

Javanese part-of-speech tagging using cross-lingual transfer learning

Gabriel Enrique, Ika Alfina, Evi Yulianti

Faculty of Computer Science, Universitas Indonesia, Depok, Indonesia

Article Info

Article history:

Received Nov 13, 2023

Revised Mar 15, 2024

Accepted Mar 31, 2024

Keywords:

Cross-lingual transfer learning

Deep learning

Low-resource language

Part-of-speech tagging

Transformer

ABSTRACT

Large datasets that are publicly available for part-of-speech (POS) tagging do not always exist for some languages. One of those languages is Javanese, a local language in Indonesia, which is considered as a low-resource language. This research aims to examine the effectiveness of cross-lingual transfer learning for Javanese POS tagging by fine-tuning the state-of-the-art transformer-based models (such as IndoBERT, mBERT, and XLM-RoBERTa) using different kinds of source languages that have a higher resource (such as Indonesian, English, Uyghur, Latin, and Hungarian languages), and then fine-tuning it again using the Javanese language as the target language. We found that the models using cross-lingual transfer learning can increase the accuracy of the models without using cross-lingual transfer learning by 14.3%–15.3% over long short-time memory (LSTM)-based models, and by 0.21%–3.95% over transformer-based models. Our results show that the most accurate Javanese POS tagger model is XLM-RoBERTa that is fine-tuned in two stages (the first one using Indonesian language as the source language, and the second one using Javanese language as the target language), capable of achieving an accuracy of 87.65%.

This is an open access article under the [CC BY-SA](#) license.



Corresponding Author:

Evi Yulianti

Faculty of Computer Science, Universitas Indonesia

Building A, Faculty of Computer Science, Universitas Indonesia, Depok, Indonesia

Email: evi.y@cs.ui.ac.id

1. INTRODUCTION

Part-of-speech (POS) tagging is the process where each word in a sentence is categorized into its respective POS categories or POS tags, such as verb, noun, adjective. An example of this process can be seen in Figure 1. The Javanese sentence "bocah kuwi seneng nggambar sesawangan sing asri" ("the boy likes to paint beautiful scenery") in the figure is taken from the UD Javanese-CSUI [1] dataset. Usually, the POS tagging process starts with tokenizing the input sentence into words using a tokenizer. Each word will then be labeled with its POS tags accordingly by a POS tagger model. The POS tagging, which falls into the lexical analysis stage of natural language processing (NLP) [2], can be used for more complex tasks in NLP, such as question answering [3], stance detection [4], and information extraction [5]. Therefore, the performance of more complex NLP tasks might benefit from an accurate POS tagging model.

The majority of research in the POS tagging task uses machine learning approaches [6] which require a large dataset [7] in order to build an accurate model. However, large datasets that are publicly available for POS tagging do not always exist for some languages. One of those languages is Javanese, a local language in Indonesia. Although Javanese is the most spoken local language in Indonesia, the amount of labeled datasets for Javanese is still relatively small [8]. This makes Javanese one of the low-resource languages.

There are a few studies that have been conducted on Javanese POS tagging. Pratama *et al.* [9] used the hidden Markov model (HMM) to perform Javanese POS tagging on a dataset consisting of 1770 words, producing a model with an accuracy of 92.6%. Askhabi *et al.* [10] used the support vector machine (SVM) on a dataset with 3000 words, producing a model with an accuracy of 77%. However, those studies did not publish the datasets that they used to develop their POS tagger models, making it difficult to reproduce the results. It is different from the work of Alfina *et al.* [1] which used a publicly-available dataset for Javanese POS tagging. Alfina *et al.* used the multilingual bidirectional encoder representations from transformers (mBERT) [11] model, implemented using UDPipe v2.0 [12], producing a model with an F1-score of 87.22%. In this work, we also utilize the same dataset as Alfina *et al.*, but using cross-lingual transfer learning with three different Transformer-based models: XLMRoBERTa, mBERT, and IndoBERT. Note that none of the previous work has investigated the use of cross-lingual transfer learning for Javanese POS tagging.

One of the approaches that can be used to overcome the low-resource language problem is cross-lingual transfer learning. Cross-lingual transfer learning is the process where a model is trained for a certain task (e.g. POS tagging) using a source language, ideally a high-resource language, so that the model can be used to do the same task in another language (i.e. target language), usually a low-resource language [13], [14]. Some recent studies have shown that cross-lingual transfer learning can be used to overcome this problem as it can increase the performance of models for low-resource languages using the help of high-resource languages [13], [15], [16]. Considering that Javanese is one of the low-resource languages, this research aims to examine the effectiveness of cross-lingual transfer learning for Javanese POS tagging.

The combination between source and target languages becomes an important factor in cross-lingual transfer learning. Vries *et al.* [13] found that a good combination for source and target languages are languages that are similar, such as those with the same language family, the same writing system, or contain many overlapping vocabularies. Lin *et al.* [14] developed the LangRank tool (<https://github.com/neulab/langrank>) to rank the best source languages for a given target language for a specific task based on dataset-dependent and -independent features. LangRank has been shown to perform well in some previous work to choose source languages in cross-lingual transfer learning, such as in [15]. Therefore, this work also uses LangRank to help in choosing source languages for Javanese language.

The selection of base model is also another important factor in cross-lingual transfer learning. According to Lauscher *et al.* [16], some multilingual state-of-the-art transformer [17]-based models, such as mBERT [11] and XLM-RoBERTa [18], are the most commonly used model in cross-lingual transfer learning because of their good performance. Therefore, in this work, we also use mBERT and XLM-RoBERTa as our POS tagging models. In addition, we also use a monolingual language model, IndoBERT, because it has been reported to perform well in Indonesian POS tagging [19]. Since Indonesian and Javanese belong to the same language family [20], so it is intriguing to investigate the effectiveness of IndoBERT for Javanese POS tagging.

Besides showing the effect of cross-lingual transfer learning for Javanese POS tagging, this research indirectly shows the performance of state-of-the-art Transformer-based models for Javanese POS tagging without using cross-lingual transfer learning. It is important to note that most of the previous studies on Javanese POS tagging still use traditional machine learning methods, instead of transformer-based models. Overall, the contributions in this work are as follows:

- We propose using cross-lingual transfer learning for Javanese POS tagging to tackle the issue of low-resource data in Javanese language.
- We propose using some state-of-the-art transformer-based models for cross-lingual transfer learning in Javanese POS tagging.
- We examine the performance of some state-of-the-art transformer-based models without using cross-lingual transfer learning and deep-learning models based on long short term memory (LSTM) for Javanese POS tagging. Different from Alfina *et al.* [1] who explored the use of a Transformer-based model, mBERT, for Javanese POS tagging without cross-lingual transfer learning, we also investigate the use of IndoBERT and XLM-RoBERTa models in this work. In addition, we also study the effectiveness of powerful deep-learning models, LSTM and bidirectional long short term memory (BiLSTM), for Javanese POS tagging to be compared against our proposed methods using cross-lingual transfer learning. Note that this study also has not been researched in previous work.

The rest of this paper is organized as follows. Section 2 presents our methods to choose the best source language dataset for our target language (i.e., Javanese) as well as our transformer-based models (i.e., IndoBERT, mBERT, and XLM-RoBERTa) to perform Javanese POS tagging using cross-lingual transfer learn-

ing. Section 3 describes the results of our models against several baseline models. Section 4 discusses the key findings of our results and relates them to those of previous work. At last, section 5 concludes this study, while section 6 suggests some possible avenues for future work.

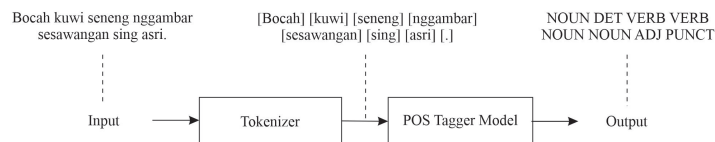


Figure 1. POS tagging process

2. METHOD

This section describes the research methodology that is applied in this research. It includes choosing the dataset (i.e., choosing the best source language for cross-lingual transfer learning in Javanese POS tagging), implementing the Javanese POS tagging models, and evaluating the models using the standard evaluation metrics. The details of each of these processes are explained in the following subsections.

2.1. Dataset

The development of POS tagger models using cross-lingual transfer learning requires the same tagset among all of the datasets [21]. Therefore, this research uses the datasets from universal dependencies (UD) v2.12 [22] as all of them uses the same tagset, consisting of 17 POS tags. UD has a lot of POS tagging datasets in various languages, making it flexible to choose the source languages. Moreover, LangRank, the framework that we use to help choosing the source languages, also evaluates its predictions using UD datasets. For Javanese language, UD only has one dataset, which is UD Javanese-CSUI [1]. It consists of 17 POS tags and 1,000 sentences with 14,344 words. Table 1 shows the list of POS tags in the UD Javanese-CSUI together with their description and word examples.

Table 1. Splitted UD Javanese-CSUI POS tag distribution

POS Tag	Description	Example w/ Translation
ADJ	Adjective	<i>apik</i> (well-made), <i>liya</i> (not-included), <i>bungah</i> (happy)
ADP	Adposition	<i>saka</i> (of), <i>karo</i> (second), <i>nalika</i> (that time)
ADV	Adverb	<i>wae</i> (only), <i>saiki</i> (now; this time), <i>maneh</i> (repeat)
AUX	Auxiliary	<i>wis</i> (done), <i>yaiku</i> (mean), <i>arep</i> (like; want)
CCONJ	Coordinating conjunction	<i>nanging</i> (but), <i>lan</i> (and), <i>karo</i> (1 also; 2 together)
DET	Determiner	<i>iku</i> (that), <i>punika</i> (that), <i>para</i> (for)
INTJ	Interjection	<i>inggih</i> (yes), <i>lha</i> , <i>oh</i>
NOUN	Noun	<i>tembang</i> (song), <i>wong</i> (human), <i>basa</i> (polite)
NUM	Numeral	<i>siji</i> (one), <i>rong</i> (two), <i>telung</i> (three)
PART	Particle	<i>ora</i> (no), <i>dudu</i> (not), <i>ya</i> (yes)
PRON	Pronoun	<i>sing</i> (which), <i>aku</i> (I; me)
PROPN	Proper noun	<i>Jawa</i> , <i>Indonesia</i> , <i>Ponorogo</i>
PUNCT	Punctuation	., ?, !
SCONJ	Subordinating conjunction	<i>kaya</i> (like), <i>supaya</i> (so)
SYM	Symbol	%, \$
VERB	Verb	<i>gawe</i> (work), <i>gelem</i> (can)
X	Other	<i>perpustakaan</i> (library), <i>rock</i>

UD Javanese-CSUI only has one split, thus we split it into train, dev, and test sets for our experiment using systematic random sampling with the proportion of 80:10:10 respectively. Table 2 shows the description of POS tags in UD Javanese-CSUI together with the word statistics of the POS tags for each split. The average number of words for each POS tag is 844. Then, the average number of words for each POS tag in each set (train, dev, test) are 677, 80, and 86 respectively.

To choose the source languages datasets, we run LangRank [14] evaluations using UD Javanese-CSUI. The five best source languages for the Javanese language according to LangRank can be seen in Table 3. We

then filter out the languages in which the dataset sizes are not significantly larger than UD Javanese-CSUI or those that have low UD scores. Here, we choose three languages to be used in our experiments: Uyghur, Tamil, and Latin. In addition to these languages, we also decide to use Indonesian and English as additional source languages. We include Indonesian because both Javanese and Indonesian belong to the same language family, which is Austronesian [20], so they are closely related. Then, we also decide to include English language because it is considered to be the common source language for cross-lingual transfer learning [13]. Finally, there are five source languages to be included in our cross-lingual transfer learning experiments for Javanese POS tagging: Indonesian, English, Uyghur, Latin, and Hungarian, as presented in Table 4.

Table 2. Splitted UD Javanese-CSUI POS tag distribution

POS Tag	#Words		
	Train	Val	Test
ADJ	579	71	85
ADP	595	75	77
ADV	633	65	100
AUX	270	24	46
CCONJ	254	31	23
DET	560	66	74
INTJ	26	1	5
NOUN	2309	287	275
NUM	293	37	32
PART	192	18	24
PRON	771	100	91
PROPN	1265	159	149
PUNCT	1804	211	217
SCONJ	241	24	49
SYM	10	1	1
VERB	1579	172	199
X	136	18	19

Table 3. Best source language ranking by LangRank

Rank	Source language dataset	LangRank score	Dataset size (words)
1	UD Uyghur-UDT	-0.26	40K
2	UD Tamil-TTB	-0.31	9K
3	UD Hungarian-Szeged	-0.33	42K
4	UD Latin-Perseus	-0.36	29K
5	UD Korean-GSD	-0.45	80K

Table 4. Datasets statistical information

Dataset	Set	Sentences	Words	UD score
UD Javanese-CSUI (target language)	train	800	11517	0.4933
	dev	100	1360	
	test	100	1466	
UD Indonesian-GSD (source language)	train	4482	97602	0.7331
	dev	559	12661	
	test	557	11756	
UD English-EWT (source language)	train	12544	204576	0.7119
	dev	2001	25149	
	test	2077	25094	
UD Uyghur-UDT (source language)	train	1656	19262	0.3552
	dev	900	10644	
	test	900	10330	
UD Latin-ITTB (source language)	train	22775	390785	0.6616
	dev	2101	29888	
	test	2101	29842	
UD Hungarian-Szeged (source language)	train	910	20166	0.7594
	dev	441	11418	
	test	449	10448	

We can see that the source language dataset with the highest number of sentences or words is Latin, followed by English and Indonesian. On the other hand, the source language dataset with the smallest size is Hungarian. The training, development, and testing split for each dataset was obtained directly from the UD dataset distribution.

2.2. Model

In this work, we investigate the use of three transformer-based models for cross-lingual transfer learning in Javanese POS tagging. They include two multilingual language models (mBERT and XLM-RoBERTa) and one monolingual language model (IndoBERT). The multilingual transformer-based models that we use are mBERT [11] and XLM-RoBERTa [18] as both models perform well for cross-lingual transfer learning scenarios [16]. Then, because we also want to see the performance of language-specific models for cross-lingual transfer learning, we also use an Indonesian-specific model, IndoBERT, because Indonesian shares the same language family as Javanese [20], which is Austronesian.

Bidirectional encoder representations from transformers (BERT) [11] uses a multi-layer bidirectional Transformer encoder from [17], providing bidirectional capabilities to understand the context from left-to-right and right-to-left. BERT can be fine-tuned to complete downstream tasks, including POS tagging, by simply adding an additional output layer and training the model using labeled task-specific data. In the original paper [11], BERT was reported to achieve superior performance in a range of NLP tasks. Because of its promising result, BERT has been pre-trained using a large corpus in various languages. mBERT is the multilingual version of BERT [11], pre-trained using 102 languages including Javanese, according to BERT's GitHub repository (<https://github.com/google-research/bert/tree/master>). While mBERT was pre-trained using various languages, IndoBERT was pre-trained using a single language only (i.e., Indonesian). IndoBERT [19] is a BERT-base model trained using a large amount of Indonesian corpus, consisting of 4 billion words. IndoBERT could achieve outstanding performance in various Indonesian NLP tasks, outperforming several state-of-the-art models in the IndoNLU [19] benchmark.

XLM-RoBERTa [18] is a multilingual transformer-based language model, pre-trained using 100 languages, including Javanese. XLM-RoBERTa performs better than previous multilingual language model such as mBERT [11] and XLM [23], and can even compete with state-of-the-art monolingual language model like RoBERTa [24]. XLM-RoBERTa can provide state-of-the-art performance by combining the XLM and RoBERTa models. XLM-RoBERTa was pre-trained using the same approach as XLM and using a larger dataset like RoBERTa, providing multilingual capabilities and increased performance.

2.3. Implementation

All of the transformer-based models that we use in this work are fine-tuned to perform Javanese POS tagging. The fine-tuning process is conducted by adding an additional output layer in the IndoBERT, mBERT, and XLM-RoBERTa architectures, and training the models using POS tagging data. We fine-tune all of these transformer-based models uniformly over three epochs with a batch size of 16, using AdamW optimizer with $5e-5$ as the initial learning rate and a 0.01 weight decay. We implement the transformer-based models using the hugging face transformers (<https://huggingface.co/docs/transformers/index>) and PyTorch (<https://pytorch.org/>) libraries.

There are three fine-tuning scenarios applied in this research, as illustrated in Figure 2. In the first scenario (baseline), all models are fine-tuned directly using the target language dataset, i.e., Javanese dataset. In the second scenario (zero-shot cross-lingual transfer learning), all models are fine-tuned using a source language dataset. In the third scenario (cross-lingual transfer learning), all models are first fine-tuned using a source language dataset, and then fine-tuned again using the target language dataset, i.e., Javanese dataset. All fine-tuned models across all scenarios will evaluate the test set of the Javanese dataset. Due to the possibility of incorrect tokenization by the tokenizers, we only use gold tokenization in the evaluation process.

To show the difference in performance between our transformer-based models using cross-lingual transfer learning and previous methods, we implement a state-of-the-art baseline method for Javanese POS tagging by Alfina *et al.* [1] that uses mBERT model in non-cross-lingual transfer learning setting (i.e., mBERT was fine-tuned directly using Javanese language), and the HMM method by Jurafsky and Martin [25] that was used in two of the previous studies on Javanese POS tagging [9]. To implement Alfina *et al.* baseline, we do not use UDPipe v2.0 [12], but using the libraries mentioned in the paragraph above. In addition, we also implement two additional baseline models using powerful deep-learning models, LSTM and BiLSTM. This is motivated by Can [26] who found that for small datasets, a small deep-learning model, such as LSTM, may perform better

than large transformer-based models, such as BERT. Considering that the Javanese dataset is relatively small in size, it is interesting to examine the effectiveness of LSTM and BiLSTM deep-learning models for Javanese POS tagging. We use Javanese fastText [27] as the word embedding for these LSTM-based models. We train these models uniformly over 20 epochs with a batch size of 16, using Adam optimizer with 1e-3 as the initial learning rate. These LSTM-based models are implemented using the Keras (<https://keras.io/>) library. Note that the HMM and LSTM-based baseline models also use scenario 1 illustrated in Figure 2, as we train the models directly using the Javanese language.

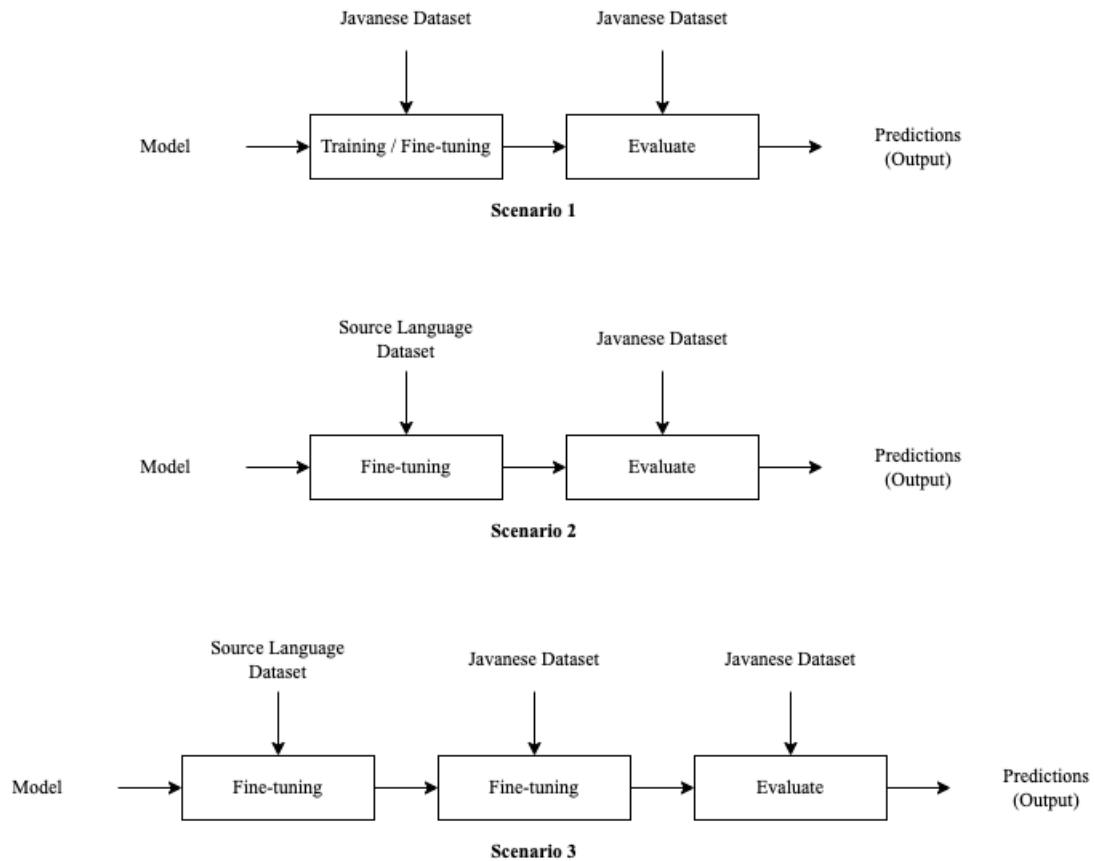


Figure 2. The flow of process for each fine-tuning scenario

2.4. Evaluation metrics

To evaluate the performance of the POS tagger models, we use F1-score and accuracy as the evaluation metrics. The F1-score is the harmonic mean between precision and recall, while the accuracy is the percentage of correctly tagged words [6]. We use macro average as the averaging method to calculate the F1-score for each POS tagger model, as F1-score can only be calculated for each POS tag [6].

3. RESULTS

The results of our experiment for all models across all scenarios can be seen in Table 5. Values displayed in bold represent the highest value for each scenario, while the underlined values represent the highest value for each model type. Models from scenario 3 perform better than models from scenarios 1 and 2. This shows that cross-lingual transfer learning implemented using a two-stage fine-tuning process (i.e., scenario 3), first using the source language and second using Javanese language, is ideal for Javanese POS tagging. The best model overall is the crosslingual-xlmroberta-id that uses a two-stage fine-tuning process using XLM-RoBERTa model and Indonesian as the source language. It gains an accuracy of 87.65%. Then,

models from scenario 2 are shown to perform worse than models from scenario 1. This shows that zero-shot cross-lingual transfer learning (i.e., scenario 2) is not ideal for Javanese POS tagging.

Table 5. Performance metrics for all models

Scenario	Model	F1-score (%)	Accuracy (%)
1	baseline-hmm (Pratama <i>et al.</i> , [9])	12.75	14.12
	baseline-lstm	63.66	75.99
	baseline-bilstm	66.19	76.74
	baseline-mbert (Alfina <i>et al.</i> [1])	69.58	85.54
	baseline-xlmroberta	67.44	83.70
	baseline-indobert	75.43	86.22
2	zeroshot-mbert-id	57.59	69.17
	zeroshot-mbert-en	49.69	62.69
	zeroshot-mbert-ug	14.06	36.77
	zeroshot-mbert-la	36.66	52.80
	zeroshot-mbert-hu	32.50	58.66
	zeroshot-xlmroberta-id	61.69	71.62
	zeroshot-xlmroberta-en	54.05	67.80
	zeroshot-xlmroberta-ug	31.49	54.37
	zeroshot-xlmroberta-la	41.24	58.59
	zeroshot-xlmroberta-hu	36.95	60.03
	zeroshot-indobert-id	33.40	47.41
	zeroshot-indobert-en	33.04	41.95
	zeroshot-indobert-ug	6.09	20.94
	zeroshot-indobert-la	18.31	39.22
	zeroshot-indobert-hu	18.88	35.88
3	crosslingual-mbert-id	79.77	87.31
	crosslingual-mbert-en	82.08	87.52
	crosslingual-mbert-ug	74.20	87.24
	crosslingual-mbert-la	72.72	87.04
	crosslingual-mbert-hu	71.84	86.97
	crosslingual-xlmroberta-id	79.96	87.65
	crosslingual-xlmroberta-en	83.63	86.63
	crosslingual-xlmroberta-ug	75.36	87.38
	crosslingual-xlmroberta-la	71.36	86.29
	crosslingual-xlmroberta-hu	71.82	86.29
	crosslingual-indobert-id	82.26	87.04
	crosslingual-indobert-en	82.33	86.02
	crosslingual-indobert-ug	75.54	86.43
	crosslingual-indobert-la	75.01	85.06
	crosslingual-indobert-hu	74.85	86.22

By averaging the accuracy scores for each source language across all models from cross-lingual transfer learning scenario (i.e., scenario 3), we can see the performance ranking between all source languages in Table 6. The best source language is Indonesian. The high similarity between Javanese and Indonesian is the most likely reason why Indonesian is the best source language for Javanese POS tagging.

Table 6. Best source language rankings based on the results

Rank	Language	Average Accuracy (%)
1	id	87.33
2	ug	87.02
3	en	86.72
4	hu	86.49
5	la	86.13

Table 7 shows two examples of the prediction results of the best-performing model in this work which uses cross-lingual transfer learning, i.e., crosslingual-xlmroberta-id. The results are compared with the non-cross-lingual transfer learning model counterparts, i.e., baseline-xlmroberta. The POS tag labels printed in red

represent the incorrect predictions. We can see from the table that the accuracy of the crosslingual-xlmroberta-id model predictions are 30% and 28.58% higher than the baseline-xlmroberta model predictions. Some POS tags that are incorrectly identified by the baseline-xlmroberta, could be predicted correctly by the crosslingual-xlmroberta-id model. This further shows that the cross-lingual transfer learning model can increase the performance of the non-cross-lingual transfer learning model in Javanese POS tagging.

Table 7. Crosslingual-xlmroberta-id performance increase from baseline-xlmroberta

Category	Grammar
English translation	In the seventh year they must be released without paying any kind of ransom
Javanese sentence	<i>Ing taun kapitu kudu diluwari tanpa mbayar tebusan apa - apa .</i>
Ground truth	ADP NOUN ADJ AUX VERB CONJ VERB NOUN PRON PUNCT
baseline-xlmroberta (Acc: 70%)	ADP NOUN NOUN AUX VERB ADP VERB NOUN ADV PUNCT
crosslingual-xlmroberta-id (Acc: 100%)	ADP NOUN ADJ AUX VERB CONJ VERB NOUN PRON PUNCT
English translation	Every morning Mr. Hamid tries to drive
Javanese sentence	<i>Saben esuk Pak Hamid nyoba nyopir .</i>
Ground truth POS tag	DET NOUN PROPN PROPN VERB VERB PUNCT
baseline-xlmroberta (Acc: 71.42%)	ADV ADV PROPN PROPN VERB VERB PUNCT
crosslingual-xlmroberta-id (Acc: 100%)	DET NOUN PROPN PROPN VERB VERB PUNCT

Next, we analyze the predictions given by the best model, i.e., crosslingual-xlmroberta-id, for each POS tag. The results are presented in Table 8. Values displayed in bold represent the highest value for each metric, while the underlined values represent the lowest value for each metric. We can see that the model perfectly predict all words with the CCONJ, PUNCT, and SYM tags. However, the model could not predict correctly for INTJ tag. This might be due to the low amount of INTJ words in the train set, making the model unable to familiarize itself with INTJ words. The POS tag with the least amount of words in the train set is SYM, with just 10 words, compared to INTJ with 26 words. But because SYM words are symbols, which is not similar at all with the other types of words, the model can still predict SYM words perfectly.

Table 8. Crosslingual-xlmroberta-id performance metric for each POS tag

POS Tag	Precision (%)	Recall (%)	F1-score (%)	Support
ADJ	69.07	78.82	73.63	85
ADP	89.33	87.01	88.16	77
ADV	80.68	71.00	75.53	100
AUX	89.80	95.65	92.63	46
CCONJ	100.00	100.00	100.00	23
DET	89.06	77.03	82.61	74
INTJ	0.00	0.00	0.00	5
NOUN	82.06	89.82	85.76	275
NUM	95.83	71.88	82.14	32
PART	90.48	79.17	84.44	24
PRON	80.77	92.31	86.15	91
PROP	96.67	97.32	96.99	149
PUNCT	100.00	100.00	100.00	217
SCONJ	93.62	89.80	91.67	49
SYM	100.00	100.00	100.00	1
VERB	86.00	86.43	86.22	199
X	80.00	21.05	33.33	19

Table 9 shows four examples of the prediction errors made by the crosslingual-xlmroberta-id model. We examine these examples to conduct an error analysis. An English translation of each Javanese sentence in the table is given in the row before the sentence. The POS tag labels printed in red represent the incorrect predictions. Some words in the dataset may have the PRON or DET tags depending on the context. Interestingly, the crosslingual-xlmroberta-id model sometimes could not distinguish between PRON and DET words. The words *e* and *ipun* are some of the examples of words that may have the PRON or DET tags depending on the context. We found that sometimes the model would predict the word *e*, which has the DET tag, as PRON; and

sometimes the model would predict the word *e*, which has the PRON tag, as DET. The same case happens with the word *ipun*. Examples of this error can be seen in the first, second, and third Javanese sentences in the table.

In the Javanese dataset, words that have the X POS tag are non-Javanese words, such as Indonesian and English. However, we analyze that the crosslingual-xlmroberta-id model still has Indonesian POS tagging capabilities because the source language of this model is Indonesian. Consequently, instead of predicting the X POS tag to a particular non-Javanese word, the model often gives out the Indonesian POS tag. Examples of this error can be seen in the third and fourth Javanese sentences in the table. In the third sentence, the model incorrectly predicts the Indonesian word *penyebab* as NOUN, while it should be X. The same case happens for Indonesian word *perpustakaan* in the fourth sentence in the table. This happens because the Indonesian POS tag for words *penyebab* and *perpustakaan* are NOUN.

Table 9. Crosslingual-xlmroberta-id prediction error examples

Category	Grammar
English translation	In the afternoon, Siti usually studies, her older brother reads the newspaper, and her younger sibling plays in the yard.
Javanese sentence	<i>Ing wayah sore biasane Siti sinau , kangmas e maca koran, lan adhi e dolan neng pekaran.</i>
Ground truth	ADP NOUN NOUN ADV PROPON VERB PUNCT NOUN PRON VERB NOUN PUNCT CONJ NOUN PRON VERB ADP NOUN PUNCT
Prediction	ADP NOUN NOUN ADV PROPON VERB PUNCT NOUN DET VERB NOUN PUNCT CONJ NOUN DET VERB ADP NOUN PUNCT
English translation	"Whereas he processes materials that are strategic for the benefit of the people,"said Endriartono.
Javanese sentence	<i>"Dene piyambakipun ngolah bakal ingkang strategis kangge kapentingan rakyat , ngendika ipun Endriartono .</i>
Ground truth	PUNCT ADV PRON VERB NOUN PRON ADJ ADP NOUN NOUN PUNCT PUNCT VERB DET PROPON
Prediction	PUNCT ADV PRON VERB ADV PRON ADJ ADP NOUN NOUN PUNCT PUNCT VERB PRON PROPON
English translation	Many unusual causes.
Javanese sentence	<i>Akeh penyebab e sing njalari mirunggan.</i>
Ground truth	DET X DET PRON VERB ADJ PUNCT
Prediction	DET NOUN PRON PRON VERB ADJ PUNCT
English translation	Lintang returned to his class, I went to the library by myself
Javanese sentence	<i>Lintang bali menyang kelase, aku dhewe mbacutne laku tekan perpustakaan.</i>
Ground truth	PROPN VERB ADP ADV PUNCT PRON DET VERB NOUN VERB X PUNCT
Prediction	PROPN VERB ADP NOUN PUNCT PRON DET VERB NOUN VERB NOUN PUNCT

4. DISCUSSION

Our results show that cross-lingual transfer learning models using two-stage fine-tuning are more effective than zero-shot cross-lingual transfer learning models, which simply use one-stage fine-tuning. The results that we obtain here agree with [16], which shows that the performance of zero-shot cross-lingual transfer learning models can be increased when fine-tuned further using the target language. Then, the performance of XLMRoBERTa in this study is shown to be the most superior, outperforming mBERT and IndoBERT. This result is also consistent with the findings reported in the original paper of XLMRoBERTa [18] which reported that this model outperforms mBERT in various NLP tasks.

IndoBERT models demonstrate relatively good performance in Javanese POS tagging, although it was only pre-trained using Indonesian language without Javanese language [19], different from mBERT and XLMRoBERTa. To analyze this case, we measure the ability of the baseline-indobert model (i.e., the IndoBERT model fine-tuned directly with the Javanese dataset) to capture semantic relationships in Javanese by computing the cosine similarity between the IndoBERT embeddings of two Javanese words that often appear together in the corpus, e.g., "*nyambut gawe*" ("to work"). We compare the results of using IndoBERT embeddings from the base model (i.e., the pre-trained IndoBERT model without any fine-tuning) and the baseline-indobert model. The cosine similarity results using IndoBERT embeddings from the base and the baseline-indobert models are 0.64 and 0.81, respectively. It shows that there is a considerably increased ability in the baseline-indobert model to better capture the semantics of the Javanese language, even though it was only fine-tuned

using a relatively small amount of Javanese data. This finding agrees with [11] which states that BERT can give good performance for downstream task even though only trained using a small amount of data.

According to our experimental results, the best source language for Javanese POS tagging is Indonesian. This could be understood because it has the closest language family to Javanese. This is consistent with the study in [13] that showed that a good criterion for a source language is its similarity with the target language, such as matching language family, matching writing systems, or overlapping vocabularies. However, looking at LangRank's ranking, Indonesian was not included in the top-ranked source languages for Javanese according to LangRank. We analyze that this could happen because we use the default LangRank's overall ranking that includes all features. Our analysis shows that when we use LangRank's per-feature ranking, namely, word overlap, genetic distance, and geographic distance, Indonesian was actually ranked first by LangRank. Therefore, if LangRank only uses those three features to rank source languages for POS tagging, then the model can give more accurate results, consistent with the finding in [14]. However, when LangRank uses all of the features of the framework, it turns out that the dominant features are dataset size and TTR distance. Consequently, because the Uyghur dataset excels in these two features, then it ranks 1st by LangRank's default ranking.

Our analysis reveals that the cross-lingual transfer learning models that were first fine-tuned using Indonesian and after that fine-tuned again using Javanese still have some problems in identifying Indonesian words. It is because the model still recognizes these words as having the Indonesian POS tags as it uses Indonesian as the source language. This finding agrees with [28] which found that language models, like BERT, might still recognize languages used in previous stages of training.

5. CONCLUSION

In this work, we propose cross-lingual transfer learning using transformer-based models for solving the issue of low-resource language in Javanese POS tagging. This study includes the investigation of the selection of the best source language, the most accurate model, and the best fine-tuning scenario. Based on our findings, we show that cross-lingual transfer learning using the XLM-RoBERTa model and Indonesian as the source language, implemented using a two-stage fine-tuning process, first using the source language (i.e., Indonesian) and second using the target language (i.e., Javanese), is the best method to implement cross-lingual transfer learning for Javanese POS tagging. This model achieves an accuracy of 87.65%. The cross-lingual transfer learning results in increasing the accuracy of Javanese POS tagging models without cross-lingual transfer learning by 14.4%–15.3% over LSTM-based models and by 0.21%–3.95% over transformer-based models. Moreover, we also show that, without cross-lingual transfer learning, the transformer-based models can outperform all other baselines, including LSTM and Bi-LSTM deep-learning methods, and the methods used in previous studies on Javanese POS tagging, such as HMM.

6. FUTURE WORK

The Javanese POS tagging dataset used in this study is limited to 1,000 sentences. Despite this limitation on dataset size, our approach using transformer-based models in cross-lingual transfer learning framework can achieve an F-1 score of up to 84% for Javanese POS tagging task. This gives significant improvements over the classic machine learning and deep learning baselines. While this already shows satisfactory results, in the future we are interested to examine to what extent the increase in the Javanese dataset size may further boost the performance of our models. Therefore, we plan to add more data to the Javanese POS tagging dataset by performing further human annotation on Javanese articles.

We found that the best source language for Javanese POS tagging using cross-lingual transfer learning is Indonesian. Comparing the best source language ranking based on our findings and LangRank's ranking, we show that LangRank's overall ranking is not fully accurate, as Indonesian was not among LangRank's overall ranking, although it is the best source language based on our findings. Our analysis highlights that we need to look at LangRank's per-feature ranking as well, as it can give more insights in choosing the optimal source language. We suggest that future research that uses LangRank can look at LangRank's per-feature ranking, in addition to the overall ranking. This aims to obtain a better source language in the cross-lingual transfer learning.

ACKNOWLEDGEMENTS

This research was funded by the Directorate of Research and Development, Universitas Indonesia, under Hibah PUTI Q2 2022 (Grant No. NKB-571/UN2.RST/HKP.05.00/2022). In addition, this research is also supported by the computing facilities provided by Tokopedia-UI AI Center, Faculty of Computer Science Universitas Indonesia.




REFERENCES

- [1] Alfina et al., "A gold standard dataset for javanese tokenization, POS tagging, morphological feature tagging, and dependency parsing," *Github*, 2023. [Online]. Available: <https://github.com/UniversalDependencies/UD-Javanese-CSUI>
- [2] N. Indurkha and F. J. Damerau, *Handbook of natural language processing*. New York: Chapman and Hall/CRC, 2010, doi: 10.1201/9781420085938.
- [3] S. P. Lende and M. M. Raghuwanshi, "Question answering system on education acts using NLP techniques," in *2016 World Conference on Futuristic Trends in Research and Innovation for Social Welfare (Startup Conclave)*, 2016, pp. 1–6, doi: 10.1109/STARTUP.2016.7583963.
- [4] G. G. Shenoy, E. H. Dsouza, and S. Kuebler, "Performing stance detection on twitter data using computational linguistics techniques," *arXiv-Computer Science*, pp. 1–8, 2017, doi: 10.48550/arXiv.1703.02019.
- [5] F. Wu and D. S. Weld, "Open information extraction using wikipedia," in *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2010, pp. 118–127.
- [6] A. Chiche and B. Yitagesu, "Part of speech tagging: a systematic review of deep learning and machine learning approaches," *Journal of Big Data*, vol. 9, no. 1, 2022, doi: 10.1186/s40537-022-00561-y.
- [7] C. Tan, F. Sun, T. Kong, W. Zhang, C. Yang, and C. Liu, "A survey on deep transfer learning," in *Artificial Neural Networks and Machine Learning – ICANN 2018*, 2018, pp. 270–279, doi: 10.1007/978-3-030-01424-7_27.
- [8] A. F. Aji et al., "One country, 700+ languages: NLP challenges for underrepresented languages and dialects in Indonesia," in *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2022, pp. 7226–7249, doi: 10.18653/v1/2022.acl-long.500.
- [9] R. A. Pratama, A. A. Suryani, and W. Maharani, "Part of speech tagging for javanese language with hidden markov model," *Journal of Computer Science and Informatics Engineering (J-Cosine)*, vol. 4, no. 1, pp. 84–91, 2020, doi: 10.29303/jcosine.v4i1.346.
- [10] F. Askhabi, A. A. Suryani, and M. A. Bijaksana, "Part of speech tagging in javanese using support vector machine method," in *eProceedings of Engineering*, 2020, pp. 8095–8102.
- [11] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, 2019, pp. 4171–4186.
- [12] M. Straka, J. Straková, and J. Hajic, "UDPipe at SIGMORPHON 2019: Contextualized embeddings, regularization with morphological categories, corpora merging," in *Proceedings of the 16th Workshop on Computational Research in Phonetics, Phonology, and Morphology*, 2019, pp. 95–103, doi: 10.18653/v1/W19-4212.
- [13] W. D. Vries, M. Wieling, and M. Nissim, "Make the best of cross-lingual transfer: Evidence from POS tagging with over 100 languages," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, 2022, pp. 7676–7685, doi: 10.18653/v1/2022.acl-long.529.
- [14] Y.-H. Lin et al., "Choosing transfer languages for cross-lingual learning," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 3125–3135, doi: 10.18653/v1/P19-1301.
- [15] A. Y. Maulana, I. Alfina, and K. Azizah, "Building Indonesian dependency parser using cross-lingual transfer learning," in *2022 International Conference on Asian Language Processing, IALP 2022*, 2022, pp. 488–493, doi: 10.1109/IALP57159.2022.9961296.
- [16] A. Lauscher, V. Ravishanker, I. Vulić, and G. Glavaš, "From zero to hero: On the limitations of zero-shot language transfer with multilingual transformers," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 4483–4499, doi: 10.18653/v1/2020.emnlp-main.363.
- [17] A. Vaswani et al., "Attention is all you need," in *31st Conference on Neural Information Processing Systems (NIPS 2017)*, 2021, pp. 1–11.
- [18] A. Conneau et al., "Unsupervised cross-lingual representation learning at scale," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 8440–8451, doi: 10.18653/v1/2020.acl-main.747.
- [19] B. Wilie et al., "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*, 2020, pp. 843–857.
- [20] M. Dryer and M. Haspelmath, "The world Atlas of language structures online (v2020.3)," *Research Dataset*, 2020, doi: 10.5281/zenodo.7385533.
- [21] R. Eskander, S. Muresan, and M. Collins, "Unsupervised cross-lingual part-of-speech tagging for truly low-resource scenarios," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 4820–4831, doi: 10.18653/v1/2020.emnlp-main.391.
- [22] M. C. D. Marneffe, C. D. Manning, J. Nivre, and D. Zeman, "Universal dependencies," *Computational Linguistics*, vol. 47, no. 2, pp. 255–308, 2021, doi: 10.1162/COLI.a.00402.
- [23] A. Conneau and G. Lample, "Cross-lingual language model pretraining," *33rd Conference on Neural Information Processing Systems (NeurIPS 2019)*, vol. 32, 2019.
- [24] Y. Liu et al., "RoBERTa: A robustly optimized BERT pretraining approach," in *ICLR 2020 Conference Program Chairs*, 2019, pp. 1–15.
- [25] D. Jurafsky and J. Martin, *Speech and language processing*. Stanford University, 2024.
- [26] A. E. -Can, "A comparison of LSTM and BERT for small corpus," *arXiv-Computer Science*, pp. 1–12, 2020, doi: 10.48550/arXiv.2005.01234.




- 10.48550/arXiv.2009.05451.
- [27] E. Grave, P. Bojanowski, P. Gupta, A. Joulin, and T. Mikolov, "Learning word vectors for 157 languages," in *LREC 2018 - 11th International Conference on Language Resources and Evaluation*, 2019, pp. 3483–3487.
- [28] W. Wongso, D. S. Setiawan, and D. Suhartono, "Causal and masked language modeling of javanese language using transformer," in *2021 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, 2021, pp. 1–7, doi: 10.1109/ICACSIS53237.2021.9631331.

BIOGRAPHIES OF AUTHORS






Gabriel Enrique    received his bachelor's degree in computer science from the Universitas Indonesia in 2023. His research interests are related to natural language processing. He can be contacted at email: gabriel.enrique@ui.ac.id.



Ika Alfina    is a lecturer and researcher at the Faculty of Computer Science, Universitas Indonesia. She received her bachelor's degree in computer science from Universitas Indonesia (UI) in 2000, her master's degree in computer science from UI in 2007, and her doctorate in computer science in 2021. Her current research interest is natural language processing. She can be contacted at email: ika.alfina@cs.ui.ac.id.



Evi Yulianti    is a lecturer and researcher at Faculty of Computer Science, Universitas Indonesia. She received the B.Comp.Sc. degree from the Universitas Indonesia in 2010, the dual M.Comp.Sc. degree from Universitas Indonesia and Royal Melbourne Institute of Technology University in 2013, and the Ph.D. degree from Royal Melbourne Institute of Technology University in 2018. Her research interests include information retrieval and natural language processing. She can be contacted at email: evi.y@cs.ui.ac.id.