NusaCrowd: Open Source Initiative for Indonesian NLP Resources

Samuel Cahyawijaya[♠,1,2], Holy Lovenia[♠,1,2], Alham Fikri Aji[♠,3], Genta Indra Winata[♠,4], Bryan Wilie[♠,1,2], Fajri Koto[♠,3,2], Rahmad Mahendra[5,2], Christian Wibisono[6], Ade Romadhony[7,2], Karissa Vincentio[8,2], Jennifer Santoso[9], David Moeljadi[10], Cahya Wirawan[12], Frederikus Hudi[11,20], Muhammad Satrio Wicaksono[13],, Ivan Halim Parmonangan[14], Ika Alfina[5], Ilham Firdausi Putra[13], Samsul Rahmadani[15], Yulianti Oenang[13], Ali Akbar Septiandri[16], James Jaya[13], Kaustubh D. Dhole[17], Arie Ardiyanti Suryani[7], Rifki Afina Putri[18], Dan Su[1], Keith Stevens[19], Made Nindyatama Nityasya[13], Muhammad Farid Adilazuarda[6], Ryan Ignatius[13], Ryandito Diandaru[6],Vito Ghifari[6], Tiezheng Yu[1], Wenliang Dai[1], Yan Xu[1], Dyah Damapuspita[5], Haryo Akbarianto Wibowo[13], Cuk Tho[14], Ichwanul Muslim Karo Karo[21], Tirana Noor Fatyanosa[22], Ziwei Ji[1], Graham Neubig[23], Timothy Baldwin[3], Sebastian Ruder[24], Pascale Fung[1], Herry Sujaini[25,2], Sakriani Sakti[26,11], Ayu Purwarianti[6,27,2]

[♠]Main Authors

[1]HKUST [2]INACL [3]MBZUAI [4]Bloomberg [5]Universitas Indonesia
[6]Institut Teknologi Bandung [7]Telkom University [8]JULO [9]University of Tsukuba
[10]Kanda University of International Studies [11]NAIST [12]AI-Research.id
[13]Independent Researcher [14]BINUS [15]Bahasa.ai [16]Universitas Al Azhar Indonesia
[17]Emory University [18]KAIST [19]Surface Data [20] Works Applications
[21]State University of Medan [22]Kumamoto University [23]CMU [24]Google
[25]Tanjungpura University [26]JAIST [27]Prosa.ai

Abstract

We present NusaCrowd, a collaborative initiative to collect and unify existing resources for Indonesian languages, including opening access to previously non-public resources. Through this initiative, we have brought together 137 datasets and 118 standardized data loaders. The quality of the datasets has been assessed manually and automatically, and their value is demonstrated through multiple experiments. NusaCrowd's data collection enables the creation of the first zero-shot benchmarks for natural language understanding and generation in Indonesian and the local languages of Indonesia. Furthermore, NusaCrowd brings the creation of the first multilingual automatic speech recognition benchmark in Indonesian and the local languages of Indonesia. Our work strives to advance natural language processing (NLP) research for languages that are underrepresented despite being widely spoken.

1 Introduction

Indonesia is one of the most linguistically diverse and populous countries in the world, with over 270 million people living across 18,000+ islands. It covers more than 700 spoken languages, making up ∼10% of all languages in the world (Grimes, 2000; Lewis, 2009; Cohn and Ravindranath, 2014). However, the progress of NLP research in Indonesian languages has been held back by factors including language diversity (Anderbeck, 2008; Haryono, 2012; Siregar et al., 2014; Fauzi and Puspitorini, 2018), orthographic variation (Soeparno, 2015), resource limitation (Wilie et al., 2020; Koto et al., 2020b), and other societal challenges (Nurjanah, 2018; Jahang and Meirina, 2021; Aji et al., 2022).

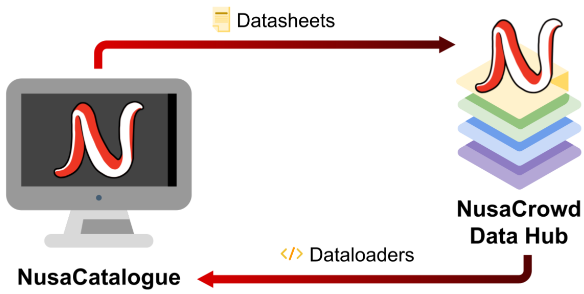

Figure 1: System architecture of NusaCrowd. Open access to the datasheets is provided through **NusaCatalogue**, while dataloader scripts to access the resources are implemented in **NusaCrowd Data Hub**.

Existing NLP research mainly focuses on high-resource languages (Wang et al., 2018; Xu et al., 2020; Ruder, 2022), while the vast majority of languages with limited data—including most languages spoken in Indonesia—are neglected (Joshi et al., 2020). Specifically, many Indonesian NLP resources are scattered, undocumented, and not publicly available. These issues cause a severe data scarcity problem, which hinders NLP research in Indonesian and other local languages spoken in Indonesia from progressing.

In this work, we introduce NusaCrowd,[1] an open collaborative effort to gather and unify existing resources in Indonesian languages for public use, and liberate non-public resources. This initiative has successfully collected a total of 137 datasheets with 118 standardized data loaders in NusaCrowd Data Hub[2]. The datasets were manually assessed for data quality by multiple native speakers and experts in NLP. Utilizing the datasets collected in NusaCrowd, we introduce the first zero-shot NLU benchmark (**NusaNLU**), zero-shot NLG benchmark (**NusaNLG**), and multilingual ASR benchmark (**NusaASR**) for Indonesian languages. We evaluate various Indonesian and multilingual models on the benchmarks.

Our contributions can be summarized as follows:

- We introduce the first large-scale resource hub of standardized Indonesian corpora, covering 100+ datasets and 200+ tasks, spanning 19 Indonesian languages in text, speech, and image modalities. As part of this, we provide first-time access to 14 previously private datasets.
- We develop the first Indonesian multilingual zero-shot benchmarks for natural language understanding (NusaNLU) and natural language generation (NusaNLG), which cover 40 NLU and NLG tasks in 12 languages.
- We conduct a comprehensive analysis of the collected datasets across various factors. Our analysis reflects the quality and diversity of existing NLP datasets in Indonesian and other languages spoken in the region.
- For speech, our initiative opens up access to a wide variety of ASR corpora ($\sim$800 hours) covering 10 Indonesian languages. Using these resources, we build NusaASR and develop various Indonesian monolingual and multilingual ASR models.

2 Related Work

Indonesian NLP Resources The lack of labeled datasets for training and evaluation has impeded the advancement of NLP research in Indonesian languages (Aji et al., 2022). As a result, research has focused on using unlabeled data by building large language models (LLMs) to enable zero-shot and few-shot transfer learning. In recent years, multiple efforts have worked on language models (LMs) in Indonesian languages by exploring and developing different LM structures. Several efforts have focused on encoder-only LMs, such as IndoBERT (Wilie et al., 2020; Koto et al., 2020b), SundaBERT (Wongso et al., 2022), and IndoBERT-Tweet (Koto et al., 2021). Elsewhere, a number of generative models have been proposed, i.e., IndoBART and IndoGPT, along with the generation task benchmark, IndoNLG (Cahyawijaya et al., 2021b).

Open and Community-based Initiatives Open source/open science initiatives are a core part of the motivation behind this paper. Large-scale collaborations have made their mark in various research areas through developing a variety of resources, e.g., LMs (Scao et al., 2022; Muennighoff et al., 2022), datasets (Ardila et al., 2020; Adelani et al., 2021; Mager et al., 2021), catalogues (Alyafeai et al., 2022; Altaher et al., 2022; McMillan-Major et al., 2022), and benchmarks (Srivastava et al., 2022; Dhole et al., 2021; Fries et al., 2022).

3 NusaCrowd

In this section, we provide an overview of NusaCrowd, a detailed description of the NusaCrowd framework, the dataset curation process, as well as a detailed summary and statistics of the datasets contained in NusaCrowd.

3.1 Overview of NusaCrowd

NusaCrowd is a crowdsourcing initiative to collect, open-source, and standardize access to datasets in Indonesian and 700+ local languages in Indonesia. NusaCrowd aims to address the resource limitation problem in Indonesian NLP across three dimensions: (1) complete datasheets for each curated, ready-to-use dataset; (2) an open-access and centralized data hub for accessing datasets through standardized data loading scripts; and (3) promoting public data access for published non-public datasets. Through promoting public data access,

[1] NusaCrowd is a portmanteau of the words **Nusantara** and **Crowd**. The word **Nusantara** is derived from an old Javanese term referring to the territories of the Majapahit empire that corresponds to present-day Indonesia.

[2] We publicly release NusaCrowd's data hub at `https://github.com/IndoNLP/nusa-crowd` and the NusaCatalogue at `https://indonlp.github.io/nusa-catalogue/`

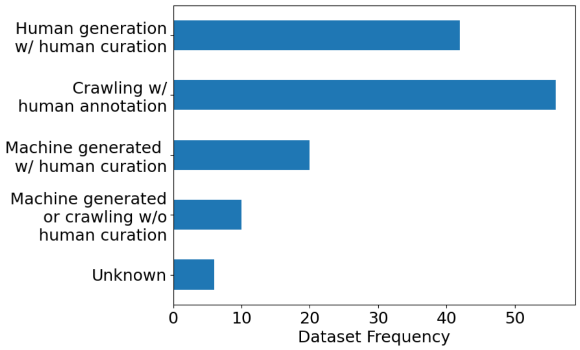

Figure 2: Distribution of dataset curation approaches used in datasets contained in NusaCrowd.

NusaCrowd provides access to 14 previously non-public datasets, some of which are multilingual, covering a total of ∼40 tasks over 12 languages. It also serves as a portal for retrieving and loading a wide variety of Indonesian NLP datasets, in text and other modalities (e.g., speech and images). NusaCrowd does not store or copy any of the hosted datasets, and control and ownership of the hosted datasets belong to the original owners.

3.2 NusaCrowd Framework

As shown in Figure 1, NusaCrowd consists of two platforms: NusaCatalogue and NusaCrowd Data Hub. The two platforms interact to support dataset registration and provide a standardized pipeline for NusaCrowd. In general, NusaCatalogue stores the datasheets (metadata) of all datasets, and NusaCrowd Data Hub stores the standardized data loaders for all of the datasets. The two systems share information about the datasheets and the data loaders, enabling users to seamlessly explore and use the datasets.[3]

NusaCrowd Workflow The dataset registration and standardization pipeline in NusaCrowd consists of four stages: (1) submission of datasheet information through an online form; (2) manual curation of the datasheet information by an expert in NLP, which, once approved (Section 3.3), is made available via the **NusaCatalogue** portal and a data loader implementation request is submitted to **NusaCrowd Data Hub**; (3) implementation of a data loader; and (4) review and approval of the implemented data loader by two maintainers, which is then published on **NusaCrowd Data Hub**. In addition to the datasheets, we also provide instructions on how to use the data in **NusaCatalogue**.

[3]All code in NusaCrowd will be made publicly available under Apache License 2.0.

Language	langid.py		FastText		CLD3
	Top-1	Top-3	Top-1	Top-3	Top-1
Eng	98.33	99.33	94.05	99.03	99.69
Ind	72.11	90.39	82.42	89.92	60.27
Sun	—	—	34.28	75.21	50.53
Jav	48.97	79.07	28.08	69.43	46.88

Table 1: Language identification accuracy based on different languages. For Sundanese and Javanese, several datasets consist of informal Indonesian utterances including Ind–Sun and Ind–Jav code-mixed sentences.

3.3 Dataset Standardization and Curation

We standardize the tasks from the datasets in NusaCrowd into several categories according to a specific schema, defined as a common set of attributes required to perform the task. We use the schema to cover similar tasks across the datasets. We define 13 schemas to cover all the tasks and modalities in the datasets, e.g., text classification, text generation, image captioning, and speech recognition. For instance, in the single-label text classification schema (`TEXT`), each example consists of three attributes `(id, text, label)`, where `id` denotes a unique row identifier, `text` denotes the input text, and `label` denotes a discriminative target variable. We elaborate on the attributes of each schema in Appendix B.

To assess the quality of the datasets in NusaCrowd, we perform manual curation for each datasheet submission based on two criteria: language correctness, and the annotation process. We provide the results as metadata for each dataset. We check the correctness of the reported language using off-the-shelf language identification (LID) tools. We perform LID in 4 languages: English, Indonesian, Sundanese, and Javanese, We measure the LID accuracy compared to the reported languages in the metadata on all tasks containing text modality in NusaCrowd. Since many datasets consist of a large number of samples, language correctness checking is done both automatically and manually.

We conduct automatic language identification for 4 languages, i.e., English, Indonesian, Sundanese, and Javanese [5] using 3 off-the-shelf language identification tools, i.e., langid.py (Lui and

[5]We only perform language identification as these are the only languages supported by most of the existing off-the-shelf language identification tools.

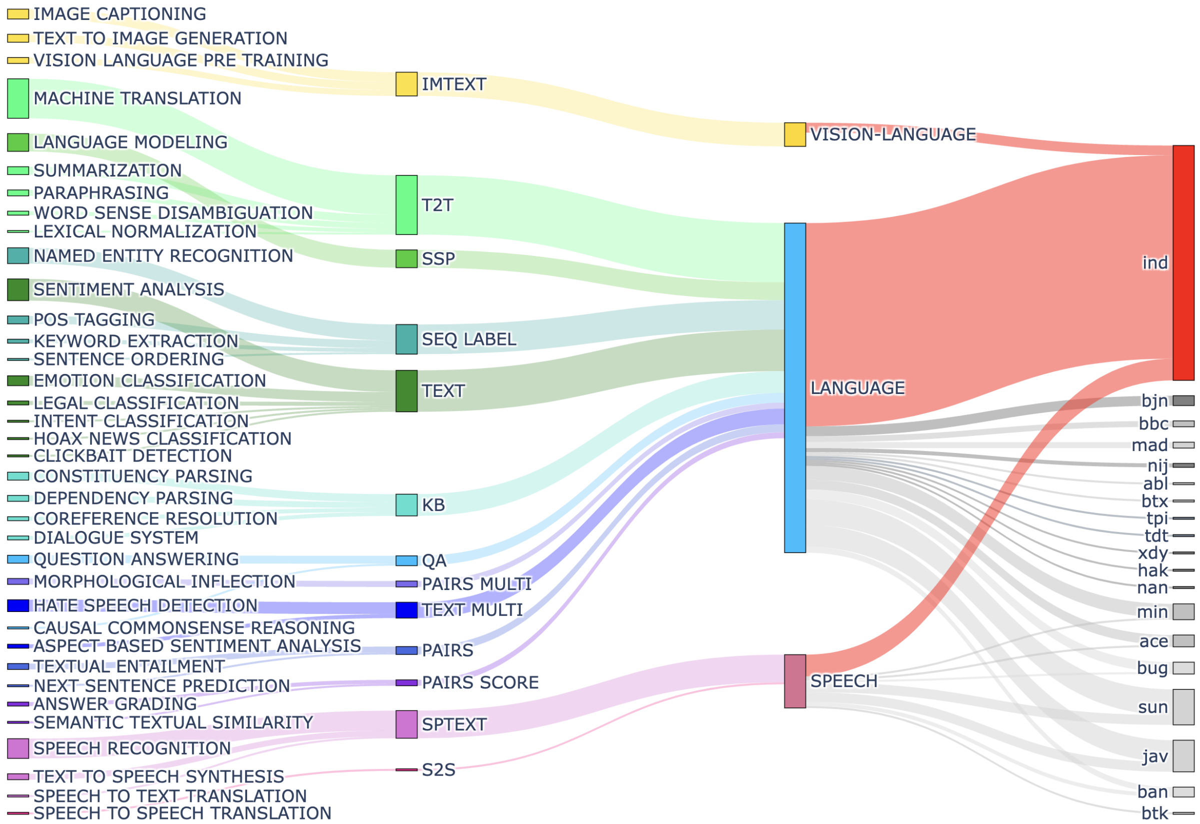

Figure 3: Summary of tasks, schemas, modalities, and languages[4] in NusaCrowd. ∼75% of the datasets are textual language data in Indonesian, with the other two modalities being vision-language and speech. Textual language data covers 19 Indonesian languages (Indonesian and 18 other languages spoken in the region), the speech data covers 8 languages (Indonesian and 7 local languages), while vision-language data only covers Indonesian.

Baldwin, 2012), FastText LID (Ooms, 2022), and Google CLD3 (Ooms, 2022). For other languages, since there is no language identification library available, the curation is done manually through sampling. Based on the automatic language identification result in Table 1, the correctness of languages is quite high, indicated by the top-3 accuracy of each language identification tools [6]. Additionally, the accuracy of Indonesian is not as high as English, we conjecture that this is caused by there are many English terms from tasks that are collected from online platforms.

For assessing the annotation process for each dataset, we manually check the dataset annotation process from relevant publications and/or other descriptions and classify them into five categories, i.e., *human-generated*, *crawling with human annotation*, *machine-generated with human curation*, *machine-generated or crawling without human curation*, and *unknown*. The statistics of the dataset annotation assessment are shown in Appendix K.

[6]we don't consider Top-1 for Sundanese and Javanese since the languages are low-resource and often mispredicted

In general, ∼90% of all the datasets listed in NusaCrowd are human-curated, showing that most of the datasets in NusaCrowd are high-quality and well-suited for building and evaluating Indonesian NLP models. Moreover, almost half of the datasets are collected through crawling and are annotated manually by humans, usually for NLP tasks such as sentiment analysis, emotion recognition, hate speech detection, named entity recognition, and machine translation. The crawling often comes from sources such as social media, news platforms, online reviews, etc.

3.4 Datasets in NusaCrowd

NusaCrowd includes 137 datasheets and 118 dataloaders, including access to 14 previously non-public datasets, and a variety of tasks and languages. We list all of the previously private datasets in Appendix I. NusaCrowd covers 36 task types, including: machine translation, summarization, sentiment analysis, part-of-speech (POS) tagging, and question answering, which are standardized into 13 different schemas. The datasets in

NusaCrowd stem from three modalities—image, text, and speech—with the majority of the data coming from the text modality. In terms of languages, NusaCrowd covers 19 Indonesian languages, i.e., Indonesian and 18 regional languages, in addition to some non-Indonesian languages such as Japanese, English, Spanish, and Russian, which come into the mix as machine translation language pairs. A summary of the datasets is shown in Figure 3. A list of language codes with the complete language name and family is provided in Appendix A. We present comprehensive details of the datasets in Appendix K, and a comparison of NusaCrowd with other initiatives in Appendix J.

Modalities NusaCrowd comprises datasets from three different modalities, i.e., image, text, and speech, all of which are related to language tasks. Most datasets contain text data for natural language understanding (e.g., sentiment analysis, named entity recognition, and parsing) and natural language generation tasks (e.g., machine translation, paraphrasing, and abstractive summarization). These account for 29 out of 36 task types in NusaCrowd. In addition, NusaCrowd covers three vision tasks: vision-language pre-training, image captioning, and text-to-image generation. For speech, NusaCrowd covers four tasks: automatic speech recognition (ASR), text-to-speech synthesis (TTS), speech-to-text translation (S2T), and speech-to-speech translation (S2S).

Languages NusaCrowd covers Indonesian and 18 regional languages. Most languages covered in NusaCrowd belong to the Austronesian language family, 14 of which are part of Malayo-Polynesian family (including Indonesian), 2 of which are creole languages, i.e., Tok Pisin (tpi) and Tetun Dili (tdt).[7] The other two languages — Hakka/Khek (hak) and Min Nan (nan) with Teochew dialect — are Sinitic and belong to the Sino-Tibetan language family. Detailed descriptions of each language are provided in Appendix A.

4 NusaCrowd Benchmarks

To showcase the benefit of NusaCrowd, we develop three different benchmarks from subsets of the datasets. Specifically, we develop benchmarks for Indonesian and other local languages including a zero-shot NLU benchmark (NusaNLU), a zero-shot NLG benchmark (NusaNLG), and a multilingual ASR benchmark (NusaASR).

4.1 NusaNLU

Existing benchmarks (Wilie et al., 2020; Koto et al., 2020b) in Indonesian NLU only cover one language, i.e., the national language, Indonesian. Moreover, these benchmarks only focus on comparing traditional machine learning approaches with the fine-tuning approaches of pre-trained LMs. Following recent work in other high-resource languages that explore zero-shot generalization of large LMs (Scao et al., 2022; Lin et al., 2022; Muennighoff et al., 2022; Fries et al., 2022), we develop NusaNLU, the first zero-shot NLU benchmark in Indonesian and regional languages to benchmark zero-shot techniques over 26 datasets using both Indonesian monolingual and multilingual LMs. NusaNLU covers 12 languages across various tasks, including 3 emotion classification tasks (Saputri et al., 2018; Yulianti et al., 2021; Riccosan et al., 2022), 18 sentiment analysis tasks (Winata et al., 2023; Nurlaila et al., 2017; Hidayatullah et al., 2020; Wongso et al., 2021; Koto et al., 2020b; Purwarianti and Crisdayanti, 2019), one review score rating task[8], one hate speech detection task (Ibrohim and Budi, 2019), one abusive language detection task (Putri et al., 2021), one next tweet prediction task (Koto et al., 2020b), and one natural language inference (NLI) task (Mahendra et al., 2021). A visual overview of the datasets in NusaNLU is provided in Figure 4.

Models We evaluate three state-of-the-art multilingual language models: XLM-R (Conneau et al., 2020), XGLM (Lin et al., 2022), and BLOOMZ (Muennighoff et al., 2022). We generally evaluate in a zero-shot cross-lingual transfer setting (Hu et al., 2020). For XLM-R, we employ intermediate-task training on NLI by predicting the entailment relation between the input text and the label (Phang et al., 2020). We explore both XLM-R fine-tuned on XNLI (Conneau et al., 2018) and Indonesian IndoNLI (Mahendra et al., 2021). For XGLM and BLOOMZ, we employ zero-shot prompt-based learning with prompts in English and Indonesian. For each language and task, we employ three different prompts and take the average score

[6]Based on ISO639-3 language codes: `https://iso639-3.sil.org/code_tables/639/data`.

[7]The two languages are not spoken in Indonesia, but instead used in neighboring countries: Papua New Guinea and Timor Leste.

[8]`https://huggingface.co/datasets/jakartaresearch/google-play-review`

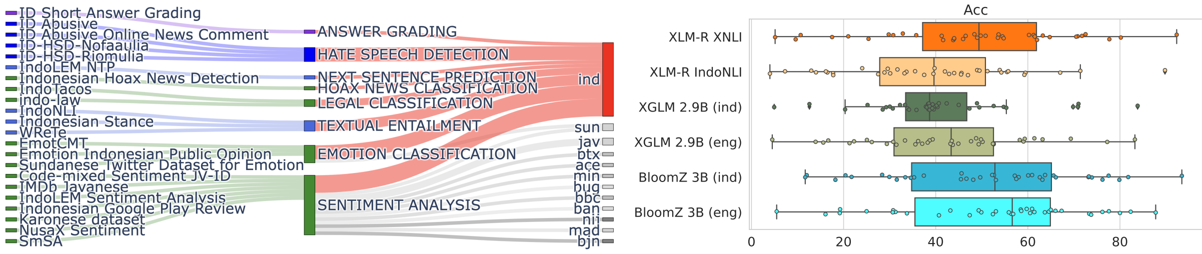

Figure 4: **(left)** The datasets used in NusaNLU and **(right)** zero-shot generalization in NusaNLU. Box plots show summary statistics of accuracy scores. For XGLM and BLOOMZ, each point denotes average per-dataset performance using three different prompts. **(ind)** and **(eng)** denote the prompt language used for prompting, i.e., Indonesian and English, respectively.

for the evaluation of each task. More details about fine-tuning hyperparameters and the prompt used in the NLU experiments are shown in Appendix C.

Results Figure 4 shows the zero-shot NLU results of all the models. Overall, the prompting performance of BLOOMZ outperforms other models. Prompting with BLOOMZ outperforms XGLM by a huge margin, providing evidence of the benefit of instruction tuning for prompting. Interestingly, zero-shot cross-task transfer using XLM-R trained on XNLI (XLM-R XNLI) outperforms prompting using XGLM and performs on par with prompting using BLOOMZ, despite the huge difference in their model sizes. This suggests that large LMs are not always needed to perform zero-shot NLU tasks and better efficiency can be achieved through cross-task transfer using much smaller models while achieving similar performance.

Comparing the performance of cross-task fine-tuning across monolingual and multilingual NLI, XLM-R XNLI (122k training instances) outperforms XLM-R IndoNLI (11k training instances) by a large margin, suggesting that using large-scale multilingual data is more effective than using smaller-scale data from closely-related or even the same language to fine-tune a multilingual model in a zero-shot cross-task setting. Comparing the language of the prompts, both BLOOMZ and XGLM with English prompts perform better than the corresponding models with Indonesian prompts. Our findings align with prior work (Muennighoff et al., 2022; Lin et al., 2022; Shi et al., 2022), which shows that, in most cases, the corresponding models perform better in English than on human-translated prompts, despite the language distance between the prompt template and the corresponding text data.

Comparing the performance across different languages, as shown in Figure 5, we can conclude that the performance of all models is generally better for Indonesian and English compared to regional Indonesian languages, suggesting that existing multilingual models are unable to generalize well on these languages, and better language representations are vital to close the gap. A full breakdown of per-task performance is provided in Appendix F.

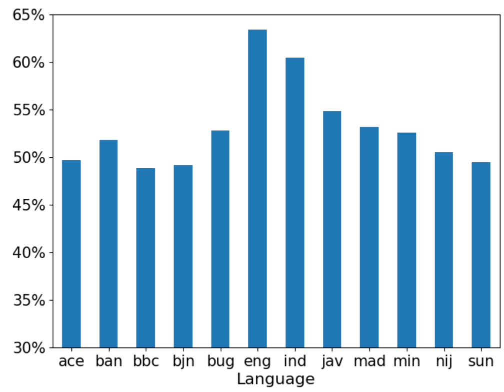

Figure 5: Average zero-shot performance per language across all models on the NusaX subset. All models achieve higher scores for Indonesian (ind) and English (eng).

4.2 NusaNLG

Recent work on Indonesian NLG benchmarks (Cahyawijaya et al., 2021b; Guntara et al., 2020) has employed transformer-based models, both decoder-only (e.g., IndoGPT) and encoder–decoder (e.g., IndoBART) architectures. To further broaden NLG research in Indonesian and other regional languages, we develop an NLG benchmark, NusaNLG, which covers NLG tasks in 12 languages including English, Indonesian, and 10 local languages. NusaNLG incorporates a total of 36 datasets across various tasks covering 33 machine translation tasks (Guntara et al., 2020;

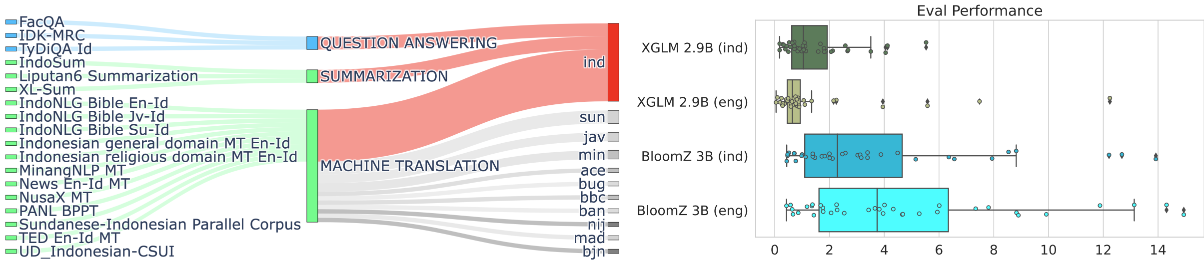

Figure 6: **(left)** The datasets used in NusaNLG and **(right)** Zero-shot generalization to machine translation and summarization tasks in NusaNLG. Box plots show summary statistics of the evaluation performance. Points are per-dataset scores from the average of performances over 3 different prompts. **(ind)** and **(eng)** denote the prompt language used for prompting, i.e., Indonesian and English, respectively.

Cahyawijaya et al., 2021b) and 3 summarization tasks (Kurniawan and Louvan, 2018; Koto et al., 2020a) (Figure 6). We use SacreBLEU for machine translation evaluation, and ROUGE-L for summarization evaluation.

Models Following recent work in prompting, we explore the possibility of zero-shot generalization of various large LMs on generation tasks through prompting on two NLG tasks, i.e., machine translation and summarization. To explore the effect of different prompt languages on the zero-shot generalization performance, we evaluate prompts in English and Indonesian. We employ two large LMs: XGLM (Lin et al., 2022), and BLOOMZ (Muennighoff et al., 2022). For each task and prompt language, we provide three different prompts and average the result. More details about the hyperparameters and the prompt used in the NLG experiments are shown in Appendix D.

Results The zero-shot NLG results of all models are shown in Figure 6. Outputs obtained by prompting BLOOMZ outperform those obtained from XGLM for both English and Indonesian prompts. The performance is better on average when prompting BLOOMZ with English prompts than when using Indonesian prompts, which aligns with the results of BLOOMZ on XNLI (Conneau et al., 2018), where BLOOMZ with English prompts performs better than the human-translated prompts (Muennighoff et al., 2022).

Prompting using XGLM yields better quality outputs using Indonesian language prompts than using English prompts. A similar result is reported in XGLM evaluation for Spanish XNLI and Chinese XCOPA (Ponti et al., 2020), which shows that prompting with the human-translated prompt to the target language produces a better score than the English one. For the BLOOMZ models, the result for English is better since we use the BLOOMZ checkpoint fine-tuned only on English prompts. Additionally, we found that the zero-shot translation quality across all models and prompt languages is poor, especially for local languages, as shown in Table 2. This is even more severe when local languages are involved, yielding $\sim$2% SacreBLEU. This finding suggests that existing large multilingual LMs still fail to learn representations for these local languages. A full breakdown of per-task results over NusaNLG is provided in Appendix G.

Language	Ind prompt	Eng prompt
eng $\rightarrow$ ind	5.11	6.04
ind $\rightarrow$ eng	4.65	7.90
local $\rightarrow$ ind	2.11	2.72
ind $\rightarrow$ local	1.66	2.96

Table 2: Average SacreBLEU performance of BLOOMZ for different language pairs. **Local** denotes all Indonesian local languages in NusaCrowd.

4.3 NusaASR

In addition to zero-shot benchmarks for textual language data, we showcase the benefit of NusaCrowd by extending the NLP benchmark in Indonesian languages to speech. We develop the first multilingual ASR benchmark for Indonesian and other local languages covering 17 ASR datasets in eight languages: 5$\times$Indonesian (ind), 3$\times$Sundanese (sun), 3$\times$Javanese (jav), 2$\times$Balinese (ban), 1$\times$Acehnese (ace), 1$\times$Batak (btk), 1$\times$Buginese (bug), and 1$\times$Minangkabau (min).

Models We employ pre-trained wav2vec 2.0 (Baevski et al., 2020) models in our experiment. We explore three training settings: single-task

Model	ace	ban	btk	bug	ind	jav	min	sun
Single-task Training								
wav2vec 2.0-pt	100.00	71.99	64.77	100.00	12.51	85.78	100.00	83.01
wav2vec 2.0-ft	49.31	28.74	40.92	90.09	2.13	32.11	24.29	26.62
Monolingual Multi-task Training								
wav2vec 2.0-pt (ind)	95.14	>100	>100	96.70	4.20	>100	46.19	>100
wav2vec 2.0-pt (jav)	>100	67.02	81.24	>100	88.87	46.97	68.10	69.89
wav2vec 2.0-pt (sun)	92.36	82.37	74.67	>100	91.22	93.43	98.57	40.42
wav2vec 2.0-ft (ind)	91.67	>100	>100	>100	**1.87**	≥100	70.48	>100
wav2vec 2.0-ft (jav)	90.28	52.63	59.79	>100	78.87	27.23	52.86	54.31
wav2vec 2.0-ft (sun)	89.58	76.52	61.34	>100	89.59	88.50	79.05	25.11
Multilingual Multi-task Training								
wav2vec 2.0-pt	40.85	**16.73**	**18.98**	**41.59**	8.05	**18.57**	**16.94**	**13.93**
wav2vec 2.0-ft	**31.94**	21.05	35.99	53.30	1.90	27.55	18.10	20.79

Table 3: Speech recognition results in terms of average word error rate (WER) per language over NusaASR (lower is better). For monolingual multi-task training, the language in brackets denotes the language used for training. **Bold** denotes the best performance across all groups. Underline denotes the best performance within the group. In monolingual multi-task training, Highlight denotes that the model is trained in the corresponding language.

monolingual training, where we fine-tune and evaluate the model on the corresponding ASR dataset; multi-task monolingual training, where we fine-tune the model using multiple ASR datasets on a single language (we evaluate three languages with the largest resources, i.e., Indonesian, Javanese, and Sundanese); and joint multi-task multilingual training, where we fine-tune the model using all 17 ASR datasets listed on NusaASR. We experiment with two wav2vec 2.0$_{LARGE}$ (∼300M parameters) checkpoints, i.e., an unsupervised pre-trained XLS-R wav2vec 2.0 (**wav2vec 2.0-pt**)[9] and an Indonesian, Javanese, and Sundanese ASR fine-tuned XLS-R wav2vec 2.0 (**wav2vec 2.0-ft**).[10] In addition to wav2vec 2.0, we also employ Whisper$_{SMALL}$ (Radford et al., 2022)[11] (∼250M parameters). Full details of the experiment setup are provided in Appendix E.

Results Table 3 shows the per-language task-averaged performances of wav2vec 2.0 models over NusaASR. The complete per-task results of NusaASR along with the performance of Whisper$_{SMALL}$ are provided in Appendix H. Based on the results, single-task training on **wav2vec 2.0-pt** performs poorly due to the limited training data to adapt from unsupervised contrastive pre-training to the ASR task, while the ASR fine-tuned **wav2vec 2.0-ft** model yields decent results in most languages, except for Buginese (bug) with 90.09% WER. This suggests limited transferability from Indonesian, Sundanese, and Javanese to Buginese, consistent with the analysis from NusaX (Winata et al., 2023) regarding the low overlap between Buginese and other local languages included in NusaCrowd. For monolingual multi-task training, all models perform well only in the languages that they were trained on. This shows that there is a large difference between vocabulary and speech features from one language to another.

For all models evaluated over NusaASR (**wav2vec 2.0-pt**, **wav2vec 2.0-ft**, and **Whisper**), the best performance is achieved through multilingual multi-task training, yielding as low as ∼20% average WER across all languages, suggesting transferability of speech features from one language to the others (Fung et al., 1998; PLU et al., 2000; Sakti et al., 2012; Nakayama et al., 2019). Unlike prior work (Winata et al., 2023), where Acehnese (ace) yields similar performance to other languages in sentiment analysis, the same behavior

[9] https://huggingface.co/facebook/wav2vec2-large-xlsr-53

[10] https://huggingface.co/indonesian-nlp/wav2vec2-indonesian-javanese-sundanese

[11] https://huggingface.co/openai/whisper-small

is not reflected in ASR. This suggests that there is a distinction between the speech of Acehnese (ace) and other regional languages, despite vocabulary overlap and shared language structure.

5 Discussion

Multilinguality for Low-Resource Languages Despite the higher pre-training cost relative to monolingual LMs (Cahyawijaya et al., 2021b), multilingual LMs are more versatile and transferable. Recent low-resource monolingual language LMs are on the scale of a hundred million parameters, while the size of multilingual LMs, within a period of three years, has increased by around 1,000$\times$ from $\sim$100M to $\geq$100B parameters (Devlin et al., 2019; Xue et al., 2021; Tang et al., 2021; Muennighoff et al., 2022; Scao et al., 2022). This benefit comes from the data scale of multilingual LMs, which is orders of magnitude larger than monolingual LMs. Additionally, multilingual LMs benefit from positive transfer between related languages, which is especially beneficial for low-resource languages. Moving forward, we expect that multilingual LMs will play a significant role in the exploration of low-resource languages.

Viability of Large Models for Indonesian Computational resources are limited among Indonesian research institutions and in industry, even among the top Indonesian universities (Indonesia, 2020; Nityasya et al., 2020). Focusing solely on large LMs will limit accessibility, and adoption will likely be low. Therefore, although larger LMs empirically offer better quality, we instead suggest investing more effort in efficiency. This includes smaller sizes LMs and modularized LMs (Pfeiffer et al., 2020; Ansell et al., 2021; Pfeiffer et al., 2022). Furthermore, more work on efficiency through factorization (Winata et al., 2020; Cahyawijaya et al., 2021a), pruning (Frankle and Carbin, 2019; Dai et al., 2021), quantization (Shen et al., 2020; Aji and Heafield, 2020), or distillation (Zhang et al., 2020; Bai et al., 2021; Dai et al., 2022) are also likely to be beneficial.

6 Conclusion

We have introduced NusaCrowd, a combined resource for Indonesian and regional languages, covering 137 datasets, 118 of which have a standardized loader. NusaCrowd covers Indonesian and 18 regional languages, encompassing 3 different data modalities. Manual and automatic curation processes were conducted to verify the quality of the collected datasets. The effectiveness of NusaCrowd is shown in three use cases: zero-shot NLU (NusaNLU), zero-shot NLG (NusaNLG), and multilingual ASR (NusaASR) benchmarks. Our experiments provide evidence regarding the efficiency of cross-tasks method over prompting for zero-shot NLU, the limited capabilities of existing large LMs for handling NLG tasks in local languages, and the potential of joint multilingual multi-task learning for Indonesian ASR. We hope NusaCrowd will benefit the research community as a data hub for Indonesian and regional languages by facilitating easy access to datasets, as well as accelerating research and development.

7 Limitations

Dataset Utilization We have collected 137 datasets, yet we have only conducted experiments over a minority of these ($\sim$40 datasets), leaving the remaining datasets unexplored. Since the datasets are already curated, future work should further explore these datasets in additional experiments. In this work, we do not experiment on image-text datasets for two reasons: (1) all of the image-text datasets are translated from English versions; and (2) there is no large LM available for zero-shot image-to-text generation.

Experiments We did not attempt few-shot or fully-supervised learning experiments in NusaCrowd since prior work has explored these approaches on some of the datasets (Wilie et al., 2020; Koto et al., 2020b; Cahyawijaya et al., 2021b; Winata et al., 2023). We specifically conduct our experiments on zero-shot methods to explore the generalization of zero-shot cross-lingual and zero-shot prompting approaches to extremely low-resource languages.

Task Diversity The tasks represented in NusaCrowd are skewed towards MT, sentiment, abusive text classification, and ASR. Many other tasks remain unexplored for Indonesian and regional languages. Furthermore, most ASR work come from the same authors or research groups. While these topics are prevalent among Indonesian researchers, it is also important to expand to other tasks.

Domain Diversity The datasets in NusaCrowd are primarily from the domains of social media, news, and other general domain sources. Despite having a huge potential, narrow-domain datasets, such as clinical, biomedical, legal, financial, and educational datasets remain underrepresented for Indonesian and regional languages. Exploration of domain-specific data and use cases for Indonesian and regional languages is critical.

Language Diversity There are 700+ languages in Indonesia. However, we have only focused on a small fraction of these languages. In addition, there are also other regional languages similar to the two Sinitic languages in NusaCrowd, i.e., Hakka (Khek) and Min Nan (Teochew). More focus on under-represented languages is an interesting future direction.

Multimodality The datasets in NusaCrowd are mainly in the text modality. Exploration of speech, image, and other modalities for Indonesian and regional languages is still limited, and there are potentially exciting opportunities to capture locally-relevant Indonesian culture in such modalities.

Utilization of Datasets There are 137 datasets contained in NusaCrowd. While we showcased three different use cases for the datasets (i.e., zero-shot NLU, zero-shot NLG, and multilingual ASR benchmarks), there is much greater potential to use the datasets in NusaCrowd. Potential areas of focus include experimenting with various approaches and analyses over multiple datasets, such as multi-task learning, continual learning, or few-shot learning.

8 Ethical Statement

Our work highlights the importance of democratizing access to Natural Language Processing (NLP) technology for underrepresented and extremely low-resource languages with a focus on the Austronesian language family specifically in Indonesian languages. Within our study, we are well aware of the ethical responsibility associated with language research and the potential impact that comes with it. Our study prioritizes diversity, inclusivity, and fairness. Within this work, the contribution of each collaborator is calculated following a fair and transparent scoring guideline that empowers the core principles of NusaCrowd. We have obtained informed consent from all dataset authors to provide publicly open-access corpora and benchmarks. 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. We encourage further collaboration and engagement with underrepresented language communities to ensure that their voices are heard and their needs are addressed in future NLP development. We remain committed to the principles of ethical research, diversity, inclusivity, and fairness, striving to promote social good through our work in the field of language technology.

References

Z Abidin and I Ahmad. 2021. Effect of mono corpus quantity on statistical machine translation Indonesian–Lampung dialect of nyo. In *Journal of Physics: Conference Series*, volume 1751. IOP Publishing.

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. 2021. MasakhaNER: Named entity recognition for African languages. *Transactions of the Association for Computational Linguistics*, 9:1116–1131.

Alham Fikri Aji, Radityo Eko Prasojo Tirana Noor Fatyanosa, Philip Arthur, Suci Fitriany, Salma Qonitah, Nadhifa Zulfa, Tomi Santoso, and Mahendra Data. 2021. Paracotta: Synthetic multilingual paraphrase corpora from the most diverse translation sample pair. In *Proceedings of the 35th Pacific Asia Conference on Language, Information and Computation*, pages 666–675.

Alham Fikri Aji and Kenneth Heafield. 2020. Compressing neural machine translation models with 4-bit precision. In *Proceedings of the Fourth Workshop on Neural Generation and Translation*, pages 35–42, 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.

Ika Alfina, Indra Budi, and Heru Suhartanto. 2020. Tree rotations for dependency trees: Converting the head-directionality of noun phrases. *Journal of Computer Science*, 16(11):1585–1597.

Ika Alfina, Arawinda Dinakaramani, Mohamad Ivan Fanany, and Heru Suhartanto. 2019. A gold standard dependency treebank for Indonesian. In *Proceedings of the 33rd Pacific Asia Conference on Language, Information and Computation*, pages 1–9. Waseda Institute for the Study of Language and Information.

Ika Alfina, Rio Mulia, Mohamad Ivan Fanany, and Yudo Ekanata. 2017a. Hate speech detection in the Indonesian language: A dataset and preliminary study. In *2017 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, pages 233–238.

Ika Alfina, Septiviana Savitri, and Mohamad Ivan Fanany. 2017b. Modified DBpedia entities expansion for tagging automatically NER dataset. In *2017 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, pages 216–221. IEEE.

Yousef Altaher, Ali Fadel, Mazen Alotaibi, Mazen Alyazidi, Mishari Al-Mutairi, Mutlaq Aldhbuiub, Abdulrahman Mosaibah, Abdelrahman Rezk, Abdulrazzaq Alhendi, Mazen Abo Shal, Emad A. Alghamdi, Maged S. Alshaibani, Jezia Zakraoui, Wafaa Mohammed, Kamel Gaanoun, Khalid N. Elmadani, Mustafa Ghaleb, Nouamane Tazi, Raed Alharbi, Maraim Masoud, and Zaid Alyafeai. 2022. Masader Plus: A new interface for exploring+ 500 Arabic NLP datasets. *arXiv preprint arXiv:2208.00932*.

Zaid Alyafeai, Maraim Masoud, Mustafa Ghaleb, and Maged S. Al-shaibani. 2022. Masader: Metadata sourcing for Arabic text and speech data resources. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6340–6351, Marseille, France. European Language Resources Association.

Antonios Anastasopoulos, Alessandro Cattelan, Zi-Yi Dou, Marcello Federico, Christian Federman, Dmitriy Genzel, Francisco Guzmán, Junjie Hu, Macduff Hughes, Philipp Koehn, Rosie Lazar, Will Lewis, Graham Neubig, Mengmeng Niu, Alp Öktem, Eric Paquin, Grace Tang, and Sylwia Tur. 2020. TICO-19: the Translation Initiative for COVID-19. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*.

Karl Ronald Anderbeck. 2008. *Malay dialects of the Batanghari river basin (Jambi, Sumatra)*. SIL International.

Alan Ansell, Edoardo Maria Ponti, Jonas Pfeiffer, Sebastian Ruder, Goran Glavaš, Ivan Vulić, and Anna Korhonen. 2021. MAD-G: Multilingual adapter generation for efficient cross-lingual transfer. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4762–4781, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4218–4222.

Arie Ardiyanti Suryani, Dwi Hendratmo Widyantoro, Ayu Purwarianti, and Yayat Sudaryat. 2022a. Postagged sundanese monolingual corpus. Telkom University Dataverse.

Arie Ardiyanti Suryani, Dwi Hendratmo Widyantoro, Ayu Purwarianti, and Yayat Sudaryat. 2022b. Sundanese-indonesian parallel corpus.

I Wayan Arka. 2003. *Balinese morphosyntax: a lexical-functional approach*. Pacific Linguistics.

Valentina Kania Prameswara Artari, Rahmad Mahendra, Meganingrum Arista Jiwanggi, Adityo Anggraito, and Indra Budi. 2021. A multi-pass sieve coreference resolution for Indonesian. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 79–85, Held Online. INCOMA Ltd.

Jessica Naraiswari Arwidarasti, Ika Alfina, and Adila Alfa Krisnadhi. 2019. Converting an Indonesian constituency treebank to the Penn Treebank format. In *23rd International Conference on Asian Language Processing, IALP 2019*, pages 331–336. Institute of Electrical and Electronics Engineers Inc.

Nofa Aulia and Indra Budi. 2019. Hate speech detection on Indonesian long text documents using machine learning approach. In *Proceedings of the 2019 5th International Conference on Computing and Artificial Intelligence*, ICCAI '19, page 164–169, New York, NY, USA. Association for Computing Machinery.

A. N. Azhar, M. L. Khodra, and A. P. Sutiono. 2019. Multi-label aspect categorization with convolutional neural networks and extreme gradient boosting. In *2019 International Conference on Electrical Engineering and Informatics (ICEEI)*, pages 35–40.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, 33:12449–12460.

Haoli Bai, Wei Zhang, Lu Hou, Lifeng Shang, Jin Jin, Xin Jiang, Qun Liu, Michael Lyu, and Irwin King. 2021. BinaryBERT: Pushing the limit of BERT quantization. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4334–4348, Online. Association for Computational Linguistics.

Anab Maulana Barik, Rahmad Mahendra, and Mirna Adriani. 2019. Normalization of Indonesian-English code-mixed Twitter data. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 417–424, Hong Kong, China. Association for Computational Linguistics.

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

Samuel Cahyawijaya, Genta Indra Winata, Holy Lovenia, Bryan Wilie, Wenliang Dai, Etsuko Ishii, and Pascale Fung. 2021a. Greenformer: Factorization toolkit for efficient deep neural networks. arXiv 2109.06762.

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. 2021b. 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.

Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. 2021. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3557–3567.

Paula Chocron and Paolo Pareti. 2018. Vocabulary alignment for collaborative agents: a study with real-world multilingual how-to instructions. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 159–165. International Joint Conferences on Artificial Intelligence Organization.

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, Édouard 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.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. Xnli: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

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

Wenliang Dai, Samuel Cahyawijaya, Zihan Liu, and Pascale Fung. 2021. Multimodal end-to-end sparse model for emotion recognition. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5305–5316, Online. Association for Computational Linguistics.

Wenliang Dai, Lu Hou, Lifeng Shang, Xin Jiang, Qun Liu, and Pascale Fung. 2022. Enabling multimodal generation on CLIP via vision-language knowledge distillation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2383–2395, Dublin, Ireland. Association for Computational Linguistics.

Robby Darwis, Herry Sujaini, and Rudy Dwi Nyoto. 2019. Peningkatan mesin penerjemah statistik dengan menambah kuantitas korpus monolingual (studi kasus: Bahasa indonesia-sunda). *JUSTIN (Jurnal Sistem dan Teknologi Informasi)*, 7(1):27–32.

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, Minneapolis, Minnesota. Association for Computational Linguistics.

Kaustubh D. Dhole, Varun Gangal, Sebastian Gehrmann, Aadesh Gupta, Zhenhao Li, Saad Mahamood, Abinaya Mahendiran, Simon Mille, Ashish Shrivastava, Samson Tan, Tongshuang Wu, Jascha Sohl-Dickstein, Jinho D. Choi, Eduard Hovy, Ondrej Dusek, Sebastian Ruder, Sajant Anand, Nagender Aneja, Rabin Banjade, Lisa Barthe, Hanna Behnke, Ian Berlot-Attwell, Connor Boyle, Caroline Brun, Marco Antonio Sobrevilla Cabezudo, Samuel Cahyawijaya, Emile Chapuis, Wanxiang Che, Mukund Choudhary, Christian Clauss, Pierre Colombo, Filip Cornell, Gautier Dagan, Mayukh Das, Tanay Dixit, Thomas Dopierre, Paul-Alexis Dray, Suchitra Dubey, Tatiana Ekeinhor, Marco Di Giovanni, Tanya Goyal, Rishabh Gupta, Rishabh Gupta, Louanes Hamla, Sang Han, Fabrice Harel-Canada, Antoine Honore, Ishan Jindal, Przemyslaw K. Joniak, Denis Kleyko, Venelin Kovatchev, Kalpesh Krishna, Ashutosh Kumar, Stefan Langer, Seungjae Ryan Lee, Corey James Levinson, Hualou Liang, Kaizhao Liang, Zhexiong Liu, Andrey Lukyanenko, Vukosi Marivate, Gerard de Melo, Simon Meoni, Maxime Meyer, Afnan Mir, Nafise Sadat Moosavi, Niklas Muennighoff, Timothy Sum Hon Mun, Kenton Murray, Marcin Namysl, Maria Obedkova, Priti Oli, Nivranshu Pasricha, Jan Pfister, Richard Plant, Vinay Prabhu, Vasile Pais, Libo Qin, Shahab Raji, Pawan Kumar Rajpoot, Vikas Raunak, Roy Rinberg, Nicolas Roberts, Juan Diego Rodriguez, Claude Roux, Vasconcellos P. H. S., Ananya B. Sai, Robin M. Schmidt, Thomas Scialom, Tshephisho Sefara, Saqib N. Shamsi, Xudong Shen, Haoyue Shi, Yiwen Shi, Anna Shvets, Nick Siegel, Damien Sileo, Jamie Simon, Chandan Singh, Roman Sitelew, Priyank Soni, Taylor Sorensen, William Soto, Aman Srivastava, KV Aditya Srivatsa, Tony Sun, Mukund Varma T, A Tabassum, Fiona Anting Tan, Ryan Teehan, Mo Tiwari, Marie Tolkiehn, Athena Wang, Zijian Wang, Gloria Wang, Zijie J. Wang, Fuxuan Wei, Bryan Wilie, Genta Indra Winata, Xinyi Wu, Witold Wydmański, Tianbao Xie, Usama Yaseen, Michael A. Yee, Jing Zhang, and Yue Zhang. 2021. NL-augmenter: A framework for task-sensitive natural language augmentation. *arXiv preprint arXiv:2112.02721*.

Arawinda Dinakaramani, Fam Rashel, Andry Luthfi, and Ruli Manurung. 2014. Designing an Indonesian part of speech tagset and manually tagged Indonesian corpus. In *2014 International Conference on Asian Language Processing (IALP)*, pages 66–69.

Sumanth Doddapaneni, Rahul Aralikatte, Gowtham Ramesh, Shreya Goyal, Mitesh M Khapra, Anoop Kunchukuttan, and Pratyush Kumar. 2022. Indicxtreme: A multi-task benchmark for evaluating indic languages. *arXiv preprint arXiv:2212.05409*.

Mark Durie. 1985. *A Grammar of Acehnese on the Basis of a Dialect of North Aceh*. Verhandelingen van het Koninklijk Instituut voor Taal-, Land- en Volkenkunde, Dordrecht-Holland.

Mark Durie. 1988. Preferred argument structure in an active language: Arguments against the category 'intransitive subject'. *Lingua*, 74(1):1–25.

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

Muhammad Dwi Etsa, Herry Sujaini, and Novi Safriadi. 2018. Pengaruh metode dictionary lookup pada cleaning korpus terhadap akurasi mesin penerjemah statistik indonesia–melayu pontianak. *Jurnal Edukasi dan Penelitian Informatika (JEPIN)*, 4(1):49.

Muhammad Fachri. 2014. Named entity recognition for Indonesian text using hidden Markov model. *Universitas Gadjah Mada*.

Manaal Faruqui and Shankar Kumar. 2015. Multilingual open relation extraction using cross-lingual projection. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1351–1356.

Andri Imam Fauzi and Dwi Puspitorini. 2018. Dialect and identity: A case study of Javanese use in WhatsApp and Line. In *IOP Conference Series: Earth and Environmental Science*, volume 175, page 012111. IOP Publishing.

Jordhy Fernando, Masayu Leylia Khodra, and Ali Akbar Septiandri. 2019. Aspect and opinion terms extraction using double embeddings and attention mechanism for Indonesian hotel reviews. In *2019 International Conference of Advanced Informatics: Concepts, Theory and Applications (ICAICTA)*, pages 1–6. IEEE.

Jonathan Frankle and Michael Carbin. 2019. The lottery ticket hypothesis: Finding sparse, trainable neural networks. In *ICLR*. OpenReview.net.

Jason Alan Fries, Leon Weber, Natasha Seelam, Gabriel Altay, Debajyoti Datta, Samuele Garda, Myungsun Kang, Ruisi Su, Wojciech Kusa, Samuel Cahyawijaya, Fabio Barth, Simon Ott, Matthias Samwald, Stephen Bach, Stella Biderman, Mario Sänger, Bo Wang, Alison Callahan, Daniel León Periñán, Théo Gigant, Patrick Haller, Jenny Chim, Jose David Posada, John Michael Giorgi, Karthik Rangasai Sivaraman, Marc Pàmies, Marianna Nezhurina, Robert Martin, Michael Cullan, Moritz Freidank, Nathan Dahlberg, Shubhanshu Mishra, Shamik Bose, Nicholas Michio Broad, Yanis Labrak, Shlok S Deshmukh, Sid Kiblawi, Ayush Singh, Minh Chien Vu, Trishala Neeraj, Jonas Golde, Albert Villanova del Moral, and Benjamin Beilharz. 2022. Bigbio: A framework for data-centric biomedical natural language processing. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Pascale Fung, Chi Shun Cheung, Kwok Leung Lam, Wai Kat Liu, and Yuen Yee Lo. 1998. SALSA version 1.0: a speech-based web browser for hong kong english. In *5th International Conference on Spoken Language Processing (ICSLP 1998)*. ISCA.

Sebastian Gehrmann, Tosin Adewumi, Karmanya Aggarwal, Pawan Sasanka Ammanamanchi, Anuoluwapo Aremu, Antoine Bosselut, Khyathi Raghavi Chandu, Miruna-Adriana Clinciu, Dipanjan Das, Kaustubh Dhole, Wanyu Du, Esin Durmus, Ondřej Dušek, Chris Chinenye Emezue, Varun Gangal, Cristina Garbacea, Tatsunori Hashimoto, Yufang Hou, Yacine Jernite, Harsh Jhamtani, Yangfeng Ji, Shailza Jolly, Mihir Kale, Dhruv Kumar, Faisal Ladhak, Aman Madaan, Mounica Maddela, Khyati Mahajan, Saad Mahamood, Bodhisattwa Prasad Majumder, Pedro Henrique Martins, Angelina McMillan-Major, Simon Mille, Emiel van Miltenburg, Moin Nadeem, Shashi Narayan, Vitaly

Nikolaev, Andre Niyongabo Rubungo, Salomey Osei, Ankur Parikh, Laura Perez-Beltrachini, Niranjan Ramesh Rao, Vikas Raunak, Juan Diego Rodriguez, Sashank Santhanam, João Sedoc, Thibault Sellam, Samira Shaikh, Anastasia Shimorina, Marco Antonio Sobrevilla Cabezudo, Hendrik Strobelt, Nishant Subramani, Wei Xu, Diyi Yang, Akhila Yerukola, and Jiawei Zhou. 2021. The GEM benchmark: Natural language generation, its evaluation and metrics. In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, pages 96–120, Online. Association for Computational Linguistics.

Sebastian Gehrmann, Abhik Bhattacharjee, Abinaya Mahendiran, Alex Wang, Alexandros Papangelis, Aman Madaan, Angelina McMillan-Major, Anna Shvets, Ashish Upadhyay, 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 D. Dhole, Khyathi Raghavi Chandu, Laura Perez-Beltrachini, Leonardo F. R. Ribeiro, Lewis Tunstall, Li Zhang, Mahima 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 Stajner, Sébastien Montella, Shailza Jolly, Simon Mille, Tahmid Hasan, Tianhao Shen, Tosin P. AMahidewumi, 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. *CoRR*, abs/2206.11249.

Barbara F Grimes. 2000. *Ethnologue*. SIL International, Dallas, TX.

Yohanes Gultom and Wahyu Catur Wibowo. 2017. Automatic open domain information extraction from Indonesian text. In *2017 International Workshop on Big Data and Information Security (IWBIS)*, pages 23–30. IEEE.

Wahyu Gunawan, Herry Sujaini, and Tursina Tursina. 2021. Analisis perbandingan nilai akurasi mekanisme attention bahdanau dan luong pada neural machine translation bahasa indonesia ke bahasa melayu ketapang dengan arsitektur recurrent neural network. *Jurnal Edukasi dan Penelitian Informatika (JEPIN)*, 7(3):488.

Tri Wahyu Guntara, Alham Fikri Aji, and Radityo Eko Prasojo. 2020. Benchmarking multidomain English-Indonesian machine translation. In *Proceedings of the 13th Workshop on Building and Using Comparable Corpora*, pages 35–43, Marseille, France. European Language Resources Association.

Ashim Gupta and Vivek Srikumar. 2021. X-fact: A new benchmark dataset for multilingual fact checking. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 675–682.

Muh Haidir and Ayu Purwarianti. 2020. Short answer grading using contextual word embedding and linear regression. *Jurnal Linguistik Komputasional*, 3(2):54–61.

Akhmad Haryono. 2012. *Perubahan dan perkembangan bahasa: Tinjauan historis dan sosiolinguistik*. Ph.D. thesis, Udayana University.

Tahmid Hasan, Abhik Bhattacharjee, Md Saiful Islam, Kazi Mubasshir, Yuan-Fang Li, Yong-Bin Kang, M Sohel Rahman, and Rifat Shahriyar. 2021. Xl-sum: Large-scale multilingual abstractive summarization for 44 languages. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4693–4703.

Muhammad Hasbiansyah, Herry Sujaini, and Novi Safriadi. 2016. Tuning for quality untuk uji akurasi mesin penerjemah statistik (mps) bahasa indonesia-bahasa dayak kanayatn. *JUSTIN (Jurnal Sistem dan Teknologi Informasi)*, 4(1):209–213.

Ahmad Fathan Hidayatullah, Siwi Cahyaningtyas, and Rheza Daffa Pamungkas. 2020. Attention-based cnn-bilstm for dialect identification on javanese text. *Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control*, pages 317–324.

Devin Hoesen and Ayu Purwarianti. 2018. Investigating bi-LSTM and CRF with POS tag embedding for Indonesian named entity tagger. In *2018 International Conference on Asian Language Processing (IALP)*, pages 35–38. IEEE.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. XTREME: A massively multilingual multitask benchmark for evaluating cross-lingual generalisation. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4411–4421. PMLR.

Muhammad Okky Ibrohim and Indra Budi. 2018. A dataset and preliminaries study for abusive language detection in Indonesian social media. *Procedia Computer Science*, 135:222–229. The 3rd International Conference on Computer Science and Computational Intelligence (ICCSCI 2018) : Empowering Smart Technology in Digital Era for a Better Life.

Muhammad Okky Ibrohim and Indra Budi. 2019. Multi-label hate speech and abusive language detection in Indonesian Twitter. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 46–57, Florence, Italy. Association for Computational Linguistics.

Arfinda Ilmania, Abdurrahman, Samuel Cahyawijaya, and Ayu Purwarianti. 2018. Aspect detection and sentiment classification using deep neural network for indonesian aspect-based sentiment analysis. In *2018 International Conference on Asian Language Processing (IALP)*, pages 62–67.

Adylan Roaffa Ilmy and Masayu Leylia Khodra. 2020. Parsing Indonesian sentence into Abstract Meaning Representation using machine learning approach. *2020 7th International Conference on Advance Informatics: Concepts, Theory and Applications (ICAICTA)*, pages 1–6.

Kercedasan Artifisal Indonesia. 2020. *National Strategy for Artificial Intelligence 2020-2045 (2020) (Indonesian)*. Kercedasan Artifisal Indonesia.

Danny Indrayana. 2016. Meningkatkan akurasi pada mesin penerjemah bahasa indonesia ke bahasa melayu pontianak dengan part of speech. *JUSTIN (Jurnal Sistem dan Teknologi Informasi)*, 4(3):476–480.

Benediktus Sridin Sulu Jahang and Zita Meirina. 2021. 1,3 juta anak di ntt belum bisa berbahasa indonesia. Last accessed on 05/10/2021.

Rini Jannati, Rahmad Mahendra, Cakra Wishnu Wardhana, and Mirna Adriani. 2018. Stance classification towards political figures on blog writing. In *2018 International Conference on Asian Language Processing (IALP)*, pages 96–101. IEEE.

Ye Jia, Michelle Tadmor Ramanovich, Quan Wang, and Heiga Zen. 2022. CVSS corpus and massively multilingual speech-to-speech translation. In *Proceedings of Language Resources and Evaluation Conference (LREC)*, pages 6691–6703.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. 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.

Ichwanul Muslim Karo Karo, Mohd Farhan Md Fudzee, Shahreen Kasim, and Azizul Azhar Ramli. 2022. Sentiment analysis in karonese tweet using machine learning. *Indonesian Journal of Electrical Engineering and Informatics (IJEEI)*, 10(1):219–231.

Dhamir Raniah Kiasati Desrul and Ade Romadhony. 2019. Abusive language detection on Indonesian online news comments. In *2019 International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)*, pages 320–325.

Oddur Kjartansson, Supheakmungkol Sarin, Knot Pipatsrisawat, Martin Jansche, and Linne Ha. 2018. Crowd-sourced speech corpora for javanese, sundanese, sinhala, nepali, and bangladeshi bengali. In *Proceedings of the 6th International Workshop on Spoken Language Technologies for Under-Resourced Languages*, pages 52–55.

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.

Fajri Koto, Jey Han Lau, and Timothy Baldwin. 2020a. Liputan6: A large-scale Indonesian dataset for text summarization. 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 598–608, Suzhou, China. 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.

Fajri Koto, Afshin Rahimi, Jey Han Lau, and Timothy Baldwin. 2020b. 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.

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.

Kemal Kurniawan. 2019. KaWAT: A word analogy task dataset for Indonesian. *arXiv preprint arXiv:10.48550*.

Kemal Kurniawan and Samuel Louvan. 2018. Indosum: A new benchmark dataset for Indonesian text summarization. In *2018 International Conference on Asian Language Processing (IALP)*, pages 215–220.

Faisal Ladhak, Esin Durmus, Claire Cardie, and Kathleen Mckeown. 2020. Wikilingua: A new benchmark dataset for cross-lingual abstractive summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4034–4048.

Septina Dian Larasati. 2012. IDENTIC corpus: Morphologically enriched Indonesian-English parallel corpus. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 902–906, Istanbul, Turkey. European Language Resources Association (ELRA).

Desi Puji Lestari. 2006. A large vocabulary continuous speech recognition system for Indonesian language. In *Proc. 15th Indonesian Scientific Conference in Japan (ISA-Japan), Hiroshima, Japan, 2006*, pages 17–22.

M. Paul Lewis, editor. 2009. *Ethnologue: Languages of the World*, sixteenth edition. SIL International, Dallas, TX, USA.

Yaobo Liang, Nan Duan, Yeyun Gong, Ning Wu, Fenfei Guo, Weizhen Qi, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, Xiaodong Fan, Ruofei Zhang, Rahul Agrawal, Edward Cui, Sining Wei, Taroon Bharti, Ying Qiao, Jiun-Hung Chen, Winnie Wu, Shuguang Liu, Fan Yang, Daniel Campos, Rangan Majumder, and Ming Zhou. 2020. XGLUE: A new benchmark datasetfor cross-lingual pre-training, understanding and generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6008–6018, Online. Association for Computational Linguistics.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona Diab, Veselin Stoyanov, and Xian Li. 2022. Few-shot learning with multilingual language models. In *Proceedings of EMNLP 2022*.

Zhaojiang Lin, Zihan Liu, Genta Indra Winata, Samuel Cahyawijaya, Andrea Madotto, Yejin Bang, Etsuko Ishii, and Pascale Fung. 2021. Xpersona: Evaluating multilingual personalized chatbot. In *Proceedings of the 3rd Workshop on Natural Language Processing for Conversational AI*, pages 102–112.

Fangyu Liu, Emanuele Bugliarello, Edoardo Maria Ponti, Siva Reddy, Nigel Collier, and Desmond Elliott. 2021a. Visually grounded reasoning across languages and cultures. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10467–10485.

Qianchu Liu, Edoardo Maria Ponti, Diana McCarthy, Ivan Vulić, and Anna Korhonen. 2021b. AM2iCo: Evaluating word meaning in context across low-resource languages with adversarial examples. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7151–7162, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Marco Lui and Timothy Baldwin. 2012. langid. py: An off-the-shelf language identification tool. In *Proceedings of the ACL 2012 system demonstrations*, pages 25–30.

Manuel Mager, Arturo Oncevay, Abteen Ebrahimi, John Ortega, Annette Rios, Angela Fan, Ximena Gutierrez-Vasques, Luis Chiruzzo, Gustavo Giménez-Lugo, Ricardo Ramos, Ivan Vladimir Meza Ruiz, Rolando Coto-Solano, Alexis Palmer, Elisabeth Mager-Hois, Vishrav Chaudhary, Graham Neubig, Ngoc Thang Vu, and Katharina Kann. 2021. Findings of the AmericasNLP 2021 shared task on open machine translation for indigenous languages of the Americas. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 202–217.

Rahmad Mahendra, Alham Fikri Aji, Samuel Louvan, Fahrurrozi Rahman, and Clara Vania. 2021. IndoNLI: A natural language inference dataset for Indonesian. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10511–10527, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Rahmad Mahendra, Heninggar Septiantri, Haryo Akbarianto Wibowo, Ruli Manurung, and Mirna Adriani. 2018. Cross-lingual and supervised learning approach for Indonesian word sense disambiguation task. In *Proceedings of the 9th Global Wordnet Conference*, pages 245–250.

Miftahul Mahfuzh, Sidik Soleman, and Ayu Purwarianti. 2019. Improving joint layer RNN based keyphrase extraction by using syntactical features. In *2019 International Conference of Advanced Informatics: Concepts, Theory and Applications (ICAICTA)*, pages 1–6. IEEE.

Olga Majewska, Evgeniia Razumovskaia, Edoardo M. Ponti, Ivan Vulić, and Anna Korhonen. 2023. Cross-Lingual Dialogue Dataset Creation via Outline-Based Generation. *Transactions of the Association for Computational Linguistics*, 11:139–156.

Ryan McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith Hall, Slav Petrov, Hao Zhang, Oscar

Täckström, Claudia Bedini, Nuria Bertomeu Castelló, and Jungmee Lee. 2013. Universal dependency annotation for multilingual parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 92–97.

Angelina McMillan-Major, Zaid Alyafeai, Stella Biderman, Kimbo Chen, Francesco De Toni, Gérard Dupont, Hady Elsahar, Chris Emezue, Alham Fikri Aji, Suzana Ilić, Nurulaqilla Khamis, Colin Leong, Maraim Masoud, Aitor Soroa, Pedro Ortiz Suarez, Zeerak Talat, Daniel van Strien, and Yacine Jernite. 2022. Documenting geographically and contextually diverse data sources: The BigScience catalogue of language data and resources. *arXiv preprint arXiv:2201.10066*.

David Moeljadi. 2012. Usage of Indonesian possessive verbal predicates: A statistical analysis based on questionnaire and storytelling surveys. In *APLL-5 conference. SOAS, University of London*.

David Moeljadi. 2017. Building Jati: A treebank for Indonesian. In *Proceedings of The 4th Atma Jaya Conference on Corpus Studies (ConCorps 4)*, pages 1–9.

David Moeljadi and Zakariya Pamuji Aminullah. 2020. Building the old Javanese Wordnet. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2940–2946, Marseille, France. European Language Resources Association.

David Moeljadi, Aditya Kurniawan, and Debaditya Goswami. 2019. Building cendana: a treebank for informal indonesian. In *Proceedings of the 33rd Pacific Asia Conference on Language, Information and Computation*, pages 156–164. Waseda Institute for the Study of Language and Information.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2022. Crosslingual generalization through multitask finetuning. *arXiv preprint arXiv:2211.01786*.

Ferdiant Joshua Muis and Ayu Purwarianti. 2020. Sequence-to-sequence learning for Indonesian automatic question generator. In *2020 7th International Conference on Advance Informatics: Concepts, Theory and Applications (ICAICTA)*, pages 1–6. IEEE.

Sahoko Nakayama, Andros Tjandra, Sakriani Sakti, and Satoshi Nakamura. 2019. Zero-shot code-switching asr and tts with multilingual machine speech chain. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 964–971.

Vivi Adryani Nasution and Niza Ayuningtyas. 2020. The language choice of chinese community in medan: A sociolinguistics study. *JOALL (Journal of Applied Linguistics And Literature)*, 5(1):11–25.

Della Widya Ningtyas, Herry Sujaini, and Novi Safriadi. 2018. Penggunaan pivot language pada mesin penerjemah statistik bahasa inggris ke bahasa melayu sambas. *Jurnal Edukasi dan Penelitian Informatika (JEPIN)*, 4(2):173.

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*.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Hiroki Nomoto. 2022. Kyokushoushugi ni motoduku heiretsu tsuriibanku no kouchiku [building a parallel treebank based on minimalism]. In *Proceedings of the Twenty-Eighth Annual Meeting of the Association for Natural Language Processing*, pages 103–107.

Hiroki Nomoto, Kenji Okano, David Moeljadi, and Hideo Sawada. 2018. TUFS Asian Language Parallel Corpus (TALPCO). In *Proceedings of the Twenty-Fourth Annual Meeting of the Association for Natural Language Processing*.

Sashi Novitasari, Andros Tjandra, Sakriani Sakti, and Satoshi Nakamura. 2020. Cross-lingual machine speech chain for Javanese, Sundanese, Balinese, and Bataks speech recognition and synthesis. In *Proc. Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, pages 131–138, Marseille, France.

Eka Qadri Nuranti, Evi Yulianti, and Husna Sarirah Husin. 2022. Predicting the category and the length of punishment in Indonesian courts based on previous court decision documents. *Computers*, 11(6):88.

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

Afifah Nurlaila, Wiranto Wiranto, and Ristu Sapton. 2017. Classification of Customers Emotion using Naive Bayes Classifier (Case Study: Natasha Skin

Care). In *ITSMART: Jurnal Teknologi dan Informasi*, volume 6.

Jeroen Ooms. 2022. *cld3: Google's Compact Language Detector 3*. Https://docs.ropensci.org/cld3/, https://github.com/ropensci/cld3 (devel) https://github.com/google/cld3 (upstream).

Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. 2017. Cross-lingual name tagging and linking for 282 languages. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1946–1958.

Seunghyun Park, Seung Shin, Bado Lee, Junyeop Lee, Jaeheung Surh, Minjoon Seo, and Hwalsuk Lee. 2019. {CORD}: A consolidated receipt dataset for post-{ocr} parsing. In *Workshop on Document Intelligence at NeurIPS 2019*.

Sungjoon Park, Jihyung Moon, Sungdong Kim, Won Ik Cho, Ji Yoon Han, Jangwon Park, Chisung Song, Junseong Kim, Youngsook Song, Taehwan Oh, Joohong Lee, Juhyun Oh, Sungwon Lyu, Younghoon Jeong, Inkwon Lee, Sangwoo Seo, Dongjun Lee, Hyunwoo Kim, Myeonghwa Lee, Seongbo Jang, Seungwon Do, Sunkyoung Kim, Kyungtae Lim, Jongwon Lee, Kyumin Park, Jamin Shin, Seonghyun Kim, Lucy Park, Alice Oh, Jung-Woo Ha, and Kyunghyun Cho. 2021. KLUE: Korean language understanding evaluation. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Anne E Peng. 2011. Head-final and head-initial relative clauses in jambi teochew. In *Online Proceedings of GLOW in Asia Workshop for Young Scholars*, volume 262, page 276.

Jonas Pfeiffer, Naman Goyal, Xi Lin, Xian Li, James Cross, Sebastian Riedel, and Mikel Artetxe. 2022. Lifting the curse of multilinguality by pre-training modular transformers. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3479–3495, Seattle, United States. Association for Computational Linguistics.

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.

Jason Phang, Iacer Calixto, Phu Mon Htut, Yada Pruksachatkun, Haokun Liu, Clara Vania, Katharina Kann, and Samuel R. Bowman. 2020. English intermediate-task training improves zero-shot cross-lingual transfer too. 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 557–575, Suzhou, China. Association for Computational Linguistics.

Tiago Pimentel, Maria Ryskina, Sabrina J. Mielke, Shijie Wu, Eleanor Chodroff, Brian Leonard, Garrett Nicolai, Yustinus Ghanggo Ate, Salam Khalifa, Nizar Habash, Charbel El-Khaissi, Omer Goldman, Michael Gasser, William Lane, Matt Coler, Arturo Oncevay, Jaime Rafael Montoya Samame, Gema Celeste Silva Villegas, Adam Ek, Jean-Philippe Bernardy, Andrey Shcherbakov, Aziyana Bayyr-ool, Karina Sheifer, Sofya Ganieva, Matvey Plugaryov, Elena Klyachko, Ali Salehi, Andrew Krizhanovsky, Natalia Krizhanovsky, Clara Vania, Sardana Ivanova, Aelita Salchak, Christopher Straughn, Zoey Liu, Jonathan North Washington, Duygu Ataman, Witold Kieraś, Marcin Woliński, Totok Suhardijanto, Niklas Stoehr, Zahroh Nuriah, Shyam Ratan, Francis M. Tyers, Edoardo M. Ponti, Grant Aiton, Richard J. Hatcher, Emily Prud'hommeaux, Ritesh Kumar, Mans Hulden, Botond Barta, Dorina Lakatos, Gábor Szolnok, Judit Ács, Mohit Raj, David Yarowsky, Ryan Cotterell, Ben Ambridge, and Ekaterina Vylomova. 2021. SIGMORPHON 2021 shared task on morphological reinflection: Generalization across languages. In *Proceedings of the 18th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*.

Femphy Pisceldo, Rahmad Mahendra, Ruli Manurung, and I Wayan Arka. 2008. A two-level morphological analyser for the Indonesian language. In *Proceedings of the Australasian Language Technology Association Workshop 2008*, pages 142–150, Hobart, Australia.

Amélie PLU, MA Chi Yuen, and Pascale Fung. 2000. Salsa version 3.0: A single recognizer-based multilingual speech-based web browser. In *Content-Based Multimedia Information Access - Volume 1*, RIAO '00, page 426–430, Paris, FRA. LE CENTRE DE HAUTES ETUDES INTERNATIONALES D'INFORMATIQUE DOCUMENTAIRE.

Edoardo M. Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020. XCOPA: A multilingual dataset for causal commonsense reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Inggrid Yanuar Risca Pratiwi, Rosa Andrie Asmara, and Faisal Rahutomo. 2017. Study of hoax news detection using naïve Bayes classifier in Indonesian language. In *2017 11th International Conference on Information & Communication Technology and System (ICTS)*, pages 73–78. IEEE.

Ayu Purwarianti and Ida Ayu Putu Ari Crisdayanti. 2019. Improving bi-LSTM performance for Indonesian sentiment analysis using paragraph vector. In *2019 International Conference of Advanced Informatics: Concepts, Theory and Applications (ICAICTA)*. IEEE.

Ayu Purwarianti, Masatoshi Tsuchiya, and Seiichi Nakagawa. 2007. A machine learning approach for Indonesian question answering system. In *Artificial Intelligence and Applications*, pages 573–578.

Oddy Virgantara Putra, Fathin Muhammad Wasmanson, Triana Harmini, and Shoffin Nahwa Utama. 2020. Sundanese Twitter dataset for emotion classification. In *2020 International Conference on Computer Engineering, Network, and Intelligent Multimedia (CENIM)*, pages 391–395. IEEE.

Rifki Afina Putri and Alice Oh. 2022. IDK-MRC: Unanswerable questions for Indonesian machine reading comprehension. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6918–6933, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Shofianina Dwi Ananda Putri, Muhammad Okky Ibrohim, and Indra Budi. 2021. Abusive language and hate speech detection for javanese and sundanese languages in tweets: Dataset and preliminary study. In *2021 11th International Workshop on Computer Science and Engineering, WCSE 2021*, pages 461–465. International Workshop on Computer Science and Engineering (WCSE).

Ye Qi, Devendra Sachan, Matthieu Felix, Sarguna Padmanabhan, and Graham Neubig. 2018. When and why are pre-trained word embeddings useful for neural machine translation? In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 529–535.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision.

Riccosan, Karen Etania Saputra, Galih Dea Pratama, and Andry Chowanda. 2022. Emotion dataset from Indonesian public opinion. *Data in Brief*, 43:108465.

Hammam Riza and Chairil Hakim. 2009. Resource report: building parallel text corpora for multi-domain translation system. In *Proceedings of the 7th Workshop on Asian Language Resources (ALR7)*, pages 92–95.

Sebastian Ruder. 2022. The State of Multilingual AI. `http://ruder.io/state-of-multilingual-ai/`.

Sakriani Sakti, Arry Akhmad Arman, Satoshi Nakamura, and Paulus Hutagaol. 2004. Indonesian speech recognition for hearing and speaking impaired people. In *Proc. International Conference on Spoken Language Processing (INTERSPEECH - ICSLP)*, pages 1037–1040, Jeju Island, Korea.

Sakriani Sakti, Eka Kelana, Hammam Riza, Shinsuke Sakai, Konstantin Markov, and Satoshi Nakamura. 2008a. Development of Indonesian large vocabulary continuous speech recognition system within A-STAR project. In *Proceedings of the Workshop on Technologies and Corpora for Asia-Pacific Speech Translation (TCAST)*.

Sakriani Sakti, Ranniery Maia, Shinsuke Sakai, Tohru Shimizu, and Satoshi Nakamura. 2008b. Development of HMM-based Indonesian speech synthesis. In *Proc. Oriental COCOSDA*, volume 1.

Sakriani Sakti and Satoshi Nakamura. 2013. Towards language preservation: Design and collection of graphemically balanced and parallel speech corpora of Indonesian ethnic languages. In *2013 International Conference Oriental COCOSDA held jointly with 2013 Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE)*, pages 1–5. IEEE.

Sakriani Sakti and Satoshi Nakamura. 2014. Recent progress in developing grapheme-based speech recognition for Indonesian ethnic languages: Javanese, Sundanese, Balinese and Bataks. In *Proc. 4th Workshop on Spoken Language Technologies for Under-Resourced Languages (SLTU 2014)*, pages 46–52, St. Petersburg, Russia.

Sakriani Sakti, Michael Paul, Andrew Finch, Xinhui Hu, Jinfu Ni, Noriyuki Kimura, Shigeki Matsuda, Chiori Hori, Yutaka Ashikari, Hisashi Kawai, Hideki Kashioka, Eiichiro Sumita, and Satoshi Nakamura. 2012. Distributed speech translation technologies for multiparty multilingual communication. *ACM Trans. Speech Lang. Process.*, 9(2).

Sakriani Sakti, Michael Paul, Andrew Finch, Shinsuke Sakai, Thang Tat Vu, Noriyuki Kimura, Chiori Hori, Eiichiro Sumita, Satoshi Nakamura, Jun Park, Chai Wutiwiwatchai, Bo Xu, Hammam Riza, Karunesh Arora, Chi Mai Luong, and Haizhou Li. 2013. A-STAR: Toward translating asian spoken languages. *Computer Speech & Language*, 27(2):509–527.

Sakriani Sakti, Shinsuke Sakai, Ryosuke Isotani, Hisashi Kawai, and Satoshi Nakamura. 2010. Quality and intelligibility assessment of indonesian hmm-based speech synthesis system. In *Proc. MALINDO*, pages 51–57, Jakarta, Indonesia.

Muhammad Saleh, Syukur Kholil, and Ahmad Tamrin Sikumbang. 2018. Chinese ethnic communication pattern in the environment of indigenous people in lhokseumawe, indonesia. *Budapest International Research and Critics Institute-Journal (BIRCI-Journal) Vol I (4)*, pages 114–123.

Nikmatun Aliyah Salsabila, Yosef Ardhito Winatmoko, Ali Akbar Septiandri, and Ade Jamal. 2018. Colloquial indonesian lexicon. In *2018 International Conference on Asian Language Processing (IALP)*, pages 226–229. IEEE.

Auliya Sani, Sakriani Sakti, Graham Neubig, Tomoki Toda, Adi Mulyanto, and Satoshi Nakamura. 2012. Towards language preservation: Preliminary collection and vowel analysis of Indonesian ethnic speech data. In *Proc. Oriental COCOSDA*, pages 118–122, Macau, China.

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

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klamm, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani, Dragomir Radev, Eduardo González Ponferrada, Efrat Levkovizh, Ethan Kim, Eyal Bar Natan, Francesco De Toni, Gérard Dupont, Germán Kruszewski, Giada Pistilli, Hady Elsahar, Hamza Benyamina, Hieu Tran, Ian Yu, Idris Abdulmumin, Isaac Johnson, Itziar Gonzalez-Dios, Javier de la Rosa, Jenny Chim, Jesse Dodge, Jian Zhu, Jonathan Chang, Jörg Frohberg, Joseph Tobing, Joydeep Bhattacharjee, Khalid Almubarak, Kimbo Chen, Kyle Lo, Leandro Von Werra, Leon Weber, Long Phan, Loubna Ben allal, Ludovic Tanguy, Manan Dey, Manuel Romero Muñoz, Maraim Masoud, María Grandury, Mario Šaško, Max Huang, Maximin Coavoux, Mayank Singh, Mike Tian-Jian Jiang, Minh Chien Vu, Mohammad A. Jauhar, Mustafa Ghaleb, Nishant Subramani, Nora Kassner, Nurulaqilla Khamis, Olivier Nguyen, Omar Espejel, Ona de Gibert, Paulo Villegas, Peter Henderson, Pierre Colombo, Priscilla Amuok, Quentin Lhoest, Rheza Harliman, Rishi Bommasani, Roberto Luis López, Rui Ribeiro, Salomey Osei, Sampo Pyysalo, Sebastian Nagel, Shamik Bose, Shamsuddeen Hassan Muhammad, Shanya Sharma, Shayne Longpre, Somaieh Nikpoor, Stanislav Silberberg, Suhas Pai, Sydney Zink, Tiago Timponi Torrent, Timo Schick, Tristan Thrush, Valentin Danchev, Vassilina Nikoulina, Veronika Laippala, Violette Lepercq, Vrinda Prabhu, Zaid Alyafeai, Zeerak Talat, Arun Raja, Benjamin Heinzerling, Chenglei Si, Elizabeth Salesky, Sabrina J. Mielke, Wilson Y. Lee, Abheesht Sharma, Andrea Santilli, Antoine Chaffin, Arnaud Stiegler, Debajyoti Datta, Eliza Szczechla, Gunjan Chhablani, Han Wang, Harshit Pandey, Hendrik Strobelt, Jason Alan Fries, Jos Rozen, Leo Gao, Lintang Sutawika, M Saiful Bari, Maged S. Al-shaibani, Matteo Manica, Nihal Nayak, Ryan Teehan, Samuel Albanie, Sheng Shen, Srulik Ben-David, Stephen H. Bach, Taewoon Kim, Tali Bers, Thibault Fevry, Trishala Neeraj, Urmish Thakker, Vikas Raunak, Xiangru Tang, Zheng-Xin Yong, Zhiqing Sun, Shaked Brody, Yallow Uri, Hadar Tojarieh, Adam Roberts, Hyung Won Chung, Jaesung Tae, Jason Phang, Ofir Press, Conglong Li, Deepak Narayanan, Hatim Bourfoune, Jared Casper, Jeff Rasley, Max Ryabinin, Mayank Mishra, Minjia Zhang, Mohammad Shoeybi, Myriam Peyrounette, Nicolas Patry, Nouamane Tazi, Omar Sanseviero, Patrick von Platen, Pierre Cornette, Pierre François Lavallée, Rémi Lacroix, Samyam Rajbhandari, Sanchit Gandhi, Shaden Smith, Stéphane Requena, Suraj Patil, Tim Dettmers, Ahmed Baruwa, Amanpreet Singh, Anastasia Cheveleva, Anne-Laure Ligozat, Arjun Subramonian, Aurélie Névéol, Charles Lovering, Dan Garrette, Deepak Tunuguntla, Ehud Reiter, Ekaterina Taktasheva, Ekaterina Voloshina, Eli Bogdanov, Genta Indra Winata, Hailey Schoelkopf, Jan-Christoph Kalo, Jekaterina Novikova, Jessica Zosa Forde, Jordan Clive, Jungo Kasai, Ken Kawamura, Liam Hazan, Marine Carpuat, Miruna Clinciu, Najoung Kim, Newton Cheng, Oleg Serikov, Omer Antverg, Oskar van der Wal, Rui Zhang, Ruochen Zhang, Sebastian Gehrmann, Shani Pais, Tatiana Shavrina, Thomas Scialom, Tian Yun, Tomasz Limisiewicz, Verena Rieser, Vitaly Protasov, Vladislav Mikhailov, Yada Pruksachatkun, Yonatan Belinkov, Zachary Bamberger, Zdeněk Kasner, Alice Rueda, Amanda Pestana, Amir Feizpour, Ammar Khan, Amy Faranak, Ana Santos, Anthony Hevia, Antigona Unldreaj, Arash Aghagol, Arezoo Abdollahi, Aycha Tammour, Azadeh HajiHosseini, Bahareh Behroozi, Benjamin Ajibade, Bharat Saxena, Carlos Muñoz Ferrandis, Danish Contractor, David Lansky, Davis David, Douwe Kiela, Duong A. Nguyen, Edward Tan, Emi Baylor, Ezinwanne Ozoani, Fatima Mirza, Frankline Ononiwu, Habib Rezanejad, Hessie Jones, Indrani Bhattacharya, Irene Solaiman, Irina Sedenko, Isar Nejadgholi, Jesse Passmore, Josh Seltzer, Julio Bonis Sanz, Karen Fort, Livia Dutra, Mairon Samagaio, Maraim Elbadri, Margot Mieskes, Marissa Gerchick, Martha Akinlolu, Michael McKenna, Mike Qiu, Muhammed Ghauri, Mykola Burynok, Nafis Abrar, Nazneen Rajani, Nour Elkott, Nour Fahmy, Olanrewaju Samuel, Ran An, Rasmus Kromann, Ryan Hao, Samira Alizadeh, Sarmad Shubber, Silas Wang, Sourav Roy, Sylvain Viguier, Thanh Le, Tobi Oyebade, Trieu Le, Yoyo Yang, Zach Nguyen, Abhinav Ramesh Kashyap, Alfredo Palasciano, Alison Callahan, Anima Shukla, Antonio Miranda-Escalada, Ayush Singh, Benjamin Beilharz, Bo Wang, Caio Brito, Chenxi Zhou, Chirag Jain, Chuxin Xu, Clémentine Fourrier, Daniel León Periñán, Daniel Molano, Dian Yu, Enrique Manjavacas, Fabio Barth, Florian Fuhrimann, Gabriel Altay, Giyaseddin Bayrak, Gully Burns, Helena U. Vrabec, Imane Bello, Ishani Dash, Jihyun Kang, John Giorgi, Jonas Golde, Jose David Posada, Karthik Rangasai Sivaraman, Lokesh Bulchandani, Lu Liu, Luisa Shinzato, Madeleine Hahn de Bykhovetz, Maiko Takeuchi, Marc Pàmies, Maria A Castillo, Marianna Nezhurina, Mario Sänger, Matthias Samwald, Michael Cullan, Michael Weinberg, Michiel De Wolf, Mina Mihaljcic, Minna Liu, Moritz Freidank, Myungsun Kang, Natasha Seelam, Nathan Dahlberg, Nicholas Michio Broad, Nikolaus Muellner, Pascale Fung, Patrick Haller, Ramya Chandrasekhar, Renata Eisenberg, Robert Martin, Rodrigo Canalli, Rosaline Su, Ruisi Su, Samuel Cahyawijaya, Samuele Garda, Shlok S Deshmukh, Shubhanshu Mishra, Sid Kiblawi, Simon Ott, Sinee Sang-aroonsiri, Srishti Kumar, Stefan Schweter, Sushil Bharati, Tanmay Laud, Théo Gigant, Tomoya Kainuma, Wojciech Kusa, Yanis Labrak, Yash Shailesh Bajaj, Yash Venkatraman, Yifan Xu, Yingxin Xu, Yu Xu, Zhe Tan, Zhongli Xie, Zifan Ye, Mathilde Bras, Younes Belkada, and Thomas Wolf. 2022. Bloom: A 176b-parameter open-access multilingual language model.

Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. 2021. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs.

Haitham Seelawi, Ibraheem Tuffaha, Mahmoud Gzawi, Wael Farhan, Bashar Talafha, Riham Badawi, Zyad Sober, Oday Al-Dweik, Abed Alhakim Freihat, and Hussein Al-Natsheh. 2021. ALUE: Arabic language understanding evaluation. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 173–184, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Ali Akbar Septiandri and Yosef Ardhito Winatmoko. 2020. Ukara 1.0 challenge track 1: automatic short-answer scoring in Bahasa Indonesia. *arXiv preprint arXiv:2002.12540*.

Ken Nabila Setya and Rahmad Mahendra. 2023. Semi-supervised textual entailment on indonesian wikipedia data. In *Computational Linguistics and Intelligent Text Processing*, pages 416–427, Cham. Springer Nature Switzerland.

Sheng Shen, Zhen Dong, Jiayu Ye, Linjian Ma, Zhewei Yao, Amir Gholami, Michael W. Mahoney, and Kurt Keutzer. 2020. Q-BERT: Hessian based ultra low precision quantization of BERT. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8815–8821.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. 2022. Language models are multilingual chain-of-thought reasoners. *arXiv preprint arXiv:2210.03057*.

Emmanuella Anggi Siallagan and Ika Alfina. 2023. Poetry Generation for Indonesian Pantun : Comparison Between SeqGAN and GPT-2. *Jurnal Ilmu Komputer dan Informasi (Journal of Computer Science and Information)*, 16(1):59–67.

Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. 2022. FLAVA: A foundational language and vision alignment model. In *CVPR*.

Ray Andrew Obaja Sinurat. 2019. Pembangkitan deskripsi gambar dalam bahasa indonesia dengan pendekatan semantic compositional networks. Master's thesis, Teknik Informatika, Institut Teknologi Bandung.

Masitowarni Siregar, Syamsul Bahri, and Dedi Sanjaya. 2014. Code switching and code mixing in Indonesia: Study in sociolinguistics. *English Language and Literature Studies*, 4(1):77–92.

James Neil Sneddon. 2003. *The Indonesian language: Its history and role in modern society*. UNSW Press, Sydney.

Keshan Sodimana, Pasindu-De Silva, Supheakmungkol Sarin, Oddur Kjartansson, Martin Jansche, Knot Pipatsrisawat, and Linne Ha. 2018. A step-by-step process for building tts voices using open source data and frameworks for bangla, javanese, khmer, nepali, sinhala, and sundanese. In *Proc. The 6th Intl. Workshop on Spoken Language Technologies for Under-Resourced Languages*, pages 52–55.

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

Carly J Sommerlot. 2020. *On the Syntax of West Kalimantan: Asymmetries and A'-Movement in Malayic and Land Dayak Languages*. Ph.D. thesis, The University of Texas at Arlington.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, Alice Xiang, Alicia Parrish, Allen Nie, Aman Hussain, Amanda Askell, Amanda Dsouza, Ambrose Slone, Ameet Rahane, Anantharaman S. Iyer, Anders Andreassen, Andrea Madotto, Andrea Santilli, Andreas Stuhlmüller, Andrew Dai, Andrew La, Andrew Lampinen, Andy Zou, Angela Jiang, Angelica Chen, Anh Vuong, Animesh Gupta, Anna Gottardi, Antonio Norelli, Anu Venkatesh, Arash Gholamidavoodi, Arfa Tabassum, Arul Menezes, Arun Kirubarajan, Asher Mullokandov, Ashish Sabharwal, Austin Herrick, Avia Efrat, Aykut Erdem, Ayla Karakaş, B. Ryan Roberts, Bao Sheng Loe, Barret Zoph, Bartłomiej Bojanowski, Batuhan Özyurt, Behnam Hedayatnia, Behnam Neyshabur, Benjamin Inden, Benno Stein, Berk Ekmekci, Bill Yuchen Lin, Blake Howald, Cameron Diao, Cameron Dour, Catherine Stinson, Cedrick Argueta, César Ferri Ramírez, Chandan Singh, Charles Rathkopf, Chenlin Meng, Chitta Baral, Chiyu Wu, Chris Callison-Burch, Chris Waites, Christian Voigt, Christopher D. Manning, Christopher Potts, Cindy

Ramirez, Clara E. Rivera, Clemencia Siro, Colin Raffel, Courtney Ashcraft, Cristina Garbacea, Damien Sileo, Dan Garrette, Dan Hendrycks, Dan Kilman, Dan Roth, Daniel Freeman, Daniel Khashabi, Daniel Levy, Daniel Moseguí González, Danielle Perszyk, Danny Hernandez, Danqi Chen, Daphne Ippolito, Dar Gilboa, David Dohan, David Drakard, David Jurgens, Debajyoti Datta, Deep Ganguli, Denis Emelin, Denis Kleyko, Deniz Yuret, Derek Chen, Derek Tam, Dieuwke Hupkes, Diganta Misra, Dilyar Buzan, Dimitri Coelho Mollo, Diyi Yang, Dong-Ho Lee, Ekaterina Shutova, Ekin Dogus Cubuk, Elad Segal, Eleanor Hagerman, Elizabeth Barnes, Elizabeth Donoway, Ellie Pavlick, Emanuele Rodola, Emma Lam, Eric Chu, Eric Tang, Erkut Erdem, Ernie Chang, Ethan A. Chi, Ethan Dyer, Ethan Jerzak, Ethan Kim, Eunice Engefu Manyasi, Evgenii Zheltonozhskii, Fanyue Xia, Fatemeh Siar, Fernando Martínez-Plumed, Francesca Happé, Francois Chollet, Frieda Rong, Gaurav Mishra, Genta Indra Winata, Gerard de Melo, Germán Kruszewski, Giambattista Parascandolo, Giorgio Mariani, Gloria Wang, Gonzalo Jaimovitch-López, Gregor Betz, Guy Gur-Ari, Hana Galijasevic, Hannah Kim, Hannah Rashkin, Hannaneh Hajishirzi, Harsh Mehta, Hayden Bogar, Henry Shevlin, Hinrich Schütze, Hiromu Yakura, Hongming Zhang, Hugh Mee Wong, Ian Ng, Isaac Noble, Jaap Jumelet, Jack Geissinger, Jackson Kernion, Jacob Hilton, Jaehoon Lee, Jaime Fernández Fisac, James B. Simon, James Koppel, James Zheng, James Zou, Jan Kocoń, Jana Thompson, Jared Kaplan, Jarema Radom, Jascha Sohl-Dickstein, Jason Phang, Jason Wei, Jason Yosinski, Jekaterina Novikova, Jelle Bosscher, Jennifer Marsh, Jeremy Kim, Jeroen Taal, Jesse Engel, Jesujoba Alabi, Jiacheng Xu, Jiaming Song, Jillian Tang, Joan Waweru, John Burden, John Miller, John U. Balis, Jonathan Berant, Jörg Frohberg, Jos Rozen, Jose Hernandez-Orallo, Joseph Boudeman, Joseph Jones, Joshua B. Tenenbaum, Joshua S. Rule, Joyce Chua, Kamil Kanclerz, Karen Livescu, Karl Krauth, Karthik Gopalakrishnan, Katerina Ignatyeva, Katja Markert, Kaustubh D. Dhole, Kevin Gimpel, Kevin Omondi, Kory Mathewson, Kristen Chiafullo, Ksenia Shkaruta, Kumar Shridhar, Kyle McDonell, Kyle Richardson, Laria Reynolds, Leo Gao, Li Zhang, Liam Dugan, Lianhui Qin, Lidia Contreras-Ochando, Louis-Philippe Morency, Luca Moschella, Lucas Lam, Lucy Noble, Ludwig Schmidt, Luheng He, Luis Oliveros Colón, Luke Metz, Lütfi Kerem Şenel, Maarten Bosma, Maarten Sap, Maartje ter Hoeve, Maheen Farooqi, Manaal Faruqui, Mantas Mazeika, Marco Baturan, Marco Marelli, Marco Maru, Maria Jose Ramírez Quintana, Marie Tolkiehn, Mario Giulianelli, Martha Lewis, Martin Potthast, Matthew L. Leavitt, Matthias Hagen, Mátyás Schubert, Medina Orduna Baitemirova, Melody Arnaud, Melvin McElrath, Michael A. Yee, Michael Cohen, Michael Gu, Michael Ivanitskiy, Michael Starritt, Michael Strube, Michał Swędrowski, Michele Bevilacqua, Michihiro Yasunaga, Mihir Kale, Mike Cain, Mimee Xu, Mirac Suzgun, Mo Tiwari, Mohit Bansal, Moin Aminnaseri, Mor Geva, Mozhdeh Gheini, Mukund Varma T, Nanyun Peng, Nathan Chi, Nayeon Lee, Neta Gur-Ari Krakover, Nicholas Cameron, Nicholas Roberts, Nick Doiron, Nikita Nangia, Niklas Deckers, Niklas Muennighoff, Nitish Shirish Keskar, Niveditha S. Iyer, Noah Constant, Noah Fiedel, Nuan Wen, Oliver Zhang, Omar Agha, Omar Elbaghdadi, Omer Levy, Owain Evans, Pablo Antonio Moreno Casares, Parth Doshi, Pascale Fung, Paul Pu Liang, Paul Vicol, Pegah Alipoormolabashi, Peiyuan Liao, Percy Liang, Peter Chang, Peter Eckersley, Phu Mon Htut, Pinyu Hwang, Piotr Miłkowski, Piyush Patil, Pouya Pezeshkpour, Priti Oli, Qiaozhu Mei, Qing Lyu, Qinlang Chen, Rabin Banjade, Rachel Etta Rudolph, Raefer Gabriel, Rahel Habacker, Ramón Risco Delgado, Raphaël Millière, Rhythm Garg, Richard Barnes, Rif A. Saurous, Riku Arakawa, Robbe Raymaekers, Robert Frank, Rohan Sikand, Roman Novak, Roman Sitelew, Ronan LeBras, Rosanne Liu, Rowan Jacobs, Rui Zhang, Ruslan Salakhutdinov, Ryan Chi, Ryan Lee, Ryan Stovall, Ryan Teehan, Rylan Yang, Sahib Singh, Saif M. Mohammad, Sajant Anand, Sam Dillavou, Sam Shleifer, Sam Wiseman, Samuel Gruetter, Samuel R. Bowman, Samuel S. Schoenholz, Sanghyun Han, Sanjeev Kwatra, Sarah A. Rous, Sarik Ghazarian, Sayan Ghosh, Sean Casey, Sebastian Bischoff, Sebastian Gehrmann, Sebastian Schuster, Sepideh Sadeghi, Shadi Hamdan, Sharon Zhou, Shashank Srivastava, Sherry Shi, Shikhar Singh, Shima Asaadi, Shixiang Shane Gu, Shubh Pachchigar, Shubham Toshniwal, Shyamolima Upadhyay, Debnath, Siamak Shakeri, Simon Thormeyer, Simone Melzi, Siva Reddy, Sneha Priscilla Makini, Soo-Hwan Lee, Spencer Torene, Sriharsha Hatwar, Stanislas Dehaene, Stefan Divic, Stefano Ermon, Stella Biderman, Stephanie Lin, Stephen Prasad, Steven T. Piantadosi, Stuart M. Shieber, Summer Misherghi, Svetlana Kiritchenko, Swaroop Mishra, Tal Linzen, Tal Schuster, Tao Li, Tao Yu, Tariq Ali, Tatsu Hashimoto, Te-Lin Wu, Théo Desbordes, Theodore Rothschild, Thomas Phan, Tianle Wang, Tiberius Nkinyili, Timo Schick, Timofei Kornev, Timothy Telleen-Lawton, Titus Tunduny, Tobias Gerstenberg, Trenton Chang, Trishala Neeraj, Tushar Khot, Tyler Shultz, Uri Shaham, Vedant Misra, Vera Demberg, Victoria Nyamai, Vikas Raunak, Vinay Ramasesh, Vinay Uday Prabhu, Vishakh Padmakumar, Vivek Srikumar, William Fedus, William Saunders, William Zhang, Wout Vossen, Xiang Ren, Xiaoyu Tong, Xinran Zhao, Xinyi Wu, Xudong Shen, Yadollah Yaghoobzadeh, Yair Lakretz, Yangqiu Song, Yasaman Bahri, Yejin Choi, Yichi Yang, Yiding Hao, Yifu Chen, Yonatan Belinkov, Yu Hou, Yufang Hou, Yuntao Bai, Zachary Seid, Zhuoye Zhao, Zijian Wang, Zijie J. Wang, Zirui Wang, and Ziyi Wu. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*.

Josh Stenberg. 2015. Multilingualism and the west kalimantan hakka. In *Multilingualism in the Chinese diaspora worldwide*, pages 123–140. Routledge.

Gilang Julian Suherik and Ayu Purwarianti. 2017. Experiments on coreference resolution for Indonesian language with lexical and shallow syntactic features. In *2017 5th International Conference on Information and Communication Technology (ICoIC7)*.

Herry Sujaini. 2019. Penggunaan bahasa indonesia sebagai pivot language pada mesin penerjemah madura-sunda dengan metode transfer dan triangulation. *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, 3(2):170–175.

Herry Sujaini. 2020. Improving the role of language model in statistical machine translation (Indonesian-Javanese). *International Journal of Electrical and Computer Engineering (IJECE)*, 10(2):2102–2109.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2021. Multilingual translation from denoising pre-training. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3450–3466, Online. Association for Computational Linguistics.

C Tho, Y Heryadi, L Lukas, and A Wibowo. 2021. Code-mixed sentiment analysis of indonesian language and javanese language using lexicon based approach. *Journal of Physics: Conference Series*, 1869(1):012084.

Bayu Distiawan Trisedya and Dyah Inastra. 2014. Creating Indonesian-Javanese parallel corpora using Wikipedia articles. In *2014 International Conference on Advanced Computer Science and Information System*, pages 239–245.

Motomitsu Uchibori and Norio Shibata. 1988. Ngaju-Dayak Language. *The Sanseido Encyclopedia of Linguistics: Languages of The World*, 1:1156–1160.

Jörgen Valk and Tanel Alumäe. 2021. Voxlingua107: a dataset for spoken language recognition. In *2021 IEEE Spoken Language Technology Workshop (SLT)*, pages 652–658. IEEE.

Rob van der Goot, Alan Ramponi, Arkaitz Zubiaga, Barbara Plank, Benjamin Muller, Iñaki San Vicente Roncal, Nikola Ljubešic, Özlem Çetinoğlu, Rahmad Mahendra, Talha Çolakoglu, Timothy Baldwin, Tommaso Caselli, and Wladimir Sidorenko. 2021a. MultiLexNorm: A shared task on multilingual lexical normalization. In *Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, pages 493–509. Association for Computational Linguistics.

Rob van der Goot, Ibrahim Sharaf, Aizhan Imankulova, Ahmet Üstün, Marija Stepanovic, Alan Ramponi, Siti Oryza Khairunnisa, Mamoru Komachi, and Barbara Plank. 2021b. From masked language modeling to translation: Non-english auxiliary tasks improve zero-shot spoken language understanding. In *NAACL-HLT*.

Yohana Veniranda. 2015. *Perfective aspect and negation in Pontianak Teochew*. Ph.D. thesis, University of Delaware.

Mirda Wahyuni, Herry Sujaini, and Hafiz Muhardi. 2019. Pengaruh kuantitas korpus monolingual terhadap akurasi mesin penerjemah statistik. *Jurnal Sistem dan Teknologi Informasi (JUSTIN)*, 7:20.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Changhan Wang, Anne Wu, Jiatao Gu, and Juan Pino. 2021. CoVoST 2 and Massively Multilingual Speech Translation. In *Proc. Interspeech 2021*, pages 2247–2251.

Sukardi Weda. 2016. Syntactic variation of buginese, a language in austronesian great family. *Kongres Internasional Masyarakat Linguistik Indonesia (KIMLI) 2016*, pages 838–841.

Haryo Akbarianto Wibowo, Made Nindyatama Nityasya, Afra Feyza Akyürek, Suci Fitriany, Alham Fikri Aji, Radityo Eko Prasojo, and Derry Tanti Wijaya. 2021. IndoCollex: A testbed for morphological transformation of Indonesian colloquial words. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3170–3183, Online. Association for Computational Linguistics.

Haryo Akbarianto Wibowo, Tatag Aziz Prawiro, Muhammad Ihsan, Alham Fikri Aji, Radityo Eko Prasojo, Rahmad Mahendra, and Suci Fitriany. 2020. Semi-supervised low-resource style transfer of Indonesian informal to formal language with iterative forward-translation. In *2020 International Conference on Asian Language Processing (IALP)*, pages 310–315. IEEE.

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.

Andika William and Yunita Sari. 2020. CLICK-ID: A novel dataset for Indonesian clickbait headlines. *Data in Brief*, 32:106231.

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.

Genta Indra Winata, Samuel Cahyawijaya, Zhaojiang Lin, Zihan Liu, and Pascale Fung. 2020. Lightweight and efficient end-to-end speech recognition using low-rank transformer. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6144–6148.

Cahya Wirawan. 2022. indonesian-nlp/librivox-indonesia.

Wilson Wongso, Henry Lucky, and Derwin Suhartono. 2022. Pre-trained transformer-based language models for sundanese. *Journal of Big Data*, 9(1):1–17.

Wilson Wongso, David Samuel Setiawan, and Derwin Suhartono. 2021. Causal and masked language modeling of javanese language using transformer-based architectures. In *2021 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, pages 1–7.

Geoffrey Woollams. 2005. Karo batak. *The Austronesian Languages of Asia and Madagascar*, pages 534–561.

Liang Xu, Hai Hu, Xuanwei Zhang, Lu Li, Chenjie Cao, Yudong Li, Yechen Xu, Kai Sun, Dian Yu, Cong Yu, Yin Tian, Qianqian Dong, Weitang Liu, Bo Shi, Yiming Cui, Junyi Li, Jun Zeng, Rongzhao Wang, Weijian Xie, Yanting Li, Yina Patterson, Zuoyu Tian, Yiwen Zhang, He Zhou, Shaoweihua Liu, Zhe Zhao, Qipeng Zhao, Cong Yue, Xinrui Zhang, Zhengliang Yang, Kyle Richardson, and Zhenzhong Lan. 2020. CLUE: A Chinese language understanding evaluation benchmark. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4762–4772, Barcelona, Spain (Online). International Committee on Computational Linguistics.

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.

Evi Yulianti, Ajmal Kurnia, Mirna Adriani, and Yoppy Setyo Duto. 2021. Normalisation of Indonesian-English code-mixed text and its effect on emotion classification. *Int. J. Adv. Comput. Sci. Appl.*, 12(11).

Wei Zhang, Lu Hou, Yichun Yin, Lifeng Shang, Xiao Chen, Xin Jiang, and Qun Liu. 2020. TernaryBERT: Distillation-aware ultra-low bit BERT. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 509–521, Online. Association for Computational Linguistics.

Appendix

Lang Code	Lang Name	Family
ace	Acehnese	MP
abl	Lampung Nyo	MP
ban	Balinese	MP
bbc	Batak Toba	MP
bjn	Banjar	MP
btk	Batak	MP
btx	Batak Karo	MP
bug	Buginese	MP
hak	Hakka/Khek	ST
ind	Indonesian	MP
jav	Javanese	MP
mad	Madura	MP
min	Minangkabau	MP
nan	Min Nan (Teochew)	ST
nij	Ngaju	MP
sun	Sundanese	MP
tpi	Tok Pisin	CR
tdt	Tetun Dili	CR
xdy	Malayic Dayak	MP

Table A: Language codes and its complete names for all 19 languages listed in NusaCrowd. **MP** denotes Malayo-Polynesian, **CR** denotes Creole, and **ST** denotes Sino-Tibetan language family.

Figure A: Language family tree for all the languages covered in NusaCrowd. Most languages are Austronesian with two Creole languages and two other languages are under Sino-Tibetan language family.

A Languages in NusaCrowd

Table A provides the language code, name, and family for all 19 languages listed in NusaCrowd. The language family information is collected from Ethnologue (Eberhard et al., 2021). We follow the ISO 639-3 standard[12] for language coding in NusaCrowd. The language tree of all languages in NusaCrowd is shown in Figure A.

Acehnese (ace) is a language spoken mainly in the Aceh province. Although it is the de facto language of provincial identity of Aceh, language use is shifting to Indonesian in urban areas. Acehnese has features typical of the Mon-Khmer languages of mainland Southeast Asia, a result of its former status as part of the early Chamic dialect continuum on the coast of Vietnam. It has at least ten contrasting vowels and as many distinct diphthongs, as well as voiceless aspirated stops and murmured voiced stops (Blust, 2013). In addition to the large number of diphthongs, it has a high percentage of monosyllabic root morphemes. Prefixes and infixes play an active role while suffixes are absent (Durie, 1985). It is of the 'active' or so-called 'Split-S' type: some intransitive verbs take arguments, which have the properties of 'transitive subjects' while others take arguments with the properties of 'transitive objects' (Durie, 1988).

Lampung Nyo (abl) is a language spoken in three enclaves east between Kanan and Seputih rivers in Lampung province. It is one of the three languages under the subgroup Lampung. The other two languages are Komering and Lampung Api. It has four dialects: Abung, Tulangbawang, Sukadana, and Melinting, with 77% of lexical similarity among dialects. It was written in Kaganga script but it is written mainly in Latin script (Eberhard et al., 2021).

Balinese (ban) is a language spoken mainly in the Bali province and in the West Nusa Tenggara province. It has three main dialects: Highland Balinese, Lowland Balinese, and Nusa Penida. It is mainly written in the Latin script since the early 20th century although it has its own Balinese script. The word order in Balinese is SVO. It is non-tonal and has 17 consonant and 6 vowel phonemes. Stress is on the penultimate syllable. It has three sociolinguistic registers. Regarding patterns of verb affixation, Balinese is an 'active' or 'split-S' language: verbs with Undergoer-like

[12] https://iso639-3.sil.org/

subject arguments are marked in one way (with a 'zero prefix'), while verbs with Actor-like subject arguments—intransitive or transitive—are marked in another (either with the nasal prefix 'N-', or with 'ma-') (Arka, 2003).

Toba Batak (bbc) is a language spoken in the North Sumatra province. Similarly to Acehnese, 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 predicate-initial, and have verb systems reminiscent of Philippine languages, although they differ from them in many details (Blust, 2013).

Banjarese (bjn) is a language spoken in Kalimantan (Central, East, South, and West Kalimantan provinces). It became a language of wider communication through trade in the market, in business, and in media. It is dominant in the South Kalimantan Province and also growing rapidly in the Central and Eastern Kalimantan provinces. It has two main dialects: Kuala and Hulu dialects. Although it is a Malayic language, it has many Javanese loanwords, probably acquired during the Majapahit period from the late thirteenth century until the fifteenth century (Blust, 2013). It has 73% of lexical similarity with Indonesian[13] and it is written in Arabic and Latin scripts.

Batak languages (btk) are a subgroup of the languages of Northwest Sumatra-Barrier Islands 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). The Batak languages were written using the Batak script, but the Latin script is now used for most writing.

Batak Karo (btx) is a language spoken in Aceh province and North Sumatra province. The language status is threatened. The lexical similarity is 81% with Batak Dairi (btd), 80% with Batak Simalungun (bts), and 76% with Batak Alas-Kluet (btz) (Woollams, 2005). It has 17 consonants and 7 vowels. The stress is on the penultimate syllable. Similar to Indonesian, it has inclusive/exclusive pronouns. The basic word order is SVO with prepositions. It is a head initial language, except for the order of quantifiers. It has two voices: actor-voice and undergoer-voice. It is written in Batak script and also Latin script.

Buginese (bug) is a language spoken mainly in the South Sulawesi, Southeast Sulawesi, Central Sulawesi, and West Sulawesi provinces. The word order is SVO. Verb affixes are used to mark persons. It is non-tonal and has 19 consonant and 6 vowel phonemes. Stress is on the penultimate syllable. It was written in the Buginese script in the past (derived from Brahmi script) but is mainly written in Latin script now (Eberhard et al., 2021). In Buginese, the pronoun 'I' has three forms: the independent form 'iyya', the ergative form '-ka', and the absolutive form/clitic 'u-'. Buginese employs sentence patterns, pronouns, and certain terms to express politeness (Weda, 2016).

Hakka (hak) is a language spoken in Southeastern China, mainly in Guangdong province, also in Fujian, Guangxi, Hainan, Hunan, south Jiangxi, and Sichuan provinces. It is also spoken by Chinese descendants in some parts in Indonesia, such as in Singkawang in West Kalimantan province (Stenberg, 2015), in Medan in North Sumatra province (Nasution and Ayuningtyas, 2020), and in Lhokseumawe in Aceh province (Saleh et al., 2018). It is a tonal language and the basic word order is SVO. It is written in Han script and also Latin script.

Indonesian (ind) is the national language of Indonesia in 1945 Constitution, Article 36. Its lexical similarity to Standard Malay is over 80%. The word order is SVO. It is non-tonal and has 19 consonants, 6 vowels, and 3 diphthongs. The stress is on the penultimate syllable. It has a rich affixation system, including a variety of prefixes, suffixes, circumfixes, and reduplication. Most of the affixes in Indonesian are derivational (Pisceldo et al., 2008). It is developed from literary 'Classical Malay' of the Riau-Johor sultanate (Sneddon, 2003) and has regional variants. It is written in Latin script.

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 20th century is mostly written in Latin script. Javanese differs from most other languages of western In-

[13] i.e., 73% of its words also occur in Indonesian.

donesia in contrasting dental and retroflex stops, and in the feature of breathy voice or murmur 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, 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).

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

Min Nan (nan) is a language spoken in Southeastern China. One of its dialects is Chaozhou-Shantou (Chao-Shan dialect) or Teochew dialect. It is spoken by Chinese descendants in some parts of Indonesia such as in Jambi (Peng, 2011) and in Pontianak in West Kalimantan province (Veniranda, 2015). While Teochew is historically Chinese, its contact with languages in Indonesia has resulted in some changes uncharacteristic of Chinese languages. For example, regarding word order, Teochew spoken in Jambi exhibits both head-final and head-initial relative clauses even though head-initial relative clauses are generally ungrammatical in Chinese languages. In addition to the head-initial word order, Jambi Teochew has also borrowed the Malay relativizer *yang* (Peng, 2011). It is a tonal language with tone sandhi. The word order is SVO (Eberhard et al., 2021).

Ngaju (nij) is a language spoken in the Central Kalimantan province. It is widely used as a language of wider communication for trade in much of Kalimantan, from the Barito to the Sampit river. It is used in many domains (church, school, village-level government, market, etc.). It has various affixes and reduplication, similar to Indonesian. The active voice is marked by prefix 'maN-' and the passive voice is marked by prefix 'iN-'. The word order is similar to the one in Indonesian. The pronouns have enclitic forms to mark possessors in a noun phrase or agents in a passive sentence (Uchibori and Shibata, 1988).

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 is 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).

Tok Pisin (tpi) is an English-based creole and de facto the national language of Papua New Guinea, a neighboring country of Indonesia. Dialect differences exist among lowlands, highlands, and islands. Highlands lexicon has more English influence. It is a non-tonal language and has 16 consonant and 5 vowel phonemes. It has inclusive/exclusive pronouns and the basic word order is SVO. It is written in Latin script (Eberhard et al., 2021).

Tetun Dili (tdt) is a Tetun-based creole spoken in Dili district, East Timor north coast as the first language and scattered in western part of East Timor as the second language. It is a statutory national language according to the 2002 Constitution, Article 13. It has heavy Portuguese (por) and Mambae (mgm) influence as well as some Indonesian (ind) or Malay influence. It is a non-tonal language with 22 consonants and 5 vowels. The stress is most commonly on the penultimate syllable. It has inclusive/exclusive pronouns. The basic word order is SVO with prepositions and tense-aspect markers. It is a head-initial language, except for possessors. The speakers of Tetun Dili also use Tetun [tet], some bilingually, but many others have significant difficulty understanding it in many domains. It is written in Latin script (Eberhard et al., 2021).

Malayic Dayak (xdy) is a language widely dispersed in Central and West Kalimantan provinces. It has many dialects and it is written in Latin script (Eberhard et al., 2021). Malayic Dayak is not a proper subgroup, but refers to the large number of unclassified but clearly Malayic languages of Borneo which have a three voice system (Sommerlot, 2020).

B Schemas in NusaCrowd

Schema serves to define and format the attributes of the dataset returned by a data loader. For each data loader, we implement a source schema, which is responsible to present the dataset in a format similar to its original structure, and a nusantara schema, which supports the standardization data structure across similar tasks.

We define the nusantara schemas as follows. Labels are in string format unless indicated otherwise.

- **Image-text (`IMTEXT`)**. This schema could be used for image captioning, text-to-image generation, and vision-language pre-training. It consists of `(id, text, image_paths, metadata)`, where `id` denotes a unique row identifier of the dataset, `text` denotes an input text, `image_paths` denotes a list of paths to the input image sources, and `metadata` denotes relevant details such as visual concepts and labels (if required).
- **Speech-text (`SPTEXT`)**. This could be used for speech recognition, text-to-speech (TTS) or speech synthesis, and speech-to-text translation. It consists of `(id, path, audio, text, speaker_id, metadata)`, where `id` denotes a unique row identifier of the dataset, `path` denotes the file path to an input audio source, `audio` denotes the audio data loaded from the corresponding `path`, `text` denotes an input text, `speaker_id` denotes a unique identifier of the speaker, `metadata` denotes relevant details such as the age and gender of the speaker (if required).
- **Speech-to-speech (`S2S`)**. This could be used for speech-to-speech translation. It consists of `(id, path_1, audio_1, text_1, metadata_1, path_2, audio_2, text_2, metadata_2)`, where `id` denotes a unique row identifier of the dataset, `path_1` and `path_2` denote the file path to a respective input audio source, `audio_1` and `audio_2` denote the audio data loaded from the corresponding `path`, `text_1` and `text_2` denote input texts, and `metadata_1` and `metadata_2` denote relevant details such as the age of the speaker and their gender (if required).
- **Unlabeled text (`SSP`)**. This schema could be used for language modeling in self-supervised pre-training. It consists of `(id, text)`, where `id` denotes a unique row identifier of the dataset and `text` denotes an input text.
- **Single-label text classification (`TEXT`)**. This schema could be used for sentiment analysis, emotion classification, legal classification, and others. It consists of `(id, text, label)`, where `id` denotes a unique row identifier of the dataset, `text` denotes an input text, and `label` denotes a deterministic target variable.
- **Multi-label text classification (`TEXT MULTI`)**. This schema could be used for hate speech detection and aspect-based sentiment analysis. It consists of `(id, text, labels)`, where `id` denotes a unique row identifier of the dataset, `text` denotes an input text, and `labels` denotes a list of deterministic target variables.
- **Text-to-text (`T2T`)**. This schema could be used for machine translation, summarization, and paraphrasing. It consists of `(id, text_1, text_2, text_1_name, text_2_name)`, where `id` denotes a unique row identifier of the dataset, `text_1` and `text_2` denote an input text pair, and `text_1_name` and `text_2_name` denote the names of the input text pair (e.g., `ind` and `jav` for translation input text pairs, or `document` and `summary` for summarization input text pairs).
- **Sequence labeling (`SEQ LABEL`)**. This schema could be used for named entity recognition (NER), POS tagging, and others. It consists of `(id, tokens, labels)`, where `id` denotes a unique row identifier of the dataset, `tokens` denotes a list of tokens of an input text, and `labels` denotes a list of targets for the tokens.
- **Question answering (`QA`)**. This schema could be used for extractive QA, multiple-choice QA, and others. It consists of `(id, question_id, document_id, question, type, choices, context, answer)`, where `id` denotes a unique row identifier of the dataset, `question_id` denotes a unique identifier of the question, `document_id` denotes a unique identifier of the context document, `question` denotes an input question to be answered, `type` denotes the type of the QA task (e.g., extractive, multiple-choice,

open-generative, closed-generative, etc.), `choices` denotes a list of answer choices (if required), `context` denotes a passage that serves as the background information of the question (if required), and `answer` denotes the gold answer to the question (if required).

- **Single-label text pair classification (`PAIRS`)**. This could be used for textual entailment and next sentence prediction. It consists of `(id, text_1, text_2, label)`, where `id` denotes a unique row identifier of the dataset, `text_1` and `text_2` denote an input text pair, and `label` denotes the target variable.
- **Single-label text pair classification with continuous values or regression (`PAIRS SCORE`)**. This could be used for answer grading and semantic textual similarity. It consists of `(id, text_1, text_2, label)`, where `id` denotes a unique row identifier of the dataset, `text_1` and `text_2` denote an input text pair, and `label` denotes a target variable as a continuous value.
- **Multi-label text pair classification (`PAIRS MULTI`)**. This could be used for morphological inflection. It consists of `(id, text_1, text_2, labels)`, where `id` denotes a unique row identifier of the dataset, `text_1` and `text_2` denote an input text pair, and `labels` denotes a list of target variables.
- **Knowledge base (`KB`)**. This schema could be used for constituency parsing, dependency parsing, coreference resolution, dialogue system, and other tasks with complex structures. It consists of `(id, passages, entities, events, coreferences, relations)`. Considering its intricate structure, we encourage readers to take a look at the implementation of the knowledge base schema.

C Details for Zero-Shot Setting Experiment in NusaNLU

Model Checkpoints For the NLU experiment, we utilize 4 model checkpoints, which are: 1) BLOOMZ fine-tuned on English prompt with 3B parameters[14], 2) XGLM with 2.9B parameters[15], 3) off-the-shelf XLM-R fine-tuned on XNLI[16], and 4) XLM-R large fine-tuned on IndoNLI. For XLM-R large fine-tuned on IndoNLI, we fine-tuned the XLM-R large model with batch size of 128 and initial learning rate of 1e-5 for 50 epochs. We use AdamW optimizer with a linear learning rate decay and apply early stopping of 5 epochs based on the validation accuracy score.

Prompts We run the prompting experiment using 3 different prompts for each task type. We cover several different task types in our NLG experiments, i.e., sentiment analysis, abusive detection, hate speech detection, emotion classification, natural language inference (NLI), and next tweet prediction. The prompt templates used for each task type are shown in Tables A and A.

D Details for Zero-Shot Setting Experiment in NusaNLG

Model Checkpoints For the NLG experiment, we utilize 2 model checkpoints, i.e., BLOOMZ fine-tuned on English prompt with 3B parameters and XGLM with 2.9B parameters. We use the same checkpoint as the one used in the zero-shot NLU experiment.

Generation Hyperparameters For generating the prediction sequence, we generate sequence using greedy decoding with sampling, using top-k of 50 and top-p of 1.0. We force the model to at least generate one token and limit the generation sequence length to 100 tokens.

Prompts We run the prompting experiment using 3 different prompts for each task type. We cover two different task types in our NLG experiments, i.e., machine translation and summarization. The prompt templates used in our NLG experiment are shown in Table A and Table A.

E Details of Speech Recognition Experiment in NusaASR

Model Checkpoints For both the monolingual and multilingual ASR experiment, we employ 2 wav2vec 2.0_{LARGE} model checkpoints (both with $\sim$300M parameters) as follows: 1) pre-trained XLSR wav2vec 2.0 model[17] and an off-the-shelf fine-tuned XLSR wav2vec 2.0 model to Indoensian,

[14] https://huggingface.co/bigscience/bloomz

[15] https://huggingface.co/facebook/xglm-2.9B

[16] https://huggingface.co/joeddav/xlm-roberta-large-xnli

[17] wav2vec2-large-xlsr-53:https://huggingface.co/facebook/wav2vec2-large-xlsr-53

Sundanese, and Javanese speech data[18]. For Whisper model we employ the Whisper$_{SMALL}$[19] model with 244M parameters. For the monolingual experiment, we explore training using the 3 largest and widely-used languages in Indonesia, i.e., Indonesian (ind), Javanese (jav), and Sundanese (sun).

Fine-Tuning Hyperparameters We apply fine-tuning to both XLSR wav2vec 2.0 and Whisper models for single-task training, monolingual multi-task training, and multilingual multi-task training settings. We fine-tune the models using the following hyperparameters, i.e., Adam optimizer with a learning rate of 5e-5 for the wav2vec 2.0 model and 1e-4 for the Whisper model, training batch size of 16, fine-tuning epoch of 30, and apply an early stopping of 5 epoch based on the validation word error rate (WER). For each model, we search for the best learning rate ranging from [5e-4 ... 1e-5]. We run all experiments on a single A100 GPU.

F Zero-Shot Results of NusaNLU

Here we elaborate further on the analysis in Section 4.1. We report the overall performances of each model in Figure A and per task performance in Table A. Predictions derived by prompting BLOOMZ outperform all the other models and perform on average on par with zero-shot cross-task prompting using the XLM-R model trained on XNLI. In detail, predictions using cross-task prompting actually are better in F1 than using BLOOMZ in 17 tasks, while it's actually worse in accuracy in 13 tasks, all out of the 26 NLU tasks sampled. One extreme example can be observed in their performance comparison on the id_abusive task, where predicting by cross-task prompting XLM-R trained on XNLI nearly triples the F1 on prompting BLOOMZ. These results suggest that methods like cross-task prompting are worth exploring, benefitting better efficiency through cross-task transfer on low-resource language tasks compared to large multilingual LMs.

Comparing the languages of the prompt, although on both XGLM and BLOOMZ it's better to use the English prompt, the difference is actually more apparent on average when prompting is done using XGLM. However, when we zoom into each of the tasks, the difference is much larger in prompting using BLOOMZ. The largest spread is observed on utilizing the English prompt when predicting for the indolem sentiment analysis task, where the accuracy differs by ∼30%, and the F1 differs by ∼37.8%. Comparing the same variables in XGLM, the largest accuracy difference of ∼24% is observed on id_google_play_review_posneg, and the largest F1 difference of ∼19.1% is observed on Madurese (mad) sentiment analysis task. Furthermore, utilizing Indonesian prompts is not always the case, worse. On Buginese (bug) sentiment analysis utilizing BLOOMZ we can get ∼23% more accuracy by using Indonesian prompt. On classifying emotion in emotcmt task utilizing XGLM, we can get ∼7% more F1 by using also the Indonesian prompt. On the indolem next-tweet-prediction task, utilizing both BLOOMZ and XGLM using also the Indonesian prompt, we can get additional ∼14% accuracy and ∼23% F1 respectively.

G Zero-Shot Results of NusaNLG

Here we elaborate further on the analysis in Section 4.2. We report the overall performances of each model in Figure A and per task performance in Table A. Generations derived by prompting BLOOMZ are better than prompting XGLM in all of the tasks except in indosum_fold0_nusantara_t2t, where the scores differ slightly. The performances in the summarization tasks are generally lower than the performances in the machine translation tasks. On the machine translation tasks, the performance in translating to the Indonesian language as the target language is generally higher than translating to the local languages, while translating from English to Indonesian is generally performing the highest.

Prompting using BLOOMZ yields better performances in most of the tasks, when prompting using English prompts than using Indonesian prompts. In general, prompting using XGLM yields better generation using Indonesian prompts than using English prompts. This is especially the case in the machine translation tasks, where most of them yield better performances except when translating to Toba Batak (bbc) and Banjarese (bjn) from Indonesian (ind), and also when translating to Minangkabau (min) to Indonesia (ind) and vice versa. In the summarization task, prompting using XGLM with English prompts produce better results than with Indonesian prompts.

It's worth noting that the translation quality is

[18] https://huggingface.co/indonesian-nlp/wav2vec2-indonesian-javanese-sundanese

[19] https://huggingface.co/openai/whisper-small

extremely poor for local languages, especially in Banjarese (bjn), Acehnese (ace), Toba Batak (bbc), Ngaju (nij), Madurese (mad), and Sundanese (sun). This is even more severe when those local languages become the target languages. This finding suggests that both BLOOMZ and XGLM still fail to learn the representation of these local languages.

H ASR Results of NusaASR

Here we elaborate further on the analysis in Section 4.3. We report the per-task performance of each model in Table A for the wav2vec 2.0-pt model, Table A for the wav2vec 2.0 model-ft, and Table A for the Whisper model. The best overall performance is achieved by **wav2vec 2.0-pt** fine-tuned in multilingual multi-task setting, achieving 17.03% average WER over all tasks. The model also performs better in most cases for languages other than Indonesian compared to the **wav2vec 2.0-ft** model. While for Indonesian, **wav2vec 2.0-ft** fine-tuned in all 3 training settings, i.e., multilingual multi-task, monolingual (ind) multi-task, and single-task settings, achieve much better scores, i.e., <5% WER over all Indonesian tasks.

Comparing the performance per language, the best Indonesian ASR model achieves very low WER on Indonesian (ind) speech corpora, i.e., indspeech_digit_cdsr, indspeech_news_lvcsr, indspeech_teldialog_lvcsr, and indspeech_teldialog_svcsr. Compared to local languages, i.e., Minangkabau (min), Sundanese (sun), Javanese (jav), Balinese (ban), Acehnese (ace), Batak (btk), and Buginese (bug), the performance of the best ASR model only achieves ∼10-30% WER. The performance is especially low for Buginese (bug), which suggests distinct speech features are required for handling speech recognition in Buginese (bug) language. This fact aligns with the result of prior work in Indonesian local languages (Winata et al., 2023), where Buginese (bug) has inferior performance in the leave-one-language-out setting.

I Private Datasets in NusaCrowd

NusaCrowd offers access to 14 previously private datasets. We provide the details of all 14 previously private datasets listed in NusaCrowd along with the task, languages, and modality in Table A.

J Comparison with Other Initiatives

To provide a broader perspective of the impact of NusaCrowd, we provide the comparison of NusaCrowd initiatives with other global, regional, and Indonesian data initiatives in Table A.

K Details of Datasets in NusaCrowd

Table A provides the details description, license, languages, dataset volume, annotation quality, and other metadata of all 137 datasets collected in NusaCrowd.

Language	Prompt in Sentiment Analysis Task
Indonesian (ind)	[INPUT]\nApakah sentimen dari teks tersebut? [LABELS_CHOICE]
	Apakah sentimen dari teks berikut?\nTeks: [INPUT]\nSentimen: [LABELS_CHOICE]
	Teks: [INPUT]\n\nTolong prediksikan sentimen dari teks diatas: [LABELS_CHOICE]
English (eng)	[INPUT]\nWhat would be the sentiment of the text above? [LABELS_CHOICE]
	What is the sentiment of this text?\nText: [INPUT]\nSentiment: [LABELS_CHOICE]
	Text: [INPUT]\n\nPlease classify the sentiment of above text: [LABELS_CHOICE]

Table A: Prompt used for Sentiment Analysis task

Language	Prompt in Emotion Classification Task
Indonesian (ind)	[INPUT]\nApakah emosi dari teks diatas? [LABELS_CHOICE]
	Apakah emosi dari teks ini?\n Teks: [INPUT]\n Emosi: [LABELS_CHOICE]
	Teks: [INPUT]\n\nTolong prediksikan emosi dari teks diatas: [LABELS_CHOICE]
English (eng)	[INPUT]\nWhat would be the emotion of the text above? [LABELS_CHOICE]
	What is the emotion of this text?\nText: [INPUT]\nEmotion: [LABELS_CHOICE]
	Text: [INPUT]\n\nPlease classify the emotion of above text: [LABELS_CHOICE]

Table A: Prompt used for Emotion Classification task

Language	Prompt in Abusive Detection Task
Indonesian (ind)	[INPUT]\nApakah teks diatas kasar? [LABELS_CHOICE]
	Apakah teks berikut ini kasar?\n[INPUT]\nJawab dengan [OPTIONS]: [LABELS_CHOICE]
	[INPUT]\nApakah menurutmu teks diatas itu [OPTIONS]? [LABELS_CHOICE]
English (eng)	[INPUT]\nIs the text abusive? [LABELS_CHOICE]
	Is the following text abusive?\n[INPUT]\nAnswer with [OPTIONS]: [LABELS_CHOICE]
	[INPUT]\nDo you think the text is [OPTIONS]? [LABELS_CHOICE]

Table A: Prompt used for Abusive Detection task

Language	Prompt in Clickbait Detection Task
Indonesian (ind)	[INPUT]\nApakah judul diatas clickbait? [LABELS_CHOICE]
	Apakah judul berikut ini clickbait?\n[INPUT]\nJawab dengan [OPTIONS]: [LABELS_CHOICE]
	[INPUT]\nApakah menurutmu teks diatas itu [OPTIONS]? [LABELS_CHOICE]
English (eng)	[INPUT]\nIs the title clickbait? [LABELS_CHOICE]
	Is the following title a clickbait?\n[INPUT]\nAnswer with [OPTIONS]: [LABELS_CHOICE]
	[INPUT]\nDo you think the text is [OPTIONS]? [LABELS_CHOICE]

Table A: Prompt used for Clickbait Detection task

Language	Prompt in Rating Review Regression Task
Indonesian (ind)	[INPUT]\nBerapa rating dari teks review tersebut, dari 1 sampai 5? [LABELS_CHOICE]
	[INPUT]\nDari 1 sampai 5, berapa rating dari review diatas? [LABELS_CHOICE]
	[INPUT]\nDari 1 sampai 5 bintang, bagaimana menurutmu rating dari review tersebut? [LABELS_CHOICE]
English (eng)	[INPUT]\nWhat is the rating of the review above, from 1 to 5? [LABELS_CHOICE]
	[INPUT]\nFrom 1 to 5, what is the rating of the review above? [LABELS_CHOICE]
	[INPUT]\nFrom 1 to 5 stars, how would you rate the previous review? [LABELS_CHOICE]

Table A: Prompt used for Rating Review Regression task

Language	Prompt in Hate Speech Detection Task
Indonesian (ind)	[INPUT]\nApakah teks diatas hatespeech? [LABELS_CHOICE]
	Apakah teks berikut ini hatespeech\n[INPUT]\nJawab dengan [OPTIONS]: [LABELS_CHOICE]
	[INPUT]\nApakah menurutmu teks diatas itu [OPTIONS]? [LABELS_CHOICE]
English (eng)	[INPUT]\nDo you think the text is hatespeech? Answer: [LABELS_CHOICE]
	Is the following text a hatespeech?\n[INPUT]\nAnswer with [OPTIONS]: [LABELS_CHOICE]
	[INPUT]\nDo you think the text is [OPTIONS]? [LABELS_CHOICE]

Table A: Prompt used for Hate Speech Detection task

Language	Prompt in Next Tweet Prediction Task
Indonesian (ind)	`Diberikan dua tweet\nA: [INPUT_A]\nB: [INPUT_B]\n\nApakah tweet B adalah sambungan dari tweet A? [LABELS_CHOICE]`
	`Apakah tweet "[INPUT_B]" adalah sambungan dari tweet "[INPUT_A]"?  [LABELS_CHOICE]`
	`Tweet pertama:  [INPUT_A].\nApakah "[INPUT_B]" merupakan sambungan dari tweet pertama?  [LABELS_CHOICE]`
English (eng)	`Given two tweets\nA: [INPUT_A]\nB: [INPUT_B]\n\nIs tweet B is a continuation of tweet A? [LABELS_CHOICE]`
	`Is tweet "[INPUT_B]" a continuation of tweet "[INPUT_A]"? [LABELS_CHOICE]`
	`First Tweet:  [INPUT_A].\nWould "[INPUT_B]" a continuation of the first tweet?  [LABELS_CHOICE]`

Table A: Prompt used for Next Tweet Prediction task

Language	Prompt in NLI Task
Indonesian (ind)	`[INPUT_A]\nBerdasarkan kutipan sebelumnya, apakah benar bahwa "[INPUT_B]"?  [OPTIONS]? [LABELS_CHOICE]`
	`[INPUT_A]\n\nPertanyaan:  Apakah kalimat tersebut mengimplikasikan bahwa "[INPUT_B]"?  [OPTIONS]? [LABELS_CHOICE]`
	`Diberikan [INPUT_A]. Apakah kalimat tersebut sesuai dengan [INPUT_B]? [OPTIONS]? [LABELS_CHOICE]`
English (eng)	`[INPUT_A]\nBased on the previous passage, is it true that "[INPUT_B]"?  Yes, no, or maybe?  [LABELS_CHOICE]`
	`[INPUT_A]\n\nQuestion:  Does this imply that "[INPUT_B]"? Yes, no, or maybe?  [LABELS_CHOICE]`
	`Given that [INPUT_A]. Does it follow that [INPUT_B]? Yes, no, or maybe?  [LABELS_CHOICE]`

Table A: Prompt used for Natural Language Inference task

Language	Prompt in Summary Task
Indonesian (ind)	`[INPUT]\n===\nTulis rangkuman dari teks diatas dalam bahasa Indonesia:`
	`Artikel dalam bahasa Indonesia: [INPUT]\nRangkuman dalam bahasa Indonesia:`
	`[SOURCE]\nBagaimana kamu merangkum teks diatas dalam bahasa Indonesia?`
English (eng)	`[INPUT]\n===\nWrite a summary of the text above in Indonesian:`
	`Article in Indonesian: [INPUT]\nSummary in Indonesian:`
	`[SOURCE]\nHow would you rephrase that briefly in Indonesian?`

Table A: Prompt used for Summary task

Language	Prompt in Translation Task
Indonesian (ind)	`Terjemahkan teks berikut dari bahasa [SOURCE] ke bahasa [TARGET].\nTeks: [INPUT]\nTerjemahan:`
	`[INPUT]\nTerjemahkan teks diatas dari bahasa [SOURCE] ke bahasa [TARGET].`
	`Teks dalah bahasa [SOURCE]: [INPUT]\nBagaimana kamu menterjemahkan teks diatas dalam bahasa [TARGET]?`
English (eng)	`Translate the following text from [SOURCE] to [TARGET].\nText: [INPUT]\nTranslation:`
	`[INPUT]\nTranslate the text above from [SOURCE] to [TARGET].`
	`Text in [SOURCE]: [INPUT]\nHow would you translate that in [TARGET]?`

Table A: Prompt used for Translation task

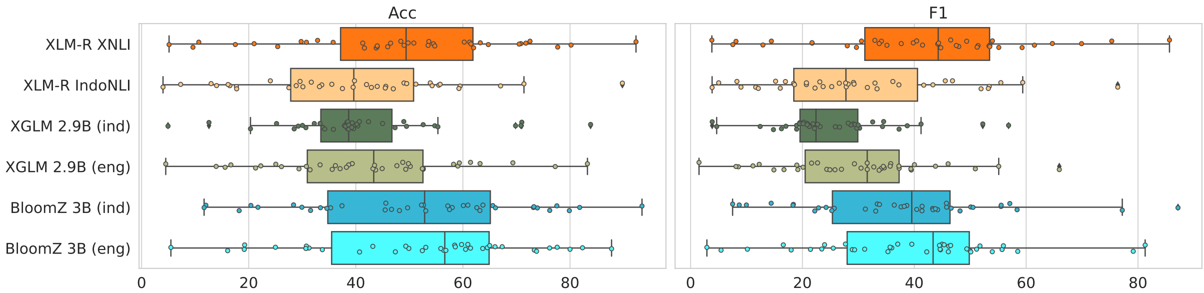

Figure A: Zero-shot generalization to NLU tasks in NusaNLU. Box plots show summary statistics on accuracy **(left)** and **F1** (right). Points are per-dataset scores from the average of performances using three different prompts.

Dataset Name	**Lang**	**Task**	**XLM-R XNLI**		**XLM-R IndoNLI**		**XGLM 2.9B (id)**		**XGLM 2.9B (en)**		**BLOOMZ 3B (id)**		**BLOOMZ 3B (en)**	
			acc	**f1**	**acc**	**f1**	**acc**	**f1**	**acc**	**f1**	**acc**	**f1**	**acc**	**f1**
code_mixed_jv_id_id	ind	SA	29.79	29.66	16.48	15.20	71.00	29.10	15.59	14.90	21.73	22.88	19.17	23.96
code_mixed_jv_id_jv	jav	SA	32.86	34.12	17.71	18.42	70.83	28.65	13.92	12.27	18.22	18.38	16.07	16.58
emot	ind	EMOT	49.77	47.95	33.64	29.14	30.68	20.82	29.39	19.18	45.61	38.66	43.18	35.29
emotcmt	ind	EMOT	43.81	41.48	28.87	25.24	32.25	25.88	26.12	19.00	45.48	33.38	56.41	44.68
emotion_id_opinion	ind	EMOT	50.83	49.37	12.98	11.65	29.25	21.90	31.46	30.48	48.15	45.79	49.90	49.98
id_abusive	ind	AD	49.01	35.11	42.91	32.97	40.38	19.72	77.24	34.08	77.54	33.11	55.87	28.92
id_abusive_news_comment	ind	AD	30.78	21.74	7.29	8.29	25.20	23.16	22.26	20.11	28.37	22.21	31.04	25.51
id_google_play_review	ind	RR	9.59	12.94	51.22	23.70	42.38	14.82	43.70	16.70	73.90	37.64	63.70	35.97
id_google_play_review_posneg	ind	SA	92.32	85.63	89.78	76.38	83.84	52.25	83.25	65.93	93.44	87.15	87.76	81.28
id_hatespeech	ind	HSD	77.70	75.31	59.47	59.39	40.21	34.32	36.19	32.76	62.60	40.44	64.56	42.23
id_hoax_news	ind	HD	48.40	46.51	45.60	45.54	37.87	30.23	39.33	34.64	53.07	38.20	53.07	44.96
id_hsd_nofaaulia	ind	HSD	72.53	51.40	67.03	51.94	30.04	28.18	30.77	29.30	63.74	41.40	76.19	51.75
id_short_answer_grading	ind	SAG	17.50	7.56	17.75	5.03	12.58	4.67	21.29	8.55	20.41	8.70	19.24	5.45
id_stance	ind	SD	10.68	8.12	24.04	16.04	48.96	26.64	63.20	25.82	31.55	14.41	58.46	29.33
imdb_jv	jav	SA	21.01	14.45	33.22	20.71	49.47	38.74	48.77	39.44	37.42	31.25	31.27	31.13
indo_law	ind	LG	53.64	53.51	53.90	53.40	52.58	34.52	52.59	34.46	52.64	43.29	47.38	35.67
indolem_ntp	ind	NTP	61.26	32.95	31.69	20.72	33.87	32.56	43.20	40.03	76.21	57.08	77.60	53.95
indolem_sentiment	ind	SA	70.82	69.91	55.49	55.48	69.83	56.86	59.05	55.10	81.80	77.19	82.33	79.15
indonli	ind	NLI	35.77	28.02	35.52	27.84	35.82	29.71	35.59	24.36	52.27	41.38	56.82	44.62
indotacos	ind	LG	5.13	3.81	16.15	9.01	4.93	3.87	4.51	1.51	12.00	7.52	5.47	2.92
jadi_ide	jav	DI	41.61	33.73	30.12	28.18	33.33	21.06	33.33	22.70	33.40	18.31	33.75	17.99
karonese_sentiment	btx	SA	41.30	39.82	27.50	16.96	34.20	17.04	37.07	21.63	35.30	25.70	40.70	21.44
nusax_senti_ace	ace	SA	53.50	44.74	39.50	22.25	38.42	19.52	44.92	33.71	60.92	46.59	60.92	46.17
nusax_senti_ban	ban	SA	54.50	44.20	45.50	31.61	39.33	21.23	50.17	37.30	60.08	45.75	61.42	46.50
nusax_senti_bjn	bjn	SA	61.50	53.97	47.00	33.34	38.33	18.97	38.67	26.98	49.67	36.42	52.42	39.56
nusax_senti_bug	bug	SA	44.00	40.43	37.50	18.58	39.58	22.15	49.33	35.79	65.58	50.09	67.33	51.14
nusax_senti_eng	eng	SA	71.75	61.48	55.75	43.33	38.08	18.95	38.25	24.91	46.67	33.78	49.33	37.04
nusax_senti_ind	ind	SA	70.50	59.28	59.25	46.83	54.75	41.21	58.25	43.79	73.17	55.51	73.33	55.70
nusax_senti_jav	jav	SA	64.75	55.11	54.25	41.69	45.00	29.42	61.42	46.05	73.25	55.54	73.75	55.93
nusax_senti_mad	mad	SA	60.25	51.24	44.00	29.23	40.67	22.70	52.58	39.47	66.08	50.50	66.00	50.11
nusax_senti_min	min	SA	62.00	53.36	49.00	36.17	38.50	19.46	47.67	35.69	58.00	44.17	60.50	46.09
nusax_senti_nij	nij	SA	54.75	47.52	42.00	26.59	38.67	20.25	52.42	39.39	63.17	48.18	65.00	49.39
nusax_senti_sun	sun	SA	63.25	53.50	49.50	37.25	38.67	20.12	43.50	31.58	57.17	43.36	59.75	45.34
nusax_senti_bbc	bbc	SA	46.50	37.67	39.75	23.11	39.25	20.85	49.83	37.25	57.50	43.77	58.58	44.50
sentiment_nathasa_review	ind	SA	25.41	21.71	14.00	12.12	20.33	12.44	16.67	8.12	29.93	24.23	30.76	27.69
smsa	ind	SA	80.20	64.69	71.40	53.24	55.33	37.28	69.33	50.92	79.87	58.38	80.00	58.50
su_emot	sun	SA	45.95	44.35	29.59	27.72	28.49	16.25	25.06	11.16	34.59	25.25	24.99	10.16
wrete	ind	ENT	44	30.56	4.00	3.85	47.33	29.97	59.33	31.59	11.67	9.98	30.67	23.46

Table A: Details of zero-shot generalization to NLU tasks in NusaNLU. **EMOT** denotes emotion classification, **AD** denotes abusive detection, **RR** denotes review rating, **SA** denotes sentiment analysis, **HD** denotes hoax detection, **HSD** denotes hate speech detection, **SAG** denotes short answer grading, **SF** denotes stance detection, **LG** denotes legal classification, **NLI** denotes natural language inference, **NTP** denotes next tweet prediction, and **ENT** denotes entailment.

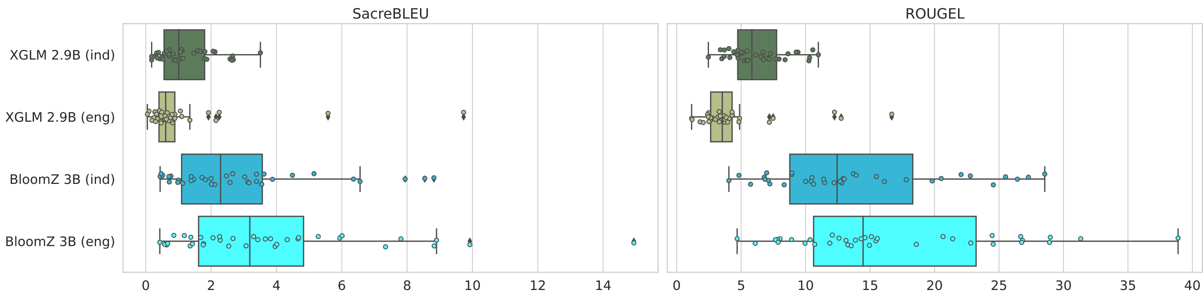

Figure A: Zero-shot generalization to NLG tasks in NusaNLG. Box plots show summary statistics on SacreBLEU **(left)** and ROUGE-L **(right)**. Points are per-dataset scores from the average of performances using 3 different prompts.

Dataset Name	Lang	Task	XGLM 2.9B (id)		XGLM 2.9B (en)		BLOOMZ 3B (id)		BLOOMZ 3B (en)	
			SacreBLEU	ROUGEL	SacreBLEU	ROUGEL	SacreBLEU	ROUGEL	SacreBLEU	ROUGEL
bible_en_id	eng-ind	MT	1.11	8.41	0.49	3.69	3.39	20.52	3.82	22.81
bible_jv_id	jav-ind	MT	0.18	5.13	0.10	1.16	0.47	7.23	0.43	8.02
bible_su_id	sun-ind	MT	0.18	4.95	0.05	1.20	0.44	7.11	0.64	7.81
id_panl_bppt	eng-ind	MT	2.58	10.56	0.48	2.77	6.36	26.43	7.81	28.98
indo_general_mt_en_id	eng-ind	MT	3.51	9.41	1.35	3.40	8.82	25.51	9.92	26.82
indo_religious_mt_en_id	eng-ind	MT	1.80	9.27	0.25	1.81	4.49	22.06	5.94	26.76
minangnlp_mt	min-ind	MT	2.12	7.14	2.25	7.19	5.15	17.81	7.34	26.68
news_en_id	eng-ind	MT	2.66	10.26	0.48	2.62	6.56	24.56	8.83	28.92
nusax_mt_ace_ind	ace-ind	MT	0.72	6.55	0.47	3.19	1.47	11.47	1.77	13.28
nusax_mt_ban_ind	ban-ind	MT	1.09	6.94	0.82	3.68	2.58	12.92	2.54	14.61
nusax_mt_bbc_ind	bbc-ind	MT	1.59	7.94	0.89	4.33	3.03	15.49	3.96	18.59
nusax_mt_bjn_ind	bjn-ind	MT	0.30	4.43	0.28	2.06	0.52	6.80	0.58	7.66
nusax_mt_bug_ind	bug-ind	MT	2.06	10.31	0.19	2.50	3.12	22.78	4.33	24.55
nusax_mt_eng_ind	eng-ind	MT	0.72	4.77	0.66	3.38	1.00	6.84	1.76	10.35
nusax_mt_ind_ace	ind-ace	MT	1.03	5.26	0.77	4.30	1.72	8.34	3.43	15.45
nusax_mt_jav_ind	jav-ind	MT	0.53	5.47	0.30	2.48	1.13	10.00	1.38	10.71
nusax_mt_mad_ind	mad-ind	MT	1.81	8.64	1.06	4.83	3.39	16.11	5.28	20.66
nusax_mt_min_ind	min-ind	MT	0.64	6.14	0.63	3.38	1.39	11.40	1.68	12.07
nusax_mt_nij_ind	nij-ind	MT	0.84	6.73	0.81	4.08	2.01	12.80	2.26	13.54
nusax_mt_sun_ind	sun-ind	MT	0.39	5.00	0.29	2.40	0.98	8.94	1.18	9.96
nusax_mt_ind_ban	ind-ban	MT	1.67	7.19	1.10	4.88	1.83	10.43	3.66	15.10
nusax_mt_ind_bbc	ind-bbc	MT	0.33	2.46	0.36	2.64	0.47	4.05	0.67	4.69
nusax_mt_ind_bjn	ind-bjn	MT	0.57	3.70	2.15	12.76	3.16	19.82	4.66	24.47
nusax_mt_ind_bug	ind-buh	MT	0.86	4.75	0.70	3.73	1.39	8.92	2.28	12.59
nusax_mt_ind_eng	ind-buh	MT	0.44	3.50	0.42	3.00	0.77	5.73	0.86	6.09
nusax_mt_ind_jav	ind-jav	MT	1.87	7.69	0.85	4.30	2.47	10.62	6.01	21.41
nusax_mt_ind_mad	ind-mad	MT	0.71	4.69	0.59	3.31	0.72	6.98	1.36	8.91
nusax_mt_ind_min	ind-min	MT	0.99	5.41	0.76	3.80	2.00	10.50	4.01	15.56
nusax_mt_ind_nij	ind-nij	MT	0.50	3.39	0.47	3.30	0.71	4.83	1.43	7.87
nusax_mt_ind_sun	ind-sun	MT	1.14	7.27	0.59	3.74	2.11	13.00	2.06	13.87
parallel_su_id	sun-ind	MT	0.91	6.70	0.89	3.84	3.88	13.71	4.68	14.95
ted_en_id	eng-ind	MT	2.68	10.99	0.49	2.44	7.94	28.55	8.90	31.34
ud_id_csui	ind-eng	MT	1.06	3.69	5.58	16.68	8.54	27.28	14.94	38.89
indosum_fold0	ind	SUM	2.64	5.53	9.73	12.24	3.55	12.21	3.07	11.88
xl_sum	ind	SUM	0.35	4.05	0.33	3.95	2.66	12.68	3.30	14.31
liputan6	ind	SUM	1.47	4.12	1.92	7.48	3.62	13.92	2.67	13.13
			EM	F1	EM	F1	EM	F1	EM	F1
facqa	ind	QA	0.21	5.61	0.21	4.72	61.95*	81.97*	61.95*	82.02*
idk_mrc	ind	QA	0.71	3.67	0.28	4.31	37.05*	45.51*	34.91*	43.57*
tydiqa_ind	ind	QA	0.31	5.52	0.27	4.67	63.08*	77.28*	62.85*	77.09*

Table A: Details of zero-shot generalization to NLG tasks in NusaNLG. **MT** denotes machine translation, **SUM** denotes summarization, and **QA** denotes question answering. * the test data might be (partially) seen during the language pre-training and prompt tuning phase.

Dataset	Lang	Multilingual Multi-task	Monolingual Multi-task ind	Monolingual Multi-task jav	Monolingual Multi-task sun	Single-task
indspeech_digit_cdsr	ind	2.36	0.22	97.51	97.65	43.84
indspeech_news_lvcsr	ind	13.04	3.31	84.88	81.39	1.41
indspeech_teldialog_lvcsr	ind	1.59	0.37	81.16	77.65	0.45
indspeech_teldialog_svcsr	ind	9.01	1.72	94.58	97.03	0.65
librivox_indonesia_ind	ind	14.24	15.41	86.23	>100	16.20
librivox_indonesia_ace	ace	40.85	95.14	>100	92.36	100
indspeech_newstra_ethnicsr_ban	ban	12.21	>100	56	59.14	43.98
librivox_indonesia_ban	ban	21.24	>100	78.04	>100	100
indspeech_newstra_ethnicsr_btk	btk	18.98	>100	81.24	74.67	64.77
librivox_indonesia_bug	bug	41.59	96.70	>100	>100	100
indspeech_news_ethnicsr_jv	jav	13.37	>100	37.50	93.32	100
indspeech_newstra_ethnicsr_jav	jav	13.30	>100	42.70	72.77	57.34
librivox_indonesia_jav	jav	29.05	81.18	60.70	>100	100
librivox_indonesia_min	min	16.94	46.19	68.10	98.57	100
indspeech_news_ethnicsr_su	sun	20.54	84.08	84.17	50.17	100
indspeech_newstra_ethnicsr_sun	sun	12.41	>100	65.40	47.38	49.03
librivox_indonesia_sun	sun	8.85	60.69	60.12	23.70	100
Average		**17.03**	**>100**	**75.98**	**82.37**	**63.39**

Table A: Per task WER (lower is better) of the **wav2vec 2.0-pt** model on all 17 ASR tasks in NusaASR.

Dataset	Lang	Multilingual Multi-task	Monolingual Multi-task ind	Monolingual Multi-task jav	Monolingual Multi-task sun	Single-task
indspeech_digit_cdsr	ind	0.38	0.18	92.86	97.46	0.22
indspeech_news_lvcsr	ind	0.82	0.65	67.33	80.12	1.37
indspeech_teldialog_lvcsr	ind	0.59	0.09	68.65	74.26	0.22
indspeech_teldialog_svcsr	ind	0.41	0.33	92.92	94.89	0.46
librivox_indonesia_ind	ind	7.32	8.11	72.57	>100	8.37
librivox_indonesia_ace	ace	31.94	91.67	90.28	89.58	49.31
indspeech_newstra_ethnicsr_ban	ban	22.95	>100	36.57	43.69	21.50
librivox_indonesia_ban	ban	19.16	>100	68.69	>100	35.98
indspeech_newstra_ethnicsr_btk	btk	35.99	>100	59.79	61.34	40.92
librivox_indonesia_bug	bug	53.30	>100	>100	>100	90.09
indspeech_news_ethnicsr_jv	jav	22.30	>100	17.90	96.73	27.13
indspeech_newstra_ethnicsr_jav	jav	21.41	>100	22.10	50.14	25.10
librivox_indonesia_jav	jav	38.93	96.49	41.70	>100	44.10
librivox_indonesia_min	min	18.10	70.48	52.86	79.05	24.29
indspeech_news_ethnicsr_su	sun	35.47	87.02	74.13	43.25	44.38
indspeech_newstra_ethnicsr_sun	sun	19.39	>100	39.67	25.71	20.45
librivox_indonesia_sun	sun	7.51	67.63	49.13	6.36	15.03
Average		**19.76**	**>100**	**61.76**	**75.76**	**26.41**

Table A: Per task WER (lower is better) of the **wav2vec 2.0-ft** modes on all 17 ASR tasks in NusaASR.

Dataset	Lang	Multilingual Multi-task	Monolingual Multi-task ind	Monolingual Multi-task jav	Monolingual Multi-task sun	Single-task
indspeech_digit_cdsr	ind	0.27	0.30	43.61	79.08	3.94
indspeech_news_lvcsr	ind	2.56	2.40	59.95	97.42	2.08
indspeech_teldialog_lvcsr	ind	0.38	0.57	58.14	94.44	0.97
indspeech_teldialog_svcsr	ind	50.85	50.91	72.47	87.69	100
librivox_indonesia_ind	ind	87.43	17.32	88.15	99.70	10.13
librivox_indonesia_ace	ace	97.92	99.31	98.61	>100	37.50
indspeech_newstra_ethnicsr_ban	ban	51.89	51.81	57.85	95.56	70.66
librivox_indonesia_ban	ban	88.32	>100	87.38	99.07	39.25
indspeech_newstra_ethnicsr_btk	btk	63.85	56.52	82.16	98.39	>100
librivox_indonesia_bug	bug	100	>100	99.06	>100	56.60
indspeech_news_ethnicsr_jv	jav	58.24	96.31	30.82	>100	53.84
indspeech_newstra_ethnicsr_jav	jav	47.55	49.45	42.77	95.99	73.60
librivox_indonesia_jav	jav	85.71	>100	46.94	>100	51.84
librivox_indonesia_min	min	80.95	82.38	78.57	90.95	36.67
indspeech_news_ethnicsr_su	sun	64.18	97.28	78.37	68.91	67.02
indspeech_newstra_ethnicsr_sun	sun	54.65	53.51	66.26	73.02	72.71
librivox_indonesia_sun	sun	77.46	88.44	68.21	41.62	27.17
Average		**59.54**	**62.83**	**68.20**	**89.98**	**49.44**

Table A: Per task WER (lower is better) of the **Whisper** model on all 17 ASR tasks in NusaASR.

Dataset	Task	Modal	Languages
Korpus Nusantara (Indrayana, 2016; Hasbiansyah et al., 2016; Ningtyas et al., 2018; Etsa et al., 2018; Darwis et al., 2019; Wahyuni et al., 2019; Sujaini, 2019, 2020; Gunawan et al., 2021)	MT	Text	ind, jav, xdy, bug, sun, mad, bjn, bbc, hak, msa, min, nan
Karonese dataset (Karo et al., 2022)	SA	Text	btx
Sundanese-Indonesian Parallel Corpus (Ardiyanti Suryani et al., 2022b)	MT	Text	sun, ind
PoSTagged Sundanese Monolingual Corpus (Ardiyanti Suryani et al., 2022a)	POS	Text	sun
Code-mixed Sentiment JV-ID (Tho et al., 2021)	SA	Text	ind
ID-HSD-Nofaaulia (Aulia and Budi, 2019)	HSD	Text	ind
Indo Wiki Paralel Corpora (Trisedya and Inastra, 2014)	MT	Text	ind, jav, sun, min
INDspeech_DIGIT_CDSR (Sakti et al., 2004)	ASR	Speech	ind
INDspeech_NEWS_EthnicSR (Sani et al., 2012)	ASR	Speech	sun, jav
INDspeech_NEWS_LVCSR (Sakti et al., 2008a, 2004, 2013)	ASR	Speech	ind
INDspeech_NEWS_TTS (Sakti et al., 2008b, 2010, 2013)	TTS	Speech	ind
INDspeech_NEWSTRA_EthnicSR (Sakti and Nakamura, 2013, 2014; Novitasari et al., 2020)	ASR	Speech	sun, jav, btk, ban
INDspeech_TELDIALOG_LVCSR (Sakti et al., 2008a, 2004, 2013)	ASR	Speech	ind
INDspeech_TELDIALOG_SVCSR (Sakti et al., 2004)	ASR	Speech	ind

Table A: List of private datasets that have been made public through NusaCrowd initiative. **MT** denotes machine translation, **SA** denotes sentiment analysis, **POS** denotes POS tagging, **HSD** denotes hate speech detection, **ASR** denotes automatic speech recognition, and **TTS** denotes text-to-speech.

Project	#Dataset	Languages	Modality	Open/crowd initiative?	Opening access to private data
		Global Data Initiative			
XTREME (Hu et al., 2020)	9	40	text	✗	✗
XGLUE (Liang et al., 2020)	11	19	text	✗	✗
GEM (Gehrmann et al., 2021)	11	18	text	✓	✗
GEMv2 (Gehrmann et al., 2022)	40	51	text	✓	✗
		Other Regional Data Initiative			
CLUE (Xu et al., 2020)	9	1 (zho)	text	✗	✗
KLUE (Park et al., 2021)	8	1 (kor)	text	✗	✗
ALUE (Seelawi et al., 2021)	9	1 (ara)	text	✗	✗
IndicGLUE (Kakwani et al., 2020)	14	12 Indian languages	text	✗	✗
IndicNLG (Kumar et al., 2022)	5	11 Indian languages	text	✗	✗
IndicXTREME (Doddapaneni et al., 2022)	103	18 Indian languages	text	✗	✗
		Indonesian Data Initiative			
IndoNLU (Wilie et al., 2020)	12	1 (ind)	text	✗	✗
IndoLEM (Koto et al., 2020b)	12	1 (ind)	text	✗	✗
IndoNLG (Cahyawijaya et al., 2021b)	10	3 (ind, sun, jav)	text	✗	✗
NusaCrowd	137	19 Indonesian languages	text, speech, image	✓	✓

Table A: Comparison of NusaCrowd with other similar initiatives.

Dataset	Description	License	Year	Language	Annotation Quality	Data Volume	Data Unit	Split
AM2iCo (Liu et al., 2021b)	We present AM2ICO (Adversarial and Multilingual Meaning in Context), a wide coverage cross-lingual and multilingual evaluation set; it aims to faithfully assess the ability of state-of-the-art (SotA) representation models to understand the identity of word meaning in cross-lingual contexts for 14 language pairs.	CC-BY 4.0	2021	ind, eng	Crawling & human annotation	3,098	examples	1589 train, 500 validation, 1000 test
Barasa	Barasa: Indonesian SentiWordNet for sentiment analysis	MIT	2015	ind	Unknown	16	MB	No dataset split
CASA (Ilmania et al., 2018)	CASA is an aspect-based sentiment analysis dataset consisting of around a thousand car reviews collected from multiple Indonesian online automobile platforms. The dataset covers six aspects of car quality, where each label represents a sentiment for a single aspect with three possible values: positive, negative, and neutral.	CC-BY-SA 4.0	2018	ind	Crawling & human annotation	1,080	sentences	810 train, 90 validation, 180 test
CC100 (Conneau et al., 2020)	CC100 comprises of monolingual data for 100+ languages and also includes data for romanized languages. This was constructed using the urls and paragraph indices provided by the CC-Net repository by processing January-December 2018 Commoncrawl snapshots. Each file comprises of documents separated by double-newlines and paragraphs within the same document separated by a newline. The data is generated using the open source CC-Net repository.	Common Crawl's license	2020	ind, sun, jav	Machine generated / Crawling w/o curation	36,052	MB	No dataset split
COCO Captions ID (Sinurat, 2019)	COCO Captions contains over one and a half million captions describing over 330,000 images. For the training and validation images, five independent human generated captions are be provided for each image. This is an Indonesian version of COCO translated using Google Translate.	CC-BY 4.0	2019	ind	Machine generated / Crawling w/o curation	123,287	sentences	113287 train, 5000 validation, 5000 test

Continued on the next page

Table A – continued from the previous page

Dataset	Description	License	Year	Language	Annotation Quality	Data Volume	Data Unit	Split
CORD (Park et al., 2019)	In this paper, we introduce a novel dataset called CORD, which stands for a Consolidated Receipt Dataset for post-OCR parsing. To the best of our knowledge, this is the first publicly available dataset which includes both box-level text and parsing class annotations. The parsing class labels are provided in two-levels. The eight superclasses include store, payment, menu, subtotal, and total. The eight superclasses are subdivided into 54 subclasses, e.g., store has nine subclasses including name, address, telephone, and fax. Furthermore, it also provides line annotations for the serialization task which is a newly emerging problem as a combination of the two tasks.	CC-BY 4.0	2019	ind	Crawling & human annotation	1,000	receipts	800 train, 100 validation, 100 test
CVSS (Jia et al., 2022)	We introduce CVSS, a massively multilingual-to-English speech-to-speech translation (S2ST) corpus, covering sentence-level parallel S2ST pairs from 21 languages into English. CVSS is derived from the Common Voice (Ardila et al., 2020) speech corpus and the CoVoST 2 (Wang et al., 2021b) speech-to-text translation (ST) corpus, by synthesizing the translation text from CoVoST 2 into speech using state-of-the-art TTS systems. Two versions of translation speech in English are provided: 1) CVSS-C: All the translation speech is in a single high-quality canonical voice; 2) CVSS-T: The translation speech is in voices transferred from the corresponding source speech. In addition, CVSS provides normalized translation text which matches the pronunciation in the translation speech.	CC-BY 4.0	2022	ind	Crawling & human annotation	6	hours	2.6 train, 1.8 validation, 1.9 test
Cendana (Moeljadi et al., 2019)	Cendana is a linguistically annotated corpus that includes some grammatical analyses, such as parts-of-speech, phrases, relations between entities, and meaning representations. Cendana is built using tools developed in the Deep Linguistic Processing with HPSG (DELPHIN) community.	GNU General Public License, version 2	2019	ind	Human generation & curation	552	sentences	No dataset split

Continued on the next page

Table A – continued from the previous page

Dataset	Description	License	Year	Language	Annotation Quality	Data Volume	Data Unit	Split
CoVoST 2 (Wang et al., 2021)	With the aim to foster research in massive multilingual ST and ST for low resource language pairs, we release CoVoST 2, a large-scale multilingual ST corpus covering translations from 21 languages into English and from English into 15 languages. This represents the largest open dataset available to date from total volume and language coverage perspective.	CC0	2020	ind, eng	Crawling & human annotation	3	hours	1 train, 1 validation, 1 test
Code-mixed Sentiment JV-ID (Tho et al., 2021)	Dataset terdiri dari 3.963 kalimat code-mixing dalam bahasa Indonesia dan bahasa Jawa yang dikumpulkan dari twitter. Label dataset terdiri dari 3 kelas sentimen, yaitu: positif, negatif, dan netral. Label sentimen dikumpulkan dengan melakukan anotasi manual untuk setiap tweet.	CC-BY 3.0	2021	ind	Human generation & curation	977	tweets	No dataset split
Cross-lingual Outline-based Dialogue (COD) (Majewska et al., 2023)	Cross-lingual Outline-based Dialogue dataset (termed COD) enables natural language understanding, dialogue state tracking, and end-to-end dialogue modelling and evaluation in 4 diverse languages: Arabic, Indonesian, Russian, and Kiswahili. The data covers multi domain instances, e.g., bank, travel, weather, movies, music.	Unknown	2022	ind	Machine generated w/ human curation	194	dialogues	0 train, 92 validation, 102 test
Customer Review (Natasha Skincare) (Nurlaila et al., 2017)	This dataset is collected from tweet costumer review from Natasha Skincare. This dataset contain of label emotion (joy, sad, angry, fear, disgust, surprise, or no emotion).	Unknown	2017	ind	Crawling & human annotation	124,263	tweets	87120 train, 37143 validation, 37143 test
EmoT (IndoNLU Split) (Saputri et al., 2018)	EmoT is an emotion classification dataset collected from the social media platform Twitter. The dataset consists of around 4000 Indonesian colloquial language tweets, covering five different emotion labels: anger, fear, appiness, love, and sadness.	CC-BY-SA 4.0	2018	ind	Human generation & curation	4,403	sentence pairs	3521 train, 440 validation, 442 test

Continued on the next page

Table A – continued from the previous page

Dataset	Description	License	Year	Language	Annotation Quality	Data Volume	Data Unit	Split
EmotCMT (Yulianti et al., 2021)	EmotCMT is a Indonesian-English code-switching data collected from Twitter for emotion classification task.	Unknown	2021	ind, eng	Crawling & Human annotation	582	sentences	No dataset split
Emotion Indonesian Public Opinion (Riccosan et al., 2022)	The dataset is formed from Indonesian tweet containing six emotion values, namely anger, fear, joy, love, sad, and neutral. The total data in this dataset is 7,080 and it is fully cleaned and fully annotated. Each label has a varied amount of data distribution, including 1,130 data for anger, 911 data for fear, 1,275 data for joy, 760 data for love, 1,003 data for sad, and 2,001 data for neutral.	CC-BY-NC-ND 4.0	2022	ind	Human generation & curation	7,080	tweets	No dataset split
FacQA (Purwarianti et al., 2007)	The goal of the FacQA dataset is to find the answer to a question from a provided short passage from a news article . Each row in the FacQA dataset consists of a question, a short passage, and a label phrase, which can be found inside the corresponding short passage. There are six categories of questions: date, location, name, organization, person, and quantitative.	CC-BY-SA 4.0	2007	ind	Human generation & curation	3,117	documents	2495 train, 311 dev, 311 test
HoASA (IndoNLU Split) (Azhar et al., 2019)	HoASA is an aspect-based sentiment analysis dataset consisting of hotel reviews collected from the hotel aggregator platform, AiryRooms	CC-BY-SA 4.0	2019	ind	Crawling & human annotation	9,450	sentences	7,560 train, 1890 test
Human Instructions - Indonesian (wikihow) (Chocron and Pareti, 2018)	Human Instructions - Indonesian (wikihow) is 39.246 Human Instructions in Indonesian Extracted from wikiHow. Step-by-step instructions in Indonesian extracted from wikiHow and decomposed into a formal graph representation in RDF. Instructions are represented in RDF following the PROHOW vocabulary and data model. For example, the category, steps, requirements and methods of each set of instructions have been extracted. This dataset has been produced as part of the The Web of Know-How project.	CC-BY-NC-SA 4.0	2017	ind	Crawling & human annotation	39,246	documents	No dataset split

Continued on the next page

Table A – continued from the previous page

Dataset	Description	License	Year	Language	Annotation Quality	Data Volume	Data Unit	Split
ID Abusive (Ibrohim and Budi, 2018)	ID Abusive is a Twitter dataset for abusive language detection in Indonesian. Pre-defined abusive words are used as queries to collect the tweets. The dataset is labeled into 3 labels: not abusive language, abusive but not offensive, and offensive language by 20 volunteer annotators.	CC-BY-NC-SA 4.0	2018	ind	Crawling & human annotation	2,016	tweets	No dataset split
ID Abusive Online News Comment (Kiasati Desrul and Romadhony, 2019)	The dataset consists of comments that are in some of the top news stories in 2019, obtained from several online news/forum, such as: kompas, kaskus, and detik. The labeling process is carried out by a total of 10 annotators and each comment is annotated by 3 annotators. Each comment was labeled with one of the following labels: not abusive, abusive but not offensive, abusive and offensive.	Unknown	2019	ind	Machine generated / Crawling w/o curation	3,184	comments	No dataset split
ID Coreference Resolution (Suherik and Purwarianti, 2017)	ID Coreference resolution is news dataset aimed for coreference resolution task. This dataset consists of 1030 manually labelled sentences derived from IDENTIC parallel corpus.	Unknown	2017	ind	Crawling & human annotation	1,030	sentences	759 train, 0 validation, 108 test

Continued on the next page

Table A – continued from the previous page

Dataset	Description	License	Year	Language	Annotation Quality	Data Volume	Data Unit	Split
ID Multilabel HS (Ibrohim and Budi, 2019)	ID Multilabel HS consists of hate speech and abusive language Twitter dataset from several previous researches consisting of (Alfina et al., 2017, 2018), (Putri, 2018), and (Ibrohim and Budi, 2018), and new tweets queried using specific abusive words/phrases. Labels used in the dataset are: HS (hate speech label), Abusive (abusive language label), HS_Individual (hate speech targeted to an individual), HS_Group (hate speech targeted to a group), HS_Religion (hate speech related to religion/creed), HS_Race (hate speech related to race/ethnicity), HS_Physical (hate speech related to physical/disability), HS_Gender (hate speech related to gender/sexual orientation), HS_Physical (hate related to other invective/slander), HS_Weak (weak hate speech), HS_Moderate (moderate hate speech), HS_Strong (strong hate speech).	CC-BY-NC-SA 4.0	2019	ind	Crawling & human annotation	13,169	tweets	No dataset split
ID Quora Question Pairs	Quora Question Pairs (QQP) dataset consists of over 400,000 question pairs, and each question pair is annotated with a binary value indicating whether the two questions are paraphrase of each other. This dataset is translated version of QQP to Indonesian Language.	Custom[20]	2021	ind	Machine generated / Crawling w/o curation	149,011	sentence pairs	134084 train, 14927 validation, 0 test

Continued on the next page

[20] https://www.quora.com/about/tos

Table A – continued from the previous page

Dataset	Description	License	Year	Language	Annotation Quality	Data Volume	Data Unit	Split
ID Short Answer Grading (Haidir and Purwarianti, 2020)	ID Short Answer Grading dataset is dataset of questions using Edukasystem platform. It used 4 exams consisting of Biology and Geography subject. Two exams are used for training data and 2 exams are used for testing data. Exam used for training data has 15 questions and exam for testing data has 3 questions. The dataset has 30 questions + 7605 short answers as training data and 6 questions + 1560 short answers as testing data. The number of respondents is 534 different respondents. Assessment was carried out by 2 experts for Biology subject and 5 experts for Geography subject. The assessment is carried out on a scale of 0 to 5.	Unknown	2020	ind	Human generation & curation	9,165	sentences	7605 train, 0 validation, 1560 test
ID-HSD-Nofaaulia (Aulia and Budi, 2019)	There have been many studies on detecting hate speech in short documents like Twitter data. But to our knowledge, research on long documents is rare, we suppose that the difficulty is increasing due to the possibility of the message of the text may be hidden. In this research, we explore in detecting hate speech on Indonesian long documents using machine learning approach. We build a new Indonesian hate speech dataset from Facebook.	Unknown	2022	ind	Crawling & human annotation	906	documents	815 train, 0 validation, 91 test
ID-HSD-Riomulia (Alfina et al., 2017a)	ID-HSD-RioMulia composed of tweets about the Jakarta Governor Election 2017, whose selection of candidates potentially triggers hate speech in relation to race, religion, and gender. Each tweet is labelled as either containing hate speech or not by 30 human annotators.	GNU General Public License v3.0	2017	ind	Crawling & human annotation	713	tweets	No dataset split

Continued on the next page

Table A – continued from the previous page

Dataset	Description	License	Year	Language	Annotation Quality	Data Volume	Data Unit	Split
IDK-MRC (Putri and Oh, 2022)	IDK-MRC is an Indonesian Machine Reading Comprehension dataset that covers answerable and unanswerable questions. Based on the combination of the existing answerable questions in TyDiQA, the new unanswerable question in IDK-MRC is generated using question generation model and human-written question. Each paragraph in the dataset have a set of answerable and unanswerable question with the corresponding answer. (Note: the paper for this dataset is still under review on EMNLP 2022 – the anonymity period ends on Oct 6, 2022)	CC-BY-SA 4.0	2022	ind	Machine generated w/ human curation	10,940	paragraph, question, and answer pairs	9332 train, 764 validation, 844 test
IMDb Javanese (Wongso et al., 2021)	Large Movie Review Dataset translated to Javanese. This is a dataset for binary sentiment classification containing substantially more data than previous benchmark datasets. We provide a set of 25,000 highly polar movie reviews for training, and 25,000 for testing.	Unknown	2021	jav	Machine generated / Crawling w/o curation	50,000	sentences	25000 train, 0 validation, 25000 test
INDspeech_ DIGIT_ CDSR (Sakti et al., 2004)	INDspeech_DIGIT_CDSR is the first Indonesian speech dataset for connected digit speech recognition (CDSR). The data was developed by TELKOMRisTI (R&D Division, PT Telekomunikasi Indonesia) in collaboration with Advanced Telecommunication Research Institute International (ATR) Japan and Bandung Institute of Technology (ITB) under the Asia-Pacific Telecommunity (APT) project in 2004 [Sakti et al., 2004]. Although it was originally developed for a telecommunication system for hearing and speaking-impaired people, it can be used for other applications, i.e., automatic call centers that recognize telephone numbers.	CC-BY-NC-SA 4.0	2004	ind	Human generation & curation	12444 [214]	utterances [speakers]	8440 train, 0 validation, 4004 test

Continued on the next page

Table A – continued from the previous page

Dataset	Description	License	Year	Language	Annotation Quality	Data Volume	Data Unit	Split
INDspeech_ NEWSTRA_ EthnicSR (Sakti and Nakamura, 2013)	INDspeech_NEWSTRA_EthnicSR is a collection of graphemically balanced and parallel speech corpora of four major Indonesian ethnic languages: Javanese, Sundanese, Balinese, and Bataks. It was developed in 2013 by the Nara Institute of Science and Technology (NAIST, Japan) [Sakti et al., 2013]. The data has been used to develop Indonesian ethnic speech recognition in supervised learning [Sakti et al., 2014] and semi-supervised learning [Novitasari et al., 2020] based on the Machine Speech Chain framework [Tjandra et al., 2020].	CC-BY-NC-SA 4.0	2013	ind, sun, jav, ban, btk	Human generation & curation	13000 [40]	utterances [speakers]	9000 train, 0 validation, 4000 test
INDspeech_ NEWS_ EthnicSR (Sani et al., 2012)	INDspeech_NEWS_EthnicSR is a collection of Indonesian ethnic speech corpora (Javanese and Sundanese) for Indonesian ethnic speech recognition. It was developed in 2012 by the Nara Institute of Science and Technology (NAIST, Japan) in collaboration with the Bandung Institute of Technology (ITB, Indonesia) [Sani et al., 2012]. Furthermore, as all speakers utter the same sentences, it can also be used for voice conversion tasks.	CC-BY-NC-SA 4.0	2012	sun, jav	Human generation & curation	2300 [20]	utterances [speakers]	2000 train, 0 validation, 300 test
INDspeech_ NEWS_ LVCSR (Sakti et al., 2008a)	INDspeech_NEWS_LVCSR is the first Indonesian speech dataset for large vocabulary continuous speech recognition (LVCSR) with more than 40 hours of speech and 400 speakers [Sakti et al., 2008]. R&D Division of PT Telekomunikasi Indonesia (TELKOMRisTI) developed the data in 2005-2006, in collaboration with Advanced Telecommunication Research Institute International (ATR) Japan, as the continuation of the Asia-Pacific Telecommunity (APT) project [Sakti et al., 2004]. It has also been successfully used for developing Indonesian LVCSR in the Asian speech translation advanced research (A-STAR) project [Sakti et al., 2013].	CC-BY-NC-SA 4.0	2008	ind	Human generation & curation	44000 [400]	utterances [speakers]	39600 train, 0 validation, 4400 test

Continued on the next page

Table A – continued from the previous page

Dataset	Description	License	Year	Language	Annotation Quality	Data Volume	Data Unit	Split
INDspeech_ NEWS_ TTS (Sakti et al., 2008b)	INDspeech_NEWS_TTS is a speech dataset for developing an Indonesian text-to-speech synthesis system [Sakti et al., 2008, Sakti et al., 2010]. The data was developed by Advanced Telecommunication Research Institute International (ATR) Japan under the Asian speech translation advanced research (A-STAR) project [Sakti et al., 2013].	CC-BY-NC-SA 4.0	2008	ind	Human generation & curation	2,012	utterances	1972 train, 0 validation, 40 test
INDspeech_ TELDIALOG_ LVCSR (Sakti et al., 2008a)	INDspeech_TELDIALOG_LVCSR is one of the first Indonesian speech datasets for large vocabulary continuous speech recognition (LVCSR) [Sakti et al., 2008]. R&D Division of PT Telekomunikasi Indonesia (TELKOMRisTI) developed the data in 2005-2006, in collaboration with Advanced Telecommunication Research Institute International (ATR) Japan, as the continuation of the Asia-Pacific Telecommunity (APT) project [Sakti et al., 2004]. It has also been successfully used for developing Indonesian LVCSR in the Asian speech translation advanced research (A-STAR) project [Sakti et al., 2013].	CC-BY-NC-SA 4.0	2008	ind	Human generation & curation	40000 [400]	utterances [speakers]	36000 train, 0 validation, 4000 test
INDspeech_ TELDIALOG_ SVCSR (Sakti et al., 2004)	INDspeech_TELDIALOG_SVCSR is the first Indonesian speech dataset for small vocabulary continuous speech recognition (SVCSR). The data was developed by TELKOMRisTI (R&D Division, PT Telekomunikasi Indonesia) in collaboration with Advanced Telecommunication Research Institute International (ATR) Japan and Bandung Institute of Technology (ITB) under the Asia-Pacific Telecommunity (APT) project in 2004 [Sakti et al., 2004]. Although it was originally developed for a telecommunication system for hearing and speaking impaired people, it can be used for other applications, i.e., automatic call centers. Furthermore, as all speakers utter the same sentences, it can also be used for voice conversion tasks.	CC-BY-NC-SA 4.0	2004	ind	Human generation & curation	20000 [200]	utterances [speakers]	10000 train, 0 validation, 10000 test

Continued on the next page

Table A – continued from the previous page

Dataset	Description	License	Year	Language	Annotation Quality	Data Volume	Data Unit	Split
Identic (Larasati, 2012) (Larasati, 2012)	IDENTIC is an Indonesian-English parallel corpus for research purposes. The corpus is a bilingual corpus paired with English. The aim of this work is to build and provide researchers a proper Indonesian-English textual dataset and also to promote research in this language pair. The corpus contains texts coming from different sources with different genres.	Unknown	2012	eng, ind	Crawling & human annotation	45,000	sentences	No dataset split
Idn-tagged-corpus-CSUI (Dinakaramani et al., 2014)	Id-tagged-corpus-CSUI is a POS tagging dataset contains about 10,000 sentences, collected from the PAN Localization Project tagged with 23 POS tag classes.	CC-BY-SA 4.0	2014	ind	Crawling & human annotation	10,000	sentences	8000 train, 1000 validation, 1029 test
InSet Lexicon (Koto and Rahmaningtyas, 2017)	Sentiment analysis from Twitter	Unknown	2017	ind	Crawling & human annotation	2,630	sentences	2630 test
IndQNER	IndQNER is a NER dataset created by manually annotating 8 chapters in the Indonesian translation of Quran text. The dataset consists of 2476 named entities from 18 categories. Each named entity is labeled using BIO (Beginning-Inside-Outside) tagging format.	Unknown	2022	ind	Human generation & curation	3,118	sentences	2494 train, 312 validation, 312 test
Indo4B (Wilie et al., 2020)	Indo4B is an Indonesian pre-training corpus collected from multiple online sources, Indo4B consists of 3.6B tokens and over more than 250M sentences. Indo4B has been used to pre-trained a large pre-trained language model called IndoBERT and IndoBERT-lite.	CC-BY-NC-SA 4.0	2020	ind	Machine generated / Crawling w/o curation	3.6B	tokens	No dataset split
Indo4B Plus (Cahyawijaya et al., 2021b)	Indo4BPlus is an Indonesian pre-training corpus derived from Indo4B. Indo4BPlus covers three languages in Indonesia, i.e., Indonesian, Javanese, and Sundanese. Indo4BPlus consists of 4B tokens with over more than 300M documents.	CC-BY-NC-SA 4.0	2021	ind, sun, jav	Machine generated / Crawling w/o curation	4B	tokens	No dataset split

Continued on the next page

Table A – continued from the previous page

Dataset	Description	License	Year	Language	Annotation Quality	Data Volume	Data Unit	Split
IndoAMR (Ilmy and Khodra, 2020)	IndoAMR is annotated Indonesia AMR parser from Indonesian simple sentences.	Unknown	2021	ind	Crawling & human annotation	1,130	sentences	700 train, 0 validation, 300 test
IndoCollex (Wibowo et al., 2021)	List of Colloquial word Transformation with its label. e.g.: makan -> mkn (shortening). Data is published on IndoCollex: A Testbed for Morphological Transformation of Indonesian Word Colloquialism Research Paper published on ACL-IJCNLP 2021. Useful for morphological research.	MIT	2021	ind	Human generation & curation	2,126	sentence pairs	1637 train, 182 validation, 193 test
IndoCoref (Artari et al., 2021)	IndoCoref is a coreference resolution dataset collected from Wikipedia. IndoCoref consists of 201 passages from wikipedia with manually labelled coreference by five annotators.	MIT	2021	ind	Human generation & curation	201	documents	No dataset split
IndoLEM NTP (Koto et al., 2020b)	IndoLEM next tweet prediction is a next tweet prediction dataset collected from tweeter	CC BY-SA 3.0	2020	ind	Crawling & human annotation	8,382	instances	5681 train, 811 validation, 1890 test
IndoLEM Sentiment Analysis (Koto et al., 2020b)	IndoLEM Sentiment Analysis is a textual sentiment analysis dataset collected from twitter	CC BY-SA 3.0	2020	ind	Crawling & human annotation	5,048	sentences	3638 train, 399 validation, 1011 test
IndoLEM Tweet Ordering (Koto et al., 2020b)	IndoLEM tweet ordering is a text ordering dataset collected from tweeter	CC BY-SA 3.0	2020	ind	Crawling & human annotation	7,608	instances	5327 train, 760 validation, 1521 test

Continued on the next page

Table A – continued from the previous page

Dataset	Description	License	Year	Language	Annotation Quality	Data Volume	Data Unit	Split
IndoNLG Bible En-Id (Cahyawijaya et al., 2021b)	Bible En-Id is a machine translation dataset containing Indonesian-English parallel sentences collected from the bible. We also add a Bible dataset to the English Indonesian translation task. Specifically, we collect an Indonesian and an English language Bible and generate a verse-aligned parallel corpus for the English-Indonesian machine translation task. We split the dataset and use 75% as the training set, 10% as the validation set, and 15% as the test set. Each of the datasets is evaluated in both directions, i.e., English to Indonesian (En $\rightarrow$ Id) and Indonesian to English (Id $\rightarrow$ En) translations.	CC-BY-SA 4.0	2021	eng, ind	Crawling & human annotation	31,078	sentences	23308 training, 3109 validation, 4661 test
IndoNLG Bible Jv-Id (Cahyawijaya et al., 2021b)	Bible Jv-Id is a machine translation dataset containing Indonesian-Javanese parallel sentences collected from the bible. Analogous to the En $\leftrightarrow$ Id and Su $\leftrightarrow$ Id datasets, we create a new dataset for Javanese and Indonesian translation generated from the verse-aligned Bible parallel corpus with the same split setting. In terms of size, both the Su $\leftrightarrow$ Id and Jv $\leftrightarrow$ Id datasets are much smaller compared to the En $\leftrightarrow$ Id dataset, because there are Bible chapters for which translations are available for Indonesian, albeit not for the local languages.	CC-BY-SA 4.0	2021	jav, ind	Crawling & human annotation	7,957	sentences	5967 train, 797 validation, 1193 test
IndoNLG Bible Su-Id (Cahyawijaya et al., 2021b)	Bible Su-Id is a machine translation dataset containing Sundanese Indonesian parallel sentences collected from the bible. Analogous to the En $\leftrightarrow$ Id and Su $\leftrightarrow$ Id datasets, we create a new dataset for Javanese and Indonesian translation generated from the verse-aligned Bible parallel corpus with the same split setting. In terms of size, both the Su $\leftrightarrow$ Id and Jv $\leftrightarrow$ Id datasets are much smaller compared to the En $\leftrightarrow$ Id dataset, because there are Bible chapters for which translations are available for Indonesian, albeit not for the local languages	CC-BY-SA 4.0	2021	sun, ind	Crawling & human annotation	7,958	sentences	5968 train, 797 validation, 1193 test

Continued on the next page

Table A – continued from the previous page

Dataset	Description	License	Year	Language	Annotation Quality	Data Volume	Data Unit	Split
IndoNLI (Mahendra et al., 2021)	IndoNLI is the first human-elicited Natural Language Inference (NLI) dataset for Indonesian. IndoNLI is annotated by both crowd workers and experts. The expert-annotated data is used exclusively as a test set. It is designed to provide a challenging test-bed for Indonesian NLI by explicitly incorporating various linguistic phenomena such as numerical reasoning, structural changes, idioms, or temporal and spatial reasoning.	CC-BY-SA 4.0	2021	ind	Human generation & curation	17,712	sentence pairs	10330 train, 2197 validation, 5183 test
IndoNLU NERGrit (Wilie et al., 2020)	NER Grit dataset is a NER dataset taken from the Grit-ID repository, and the labels are spans in IOB chunking representation. The dataset consists of three kinds of named entity tags, PERSON (name of person), PLACE (name of location), and ORGANIZATION (name of organization).	CC-BY-SA 4.0	2020	ind	Unknown	2,090	sentences	1672 train, 209 validation, 209 test
IndoPuisi	Puisi is an Indonesian poetic form. The dataset contains 7223 Indonesian puisi (poem) with its title and author. The data was scraped online using BeautifulSoup. The title and author column was produced using regex.	MIT	2020	ind	Machine generated / Crawling w/o curation	7,223	documents	No dataset split
IndoSum (Kurniawan and Louvan, 2018)	The Indosum dataset was collected from news aggregators covering six topics: entertainment, inspiration, sport, showbiz, headline, and technology. Compared to Liputan6, the summary label of Indosum is less abstractive, with novel 1-gram and novel 4-gram rates of 3.1% and 20.3%, respectively (Koto et al., 2020a).	CC-BY-SA 4.0	2021	ind	Crawling & human annotation	18,773	sentences	14083 train, 1880 validation, 2810 test
IndoTacos	IndoTacos dataset is tax court verdict summary collected from perpajakan.ddtc,co.id. It contains 12k tax court summary with its verdict: mengabulkan seluruhnya, mengabulkan sebagian, menolak, mengabulkan, lain-lain. This legal document is spesific for Indonesia tax cases.	CC-BY-NC-SA 4.0	2021	ind	Machine generated / Crawling w/o curation	12,291	documents	No dataset split

Continued on the next page

Table A – continued from the previous page

Dataset	Description	License	Year	Language	Annotation Quality	Data Volume	Data Unit	Split
Indo_ MultiModal_ CC_ 12M (Changpinyo et al., 2021)	Conceptual 12M (CC12M) is a dataset with 12 million image-text pairs specifically meant to be used for visionand-language pre-training. Its data collection pipeline is a relaxed version of the one used in Conceptual Captions 3M (CC3M). Indo_MultiModal_CC_ 12M is the Indonesian language version.	Custom[21]	2022	ind	Machine generated / Crawling w/o curation	1	GB	No dataset split
Indo_ MultiModal_ LAION (Schuhmann et al., 2021)	Indo_MultiModal_LAION is a translated subset of the LAION-400M dataset with 70M image-text pairs specifically meant to be used for vision-language pre-training in Indonesian language. LAION-400M is a dataset with 400M English (image, text) pairs, filtered with OpenAI's CLIP by calculating the cosine similarity between the text and image embeddings and dropping those with a similarity below 0.3. The threshold of 0.3 had been determined through human evaluations and seemed to be a good heuristic for estimating semantic image-text-content matching. The image-text-pairs have been extracted from the Common Crawl web data dump and are from random web pages crawled between 2014 and 2021. More info for LAION-400M: https://laion.ai/blog/laion-400-open-dataset/.	CC-BY 4.0	2022	ind	Machine generated / Crawling w/o curation	7	GB	No dataset split
Indo_ MultiModal_ PMD_ ID (Singh et al., 2022)	Introduced in the FLAVA paper, Public Multimodal Dataset (PMD) is a collection of publicly-available image-text pair datasets. PMD contains 70M image-text pairs in total with 68M unique images. The dataset contains pairs from Conceptual Captions, Conceptual Captions 12M, WIT, Localized Narratives, RedCaps, COCO, SBU Captions, Visual Genome and a subset of YFCC100M dataset. Indo_MultiModal_PMD_ Indonesia is the Indonesian language version.	CC-BY-4.0	2022	ind	Machine generated / Crawling w/o curation	15	GB	0 train, 0 validation, 0 test

Continued on the next page

[21] https://github.com/google-research-datasets/conceptual-12m/blob/main/LICENSE

Dataset	Description	License	Year	Language	Annotation Quality	Data Volume	Data Unit	Split
Indonesian Clickbait (William and Sari, 2020)	The CLICK-ID dataset is a collection of Indonesian news headlines that was collected from 12 local online news publishers; detikNews, Fimela, Kapanlagi, Kompas, Liputan6, Okezone, Posmetro-Medan, Republika, Sindonews, Tempo, Tribunnews, and Wowkeren. This dataset is comprised of mainly two parts; (i) 46,119 raw article data, and (ii) 15,000 clickbait annotated sample headlines. Annotation was conducted with 3 annotator examining each headline. Judgment were based only on the headline. The majority then is considered as the ground truth. In the annotated sample, our annotation shows 6,290 clickbait and 8,710 non-clickbait.	CC-BY 4.0	2020	ind	Crawling & human annotation	15,000	headlines	No dataset split
Indonesian Frog Storytelling Corpus (Moeljadi, 2012)	Indonesian written and spoken storytelling corpus, based on the twenty-eight pictures.	Unknown	2014	ind	Unknown	0	documents	No dataset split
Indonesian Google Play Review	Indonesian Google Play Review, dataset scrapped from e-commerce app on Google Play for sentiment analysis.	CC-BY 4.0	2022	ind	Machine generated / Crawling w/o curation	10,041	sentences	7028 train, 3012 validation, 0 test
Indonesian Hoax News Detection (Pratiwi et al., 2017)	Indonesian Hoax News Detection is a dataset for hoax news detection. 600 Data are retrieved in Indonesian language with 372 valid news and 228 fake news.All data are manually labelled.	CC-BY 4.0	2018	ind	Human generation & curation	600	documents	No dataset split
Indonesian Poem Tweets	Indonesian Poem tweets is dataset crawled from Twitter. The purpose of this data is to create text generation model for short text and make sure they are all coherence and rhythmic	CC-BY 4.0	2022	ind	Machine generated / Crawling w/o curation	16,427	tweets	No dataset split

Continued on the next page

Table A – continued from the previous page

Dataset	Description	License	Year	Language	Annotation Quality	Data Volume	Data Unit	Split
Indonesian Stance (Jannati et al., 2018)	ID Stance is a collection of Kompasiana articles that match with a pre-defined list of Indonesian politician names. Each article possesses a stance towards a candidate entity and election event, determined by annotators. Since the task is framed as a binary classification task, articles with no stance (neutral) are excluded from the gold-standard set.	CC-BY-NC-SA 4.0	2018	ind	Crawling & human annotation	337	documents	No dataset split
Indonesian WSD (Mahendra et al., 2018)	Indonesian WSD is a word sense dissambiguation dataset automatically collected using CrossLingual WSD (CLWSD) approach by utilizing WordNet and parallel corpus GIZA++. The monolingual WSD model is built from training data and it is used to assign the correct sense to any previously unseen word in a new context. The dataset covers 6 commonly ambiguous words, i.e., alam, atas, kayu, anggur, perdana, and dasar, with a total of 2416 sentences.	Unknown	2018	ind	Machine generated / Crawling w/o curation	2,416	sentences	No dataset split

Continued on the next page

Table A – continued from the previous page

Dataset	Description	License	Year	Language	Annotation Quality	Data Volume	Data Unit	Split
Indonesian general domain MT En-Id (Guntara et al., 2020)	For the general domain, both Tatoeba and TALPCo are manually curated, but their sentences (especially Tatoeba) are very short compared to Wikimatrix. Therefore, for these two datasets, we do a random split involving all datasets in the domain for validation and testing, each having 2000 unique pairs not present in the training set. For the general domain, we mix shorter sentences from TALPCo and the longer ones from Wikimatrix as our validation and test data. We observe that Tatoeba has similar types of high-quality sentences like TALPCo has, albeit shorter. Therefore we choose TALPCo to be in the validation and test sets instead, because longer sentences mean more difficult and meaningful evaluation. Tatoeba dataset contains short sentences. However, they contain high-quality full-sentence pairs with precise translation and is widely used in previous work in other languages (Artetxe and Schwenk, 2019b). Due to its simplicity, we do not use Tatoeba as our test and validation sets. We find that the Wikipedia scraper for Wikimatrix is faulty in some cases, causing some noise coming from unfiltered markup tags.	CC-BY-SA 4.0	2020	eng, ind	Human generation & curation	1,811,300	sentences	1729472 train, 2000 validation, 2000 test

Continued on the next page

Table A – continued from the previous page

Dataset	**Description**	**License**	**Year**	**Language**	**Annotation Quality**	**Data Volume**	**Data Unit**	**Split**
Indonesian religious domain MT En-Id (Guntara et al., 2020)	Religious domain consists of religious manuscripts or articles. These articles are different from news as they are not in a formal, informative style. Instead, they are written to advocate and inspire religious values, often times citing biblical or quranic anecdotes. The Tanzil dataset is a Quran translation dataset which has a relatively-imbalanced sentence length between the two languages, evidenced in Table 2, where an average Indonesian sentence in this dataset is about 50% longer than an average English one. Furthermore, an average pair of sentences in this dataset would, on average, have one of them twice as long as the other. However, we still decide to include the dataset in the domain to avoid overfitting because the remaining datasets are all about Christianity. Another interesting property in the religion domain corpus is the localized names, for example, David to Daud, Mary to Maryam, Gabriel to Jibril, and more. In contrast, entity names are usually kept unchanged in other domains. We also find quite a handful of Indonesian translations of JW300 are missing the end sentence dot (.), even though the end sentence dot is present in their English counterpart. Lastly, we also find some inconsistency in the transliteration, for example praying is sometimes written as *salat* or *shalat*, or repentance as *tobat* or *taubat*.	CC-BY-SA 4.0	2020	eng, ind	Human generation & curation	1,068,400	sentences	579544 train, 5000 validation, 4823 test

Continued on the next page

Table A – continued from the previous page

Dataset	Description	License	Year	Language	Annotation Quality	Data Volume	Data Unit	Split
JATI (Moeljadi, 2017)	JATI is a treebank built from a subset of parsed dictionary definition sentences. The main data for this study comes from the fifth edition of Kamus Besar Bahasa Indonesia (KBBI) (Amalia, 2016), the official and the most comprehensive dictionary for the Indonesian language. The dictionary definition sentences are parsed using the Indonesian Resource Grammar (INDRA) (Moeljadi, Bond, and Song 2015), a computational grammar for Indonesian in the Head-Driven Phrase Structure Grammar (HPSG) framework (Sag, Wasow, and Bender, 2003). JATI will be employed to build an ontology, in which knowledge is extracted from the semantic representation in Minimal Recursion Semantics (MRS) (Copestake et al., 2005).	Unknown	2017	ind	Human generation & curation	1,253	sentences	No dataset split
JV-ID ASR (Kjartansson et al., 2018)	This dataset contains transcribed audio data for Javanese. The dataset consists of wave files, and a TSV file. The file utt_spk_text.tsv contains a FileID, UserID and the transcription of audio in the file. The dataset has been manually quality checked, but there might still be errors. This dataset was collected by Google in collaboration with Reykjavik University and Universitas Gadjah Mada in Indonesia.	CC-BY-SA 4.0	2018	jav	Human generation & curation	185,076	utterances	No dataset split
JV-ID TTS (Sodimana et al., 2018)	This dataset contains high-quality transcribed audio data for Javanese. The dataset consists of wave files, and a TSV file. The file line_index.tsv contains a filename and the transcription of audio in the file. Each filename is prepended with a speaker identification number. The dataset has been manually quality checked, but there might still be errors.This dataset was collected by Google in collaboration with Gadjah Mada University in Indonesia.	CC-BY-SA 4.0	2018	ind, jav	Human generation & curation	5,800	sentences	No dataset split

Continued on the next page

Table A – continued from the previous page

Dataset	Description	License	Year	Language	Annotation Quality	Data Volume	Data Unit	Split
JaDi-Ide (Hidayatullah et al., 2020)	The dataset is collected from Twitter. We named the dataset as Javanese dialect identification (JaDi-Ide). The dialect is classified into Standard Javanese, Ngapak Javanese, and East Javanese dialects.	Unknown	2020	jav	Crawling & human annotation	16,000	tweets	No dataset split
KEPS (Mahfuzh et al., 2019)	KEPS is a keyphrase extraction dataset consists of text from Twitter discussing banking products and services and is written in the Indonesian language. A phrase containing important information is considered a keyphrase. Text may contain one or more keyphrases since important phrases can be located at different positions. The dataset follows the IOB chunking format, which represents the position of the keyphrase.	CC-BY-SA 4.0	2019	ind	Crawling & human annotation	1,247	documents	1000 train, 247 test
KaWAT (Kurniawan, 2019)	We introduced KaWAT (Kata Word Analogy Task), a new word analogy task dataset for Indonesian. We evaluated on it several existing pretrained Indonesian word embeddings and embeddings trained on Indonesian online news corpus. We also tested them on two downstream tasks and found that pretrained word embeddings helped either by reducing the training epochs or yielding significant performance gains.	Apache 2.0	2019	ind	Human generation & curation	34,000	sentence pairs	No dataset split
Kamus Alay (Salsabila et al., 2018)	We provide a lexicon for text normalization of Indonesian colloquial words. We gathered 3,592 unique colloquial words-also known as *bahasa alay* -and manually annotated them with the normalized form. We built this lexicon from Instagram comments provided by Septiandri & Wibisono (2017).	Unknown	2018	ind	Human generation & curation	3,592	tokens	No dataset split
Karonese dataset (Karo et al., 2022)	Karonese dataset consist karonese text and the label (positive, negaitive or neutra). karonese text comes from multi domain social media, such us facebook, twitter, Instagram and Youtube	Unknown	2022	btx	Crawling & human annotation	1,001	sentences	0 train, 0 validation, 0 test

Continued on the next page

Table A – continued from the previous page

Dataset	Description	License	Year	Language	Annotation Quality	Data Volume	Data Unit	Split
Kethu (Arwidarasti et al., 2019)	Kethu is a constituency treebank derived from Universitas Indonesia Constituency Treebank (UI-CTB) corpus Kethu converts UI-CTB treebank format into the widely accepted Penn Treebank format by adjusting the bracketing format for compound words as well as the POS tagset according to the Penn Treebank format. In addition, word segmentation and POS tagging of a number of tokens are also revised from the original UI-CTB corpus.	Unknown	2019	ind	Machine generated w/ human curation	1,030	sentences	No dataset split
KoPI-CC (Korpus Perayapan Indonesia)	KoPI-CC (Korpus Perayapan Indonesia)-CC is Indonesian Only Extract from Common Crawl snapshots ,each snapshots get extracted using ungoliant oscar tools and get extra filtering using deduplication technique (Exact Hash Dup and Minhash LSH)	CC0	2022	ind	Machine generated / Crawling w/o curation	106	GB	No dataset split
KoPI-CC_ News	KoPI(Korpus Perayapan Indonesia)-CC_News is Indonesian Only Extract from CC NEWS Common Crawl from 2016-2022(july) ,each snapshots get extracted using warcio and filter using fasttext	CC0	2022	ind	Machine generated / Crawling w/o curation	4	GB	No dataset split
KoPI-NLLB	KopI(Korpus Perayapan Indonesia)-NLLB, is Indonesian family language(aceh,bali,banjar,indonesia,jawa,minang,sunda) only extracted from NLLB Dataset each language set also filtered using some deduplicate technique such as exact hash(md5) dedup technique and minhash LSH neardup	ODC-BY	2022	ind, sun, jav, min, ban, bjn, ace	Machine generated / Crawling w/o curation	18	GB	No dataset split

Continued on the next page

Table A – continued from the previous page

Dataset	Description	License	Year	Language	Annotation Quality	Data Volume	Data Unit	Split
Korpus Nusantara (Sujaini, 2020)	The dataset is a combination of multiple machine translation works from the author, Herry Sujaini, covering Indonesian to 25 local dialects in Indonesia. Since not all dialects have ISO639-3 standard coding, as agreed with Pak Herry , we decided to group the dataset into the closest language family, i.e., Javanese, Dayak, Buginese, Sundanese, Madurese, Banjar, Batak Toba, Hakka (Khek), Malay, Minangkabau, and Min Nan (Teochew).	Unknown	2022	ind, sun, jav, min, mad, bbc, bug, msa, xdy, hak, nan	Human generation & curation	68,856	sentence pairs	No dataset split
LibriVox-Indonesia (Wirawan, 2022)	The LibriVox Indonesia dataset consists of MP3 audio and a corresponding text file we generated from the public domain audiobooks LibriVox. We collected only languages in Indonesia for this dataset. The original LibriVox audiobooks or sound files' duration varies from a few minutes to a few hours. Each audio file in the speech dataset now lasts from a few seconds to a maximum of 20 seconds.	CC0	2022	ind, sun, jav, min, bug, ban, ace	Machine generated / Crawling w/o curation	7,815	utterances	No dataset split
Liputan6 Summarization (Koto et al., 2020a)	The Liputan6 dataset was crawled from an online Indonesian news portal, which covers a wide range of topics, such as politics, sport, technology, business, health, and entertainment. There are two different experimental settings for Liputan6: Canonical, which includes all the test samples, and Xtreme, which only includes test samples with more than 90% novel 4-grams in the summary label.	CC-BY-SA 4.0	2021	ind	Crawling & human annotation	224,637	sentences	193883 train, canonical: 10972 validation, 10972 test), extreme: 4948 validation, 3862 test),

Continued on the next page

Table A – continued from the previous page

Dataset	Description	License	Year	Language	Annotation Quality	Data Volume	Data Unit	Split
Local ID Abusive (Putri et al., 2021)	Local ID Abusive is dataset aimed to be used for abusive and hate speech detection available in Javanese and Sundanese. The Javanese and Sundanese were annotated manually by annotator from each region. The annotation process involved multiple-step processes. It was carried by two annotators for each language, after an initial step where the guidelines were discussed and refined to reach unanimous comprehension. The annotation process gives 3449 and 2207 tweets for Javanese and Sundanese dataset respectively with 100% agreement.	Unknown	2021	sun, jav	Crawling & human annotation	5,656	sentences	No dataset split
MaRVL (Liu et al., 2021a)	Multicultural Reasoning over Vision and Language (MaRVL) is a dataset based on an ImageNet-style hierarchy representative of many languages and cultures (Indonesian, Mandarin Chinese, Swahili, Tamil, and Turkish). The selection of both concepts and images is entirely driven by native speakers. Afterwards, we elicit statements from native speakers about pairs of images. The task consists in discriminating whether each grounded statement is true or false. The present file contains all the dataset images and annotations.	CC-BY 4.0	2021	ind	Crawling & human annotation	1,128	image, image, concept, caption	No dataset split
MinangNLP MT (Koto and Koto, 2020)	MinangNLP MT is a machine translation dataset generated from two sentiment analysis dataset by manual translation to Minangkabau language	MIT	2020	ind, min	Crawling & human annotation	5,000	sentences	11,571 train, 1600 validation, 3200 test
MultiLexNorm (van der Goot et al., 2021a)	MultiLexNorm is multilingual benchmark dataset for lexical normalization task for 12 languages, including Indonesian-English (code-mixed). Lexical normalization is the task of transforming an utterance into its standard form, word by word, including both one-to-many (1-n) and many-to-one (n-1) replacements. ID-EN dataset actually originates from Barik et.al. (2019) work. However, there is preprocessing work upon the original dataset.	CC-BY-NC-SA 4.0	2021	ind, eng	Crawling & human annotation	13,949	tokens	13950 train, 4810 validation, 4367 test

Continued on the next page

Table A – continued from the previous page

Dataset	Description	License	Year	Language	Annotation Quality	Data Volume	Data Unit	Split
Multilingual Open Relations (Faruqui and Kumar, 2015)	This dataset provides the set of automatically extracted relations obtained using cross-lingual annotation projection method. The data covers 61 languages, including Indonesian. Relation extraction is the task of assigning a semantic relationship between a pair of arguments. For example, from the sentence Soekarno lahir di Jawa Timur, the relation <Soekarno, born_in, Jawa Timur> is expected to be extracted.	Unknown	2015	ind	Machine generated / Crawling w/o curation	1,876	relations	No dataset split
NER UGM (IndoLEM split) (Fachri, 2014)	NER UGM is a named entity recognition dataset collected by UGM. We use IndoLEM split for the dataset.	CC BY-SA 3.0	2014	ind	Crawling & human annotation	2,343	sentences	1687 train, 187 validation, 469 test
NER UI (IndoLEM split) (Gultom and Wibowo, 2017)	NER UI is a named entity recognition dataset collected by UI. We use IndoLEM split for the dataset.	CC BY-SA 3.0	2017	ind	Human generation & curation	2,125	sentences	1530 train, 170 validation, 425 test
NERGrit	NER Grit dataset is a NER dataset taken from the Grit-ID repository, and the labels are spans in IOB chunking representation. The dataset consists of three kinds of named entity tags, PERSON (name of person), PLACE (name of location), and ORGANIZATION (name of organization).	custom	2020	ind	Unknown	17,437	sentences	12518 train, 2521 validation, 2398 test
NERP (IndoNLU Split) (Hoesen and Purwarianti, 2018)	NERP is a NER dataset which is collected from several Indonesian news websites, labelled with 5 entity classes: PER (name of person), LOC (name of location), IND (name of product or brand), EVT (name of the event), and FNB (name of food and beverage).	CC-BY-SA 4.0	2018	ind	Human generation & curation	8,400	sentences	6720 train, 840 validation, 840 test

Continued on the next page

Table A – continued from the previous page

Dataset	Description	License	Year	Language	Annotation Quality	Data Volume	Data Unit	Split
NLLB Seed (NLLB Team et al., 2022)	NLLB Seed is a set of professionally-translated sentences in the Wikipedia domain. Data for NLLB-Seed was sampled from Wikimedia's List of articles every Wikipedia should have, a collection of topics in different fields of knowledge and human activity. NLLB-Seed consists of around six thousand sentences in 39 languages. NLLB-Seed is meant to be used for training rather than model evaluation. Due to this difference, NLLB-Seed does not go through the human quality assurance process present in FLORES-200.	CC-BY-NC 4.0	2022	eng, ace, bjn, bug	Crawling & human annotation	30,965	sentence pairs	No dataset split
Netifier	Netifier dataset is a collection of scraped posts on famous social media sites in Indonesia, such as Instagram, Twitter, and Kaskus aimed to do multi-label toxicity classification. The dataset consists of 7,773 texts. The author manually labelled 7k samples into 4 categories: pornography, hate speech, racism, and radicalism.	CC-BY-NC-SA 4.0	2018	ind	Crawling & human annotation	7,773	tweets	6995 train, 0 validation, 778 test
News En-Id MT (Moeljadi and Aminullah, 2020)	News En-Id is a machine translation dataset containing Indonesian-English parallel sentences collected from news translation dataset (Guntara et al., 2020). The news dataset (Guntara et al., 2020) is collected from multiple sources: Pan Asia Networking Localization (PANL), Bilingual BBC news articles, Berita Jakarta, and GlobalVoices	CC-BY-SA 4.0	2021	eng, ind	Crawling & human annotation	44,325	sentences	38469 train, 1953 validation, 1954 test
NusaX MT (Winata et al., 2023)	The first-ever parallel resource for 10 low-resource languages in Indonesia.	CC-BY-SA 4.0	2022	ind, ace, jav, sun, min, bug, bbc, ban, nij, mad, bjn, eng	Human generation & curation	132,000	sentence pairs	500 train, 100 dev, 400 test

Continued on the next page

Table A – continued from the previous page

Dataset	Description	License	Year	Language	Annotation Quality	Data Volume	Data Unit	Split
NusaX Sentiment (Winata et al., 2023)	The first-ever parallel resource for 10 low-resource languages in Indonesia.	CC-BY-SA 4.0	2022	ind, ace, jav, sun, min, bug, bbc, ban, nij, mad, bjn, eng	Human generation & curation	12,000	sentences	500 train, 100 dev, 400 test
OJW (Moeljadi and Aminullah, 2020)	OJW is abbreviation of Old Javanese Wordnet.	Unknown	2020	kaw	Unknown	5,038	tokens	No dataset split
PANL BPPT (Riza and Hakim, 2009)	Parallel Text Corpora for Multi-Domain Translation System created by BPPT (Indonesian Agency for the Assessment and Application of Technology) for PAN Localization Project (A Regional Initiative to Develop Local Language Computing Capacity in Asia). The dataset contains around 24K sentences divided in 4 difference topics (Economic, international, Science and Technology and Sport).	Unknown	2009	eng, ind	Crawling & human annotation	24,000	sentences	No dataset split
POSP (IndoNLU Split) (Hoesen and Purwarianti, 2018)	POSP is an Indonesian part-of-speech tagging (POS) dataset collected from Indonesian news websites. The dataset consists of around 8000 sentences with 26 POS tags following the Indonesian Association of Computational Linguistics (INACL) POS Tagging Convention.	CC-BY-SA 4.0	2018	ind	Human generation & curation	8,400	sentences	6720 train, 840 validation, 840 test
ParaCotta (Aji et al., 2021)	ParaCotta is a synthetic parallel paraphrase corpus generated from monolingual data and a neural machine translation system. Multiple translations were generated using beam search, and then paraphrase pairs were selected based on the lexical difference determined by their sentence BLEU.	Unknown	2021	ind, eng	Machine generated / Crawling w/o curation	6,000,000	sentence pairs	No dataset split

Continued on the next page

Table A – continued from the previous page

Dataset	Description	License	Year	Language	Annotation Quality	Data Volume	Data Unit	Split
Parallel: Indonesian - Lampung Nyo (Abidin and Ahmad, 2021)	Parallel Indonesian - Lampung Nyo corpus is constructed from documents taken from the Lampung language book for elementary and junior high school levels in the province of Lampung. The document data that has been collected will then be manually typed to be made into a parallel corpus in Indonesian – Lampung dialect of nyo and mono corpus in Lampung dialect of nyo. There are 3000 parallel corpus sentences collected in Indonesian - Lampung dialect of nyo and 3000 mono corpus sentences in Indonesian - Lampung dialect of nyo.	Unknown	2021	abl, ind	Crawling & human annotation	3,000	sentences	No dataset split
PoSTagged Sundanese Monolingual Corpus (Ardiyanti Suryani et al., 2022a)	The dataset consists of 3616 Sundanese sentences collected from several Sundanese online magazine (Mangle, Dewan Dakwah Jabar, and Balebat). The dataset is mannualy annotated with Part of Speech label by several Sundanese Language faculty students from UPI Bandung.	CC0	2015	sun	Crawling & human annotation	3,616	sentences	No dataset split
SQuAD ID (Muis and Purwarianti, 2020)	Stanford Question Answering Dataset (SQuAD) is a reading comprehension dataset, consisting of questions posed by crowdworkers on a set of Wikipedia articles, where the answer to every question is a segment of text, or span, from the corresponding reading passage, or the question might be unanswerable. This version is translated version of SQuAD to Indonesian Language.	Unknown	2020	ind	Machine generated / Crawling w/o curation	124,631	documents	102657 train, 11407 validation, 10567 test
STIF-Indonesia (Wibowo et al., 2020)	STIF-Indonesia is an Indonesian text style transfer dataset collected from Twitter. The style transfer is done from informal style to a formal style. STIF-Indonesia consists of 52.5k sentences with 2.5k of which is manually annotated.	MIT	2020	ind	Crawling & human annotation	52.5k	sentences	1922 train, 214 validation, 363 test

Continued on the next page

Table A – continued from the previous page

Dataset	Description	License	Year	Language	Annotation Quality	Data Volume	Data Unit	Split
SU-ID ASR (Kjartansson et al., 2018)	This dataset contains transcribed audio data for Sundanese. The dataset consists of wave files, and a TSV file. The file utt_spk_text.tsv contains a FileID, UserID and the transcription of audio in the file. The dataset has been manually quality checked, but there might still be errors. This dataset was collected by Google in Indonesia.	CC-BY-SA 4.0	2018	sun	Human generation & curation	220,000	sentences	No dataset split
SU-ID TTS (Sodimana et al., 2018)	This dataset contains high-quality transcribed audio data for Sundanese. The dataset consists of wave files, and a TSV file. The file line_index.tsv contains a filename and the transcription of audio in the file. Each filename is prepended with a speaker identification number. The dataset has been manually quality checked, but there might still be errors. This dataset was collected by Google in collaboration with Universitas Pendidikan Indonesia.	CC-BY-SA 4.0	2018	sun	Human generation & curation	4,213	utterances	No dataset split
Sampiran (Siallagan and Alfina, 2023)	Sampiran is a dataset for pantun generation. It consists of 7.8K Indonesian pantun, collected from various sources (online). Pantun is a traditional Malay poem consisting of four lines: two lines of deliverance and two lines of message. This dataset filtered the gathered Pantun to follow the general rules of Pantun; four lines with ABAB rhyme and eight to twelve syllables per line.	AGPL-3.0	2023	ind	Crawling & human annotation	7,879	sentences	No dataset split
Semeval STS Indo	SemEval is a series of international natural language processing (NLP) research workshops whose mission is to advance the current state of the art in semantic analysis and to help create high-quality annotated datasets in a range of increasingly challenging problems in natural language semantics. This is a translated version of SemEval Dataset from 2012-2016 for Semantic Textual Similarity Task to Indonesian language.	Unknown	2019	ind, eng	Machine generated / Crawling w/o curation	12,901	sentence pairs	10293 train, 0 validation, 2608 test

Continued on the next page

Table A – continued from the previous page

Dataset	Description	License	Year	Language	Annotation Quality	Data Volume	Data Unit	Split
Singgalang (Alfina et al., 2017b)	Singgalang is an automatically tagged Indonesian NER dataset collected from Wikipedia articles which uses DBpedia as the reference of the entity type. The dataset used expanded DBpedia of MDEE_Gazetteer as the reference to label the token and contains 48,957 sentences.	Unknown	2017	ind	Machine generated / Crawling w/o curation	48,957	sentences	No dataset split
SmSA (Purwarianti and Crisdayanti, 2019)	SmSA is a sentence-level sentiment analysis dataset consisting of of comments and reviews in Indonesian obtained from multiple online platforms with 3 possible sentiments:positive, negative, and neutral. The text was crawled and then annotated by several Indonesian linguists to construct this dataset.	CC-BY-SA 4.0	2020	ind	Crawling & human annotation	12,760	sentences	11000 train, 1260 validation, 500 test
Sundanese Twitter Dataset for Emotion (Putra et al., 2020)	Sunda Emotion dataset gathered from Twitter API between January and March 2019 with 2518 tweets in total. The tweets filtered by using some hashtags which are represented Sun-danese emotion, for instance, #persib, #corona, #saredih, #nyakakak, #garoblog, #sangsara, #gumujeng, #bungah, #sararieun, #ceurik, and #hariwang. This dataset contains four distinctive emotions: anger, joy, fear, and sadness.Each tweet is annotated using related emotion. For data validation, we consulted a Sundanese language teacher for expert validation	Unknown	2020	sun	Crawling & human annotation	2,518	sentences	No dataset split
Sundanese-Indonesian Parallel Corpus (Ardiyanti Suryani et al., 2022b)	The dataset consists of 3616 Sundanese sentences taken from a Sundanese online magazing (Mangle), Dewan Dakwah Jabar, and Balebat. The dataset is manually translated to Indonesian language by several Sundanese Language faculty students from UPI Bandung.	CC0	2015	ind, sun	Human generation & curation	3,616	sentence pairs	No dataset split

Continued on the next page

Table A – continued from the previous page

Dataset	Description	License	Year	Language	Annotation Quality	Data Volume	Data Unit	Split
TALPCo (Nomoto et al., 2018)	The TUFS Asian Language Parallel Corpus (TALPCo) is an open parallel corpus consisting of Japanese sentences and their translations into Korean, Burmese (Myanmar; the official language of the Republic of the Union of Myanmar), Malay (the national language of Malaysia, Singapore and Brunei), Indonesian, Thai, Vietnamese and English	CC-BY 4.0	2018	ind, eng, kor, jpn, vie, tha, msa	Unknown	0	sentences	No dataset split
TED En-Id MT (Qi et al., 2018)	TED En-Id is a machine translation dataset containing Indonesian-English parallel sentences collected from TED translation dataset (Qi et al., 2018), collected from TED talk transcripts	CC-BY-SA 4.0	2021	eng, ind	Crawling & human annotation	93,262	sentences	87406 train, 2677 validation, 3179 test
TICO-19 (Anastasopoulos et al., 2020)	TICO-19 (Translation Initiative for COVID-19) is sampled from a variety of public sources containing COVID-19 related content, representing different domains (e.g., news, wiki articles, and others). TICO-19 includes 30 documents (3071 sentences, 69.7k words) translated from English into 36 languages: Amharic, Arabic (Modern Standard), Bengali, Chinese (Simplified), Dari, Dinka, Farsi, French (European), Hausa, Hindi, Indonesian, Kanuri, Khmer (Central), Kinyarwanda, Kurdish Kurmanji, Kurdish Sorani, Lingala, Luganda, Malay, Marathi, Myanmar, Nepali, Nigerian Fulfulde, Nuer, Oromo, Pashto, Portuguese (Brazilian), Russian, Somali, Spanish (Latin American), Swahili, Congolese Swahili, Tagalog, Tamil, Tigrinya, Urdu, Zulu.	CC0	2020	ind, eng, ara, spa, fra, hin, por, rus, zho	Machine generated w/ human curation	21,497	sentence pairs	0 train, 6797 validation, 14700 test

Continued on the next page

Table A – continued from the previous page

Dataset	Description	License	Year	Language	Annotation Quality	Data Volume	Data Unit	Split
TITML-IDN (Lestari, 2006)	TITML-IDN (Tokyo Institute of Technology Multilingual - Indonesian) is collected to build a pioneering Indonesian Large Vocabulary Continuous Speech Recognition (LVCSR) System. In order to build an LVCSR system, high accurate acoustic models and large-scale language models are essential. Since Indonesian speech corpus was not available yet, we tried to collect speech data from 20 Indonesian native speakers (11 males and 9 females) to construct a speech corpus for training the acoustic model based on Hidden Markov Models (HMMs). A text corpus which was collected by ILPS, Informatics Institute, University of Amsterdam, was used to build a 40K-vocabulary dictionary and a n-gram language model.	For research purposes only. If you use this corpus, you have to cite (Lestari et al, 2006).	2006	ind	Human generation & curation	6,679	utterances	No dataset split
TUFS Indonesia Constituency Tree (Nomoto, 2022)	TUFS Indonesia Constituency Tree is annotated dataset for Indonesian language constituency tree.	CC-BY 4.0	2022	ind, msa	Human generation & curation	1,385	sentences	No dataset split
TermA (Fernando et al., 2019)	The TermA span-extraction dataset is collected from the hotel aggregator platform, AiryRooms. The dataset consists of thousands of hotel reviews, which each contain a span label for aspect and sentiment words representing the opinion of the reviewer on the corresponding aspect. The labels use Inside-Outside-Beginning (IOB) tagging representation with two kinds of tags, aspect and sentiment.	CC-BY-SA 4.0	2019	ind	Human generation & curation	5,000	documents	3000 train, 1000 validation, 1000 test

Continued on the next page

Table A – continued from the previous page

Dataset	Description	License	Year	Language	Annotation Quality	Data Volume	Data Unit	Split
Toxicity-200 (NLLB Team et al., 2022)	Toxicity-200 is a wordlist to detect toxicity in 200 languages. It contains files that include frequent words and phrases generally considered toxic because they represent: 1) frequently used profanities; 2) frequently used insults and hate speech terms, or language used to bully, denigrate, or demean; 3) pornographic terms; and 4) terms for body parts associated with sexual activity.	CC-BY-NC 4.0	2022	ind, ace, bjn, bug, jav	Human generation & curation	226	phrases	No dataset split
TyDiQA Id (Cahyawijaya et al., 2021b)	For the question answering task, we use the TyDiQA (Clark et al., 2020) dataset. This dataset is collected from Wikipedia articles with human-annotated question and answer pairs covering 11 languages. The question-answer pairs are collected for each language without using translation services. We use the Indonesian data from the secondary Gold passage task of the TyDiQA dataset. As the original dataset only provides training and validation sets, we randomly split off 15% of the training data and use it as the test set.	CC-BY-SA 4.0	2021	ind	Human generation & curation	6,267	sentences	4847 train, 565 validation, 855 test
UD_Indonesian-CSUI (Alfina et al., 2020)	The UD_Indonesian-CSUI is a dependency treebank in Indonesian in the CoNLL-U format. It was converted from a consituency treebank (Kethu) while Kethu was also converted from another consituency treebank (IDN treebank). Currently, this treebank, consist of 1030 sentences.	MIT	2020	ind	Machine generated w/ human curation	1,030	sentences	656 train, 0 validation, 374 test
UD_Indonesian-GSD (McDonald et al., 2013)	UD_Indonesian-GSD is is an Indonesian-GSD treebank dataset originally converted from the content head version of the universal dependency treebank v2.0 (legacy) in 2015. In order to comply with the latest Indonesian annotation guidelines, the treebank has undergone a major revision between UD releases v2.8 and v2.9 (2021).	CC BY-SA 3.0	2013	ind	Human generation & curation	5,593	sentences	4477 train, 559 validation, 557 test

Continued on the next page

Table A – continued from the previous page

Dataset	Description	License	Year	Language	Annotation Quality	Data Volume	Data Unit	Split
UD_ Indonesian-PUD (Alfina et al., 2019)	An Indonesian dependency treebank that is part of a collection of 18 Parallel Universal Dependencies (PUD) treebanks.	CC BY-SA 3.0	2019	ind	Human generation & curation	1,000	sentences	10 fold cross validation according to UD standard
UKARA 1.0 Challenge (Septiandri and Winatmoko, 2020)	Ukara 1.0 Challenge dataset is a dataset for automatic short answer scoring system which is a collaboration project between FMIPA UGM and PUSPENDIK, Ministry of Education and Culture of Indonesia. It was intended to build supervised machine learning approach which is able to assign a score to student's answer. The student's answer usually consists of maximum 2-3 sentences.	Unknown	2020	ind	Human generation & curation	2,861	sentences	268 train, 215 validation, 855 test
Unimorph ID (Pimentel et al., 2021)	The Universal Morphology (UniMorph) project is a collaborative effort to improve how NLP handles complex morphology in the world's languages. The goal of UniMorph is to annotate morphological data in a universal schema that allows an inflected word from any language to be defined by its lexical meaning, typically carried by the lemma, and by a rendering of its inflectional form in terms of a bundle of morphological features from our schema	Creative Commons Attribution-ShareAlike 3.0 Unported (CC BY-SA 3.0)	2021	ind	Crawling & human annotation	27,714	forms	70% train, 10% validation, 20% test
VoxLingua107 (Valk and Alumäe, 2021)	VoxLingua107 is a speech dataset for training spoken language identification models. The dataset consists of short speech segments automatically extracted from YouTube videos and labeled according the language of the video title and description, with some post-processing steps to filter out false positives. VoxLingua107 contains data for 107 languages, including Indonesian, Javanese, and Sundanese.	CC-BY 4.0	2021	ind, sun, jav	Crawling & human annotation	157	hours	No dataset split

Continued on the next page

Table A – continued from the previous page

Dataset	Description	License	Year	Language	Annotation Quality	Data Volume	Data Unit	Split
WReTe (Setya and Mahendra, 2023)	WReTe is an entailment dataset which consists of 450 sentence pairs constructed from Wikipedia revision history. The dataset contains pairs of sentences and binary semantic relations between the pairs. The data are labeled as entailed when the meaning of the second sentence can be derived from the first one, and not entailed otherwise.	CC-BY-SA 4.0	2018	ind	Crawling & human annotation	450	sentence pairs	300 train, 50 validation, 100 test
WikiAnn (Pan et al., 2017)	We developed a simple yet effective framework that can extract names from 282 languages and link them to an English KB. This framework follows a fully automatic training and testing pipeline, without the needs of any manual annotations or knowledge from native speakers. We release the following resources for each of these 282 languages: "silver-standard" name tagging and linking annotations with multiple levels of granularity, morphology analyzer if it's a morphologically-rich language, and an end to-end name tagging and linking system.	Attribution License (ODC-By)	2017	ind, eng, sun, jav, min, bug, bjn, tpi, ace, tdt, msa, jav-bms	Machine generated / Crawling w/o curation	254,240	Number of name mentions	No dataset split
WikiLingua (Ladhak et al., 2020)	We introduce WikiLingua, a large-scale, multilingual dataset for the evaluation of crosslingual abstractive summarization systems. We extract article and summary pairs in 18 languages from WikiHow12, a high quality, collaborative resource of how-to guides on a diverse set of topics written by human authors. We create gold-standard article summary alignments across languages by aligning the images that are used to describe each how-to step in an article.	CC-BY-NC-SA 3.0	2020	ind, eng	Crawling & human annotation	47,511	article-summary pairs	No dataset split

Continued on the next page

Table A – continued from the previous page

Dataset	Description	License	Year	Language	Annotation Quality	Data Volume	Data Unit	Split
X-FACT (Gupta and Srikumar, 2021)	The largest publicly available multilingual dataset for factual verification of naturally existing real world claims. The dataset contains short statements in 25 languages and is labeled for veracity by expert fact-checkers. The dataset includes a multilingual evaluation benchmark that measures both out-of-domain generalization, and zero-shot capabilities of the multilingual models.	MIT	2021	ind, eng	Crawling & human annotation	3,548	evidences-links-claim	2231 train, 297 validation, 448 test
XCOPA (Ponti et al., 2020)	Cross-lingual Choice of Plausible Alternatives (XCOPA), a typologically diverse multilingual dataset for causal commonsense reasoning in 11 languages, including Indonesian. The causal commonsense reasoning task consists two task variations, forward causal reasoning and backward reasoning. In forward causal reasoning, given the premise sentence, the model is asked to predict most reasonable result from two alternative. In backward causal reasoning, the model is asked to predict what causes the premise happens	CC-BY 4.0	2021	ind	Human generation & curation	600	sentences	0 train, 100 validation, 500 test
XL-Sum (Hasan et al., 2021)	XL-Sum is a comprehensive and diverse dataset comprising 1 million professionally annotated article-summary pairs from BBC, extracted using a set of carefully designed heuristics. The dataset covers 44 languages, including Indonesian.	CC-BY-NC-SA 4.0	2021	ind	Crawling & human annotation	47,802	document-summary pairs	38242 train, 4780 validation, 4780 test
XPersona Id (Lin et al., 2021)	XPersona is a open-domain dialogue system on 7 languages including Indonesia. The test set is manually translated by exper annotators, while the training and validation set isareautomatically translated from the persona chat dataset with an additional manual keyword correction phase.	CC-BY-SA 4.0	2021	ind	Machine generated w/ human curation	17,866	utterances	16878 train, 484 validation, 484 test

Continued on the next page

Table A – continued from the previous page

Dataset	Description	License	Year	Language	Annotation Quality	Data Volume	Data Unit	Split
id-en-code-mixed (Barik et al., 2019)	This dataset contain 825 tweet instances of Indonesian-English, corresponding to four NLP tasks, i.e., tokenization, language identification, lexical normalization, and word translation. Data for lexical normalization task is curated in MultiLexNorm (already in Nusa Catalogue), but other tasks are not. Tokenization for social media data is not as trivial as splitting the token using white space delimiter. In this data, language identification is performed in token-level granularity.	CC-BY-NC-SA 4.0	2019	ind, eng	Crawling & human annotation	22,736	tokens	No dataset split
Indo Wiki Paralel Corpora (Trisedya and Inastra, 2014)	Manually aligned parallel corpora from Wikipedia	Unknown	2014	ind, sun, jav, min	Crawling & human annotation	2,422	sentence pairs	No dataset split
indo-law (Nuranti et al., 2022)	This dataset consists of Indonesian court decision documents for general criminal cases that have been annotated for the document sections. The documents were taken from the website of the Indonesian Supreme Court Decision. There are 22,630 documents with xml format in this dataset, which each contains 11 tags that enclose the annotated sections of the court decision documents.	Unknown	2022	ind	Crawling & human annotation	22,630	documents	No dataset split
xSID (van der Goot et al., 2021b)	We introduce XSID, a new benchmark for cross-lingual (X) Slot and Intent Detection in 13 languages from 6 language families, including a very low-resource dialect.	CC-BY-SA 4.0	2021	ind, eng	Crawling & human annotation	5,370,460	sentences	No dataset split

Table A: Overview of all datasets in NusaCrowd.