IndoNLU: Benchmark and Resources for Evaluating Indonesian Natural Language Understanding

Bryan Wilie[1]*, **Karissa Vincentio**[2]*, **Genta Indra Winata**[3]*, **Samuel Cahyawijaya**[3]*, **Xiaohong Li**[4], **Zhi Yuan Lim**[4], **Sidik Soleman**[5], **Rahmad Mahendra**[6], **Pascale Fung**[3], **Syafri Bahar**[4], **Ayu Purwarianti**[1,5]

[1]Institut Teknologi Bandung [2]Universitas Multimedia Nusantara
[3]The Hong Kong University of Science and Technology
[4]Gojek [5]Prosa.ai [6]Universitas Indonesia

{bryanwilie92, karissavin}@gmail.com, {giwinata, scahyawijaya}@connect.ust.hk

Abstract

Although Indonesian is known to be the fourth most frequently used language over the internet, the research progress on this language in natural language processing (NLP) is slow-moving due to a lack of available resources. In response, we introduce the first-ever vast resource for training, evaluation, and benchmarking on Indonesian natural language understanding (IndoNLU) tasks. IndoNLU includes twelve tasks, ranging from single sentence classification to pair-sentences sequence labeling with different levels of complexity. The datasets for the tasks lie in different domains and styles to ensure task diversity. We also provide a set of Indonesian pre-trained models (IndoBERT) trained from a large and clean Indonesian dataset (Indo4B) collected from publicly available sources such as social media texts, blogs, news, and websites. We release baseline models for all twelve tasks, as well as the framework for benchmark evaluation, thus enabling everyone to benchmark their system performances.

1 Introduction

Following the notable success of contextual pre-trained language methods (Peters et al., 2018; Devlin et al., 2019), several benchmarks to gauge the progress of general-purpose NLP research, such as GLUE (Wang et al., 2018), SuperGLUE (Wang et al., 2019), and CLUE (Xu et al., 2020), have been proposed. These benchmarks cover a large range of tasks to measure how well pre-trained models achieve compared to humans. However, these metrics are limited to high-resource languages, such as English and Chinese, that already have existing datasets available and are accessible to the research community. Most languages, by contrast, suffer from limited data collection and low awareness of published data for research. One of the languages which suffer from this resource scarcity problem is Indonesian.

Indonesian is the fourth largest language used over the internet, with around 171 million users across the globe.[1] Despite a large amount of Indonesian data available over the internet, the advancement of NLP research in Indonesian is slow-moving. This problem occurs because available datasets are scattered, with a lack of documentation and minimal community engagement. Moreover, many existing studies in Indonesian NLP do not provide codes and test splits, making it impossible to reproduce results.

To address the data scarcity problem, we propose the first-ever Indonesian natural language understanding benchmark, IndoNLU, a collection of twelve diverse tasks. The tasks are mainly categorized based on the input, such as single-sentences and sentence-pairs, and objectives, such as sentence classification tasks and sequence labeling tasks. The benchmark is designed to cater to a range of styles in both formal and colloquial Indonesian, which are highly diverse. We collect a range of datasets from existing works: an emotion classification dataset (Saputri et al., 2018), QA factoid dataset (Purwarianti et al., 2007), sentiment analysis dataset (Purwarianti and Crisdayanti, 2019), aspect-based sentiment analysis dataset (Ilmania et al., 2018; Azhar et al., 2019), part-of-speech (POS) tag dataset (Dinakaramani et al., 2014; Hoesen and Purwarianti, 2018), named entity recognition (NER) dataset (Hoesen and Purwarianti, 2018), span extraction dataset (Mahfuzh et al., 2019; Septiandri and Sutiono, 2019; Fernando et al., 2019), and textual entailment dataset (Setya and Mahendra, 2018). It is difficult to compare model performance since there is no official

* These authors contributed equally.

[1] https://www.internetworldstats.com/stats3.htm

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

split of information for existing datasets. Therefore we standardize the benchmark by resplitting the datasets on each task for reproducibility purposes. To expedite the modeling and evaluation processes for this benchmark, we present samples of the model pre-training code and a framework to evaluate models in all downstream tasks. We will publish the score of our benchmark on a publicly accessible leaderboard to provide better community engagement and benchmark transparency.

To further advance Indonesian NLP research, we collect around four billion words from Indonesian preprocessed text data (≈ 23 GB), as a new standard dataset, called `Indo4B`, for self-supervised learning. The dataset comes from sources like online news, social media, Wikipedia, online articles, subtitles from video recordings, and parallel datasets. We then introduce an Indonesian BERT-based model, IndoBERT, which is trained on our `Indo4B` dataset. We also introduce another IndoBERT variant based on the ALBERT model (Lan et al., 2020), called IndoBERT-lite. The two variants of IndoBERT are used as baseline models in the `IndoNLU` benchmark. In this work, we also extensively compare our IndoBERT models to different pre-trained word embeddings and existing multilingual pre-trained models, such as Multilingual BERT (Devlin et al., 2019) and XLM-R (Conneau et al., 2019), to measure their effectiveness. Results show that our pre-trained models outperform most of the existing pre-trained models.

2 Related Work

Benchmarks GLUE (Wang et al., 2018) is a multi-task benchmark for natural language understanding (NLU) in the English language. It consists of nine tasks: single-sentence input, semantic similarity detection, and natural language inference (NLI) tasks. GLUE's harder counterpart SuperGLUE (Wang et al., 2019) covers question answering, NLI, co-reference resolution, and word sense disambiguation tasks. CLUE (Xu et al., 2020) is a Chinese NLU benchmark that includes a test set designed to probe a unique and specific linguistic phenomenon in the Chinese language. It consists of eight diverse tasks, including single-sentence, sentence-pair, and machine reading comprehension tasks. FLUE (Le et al., 2019) is an evaluation NLP benchmark for the French language which is divided into six different task categories: text classification, paraphrasing, NLI, parsing, POS tagging, and word sense disambiguation.

Contextual Language Models In recent years, contextual pre-trained language models have shown a major breakthrough in NLP, starting from ELMo (Peters et al., 2018). With the emergence of the transformer model (Vaswani et al., 2017), Devlin et al. (2019) proposed BERT, a faster architecture to train a language model that eliminates recurrences by applying a multi-head attention layer. Liu et al. (2019) later proposed RoBERTa, which improves the performance of BERT by applying dynamic masking, increasing the batch size, and removing the next-sentence prediction. Lan et al. (2020) proposed ALBERT, which extends the BERT model by applying factorization and weight sharing to reduce the number of parameters and time.

Many research studies have introduced contextual pre-trained language models on languages other than English. Cui et al. (2019) introduced the Chinese BERT and RoBERTa models, while Martin et al. (2019) and Le et al. (2019) introduced CamemBERT and FLAUBert respectively, which are BERT-based models for the French language. Devlin et al. (2019) introduced the Multilingual BERT model, a BERT model trained on monolingual Wikipedia data in many languages. Meanwhile, Lample and Conneau (2019) introduced XLM, a cross-lingual pre-trained language model that uses parallel data as a new translation masked loss to improve the cross-linguality. Finally, Conneau et al. (2019) introduced XLM-R, a RoBERTa-based XLM model.

3 `IndoNLU` Benchmark

In this section, we describe our benchmark as four components. Firstly, we introduce the 12 tasks in `IndoNLU` for Indonesian natural language understanding. Secondly, we introduce a large-scale Indonesian dataset for self-supervised pre-training models. Thirdly, we explain the various kinds of baseline models used in our `IndoNLU` benchmark. Lastly, we describe the evaluation metric used to standardize the scoring over different models in our `IndoNLU` benchmark.

3.1 Downstream Tasks

The `IndoNLU` downstream tasks covers 12 tasks divided into four categories: (a) single-sentence classification, (b) single-sentence sequence-tagging, (c) sentence-pair classification, and (d)

Dataset	|Train|	|Valid|	|Test|	Task Description	#Label	#Class	Domain	Style
Single-Sentence Classification Tasks								
EmoT†	3,521	440	442	emotion classification	1	5	tweets	colloquial
SmSA	11,000	1,260	500	sentiment analysis	1	3	general	colloquial
CASA	810	90	180	aspect-based sentiment analysis	6	3	automobile	colloquial
HoASA†	2,283	285	286	aspect-based sentiment analysis	10	4	hotel	colloquial
Sentence-Pair Classification Tasks								
WReTE†	300	50	100	textual entailment	1	2	wiki	formal
Single-Sentence Sequence Labeling Tasks								
POSP†	6,720	840	840	part-of-speech tagging	1	26	news	formal
BaPOS	8,000	1,000	1,029	part-of-speech tagging	1	41	news	formal
TermA	3,000	1,000	1,000	span extraction	1	5	hotel	colloquial
KEPS	800	200	247	span extraction	1	3	banking	colloquial
NERGrit†	1,672	209	209	named entity recognition	1	7	wiki	formal
NERP†	6,720	840	840	named entity recognition	1	11	news	formal
Sentence-Pair Sequence Labeling Tasks								
FacQA	2,495	311	311	span extraction	1	3	news	formal

Table 1: Task statistics and descriptions. †We create new splits for the dataset.

sentence-pair sequence labeling. The data samples for each task are shown in Appendix A.

3.1.1 Single-Sentence Classification Tasks

EmoT An emotion classification dataset collected from the social media platform Twitter (Saputri et al., 2018). The dataset consists of around 4000 Indonesian colloquial language tweets, covering five different emotion labels: anger, fear, happiness, love, and sadness.

SmSA This sentence-level sentiment analysis dataset (Purwarianti and Crisdayanti, 2019) is a collection of comments and reviews in Indonesian obtained from multiple online platforms. The text was crawled and then annotated by several Indonesian linguists to construct this dataset. There are three possible sentiments on the SmSA dataset: positive, negative, and neutral.

CASA An aspect-based sentiment analysis dataset consisting of around a thousand car reviews collected from multiple Indonesian online automobile platforms (Ilmania et al., 2018). The dataset covers six aspects of car quality. We define the task to be a multi-label classification task, where each label represents a sentiment for a single aspect with three possible values: positive, negative, and neutral.

HoASA An aspect-based sentiment analysis dataset consisting of hotel reviews collected from the hotel aggregator platform, AiryRooms (Azhar et al., 2019).[2] The dataset covers ten different aspects of hotel quality. Similar to the CASA dataset, each review is labeled with a single sentiment label for each aspect. There are four possible sentiment classes for each sentiment label: positive, negative, neutral, and positive-negative. The positive-negative label is given to a review that contains multiple sentiments of the same aspect but for different objects (e.g., cleanliness of bed and toilet).

3.1.2 Sentence-Pair Classification Task

WReTE The Wiki Revision Edits Textual Entailment dataset (Setya and Mahendra, 2018) consists of 450 sentence pairs constructed from Wikipedia revision history. The dataset contains pairs of sentences and binary semantic relations between the pairs. The data are labeled as entailed when the meaning of the second sentence can be derived from the first one, and not entailed otherwise.

3.1.3 Single-Sentence Sequence Labeling Tasks

POSP This Indonesian part-of-speech tagging (POS) dataset (Hoesen and Purwarianti, 2018) is collected from Indonesian news websites. The dataset consists of around 8000 sentences with 26 POS tags. The POS tag labels follow the Indonesian Association of Computational Linguistics (INACL) POS Tagging Convention.[3]

[2] https://github.com/annisanurulazhar/absa-playground

[3] http://inacl.id/inacl/wp-content/uploads/2017/06/INACL-POS-Tagging-Convention-26-Mei.pdf

Model	#Params	#Layers	#Heads	Emb. Size	Hidden Size	FFN Size	Language Type	Pre-train Emb. Type
Scratch	15.1M	6	10	300	300	3072	Mono	-
fastText-cc-id	15.1M	6	10	300	300	3072	Mono	Word Emb.
fastText-indo4b	15.1M	6	10	300	300	3072	Mono	Word Emb.
IndoBERT-lite$_{\text{BASE}}$	11.7M	12	12	128	768	3072	Mono	Contextual
IndoBERT$_{\text{BASE}}$	124.5M	12	12	768	768	3072	Mono	Contextual
IndoBERT-lite$_{\text{LARGE}}$	17.7M	24	16	128	1024	4096	Mono	Contextual
IndoBERT$_{\text{LARGE}}$	335.2M	24	16	1024	1024	4096	Mono	Contextual
mBERT	167.4M	12	12	768	768	3072	Multi	Contextual
XLM-R$_{\text{BASE}}$	278.7M	12	12	768	768	3072	Multi	Contextual
XLM-R$_{\text{LARGE}}$	561.0M	24	16	1024	1024	4096	Multi	Contextual
XLM-MLM$_{\text{LARGE}}$	573.2M	16	16	1280	1280	5120	Multi	Contextual

Table 2: The details of baseline models used in IndoNLU benchmark

BaPOS This POS tagging dataset (Dinakaramani et al., 2014) contains about 1000 sentences, collected from the PAN Localization Project.[4] In this dataset, each word is tagged by one of 23 POS tag classes.[5] Data splitting used in this benchmark follows the experimental setting used by Kurniawan and Aji (2018).

TermA This span-extraction dataset is collected from the hotel aggregator platform, AiryRooms (Septiandri and Sutiono, 2019; Fernando et al., 2019).[6] The dataset consists of thousands of hotel reviews, which each contain a span label for aspect and sentiment words representing the opinion of the reviewer on the corresponding aspect. The labels use Inside-Outside-Beginning (IOB) tagging representation with two kinds of tags, aspect and sentiment.

KEPS This keyphrase extraction dataset (Mahfuzh et al., 2019) consists of text from Twitter discussing banking products and services and is written in the Indonesian language. A phrase containing important information is considered a keyphrase. Text may contain one or more keyphrases since important phrases can be located at different positions. The dataset follows the IOB chunking format, which represents the position of the keyphrase.

NERGrit This NER dataset is taken from the Grit-ID repository,[7] and the labels are spans in IOB chunking representation. The dataset consists of three kinds of named entity tags, PERSON (name of person), PLACE (name of location), and ORGANIZATION (name of organization).

NERP This NER dataset (Hoesen and Purwarianti, 2018) contains texts collected from several Indonesian news websites. There are five labels available in this dataset, PER (name of person), LOC (name of location), IND (name of product or brand), EVT (name of the event), and FNB (name of food and beverage). Similar to the TermA dataset, the NERP dataset uses the IOB chunking format.

3.1.4 Sentence-Pair Sequence Labeling Task

FacQA The goal of the FacQA dataset is to find the answer to a question from a provided short passage from a news article (Purwarianti et al., 2007). Each row in the FacQA dataset consists of a question, a short passage, and a label phrase, which can be found inside the corresponding short passage. There are six categories of questions: date, location, name, organization, person, and quantitative.

3.2 Indo4B Dataset

Indonesian NLP development has struggled with the availability of data. To cope with this issue, we provide a large-scale dataset called Indo4B for building a self-supervised pre-trained model. Our self-supervised dataset consists of around 4B words, with around 250M sentences. The Indo4B dataset covers both formal and colloquial Indonesian sentences compiled from 12 datasets, of which two cover Indonesian colloquial language, eight cover formal Indonesian language, and the rest have a mixed style of both colloquial and formal. The statistics of our large-scale dataset can be

[4] http://www.panl10n.net/
[5] http://bahasa.cs.ui.ac.id/postag/downloads/Tagset.pdf
[6] https://github.com/jordhy97/final_project
[7] https://github.com/grit-id/nergrit-corpus

Dataset	# Words	# Sentences	Size	Style	Source
OSCAR (Ortiz Suárez et al., 2019)	2,279,761,186	148,698,472	14.9 GB	mixed	OSCAR
CoNLLu Common Crawl (Ginter et al., 2017)	905,920,488	77,715,412	6.1 GB	mixed	LINDAT/CLARIAH-CZ
OpenSubtitles (Lison and Tiedemann, 2016)	105,061,204	25,255,662	664.8 MB	mixed	OPUS OpenSubtitles
Twitter Crawl[2]	115,205,737	11,605,310	597.5 MB	colloquial	Twitter
Wikipedia Dump[1]	76,263,857	4,768,444	528.1 MB	formal	Wikipedia
Wikipedia CoNLLu (Ginter et al., 2017)	62,373,352	4,461,162	423.2 MB	formal	LINDAT/CLARIAH-CZ
Twitter UI[2] (Saputri et al., 2018)	16,637,641	1,423,212	88 MB	colloquial	Twitter
OPUS JW300 (Agić and Vulić, 2019)	8,002,490	586,911	52 MB	formal	OPUS
Tempo[3]	5,899,252	391,591	40.8 MB	formal	ILSP
Kompas[3]	3,671,715	220,555	25.5 MB	formal	ILSP
TED	1,483,786	111,759	9.9 MB	mixed	TED
BPPT	500,032	25,943	3.5 MB	formal	BPPT
Parallel Corpus	510,396	35,174	3.4 MB	formal	PAN Localization
TALPCo (Nomoto et al., 2018)	8,795	1,392	56.1 KB	formal	Tokyo University
Frog Storytelling (Moeljadi, 2012)	1,545	177	10.1 KB	mixed	Tokyo University
TOTAL	3,581,301,476	275,301,176	23.43 GB		

Table 3: `Indo4B` dataset statistics. [1] https://dumps.wikimedia.org/backup-index.html. [2] We crawl tweets from Twitter. The Twitter data will not be shared publicly due to restrictions of the Twitter Developer Policy and Agreement. [3] https://ilps.science.uva.nl/.

found in Table 3. We share the datasets that are listed in the table, except for those from Twitter due to restrictions of the Twitter Developer Policy and Agreement. The details of `Indo4B` dataset sources are shown in Appendix B.

3.3 Baselines

In this section, we explain the baseline models and the fine-tuning settings that we use in the `IndoNLU` benchmark.

3.3.1 Models

We provide a diverse set of baseline models, from a non-pre-trained model (scratch), to a word-embedding-based model, to contextualized language models. For the word-embeddings-based model, we use an existing fastText model trained on the Indonesian Common Crawl (CC-ID) dataset (Joulin et al., 2016; Grave et al., 2018).

fastText We build a fastText model with our large-scale self-supervised dataset, `Indo4B`, for comparison with the CC-ID fastText model and contextualized language model. For the models above and the fastText model, we use the transformer architecture (Vaswani et al., 2017). We experiment with different numbers of layers, 2, 4, and 6, for the transformer encoder. For the fastText model, we first pre-train the fastText embeddings with skipgram word representation and produce a 300-dimensional embedding vector. We then generate all required embeddings for each downstream task from the pre-trained fastText embeddings and cover all words in the vocabulary.

Contextualized Language Models We build our own Indonesian BERT and ALBERT models, named IndoBERT and IndoBERT-lite, respectively, in both base and large sizes. The details of our IndoBERT and IndoBERT-lite models are explained in Section 4. Aside from a monolingual model, we also provide multilingual model baselines such as Multilingual BERT (Devlin et al., 2019), XLM (Lample and Conneau, 2019), and XLM-R (Conneau et al., 2019). The details of each model are shown in Table 2.

3.3.2 Fine-tuning Settings

We fine-tune a pre-trained model for each task with initial learning with a range of learning rates [1e-5, 4e-5]. We apply a decay rate of [0.8, 0.9] for every epoch, and sample each batch with a size of 16 for all datasets except FacQA and POSP, for which we use a batch size of 8. To establish a benchmark, we keep a fixed setting, and we use an early stop on the validation score to choose the best model. The details of the fine-tuning hyperparameter settings used are shown in Appendix D.

3.4 Evaluation Metrics

We use the F1 score to measure the evaluation performance of all tasks. For the binary and multi-label classification tasks, we measure the macro-averaged F1 score by taking the top-1 prediction from the model. For the sequence labeling task, we calculate word-level sequence labeling macro-

Model	Maximum Sequence Length = 128				Maximum Sequence Length = 512			
	Batch Size	Learning Rate	Steps	Duration (Hr.)	Batch Size	Learning Rate	Steps	Duration (Hr.)
IndoBERT-lite$_{\text{BASE}}$	4096	0.00176	112.5 K	38	1024	0.00088	50 K	23
IndoBERT$_{\text{BASE}}$	256	0.00002	1 M	35	256	0.00002	68 K	9
IndoBERT-lite$_{\text{LARGE}}$	1024	0.00044	500 K	134	256	0.00044	129 K	45
IndoBERT$_{\text{LARGE}}$	256	0.0001	1 M	89	128	0.00008	120 K	32

Table 4: Hyperparameters and training duration for IndoBERT model pre-training.

averaged F1-score for all models by following the sequence labeling evaluation method described in the CoNLL evaluation script. We calculate two mean F1-scores separately for classification and sequence labeling tasks to evaluate models on our IndoNLU benchmark.

4 IndoBERT

In this section, we describe the details of our Indonesian contextualized models, IndoBERT and IndoBERT-lite, which are trained using our Indo4B dataset. We elucidate the extensive details of the models' development, first the dataset preprocessing, followed by the pre-training setup.

4.1 Preprocessing

Dataset Preparation To get the most beneficial next sentence prediction task training from the Indo4B dataset, we do either a paragraph separation or line separation if we notice document separator absence in the dataset. This document separation is crucial as it is used in the BERT architecture to extract long contiguous sequences (Devlin et al., 2019). A separation between sentences with a new line is also required to differentiate each sentence. These are used by BERT to create input embeddings out of sentence pairs that are compacted into a single sequence. We specify the number of duplication factors for each of the datasets differently due to the various formats of the datasets that we collected. We create duplicates on datasets with the end of document separators with a higher duplication factor. The preprocessing method is applied in both the IndoBERT and IndoBERT-lite models.

We keep the original form of a word to hold its contextual information since Indonesian words are built with rich morphological operations, such as compounding, affixation, and reduplication (Pisceldo et al., 2008). In addition, this setting is also suitable for contextual pre-training models that leverage inflections to improve the sentence-level representations.(Kutuzov and Kuzmenko, 2019)

Twitter data contains specific details, such as usernames, hashtags, emails, and URL hyperlinks. To preserve privacy and also to reduce noise, this private information in the Twitter UI dataset (Saputri et al., 2018) is masked into generics tokens such as <username>, <hashtag>, <email> and <links>. On the other hand, this information is discarded in the larger Twitter Crawl dataset.

Vocabulary For both the IndoBERT and the IndoBERT-lite models, we utilize SentencePiece (Kudo and Richardson, 2018) with a byte pair encoding (BPE) tokenizer as the vocabulary generation method. We use a vocab size of 30.522 for the IndoBERT models and vocab size of 30.000 for the IndoBERT-lite models.

4.2 Pre-training Setup

All IndoBERT models are trained on TPUv3-8 in two phases. In the first phase, we train the models with a maximum sequence length of 128. The training takes around 35, 89, 38 and 134 hours on IndoBERT$_{\text{BASE}}$, IndoBERT$_{\text{LARGE}}$, IndoBERT-lite$_{\text{BASE}}$, and IndoBERT-lite$_{\text{LARGE}}$, respectively. In the second phase, we continue the training of the IndoBERT models with a maximum sequence length of 512. It takes 9, 32, 23 and 45 hours on IndoBERT$_{\text{BASE}}$, IndoBERT$_{\text{LARGE}}$, IndoBERT-lite$_{\text{BASE}}$, and IndoBERT-lite$_{\text{LARGE}}$, respectively. The details of the pre-training hyperparameter settings are shown in Appendix D.

IndoBERT We use a batch size of 256 and a learning rate of 2e-5 in both training phases for IndoBERT$_{\text{BASE}}$, and we adjust the learning rate to 1e-4 for IndoBERT$_{\text{LARGE}}$ to stabilize the training. Due to memory limitation, we scale down the batch size to 128 and the learning rate to 8e-5 in the second phase of the training, with a number of training steps adapted accordingly. The base and large models are trained using the masked language modeling loss. We limit the maximum prediction per sequence into 20 tokens.

Model	Classification						Sequence Labeling							
	EmoT	SmSA	CASA	HoASA	WReTE	AVG	POSP	BaPOS	TermA	KEPS	NERGrit	NERP	FacQA	AVG
Scratch	57.31	67.35	67.15	76.28	64.35	66.49	86.78	70.24	70.36	39.40	5.80	30.66	5.00	44.03
fastText-cc-id	65.36	76.92	79.02	85.32	67.36	74.79	94.35	79.85	76.12	56.39	37.32	46.46	15.29	57.97
fastText-indo4b	69.23	82.13	82.20	85.88	60.42	75.97	94.94	81.77	74.43	56.70	38.69	46.79	14.65	58.28
mBERT	67.30	84.14	72.23	84.63	84.40	78.54	91.85	83.25	89.51	64.31	75.02	69.27	61.29	76.36
XLM-MLM	65.75	86.33	82.17	88.89	64.35	77.50	**95.87**	88.40	90.55	65.35	74.75	75.06	62.15	78.88
XLM-R_{BASE}	71.15	91.39	91.71	91.57	79.95	85.15	95.16	84.64	90.99	68.82	**79.09**	75.03	64.58	79.76
XLM-R_{LARGE}	78.51	92.35	92.40	**94.27**	83.82	88.27	92.73	87.03	91.45	70.88	78.26	78.52	**74.61**	**81.92**
IndoBERT-lite$_{BASE}$†	73.88	90.85	89.68	88.07	82.17	84.93	91.40	75.10	89.29	69.02	66.62	46.58	54.99	70.43
+ phase two	72.27	90.29	87.63	87.62	83.62	84.29	90.05	77.59	89.19	69.13	66.71	50.52	49.18	70.34
IndoBERT$_{BASE}$†	75.48	87.73	93.23	92.07	78.55	85.41	95.26	87.09	90.73	70.36	69.87	75.52	53.45	77.47
+ phase two	76.28	87.66	93.24	92.70	78.68	85.71	95.23	85.72	91.13	69.17	67.42	75.68	57.06	77.34
IndoBERT-lite$_{LARGE}$	75.19	88.66	90.99	89.53	78.98	84.67	91.56	83.74	90.23	67.89	71.19	74.37	65.50	77.78
+ phase two	70.80	88.61	88.13	91.05	**85.41**	84.80	94.53	84.91	90.72	68.55	73.07	74.89	62.87	78.51
IndoBERT$_{LARGE}$	77.08	**92.72**	**95.69**	93.75	82.91	**88.43**	95.71	**90.35**	91.87	71.18	77.60	**79.25**	62.48	81.21
+ phase two	**79.47**	92.03	94.94	93.38	80.30	88.02	95.34	87.36	**92.14**	**71.27**	76.63	77.99	68.09	81.26

Table 5: Results of baseline models with best performing configuration on the IndoNLU benchmark. Extensive experimental results are shown in Appendix E. Bold numbers are the best results among all. †The IndoBERT models are trained using two training phases.

IndoBERT-lite We follow the ALBERT pre-training hyperparameters setup (Lan et al., 2020) to pre-train the IndoBERT-lite models. We limit the maximum prediction per sequence into 20 tokens on the models, pre-training with whole word masked loss. We train the base model with a batch size of 4096 in the first phase, and 1024 in the second phase. Since we have a limitation in computation power, we use a smaller batch size of 1024 in the first phase and 256 in the second phase in training our large model.

5 Results and Analysis

In this section, we show the results of the IndoNLU benchmark and analyze the performance of our models in terms of downstream tasks score and performance-space trade-off. In addition, we show an analysis of the effectiveness of using our collected data compared to existing baselines.

5.1 Benchmark Results

Overall Performance As mentioned in Section 3, we fine-tune all baseline models mentioned in Section 3.3, and evaluate the model performance over all tasks, grouped into two categories, classification and sequence labeling. We can see in Table 5, that IndoBERT$_{LARGE}$, XLM-R$_{LARGE}$, and IndoBERT$_{BASE}$ achieve the top-3 best performance results on the classification tasks, and XLM-R$_{LARGE}$, IndoBERT$_{LARGE}$, and XLM-R$_{BASE}$ achieve the top-3 best performance results on the sequence labeling tasks. The experimental results also suggest that larger models have a performance advantage over smaller models. It is also evident that all pre-trained models outperform the scratch model, which shows the effectiveness of model pre-training. Another interesting observation is that all contextualized pre-trained models outperform word embeddings-based models by significant margins. This shows the superiority of the contextualized embeddings approach over the word embeddings approach.

5.2 Performance-Space Trade-off

Figure 1 shows the model performance with respect to the number of parameters. We can see two large clusters. On the bottom left, the scratch and fastText models appear, and they have the lowest F1 scores and the least floating points in the inference time. On the top right, we can see that the pre-trained models achieve decent performance, but in the inference time, they incur a high computation cost. Interestingly, in the top-left region, we can see the IndoBERT-lite models, which achieve similar performance to the IndoBERT models, but with many fewer parameters and a slightly lower computation cost.

5.3 Multilingual vs. Monolingual Models

Based on Table 5, we can conclude that contextualized monolingual models outperform contextualized multilingual models on the classification tasks by a large margin, but on the sequence labeling tasks, multilingual models tend to perform better compared to monolingual models and even perform much better on the NERGrit and FacQA tasks. As shown in Appendix A, both the NERGrit and FacQA tasks contain many entity names which

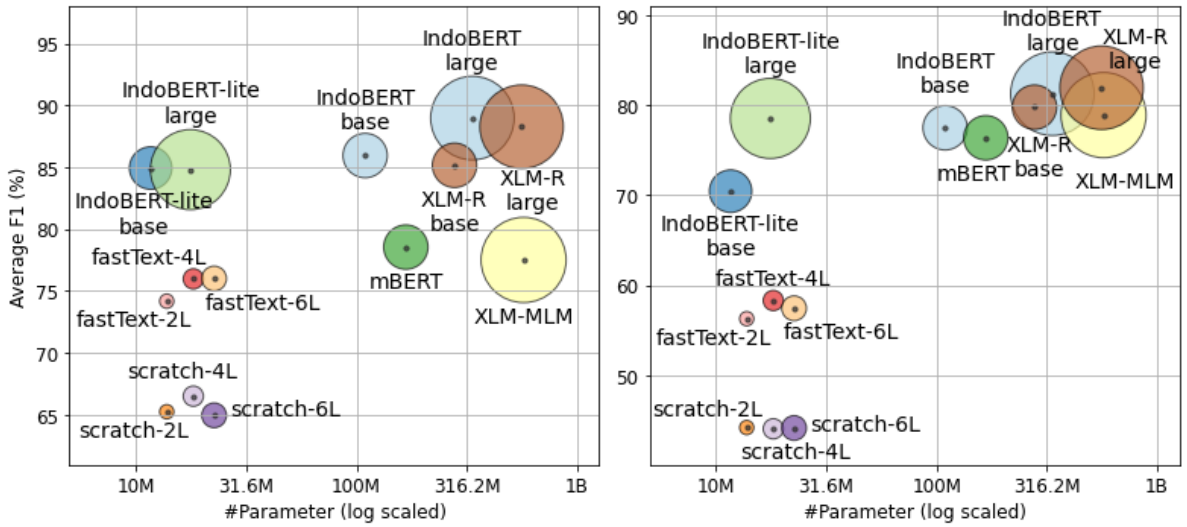

Figure 1: Performance-space trade-off for all baseline models on classification tasks (left) and sequence labeling tasks (right). We take the best model for each model size. 2L, 4L, and 6L denote the number of layers used in the model. The size of the dots represents the number of FLOPs of the model. We use python package `thop` taken from https://pypi.org/project/thop/ to calculate the number of FLOPs.

come from other languages, especially English. These facts suggest that monolingual models capture the semantic meaning of a word better than multilingual models, but multilingual models identify foreign terms better than monolingual models.

5.4 Effectiveness of `Indo4B` Dataset

Tasks	#Layer	fastText-cc-id	fastText-indo4b
Classification	2	72.00	**74.17**
	4	74.79	**75.97**
	6	74.80	**76.00**
Sequence Labeling	2	**56.26**	55.55
	4	57.97	**58.28**
	6	56.82	**57.42**

Table 6: Experiment results on fastText embeddings on IndoNLU tasks with different number of transformer layers

According to Grave et al. (2018), Common Crawl is a corpus containing over 24 TB.[8] We estimate the size of the `CC-ID` dataset to be around ≈ 180 GB uncompressed. Although the `Indo4B` dataset size is much smaller (≈ 23 GB), Table 6 shows us that the fastText models trained on the `Indo4B` dataset (fastText-indo4b) consistently outperform fastText models trained on the `CC-ID` dataset (fastText-cc-id) in both classification and sequence labeling tasks in all model settings. Based on Table 5, the fact that fastText-indo4b outperforms fastText-cc-id with a higher score on 10 out of 12 tasks suggests that a relatively smaller dataset (≈ 23 GB) can significantly outperform its larger counterpart (≈ 180 GB). We conclude that even though our `Indo4B` dataset is smaller, it covers more variety of the Indonesian language and has better text quality compared to the `CC-ID` dataset.

5.5 Effectiveness of IndoBERT and IndoBERT-lite

Table 5 shows that the IndoBERT models outperform the multilingual models on 8 out of 12 tasks. In general, the IndoBERT models achieve the highest average score on the classification task. We conjecture that monolingual models learn better sentiment-level semantics on both colloquial and formal language styles than multilingual models, even though the IndoBERT models' size is 40%–60% smaller. On sequence labeling tasks, the IndoBERT models cannot perform as well as the multilingual models (XLM-R) in three sequence labeling tasks: POSP, NERGrit, and FacQA. One of the possible explanations is that these datasets have many borrowed words from English, and multilingual models have the advantage in transferring learning from English.

Meanwhile, the IndoBERT-lite models achieve a decent performance on both classification and sequence labeling tasks with the advantage of compact size. Interestingly, the IndoBERT-lite$_{\text{LARGE}}$

[8] https://commoncrawl.github.io/cc-crawl-statistics/plots/languages

model performance is on par with that of XLM-R_{BASE} while having 16x fewer parameters. We also observe that increasing the maximum sequence length to 512 in phase two improves the performance on the sequence labeling tasks. Moreover, training the model with longer input sequences enables it to learn temporal information from a given text input.

6 Conclusion

We introduce the first Indonesian benchmark for natural language understanding, `IndoNLU`, which consists of 12 tasks, with different levels of difficulty, domains, and styles. To establish a strong baseline, we collect large clean Indonesian datasets into a dataset called `Indo4B`, which we use for training monolingual contextual pre-trained language models, called IndoBERT and IndoBERT-lite. We demonstrate the effectiveness of our dataset and our pre-trained models in capturing sentence-level semantics, and apply them to the classification and sequence labeling tasks. To help with the reproducibility of the benchmark, we release the pre-trained models, including the collected data and code. In order to accelerate the community engagement and benchmark transparency, we have set up a leaderboard website for the NLP community. We publish our leaderboard website at `https://indobenchmark.com/`.

Acknowledgments

We want to thank Cahya Wirawan, Pallavi Jain, Irene Gianni, Martijn Wieriks, Ade Romadhony, and Andrea Madotto for insightful discussions about this project. We sincerely thank the three anonymous reviewers for their insightful comments on our paper.

References

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.

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

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Ziqing Yang, Shijin Wang, and Guoping Hu. 2019. Pre-training with whole word masking for chinese bert.

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.

Arawinda Dinakaramani, Fam Rashel, Andry Luthfi, and Ruli Manurung. 2014. Designing an indonesian part of speech tagset and manually tagged indonesian corpus. In *2014 International Conference on Asian Language Processing, IALP 2014, Kuching, Malaysia, October 20-22, 2014*, pages 66–69. IEEE.

Jordhy Fernando, Masayu Leylia Khodra, and Ali Akbar Septiandri. 2019. Aspect and opinion terms extraction using double embeddings and attention mechanism for indonesian hotel reviews.

Filip Ginter, Jan Hajič, Juhani Luotolahti, Milan Straka, and Daniel Zeman. 2017. CoNLL 2017 shared task - automatically annotated raw texts and word embeddings. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

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

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

Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. 2016. Fasttext.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Kemal Kurniawan and Alham Fikri Aji. 2018. Toward a standardized and more accurate indonesian part-of-speech tagging. In *2018 International Conference on Asian Language Processing (IALP)*, pages 303–307. IEEE.

Andrey Kutuzov and Elizaveta Kuzmenko. 2019. To lemmatize or not to lemmatize: How word normalisation affects ELMo performance in word sense disambiguation. In *Proceedings of the First NLPL Workshop on Deep Learning for Natural Language Processing*, pages 22–28, Turku, Finland. Linköping University Electronic Press.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *CoRR*, abs/1901.07291.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. Albert: A lite bert for self-supervised learning of language representations. In *International Conference on Learning Representations*.

Hang Le, Loïc Vial, Jibril Frej, Vincent Segonne, Maximin Coavoux, Benjamin Lecouteux, Alexandre Allauzen, Benoît Crabbé, Laurent Besacier, and Didier Schwab. 2019. Flaubert: Unsupervised language model pre-training for french.

Pierre Lison and Jörg Tiedemann. 2016. Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles. In *Proceedings of the 10th International Conference on Language Resources and Evaluation*.

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

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

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric Villemonte de la Clergerie, Djamé Seddah, and Benoît Sagot. 2019. Camembert: a tasty french language model.

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

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.

Pedro Javier Ortiz Suárez, Benoît Sagot, and Laurent Romary. 2019. Asynchronous Pipeline for Processing Huge Corpora on Medium to Low Resource Infrastructures. In *7th Workshop on the Challenges in the Management of Large Corpora (CMLC-7)*, Cardiff, United Kingdom. Leibniz-Institut für Deutsche Sprache.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237.

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.

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.

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

Mei Silviana Saputri, Rahmad Mahendra, and Mirna Adriani. 2018. Emotion classification on indonesian twitter dataset. In *2018 International Conference on Asian Language Processing (IALP)*, pages 90–95. IEEE.

Ali Akbar Septiandri and Arie Pratama Sutiono. 2019. Aspect and opinion term extraction for aspect based sentiment analysis of hotel reviews using transfer learning.

Ken Nabila Setya and Rahmad Mahendra. 2018. Semi-supervised textual entailment on indonesian wikipedia data. In *2018 International Conference on Computational Linguistics and Intelligent Text Processing (CICLing)*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems*, pages 3261–3275.

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

Liang Xu, Xuanwei Zhang, Lu Li, Hai Hu, Chenjie Cao, Weitang Liu, Junyi Li, Yudong Li, Kai Sun, Yechen Xu, et al. 2020. Clue: A chinese language understanding evaluation benchmark. *arXiv preprint arXiv:2004.05986*.

A Data Samples

In this section, we show examples for downstream tasks in the IndoNLU benchmark.

- The examples of SmSA task are shown in Table 7.
- The examples of EmoT task are shown in Table 8.
- The examples of KEPS task are shown in Table 9.
- The examples of HoASA task are shown in Table 10.
- The examples of CASA task are shown in Table 11.
- The examples of WReTE task are shown in Table 12.
- The examples of NERGrit task are shown in Table 13.
- The examples of NERP task are shown in Table 14.
- The examples of BaPOS task are shown in Table 15.
- The examples of POSP task are shown in Table 16.
- The examples of FacQA task are shown in Table 17.
- The examples of TermA task are shown in Table 18.

B Indo4B Data Sources

In this section, we show the source of each dataset that we use to build our Indo4B dataset. The source of each corpus is shown in Table 19.

Sentence	Sentiment
pengecut dia itu , cuma bisa nantangin dari belakang saja	neg
wortel mengandung vitamin a yang bisa jaga kesehatan mata	neut
mocha float kfc itu minuman terenak yang pernah gue rasain	pos

Table 7: Sample data on task SmSA

Tweet	Emotion
Masalah ga akan pernah menjauh, hadapi Selasamu dengan penuh semangat!	happy
Sayang seribu sayang namun tak ada satupun yg nyangkut sampai sekarang	sadness
cewek suka bola itu dimata cowok cantiknya nambah, biarpun matanya panda	love

Table 8: Sample data on task EmoT

C Pre-Training Hyperparameters

In this section, we show all hyperparameters used in our IndoBERT and IndoBERT-lite training process. The hyperparameters is shown in Table 20.

D Fine-Tuning Hyperparameters

In this section, we show all hyperparameters used in the fine-tuning process of each baseline model. The hyperparameter configuration is shown in Table 21.

E Extensive Experiment Results on IndoNLU Benchmark

In this section, we show all experiments conducted in the IndoNLU benchmark. We use a batch size of 16 for all datasets except FacQA and POSP, for which we use a batch size of 8. The results of the full experiments are shown in Table 22.

Word	Layanan	BCA	Mobile	Banking	Bermasalah
Keyphrase	O	B	I	I	B
Word	Tidak	mengecewakan	pakai	BCA	Mobile
Keyphrase	O	O	B	B	I
Word	nggak	ada	tandingannya	e-channel	BCA
Keyphrase	B	I	I	B	I

Table 9: Sample data on task KEPS

Sentence	Aspect									
	AC	Air Panas	Bau	General	Kebersihan	Linen	Service	Sunrise Meal	TV	WiFi
air panas kurang berfungsi dan handuk lembab.	neut	neg	neut	neut	neut	neg	neut	neut	neut	neut
Shower zonk, resepsionis yang wanita judes	neut	neut	neut	neut	neut	neut	neg	neut	neut	neut
Kamar kurang bersih, terutama kamar mandi.	neut	neut	neut	neut	neg	neut	neut	neut	neut	neut

Table 10: Sample data on task HoASA

Sentence	Aspect					
	Fuel	Machine	Others	Part	Price	Service
bodi plus tampilan nya Avanza baru mantap juragan	neut	neut	neut	pos	neut	neut
udah gaya nya stylish ekonomis pula, beli calya deh	neut	neut	neut	pos	pos	neut
Mobil kualitas jelek kayak wuling saja masuk Indonesia	neut	neut	neg	neut	neut	neut

Table 11: Sample data on task CASA

Sentence A	Sentence B	Label
Anak sebaiknya menjalani tirah baring	Anak sebaiknya menjalani istirahat	Entail or Paraphrase
Kedua kata ini ditulis dengan huruf kanji yang sama	Jepang disebut Nippon atau Nihon dalam bahasa Jepang	Not Entail
Elektron hanya menduduki 0,06% massa total atom	Elektron hanya mengambil 0,06% massa total atom	Entail or Paraphrase

Table 12: Sample data on task WReTE

Word	Produser	David	Heyman	dan	sutradara	Mark	Herman	sedang	mencari	seseorang
Entity	O	B-PER	I-PER	O	O	B-PER	I-PERS	O	O	O
Word	Pada	tahun	1996	Williams	pindah	ke	Sebastopol	,	California	di
Entity	O	O	O	B-PER	O	O	B-PLA	O	B-PLA	O
Word	bekerja	untuk	penerbitan	perusahaan	teknologi	O	,	Reilly	Media	.
Entity	O	O	O	O	O	B-ORG	I-ORG	I-ORG	I-ORG	O

Table 13: Sample data on task NERGrit. PER = PERSON, ORG = ORGANIZATION, PLA = PLACE

Word	kepala	dinas	tata	kota	manado	amos	kenda	menyatakan	tidak	tahu
Entity	O	O	O	O	B-PLC	B-PPL	I-PPL	O	O	O
Word	telah	mendaftar	untuk	menjadi	official	merchant	bandung	great	sale	2017
Entity	O	O	O	O	O	O	B-EVT	I-EVT	I-EVT	I-EVT
Word	sekitar	timur	dan	barat	arnhem	,	katherine	dan	daerah	sekitar
Entity	O	B-PLC	O	B-PLC	I-PLC	O	B-PLC	O	O	O

Table 14: Sample data on task NERP. PLC = PLACE, PPL = PEOPLE, EVT = EVENT

Word	Pemerintah	kota	Delhi	mengerahkan	monyet	untuk	mengusir	monyet-monyet	lain	yang
Tag	B-NNP	B-NNP	B-NNP	B-VB	B-NN	B-SC	B-VB	B-NN	B-JJ	B-SC
Word	Beberapa	laporan	menyebutkan	setidaknya	10	monyet	ditempatkan	di	luar	arena
Tag	B-CD	B-NN	B-VB	B-RB	B-CD	B-NN	B-VB	B-IN	B-NN	B-NN
Word	berencana	mendatangkan	10	monyet	sejenis	dari	negara	bagian	Rajasthan	.
Tag	B-VB	B-VB	B-CD	B-NN	B-NN	B-IN	B-NNP	I-NNP	B-NNP	B-Z

Table 15: Sample data on task BaPOS. POS tag labels follow Universitas Indonesia POS Tag Standard. [9]

Word	kepala	dinas	tata	kota	manado	amos	kenda	menyatakan	tidak	tahu
Tag	B-NNO	B-VBP	B-NNO	B-NNO	B-NNP	B-NNP	B-NNP	B-VBT	B-NEG	B-VBI
Word	telah	mendaftar	untuk	menjadi	official	merchant	bandung	great	sale	2017
Tag	B-ADK	B-VBI	B-PPO	B-VBL	B-NNO	B-NNP	B-NNP	B-NNP	B-NNP	B-NUM
Word	sekitar	timur	dan	barat	arnhem	,	katherine	dan	daerah	sekitar
Tag	B-PPO	B-NNP	B-CCN	B-NNP	B-NNP	B-SYM	B-NNP	B-CCN	B-NNO	B-ADV

Table 16: Sample data on task POSP POS tag labels follow INACL POS Tagging Convention. [10]

Question	”Siapakah penasihat utama Presiden AS George W Bush?”					
Passage	Nasib	Karl	Rove	Akan	Segera	Diputuskan
Label	O	B	I	O	O	O
Question	”Dimana terjadinya letusan gunung berapi dahsyat tahun 1883?”					
Passage	Di	Kepulauan	Krakatau	Terdapat	400	Tanaman
Label	O	B	I	O	O	O
Question	”Perusahaan apakah yang sejak 1 Januari 2006, menurunkan harga pertamax dan pertamax plus?”					
Passage	Pesaing	Semakin	Banyak	,	Pertamina	Berusaha	Kompetitif
Label	O	O	O	O	B	O	O

Table 17: Sample data on task FacQA

Word	sayang	wifi	tidak	bagus	harus	keluar	kamar	.	fasilitas	lengkap
Entity	O	B-ASP	B-SEN	I-SEN	O	O	O	O	B-ASP	B-SEN
Word	pelayanan	nya	sangat	bagus	.	kamar	nya	juga	oke	.
Entity	B-ASP	I-ASP	B-SEN	I-SEN	O	B-ASP	I-ASP	O	B-SEN	O
Word	kamar	cukup	luas	,	interior	menarik	dan	unik	sekali	,
Entity	B-ASP	B-SEN	I-SEN	O	B-ASP	B-SEN	O	B-SEN	I-SEN	O

Table 18: Sample data on task TermA. SEN = SENTIMENT, ASP = ASPECT

Corpus Name	**Source**	**Public URL**
OSCAR	OSCAR	https://oscar-public.huma-num.fr/compressed/id_dedup.txt.gz
CoNLLu Common Crawl	LINDAT/CLARIAH-CZ	https://lindat.mff.cuni.cz/repository/xmlui/bitstream/handle/11234/1-1989/Indonesian-annotated-conll17.tar
OpenSubtitles	OPUS OpenSubtitles	http://opus.nlpl.eu/download.php?f=OpenSubtitles/v2016/mono/OpenSubtitles.raw.id.gz
Wikipedia Dump	Wikipedia	https://dumps.wikimedia.org/idwiki/20200401/idwiki-20200401-pages-articles-multistream.xml.bz2
Wikipedia CoNLLu	LINDAT/CLARIAH-CZ	https://lindat.mff.cuni.cz/repository/xmlui/bitstream/handle/11234/1-1989/Indonesian-annotated-conll17.tar
Twitter Crawl	Twitter	Not publicly available
Twitter UI	Twitter	Not publicly available
OPUS JW300	OPUS	http://opus.nlpl.eu/JW300.php
Tempo	ILSP	http://ilps.science.uva.nl/ilps/wp-content/uploads/sites/6/files/bahasaindonesia/tempo.zip
Kompas	ILSP	http://ilps.science.uva.nl/ilps/wp-content/uploads/sites/6/files/bahasaindonesia/kompas.zip
TED	TED	https://github.com/ajinkyakulkarni14/TED-Multilingual-Parallel-Corpus/tree/master/Monolingual_data
BPPT	BPPT	http://www.panl10n.net/english/outputs/Indonesia/BPPT/0902/BPPTIndToEngCorpusHalfM.zip
Parallel Corpus	PAN Localization	http://panl10n.net/english/outputs/Indonesia/UI/0802/Parallel/%20Corpus.zip
TALPCo	Tokyo University	https://github.com/matbahasa/TALPCo
Frog Storytelling	Tokyo University	https://github.com/davidmoeljadi/corpus-frog-storytelling

Table 19: `Indo4B` Corpus

Hyperparameter	IndoBERT$_{\text{BASE}}$	IndoBERT$_{\text{LARGE}}$	IndoBERT-lite$_{\text{BASE}}$	IndoBERT-lite$_{\text{LARGE}}$
attention_probs_dropout_prob	0.1	0.1	0	0
hidden_act	gelu	gelu	gelu	gelu
hidden_dropout_prob	0.1	0.1	0	0
embedding_size	768	1024	128	128
hidden_size	768	1024	768	1024
initializer_range	0.02	0.02	0.02	0.02
intermediate_size	3072	4096	3072	4096
max_position_embeddings	512	512	512	512
num_attention_heads	12	16	12	16
num_hidden_layers	12	24	12	24
type_vocab_size	2	2	2	2
vocab_size	30522	30522	30000	30000
num_hidden_groups	-	-	1	1
net_structure_type	-	-	0	0
gap_size	-	-	0	0
num_memory_blocks	-	-	0	0
inner_group_num	-	-	1	1
down_scale_factor	-	-	1	1

Table 20: Hyperparameter configurations for IndoBERT and IndoBERT-lite pre-trained models.

	batch_size	n_layers	n_epochs	lr	early_stop	gamma	max_norm	seed
Scratch	[8,16]	[2,4,6]	25	1e-4	12	0.9	10	42
fastText-cc-id	[8,16]	[2,4,6]	25	1e-4	12	0.9	10	42
fastText-indo4B	[8,16]	[2,4,6]	25	1e-4	12	0.9	10	42
mBERT	[8,16]	12	25	1e-5	12	0.9	10	42
XLM-MLM	[8,16]	16	25	1e-5	12	0.9	10	42
XLM-R_{BASE}	[8,16]	12	25	2e-5	12	0.9	10	42
XLM-R_{LARGE}	[8,16]	24	25	1e-5	12	0.9	10	42
IndoBERT-lite$_{BASE}$	[8,16]	12	25	1e-5	12	0.9	10	42
+ phase 2	[8,16]	12	25	1e-5	12	0.9	10	42
IndoBERT-lite$_{LARGE}$	[8,16]	24	25	[1e-5,2e-5]	12	0.9	10	42
+ phase 2	[8,16]	24	25	2e-5	12	0.9	10	42
IndoBERT$_{BASE}$	[8,16]	12	25	[1e-5,4e-5]	12	0.9	10	42
+ phase 2	[8,16]	12	25	4e-5	12	0.9	10	42
IndoBERT$_{LARGE}$	[8,16]	24	25	4e-5	12	0.9	10	42
+ phase 2	[8,16]	24	25	[3e-5,4e-5]	12	0.9	10	42

Table 21: Hyperparameter configurations for fine-tuning in IndoNLU benchmark. We use a batch size of 8 for POSP and FacQA, and a batch size of 16 for EmoT, SmSA, CASA, HoASA, WReTE, BaPOS, TermA, KEPS, NERGrit, and NERP.

Model	LR	# Layer	Param	Classification						Sequence Labeling							
				EmoT	SmSA	CASA	HoASA	WReTE	AVG	POSP	BaPOS	TermA	KEPS	NERGrit	NERP	FacQA	AVG
scratch	1e-4	2	38.6M	58.51	64.22	65.58	78.31	59.54	65.23	85.69	66.30	69.67	47.71	4.62	31.14	4.08	44.17
scratch	1e-4	4	52.8M	57.31	67.35	67.15	76.28	64.35	66.49	86.78	70.24	70.36	39.40	5.80	30.66	5.00	44.03
scratch	1e-4	6	67.0M	52.84	67.07	69.88	76.83	58.06	64.94	86.16	68.18	70.64	45.65	5.14	27.88	5.21	44.12
fasttext-cc-id-300-no-oov-uncased	1e-4	6	15.1M	67.43	78.84	81.61	85.01	61.13	74.80	94.36	78.45	77.26	57.28	26.70	46.36	17.3	56.82
fasttext-cc-id-300-no-oov-uncased	1e-4	4	10.7M	65.36	76.92	79.02	85.32	67.36	74.79	94.35	79.85	76.12	56.39	37.32	46.46	15.29	57.97
fasttext-cc-id-300-no-oov-uncased	1e-4	2	6.3M	64.74	76.71	75.39	78.05	65.11	72.00	94.42	78.12	73.45	55.22	33.27	45.44	13.89	56.26
fasttext-4B-id-300-no-oov-uncased	1e-4	6	15.1M	68.47	83.07	81.96	86.20	60.33	76.00	95.15	80.61	75.26	44.71	40.83	47.02	18.39	57.42
fasttext-4B-id-300-no-oov-uncased	1e-4	4	10.7M	69.23	82.13	82.20	85.88	60.42	75.97	94.94	81.77	74.43	56.70	38.69	46.79	14.65	58.28
fasttext-4B-id-300-no-oov-uncased	1e-4	2	6.3M	70.97	83.63	78.97	80.16	57.11	74.17	94.93	80.11	71.92	56.67	31.46	45.08	8.65	55.55
indobert-lite-base-128-112.5k	1e-5	12	11.7M	73.88	90.85	89.68	88.07	82.17	84.93	91.40	75.10	89.29	69.02	66.62	46.58	54.99	70.43
indobert-lite-base-128-191.5k	1e-5	12	11.7M	71.95	89.87	84.71	87.57	80.30	82.88	87.27	67.33	89.15	65.84	67.67	49.32	51.76	68.33
indobert-lite-base-512-162.5k	1e-5	12	11.7M	72.27	90.29	87.63	87.62	83.62	84.29	90.05	77.59	89.19	69.13	66.71	50.52	49.18	70.34
indobert-base-128	4e-5	12	124.5M	75.48	87.73	93.23	92.07	78.55	85.41	95.26	87.09	90.73	70.36	69.87	75.52	53.45	77.47
indobert-base-512	1e-5	12	124.5M	76.61	90.90	91.77	90.70	79.73	85.94	95.10	86.25	90.58	69.39	63.67	75.36	53.14	76.21
indobert-base-512	4e-5	12	124.5M	76.28	87.66	93.24	92.70	78.68	85.71	95.23	85.72	91.13	69.17	67.42	75.68	57.06	77.34
indobert-lite-large-128	1e-5	24	17.7M	75.19	88.66	90.99	89.53	78.98	84.67	91.56	83.74	90.23	67.89	71.19	74.37	65.50	77.78
indobert-lite-large-512	1e-5	24	17.7M	71.67	90.13	88.88	88.80	81.19	84.13	91.53	83.51	90.07	67.36	73.27	74.34	69.47	78.51
indobert-lite-large-512	2e-5	24	17.7M	70.80	88.61	88.13	91.05	85.41	84.80	94.53	84.91	90.72	68.55	73.07	74.89	62.87	78.51
indobert-large-128-1100k	4e-5	24	335.2M	77.04	93.71	96.64	93.27	84.17	88.97	95.71	89.74	91.97	70.82	70.76	77.54	67.27	80.55
indobert-large-128-1000k	4e-5	24	335.2M	77.08	92.72	95.69	93.75	82.91	88.43	95.71	90.35	91.87	71.18	77.60	79.25	62.48	81.21
indobert-large-512-1100k	4e-5	24	335.2M	77.39	92.90	95.90	93.77	81.62	88.32	95.25	86.05	91.92	69.71	75.20	77.53	69.86	80.79
indobert-large-512-1100k	3e-5	24	335.2M	79.47	92.03	94.94	93.38	80.30	88.02	95.34	87.36	92.14	71.27	76.63	77.99	68.09	81.26
bert-base-multilingual-uncased	1e-5	12	167.4M	67.30	84.14	72.23	84.63	84.40	78.54	91.85	83.25	89.51	64.31	75.02	69.27	61.29	76.36
xlm-mlm-100-1280	1e-5	16	573.2M	65.75	86.33	82.17	88.89	64.35	77.50	95.87	88.40	90.55	65.35	74.75	75.06	62.15	78.88
xlm-roberta-base	2e-5	12	278.7M	71.15	91.39	91.71	91.57	79.95	85.15	95.16	84.64	90.99	68.82	79.09	75.03	64.58	79.76
xlm-roberta-large	1e-5	24	561.0M	78.51	92.35	92.40	94.27	83.82	88.27	92.73	87.03	91.45	70.88	78.26	78.52	74.61	81.92

Table 22: Results of all experiments conducted in IndoNLU benchmark. We sample each batch with a size of 16 for all datasets except FacQA and POSP, for which we use a batch size of 8.