NusaDialogue: Dialogue Summarization and Generation for Underrepresented and Extremely Low-Resource Languages

Ayu Purwarianti[*2,6] **Dea Adhista**[*1] **Agung Baptiso**[*1]
Miftahul Mahfuzh[*1] **Yusrina Sabila**[*1] **Aulia Adila**[*5]
Samuel Cahyawijaya[*3,6,7] **Alham Fikri Aji**[*4,6,7]

[1]Prosa.ai [2]ITB [3]HKUST [4]MBZUAI [5]JAIST [6]SEACrowd [7]IndoNLP
ayu@informatika.org; dea.adhista@prosa.ai; agung.sorlawan@prosa.ai;
miftahul.mahfuzh@prosa.ai; yusrina.sabila@prosa.ai; adila@jaist.ac.jp;
samuelcahyawijaya@cohere.com[*]; alham.fikri@mbzuai.ac.ae

Abstract

Developing dialogue summarization for extremely low-resource languages is a challenging task. We introduce NusaDialogue, a dialogue summarization dataset for three underrepresented languages in the Malayo-Polynesian language family: Minangkabau, Balinese, and Buginese. NusaDialogue covers 17 topics and 185 subtopics, with annotations provided by 73 native speakers. Additionally, we conducted experiments using fine-tuning on medium-sized Indonesian-specific language models (LMs), as well as zero- and few-shot learning on various multilingual large language models (LLMs). The results indicate that, for extremely low-resource languages such as Minangkabau, Balinese, and Buginese, the fine-tuning approach yields significantly higher performance compared to zero- and few-shot prompting, even when applied to LLMs with considerably larger parameter sizes. We publicly release the NusaDialogue dataset in https://huggingface.co/datasets/prosa-text/nusa-dialogue under CC-BY-SA 4.0 license.

1 Introduction

Large language models (LLMs) have brought remarkable progress in language processing technology attaining a high-quality language understanding and generation capability (Workshop et al., 2023; Muennighoff et al., 2023; Bang et al., 2023; OpenAI et al., 2024; Cahyawijaya et al., 2024b; Üstün et al., 2024; Aryabumi et al., 2024). Nonetheless, the generalization toward low-resource languages is still lacking causing a huge disparity in the applicability and accessibility of LLMs in numerous underrepresented languages such as languages spoken in Africa (Adelani et al., 2022, 2023; Muhammad et al., 2023; Adelani et al., 2024; Winata et al., 2024), and South-East Asia (Cahyawijaya et al., 2023b; Yong et al., 2023b; Lovenia et al., 2024; Singh et al., 2024; Cahyawijaya et al., 2024c; Winata et al., 2024; Urailertprasert et al., 2024; Romero et al., 2024). Various efforts provide solutions to this problem by developing novel resources on these underrepresented languages (Adilazuarda et al., 2022; Yong et al., 2023a; Cahyawijaya et al., 2023d, 2024a; Adilazuarda et al., 2024).

Despite the incredible progress, most works focus on machine translation and simple language understanding tasks, such as sentiment analysis and topic classification. More complex tasks such as open-domain dialogue, task-oriented dialogue, and dialogue summarization, are still left behind for these underrepresented languages. The task coverage limitation leads to a poor evaluation suite for assessing the capability of LLMs in these underrepresented languages. Moreover, most datasets on these underrepresented languages are developed through translating text from other higher-resource languages resulting in a translationese corpus (Winata et al., 2023; Cahyawijaya et al., 2023c; Cahyawijaya, 2024) which is not ideal for representing these underrepresented languages.

In this work, we develop NusaDialogue, the first dialogue summarization dataset covering 3 underrepresented languages under the Malayo-Polynesian languages group, i.e., Minangkabau (min), Balinese (ban), and Buginese (bug). NusaDialogue is a human-annotated colloquial-styled dialogue summarization dataset covering 17 topics and 185 subtopics. The colloquial and non-translationese annotation nature of NusaDialogue makes it suitable for representing the actual day-to-day use of these underrepresented languages. We ensure that the dataset is annotated by a balanced number of male and female annotators to make the dataset represent a more balanced demography.

We further analyze the annotator bias based on

Proceedings of the Second Workshop in South East Asian Language Processing, pages 82–100
January 20, 2025.

the choice of topics and the gender of speakers within a conversation and find out that, despite being regionally diverse, the gender bias in the languages contains huge similarities. This showcases that gender bias is not only affected by local cultural values but also by broader values such as shared geopolitical and historical values. Additionally, when comparing with prior work on bias in high-resource languages such as English (Caliskan et al., 2017; Guo and Caliskan, 2021; Orgad et al., 2022; Sant et al., 2024; Stewart and Mihalcea, 2024), despite having a smaller correlation, we still find numerous amount of similarities. This showcases the potential of extracting a different scope of bias, i.e., regional, national, or global, by analyzing the bias behavior of multilingual corpora. We summarize our contribution in four-fold:

- We introduce NusaDialogue, the first dialogue summarization datasets for three underrepresented and extremely low-resource languages, which is a suitable resource for the evaluation of language understanding and generation capabilities in these languages.
- We are the first to conduct a gender bias analysis on these languages and find out that, despite having no gendered pronoun or other masculine-feminine word variation, bias in terms of gender can still be perceived in **annotation-level**, i.e., the gender of the annotator, and **topic-level**, i.e., the gender of the individual named entities in the text.
- We introduce the potential of NusaDialogue for training and benchmarking the understanding and generation capability of LLMs on three extremely low-resource languages through a dialogue summarization task.
- We develop the first gender bias analysis of LLMs in three extremely low-resource languages. In addition, we showcase a simple augmentation method through name-swapping which effectively reduces the gender bias of LMs in these languages.

2 Related Work

NLP Resources for Underrepresented Languages Most research works in today's NLP technology are culturally Anglocentric with English as the main language (Søgaard, 2022; Talat et al., 2022). While many languages, such as thousands of Austronesian languages, remain underrepresented and are over-dominated by other few high-resource languages. Prior works (Cahyawijaya et al., 2023b; Kakwani et al., 2020; Koto et al., 2020; Koto and Koto, 2020; Wilie et al., 2020; Adelani et al., 2021; Cahyawijaya et al., 2021; Ebrahimi et al., 2022; Park et al., 2021; Kumar et al., 2022; Winata et al., 2023; Adilazuarda et al., 2022; Ogundepo et al., 2023; Kabra et al., 2023; Song et al., 2023) have developed corpora for these languages mainly through document translation (Winata et al., 2023) and online scraping (Koto et al., 2021, 2022). Although such data collection methods could be effective in high-resource languages, applying the methods in underrepresented languages requires further investigation.

NLP Evaluations for Underrepresented Languages The rapid development of language technologies has enhanced accessibility across diverse linguistic communities, enabling various language understanding and generation capabilities. The evaluation processes for assessing the performance and effectiveness of these technologies to address the unique challenges posed by target languages (Aji et al., 2022; Khanuja et al., 2023; Lai et al., 2023; Cahyawijaya, 2024) has also been refined. These evaluation has also gone beyond language modality alone, but also extending to multimodality (Lovenia et al., 2024; Winata et al., 2024; Romero et al., 2024; Urailertprasert et al., 2024).

3 NusaDialogue Corpus

3.1 Corpus Coverage

3.1.1 Languages

NusaDialogue covers three extremely low-resource languages under the Austronesian language family that is spoken in Indonesia, i.e., Minangkabau (min), Balinese (ban), and Buginese (bug). All these languages are not covered in most multilingual pre-training and instruction-tuning corpora such as mC4 (Xue et al., 2021), ROOTS (Laurençon et al., 2023), XP3 (Muennighoff et al., 2023), PaLM (Chowdhery et al., 2022), PaLM2 (Anil et al., 2023), XGLM (Lin et al., 2022) etc; and in various off-the-shelf language identification models such as LangDetect (Nakatani, 2011), langid.py (Lui and Baldwin, 2012), CLD2 (Sites, 2013), FastText LID (Joulin et al., 2017), and CLD3 (Salcianu et al., 2020). A handful amount of data on these languages is covered in Wikipedia and recent works focusing on Indonesian local languages (Winata et al., 2023;

Language	Dialects
Balinese	Badung, Bali, Bali Aga, Bangli, Buleleng, Dataran, Denpasar, Gianyar, Karangasem, Klungkung, Singaraja, Tabanan
Buginese	Barru, Bone, Bugis, Bulukumba, Magai Io, Makassar, Maros, Pangkep, Pinrang, Sengkang, Sidenreng Rappang, Sinjai, Soppeng, Wajo
Minangkabau	Agam, Bukittinggi, Minangkabau, Padang, Padang Panjang, Pariaman, Pasaman, Payakumbuh, Sijunjung, Tanah Datar

Table 1: The dialect coverage of all annotators for each language under study in NusaDialogue.

Cahyawijaya et al., 2023a,c).

Minangkabau (min), primarily spoken in West Sumatra and other Sumatra Island provinces like Bengkulu and Riau, is classified as Malay but lacks mutual intelligibility with Indonesian. Expressed in the Latin script, it adheres to an SVO word order. Standard Minangkabau exhibits an Indonesian-type voice, while colloquial Minangkabau is characterized as a Sundic-type system (Crouch, 2009). Balinese (ban), spoken mainly in Bali and West Nusa Tenggara provinces, features Highland Balinese, Lowland Balinese, and Nusa Penida dialects. Despite having its own script, it is predominantly written in Latin, maintaining an SVO order, lacking tonality, and comprising 17 consonants and 6 vowels. Stress is on the penultimate syllable, and it employs an 'active' or 'split-S' verb affixation system (Arka, 2003). Buginese (bug), spoken in South Sulawesi, Southeast Sulawesi, Central Sulawesi, and West Sulawesi, adheres to SVO word order, using verb affixes for person marking. Lacking tonality, it consists of 19 consonants and 6 vowels, historically using the Buginese script but now predominantly using the Latin script (Eberhard et al., 2021). Buginese features three forms for the pronoun 'I': 'iyya,' '-ka,' and 'u-.' Politeness in Buginese is conveyed through sentence patterns, pronouns, and specific terms (Weda, 2016).

3.1.2 Tasks

NusaDialogue supports two distinct tasks aimed at advancing natural language processing capabilities across underrepresented languages. The first task is Abstractive Dialogue Summarization, inspired by the work of Goo and Chen (2018). This task focuses on generating concise summaries from given conversations, providing a valuable tool for summarizing multi-party discussions, including meetings. NusaDialogue expands on existing efforts in abstractive dialogue summarization by incorporating three underrepresented languages. Notably, the dataset maintains cultural relevance through a meticulous manual annotation process carried out by native speakers of each language.

The second task within NusaDialogue is the Open-domain Dialogue System, building upon the foundational work of Sordoni et al. (2015). In this task, the objective is to generate appropriate responses based on the context provided by the dialogue history. NusaDialogue extends the scope of open-domain dialogue systems to three underrepresented languages, differentiating itself from other multilingual datasets such as XPersona (Lin et al., 2021) by avoiding translation in the annotation process. This ensures that the content remains culturally relevant to each language without compromising linguistic nuances.

3.2 Corpus Collection

3.2.1 Annotator Selection

We conduct corpus construction through human annotation by expert annotators. All expert annotators are native speakers of each target language who have gone through a selection process. In the process of developing data in a local language, a competent and experienced team in the required local language is certainly needed. Annotators play a crucial role in compiling high-quality local language data. Therefore, strict qualifications are required for the candidate annotators who will be recruited. The qualifications include educational background and experience related to language. Annotator candidates must have good knowledge of the language and the sentence structure of the local language they are proficient in, assessed through a selection process involving two tasks: 1) translating several Indonesian sentences into local languages, and 2) writing a paragraph in their local language for specific topics. Additionally, annotators are expected to have resilience in working with a large amount of data, so commitment from annotators is also required.

The recruitment process has successfully gathered a total of 462 annotator candidates for 3 different languages. There are 88 candidates for the

Language	#Data	#Word	#Train	#Valid	#Test
Balinese	10255	3.63M	8205	1025	1025
Buginese	10277	3.68M	8220	1028	1028
Minangkabau	10355	3.70M	8283	1036	1036

Table 2: Statistics of the NusaDialogue corpus.

Balinese language, 174 candidates for the Buginese language, and 200 candidates for the Minangkabau language. Out of a total of 462 applicants, there are 118 candidates, or approximately 25%, who were eligible to participate in the annotation process. Out of that number, only 73 people persevered until the annotation process was completed, while the rest withdrew from the project midway through. The distribution of dialect diversity from the annotators is shown in Table 1.

3.2.2 Annotation Process

Our goal is to collect a diverse set of dialogue-paragraph data that has a large coverage of lexical variations for covering all the languages under study. To maximize the diversity, we first define a wide coverage of topics and subtopics for the dialogue-paragraph annotation. In total we cover 17 topics ranging from general day-to-day conversation such as hobbies, activities, leisure, food and beverages, etc; while also covering a more domain-specific conversation such as history, politics, electronics, science, etc. We further break each topic into multiple subtopics, resulting in a total of 185 subtopics. We list all the topics and subtopics covered in the NusaDialogue corpus in Appendix A.

We conduct dialogue-paragraph writing by instructing the annotators to write a pair of 200-word dialogue and 100-word paragraphs given a certain topic. In paragraph writing, we also define the types of paragraph development from the start. There are 5 types of paragraphs that annotators must develop; (1) description, (2) narration, (3) exposition, (4) argumentation, and (5) persuasion. Determining this type of paragraph development also aims to maximize variations in the use of diction in the corpus. To ensure a high-quality and standardized dialogue-paragraph annotation, we provide a specific guideline during the annotation process. The detailed criteria for writing dialogue-paragraph data are shown in Appendix B.

Throughout the data creation process, we held biweekly meeting evaluation with all annotators. In every meeting, we provide a personal evaluation regarding the data created. The meeting also becomes a forum for annotators to convey issues or constraints during the data creation process (apart from through written documentation that can be accessed together). At its essence, this meeting is aimed at maximizing the quality of the data created and minimizing errors that may occur.

During the annotation process, quality assurance (QA) is also performed with additional human annotators to ensure the data quality. We conduct QA to ensure the data correctness through automatic and manual human validation. The first step taken in the QA process is to check data duplication automatically. Checks were carried out to look for similarities by comparing the string distance between two data points divided by the length of the longest sentences. This yields a similarity information in a range of $[0 \dots 1]$. All data with similarity score ≥ 0.3 were revised by the annotator.

Human validation is carried out to ensure the completeness of the data worksheet components being worked on. Things that are also ensured in this process are the suitability of the data to the topic and subtopic, the similarity of dialogue and paragraph information, the suitability of the type of paragraph being developed, and the rules for good and correct writing. Based on the QA results of the entire data, it is known that less than 10 percent of the data from each corpus needs to undergo revision. The errors that occur vary, from minor errors such as writing errors or missing filling in the worksheet completeness column, to major errors such as the use of Indonesian in the data which still dominates and data duplication.

3.3 Corpus Statistics

We initially aimed to collect 10,000 pairs of dialogue-paragraph, with a total of 3 million words for each language. At the end of the annotation, we collected a slightly larger amount of data that exceeded the initial target, reaching 10,255, 10,277, and 10,355 dialogue-paragraph data for Balinese, Buginese, and Minangkabau, respectively. We then split the data into training, validation, and test sets. The detailed quantity of the NusaDialogue corpus is shown in Table 2.

3.4 Gender Bias on Languages with Non-Gendered Pronouns

To combat the prevalent issue of dataset bias against different genders, we take special care to conduct our annotation process in a gender-balanced manner, striving for an equally distributed

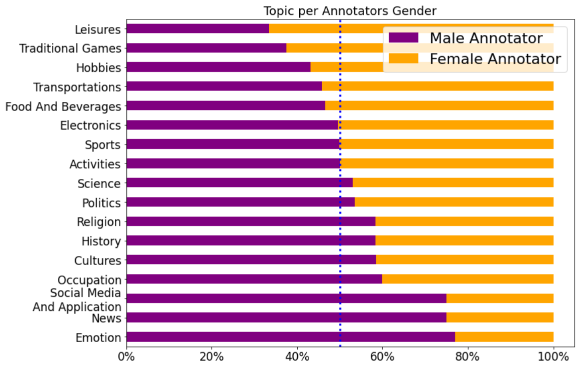

Figure 1: Topic distribution per **annotator gender** for male and female annotators.

representation of genders. Even with these measures in place, we have identified that biases still exist across various topics. This issue of bias in the languages being studied has yet to be sufficiently addressed, making our analysis of these biases all the more crucial to ensuring equitable and non-discriminatory practices in future research. These findings are visualized in Figure 1, which illustrates the topic distribution per annotator gender. The result suggests that there are different tendencies of topic choice between male and female annotators, where male annotators tend to write more dialogue-summarization data on topics such as **social media and application**, **news**, and **emotion**, while female annotators tend to write more dialogue-summarization data on topics such as **leisures**, **traditional games**, and **hobbies**. Therefore, it is evident that addressing these biases in future research involving genders is paramount to ensuring equitable representation and avoidance of discrimination. By understanding annotator biases in NusaDialogue, future research can improve the quality and applicability of language models for these languages by considering the role of annotator biases into account.

Given that NusaDialogue consists of a dialogue between two people, we further analyze the choice of actor for each annotator's gender. The distribution of the gender choice of the actors for each annotator's gender is shown in Figure 2. The result suggests that there is a tendency for annotators to select actors of the same gender on most topics. This phenomenon varies in degree and is topic-dependent. For example, male annotators tend to use female actors when discussing **transportation** and **religion**, then switch to using male actors when the topics of discussion move to **history** and **leisure**. There is also a discrepancy in that female annotators tend to use male actors when discussing **traditional games** and **sports**, and then switch to using female actors when the topics of conversation involve **food and beverages** or **emotions**. Overall, the data indicates that while there is a tendency to select actors of the same gender, but the tendency varies across different topics.

4 Experiment Settings

4.1 Models

For finetuning experiment, we use IndoNLU's (Cahyawijaya et al., 2021) IndoBART and IndoGPT, and mT5-Large (Xue et al., 2021). IndoBART and IndoGPT are language models specifically designed for Indonesian, pre-trained on a dataset comprising 25 GB of text. They utilize the architectures and pre-training objectives of BART (Lewis et al., 2019) and GPT (Brown et al., 2020) respectively. Additionally, mT5 is a multilingual T5 model (Raffel et al., 2020) pre-trained on a new Common Crawl-based dataset covering 101 languages.

In terms of their architectural design, BART, GPT, and mT5 exhibit distinct characteristics that make them uniquely suited for a range of natural language processing tasks. BART adopts an encoder-decoder structure where the encoder processes the input text and the decoder generates the output. This bidirectional nature of the encoder allows for a deep understanding of context, making BART particularly effective for tasks requiring text reconstruction and comprehension. In contrast, GPT, built on a decoder-only architecture, excels in generative tasks, leveraging its unidirectional training to predict subsequent text sequences effectively. mT5, as a multilingual extension of the T5 model, also uses an encoder-decoder framework, but it stands out for its text-to-text approach. This approach reframes all tasks as a conversion from one form of text to another, offering unparalleled flexibility in handling a wide variety of language tasks across multiple languages.

For prompting experiment, we use Llama-2's (Touvron et al., 2023) 13b and 7b variants, Merak-7B-v1 (Ichsan, 2023), Mistral-7B (Jiang et al., 2023) variants, Wizard-Vicuna-13B (Hartford, 2023), bloom-7b1 (Workshop et al., 2023), bloomz-7b1-mt (Muennighoff et al., 2023), gpt-3.5-turbo (OpenAI, 2023), zephyr-7b-alpha (Team, 2023a) and zephyr-7b-beta (Team, 2023b).

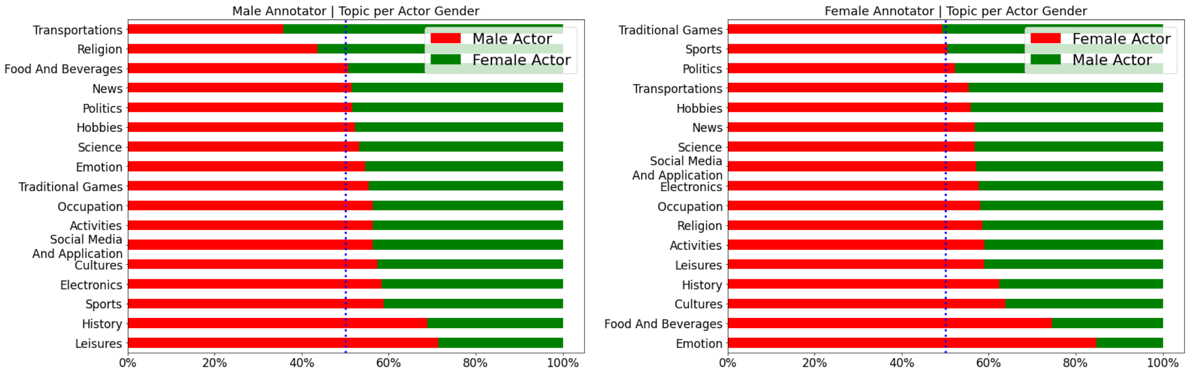

Figure 2: Topic Distribution per **actors gender** for (left) male and (right) female annotators

Lang	Prompt
Id	Simpulkan dialog berikut kedalam 1 paragraf
Id	Gabungkan obrolan di bawah menjadi satu paragraf
En	Summarize the following dialogue into one paragraph

Table 3: The prompts used within our experiments.

4.2 Training and Inference Strategies

Fine-tuned Models In the experiments, we employed Monolingual and Cross-Lingual Training. The Cross-lingual was trained with leave-one-language-out (LOLO) fine-tuning strategy. In the monolingual training setting, each of the three languages in the NusaDialogue corpus (Balinese, Buginese, Minangkabau) is treated as a separate entity. The model is trained and evaluated on the same language. This approach allows for a focused understanding of the nuances and idiosyncrasies of each language. This method can highlight the effectiveness of the models (IndoGPT, IndoBART, and mT5-large) in understanding and generating summaries specific to each language. It can reveal the strengths or weaknesses in dealing with the linguistic features inherent to each language. In the leave-one-language-out (LOLO) setting, the LM is trained in two of the three languages and tested in the unseen language. This cycle is repeated such that each language gets left out in one of the training phases. This strategy assesses the cross-lingual transfer learning ability of the LMs. It is a stringent test of the generalizability of LMs to apply learned concepts across different linguistic contexts.

Prompting Models In our experiments, we engaged in both zero-shot and few-shot prompting, employing the number of few-shot samples (k) of 2. We opted for two variations of Indonesian prompts to assess model performance when prompted in the Indonesian language. Given that the models were predominantly pre-trained using English data, we included another variation of an English prompt (Version 2) to leverage the models' familiarity with the English language. This strategic choice allows for a comparative analysis of how models respond to prompts in both languages. The list of prompts used in our study is shown in Table 3.

4.3 Evaluation

Dialogue-Summarization Benchmark for Underrepresented Languages We develop a dialogue-summarization benchmark from NusaDialogue showcasing the understanding and generation capability of existing LMs and LLMs. For smaller-scale LMs, we conduct fine-tuning to the training data and evaluate on the test data of NusaDialogue, while for LLMs, we evaluate the zero-shot and few-shot generalization capability to these languages through zero-shot and few-shot prompting. For the evaluation metric, we calculate 4 commonly used summarization metrics, i.e., ROUGE1, ROUGE2, ROUGEL, and ROUGELsum. We use the same generation configuration for all models.

Gender Benchmark for Languages with Non-gendered Pronouns We develop the first gender benchmark for languages with non-gendered pronouns using the NusaDialogue corpus. Unlike previous gender benchmark which focuses on gendered-pronoun languages especially English (Havaldar et al., 2023; Yong et al., 2023b), we focus on 3 Austronesian languages, i.e., Minangkabau, Balinese, and Buginese, of which none of them pronominal gender distinctions (Andrew Blust, 2023; Chen and Polinsky, 2019), In this matter, gender bias needs to be detected through other means, such as from the honorific or name of the person.

Models	min		ban		bug	
	R2	RL	R2	RL	R2	RL
Fine-tuning						
IndoNLU IndoBART	0	**45.27**	0	**34.38**	0	**41.87**
IndoNLU IndoGPT	0	12.27	0	12.00	0	14.26
$mT5_{large}$	0	21.48	0	21.06	0	28.43
Zero-shot						
Llama-2-13b-chat-hf	0.59	2.97	0.17	2.84	0.43	2.27
Llama-2-7b-chat-hf	0.21	1.40	0.05	2.00	0.14	1.39
Merak-7B-v1	0.14	1.20	0.02	0.76	0.02	0.70
Mistral-7B-Instruct-v0.1	0.30	2.05	0.03	1.83	0.17	1.68
Wizard-Vicuna-13B	0.11	0.71	0.03	1.36	0.07	0.73
bloomz-7b1-mt	0.31	2.03	0.07	1.66	0.08	1.36
zephyr-7b-alpha	0.21	1.31	0.03	2.03	0.14	1.23
zephyr-7b-beta	0.34	1.84	0.05	1.91	0.11	0.97
gpt-3.5-turbo	3.99	10.82	3.20	12.04	5.83	11.54
Few-shot						
Llama-2-13b-chat-hf	0.88	4.59	1.08	4.85	1.06	3.58
Llama-2-7b-chat-hf	0.29	1.84	0.22	2.25	0.27	1.79
Merak-7B-v1	0.17	1.16	0.07	0.98	0.10	0.90
Mistral-7B-Instruct-v0.1	0.37	3.26	0.15	1.40	0.28	1.43
Wizard-Vicuna-13B	0.00	0.23	0.01	0.36	0.00	0.06
bloomz-7b1-mt	0.15	1.27	0.04	0.92	0.01	0.34
zephyr-7b-alpha	0.24	2.53	0.51	2.50	0.12	1.14
zephyr-7b-beta	0.50	4.08	0.78	2.96	0.20	1.58
gpt-3.5-turbo	5.21	14.45	8.78	21.48	5.65	13.41

Table 4: Overall performance on all tasks in the Nusa-Dialogue benchmark. We report the ROUGE-2 (**R2**) and summarization ROUGEL (**RL**) for the dialogue-summarization evaluation, and ΔPPL for gender bias benchmark for each language under study. The best performances in each section are **bolded**, while the best overall performance is underlined.

In our experiment, we specifically measure gender bias by controlling the names of the speakers in each of the dialogue-summarization data. We create 3 different name lists, i.e., common male names, common female names, and common neutral names (can be both male and female), and we compute the log probability of each dialogue-summarization pair using the models. The higher log probability on female/male names indicates model biases toward the corresponding gender, while the log probability differences between the female and male names indicate the degree of gender bias of a model. For instance, a higher difference in log probability between the female and male names implies that the model has a higher degree of gender bias, and a lower degree of gender bias otherwise. Following Nangia et al. (2020) and Reusens et al. (2023), we ignore the effect of the name when computing the log probability of the sentences to avoid the perplexity bias from generating the corresponding name itself.

model	setting	ban	bug	min
IndoBART-v2	Monolingual	34.38	41.87	45.27
	LOLO	36.97	36.97	41.89
IndoGPT	Monolingual	12.00	14.26	12.27
	LOLO	2.84	3.80	2.92
$mT5_{large}$	Monolingual	21.06	28.43	21.48
	LOLO	15.20	19.83	18.29

Table 5: Monolingual and LOLO results of fine-tuned models on Balinese, Buginese, and Minangkabau.

5 Result and Discussion

5.1 LMs and LLMs Capabilities on Underrepresented Languages

LLM Benchmark for Extremely Low-Resource Languages As shown in Table 4, the fine-tuning model performances are much higher compared to zero-shot and few-shot prompting models. Most zero-shot prompting models yield very low scores, indicating the inability of these models to understand and generate extremely low-resource languages under study. Furthermore, although few-shot can help to improve the performance, the performance is still very low. In terms of open-source LLMs, Zephyr 7B Beta (zephyr-7b-beta) yields the best performance for the 7B parameter models, while LLaMA-2 (Llama-2-13b-chat) yields the highest score for the 13B parameter models. Interestingly, the zero-shot and few-shot performances of ChatGPT (gpt-3.5-turbo) model are comparable to the fine-tuned IndoGPT model, while it fails to outperform both IndoBART and $mT5_{large}$ models. This result indicates that the IndoGPT model is not as well-trained as the other fine-tuned models, while ChatGPT, despite its extremely large scale and closed-source nature, shows a strong and promising prompting capability as an alternative to fully fine-tuned models.

Limited Cross-Lingual Capability We also explore the cross-lingual capability in the languages under study by conducting leave-one-language-out (LOLO) experiments. As shown in Table 5, the cross-lingual performance of all LMs is still much lower compared to the monolingual counterpart, which is especially harmful to Buginese. These results showcase the limited linguistic transferability from Balinese and Minangkabau to Buginese, which aligns with the findings in NusaX (Winata et al., 2023) and InstructAlign (Cahyawijaya et al., 2023d). This also suggests that, despite having

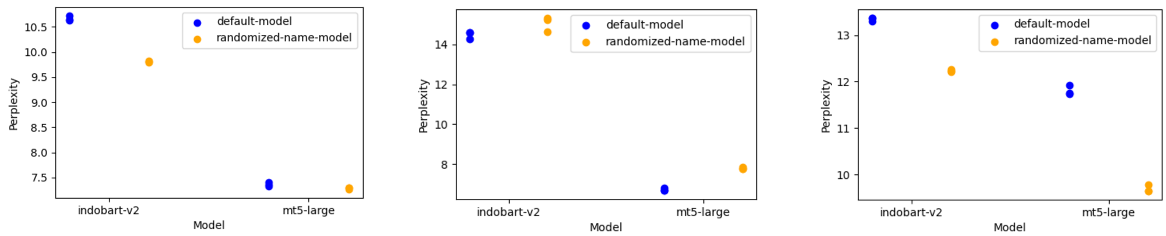

Figure 3: Perplexity score using the original data and augmented data with randomized named for IndoBart-v2 and mT5-large models in **(left)** Minangkabau, **(center)** Balinese, **(right)** Buginese.

more common entities, without proper learning of the related language, the model wouldn't be able to generalize toward relatively distal languages. Interestingly, IndoBART-v2 (Cahyawijaya et al., 2021) and mT5$_{large}$ (Xue et al., 2021) showcase a smaller drop compared to IndoGPT, these two LMs are trained on larger pretraining corpora than IndoGPT. This suggests the effect of larger pretraining corpora and, potentially, different model architecture – with IndoBART-v2 and mT5$_{large}$ utilize the encoder-decoder architecture, while IndoGPT utilizes the decoder-only architecture – in maintaining the cross-linguality of the LMs.

Zero/Few-Shot Generalization of Large Language Models We further evaluate the zero-shot and few-shot generalization capabilities of LLMs in the languages under study. As shown in Table 4, all LLMs achieve a very low ROUGEL performance, way lower compared to the worst fine-tuned LMs (IndoGPT) which achieve 12.27, 12.00, and 14.26 ROUGEL scores on Minangkabau, Balinese, and Buginese, respectively. While gpt-3.5-turbo can outcompete this performance, but it is nowhere near the best fine-tuned LMs, i.e., IndoBART-v2, with ∼35-45 ROUGEL scores on all the languages under study. This result signifies that LLMs are unable to perform dialogue summarization in these languages. This limitation occurs due to the lack of out-of-language and out-of-task generalization ability of the LLMs where neither of them has never seen both the dialogue summarization task during instruction-tuning and the languages under study during both pre-training and instruction-tuning. Furthermore, even with few-shot in-context learning, the dialogue summarization performance does not increase. This showcases that despite having a better understanding of the dialogue summarization task, the limited language capability of the languages under study still becomes the main bottleneck of the dialogue summarization quality.

5.2 Gender Name Bias in LMs and LLMs

Developing language technologies for underrepresented languages carries ethical implications that must be carefully considered. While the goal is to empower these languages and their speakers, there are potential biases and unintended consequences that could arise. One key issue is the lack of diverse and representative data for training language models. The limited availability of data for these languages may lead to biased or inaccurate representations, especially when it comes to gender. Models trained on insufficient data may perpetuate and amplify existing societal biases, such as gender stereotypes or discrimination which potentially brings inappropriate or offensive content that can be harmful or discriminatory, particularly for marginalized communities.

Through NusaDialogue, we take a step further on understanding the potential ethical implications by developing the first gender benchmark for the languages under study. The results of our gender benchmark are shown by the blue dots in Figure 3. We found that both IndoBART and mT5 models achieve low ΔPPL, indicating that both models show a minimal bias in terms of name. To provide supporting evidence that the model has only minimal bias, we introduce a simple method for gender debiasing by swapping the actor name in the training data. The results are shown in the yellow dots in Figure 3. In general, we observe no significant difference in terms of ΔPPL over different experiments in all languages, indicating that the original IndoBART and mT5$_{large}$ models have minimal bias towards different local names in all three languages. We conjecture that this may happen due to the limited amount of representation on these languages.

6 Conclusion

We introduce NusaDialogue, the first high-quality dialogue summarization corpus covering three extremely low-resource languages: Minangkabau (min), Balinese (ban), and Buginese (bug). NusaDialogue covers a diverse set of topics from general day-to-day conversation to specific topics such as science, history, and politics. Using NusaDialogue, we showcase that, despite having non-gendered pronouns, annotators still reflect gender bias in terms of role and topic selection which is propagated through person names and courtesy titles. Furthermore, we develop the first dialogue-summarization benchmark for these languages, showcasing the inability of LLMs to generalize to these languages. Lastly, we demonstrate a gender benchmark which showcases that LLMs do not have name bias o the languages under study due to the lack of representation of these languages.

Limitations

Language Coverage Due to the difficulties of finding the suitable annotators for other languages, we only cover three underrepresented languages in the Malayo-Polynesian language family: Minangkabau, Balinese, and Buginese. We encourage future work to address this limitation in future by expanding the language coverage and collaborating with a more diverse range of annotators.

Task Coverage Despite there is various type of language generation tasks, in this work only focus on the dialogue summarization task. Although prior works (Cahyawijaya et al., 2023a,c; Lovenia et al., 2024) have also explored other tasks such as machine translation, sentiment analysis, emotion recognition, etc, there is still a huge gap between language evaluation on these languages and high-resource languages, e.g., English, Chinese, French, etc. This highlights the need for further research to ensure an extensive evaluation of language generation tasks for a wider range of languages.

Ethics Statement

In the process of defining topics of NusaDialogue, several topics have the potential to cause opinion bias among annotators. These topics are usually related to emotions, for instance, liking or disliking something. It should be understood that this is the annotator's subjectivity and has nothing to do with the organization's values.

Acknowledgement

This research work is funded and supported by The Deutsche Gesellschaft für Internationale Zusammenarbeit (GIZ) GmbH and FAIR Forward - Artificial Intelligence for all. We thank Direktorat Jenderal Pendidikan Tinggi, Riset, dan Teknologi Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi (Ditjen DIKTI) for providing the computing resources for this project.

References

David Adelani, Hannah Liu, Xiaoyu Shen, Nikita Vassilyev, Jesujoba Alabi, Yanke Mao, Haonan Gao, and En-Shiun Lee. 2024. SIB-200: A simple, inclusive, and big evaluation dataset for topic classification in 200+ languages and dialects. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 226–245, St. Julian's, Malta. Association for Computational Linguistics.

David Adelani, Graham Neubig, Sebastian Ruder, Shruti Rijhwani, Michael Beukman, Chester Palen-Michel, Constantine Lignos, Jesujoba Alabi, Shamsuddeen Muhammad, Peter Nabende, Cheikh M. Bamba Dione, Andiswa Bukula, Rooweither Mabuya, Bonaventure F. P. Dossou, Blessing Sibanda, Happy Buzaaba, Jonathan Mukiibi, Godson Kalipe, Derguene Mbaye, Amelia Taylor, Fatoumata Kabore, Chris Chinenye Emezue, Anuoluwapo Aremu, Perez Ogayo, Catherine Gitau, Edwin Munkoh-Buabeng, Victoire Memdjokam Koagne, Allahsera Auguste Tapo, Tebogo Macucwa, Vukosi Marivate, Mboning Tchiaze Elvis, Tajuddeen Gwadabe, Tosin Adewumi, Orevaoghene Ahia, and Joyce Nakatumba-Nabende. 2022. MasakhaNER 2.0: Africa-centric transfer learning for named entity recognition. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4488–4508, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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 Muhammad, Chris Chinenye Emezue, Joyce Nakatumba-Nabende, Perez Ogayo, Anuoluwapo Aremu, Catherine Gitau, Derguene Mbaye, Jesujoba Alabi, Seid Muhie Yimam, Tajuddeen 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.

David Ifeoluwa Adelani, Marek Masiak, Israel Abebe Azime, Jesujoba Alabi, Atnafu Lambebo Tonja, Christine Mwase, Odunayo Ogundepo, Bonaventure F. P. Dossou, Akintunde Oladipo, Doreen Nixdorf, Chris Chinenye Emezue, Sana Al-azzawi, Blessing Sibanda, Davis David, Lolwethu Ndolela, Jonathan Mukiibi, Tunde Ajayi, Tatiana Moteu, Brian Odhiambo, Abraham Owodunni, Nnaemeka Obiefuna, Muhidin Mohamed, Shamsuddeen Hassan Muhammad, Teshome Mulugeta Ababu, Saheed Abdullahi Salahudeen, Mesay Gemeda Yigezu, Tajuddeen Gwadabe, Idris Abdulmumin, Mahlet Taye, Oluwabusayo Awoyomi, Iyanuoluwa Shode, Tolulope Adelani, Habiba Abdulganiyu, Abdul-Hakeem Omotayo, Adetola Adeeko, Abeeb Afolabi, Anuoluwapo Aremu, Olanrewaju Samuel, Clemencia Siro, Wangari Kimotho, Onyekachi Ogbu, Chinedu Mbonu, Chiamaka Chukwuneke, Samuel Fanijo, Jessica Ojo, Oyinkansola Awosan, Tadesse Kebede, Toadoum Sari Sakayo, Pamela Nyatsine, Freedmore Sidume, Oreen Yousuf, Mardiyyah Oduwole, Kanda Tshinu, Ussen Kimanuka, Thina Diko, Siyanda Nxakama, Sinodos Nigusse, Abdulmejid Johar, Shafie Mohamed, Fuad Mire Hassan, Moges Ahmed Mehamed, Evrard Ngabire, Jules Jules, Ivan Ssenkungu, and Pontus Stenetorp. 2023. MasakhaNEWS: News topic classification for African languages. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 144–159, Nusa Dua, Bali. Association for Computational Linguistics.

Muhammad Farid Adilazuarda, Samuel Cahyawijaya, Alham Fikri Aji, Genta Indra Winata, and Ayu Purwarianti. 2024. Lingualchemy: Fusing typological and geographical elements for unseen language generalization.

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

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

Robert Andrew Blust. 2023. Austronesian languages.

Rohan Anil, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, Eric Chu, Jonathan H. Clark, Laurent El Shafey, Yanping Huang, Kathy Meier-Hellstern, Gaurav Mishra, Erica Moreira, Mark Omernick, Kevin Robinson, Sebastian Ruder, Yi Tay, Kefan Xiao, Yuanzhong Xu, Yujing Zhang, Gustavo Hernandez Abrego, Junwhan Ahn, Jacob Austin, Paul Barham, Jan Botha, James Bradbury, Siddhartha Brahma, Kevin Brooks, Michele Catasta, Yong Cheng, Colin Cherry, Christopher A. Choquette-Choo, Aakanksha Chowdhery, Clément Crepy, Shachi Dave, Mostafa Dehghani, Sunipa Dev, Jacob Devlin, Mark Díaz, Nan Du, Ethan Dyer, Vlad Feinberg, Fangxiaoyu Feng, Vlad Fienber, Markus Freitag, Xavier Garcia, Sebastian Gehrmann, Lucas Gonzalez, Guy Gur-Ari, Steven Hand, Hadi Hashemi, Le Hou, Joshua Howland, Andrea Hu, Jeffrey Hui, Jeremy Hurwitz, Michael Isard, Abe Ittycheriah, Matthew Jagielski, Wenhao Jia, Kathleen Kenealy, Maxim Krikun, Sneha Kudugunta, Chang Lan, Katherine Lee, Benjamin Lee, Eric Li, Music Li, Wei Li, YaGuang Li, Jian Li, Hyeontaek Lim, Hanzhao Lin, Zhongtao Liu, Frederick Liu, Marcello Maggioni, Aroma Mahendru, Joshua Maynez, Vedant Misra, Maysam Moussalem, Zachary Nado, John Nham, Eric Ni, Andrew Nystrom, Alicia Parrish, Marie Pellat, Martin Polacek, Alex Polozov, Reiner Pope, Siyuan Qiao, Emily Reif, Bryan Richter, Parker Riley, Alex Castro Ros, Aurko Roy, Brennan Saeta, Rajkumar Samuel, Renee Shelby, Ambrose Slone, Daniel Smilkov, David R. So, Daniel Sohn, Simon Tokumine, Dasha Valter, Vijay Vasudevan, Kiran Vodrahalli, Xuezhi Wang, Pidong Wang, Zirui Wang, Tao Wang, John Wieting, Yuhuai Wu, Kelvin Xu, Yunhan Xu, Linting Xue, Pengcheng Yin, Jiahui Yu, Qiao Zhang, Steven Zheng, Ce Zheng, Weikang Zhou, Denny Zhou, Slav Petrov, and Yonghui Wu. 2023. Palm 2 technical report.

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

Viraat Aryabumi, John Dang, Dwarak Talupuru, Saurabh Dash, David Cairuz, Hangyu Lin, Bharat Venkitesh, Madeline Smith, Jon Ander Campos, Yi Chern Tan, Kelly Marchisio, Max Bartolo, Sebastian Ruder, Acyr Locatelli, Julia Kreutzer, Nick Frosst, Aidan Gomez, Phil Blunsom, Marzieh Fadaee, Ahmet Üstün, and Sara Hooker. 2024. Aya 23: Open weight releases to further multilingual progress.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023. A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 675–718, Nusa Dua, Bali. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Samuel Cahyawijaya. 2024. Llm for everyone: Representing the underrepresented in large language models.

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

Samuel Cahyawijaya, Holy Lovenia, Alham Fikri Aji, Genta Indra Winata, Bryan Wilie, Rahmad Mahendra, Christian Wibisono, Ade Romadhony, Karissa Vincentio, Fajri Koto, Jennifer Santoso, David Moeljadi, Cahya Wirawan, Frederikus Hudi, Ivan Halim Parmonangan, Ika Alfina, Muhammad Satrio Wicaksono, Ilham Firdausi Putra, Samsul Rahmadani, Yulianti Oenang, Ali Akbar Septiandri, James Jaya, Kaustubh D. Dhole, Arie Ardiyanti Suryani, Rifki Afina Putri, Dan Su, Keith Stevens, Made Nindyatama Nityasya, Muhammad Farid Adilazuarda, Ryan Ignatius, Ryandito Diandaru, Tiezheng Yu, Vito Ghifari, Wenliang Dai, Yan Xu, Dyah Damapuspita, Cuk Tho, Ichwanul Muslim Karo Karo, Tirana Noor Fatyanosa, Ziwei Ji, Pascale Fung, Graham Neubig, Timothy Baldwin, Sebastian Ruder, Herry Sujaini, Sakriani Sakti, and Ayu Purwarianti. 2023b. Nusacrowd: Open source initiative for indonesian nlp resources.

Samuel Cahyawijaya, Holy Lovenia, and Pascale Fung. 2024a. LLMs are few-shot in-context low-resource language learners. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 405–433, Mexico City, Mexico. Association for Computational Linguistics.

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

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

Samuel Cahyawijaya, Holy Lovenia, Tiezheng Yu, Willy Chung, and Pascale Fung. 2023d. InstructAlign: High-and-low resource language alignment via continual crosslingual instruction tuning. In *Proceedings of the First Workshop in South East Asian Language Processing*, pages 55–78, Nusa Dua, Bali, Indonesia. Association for Computational Linguistics.

Samuel Cahyawijaya, Genta Indra Winata, Bryan Wilie, Karissa Vincentio, Xiaohong Li, Adhiguna Kuncoro, Sebastian Ruder, Zhi Yuan Lim, Syafri Bahar, Masayu Leylia Khodra, Ayu Purwarianti, and Pascale Fung. 2021. Indonlg: Benchmark and resources for evaluating indonesian natural language generation.

Samuel Cahyawijaya, Ruochen Zhang, Holy Lovenia, Jan Christian Blaise Cruz, Elisa Gilbert, Hiroki Nomoto, and Alham Fikri Aji. 2024c. Thank you, stingray: Multilingual large language models can not (yet) disambiguate cross-lingual word sense.

Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Victoria Chen and Maria Polinsky. 2019. Gender distinctions and classifiers in austronesian languages.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. Palm: Scaling language modeling with pathways.

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

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

Abteen Ebrahimi, Manuel Mager, Arturo Oncevay, Vishrav Chaudhary, Luis Chiruzzo, Angela Fan, John Ortega, Ricardo Ramos, Annette Rios, Ivan Meza-Ruiz, Gustavo A. Giménez-Lugo, Elisabeth Mager, Graham Neubig, Alexis Palmer, Rolando Coto-Solano, Ngoc Thang Vu, and Katharina Kann. 2022. Americasnli: Evaluating zero-shot natural language understanding of pretrained multilingual models in truly low-resource languages.

Chih-Wen Goo and Yun-Nung Chen. 2018. Abstractive dialogue summarization with sentence-gated modeling optimized by dialogue acts.

Wei Guo and Aylin Caliskan. 2021. Detecting emergent intersectional biases: Contextualized word embeddings contain a distribution of human-like biases. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '21, page 122–133, New York, NY, USA. Association for Computing Machinery.

Eric Hartford. 2023. Wizard-vicuna-13b-uncensored-gptq. `https://huggingface.co/TheBloke/Wizard-Vicuna-13B-Uncensored-GPTQ`.

Shreya Havaldar, Bhumika Singhal, Sunny Rai, Langchen Liu, Sharath Chandra Guntuku, and Lyle Ungar. 2023. Multilingual language models are not multicultural: A case study in emotion. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 202–214, Toronto, Canada. Association for Computational Linguistics.

Muhammad Ichsan. 2023. Merak-7b: The llm for bahasa indonesia. *Hugging Face Repository*.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431, Valencia, Spain. Association for Computational Linguistics.

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

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

Simran Khanuja, Sebastian Ruder, and Partha Talukdar. 2023. Evaluating the diversity, equity, and inclusion of NLP technology: A case study for Indian languages. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1763–1777, Dubrovnik, Croatia. Association for Computational Linguistics.

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

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

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

Fajri Koto, Afshin Rahimi, Jey Han Lau, and Timothy Baldwin. 2020. Indolem and indobert: A benchmark dataset and pre-trained language model for indonesian nlp.

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.

Viet Lai, Nghia Ngo, Amir Pouran Ben Veyseh, Hieu Man, Franck Dernoncourt, Trung Bui, and Thien Nguyen. 2023. ChatGPT beyond English: Towards a comprehensive evaluation of large language models in multilingual learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13171–13189, Singapore. Association for Computational Linguistics.

Hugo Laurençon, Lucile Saulnier, Thomas Wang, Christopher Akiki, Albert Villanova del Moral, Teven Le Scao, Leandro Von Werra, Chenghao Mou, Eduardo González Ponferrada, Huu Nguyen, Jörg Frohberg, Mario Šaško, Quentin Lhoest, Angelina McMillan-Major, Gerard Dupont, Stella Biderman, Anna Rogers, Loubna Ben allal, Francesco De Toni, Giada Pistilli, Olivier Nguyen, Somaieh Nikpoor, Maraim Masoud, Pierre Colombo, Javier de la Rosa, Paulo Villegas, Tristan Thrush, Shayne Longpre, Sebastian Nagel, Leon Weber, Manuel Muñoz, Jian Zhu, Daniel Van Strien, Zaid Alyafeai, Khalid Almubarak, Minh Chien Vu, Itziar Gonzalez-Dios, Aitor Soroa, Kyle Lo, Manan Dey, Pedro Ortiz Suarez, Aaron Gokaslan, Shamik Bose, David Adelani, Long Phan, Hieu Tran, Ian Yu, Suhas Pai, Jenny Chim, Violette Lepercq, Suzana Ilic, Margaret Mitchell, Sasha Alexandra Luccioni, and Yacine Jernite. 2023. The bigscience roots corpus: A 1.6tb composite multilingual dataset.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension.

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.

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, Online. Association for Computational Linguistics.

Holy Lovenia, Rahmad Mahendra, Salsabil Akbar, Lester James Miranda, Jennifer Santoso, Elyanah Aco, Akhdan Fadhilah, Jonibek Mansurov, Joseph Marvin Imperial, Onno Kampman, Joel Moniz, Muhammad Habibi, Frederikus Hudi, Jann Montalan, Ryan Hadiwijaya, Joanito Lopo, William Nixon, Börje Karlsson, James Jaya, Ryandito Diandaru, Yuze Gao, Patrick Irawan, Bin Wang, Jan Christian Blaise Cruz, Chenxi Whitehouse, Ivan Parmonangan, Maria Khelli, Wenyu Zhang, Lucky Susanto, Reynard Ryanda, Sonny Hermawan, Dan Velasco, Muhammad Kautsar, Willy Hendria, Yasmin Moslem, Noah Flynn, Muhammad Adilazuarda, Haochen Li, Johanes Lee, R. Damanhuri, Shuo Sun, Muhammad Qorib, Amirbek Djanibekov, Wei Qi Leong, Quyet V. Do, Niklas Muennighoff, Tanrada Pansuwan, Ilham Firdausi Putra, Yan Xu, Tai Chia, Ayu Purwarianti, Sebastian Ruder, William Tjhi, Peerat Limkonchotiwat, Alham Aji, Sedrick Keh, Genta Winata, Ruochen Zhang, Fajri Koto, Zheng Xin Yong, and Samuel Cahyawijaya. 2024. SEACrowd: A multilingual multimodal data hub and benchmark suite for Southeast Asian languages. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5155–5203, Miami, Florida, USA. 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, Jeju Island, Korea. Association for Computational Linguistics.

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. 2023. Crosslingual generalization through multitask finetuning.

Shamsuddeen Muhammad, Idris Abdulmumin, Abinew Ayele, Nedjma Ousidhoum, David Adelani, Seid Yimam, Ibrahim Ahmad, Meriem Beloucif, Saif Mohammad, Sebastian Ruder, Oumaima Hourrane, Alipio Jorge, Pavel Brazdil, Felermino Ali, Davis David, Salomey Osei, Bello Shehu-Bello, Falalu Lawan, Tajuddeen Gwadabe, Samuel Rutunda, Tadesse Destaw Belay, Wendimu Messelle, Hailu Balcha, Sisay Chala, Hagos Gebremichael, Bernard

Opoku, and Stephen Arthur. 2023. AfriSenti: A Twitter sentiment analysis benchmark for African languages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13968–13981, Singapore. Association for Computational Linguistics.

Shuyo Nakatani. 2011. Language detection library for java.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020. CrowS-pairs: A challenge dataset for measuring social biases in masked language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1953–1967, Online. Association for Computational Linguistics.

Odunayo Ogundepo, Tajuddeen Gwadabe, Clara Rivera, Jonathan Clark, Sebastian Ruder, David Adelani, Bonaventure Dossou, Abdou Diop, Claytone Sikasote, Gilles Hacheme, Happy Buzaaba, Ignatius Ezeani, Rooweither Mabuya, Salomey Osei, Chris Emezue, Albert Kahira, Shamsuddeen Muhammad, Akintunde Oladipo, Abraham Owodunni, Atnafu Tonja, Iyanuoluwa Shode, Akari Asai, Anuoluwapo Aremu, Ayodele Awokoya, Bernard Opoku, Chiamaka Chukwuneke, Christine Mwase, Clemencia Siro, Stephen Arthur, Tunde Ajayi, Verrah Otiende, Andre Rubungo, Boyd Sinkala, Daniel Ajisafe, Emeka Onwuegbuzia, Falalu Lawan, Ibrahim Ahmad, Jesujoba Alabi, Chinedu Mbonu, Mofetoluwa Adeyemi, Mofya Phiri, Orevaoghene Ahia, Ruqayya Iro, and Sonia Adhiambo. 2023. Cross-lingual open-retrieval question answering for African languages. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14957–14972, Singapore. Association for Computational Linguistics.

OpenAI. 2023. Gpt-3.5-turbo. `https://platform.openai.com/docs/models/gpt-3-5-turbo`.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao

Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report.

Hadas Orgad, Seraphina Goldfarb-Tarrant, and Yonatan Belinkov. 2022. How gender debiasing affects internal model representations, and why it matters. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2602–2628, Seattle, United States. Association for Computational Linguistics.

Sungjoon Park, Jihyung Moon, Sungdong Kim, Won Ik Cho, Jiyoon Han, Jangwon Park, Chisung Song, Junseong Kim, Yongsook 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.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Manon Reusens, Philipp Borchert, Margot Mieskes, Jochen De Weerdt, and Bart Baesens. 2023. Investigating bias in multilingual language models: Cross-lingual transfer of debiasing techniques. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2887–2896, Singapore. Association for Computational Linguistics.

David Romero, Chenyang Lyu, Haryo Akbarianto Wibowo, Teresa Lynn, Injy Hamed, Aditya Nanda Kishore, Aishik Mandal, Alina Dragonetti, Artem Abzaliev, Atnafu Lambebo Tonja, Bontu Fufa Balcha, Chenxi Whitehouse, Christian Salamea, Dan John Velasco, David Ifeoluwa Adelani, David Le Meur, Emilio Villa-Cueva, Fajri Koto, Fauzan Farooqui, Frederico Belcavello, Ganzorig Batnasan, Gisela Vallejo, Grainne Caulfield, Guido Ivetta, Haiyue Song, Henok Biadglign Ademtew, Hernán Maina, Holy Lovenia, Israel Abebe Azime, Jan Christian Blaise Cruz, Jay Gala, Jiahui Geng, Jesus-German Ortiz-Barajas, Jinheon Baek, Jocelyn Dunstan, Laura Alonso Alemany, Kumaranage Ravindu Yasas Nagasinghe, Luciana Benotti, Luis Fernando D'Haro, Marcelo Viridiano, Marcos Estecha-Garitagoitia, Maria Camila Buitrago Cabrera, Mario Rodríguez-Cantelar, Mélanie Jouitteau, Mihail Mihaylov, Mohamed Fazli Mohamed Imam, Muhammad Farid Adilazuarda, Munkhjargal Gochoo, Munkh-Erdene Otgonbold, Naome Etori, Olivier Niyomugisha, Paula Mónica Silva, Pranjal Chitale, Raj Dabre, Rendi Chevi, Ruochen Zhang, Ryandito Diandaru, Samuel Cahyawijaya, Santiago Góngora, Soyeong Jeong, Sukannya Purkayastha, Tatsuki Kuribayashi, Teresa Clifford, Thanmay Jayakumar, Tiago Timponi Torrent, Toqeer Ehsan, Vladimir Araujo, Yova Kementchedjhieva, Zara Burzo, Zheng Wei Lim, Zheng Xin Yong, Oana Ignat, Joan Nwatu, Rada Mihalcea, Thamar Solorio, and Alham Fikri Aji. 2024. Cvqa: Culturally-diverse multilingual visual question answering benchmark.

Alex Salcianu, Andy Golding, Anton Bakalov, Chris Alberti, Daniel Andor, David Weiss, Emily Pitler, Greg Coppola, Jason Riesa, Kuzman Ganchev, Michael Ringgaard, Nan Hua, Ryan McDonald, Slav Petrov, Stefan Istrate, and Terry Koo. 2020. Compact language detector v3 (cld3).

Aleix Sant, Carlos Escolano, Audrey Mash, Francesca De Luca Fornaciari, and Maite Melero. 2024. The power of prompts: Evaluating and mitigating gender bias in MT with LLMs. In *Proceedings of the 5th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 94–139, Bangkok, Thailand. Association for Computational Linguistics.

Shivalika Singh, Freddie Vargus, Daniel D'souza, Börje Karlsson, Abinaya Mahendiran, Wei-Yin Ko, Herumb Shandilya, Jay Patel, Deividas Mataciunas, Laura O'Mahony, Mike Zhang, Ramith Hettiarachchi, Joseph Wilson, Marina Machado, Luisa Moura, Dominik Krzemiński, Hakimeh Fadaei, Irem Ergun, Ifeoma Okoh, Aisha Alaagib, Oshan Mudannayake, Zaid Alyafeai, Vu Chien, Sebastian Ruder, Surya Guthikonda, Emad Alghamdi, Sebastian Gehrmann, Niklas Muennighoff, Max Bartolo, Julia Kreutzer, Ahmet Üstün, Marzieh Fadaee, and Sara Hooker. 2024. Aya dataset: An open-access collection for multilingual instruction tuning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11521–11567, Bangkok, Thailand. Association for Computational Linguistics.

Richard Sites. 2013. Compact language detector v2 (cld2).

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

Yueqi Song, Simran Khanuja, Pengfei Liu, Fahim Faisal, Alissa Ostapenko, Genta Winata, Alham Fikri Aji, Samuel Cahyawijaya, Yulia Tsvetkov, Antonios Anastasopoulos, and Graham Neubig. 2023. Global-Bench: A benchmark for global progress in natural language processing. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14157–14171, Singapore. Association for Computational Linguistics.

Alessandro Sordoni, Michel Galley, Michael Auli, Chris Brockett, Yangfeng Ji, Margaret Mitchell, Jian-Yun Nie, Jianfeng Gao, and Bill Dolan. 2015. A neural network approach to context-sensitive generation of conversational responses. In *Proceedings of the 2015 Conference of the North American Chapter of the*

Association for Computational Linguistics: Human Language Technologies, pages 196–205, Denver, Colorado. Association for Computational Linguistics.

Ian Stewart and Rada Mihalcea. 2024. Whose wife is it anyway? assessing bias against same-gender relationships in machine translation. In *Proceedings of the 5th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 365–375, Bangkok, Thailand. Association for Computational Linguistics.

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

Huggingface Team. 2023a. zephyr-7b-alpha. `https://huggingface.co/HuggingFaceH4/zephyr-7b-alpha`.

Huggingface Team. 2023b. zephyr-7b-beta. `https://huggingface.co/HuggingFaceH4/zephyr-7b-beta`.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Norawit Urailertprasert, Peerat Limkonchotiwat, Supasorn Suwajanakorn, and Sarana Nutanong. 2024. SEA-VQA: Southeast Asian cultural context dataset for visual question answering. In *Proceedings of the 3rd Workshop on Advances in Language and Vision Research (ALVR)*, pages 173–185, Bangkok, Thailand. Association for Computational Linguistics.

Ahmet Üstün, Viraat Aryabumi, Zheng Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, Freddie Vargus, Phil Blunsom, Shayne Longpre, Niklas Muennighoff, Marzieh Fadaee, Julia Kreutzer, and Sara Hooker. 2024. Aya model: An instruction finetuned open-access multilingual language model. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15894–15939, Bangkok, Thailand. Association for Computational Linguistics.

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

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

Genta Indra Winata, Alham Fikri Aji, Samuel Cahyawijaya, Rahmad Mahendra, Fajri Koto, Ade Romadhony, Kemal Kurniawan, David Moeljadi, Radityo Eko Prasojo, Pascale Fung, Timothy Baldwin, Jey Han Lau, Rico Sennrich, and Sebastian Ruder. 2023. NusaX: Multilingual parallel sentiment dataset for 10 Indonesian local languages. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 815–834, Dubrovnik, Croatia. Association for Computational Linguistics.

Genta Indra Winata, Frederikus Hudi, Patrick Amadeus Irawan, David Anugraha, Rifki Afina Putri, Yutong Wang, Adam Nohejl, Ubaidillah Ariq Prathama, Nedjma Ousidhoum, Afifa Amriani, Anar Rzayev, Anirban Das, Ashmari Pramodya, Aulia Adila, Bryan Wilie, Candy Olivia Mawalim, Ching Lam Cheng, Daud Abolade, Emmanuele Chersoni, Enrico Santus, Fariz Ikhwantri, Garry Kuwanto, Hanyang Zhao, Haryo Akbarianto Wibowo, Holy Lovenia, Jan Christian Blaise Cruz, Jan Wira Gotama Putra, Junho Myung, Lucky Susanto, Maria Angelica Riera Machin, Marina Zhukova, Michael Anugraha, Muhammad Farid Adilazuarda, Natasha Santosa, Peerat Limkonchotiwat, Raj Dabre, Rio Alexander Audino, Samuel Cahyawijaya, Shi-Xiong Zhang, Stephanie Yulia Salim, Yi Zhou, Yinxuan Gui, David Ifeoluwa Adelani, En-Shiun Annie Lee, Shogo Okada, Ayu Purwarianti, Alham Fikri Aji, Taro Watanabe, Derry Tanti Wijaya, Alice Oh, and Chong-Wah Ngo. 2024. Worldcuisines: A massive-scale benchmark for multilingual and multicultural visual question answering on global cuisines.

BigScience Workshop, :, 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, Davut Emre Taşar, 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, Shachar Mirkin, 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, Daniel McDuff, 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, 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. 2023. Bloom: A 176b-parameter open-access multilingual language model.

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

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

Zheng Xin Yong, Ruochen Zhang, Jessica Forde, Skyler Wang, Arjun Subramonian, Holy Lovenia, Samuel Cahyawijaya, Genta Winata, Lintang Sutawika, Jan Christian Blaise Cruz, Yin Lin Tan, Long Phan, Long Phan, Rowena Garcia, Thamar Solorio, and Alham Aji. 2023b. Prompting multilingual large language models to generate code-mixed texts: The case of south East Asian languages. In *Proceedings of the 6th Workshop on Computational Approaches to Linguistic Code-Switching*, pages 43–63, Singapore. Association for Computational Linguistics.

A List of Topic in NusaDialogue

B Annotation Criteria

Topic	Subtopic
Activities	Gardening, Roof Fixing, Shopping, Debating, Fish Tank Cleaning, Others, Helping Others, House Painting, Child Parenting, Working, House Cleaning, Car Washing, Reading
Cultures	Traditional Food, Folk Songs, Traditional Houses, Folklore, Traditional Ceremonies, Traditional Souvenirs
Electronics	Electronic Store, Beauty Electronics, Office Electronics, Carpentry Electronics, Communication Electronics, Household Electronics
Emotion	Angry, Disguised, Fear, Confused, Curious, Sad, Jealous, Embarrassed, Excited, Happy, Surprising, Trust, Hate, Danger, Disappointed
Food And Beverages	Favorite Drinks, Disliked Food, Disliked Drinks, Disliked Snacks, Cooking Recipe, Cooking Utensils And Electronics, Favorite Food, Favorite Snacks, Restaurant Review
History	Historical Incident, Historic Buildings In The World, National/Regional Heroes, Origin Story
Hobbies	Fishing, Motorcycle Touring, Sewing, Hunting, Others, Hiking, Make Up, Journaling, Watching Movies, Dancing, Reading, Vehicle Modification, Playing Instrument
Leisures	Tourist Attraction, Popular/Viral Tourist Spot, Online Games, Holidays Tips, Traveling Application, Holidays Experiences, Natural Attraction
News	Online News Portal, Viral News, Magazine, Newspaper
Occupation	Secretary, Artist, Nurse, Technician, Trader, Doctor, Others, Security, Pilot, Teacher, Maid, Police, Florist
Politics	Liked Political Figures, Disliked Political Parties, Liked Political Parties, Pemilu, Political Terms/Ideologies, Election
Religion	Religious Holidays/Ceremonies, Routine Worship, Stories In The Scriptures, Religious Terms, House Of Worship
Science	A Scientific Experiment At School, Favorite Subject At School, Energy Sources, Favorite Teacher, Disliked Subject At School, Inventions, Environmental Issues, Inventors Or Scientist
Social Media	Dating Application, Learning/Educational Application, Streaming App, Editing App, Blogging Platforms
Sports	Cycling, Swimming, Yoga, Zumba, Others, Chess, Pole Dance, Badminton, Soccer, Ballet, Motorcycle/Car Racing, Boxing, Running/Jogging
Traditional Games	Cooking/House Games, Congklak, Knucklebones, Marbles, Others, Dragon Snake, Hide And Seek, Kite, Hopscotch, Yoyo, Rubber/Rope Jump, Tamiya, Tug Of War
Transportations	Water Transportation, Land Transportation, Public Transportation Experience, Online Transportation, Vehicle Car Tips, Air Transportation, Traditional Transportation, Private Transportation Recommendation

Table 6: The list of all topics and subtopics used during the annotation process of the NusaDialogue corpus.

Dialogue	Paragraph
Dialogue consists of two speakers	Paragraph follows the topic of the corresponding dialogue
Each speaker has >5 conversation turns	Paragraph covers all the important information in the dialogue
Dialogue focuses on a given conversation topic	Paragraph follows a specified rhetoric mode
Dialogue consists of $\geq$200 words.	Paragraph consists of $\geq$100 words.

Table 7: The annotation criteria for writing the dialogue-paragraph dataset in NusaDialogue.