evaluation results in Table 5. We also provided the full results in Appendix F.

Interestingly, the robustness of NusaBERT depends highly on the downstream task being tested, similar to the findings of Adilazuarda et al. (2022). On sentiment analysis (SmSA), NusaBERT$_{\text{BASE}}$ is the most robust, significantly improving the robustness of IndoBERT$_{\text{BASE}}$. However, this doesn't apply to emotion classification (EmoT) where NusaBERT$_{\text{LARGE}}$ is more robust than its NusaBERT$_{\text{BASE}}$. Further, both NusaBERT models are more prone to code-mixing on emotion classification compared to IndoBERT, but the opposite is true for sentiment analysis. Additionally, parallel to what was conjectured by Adilazuarda et al. (2022), NusaBERT is generally more robust against Indonesian-English code-mixing. We agree with their suggestion that this stems from the source bias found in most online pre-training corpora that often mix these two languages. In the same light, Wikipedia texts that we pre-trained on also contain a high ratio of English loan words (Cahyawijaya et al., 2023b), thereby explaining these findings.

Model	ind	eng	jav	msa	sun	μ
EmoT						
mBERT	61.14	12.50	14.02	12.73	12.50	12.94
XLM-R$_{\text{BASE}}$	72.88	10.98	13.94	13.18	12.50	12.65
XLM-R$_{\text{LARGE}}$	<u>78.26</u>	12.27	13.03	12.42	11.74	12.37
IndoBERT$_{\text{BASE}}$	72.42	9.55	12.35	**9.47**	9.39	**10.19**
IndoBERT$_{\text{LARGE}}$	75.53	**9.24**	12.12	10.23	**9.32**	10.23
NusaBERT$_{\text{BASE}}$	75.23	14.09	14.77	13.64	13.64	14.03
NusaBERT$_{\text{LARGE}}$	78.18	10.45	**10.45**	10.45	12.05	10.85
SmSA						
mBERT	83.00	2.20	3.00	2.93	2.47	2.65
XLM-R$_{\text{BASE}}$	91.53	3.40	3.80	4.27	4.27	3.94
XLM-R$_{\text{LARGE}}$	<u>94.07</u>	2.13	3.20	2.60	2.73	2.67
IndoBERT$_{\text{BASE}}$	91.00	1.33	5.07	3.20	2.40	3.00
IndoBERT$_{\text{LARGE}}$	94.20	2.47	4.13	4.00	2.20	3.20
NusaBERT$_{\text{BASE}}$	91.00	**0.60**	**2.80**	2.40	**1.80**	**1.90**
NusaBERT$_{\text{LARGE}}$	91.00	1.80	3.80	**2.20**	2.20	2.50

Table 5: Evaluation results on code-mixed downstream tasks, measured in delta accuracy with $R = 0.4$. Baseline results are obtained from Adilazuarda et al. (2022). The lowest delta accuracy on each task is **bolded** for clarity. The best-performing model on the originally Indonesian (`ind`) fine-tuning task has also been <u>underlined</u>.

5 Conclusion

In this study, we introduced NusaBERT, a multilingual language model specifically tailored to the linguistic diversity of Indonesia. Basing our model on IndoBERT, we applied vocabulary expansion and continued pre-training on a multilingual corpus that introduces the regional languages of Indonesia. NusaBERT achieves state-of-the-art results when evaluated on Indonesian and multilingual NLU benchmarks such as IndoNLU, NusaX, and NusaWrites. These findings highlight the effectiveness of our proposed approach in enhancing the multilingual and multicultural capabilities of IndoBERT to address Indonesia's unique linguistic framework. We also discussed several limitations of NusaBERT and how to potentially resolve them. We hope NusaBERT will enable further research in the under-represented languages of Indonesia.

Limitations

Code-mixing NusaBERT demonstrates proficiency in handling low-resource languages while surpassing or remaining competitive with monolingual models on downstream tasks. Despite this efficacy, it has yet to address the intricate challenge of intra-sentential code-mixing. While the issue of code-mixing is not explicitly tackled in the context of NusaBERT, results in Table 5 indicate potential room for improvements that can be done to enhance NusaBERT's performance in handling code-mixing scenarios. Moreover, it is important to mention that the language model's performance on IndoRobusta-Blend does not definitively represent its robustness against code-mixing as it uses synthetically generated code-mixed examples instead of human-curated code-mixed data, and is limited to only four L2 languages. Having an expert-curated code-mixing benchmark would be valuable for future evaluations.

To tackle code-mixing adversarial attacks, Adilazuarda et al. (2022) proposed a code-mixing adversarial training technique called IndoRobusta-Shot that suggests three different fine-tuning techniques: code-mixed-only tuning, two-step tuning, and joint training. Among the three examined methods, joint training shows the best results which implies that training code-mixed data with monolingual data increases the robustness of language models while maintaining its monolingual downstream capabilities.

Adapting NusaBERT to New Languages In our study, we introduced a multilingual language model designed for Indonesian and its 12 regional languages. Although 12 languages is considerably a large number, it is considered comparatively modest compared to Indonesia's boasting rich linguistic landscape with over 700 languages and dialects. This arises from the significant difference in the amount of available text corpus of regional languages and the lack of quality data.

Several endeavors have successfully extended new languages to a base language model. For example, the BLOOM language model (BigScience Workshop et al., 2023), a comprehensive multilingual language model trained on 46 languages, effectively extended its applicability to 8 previously unseen languages (Yong et al., 2023) through continued pretraining, implementation of language adapters (Pfeiffer et al., 2020), and parameter-efficient finetuning techniques (Liu et al., 2022). These strategies facilitated the inclusion of new languages while preserving existing capabilities and mitigating catastrophic forgetting. Despite the demonstrated feasibility of extending language models to existing language models, the data on these new languages are abundant in comparison to Indonesian regional languages.

A recent approach proposed by Wang et al. (2022) seeks to leverage bilingual lexicons which are widely available even for extremely low-resource languages. We can thereby potentially generate synthetic low-resource language texts by translating from Indonesian texts using these lexicons. This approach, coupled with gold few-texts of the target language, if available, is one way to possibly extend NusaBERT to extremely low-resource languages where resources are scarce.

Corpus Domain Diversity One significant limitation in our study is the lack of corpus domain diversity, particularly evident in the performance discrepancies between NusaParagraph and the other tasks (NusaX and NusaTranslation). The underpinning challenge with NusaParagraph, which diverges from the social media domain to include paragraph writing by human annotators, is its richer cultural and lexical diversity, indicative of the nuanced and colloquial language use in very low-resource and linguistically distinct local languages (Cahyawijaya et al., 2023b). This complexity is inherently difficult for models like NusaBERT, which, despite their robustness, are pre-trained predominantly on social media texts and online documents similar to the datasets used for NusaX and NusaTranslation.

Despite the apparent scarcity of directly applicable, culturally rich, and linguistically aligned corpora for very low-resource local languages, there exists an opportunity to leverage alternative texts during model pre-training. For instance, texts such as the Bible, which are often translated into numerous languages, including many under-represented ones, could provide a valuable resource (Wongso et al., 2023). These texts offer a range of linguistic structures and vocabularies that, while not entirely reflective of colloquial use, could serve as a foundational step towards bridging the gap in language representation. This approach underscores the necessity for creative solutions in the absence of conventional data sources, aiming to enhance the model's performance across a wider array of

linguistic contexts.

This strategy invites further research to not only incorporate existing texts from under-represented languages into pre-training processes but also to innovate methods such as leveraging and exploring the use of non-text data. Specifically, transcribing conversation audio through speech recognition, especially for local Indonesian languages that are rarely ever written (Aji et al., 2022), presents a novel avenue to enrich the language's resources. This approach can capture the authentic linguistic nuances and cultural richness of spoken language, offering a more comprehensive representation of these languages (Besacier et al., 2014).

This direction not only underscores the ongoing effort to fully leverage the linguistic diversity of Indonesia and similar regions but also aims to expand the applicability and inclusivity of language models by incorporating the rich, oral traditions of under-represented communities into the digital linguistic landscape.

References

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 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. 2022. 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.

Mehdi Ali, Michael Fromm, Klaudia Thellmann, Richard Rutmann, Max Lübbering, Johannes Leveling, Katrin Klug, Jan Ebert, Niclas Doll, Jasper Schulze Buschhoff, et al. 2023. Tokenizer choice for llm training: Negligible or crucial? *arXiv preprint arXiv:2310.08754*.

Badan Pusat Statistik. 2010. Kewarganegaraan suku bangsa, agama, bahasa 2010.

Laurent Besacier, Etienne Barnard, Alexey Karpov, and Tanja Schultz. 2014. Automatic speech recognition for under-resourced languages: A survey. *Speech Communication*, 56:85–100.

BigScience Workshop, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, et al. 2023. Bloom: A 176b-parameter open-access multilingual language model.

Samuel Cahyawijaya, Holy Lovenia, Alham Fikri Aji, Genta Winata, Bryan Wilie, Fajri Koto, Rahmad Mahendra, Christian Wibisono, Ade Romadhony, Karissa Vincentio, Jennifer Santoso, David Moeljadi, Cahya Wirawan, Frederikus Hudi, Muhammad Satrio Wicaksono, Ivan Parmonangan, Ika Alfina, Ilham Firdausi Putra, Samsul Rahmadani, Yulianti Oenang, Ali Septiandri, James Jaya, Kaustubh Dhole, Arie Suryani, Rifki Afina Putri, Dan Su, Keith Stevens, Made Nindyatama Nityasya, Muhammad Adilazuarda, Ryan Hadiwijaya, Ryandito Diandaru, Tiezheng Yu, Vito Ghifari, Wenliang Dai, Yan Xu, Dyah Damapuspita, Haryo Wibowo, Cuk Tho, Ichwanul Karo Karo, Tirana Fatyanosa, Ziwei Ji, Graham Neubig, Timothy Baldwin, Sebastian Ruder, Pascale Fung, Herry Sujaini, Sakriani Sakti, and Ayu Purwarianti. 2023a. NusaCrowd: Open source initiative for Indonesian NLP resources. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13745–13818, Toronto, Canada. Association for Computational Linguistics.

Samuel Cahyawijaya, Holy Lovenia, Fajri Koto, Dea Adhista, Emmanuel Dave, Sarah Oktavianti, Salsabil Akbar, Jhonson Lee, Nuur Shadieq, Tjeng Wawan Cenggoro, Hanung Linuwih, Bryan Wilie, Galih Muridan, Genta Winata, David Moeljadi, Alham Fikri Aji, Ayu Purwarianti, and Pascale Fung. 2023b. NusaWrites: Constructing high-quality corpora for underrepresented and extremely low-resource languages. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 921–945, Nusa Dua, Bali. Association for Computational Linguistics.

Samuel Cahyawijaya, Holy Lovenia, Fajri Koto, Rifki Putri, Wawan Cenggoro, Jhonson Lee, Salsabil Akbar, Emmanuel Dave, Nuurshadieq Nuurshadieq, Muhammad Mahendra, Rr Putri, Bryan Wilie, Genta Winata, Alham Aji, Ayu Purwarianti, and Pascale Fung. 2024. Cendol: Open instruction-tuned generative large language models for Indonesian languages. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14899–14914, Bangkok, Thailand. Association for Computational Linguistics.

Samuel Cahyawijaya, Genta Indra Winata, Bryan Wilie, Karissa Vincentio, Xiaohong Li, Adhiguna Kuncoro, Sebastian Ruder, Zhi Yuan Lim, Syafri Bahar, Masayu Khodra, Ayu Purwarianti, and Pascale Fung. 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, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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 and Guillaume Lample. 2019. Cross-lingual language model pretraining. *Advances in neural information processing systems*, 32.

M.C. Corporation. 2007. *World and Its Peoples: Eastern and Southern Asia*. Number v. 10 in World and Its Peoples Series. Marshall Cavendish.

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*.

Zoltan Csaki, Pian Pawakapan, Urmish Thakker, and Qiantong Xu. 2023. Efficiently adapting pretrained language models to new languages.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

David M Eberhard, Gary F Simons, and Charles D Fennig. 2022. Ethnologue: Languages of the world.

John Hewitt. 2021. Initializing new word embeddings for pretrained language models.

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, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Haokun Liu, Derek Tam, Mohammed Muqeeth, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin A Raffel. 2022. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. *Advances in Neural Information Processing Systems*, 35:1950–1965.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Lalita Lowphansirikul, Charin Polpanumas, Nawat Jantrakulchai, and Sarana Nutanong. 2021. Wangchanberta: Pretraining transformer-based thai language models.

Thuat Nguyen, Chien Van Nguyen, Viet Dac Lai, Hieu Man, Nghia Trung Ngo, Franck Dernoncourt, Ryan A Rossi, and Thien Huu Nguyen. 2023. Culturax: A cleaned, enormous, and multilingual dataset for large language models in 167 languages. *arXiv preprint arXiv:2309.09400*.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2020. MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7654–7673, Online. Association for Computational Linguistics.

Kunat Pipatanakul, Phatrasek Jirabovonvisut, Potsawee Manakul, Sittipong Sripaisarnmongkol, Ruangsak Patomwong, Pathomporn Chokchainant, and Kasima Tharnpipitchai. 2023. Typhoon: Thai large language models.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2023. Exploring the limits of transfer learning with a unified text-to-text transformer.

Panyut Sriwirote, Jalinee Thapiang, Vasan Timtong, and Attapol T. Rutherford. 2023. Phayathaibert: Enhancing a pretrained thai language model with unassimilated loanwords.

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.

S.K.N.M.H. Wahyudi, S.P.M.P. Bivit Anggoro Prasetyo Nugroho, M.P. Dra. Isnaeni Praptanti, G. Rizky, G. Dullosa, M. Kika, and A. Offset. *Bahasa Indonesia Kesehatan*. Penerbit Andi.

Xinyi Wang, Sebastian Ruder, and Graham Neubig. 2022. Expanding pretrained models to thousands more languages via lexicon-based adaptation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 863–877, Dublin, Ireland. Association for Computational Linguistics.

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*, pages 815–834, Dubrovnik, Croatia. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Wilson Wongso, Ananto Joyoadikusumo, Brandon Scott Buana, and Derwin Suhartono. 2023. Many-to-many multilingual translation model for languages of indonesia. *IEEE Access*, 11:91385–91397.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation.

Zheng Xin Yong, Hailey Schoelkopf, Niklas Muennighoff, Alham Fikri Aji, David Ifeoluwa Adelani, Khalid Almubarak, M Saiful Bari, Lintang Sutawika, Jungo Kasai, Ahmed Baruwa, Genta Winata, Stella Biderman, Edward Raff, Dragomir Radev, and Vassilina Nikoulina. 2023. BLOOM+1: Adding language support to BLOOM for zero-shot prompting. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11682–11703, Toronto, Canada. Association for Computational Linguistics.

A Evaluation Benchmarks

The list of downstream benchmarks/datasets used to evaluate NusaBERT is shown in Table 6.

Dataset	Task
Single-sentence Classification	
EmoT	Emotion Classification
SmSA	Sentiment Analysis
NusaX	Sentiment Analysis
NusaT Sentiment	Sentiment Analysis
NusaT Emotion	Emotion Classification
NusaP Emotion	Emotion Classification
NusaP Rhetorical	Rhetorical Mode Classification
NusaP Topic	Topic Modeling
Single-sentence Multi-label Classification	
CASA	Aspect-based Sentiment Analysis
HoASA	Aspect-based Sentiment Analysis
Sequence-pair Classification	
WReTE	Textual Entailment
Token Classification	
POSP	Part-of-Speech Tagging
BaPOS	Part-of-Speech Tagging
TermA	Span Extraction
KEPS	Span Extraction
NERGrit	Named Entity Recognition
NERP	Named Entity Recognition
Sequence-Pair Token Classification	
FacQA	Span Extraction

Table 6: List of downstream evaluation benchmarks for NusaBERT fine-tuning.

B Statistics

A statistical summary of the number of documents per language included in the pre-training corpus is

shown in Table 7, while the list of languages and dialects included in this study and their statistics are shown in Table 8.

Language (ISO 639-3)	#documents
Indonesian (ind)	23,905,655
Javanese (jav)	1,229,867
Sundanese (sun)	957,674
Acehnese (ace)	805,498
Malay (msa)	584,186
Minangkabau (min)	339,181
Banjarese (bjn)	306,751
Balinese (ban)	264,382
Gorontalo (gor)	14,514
Banyumasan (jav)	11,832
Buginese (bug)	9,793
Nias (nia)	1,650
Tetum (tet)	1,465
Total	28,432,448

Table 7: A summary of the number of documents per language in the pre-training corpus of NusaBERT.

C Hyperparameters

We provide the hyperparameters used for continued pre-training and downstream tasks in Table 9 and Table 10, respectively.

D NusaWrites Evaluation Results

We included the non-aggregated, per-task, and per-language evaluation results of NusaWrites. NusaTranslation results are shown in Table 12 and Table 13. NusaParagraph results are shown in Table 14, Table 15, and Table 16.

E Tokenizer Fertility

Fertility is a widely used metric for assessing tokenizer performance and is defined as the average number of tokens per word (Ali et al., 2023; Csaki et al., 2023; Cahyawijaya et al., 2024). A higher fertility score indicates lower compression efficiency, as more tokens are needed per word. To evaluate and compare the fertility of NusaBERT's extended tokenizer and IndoBERT's original tokenizer, we applied both to texts from downstream tasks. Fertility was calculated as the ratio of the total number of tokens to the total number of words, with words identified using whitespace splitting, following Ali et al. (2023). The results are summarized in Table 11, showing that NusaBERT's tokenizer has a lower tokenizer fertility and is thus more efficient than that of IndoBERT.

Language	Primary Region	#speakers
Acehnese (ace)	Aceh	2,840,000
Ambon (abs)	Maluku	1,650,900
Balinese (ban)	Bali	3,300,000
Banjarese (bjn)	Kalimantan	3,650,000
Banyumasan (jav)	Banyumasan	N/A
Batak (btk)	North Sumatra	3,320,000[†]
Betawi (bew)	Banten, Jakarta	5,000,000
Bima (bhp)	Sumbawa	500,000
Buginese (bug)	South Sulawesi	4,370,000
Gorontalo (gor)	Gorontalo	505,000
Indonesian (ind)	Indonesia	198,000,000
Javanese (jav)	Java	68,200,000
Madurese (mad)	Madura	7,790,000
Makassarese (mak)	South Sulawesi	2,110,000
Malay (msa)	Malaysia	82,285,706
Minangkabau (min)	West Sumatra	4,880,000
Musi (mui)	South Sumatra	3,105,000
Ngaju (nij)	Central Kalimantan	890,000
Nias (nia)	Nias	867,000
Rejang (rej)	Bengkulu	350,000
Sundanese (sun)	West Java	32,400,000
Tetum (tet)	East Timor	91,200
Toba Batak (bbc)	North Sumatra	1,610,000

Table 8: Statistics of languages included in this study, with data obtained from Eberhard et al. (2022) and [†]Badan Pusat Statistik (2010).

F IndoRobusta Evaluation Results

The evaluation results of baseline models and NusaBERT on IndoRobusta-Blend are shown in Table 17.

Hyperparameter	Value
Sequence length	128
Batch size	256
Peak learning rate	3e-4/3e-5†
#warmup steps	24,000
#optimization steps	500,000
Learning rate scheduler	Linear
Optimizer	AdamW
Adam (β_1, β_2)	(0.9, 0.999)
Adam ϵ	1e-8
Weight decay	0.01
PyTorch data type	bfloat16

Table 9: Continued pre-training hyperparameters. † indicates the differing values for NusaBERT$_{\text{BASE}}$ and NusaBERT$_{\text{LARGE}}$, respectively.

Classification task type	#epochs	Learning rate	Batch size	Weight decay
Sentence	100	1e-5/2e-5†	32/16†	0.01
Multi-label	100	1e-5	32	0.01
Token	10	2e-5	16	0.01

Table 10: Downstream fine-tuning hyperparameters. † indicates the differing values for NusaBERT$_{\text{BASE}}$ and NusaBERT$_{\text{LARGE}}$, respectively.

Dataset	Tokenizer Fertility	
	NusaBERT	IndoBERT
NusaX	1.770	1.787
NusaTranslation Emotion	1.910	1.924
NusaTranslation Sentiment	2.150	2.150
NusaParagraph Topic	1.743	1.761
NusaParagraph Rhetoric	1.724	1.750
NusaParagraph Emotion	1.747	1.771

Table 11: Tokenizer fertility comparison between NusaBERT's extended tokenizer and IndoBERT's original tokenizer. Higher fertility indicates lower tokenization efficiency.

NusaTranslation EmoT												
Model	**abs**	**bew**	**bhp**	**btk**	**jav**	**mad**	**mak**	**min**	**mui**	**rej**	**sun**	μ
Logistic Regression (Bag of Words)	46.77	62.31	44.62	59.38	60.66	58.05	55.63	61.73	45.33	45.61	62.90	56.18
Logistic Regression (TF-IDF)	**51.20**	63.59	**50.06**	61.25	61.47	60.42	56.39	63.94	50.98	50.61	62.99	
Naive Bayes (Bag of Words)	48.16	59.76	48.02	57.12	58.39	55.22	54.93	61.41	51.49	47.53	61.32	52.71
Naive Bayes (TF-IDF)	49.95	55.54	40.12	54.64	54.85	52.76	52.03	56.61	48.93	32.87	57.86	
SVM (Bag of Words)	44.56	61.30	43.59	58.43	58.97	55.97	52.60	61.02	48.80	41.81	60.58	55.08
SVM (TF-IDF)	48.23	61.74	48.68	61.02	63.34	59.43	58.09	62.34	51.40	48.27	61.58	
mBERT	26.05	59.75	12.65	59.28	62.80	57.30	54.92	61.50	16.48	12.24	62.49	44.13
XLM-R_{BASE}	35.79	63.54	12.44	59.95	62.86	59.87	60.54	63.39	13.94	19.75	65.14	47.02
XLM-R_{LARGE}	49.58	70.43	8.53	**65.83**	68.70	61.27	58.85	**70.84**	55.83	23.12	70.24	54.84
IndoLEM IndoBERT$_{BASE}$	35.03	67.86	25.40	59.86	64.47	59.40	58.23	61.48	45.00	39.20	62.56	52.59
IndoNLU IndoBERT$_{BASE}$	41.04	66.61	32.13	62.81	66.91	61.52	**61.81**	67.95	42.78	33.54	62.38	54.50
IndoNLU IndoBERT$_{LARGE}$	48.54	72.55	28.43	63.09	69.34	61.84	60.48	67.55	53.22	40.19	**70.53**	57.80
NusaBERT$_{BASE}$	45.21	66.09	39.03	61.72	67.41	61.10	60.54	67.11	50.98	37.36	65.34	56.54
NusaBERT$_{LARGE}$	47.75	**73.68**	36.31	62.87	**73.63**	**65.48**	60.58	70.27	**60.06**	**54.47**	70.34	**61.40**

Table 12: Evaluation results of baseline models and NusaBERT on the NusaTranslation emotion classification task, measured in macro-F1 (%). Baseline results are obtained from Cahyawijaya et al. (2023b). The best performance on each task is **bolded** for clarity.

NusaTranslation Senti												
Model	**abs**	**bew**	**bhp**	**btk**	**jav**	**mad**	**mak**	**min**	**mui**	**rej**	**sun**	μ
Logistic Regression (Bag of Words)	69.23	81.88	41.86	79.13	81.87	81.48	78.39	82.53	70.35	60.79	84.43	74.96
Logistic Regression (TF-IDF)	69.50	81.04	**70.10**	79.67	77.85	74.50	78.27	82.18	72.31	68.00	83.73	
Naive Bayes (Bag of Words)	69.67	79.12	69.36	78.05	79.88	78.38	76.77	80.10	72.20	69.05	80.51	74.89
Naive Bayes (TF-IDF)	67.71	77.03	64.51	76.56	75.71	77.70	76.41	80.11	71.41	66.90	80.34	
SVM (Bag of Words)	69.87	81.94	69.89	79.77	78.18	80.44	79.25	82.68	68.02	66.45	84.21	76.04
SVM (TF-IDF)	70.28	82.26	68.94	76.20	78.16	75.28	77.67	81.66	72.20	66.36	83.10	
mBERT	67.47	79.56	41.86	72.81	80.55	76.44	69.08	79.43	64.07	46.03	78.56	68.71
XLM-R_{BASE}	67.28	85.11	41.86	77.22	79.73	78.40	75.90	83.39	40.90	40.97	84.08	68.62
XLM-R_{LARGE}	**72.55**	86.54	65.52	80.62	86.13	78.58	**81.86**	86.04	**78.80**	65.18	**87.87**	79.06
IndoLEM IndoBERT$_{BASE}$	59.39	81.57	44.66	74.50	81.89	72.28	66.12	80.95	65.52	51.25	81.74	69.08
IndoNLU IndoBERT$_{BASE}$	70.45	86.09	62.80	72.64	84.34	75.16	76.80	82.62	71.32	66.59	78.82	75.24
IndoNLU IndoBERT$_{LARGE}$	72.16	87.92	59.91	78.39	81.61	79.84	78.96	81.99	75.98	68.79	85.83	77.40
NusaBERT$_{BASE}$	70.71	86.02	63.72	80.63	84.04	80.47	80.73	84.75	66.14	64.80	85.74	77.07
NusaBERT$_{LARGE}$	68.94	**90.11**	66.46	**83.09**	**86.71**	**83.66**	81.35	**86.42**	70.66	**69.74**	87.83	**79.54**

Table 13: Evaluation results of baseline models and NusaBERT on the NusaTranslation sentiment analysis task, measured in macro-F1 (%). Baseline results are obtained from Cahyawijaya et al. (2023b). The best performance on each task is **bolded** for clarity.

NusaParagraph Topic											
Model	**bew**	**btk**	**bug**	**jav**	**mad**	**mak**	**min**	**mui**	**rej**	**sun**	μ
Logistic Regression (Bag of Words)	90.20	88.95	68.87	90.65	88.87	87.50	90.70	85.71	82.22	89.67	**87.67**
Logistic Regression (TF-IDF)	92.63	**91.09**	**73.92**	91.49	**92.32**	91.21	92.10	88.02	86.39	90.87	
Naive Bayes (Bag of Words)	87.72	84.55	62.88	87.32	82.40	89.27	90.64	86.21	**88.09**	89.45	85.06
Naive Bayes (TF-IDF)	89.11	85.38	60.06	89.55	83.44	90.26	89.96	**88.20**	86.58	90.10	
SVM (Bag of Words)	89.48	85.59	61.46	87.79	86.49	84.85	89.55	82.51	78.36	88.28	85.86
SVM (TF-IDF)	91.76	90.25	73.57	90.64	90.61	**91.34**	**92.56**	86.06	84.88	91.19	
mBERT	89.22	86.66	43.26	87.41	77.40	84.61	88.75	83.30	9.54	88.00	73.82
XLM-R_{BASE}	90.11	86.84	46.11	89.82	83.59	84.22	88.19	3.45	54.23	90.26	71.68
XLM-R_{LARGE}	92.33	85.75	43.18	91.07	85.81	85.60	89.06	85.69	81.04	91.00	83.05
IndoLEM IndoBERT$_{BASE}$	91.74	87.23	61.53	90.52	86.50	87.96	90.82	85.00	78.77	88.59	84.87
IndoNLU IndoBERT$_{BASE}$	91.64	87.26	67.72	90.59	85.00	85.30	90.50	86.52	85.74	88.43	85.87
IndoNLU IndoBERT$_{LARGE}$	92.17	85.95	66.79	90.05	87.11	87.11	91.30	86.16	78.06	89.39	85.41
NusaBERT$_{BASE}$	91.81	87.27	52.45	91.45	87.48	87.61	91.97	83.05	77.57	91.03	83.32
NusaBERT$_{LARGE}$	**93.18**	87.20	60.97	**93.44**	85.80	88.93	92.25	87.15	77.48	**92.48**	85.08

Table 14: Evaluation results of baseline models and NusaBERT on the NusaParagraph topic classification task, measured in macro-F1 (%). Baseline results are obtained from Cahyawijaya et al. (2023b). The best performance on each task is **bolded** for clarity.

NusaParagraph Rhetoric											
Model	**bew**	**btk**	**bug**	**jav**	**mad**	**mak**	**min**	**mui**	**rej**	**sun**	μ
Logistic Regression (Bag of Words)	39.40	33.85	61.77	64.52	47.97	23.87	59.09	53.82	28.46	49.39	45.21
Logistic Regression (TF-IDF)	40.10	33.10	57.11	64.85	48.56	24.08	57.68	44.67	22.70	49.28	
Naive Bayes (Bag of Words)	37.78	28.23	51.29	56.94	42.62	22.78	46.92	35.55	20.95	44.79	37.73
Naive Bayes (TF-IDF)	36.79	26.06	44.02	53.68	42.89	22.98	44.67	32.65	20.72	42.22	
SVM (Bag of Words)	41.51	32.04	60.55	67.12	48.21	23.25	59.50	50.09	31.76	49.98	45.44
SVM (TF-IDF)	40.76	32.60	57.29	65.07	48.28	22.22	57.79	45.51	26.13	49.18	
mBERT	43.21	24.92	70.26	74.29	53.02	17.52	67.37	61.67	32.85	54.94	50.01
XLM-R_{BASE}	48.75	23.08	70.03	78.04	52.09	8.28	68.60	61.80	22.83	58.17	49.17
XLM-R_{LARGE}	50.52	29.07	68.62	**78.43**	53.78	16.47	**72.80**	**64.81**	21.91	**59.29**	51.57
IndoLEM IndoBERT$_{BASE}$	48.73	31.48	65.72	74.23	51.80	**24.87**	68.66	64.07	**36.45**	53.32	51.93
IndoNLU IndoBERT$_{BASE}$	47.40	29.14	53.40	69.24	51.59	20.42	64.75	57.11	34.07	52.11	47.92
IndoNLU IndoBERT$_{LARGE}$	6.64	7.62	6.80	73.59	48.13	11.80	66.32	17.37	25.38	53.91	31.76
NusaBERT$_{BASE}$	48.76	**34.61**	60.05	74.74	52.43	24.73	68.02	60.83	31.57	57.65	51.34
NusaBERT$_{LARGE}$	**50.25**	33.38	**72.52**	78.23	**54.47**	18.38	69.18	64.71	32.55	56.89	**53.06**

Table 15: Evaluation results of baseline models and NusaBERT on the NusaParagraph rhetoric mode classification task, measured in macro-F1 (%). Baseline results are obtained from Cahyawijaya et al. (2023b). The best performance on each task is **bolded** for clarity.

NusaParagraph EmoT											
Model	**bew**	**btk**	**bug**	**jav**	**mad**	**mak**	**min**	**mui**	**rej**	**sun**	μ
Logistic Regression (Bag of Words)	82.03	78.33	55.89	84.77	**75.20**	72.90	89.52	71.82	72.43	83.09	**78.23**
Logistic Regression (TF-IDF)	84.52	**83.68**	64.53	**88.04**	69.87	**79.55**	**91.01**	71.84	**80.67**	84.93	
Naive Bayes (Bag of Words)	78.28	71.84	**68.08**	81.37	66.53	71.43	87.07	75.39	75.72	79.42	75.51
Naive Bayes (TF-IDF)	77.97	75.06	62.92	83.15	68.27	75.80	85.98	71.34	75.95	78.57	
SVM (Bag of Words)	80.45	76.61	53.76	82.26	73.26	71.90	87.05	69.06	69.42	81.36	76.36
SVM (TF-IDF)	84.51	82.50	65.27	86.96	70.64	78.74	89.09	71.85	66.66	85.82	
mBERT	80.60	65.35	26.49	78.90	58.84	58.40	82.56	63.66	39.97	76.74	63.15
XLM-R_{BASE}	81.38	64.15	11.17	83.28	53.25	51.98	83.79	61.12	22.38	78.94	59.14
XLM-R_{LARGE}	**86.92**	70.39	30.84	85.50	57.31	60.45	84.40	78.59	32.11	**87.74**	67.43
IndoLEM IndoBERT$_{BASE}$	86.59	66.80	36.81	84.58	54.75	59.39	82.99	63.76	57.31	76.39	66.94
IndoNLU IndoBERT$_{BASE}$	83.04	67.59	31.83	82.01	59.35	62.00	84.08	74.60	49.40	77.27	67.12
IndoNLU IndoBERT$_{LARGE}$	85.49	71.92	27.88	84.52	43.55	66.51	81.75	74.87	13.06	76.89	62.64
NusaBERT$_{BASE}$	84.44	74.19	36.44	84.18	59.16	66.70	85.61	66.37	36.54	78.13	67.18
NusaBERT$_{LARGE}$	86.57	74.06	44.94	85.86	72.31	73.14	86.83	**82.96**	30.19	81.36	71.82

Table 16: Evaluation results of baseline models and NusaBERT on the NusaParagraph emotion classification task, measured in macro-F1 (%). Baseline results are obtained from Cahyawijaya et al. (2023b). The best performance on each task is **bolded** for clarity.

Model	**Original (ind)**	**eng**	**jav**	**msa**	**sun**	μ
EmoT						
IndoBERT$_{BASE}$	72.42	62.87	60.07	62.95	63.03	64.27
IndoBERT$_{LARGE}$	75.53	66.29	63.41	65.30	66.21	67.35
mBERT	61.14	48.64	47.12	48.41	48.64	50.79
XLM-R_{BASE}	72.88	61.90	58.94	59.70	60.38	62.76
XLM-R_{LARGE}	**78.26**	65.99	65.23	65.84	**66.52**	68.37
NusaBERT$_{BASE}$	75.23	61.14	60.45	61.59	61.59	64.00
NusaBERT$_{LARGE}$	78.18	**67.73**	**67.73**	**67.73**	66.14	**69.50**
SmSA						
IndoBERT$_{BASE}$	91.00	89.67	85.93	87.80	88.60	88.60
IndoBERT$_{LARGE}$	**94.20**	91.73	90.07	90.20	**92.00**	91.64
mBERT	83.00	80.80	80.00	80.07	80.53	80.88
XLM-R_{BASE}	91.53	88.13	87.73	87.26	87.26	88.38
XLM-R_{LARGE}	94.07	**91.94**	**90.87**	**91.47**	91.34	**91.94**
NusaBERT$_{BASE}$	91.00	90.40	88.20	88.60	89.20	89.48
NusaBERT$_{LARGE}$	91.00	89.20	87.20	88.80	88.80	89.00

Table 17: Code-mixing robustness evaluation results of baseline models and NusaBERT on IndoRobusta-Blend, measured in accuracy (%). Baseline results are inferred from the delta accuracies reported by Adilazuarda et al. (2022). The best performance on each task is **bolded** for clarity.