NusaX: Multilingual Parallel Sentiment Dataset for 10 Indonesian Local Languages

Genta Indra Winata[1*], **Alham Fikri Aji**[2*], **Samuel Cahyawijaya**[3*], **Rahmad Mahendra**[4,5*], **Fajri Koto**[2,6*], **Ade Romadhony**[5,7*], **Kemal Kurniawan**[5,6*], **David Moeljadi**[8], **Radityo Eko Prasojo**[9], **Pascale Fung**[3], **Timothy Baldwin**[2,6], **Jey Han Lau**[6], **Rico Sennrich**[10], **Sebastian Ruder**[11]

[1]Bloomberg [2]MBZUAI [3]HKUST [4]Universitas Indonesia [5]INACL
[6]The University of Melbourne [7]Telkom University [8]Kanda University of International Studies
[9]Kata.ai [10]University of Zurich [11]Google Research

Abstract

Natural language processing (NLP) has significant impact on society via technologies such as machine translation and search engines. Despite its success, NLP technology is only widely available for high-resource languages such as English and Mandarin Chinese, and remains inaccessible to many languages due to the unavailability of data resources and benchmarks. In this work, we focus on developing resources for languages of Indonesia. Despite being the second most linguistically-diverse country, most languages in Indonesia are categorized as endangered and some are even extinct. We develop the first-ever parallel resource for 10 low-resource languages in Indonesia. Our resource includes sentiment and machine translation datasets, and bilingual lexicons. We provide extensive analysis, and describe challenges for creating such resources. Our hope is that this work will spark more NLP research on Indonesian and other underrepresented languages.

1 Introduction

Indonesia is one of the most populous and linguistically-diverse countries in the world, with more than 700 languages spoken across the country (Aji et al., 2022; Eberhard et al., 2021). However, while many of these languages are spoken by millions of people they have received little attention from the NLP community. There are very few public datasets, preventing the global research community from exploring these languages. To this end, we introduce **NusaX**,[1] a high-quality multilingual parallel corpus that covers 10 local languages from Indonesia: Acehnese, Balinese, Banjarese, Buginese, Madurese, Minangkabau, Javanese, Ngaju, Sundanese, and Toba Batak.

The **NusaX** dataset was created by translating SmSA (Purwarianti and Crisdayanti, 2019) — an existing Indonesian sentiment analysis dataset containing comments and reviews from the IndoNLU benchmark (Wilie et al., 2020) — using competent bilingual speakers, coupled with additional human-assisted quality assurance. Sentiment analysis is one of the most popular NLP tasks, and has been explored in many applications in Indonesia, including presidential elections (Ibrahim et al., 2015; Budiharto and Meiliana, 2018), product reviews (Fauzi, 2019), stock forecasting (Cakra and Trisedya, 2015; Sagala et al., 2020), and COVID-19 monitoring (Nurdeni et al., 2021). By translating an existing text, we additionally produce a parallel corpus, which is useful for building and evaluating translation systems. As we translate from a regional high-resource language (Indonesian), we ensure that the topics and entities reflected in the data are culturally relevant to the other languages, which is generally not the case when translating an English dataset (Conneau et al., 2018; Ponti et al., 2020). We apply the corpus to two downstream tasks: sentiment analysis and machine translation. We use the new benchmark to assess the performance of existing Indonesian language models (LMs), multilingual LMs, and classical machine learning methods.

Our contributions are as follows:

- We propose NusaX, the first high-quality human annotated parallel corpus in 10 languages from Indonesia, and corresponding parallel data in Indonesian and English, covering the tasks of sentiment analysis and machine translation.
- We provide an extensive evaluation of deep learning and classical NLP/machine learning methods on downstream tasks in few-shot and full-data settings.
- We conduct comprehensive analysis of the languages under study both from linguistic

[*] These authors contributed equally.

[1]The dataset is released at https://github.com/IndoNLP/nusax.

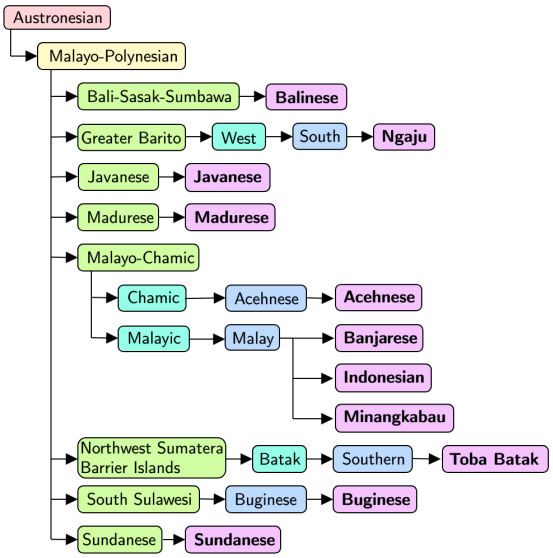

Figure 1: Language taxonomy of the 10 focus languages and Indonesian, according to Ethnologue (Eberhard et al., 2021). The color represents the language category level in the taxonomy. Purple denotes language, and other colors denote language family.

and empirical perspectives, the cross-lingual transferability of existing monolingual and multilingual LMs, and an efficiency analysis of various methods for NLP tasks in extremely low-resource languages.

2 Focus Languages

We work on 10 local languages in Indonesia: Acehnese, Balinese, Banjarese, Buginese, Madurese, Minangkabau, Javanese, Ngaju, Sundanese, and Toba Batak. Most of these languages have a population of over 2 million speakers (van Esch et al., 2022; Aji et al., 2022), but are underrepresented in NLP research. Figure 1 shows the taxonomy of these languages and Indonesian. Geographically, these languages are spoken on different big islands in Indonesia, including Sumatra, Borneo, Java, Madura, and Sulawesi. The languages belong to the Austronesian language family under the Malayo-Polynesian subgroup. While some of the covered languages are written in multiple scripts, we use the Latin script in NusaX, which has become predominant for all covered languages.

Indonesian (ind) is the national language of Indonesia based on the 1945 Constitution of the Republic of Indonesia (article 36). It is written in Latin script, and was developed from literary "Classical Malay" of the Riau-Johor sultanate (Sneddon, 2003), with regional variants. Its lexical similarity to Standard Malay is over 80%. 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).

Acehnese (ace) is a language spoken mainly in the Aceh province. Although it is the de facto language 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. In addition to the large number of diphthongs, it has a high percentage of monosyllabic root morphemes.

Balinese (ban) is a language spoken mainly in the Bali province. It has three main dialects: Highland Balinese, Lowland Balinese, and Nusa Penida. Since the early 20th century, it has mainly been written in the Latin script, but also has its own Balinese script. The word order in Balinese is SVO. Balinese has three sociolinguistic registers (Arka, 2003).

Banjarese (bjn) is a language spoken in Kalimantan (Central, East, South, and West Kalimantan provinces). It is dominant in the South Kalimantan Province and is also growing rapidly in the Central and Eastern Kalimantan provinces. It has two main dialects: Kuala and Hulu. 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 et al., 2013). It has 73% of lexical similarity with Indonesian and is written in Arabic and Latin scripts (Eberhard et al., 2021).

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. Historically, it was written in the Buginese script (derived from Brahmi script), but is mainly written in Latin script now (Eberhard et al., 2021). Buginese employs sentence patterns, pronouns, and other terms to express politeness (Weda, 2016).

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, 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. 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).

Javanese (jav) is a language spoken mainly on Java Island. It is the de facto language of provincial identity in central and eastern Java. The number of native Javanese speakers is greater than the number of Indonesian L1 speakers (Eberhard et al., 2021). Javanese consists of several regional dialects, which differ primarily in pronunciation and vocabulary. Javanese has an elaborate system of speech levels related to the relation of the speaker to the interlocutor that depend on social status, age, kinship distance, and familiarity (Wedhawati et al., 2001). It used to be written in Javanese script, but since the 20th century has mostly been written in Latin script.

Ngaju (nij) is a language spoken in the Central Kalimantan province. It is widely used as a language for trade in much of Kalimantan, from the Barito to the Sampit River. It has various affixes and reduplication, and its word order is similar to Indonesian. Pronouns have enclitic forms to mark possessors in a noun phrase or passive agents (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 has elaborate coding of respect levels. It has been written in Latin script since the mid-19th century but was previously written in Arabic, Javanese, and Sundanese scripts. Sundanese is a predominantly SVO language, and has voice marking and incorporates some (optional) actor-verb agreement, i.e., number and person (Kurniawan, 2013).

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 verb-initial, and have verb systems reminiscent of Philippine languages, although they differ from them in many details (Blust et al., 2013).

3 Data Construction

Our data collection process consists of several steps. First, we take an existing dataset in a high-resource local language (Indonesian) as a base for expansion to the other ten languages, and ask human annotators to translate the text. To ensure the quality of the final translation, we run quality assurance with additional human annotators.

3.1 Annotator Recruitment

Eliciting or annotating data in underrepresented languages generally requires working with local language communities in order to identify competent bilingual speakers (Nekoto et al., 2020). In the Indonesian setting, this challenge is compounded by the fact that most languages have several dialects. As dialects in Indonesian languages may have significant differences in word usage and meaning (Aji et al., 2022), it is important to recruit annotators who speak the same or similar dialects to ensure that translations are mutually intelligible.

In this work, we employ at least 2 expert annotators who are native speakers of each local language and Indonesian. To filter the recruited annotators, we first ask annotator candidates to translate three samples. We then conduct a peer review by asking whether they can understand the translations of other annotators for the same language, using the hired annotators as translators as well as translation validators. We also conducted 2 hours of training to introduce the user interface of the annotation system for selected workers. For English translations, we hire annotators based on their English proficiency test scores with an IELTS score ≥ 6.5 or TOEFL PBT score ≥ 600.

3.2 Data Filtering and Sampling

We base our dataset on SmSA, the largest publicly available Indonesian sentiment analysis dataset from the IndoNLU benchmark (Purwarianti and Crisdayanti, 2019; Wilie et al., 2020). SmSA is an expert-annotated sentence-level multi-domain sentiment analysis dataset consisting of more than 11,000 instances of comments and reviews collected from several online platforms such as Twitter, Zomato, and TripAdvisor. We filter the data to remove abusive language and personally-identifying information by manually inspecting all sentences. We randomly select 1,000 samples via stratified sampling for translation, ensuring that the label distribution is balanced.

3.3 Human Translation

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

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

3.4 Human-Assisted Quality Assurance

We conduct quality control (QC) between two annotators by having annotator A check the translations of annotator B, and vice versa. We include the corrected translations in our dataset. To ensure the quality assurance is performed well, we randomly perturb 5% of the sentences by removing a random sequence of words. The quality assurance annotators are then expected to notice the perturbed sentences and fix them.

We analyze the quality assurance edits for Balinese, Sundanese, and Javanese, which are spoken by the authors of this paper. For each language, we randomly sample 100 translations that have been edited by a QC annotator. We classify edits as follows:

Typos and Mechanics: Edit that involves correcting typos, punctuation, casing, white spaces/dashes, and numerical formatting.

Orthography: Edit that changes the spelling of words due to orthographic variation in local languages without a standard orthography. The word sounds and means the same before and after editing, and both are used by natives. The QC annotator might feel that one writing variant is more natural/commonly used, and hence make this change.

Translation: The words used by the translator are still in Indonesian and the QC annotator translates them to the local language.

Word edit: The QC annotator paraphrases a word/phrase. This also includes adding/removing words and morpheme changes.

Major changes: Other edits that significantly alter the original translation.

The results are shown in Table 1. Generally, word edits make up the majority of QC modifications, which involve replacing a word/phrase with a synonym or altering a morpheme slightly. In contrast, major changes are extremely rare. We also see changes to the orthography around 10% of the time. Other types of edits vary between languages. Sundanese has significantly less typos compared to other languages, but a considerably higher number of translation edits. We suspect this is because code-switching with Indonesian happens regularly in Sundanese, which results in many Indonesian words being adopted despite the existence of equivalent Sundanese translations.

Category	ban	sun	jav
Typos & Mechanic	31	14	42
Orthography	14	6	12
Translation	22	55	10
Word edit	67	65	61
Major changes	3	0	1

Table 1: Statistics of QC edits per category over 100 samples.

3.5 Bilingual Lexicon Creation

Bilingual lexicons are useful for data augmentation (Wang et al., 2022) and evaluating cross-lingual representations (Artetxe et al., 2018). We select 400 words from an Indonesian lexicon[2] to be translated into the 10 local languages and English. For each language, we employ two annotators and ask them to translate the word into all possible lexemes. The translations from both annotators are combined. We obtain 800–1,600 word pairs for each of our 11 language pairs (from Indonesian to the remaining languages). We augment the bilingual lexicon with data from PanLex (Kamholz et al., 2014).

[2] https://github.com/andria009/IndonesianSentimentLexicon

4 NusaX Benchmark

4.1 Tasks

We develop two tasks — sentiment analysis and machine translation — based on the datasets covering 12 languages, including Indonesian, English, and the 10 local languages. For the NusaX sentiment dataset, each language has the same label distribution and we show the label distribution of each dataset subset in Table 2. We maintain the label ratio in each dataset subset to ensure a similar distribution. More details of the dataset are provided in Appendix C.

4.1.1 Sentiment Analysis

Sentiment analysis is an NLP task that aims to identify the sentiment of a given text document. The sentiment is commonly categorized into 3 classes: positive, negative, and neutral. We focus our dataset construction on sentiment analysis because it is one of the most widely explored tasks in Indonesia (Aji et al., 2022) due to broad industrial relevance, such as for competitor and marketing analysis, and detection of unfavorable rumors for risk management (Socher et al., 2013). After translating 1,000 instances from the sentiment analysis dataset (SmSA), we have a sentiment analysis dataset for each translated language. For each language, we split the dataset into 500 train, 100 validation, and 400 test examples. In total, our dataset contains 6,000 train, 1,200 validation, and 4,800 test instances across 12 languages (Indonesian, English and the 10 local languages).

4.1.2 Machine Translation

Indonesia consists of 700+ languages covering three different language families (Aji et al., 2022). Despite its linguistic diversity, existing machine translation systems only cover a small fraction of Indonesian languages, mainly Indonesian (the national language), Sundanese, and Javanese. To broaden the coverage of existing machine translation systems for underrepresented local languages, we construct a machine translation dataset using our translated sentiment corpus, which results in a parallel corpus between all language pairs. In other words, we have 132 possible parallel corpora, each with 1,000 samples (500 train, 100 validation, and 400 test instances) which can be used to train machine translation models. Compared to many other MT evaluation datasets, our data is in the review domain and is not English-centric.

Subset	Negative	Neutral	Positive
Train	192	119	189
Valid	38	24	38
Test	153	96	151

Table 2: Label distribution of NusaX Sentiment dataset.

4.2 Baselines

4.2.1 Classical Machine Learning

Classical machine learning approaches are still widely used by local Indonesian researchers and institutions due to their efficiency (Nityasya et al., 2021). The trade-off between performance and compute cost is particularly important in situations with limited compute, which are common for low-resource languages. We therefore use classical methods as baselines for our comparison. Namely, we use naive Bayes, SVM, and logistic regression for the classification tasks. For MT, we employ a naive baseline that copies the original Indonesian text, a dictionary-based substitution method using the bilingual lexicon, and a phrase-based MT system based on Moses (Koehn et al., 2007).

4.2.2 Pre-trained Local Language Models

Recent developments in neural pre-trained LMs have brought substantial improvements in various NLP tasks. Despite the lack of resources in Indonesian and local languages, there have been some efforts in developing large pre-trained LMs for Indonesian and major local languages. IndoBERT (Wilie et al., 2020) and SundaneseBERT (Wongso et al., 2022) are two popular LMs for natural language understanding (NLU) tasks in Indonesian and Sundanese. IndoBART and IndoGPT have also been introduced for natural language generation (NLG) tasks in Indonesian, Sundanese, and Javanese (Cahyawijaya et al., 2021). We employ these LMs as baselines to assess their adaptability to other languages.

4.2.3 Massively Multilingual LMs

We consider large pre-trained multilingual LMs to further understand their applicability to low-resource languages. Specifically, we experiment with mBERT (Devlin et al., 2019) and XLM-R (Conneau et al., 2020) for sentiment analysis, and mBART (Liu et al., 2020) and mT5 (Xue et al., 2021) for machine translation. We provide the hyper-parameters of all models in Appendix B.

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

Table 3: Sentiment analysis results in macro-F1 (%). Models were trained and evaluated on each language.

5 Results

5.1 Overall Results

Sentiment Analysis Table 3 shows the sentiment analysis performance of various models across different local languages, trained and evaluated using data in the same language. Fine-tuned large LMs such as $\text{IndoBERT}_{\text{LARGE}}$ and $\text{XLM-R}_{\text{LARGE}}$ generally achieve the best performance. XLM-R models achieve strong performance on some languages, such as Indonesian (ind), Banjarese (bjn), English (eng), Javanese (jav), and Minangkabau (min). Many of these languages are included in XLM-R's pre-training data while others may benefit from positive transfer from related languages. For instance, Banjarese is similar to Malay and Indonesian (Nasution et al., 2021), while Minangkabau shares some words and syntax with Indonesian (Koto and Koto, 2020). IndoBERT models, despite only being pre-trained on Indonesian, also show good performance across some local languages, suggesting transferability from Indonesian to the local languages.

The classic approaches are surprisingly competitive with the neural methods, with logistic regression even outperforming $\text{IndoBERT}_{\text{LARGE}}$ and XLM-R on Acehnese (ace), Buginese (bug), and Toba Batak (bbc). These results indicate that both Indonesian and multilingual pre-trained LMs cannot transfer well to these languages, which is supported by the fact that these languages are very distinct from Indonesian, Sundanese, Javanese, or Minangkabau — the languages covered by IndoBERT and XLM-R.

Machine Translation We show the results on machine translation in Table 4 (x $\rightarrow$ ind) based on SacreBLEU (Post, 2018). As some local languages are similar to Indonesian, we observe that the Copy baseline (which does not do any translation) performs quite well. Minangkabau (min) and Banjarese (bjn) achieve high BLEU without any translation despite not being included in the LM pre-training data, due to their similarity with Indonesian (Koto and Koto, 2020; Nasution et al., 2021). Since these local languages share grammatical structure with Indonesian, dictionary-based word substitution yields a reasonable improvement.

Both PBSMT and fine-tuned LMs reach encouraging performance levels despite the limited training data, which we again attribute to the target languages' similarity to Indonesian. In contrast, the performance for translating Indonesian languages from/to English is extremely poor as shown in Table 5, demonstrating the importance of non-English-centric translation. Overall, we observe good translation performance across local languages. Thus, there is an opportunity to utilize translation models to create new synthetic datasets in local languages via translation from a related high-resource language, not only for Indonesian local languages but also other underrepresented languages. However, note that even for language pairs where the SacreBLEU score is very high, we observe translation deficiencies stemming from the small amount of training data: rare words may just be copied with PBSMT, and mistranslated with NMT.

Similar effects are also observed for (ind $\rightarrow$ x) translation, as shown in Table 6. Similar to (x $\rightarrow$ ind) translations, we observe that the Copy baseline performs quite well on Minangkabau (min) and Banjarese (bjn) due to their similarity with Indonesian (Koto and Koto, 2020; Nasution et al., 2021). Dictionary-based word substitution also yields a reasonable improvement especially for Ja-

	x → ind											
Model	ace	ban	bbc	bjn	bug	eng	jav	mad	min	nij	sun	**avg**
Copy	5.88	9.99	4.28	15.99	3.44	0.57	9.29	5.11	18.10	7.51	9.24	8.13
Word Substitution	7.33	12.30	5.02	16.17	3.52	1.67	17.34	7.89	24.17	12.07	15.38	11.17
PBSMT	**25.17**	**41.22**	**20.94**	47.80	**15.21**	6.68	46.99	**38.39**	**60.56**	32.86	41.79	34.33
IndoGPT	7.01	13.23	5.27	19.53	1.98	4.26	27.31	13.75	23.03	10.83	23.18	13.58
IndoBARTv2	24.44	40.49	19.94	**47.81**	12.64	11.73	**50.64**	36.10	58.38	**33.50**	**45.96**	**34.69**
mBART-50	18.45	34.23	17.43	41.73	10.87	**17.92**	39.66	32.11	59.66	29.84	35.19	30.64
$mT5_{BASE}$	18.59	21.73	12.85	42.29	2.64	12.96	45.22	32.35	58.65	25.61	36.58	28.13

Table 4: Results of the machine translation task from other languages to Indonesian (x → ind) based on SacreBLEU.

Model	**avg. SacreBLEU**			
	ind → x	x → ind	eng → x	x → eng
PBSMT	**28.72**	34.33	4.56	5.84
IndoBARTv2	28.21	**34.69**	6.36	**7.46**
mBART-50	24.69	30.64	**7.20**	6.45

Table 5: MT performance from / to Indonesian compared to from / to English.

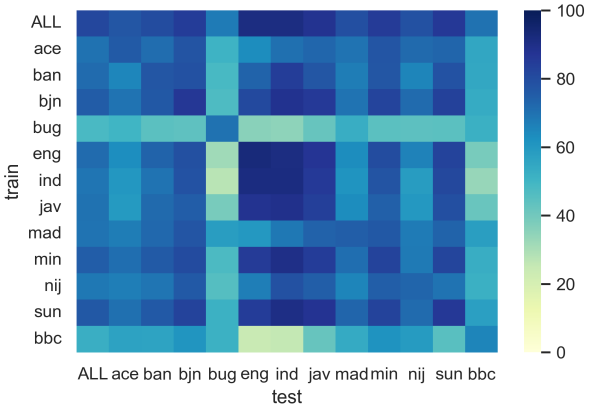

Figure 2: Zero-shot cross-lingual results for the sentiment analysis task with $XLM\text{-}R_{LARGE}$. The model is trained on the language indicated on the x-axis and evaluated on all languages.

vanese (jav), Minangkabau (min), and Sundanese (sun) due to the high similarity of the grammatical structure with Indonesian. PBSMT and fine-tuned IndoBARTv2 models achieve the best scores over multiple local languages despite the limited training data, which is also attributed to the target languages' similarity to Indonesian.

5.2 Cross-lingual Capability of LMs

From a linguistic perspective, local languages in Indonesia share similarities according to language family. Many local languages share a similar grammatical structure and have some vocabulary overlap. Following prior work that demonstrates positive transfer between closely-related languages (Cahyawijaya et al., 2021; Hu et al., 2020; Aji et al., 2020; Khanuja et al., 2020; Winata et al., 2021, 2022), we analyze the transferability between closely-related languages in the Malayo-Polynesian language family.

Empirically, we show the cross-lingual capability of the best performing model ($XLM\text{-}R_{LARGE}$) in the zero-shot cross-lingual setting for sentiment analysis. The heatmap is shown in Figure 2. In general, most languages, except for Buginese (bug) and Toba Batak (bbc), can be used effectively as the source language, reaching ∼70–75% F1 on average, compared to an average of 80% F1 in the monolingual setting (cf. $XLM\text{-}R_{LARGE}$ in Table 3). This empirical result aligns with the fact that both Buginese (bug) and Toba Batak (bbc) have very low vocabulary overlap with Indonesian (cf. Copy in Tables 4 and 6). Interestingly, despite coming from a completely different language family, English can also be effectively used as the source language for all 10 local languages, likely due to its prevalence during pre-training.

These results demonstrate that we can take advantage of language similarity by transferring knowledge from Indonesian and other local languages to perform zero-shot or few-shot classification in closely-related languages. New datasets for underrepresented languages that are closely related to high-resource languages thus do not necessarily need to be large, which make the development of NLP datasets in low-resource languages more affordable than may initially appear to be the case.

5.3 Multilingual Capability

We explore training multilingual models, as most Indonesian local languages share similarities. For

	ind → x											
Model	ace	ban	bbc	bjn	bug	eng	jav	mad	min	nij	sun	**avg**
Copy	5.89	10.00	4.28	15.99	3.45	0.56	9.29	5.11	18.10	7.52	9.24	8.13
Word Substitution	7.60	10.31	5.99	17.51	3.57	0.76	14.75	7.58	22.34	9.76	12.38	10.23
PBSMT	**20.47**	26.48	18.18	**42.08**	10.84	7.73	39.08	**33.26**	**52.21**	**29.58**	36.04	**28.72**
IndoGPT	9.60	14.17	8.20	22.23	5.18	5.89	24.05	14.44	26.95	17.56	23.15	15.58
IndoBARTv2	19.21	**27.08**	**18.41**	40.03	**11.06**	**11.53**	**39.97**	28.95	48.48	27.11	**38.46**	28.21
mBART-50	17.21	22.67	17.79	34.26	10.78	3.90	35.33	28.63	43.87	25.91	31.21	24.69
$mT5_{BASE}$	14.79	18.07	18.22	38.64	6.68	11.21	33.48	0.96	45.84	13.59	33.79	21.39

Table 6: Results of the machine translation task from Indonesian to other languages (ind → x) in SacreBLEU.

Language	**Single**	**Multi**	**LOLO**
Acehnese	75.9	76.96	75.79
Balinese	77.1	80.13	77.83
Banjarese	86.3	84.85	82.68
Buginese	70.0	67.86	63.67
English	92.6	91.05	89.88
Indonesian	91.6	91.13	90.62
Javanese	84.2	88.19	87.39
Madurese	74.9	79.41	78.52
Minangkabau	83.1	85.29	84.45
Ngaju	73.3	78.82	76.31
Sundanese	86.0	86.02	84.41
Toba batak	65.5	70.00	68.76
Average	80.04	81.64	80.03

Table 7: Sentiment analysis results for macro-F1 (%) of $XLM\text{-}R_{LARGE}$ in the multilingual setting.

sentiment analysis, we concatenate the training data of all languages. Additionally, we also explore Leave-One-Language-Out (LOLO), where we train on all data except for the test language. The LOLO setting arguably reflects the most realistic scenario where we do not have training data for a particular language, but we do have access to data in other local languages. The multilingual results for sentiment analysis are shown in Table 7. Multilingual training outperforms monolingual training, while LOLO matches the performance of training on target language data. Related language data is thus often sufficient for good cross-lingual results.

6 Data Collection Challenges

In this section, we discuss challenges faced during data collection.

Finding annotators We found collecting the NusaX dataset challenging. First of all, finding local language-speaking annotators is not easy, and popular platforms such as MTurk do not support these languages. Instead, we looked for annotators through local Indonesian networks and forums, such as the INACL forum, local campus forums, or the Indonesian polyglot community. We intended to cover as many local languages as possible, but based on the available annotators, only the 10 languages presented in this paper were possible, as we needed at least 2 annotators for each language. Searching for annotators online is not easy, due to disparities in Internet penetration in different parts of Indonesia. Hence, we might not reach potential annotators through online communities alone. However, holding an in-person workshop for data collection is also not practical; Indonesia is an archipelago and traveling between islands is costly. Similar challenges occur in many other regions, including Africa and South America.

Communication with annotators Communication between the authors and annotators was done through WhatsApp, as the most popular communication tool in Indonesian (Mulyono et al., 2021). Annotation was conducted through spreadsheets. We found that some of the annotators use mobile apps instead of a desktop for annotation. Their reasons include ease of use, no access to a laptop, and better keyboard support for typing diacritics. In the most extreme case, one annotator printed out the sheet and performed the annotation on paper, then took a picture of the paper and sent it back to us. We found some annotators to be difficult to contact, due to other commitments such as college or work. Some of them were not responsive and had to be replaced by new annotators.

7 Related Work

Multilingual Parallel Corpora Several multilingual parallel corpora have been developed to support studies on machine translation such as

GCP (Imamura and Sumita, 2018), Leipzig (Goldhahn et al., 2012), JRC Acquis (Steinberger et al., 2006), TUFS Asian Language Parallel (Nomoto et al., 2018), Intercorp (ek Čermák and Rosen, 2012), DARPA LORELEI (Strassel and Tracey, 2016), Asian Language Treebank (Riza et al., 2016), FLORES (Guzmán et al., 2019), the Bible Parallel Corpus (Resnik et al., 1999; Black, 2019), JW-300 (Agić and Vulić, 2019), BiToD (Lin et al., 2021), and WikiMatrix (Schwenk et al., 2021). Guzmán et al. (2019) describe the procedure to generate high-quality translations as part of FLORES. Similar to FLORES, we also conducted QC of the translations.

Emerging Language Benchmarks Recently, benchmarks in underrepresented languages have emerged, such as MasakhaNER (Adelani et al., 2021), AmericasNLI (Ebrahimi et al., 2022), PMIndia (Haddow and Kirefu, 2020), Samanantar (Ramesh et al., 2022), and NaijaSenti (Muhammad et al., 2022). Particularly, for Indonesian languages, NLP benchmarks have been developed such as IndoNLU (Wilie et al., 2020), IndoLEM (Koto et al., 2020), IndoNLG (Cahyawijaya et al., 2021), IndoNLI (Mahendra et al., 2021), and English–Indonesian machine translation (Guntara et al., 2020).

Datasets for Indonesian Local Languages Only a limited number of labeled datasets exist for local languages in Indonesia. WikiAnn (Pan et al., 2017) — a weakly-supervised named entity recognition dataset — covers Acehnese, Javanese, Minangkabau, and Sundanese. Putri et al. (2021) built a multilingual dataset for abusive language and hate speech detection involving Javanese, Sundanese, Madurese, Minangkabau, and Musi languages. Sakti and Nakamura (2013) constructed speech corpora for Javanese, Sundanese, Balinese, and Toba Batak. Few datasets exist for individual languages, e.g., sentiment analysis and machine translation in Minangkabau (Koto and Koto, 2020) and emotion classification in Sundanese (Putra et al., 2020). Finally, some datasets focus on colloquial Indonesian mixed with local languages in the scope of morphological analysis (Wibowo et al., 2021) and style transfer (Wibowo et al., 2020).

8 Conclusion

In this paper, we propose NusaX, the first parallel corpus for 10 low-resource Indonesian languages. We create a new benchmark for sentiment analysis and machine translation in zero-shot and full-data settings. We present a comprehensive analysis of the language similarity of these languages from both linguistic and empirical perspectives by assessing the cross-lingual transferability of existing Indonesian and multilingual pre-trained models.

We hope NusaX can enable NLP research for under-represented languages, and can be used as a testbed for adaptation or few-shot learning methods that take advantage of similarities between languages. NusaX opens up the possibility for future research that focuses on covering more local languages, and additionally, further extension to other tasks and domains. Our study on cross-lingual transfer enables further exploration on cross-lingual zero-shot learning for more diverse tasks in local languages. Our guidelines and discussion of data collection issues may also motivate future work on more efficient high-quality data collection for extremely low-resource languages.

Acknowledgments

We thank Dea Adhista and all annotators who helped us in building the corpus. We are grateful to Alexander Gutkin and Xinyu Hua for feedback on a draft of this manuscript. This work has been partially funded by Kata.ai (001/SD/YGI-NLP/1/2022) and PF20-43679 Hong Kong PhD Fellowship Scheme, Research Grant Council, Hong Kong.

Limitations

We created data for low-resource languages, which increases the accessibility of NLP research for marginalized communities. However, we were only able to cover 10 languages with only 1000 samples each, due to cost and the number of available annotators. This dataset has limited domain coverage and may also contain biases towards certain groups or entities. We tried our best to eliminate negative biases based on a manual inspection of the data. As our dataset was translated, there may be some translationese artifacts in the resulting corpus. We invited annotators based on their fluency level on a particular language. However, the fluency level is self-declared, and there is no mechanism to verify it, except for several languages that are spoken by authors of this paper. The dialect used in the dataset also depends on the annotator, for languages with multiple dialects.

References

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

Željko Agić and Ivan Vulić. 2019. JW300: A wide-coverage parallel corpus for low-resource languages. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3204–3210, Florence, Italy. Association for Computational Linguistics.

Alham Fikri Aji, Nikolay Bogoychev, Kenneth Heafield, and Rico Sennrich. 2020. In neural machine translation, what does transfer learning transfer? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7701–7710.

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.

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

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 789–798, Melbourne, Australia. Association for Computational Linguistics.

Annisa Nurul Azhar, Masayu Leylia Khodra, and Arie Pratama 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. IEEE.

Alan W Black. 2019. Cmu wilderness multilingual speech dataset. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5971–5975. IEEE.

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

Widodo Budiharto and Meiliana Meiliana. 2018. Prediction and analysis of indonesia presidential election from twitter using sentiment analysis. *Journal of Big data*, 5(1):1–10.

Samuel Cahyawijaya, Genta Indra Winata, Bryan Wilie, Karissa Vincentio, Xiaohong Li, Adhiguna Kuncoro, Sebastian Ruder, Zhi Yuan Lim, Syafri Bahar, Masayu Khodra, Ayu Purwarianti, and Pascale Fung. 2021. IndoNLG: Benchmark and resources for evaluating Indonesian natural language generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8875–8898, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yahya Eru Cakra and Bayu Distiawan Trisedya. 2015. Stock price prediction using linear regression based on sentiment analysis. In *2015 international conference on advanced computer science and information systems (ICACSIS)*, pages 147–154. IEEE.

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

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel 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*, pages 2475–2485, Brussels, Belgium. 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.

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

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

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

Abteen Ebrahimi, Manuel Mager, Arturo Oncevay, Vishrav Chaudhary, Luis Chiruzzo, Angela Fan, John Ortega, Ricardo Ramos, Annette Rios, Ivan Vladimir Meza Ruiz, Gustavo Giménez-Lugo, Elisabeth Mager, Graham Neubig, Alexis Palmer, Rolando Coto-Solano, Thang Vu, and Katharina Kann. 2022. AmericasNLI: Evaluating zero-shot natural language understanding of pretrained multilingual models in truly low-resource languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6279–6299, Dublin, Ireland. Association for Computational Linguistics.

Franti ek Čermák and Alexandr Rosen. 2012. The case of intercorp, a multilingual parallel corpus. *International Journal of Corpus Linguistics*, 17(3):411–427.

M Ali Fauzi. 2019. Word2vec model for sentiment analysis of product reviews in indonesian language. *International Journal of Electrical and Computer Engineering*, 9(1):525.

Dirk Goldhahn, Thomas Eckart, and Uwe Quasthoff. 2012. Building large monolingual dictionaries at the Leipzig corpora collection: From 100 to 200 languages. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 759–765, Istanbul, Turkey. European Language Resources Association (ELRA).

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.

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

Barry Haddow and Faheem Kirefu. 2020. Pmindia–a collection of parallel corpora of languages of india. *arXiv preprint arXiv:2001.09907*.

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

Mochamad Ibrahim, Omar Abdillah, Alfan F Wicaksono, and Mirna Adriani. 2015. Buzzer detection and sentiment analysis for predicting presidential election results in a twitter nation. In *2015 IEEE international conference on data mining workshop (ICDMW)*, pages 1348–1353. IEEE.

Kenji Imamura and Eiichiro Sumita. 2018. Multilingual parallel corpus for global communication plan. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

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

Simran Khanuja, Sandipan Dandapat, Anirudh Srinivasan, Sunayana Sitaram, and Monojit Choudhury. 2020. Gluecos: An evaluation benchmark for code-switched nlp. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3575–3585.

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

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

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

Zhaojiang Lin, Andrea Madotto, Genta Indra Winata, Peng Xu, Feijun Jiang, Yuxiang Hu, Chen Shi, and Pascale Fung. 2021. BiToD: A bilingual multi-domain dataset for task-oriented dialogue modeling. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*.

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

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.

Shamsuddeen Hassan Muhammad, David Adelani, Anuoluwapo Aremu, and Idris Abdulmumin. 2022. NaijaSenti: A nigerian Twitter sentiment corpus for multilingual sentiment analysis. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 590–602, Marseille, France. European Language Resources Association.

Herri Mulyono, Gunawan Suryoputro, and Shafa Ramadhanya Jamil. 2021. The application of whatsapp to support online learning during the covid-19 pandemic in indonesia. *Heliyon*, 7(8):e07853.

Arbi Haza Nasution, Yohei Murakami, and Toru Ishida. 2021. Plan optimization to bilingual dictionary induction for low-resource language families. *ACM Transactions Asian Low-Resource Language Information Processing*, 20(2).

Wilhelmina Nekoto, Vukosi Marivate, Tshinondiwa Matsila, Timi Fasubaa, Tajudeen Kolawole, Taiwo Fagbohungbe, Solomon Oluwole Akinola, Shamsuddee Hassan Muhammad, Salomon Kabongo, Salomey Osei, Sackey Freshia, Rubungo Andre Niyongabo, Ricky Macharm, Perez Ogayo, Orevaoghene Ahia, Musie Meressa, Mofe Adeyemi, Masabata Mokgesi-Selinga, Lawrence Okegbemi, Laura Jane Martinus, Kolawole Tajudeen, Kevin Degila, Kelechi Ogueji, Kathleen Siminyu, Julia Kreutzer, Jason Webster, Jamiil Toure Ali, Jade Abbott, Iroro Orife, Ignatius Ezeani, Idris Abdulkabir Dangana, Herman Kamper, Hady Elsahar, Goodness Duru, Ghollah Kioko, Espoir Murhabazi, Elan van Biljon, Daniel Whitenack, Christopher Onyefuluchi, Chris Emezue, Bonaventure Dossou, Blessing Sibanda, Blessing Itoro Bassey, Ayodele Olabiyi, Arshath Ramkilowan, Alp Öktem, Adewale Akinfaderin, and Abdallah Bashir. 2020. Participatory Research for Low-resourced Machine Translation: A Case Study in African Languages. In *Findings of EMNLP 2020*.

Made Nindyatama Nityasya, Haryo Akbarianto Wibowo, Radityo Eko Prasojo, and Alham Fikri Aji. 2021. Costs to consider in adopting NLP for your business.

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*, pages 436–439.

Deden Ade Nurdeni, Indra Budi, and Aris Budi Santoso. 2021. Sentiment analysis on covid19 vaccines in indonesia: From the perspective of sinovac and pfizer. In *2021 3rd East Indonesia Conference on Computer and Information Technology (EIConCIT)*, pages 122–127. IEEE.

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, Vancouver, Canada. Association for Computational Linguistics.

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.

Edoardo Maria 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)*, pages 2362–2376, Online. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.

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)*, pages 1–5. IEEE.

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.

Shofianina Dwi Ananda Putri, Muhammad Okky Ibrohim, and Indra Budi. 2021. Abusive language and hate speech detection for Indonesian-local language

in social media text. In *Recent Advances in Information and Communication Technology 2021*, pages 88–98, Cham. Springer International Publishing.

Gowtham Ramesh, Sumanth Doddapaneni, Aravinth Bheemaraj, Mayank Jobanputra, Raghavan AK, Ajitesh Sharma, Sujit Sahoo, Harshita Diddee, Mahalakshmi J, Divyanshu Kakwani, Navneet Kumar, Aswin Pradeep, Srihari Nagaraj, Kumar Deepak, Vivek Raghavan, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh Shantadevi Khapra. 2022. Samanantar: The Largest Publicly Available Parallel Corpora Collection for 11 Indic Languages. *Transactions of the Association for Computational Linguistics*, 10:145–162.

Philip Resnik, Mari Broman Olsen, and Mona Diab. 1999. The bible as a parallel corpus: Annotating the 'book of 2000 tongues'. *Computers and the Humanities*, 33(1):129–153.

Hammam Riza, Michael Purwoadi, Teduh Uliniansyah, Aw Ai Ti, Sharifah Mahani Aljunied, Luong Chi Mai, Vu Tat Thang, Nguyen Phuong Thai, Vichet Chea, Sethserey Sam, et al. 2016. Introduction of the asian language treebank. In *2016 Conference of The Oriental Chapter of International Committee for Coordination and Standardization of Speech Databases and Assessment Techniques (O-COCOSDA)*, pages 1–6. IEEE.

Tommy Wijaya Sagala, Mei Silviana Saputri, Rahmad Mahendra, and Indra Budi. 2020. Stock price movement prediction using technical analysis and sentiment analysis. In *Proceedings of the 2020 2nd Asia Pacific Information Technology Conference*, pages 123–127.

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.

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

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

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642.

Ralf Steinberger, Bruno Pouliquen, Anna Widiger, Camelia Ignat, Tomaz Erjavec, Dan Tufis, and Dániel Varga. 2006. The jrc-acquis: A multilingual aligned parallel corpus with 20+ languages. *arXiv preprint cs/0609058*.

Stephanie Strassel and Jennifer Tracey. 2016. Lorelei language packs: Data, tools, and resources for technology development in low resource languages. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3273–3280.

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

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

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

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

Wedhawati Wedhawati, Wiwin E.S.N., Sri Nardiati, Herawati Herawati, Restu Sukesti, Marsono Marsono, Edi Setiyanto, Dirgo Sabariyanto, Syamsul Arifin, Sumadi Sumadi, and Laginem Laginem. 2001. *Tata bahasa Jawa mutakhir*. Badan Pengembangan dan Pembinaan Bahasa.

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 word colloquialism. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3170–3183.

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, Suzhou, China. Association for Computational Linguistics.

Genta Winata, Shijie Wu, Mayank Kulkarni, Thamar Solorio, and Daniel Preoţiuc-Pietro. 2022. Cross-lingual few-shot learning on unseen languages. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing*, pages 777–791.

Genta Indra Winata, Andrea Madotto, Zhaojiang Lin, Rosanne Liu, Jason Yosinski, and Pascale Fung. 2021. Language models are few-shot multilingual learners. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 1–15.

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

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.

A Data Statement for NusaX

A.1 General Information

Dataset title NusaX

Dataset curators Alham Fikri Aji (MBZUAI), Rahmad Mahendra (Universitas Indonesia), Samuel Cahyawijaya (HKUST), Ade Romadhony (Telkom University, Indonesia), Genta Indra Winata (Bloomberg), Fajri Koto (University of Melbourne), Kemal Kurniawan (University of Melbourne)

Dataset version 1.0 (May 2022)

Data statement author Kemal Kurniawan (University of Melbourne)

Data statement version 1.0 (February 2022)

A.2 Executive Summary

NusaX is a multilingual parallel corpus across 10 local languages in Indonesia: Acehnese, Balinese, Banjarese, Buginese, Madurese, Minangkabau, Javanese, Ngaju, Sundanese, and Toba Batak. The data was translated obtained by human translation from Indonesian and human-assisted quality assurance.

A.3 Curation Rationale

The goal of the dataset creation process is to provide gold-standard sentiment analysis corpora for Indonesian local languages. The Indonesian data is sampled from SmSA (Purwarianti and Crisdayanti, 2019), an Indonesian sentiment analysis corpus. SmSA is chosen among other corpora (e.g., HoASA (Azhar et al., 2019) based on (1) the agreement of our manual re-annotation of a small and randomly selected samples and (2) manual inspection to ensure that the topics are diverse. After sampling, the data is edited and/or filtered to remove harmful contents and maintain quality. Several criteria are used in this process:

1. Is the sentiment label correct?
2. Does the sentence contain multiple sentiments?
3. Does the sentence contain harmful content that discriminates against race, religion, or other protected groups?
4. Does the sentence contain an attack toward an individual or is abusive?
5. Is the sentence politically charged?
6. Is the sentence overly Bandung/Sunda-centric?[3]
7. Will the sentence be difficult to translate into local languages?
8. Are there any misspellings?

A.4 Documentation for Source Datasets

NusaX is obtained by translating SmSA (Purwarianti and Crisdayanti, 2019), an Indonesian sentiment analysis dataset.

A.5 Language Variety

NusaX covers a total of 10 local languages spoken in Indonesia (ID) as shown in Table 8.

A.6 Speaker Demographic

The SmSA dataset was obtained from social media and online forums: Twitter, Zomato, TripAdvisor, Facebook, Instagram, Qraved. We can assume the users' age ranges from 25 to 34 years, which is the age range of the majority of Indonesian social media users[4].

A.7 Annotator Demographic

A total of 28 translators are employed in the translation process. All translators are Indonesian and recruited by via either online surveys or personal contacts. They are then selected based on (1) the self-reported fluency in the local language into which they would be translating and (2) the highest education level achieved. Those who (a) are native speakers of or fluent in the target local language and (b) finished at least high school education (id: *SMA/sederajat*) are selected.

Acehnese There are 3 translators for Acehnese, but only 2 of them responded when asked for demographic information. Thus, what follows is the demographic information of only those 2 translators. One has some experience in translation work, while the other does not. One identifies as male, and the other as female. Both are in their 20s. Lastly, one works as a freelancer, while the other is a farmer.

Balinese Three people translate into Balinese. Two of them have previous experience in translation work, and both identify as female. The other one, who identifies as male, does not have such

[3] Bandung is the capital city of West Java, in which Sunda is the ethnic group.

[4] https://www.statista.com/statistics/997297/indonesia-breakdown-social-media-users-age-gender/

Language	ISO 639-3	Annotators' Dialect	Example
Acehnese	ace	Banda Aceh	Meureutoh rumoh di Medan keunong ie raya
Balinese	ban	Lowland	Satusan umah ring medan merendem banjir
Toba Batak	bbc	Toba, Humbang	Marratus jabu di medan na hona banji
Banjarese	bjn	Hulu, Kuala	Ratusan rumah di medan tarandam banjir
Buginese	bug	Sidrap	Maddatu bola okko medan nala lempe
Javanese	jav	Matraman	Atusan omah ing medan kebanjiran
Madurese	mad	Situbondo	Ratosan bangko e medan tarendem banjir
Minangkabau	min	Padang, Agam	Ratuihan rumah di medan tarandam banjir
Ngaju	nij	Kapuas, Kahayan	Ratusan huma hong medan lelep awi banjir
Sundanese	sun	Priangan	Ratusan bumi di medan karendem banjir

Table 8: Local languages spoken in Indonesia (ID) that are covered in NusaX.

experience. Two of them are aged 20-29 years old, while the other is in their 30s. Their occupations are university lecturer, school teacher, and civil employee respectively.

Banjarese Two translators are employed for Banjarese, but only one responded when asked for demographic information. The translator has prior experience in translation work, identifies as male, is in his 40s, and works as a university lecturer.

Buginese Buginese is translated by 2 people, but only one responded when asked for demographic information. The person has prior translation experience, identifies as male, is aged 30-39 years old, and runs an Islamic boarding school as a living.

Javanese Four translators are employed for Javanese, but one did not respond when asked for demographic information. The other three have prior experience in translation work. Among them, two identify as female, and one as male. All of them are in their 20s. Two of them are university students, and the other one works as a freelance assistant editor.

Madurese There are 3 translators for Madurese. Only one of them has previous experience in translation work. Two of them identify as female, while the other as male. One person is aged under 20 years old and is a university student. The others are 20-29 years old and work as a school teacher and an employee in a private company respectively.

Minangkabau Three people translate into Minangkabau. Two of them have previous translation experience. All three identify as female and are aged 20-29 years old. They work as a civil employee, a university student, and a senior data annotator respectively.

Ngaju Two translators work on Ngaju, but only one responded when asked for demographic information. The translator has prior experience, identifies as female, is aged no less than 50 years old, and is a stay-at-home mother.

Sundanese There are 5 translators for Sundanese, four of which identify as female, and the other one as male. Three translators are in their 20s, one is younger than 20 years old, and the remaining one is in their 30s. The translators work as a school teacher, a university student, a university lecturer, and the remaining two as employees in a private company.

Toba Batak Three translators are employed for Toba Batak. One has prior translation experience. Two translators identify as male while the other as female. All three are in their 20s. One works for a private company, and the others are university students.

B Hyperparameters

B.1 Sentiment Analysis

Hyperparams	NB	SVM	LR
feature	{BoW, tfidf}	{BoW, tfidf}	{BoW, tfidf}
alpha	(0.001 - 1)	–	–
C	–	(0.01 - 100)	(0.001 - 100)
kernel	–	{rbf, linear}	–

Table 9: Hyperparameters of statistical models on sentiment analysis.

For statistical models, we use a spaCy as our toolkit, and we perform grid-search over the parameter ranges shown in Table 9 and select the

Hyperparams	Values
learning rate	[1e-4, 5e-5, **1e-5**, 5e-6, 1e-6]
batch size	[4, 8, 16, **32**]
num epochs	100
early stop	3
max norm	10
optimizer	Adam
Adam β	(0.9, 0.999)
Adam γ	0.9
Adam ϵ	1e-8

Table 10: Hyperparameters of pre-trained LMs on sentiment analysis. **Bold** denotes the best hyperparameter setting.

best performing model over the devset. For all pretrained LMs, we perform grid-search over batch size and learning rate while keeping the other hyperparameters fixed. The list of hyperparameters is shown in Table 10.

B.2 Machine Translation

Table 11 shows the hyperparameters of deep learning models on machine translation.

Hyperparams	IndoGPT	IndoBARTv2	mBART-50	mT5_{BASE}
learning rate	1e-4	1e-4	2e-5	5e-4
batch size		16		
gamma	0.98	0.98	0.98	0.95
max epochs		20		
early stop		10		
seed		{1...5}		

Table 11: Hyperparameters of pretrained LMs on machine translation.

C Dataset Statistics

In this section, we present more detail statistics of our NusaX datasets. To evaluate the difference between each language in the NusaX dataset, we analyze the vocabulary characteristic for each language. We collect the vocabulary for each language by removing all the punctuation in the sentence and tokenize the sentence with the spaCy tokenizer. [5] We show the vocabulary size and the top-10 words for each language on Table 12, and the vocabulary histogram for each language in Figure 4. We can see that the most common words between Indonesian and other local languages vary a lot, despite having a similar vocabulary size and histogram pattern. This shows the intuitive difference between Indonesian and local languages in Indonesia.

We further measure the vocabulary overlap over different language pairs. We measure the vocabulary overlap for each pair of languages by measuring the intersection over union (IoU) of the two vocabularies. We show the vocabulary overlap in Figure 3. From the results, we can conclude that English has the smallest vocabulary overlap with the other languages. This makes sense since English comes from a different language family, i.e., Indo-European language under the Germanic language branch, while the others are from the Austronesian language family under the Malayo-Polynesian branch. Other languages that have low vocabulary overlap are Buginese (bug) and Toba Batak (bbc). This aligns with our discussion in §5, which shows the distinction between these languages and the other languages in the NusaX dataset.

[5] https://github.com/explosion/spaCy

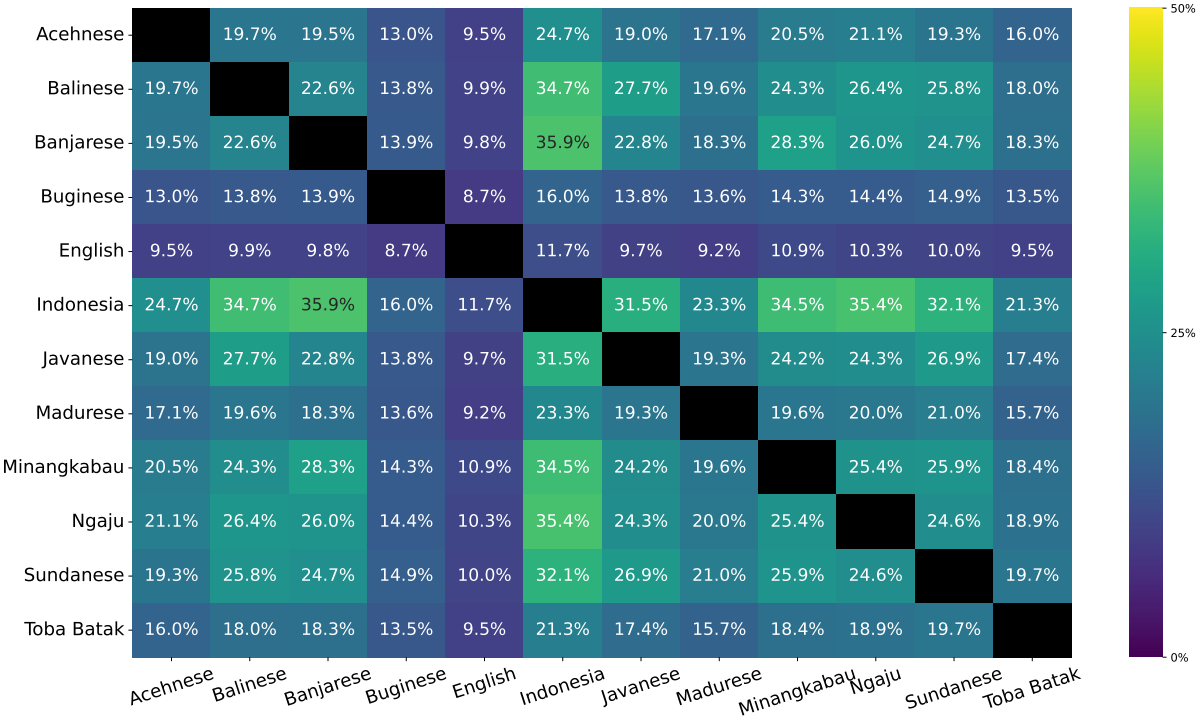

Figure 3: Vocabulary overlap between language pairs in NusaX dataset.

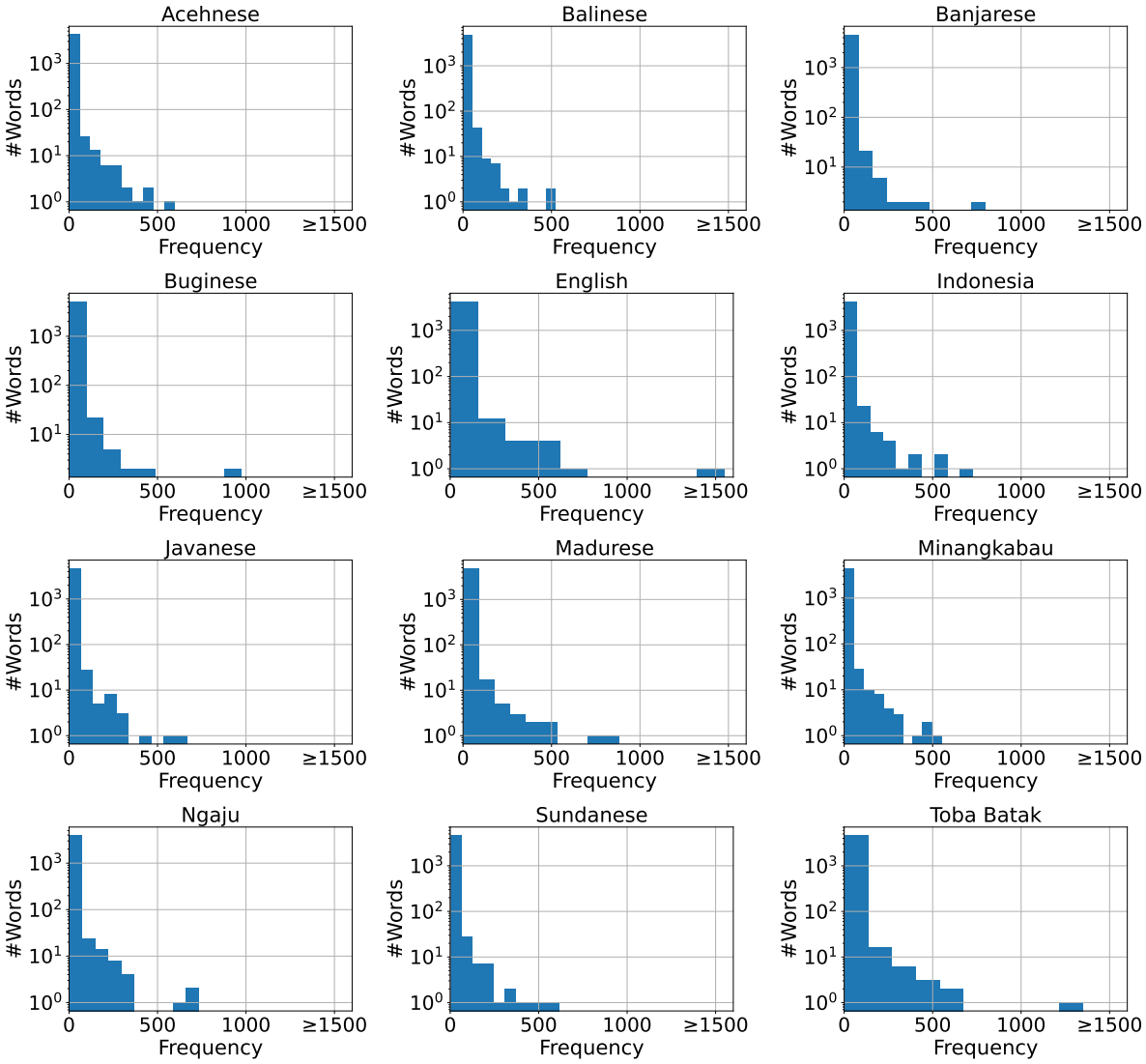

Figure 4: Word frequency histogram for each language in NusaX.

Toba Batak (4681 words)	**Balinese (4927 words)**	**Banjarese 4631 words**
na	lan	nang
di	sane	wan
dohot	ring	di
ni	ane	kada
tu	ne	ulun
do	sajan	nyaman
dang	tiyang	gasan
pe	tiang	banar
tabo	ajak	makan

Minangkabau (446 words)	**English (4233 words)**	**Ngaju (4005 words)**
di	the	te
nan	and	dengan
dan	to	ji
jo	is	mangat
untuak	a	eka
awak	of	akan
yang	for	aku
lamak	in	jadi
ka	I	diak

Sundanese (4693 words)	**Buginese (5118 words)**	**Indonesian (4269 words)**
nu	e	yang
sareng	na	di
di	okko	dan
teu	sibawa	tidak
pisan	iya	saya
abdi	de	dengan
ka	i	ini
ieu	ko	enak
aya	ladde	untuk

Javanese (4719 words)	**Acehnese (4250 words)**	**Maduranese (4846 words)**
sing	nyang	se
lan	ngon	e
ora	hana	bik
karo	lon	engkok
aku	that	ben
ing	mangat	tak
iki	nyoe	nyaman
ning	dan	ka
enak	bak	ghebey

Table 12: Vocabulary size (in bracket) and top-10 words on each language in the NusaX dataset.