5-31-2022

NusaX: Multilingual Parallel Sentiment Dataset for 10 Indonesian Local Languages

Genta Indra Winata
Bloomberg, United States

Alham Fikri Aji
Amazon, United States

Samuel Cahyawijaya
HKUST, Hong Kong

Rahmad Mahendra
Universitas Indonesia, Indonesia & INACL, Indonesia

Fajri Koto
The University of Melbourne, Australia

Recommended Citation

Authors

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, and Jey Han Lau

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[6*], **Ade Romadhony**[5,7*], **Kemal Kurniawan**[5,6*], **David Moeljadi**[8],
Radityo Eko Prasojo[9], **Pascale Fung**[3], **Timothy Baldwin**[6,10], **Jey Han Lau**[6],
Rico Sennrich[11], **Sebastian Ruder**[12]

[1]Bloomberg   [2]Amazon   [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]MBZUAI   [11]University of Zurich   [12]Google Research

Abstract

Natural language processing (NLP) has a 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 Chinese, while it remains inaccessible to many languages due to the unavailability of data resources and benchmarks. In this work, we focus on developing resources for languages in 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 datasets, a multi-task benchmark, and lexicons, as well as a parallel Indonesian-English dataset. We provide extensive analyses and describe the challenges when creating such resources. We hope that our work can spark NLP research on Indonesian and other underrepresented languages.

1 Introduction

Indonesia is one of the most populous and diverse countries, 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 receive little attention by the NLP community. Public datasets are practically non-existent, which prevents the global research community from exploring these languages. To this end, we introduce **NusaX**, a high-quality multilingual parallel corpus that covers 10 local languages in Indonesia, namely Acehnese, Balinese, Banjarese, Buginese, Madurese, Minangkabau, Javanese, Ngaju, Sundanese, and Toba Batak.

The **NusaX** dataset is 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 to ensure the quality of the translation. Sentiment analysis is one of the most popular NLP tasks (Muhammad et al., 2022) and has many practical uses in industry (Cambria et al., 2017). Sentiment analysis has been explored in many applications in Indonesia, including but not limited to presidential elections (Ibrahim et al., 2015; Budiharto and Meiliana, 2018), product reviews (Fauzi, 2019), stock prices (Cakra and Trisedya, 2015; Sagala et al., 2020), and COVID-19 (Nurdeni et al., 2021). By translating existing text, we additionally produce a parallel corpus, which is useful for building and evaluating translation systems. As we translate from a related high-resource language (Indonesian), we ensure that the topics and entities reflected in the data are culturally relevant to the other languages, which is 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 create a new benchmark and evaluate the performance of existing Indonesian language models, multilingual language models, and classical machine learning methods in few-shot and full-data settings.

Our contributions are as follows:

- We propose NusaX[1], the first high-quality human annotated parallel corpus in 10 languages from Indonesia, and corresponding parallel data in Indonesian and English.

- We introduce datasets in these local languages covering two tasks: sentiment analysis and machine translation.

---

* These authors contributed equally.
† Work done prior to joining Amazon.

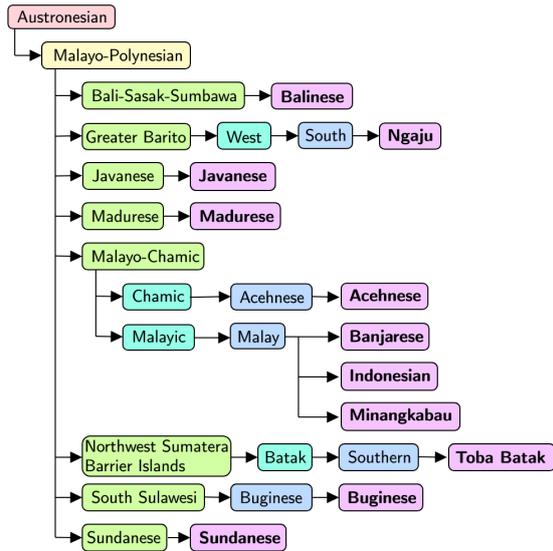[1]The code and dataset is released at https://github.com/IndoNLP/nusax

Figure 1: Language taxonomy of the studied Indonesian languages 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.

- We provide an extensive evaluation covering deep learning and classical NLP and machine learning methods on downstream tasks in few-shot and full-data settings.
- We conduct comprehensive analyses regarding the similarity of the languages under study from both a linguistics and empirical perspective, the cross-lingual transferability of existing monolingual and multilingual pre-trained models, and an efficiency analysis of different approaches for NLP tasks in extremely low-resource languages.

2 Focus Languages

In this work, we focus on 10 languages, namely Acehnese, Balinese, Banjarese, Buginese, Madurese, Minangkabau, Javanese, Ngaju, Sundanese, and Toba Batak. Figure 1 shows the taxonomy of these languages (including Indonesian) according to Ethnologue (Eberhard et al., 2021). The languages covered belong to the Austronesian language family under the Malayo-Polynesian subgroup. Table 1 provides additional information such as their status, number of L1 speakers, number of dialects, and writing systems.

Acehnese (ace) is a language spoken mainly in the Aceh province. Although it is the de facto language of provincial identity of Aceh, language use

Language name	Language status	No. of L1 speakers	No. of dialects	Writing system
Acehnese (ace)	Threatened	3,500,000 (2000 census)	7	Latin
Balinese (ban)	Developing	3,300,000 (2010 census)	3	Latin, Balinese
Banjarese (bjn)	Wider communication	3,650,000 (2015 UNSD)	2	Latin, Arabic
Toba Batak (bbc)	Threatened	2,000,000 (1991 UBS)	Not known	Latin, Batak
Buginese (bug)	Wider communication	3,870,000 (2010 census)	11	Latin, Buginese
Indonesian (ind)	National	42,800,000 (2015 UNSD)	Not known	Latin, Arabic
Javanese (jav)	Educational	68,200,000 (2015 UNSD)	11	Latin, Javanese
Madurese (mad)	Developing	7,790,000 (2015 UNSD)	6	Latin
Minangkabau (min)	Developing	4,240,000 (2015 UNSD)	12	Latin
Ngaju (nij)	Wider communication	890,000 (2003)	7	Latin
Sundanese (sun)	Developing	32,400,000 (2015 UNSD)	3	Latin, Arabic, Sundanese, Javanese

Table 1: Language data according to Ethnologue (Eberhard et al., 2021). We list the canonical language name, ISO 639-2 code, status, number of L1 speakers in Indonesia, number of dialects, and writing systems. The language status information can be found at https://www.ethnologue.com/about/language-status.

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

Balinese (ban) is a language spoken mainly in the Bali province and in the West Nusa Tenggara province. It has three main dialects: Highland Balinese, Lowland Balinese, and Nusa Penida. It is mainly written in the Latin script since the early 20th century although it has its own Balinese script. The word order in Balinese is SVO. It is non-tonal and has 17 consonant and 6 vowel

phonemes. Stress is on the penultimate syllable. It has three sociolinguistic registers. Regarding patterns of verb affixation, Balinese is an 'active' or 'split-S' language: verbs with Undergoer-like subject arguments are marked in one way (with a 'zero prefix'), while verbs with Actor-like subject arguments—intransitive or transitive—are marked in another (either with the nasal prefix 'N-', or with 'ma-') (Arka, 2003).

Banjarese (bjn) is a language spoken in Kalimantan (Central, East, South, and West Kalimantan provinces). It became a language of wider communication through trade in the market, in business, and in media. It is dominant in the South Kalimantan Province and also growing rapidly in the Central and Eastern Kalimantan provinces. It has two main dialects: Kuala and Hulu dialects. Although it is a Malayic language, it has many Javanese loanwords, probably acquired during the Majapahit period from the late thirteenth century until the fifteenth century (Blust et al., 2013). It has 73% of lexical similarity with Indonesian[2] and it is written in Arabic and Latin scripts (Eberhard et al., 2021).

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

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

Indonesian (ind) is the national language of Indonesia in 1945 Constitution, Article 36 (Indonesia, 2002). Its lexical similarity to Standard Malay

is over 80%. The word order is SVO. It is non-tonal and has 19 consonants, 6 vowels, and 3 diphthongs. The stress is on the penultimate syllable. It has a rich affixation system, including a variety of prefixes, suffixes, circumfixes, and reduplication. Most of the affixes in Indonesian are derivational. It is developed from literary 'Classical Malay' of the Riau-Johor sultanate (Sneddon, 2003) and has regional variants. It is written mainly in Latin script.

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

Minangkabau (min) is a language spoken mainly in West Sumatra and other provinces on Sumatra Island such as Bengkulu and Riau. Although it is classified as Malay, it is not intelligible with Indonesian. The word order is SVO written in Latin script. Standard Minangkabau voice can be characterised as an Indonesian-type system whereas colloquial Minangkabau voice is more effectively characterised as a Sundic-type system (Crouch, 2009).

Javanese (jav) is a language spoken mainly in Java island. It is the de facto language of provincial identity in central and eastern Java. The word order is SVO. It has 21 consonants and 8 vowels. It used to be written in Javanese script but since 20th century is mostly written in Latin script. Javanese differs from most other languages of western Indonesia in contrasting dental and retroflex stops, and in the feature of breathy voice or murmur as a phonetic property of its voiced obstruents. Javanese also differs from most languages of the Philippines and western Indonesia in allowing a number of word-initial consonant clusters. It has an elaborate system of speech levels (Blust et al., 2013).

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

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

bori and Shibata, 1988).

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

3 Data Construction

Our data collection process consists of several steps. First, we take an existing dataset in a related high-resource language (Indonesian) as a base for expansion to the other ten languages. Then, we ask human annotators to translate the data. 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 that by the fact that most languages have several dialects (cf. Table 1). 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 native speaker annotators for each local language who still practice the local language in daily communication in addition to Indonesian (see Appendix A). We run a preliminary test to filter the annotators to be recruited: We first ask annotators to translate three samples. We then conduct a peer review by asking asking annotators whether they can understand the translations of other annotators for the same language. The hired annotators work as translators, as well as translation validators. For English translations, we hire annotators based on their English proficiency test with IELTS $\geq$ 6.5 or TOEFL PBT

$\geq$ 600.

3.2 Data Filtering and Sampling

We take an existing sentence-level Indonesian sentiment analysis data, SMSA, the largest publicly available Indonesian sentiment analysis dataset containing comments and reviews from the IndoNLU benchmark (Purwarianti and Crisdayanti, 2019; Wilie et al., 2020) as the data source to build the multilingual parallel corpus. SMSA is a sentence-level multi-domain expertly-annotated sentiment analysis dataset consisting of more than 11,000 data instances of comments and reviews collected from several online platforms such as Twitter, Zomato, TripAdvisor, etc. We filter the data that contains abusive language and personal information by manually inspecting all sentences. Finally, we randomly select 1,000 samples via stratified sampling for translation, ensuring that the label distribution is balanced.

3.3 Human Translation

We instruct 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 instruct them to (1) retain the entity; (2) maintain the complete information content of the original sentence; (3) maintain the sentence's sentiment polarity. Initially, we also 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 vowel and punctuations. When the sentence is translated into the target local language, direct translation can sound unnatural. For example: 'kangeeeen' (ind) (translation in eng: 'miss'). The translation in Minangkabau is 'taragak' (as it is quite unnatural to express it as 'taragaaak'). Similarly, the original sentence may also contain typos. Due to the difficulty of accurately assessing typographical consistency of translations, we remove this as a criterion. We instruct annotators to use the Latin script as it is the most used writing system among locals.

3.4 Human-Assisted Quality Assurance

We conduct quality control (QC) between two annotators where annotator A is asked to check the translation result 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 to fix them.

We analyze the quality assurance edits for Balinese (ban), Minangkabau (min), Sundanese (sun), and Javanese (jav), 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 along the following categories:

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 being used by natives. The QC annotator might feel that one writing variant is more natural/commonly used, hence making 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.

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 2: Statistics of QC edits per category.

The result is shown in Table 2. Generally, word edits make up the majority of QC modifications, which involves changing a word/phase to its synonym or altering the morpheme slightly. In contrast, major changes are extremely rare. We also see some changes of 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.

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 10 local languages.

4.1.1 Sentiment Analysis

Sentiment analysis is an NLP task that aims to extract the sentiment from a given text document. The sentiment is commonly categorized into 3 classes, i.e., 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 its broad industry applications such as competitive 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, 100, and 400 train, validation, and test examples. In total, our dataset contains 6,000 train, 1,200 validation, and 4,800 test instances for all 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.

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
IndoBERT$_{BASE}$	75.4	74.8	70.0	83.1	73.9	79.5	90.0	81.7	**77.8**	82.5	**75.8**	77.5	78.5
IndoBERT$_{LARGE}$	76.3	**79.5**	74.0	83.2	70.9	87.3	90.2	**85.6**	77.2	82.9	**75.8**	77.2	**80.0**
IndoLEM$_{BASE}$	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
mBERT$_{BASE}$	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
XLM-R$_{BASE}$	73.9	72.8	62.3	76.6	66.6	90.8	88.4	78.9	69.7	79.1	75.0	80.1	76.2
XLM-R$_{LARGE}$	75.9	77.1	65.5	**86.3**	70.0	**92.6**	**91.6**	84.2	74.9	**83.1**	73.3	**86.0**	**80.0**

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

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 that are common for work with 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 simply copies the original Indonesian text, a dictionary-based substitution method, 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 language models (LMs) have brought substantial improvements in various NLP tasks. Despite the lack of resources in Indonesian and local languages, there are several efforts in developing large pre-trained language models for Indonesian and major local languages. IndoBERT (Wilie et al., 2020) and SundaneseBERT (Wongso et al., 2022) are two popular models for natural language understanding (NLU) tasks in Indonesian and Sundanese. Indo-BART and IndoGPT have also been introduced for natural language generation (NLG) tasks in Indonesian, Sundanese, and Javanese (Cahyawijaya et al., 2021). We employ these pretrained models as baselines to assess their adaptability to other languages.

4.2.3 Pre-trained Massively Multilingual LMs

We additionally consider large pre-trained multilingual language models 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 NLU tasks and mBART (Liu et al., 2020) and mT5 (Xue et al., 2021) models for NLG tasks. We provide the hyper-parameters of all models in Appendix B.

5 Results and Discussion

5.1 Overall Results

Sentiment Analysis Table 3 shows the sentiment analysis performance of various models across different local languages trained and evaluated using the same language data. Fine-tuned large LMs such as IndoBERT$_{LARGE}$ or XLM-R$_{LARGE}$ generally achieve the best performance. XLM-R models achieve strong performance on some languages, such as Indonesian (idn), Banjarese (bjn), English (eng), Javanese (jav), Minangkabau (min). We conjecture that this might be because these languages are included among the supported languages in the multilingual model. An exception is Banjarese, although it is very similar to Indonesian and Minangkabau. 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 compared to the neural methods. Logistic regression even outperforms both IndoBERT$_{LARGE}$ and XLM-R on Acehnese (ace), Buginese (bug) and Toba Batak (bbc). These results indicate that

						idn → 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 4: Results of the machine translation task from Indonesian to other languages (idn → x) in SacreBLEU.

						x → idn						
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 5: Results of the machine translation task on from other languages to Indonesian (x → idn) in SacreBLEU.

both Indonesian and multilingual pre-trained language models 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 XLR-R.

Machine Translation We show the results on machine translation in Table 4 (idn → x) and Table 5 (x → idn) in 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. Since these local languages also share similar grammatical structure with Indonesian, when we perform dictionary-based substitution (Word Substitution) we also see an improvement.

Both PBSMT and fine-tuned LMs demonstrate encouraging performance despite the limited training data, which we attribute to the target languages' similarity to Indonesian. We generally see a 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 BLEU

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.

5.2 Cross-lingual Capability of LMs

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

Empirically, Figure 2 shows the cross-lingual capability of the best performing model (XLM-R$_{LARGE}$) in the zero-shot cross-lingual setting for sentiment analysis. In general, most languages, except for Buginese (bug) and Toba Batak (bbc), can be used effectively as the source language, reaching on average ∼70-75% F1 compared to 80% average F1 in the monolingual setting (cf. XLM-R$_{LARGE}$ in Table 3). This empirical result aligns with the fact that both Buginese (bug) and Toba Batak (bbc) possess a very low vocabulary overlap with Indonesian (Aji et al., 2022). Interestingly, despite coming

from a completely different language family, English can also be effectively used as the source language for all 10 local languages.

In addition, we observe high BLEU scores across languages in the MT task for Banjarese (bjn) and Minangkabau (min) languages despite not being included in XLM-R's pre-training data. This may be due to Banjarese being similar to Malay (Nasution et al., 2021) while Minangkabau shares some words and syntax with Indonesian (Koto and Koto, 2020), suggesting positive transfer between closely related languages in the Malayo-Polynesian language family.

We can take advantage of language similarity by transferring knowledge from Indonesian and other local languages to perform zero-shot or a 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 may make the development of NLP datasets in low-resource languages more affordable than it may appear initially.
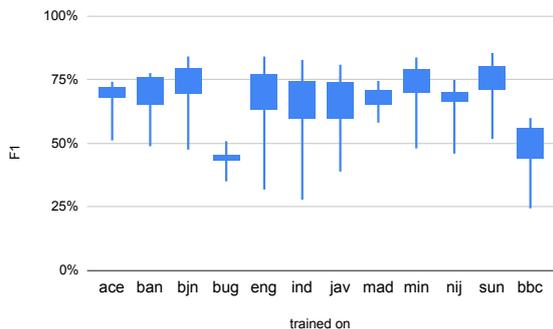

Figure 2: Zero-shot cross-lingual task results in the sentiment analysis task with XLM-R$_{\text{LARGE}}$. The model is trained with the language data on the x-axis and evaluated in all languages. The score represents the average among all languages.

6 Related Work

Multilingual Parallel Corpus Several multilingual parallel corpora have been developed to support studies on machine translation such as: the Global Communication Plan (GCP) corpus (Imamura and Sumita, 2018), Leipzig corpora (Goldhahn et al., 2012), JRC Acquis (Steinberger et al., 2006), TUFS Asian Language Parallel Corpus (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), and JW-300 (Agić and Vulić, 2019). The languages covered in the GCP corpus are Japanese, English, Spanish, French, and several Asian languages including Indonesian. While the GCP corpus covers sentence pairs in tourism related domains, JRC Acquis contains mostly legal documents in 20 official EU languages. Similar to the GCP corpus, TUFS Asian Corpus (TALPCo) covers Japanese sentences and their translations to English, Burmese, Malay, and Indonesian. Intercorp is a multilingual corpus based on Czech covering 27 other languages. DARPA LORELEI and the Asian Language Treebank cover several low resources languages, while FLORES contains sentences in two low resource languages, parallel with English. Building high-quality parallel corpora is expensive and time-consuming, since it requires translation by native or fluent translators. Guzmán et al. (2019) describe the procedure to generate high-quality translations as part of FLORES. The translation quality checks consist of two types of filtering: automatic and manual. Similar to FLORES, we also conducted quality checks of the translation.

Emerging Language Benchmarks Recently, datasets and benchmarks in underrepresented languages have emerged, such as MasakhaNER (Adelani et al., 2021), AmericasNLI (Ebrahimi et al., 2021), PMIndia (Haddow and Kirefu, 2020), Samanantar (Ramesh et al., 2022), and NaijaSenti (Muhammad et al., 2022). MasakhaNER covers ten African languages, while AmericasNLI covers ten indigenous languages of Americas and NaijaSenti covers four Nigerian languages. Particularly, for Indonesian languages, we have seen NLP benchmarks, such as IndoNLU (Wilie et al., 2020), IndoLEM (Koto et al., 2020), IndoNLG (Cahyawijaya et al., 2021), and IndoNLI (Mahendra et al., 2021). Other task-specific benchmarks include En-Id machine translation (Guntara et al., 2020), paraphrasing (Aji et al., 2021), and low-cost sequence classification and labeling in (Nityasya et al., 2021, 2022). These benchmarks mostly focus on Indonesian, except for IndoNLG, which also includes Javanese and Sundanese. PMIndia covers 13 Indic languages where each language pair (English with one of languages of India) contains a maximum of 56,000 sentence pairs, sourced from the website of India's prime minister. Samanantar covers 11 Indic languages, with a total of 49.7 million sentence

pairs where one source corpus is PMIndia.

Datasets for Indonesian 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 task involving Javanese, Sundanese, Madurese, Minangkabau, and Musi languages. 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).

7 Conclusion

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

8 Ethical Considerations

We cover low-resource languages which increase the accessibility of NLP research towards marginalize community. NusaX is created based on sentiment sentences taken from social media, reviews, and forums, therefore may contain biases towards certain groups or entities. Similarly, as our dataset was translated, there may be some translationese artifact on the resulting corpus. We argue that further study on the potential bias and the translationese effect of our dataset is needed.

Most of our best-performing baselines are based on large language models which require massive computational power to train and deploy. This computational cost is more significant given the lack of resources for local communities. We provide alternative classical-based approaches that are significantly more efficient, which still achieve competitive results.

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

Alham Fikri Aji, Genta Indra Winata, Fajri Koto, Samuel Cahyawijaya, Ade Romadhony, Rahmad Mahendra, Kemal Kurniawan, David Moeljadi, Raditityo 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.

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.

Erik Cambria, Dipankar Das, Sivaji Bandyopadhyay, and Antonio Feraco. 2017. Affective computing and sentiment analysis. In *A practical guide to sentiment analysis*, pages 1–10. Springer.

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

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

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.

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

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

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

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. 2021. Americasnli: Evaluating zero-shot natural language understanding of pretrained multilingual models in truly low-resource languages.

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. *arXiv preprint arXiv:1902.01382*.

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 Multitask 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)*.

Republik Indonesia. 2002. *Undang-Undang Dasar Negara Republik Indonesia Tahun 1945*. Sekretariat Jenderal MPR RI.

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.

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.

Shamsuddeen Hassan Muhammad, David Ifeoluwa Adelani, Sebastian Ruder, Ibrahim Said Ahmad, Idris Abdulmumin, Bello Shehu Bello, Monojit Choudhury, Chris Chinenye Emezue, Saheed Salahudeen Abdullahi, Anuoluwapo Aremu, Alipio Jeorge, and Pavel Brazdil. 2022. Naijasenti: A nigerian twitter sentiment corpus for multilingual sentiment analysis.

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, Rendi Chevi, Radityo Eko Prasojo, and Alham Fikri Aji. 2022. Which student is

best? a comprehensive knowledge distillation exam for task-specific bert models. *arXiv preprint arXiv:2201.00558*.

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.

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.

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.

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

Haryo Akbarianto Wibowo, Made Nindyatama Nityasya, Afra Feyza Akyürek, Suci Fitriany, Alham Fikri Aji, Radityo Eko Prasojo, and Derry Tanti Wijaya. 2021. IndoCollex: A testbed for morphological transformation of Indonesian 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, et al. 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.

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   Genta Indra Winata (Bloomberg), Alham Fikri Aji (Amazon), Ade Romadhony (Telkom University, Indonesia), Rahmad Mahendra (Universitas Indonesia), Fajri Koto (University of Melbourne), Samuel Cahyawijaya (Hong Kong University of Science and Technology), Kemal Kurniawan (University of Melbourne)

Dataset version   1.0 (May 2022)

Dataset citation   TODO

Data statement author   Kemal Kurniawan (University of Melbourne)

Data statement version   1.0 (February 2022)

Data statement citation   Kurniawan, Kemal. (2022). *Data Statement for NusaX*. Version 1.0. University of Melbourne. TODO url.

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?

---

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

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 6.

Language	ISO 639-3	Dialect
Acehnese	ace	Banda Aceh
Balinese	ban	Lowland
Toba Batak	bbc	Toba, Humbang
Banjarese	bjn	Hulu, Kuala
Buginese	bug	Sidrap
Javanese	jav	Matraman
Madurese	mad	Situbondo
Minangkabau	min	Padang, Agam
Ngaju	nij	Kapuas, Kahayan
Sundanese	sun	Priangan

Table 6: Local languages spoken in Indonesia (ID) that are covered in NusaX.

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 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 7: Hyperparameters of statistical models on sentiment analysis.

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 8: Hyperparameters of pre-trained language models on sentiment analysis. **Bold** denotes the best hyperparameter setting.

For statistical models, we use a spaCy as our toolkit, and we perform grid-search over the parameter ranges shown in Table 7 and select the best performing model over the devset. For all pretrained language models, we perform grid-search over batch size and learning rate while keeping the other hyperparameters fixed. The list of hyperparameters is shown in Table 8.

B.2 Machine Translation

Table 9 shows the hyperparameters of deep learning models on machine translation.

Hyperparams	IndoGPT	IndoBARTv2	mBART-50	mT5$_{\text{BASE}}$
learning rate	1e-4	1e-4	2e-5	1e-3
batch size		16		
gamma		0.98		
max epochs		20		
early stop		10		
seed		{1...5}		

Table 9: Hyperparameters of pretrained language models on machine translation.

C Examples

In Table 10, we provide several examples of the translated parallel data of NusaX corpus.

Language	Text
Indonesian	Ratusan rumah di medan terendam banjir
Translations	
English	Hundreds of houses in Medan were submerged by the flood
Acehnese	Meureutoh rumoh di Medan keunong ie raya
Balinese	Satusan umah ring medan merendem banjir
Banjarese	Ratusan rumah di medan tarandam banjir
Buginese	Maddatu bola okko medan nala lempe
Javanese	Atusan omah ing medan kebanjiran
Madurese	Ratosan bangko e medan tarendem banjir
Minangkabau	Ratuihan rumah di medan tarandam banjir
Ngaju	Ratusan huma hong medan lelep awi banjir
Sundanese	Ratusan bumi di medan karendem banjir
Toba Batak	Marratus jabu di medan na hona banji

Table 10: Example of the translation data.