Pre-trained Language Models for Tagalog with Multi-source Data

Chapter · October 2021

DOI: 10.1007/978-3-030-88480-2_17

4 authors, including:

Shengyi Jiang
Guangdong University of Foreign Studies
128 PUBLICATIONS **1,628** CITATIONS
SEE PROFILE

Yingwen Fu
21 PUBLICATIONS **36** CITATIONS
SEE PROFILE

Nankai Lin
Guangdong University of Foreign Studies
54 PUBLICATIONS **104** CITATIONS
SEE PROFILE

Pre-trained Language Models for Tagalog with Multi-source Data

Shengyi Jiang[1,2], Yingwen Fu[1], Xiaotian Lin[1], and Nankai Lin[1,2](✉)

[1] School of Information Science and Technology, Guangdong University of Foreign Studies, Guangzhou, China

[2] Guangzhou Key Laboratory of Multilingual Intelligent Processing, Guangdong University of Foreign Studies, Guangzhou, China

Abstract. Pre-trained language models (PLMs) for Tagalog can be categorized into two kinds: monolingual models and multilingual models. However, existing monolingual models are only trained in small-scale Wikipedia corpus and multilingual models fail to deal with Tagalog-specific knowledge needed for various downstream tasks. We train three existing models on a much larger corpus: ***BERT-uncased-base***, ***ELECTRA-uncased-base*** and ***RoBERTa-base***. At the pre-training stage, we construct a large-scale news text corpus for pre-training in addition to the existing open-source corpora. Experimental results show that our pre-trained models achieve consistently competitive results in various Tagalog-specific natural language processing (NLP) tasks including part-of-speech (POS) tagging, hate speech classification, dengue classification and natural language inference (NLI). Among them, POS tagging dataset is a self-constructed dataset aiming to alleviate the insufficient labeled resource for Tagalog. We will release all pre-trained models and datasets to the community, hoping to facilitate the future development of Tagalog NLP applications.

Keywords: Pre-trained language model · Tagalog · POS tagging

1 Introduction

Pre-trained language models (PLMs) represented by BERT [1] have been proven to significantly improve the performance of various downstream natural language processing (NLP) tasks and thus become extremely popular for many NLP researches. Despite of success of pre-trained BERT and its variants, they have largely limited to high-resource languages such as English. For a new language, one could pre-train a new language-specific model based on BERT architecture and training method [2–5] or utilize existing pre-trained multilingual BERT-based models [1, 6, 7].

In terms of PLMs for Tagalog, monolingual models [8–10] and multilingual models [7, 11] are both publicly available. However, there are two main concerns about these two kinds of models:

(1) **Monolingual models**: All the existing monolingual models for Tagalog are only pre-trained on the Tagalog Wikipedia corpus [8]. While Wikipedia data is not representative of a general language use, and the Tagalog Wikipedia data size is relatively

L. Wang et al. (Eds.): NLPCC 2021, LNAI 13028, pp. 210–223, 2021.
https://doi.org/10.1007/978-3-030-88480-2_17

small (283M in size uncompressed), pre-trained language models can be significantly improved by using more pre-training data [12] from different data sources such as news.

(2) **Multilingual models**: Multilingual pre-trained models struggle to explain their applicability in acquiring language-invariant knowledge for downstream tasks of various languages. As different languages have different sequence structures, multilingual pre-trained models are more suitable for cross-language applications than in monolingual applications. As an agglutinative language, Tagalog shows some characteristics of inflectional languages. It also has a variety of lexical morphology, complex syntactic structure and relatively free sequence order. It is necessary to pre-train monolingual models for Tagalog to improve the performance of downstream tasks.

To tackle the two issues above, we train three monolingual BERT-based models using 1444M Tagalog corpus (four times more than Wikidata used in previous works) from multiple data sources. At the pre-training stage, we construct a large-scale news text corpus for pre-training in addition to the existing open-source corpora. We evaluate our models on three benchmark Tagalog text classification datasets: Hate Speech classification, Dengue classification and natural language inference (NLI) [9, 10]. In addition to text classification, pre-trained models should be evaluated in more kinds of NLP tasks such as sequence labeling tasks. However, the recent sequence-labeled resources in Tagalog are scare that they cannot meet the development of deep learning technology in terms of scale and quality. Therefore, we construct a Tagalog part-of-speech (POS) tagging (referred as a common sequence labeling task) dataset consisting of 14438 sentences. Experimental results show that our models obtain competitive results on all these tasks.

The contributions in this paper are summarized as follows:

(1) We present a series of large-scale monolingual language models pre-trained for Tagalog on a much larger size of corpus.
(2) We construct a large-size news corpus for Tagalog language, which could make up for the gap in scare Tagalog NLP resources.
(3) We construct a large-scale and high-quality Tagalog POS tagging dataset to alleviate the current situation of insufficient language resources.
(4) Our models achieve competitive performances on four downstream datasets, showing the effectiveness of BERT-based monolingual language models for Tagalog.
(5) The pre-trained models and the POS tagging dataset would be publicly available serving as strong baselines.

2 Related Previous Research

2.1 Natural Language Processing for Tagalog

Part-of-Speech Tagging. Part-of-speech (POS) is a fundamental grammatical attribute of tokens that signifies the morphological and syntactic behaviors of a lexical item. It is

designed as one of the sequence labeling tasks. Cheng and Rabo [13] construct a POS tagging corpus comprised of 141 sentences and 59 tags and propose a template-based n-gram POS tagger. Reyes et al. [14] develop a Tagalog POS tagger (SVPOST) using support vector machines (SVMs) and their corpus consists of 122318 tokens and 64 tags. Olivo et al. [15] are the first to use conditional random field (CRF) for Tagalog POS tagging. There are two main concerns about the POS tagging research in Tagalog: (1) Tagalog is represented as a low-resource language that most of the Tagalog POS taggers are still based on rules and machine learning (ML). (2) The corpora above are not publicly available, which makes it impossible for us to properly compare performance of different models and techniques. In this work, we build and release a high-quality POS corpus and use neural methods to construct baseline POS tagger.

Text Classification. Cruz et al. [10] create and release News PH-NLI, the first Natural Language Inference (NLI) benchmark dataset in Tagalog. Moreover, they produce new pre-trained transformers to further alleviate the resource scarcity in Tagalog. Cruz and Cheng [9] release two text classification datasets, namely Hate Speech Dataset (binary classification) and Dengue Dataset (multilabel text classification). They also pre-train transformer-based language models for use within Tagalog setting. Our pre-trained models are evaluated in these three benchmark datasets for comparison.

2.2 Pre-trained Language Model for Tagalog

Monolingual Pre-trained Language Model. Cruz and Cheng [8] pre-train a new Tagalog BERT model using the WikiTextTL-39 dataset. In order to cater to low-resource settings in an equipment perspective, they also construct a smaller version of the BERT model via model distillation, producing a DistilBERT model. Cruz et al. produce four ELECTRA models: a cased and an uncased model respectively in the base size and small size, using the WikiText TL-39 dataset [10].

Multilingual Pre-trained Language Model. Publicly transformer-based multilingual PLMs represented by multilingual BERT (mBERT) [1], XLM [9] and mt5 [11] are trained in a large dataset including multiple language datasets to obtain language-invariant information. It is notable that XLM-100 and mt5 support Tagalog language while mBERT does not support Tagalog language.

3 Model

Three model for Tagalog are introduced in this paper: an uncased BERT model[1], an uncased ELECTRA[2] model and a RoBERTa[3] model. They are all in the base size (12 layers, 768 hidden units, 12 attention heads).

[1] https://huggingface.co/GKLMIP/bert-tagalog-base-uncased.
[2] https://huggingface.co/GKLMIP/electra-tagalog-base-uncased.
[3] https://huggingface.co/GKLMIP/roberta-tagalog-base.

3.1 BERT

BERT (Bidirectional Encoder Representations for Transformers) [1], is designed to learn deep bidirectional representations from unlabeled text by jointly modeling context from both forward and backward directions in all layers. It consists of multiple bidirectional transformer encoders [17].

BERT is comprised of two unsupervised subtasks, namely Mask Language Model (MLM) and Next Sentence Prediction (NSP): (1) MLM refers to masking some words from the input sequence and then predicting the masked word through the context; (2) NSP is designed to enhance the relationship between a sentence pair. Its objective is to predict whether the sentence pair are continuous. Pre-trained BERT can be fine-tuned for a variety of downstream tasks such as text classification, named entity recognition (NER) and question answering (QA) tasks (Fig. 1).

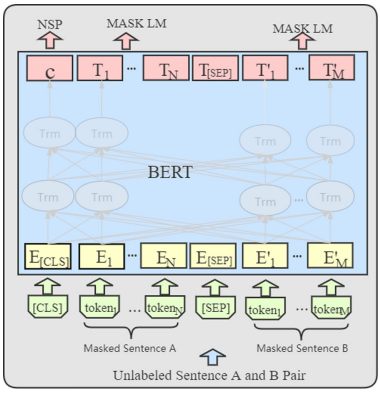

Fig. 1. BERT model.

3.2 RoBERTa

Being a variant of BERT, RoBERTa [12] aims to make full use of BERT architecture and training methods. There are three improvements in RoBERTa compared with BERT: (1) **More training data:** RoBERTa leverages more unlabeled data to pre-train the model for a more robust performance in downstream tasks; (2) **Abondance of NSP task:** Liu et al. [12] verified the invalidity of the NSP task and removed this task; (3) **Dynamic word masking:** RoBERTa uses dynamic word masking to train the MLM task instead of the static word masking proposed by BERT model, which allows the parameters of the pre-trained model to be more fully optimized and the model can better capture sequence features (Fig. 2).

3.3 ELECTRA

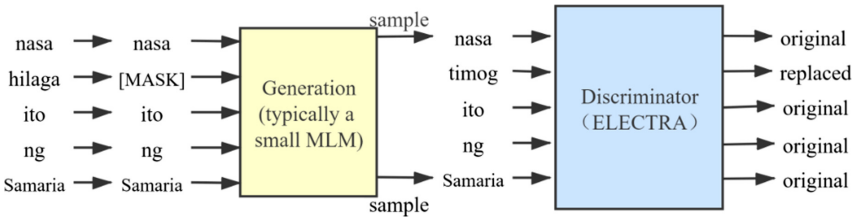

Fig. 2. ELECTRA model.

Apart from BERT model, a new pre-trained framework, ELECTRA [16], uses the combination of generator and discriminator.

Compared with BERT, the innovations of ELECTRA are as follows: (1) It proposes replaced token detection (RTD), a pre-training task in which the model learns to distinguish real input tokens from plausible but synthetically generated replacements. Instead of masking, ELECTRA corrupts the input by replacing some tokens with samples from a proposal distribution, which is typically the output of a small masked language model. (2) Instead of training a model that predicts the original identities of the corrupted tokens, ELECTRA trains a discriminative model that predicts whether each token in the corrupted input was replaced by a generator sample or not. (3) In order to effectively learn context information, it uses weight sharing to share the generator's embedding information with the discriminator. (4) The model jointly trains a small generator and a discriminator to ease the training difficulty of the discriminator.

4 Pre-training Corpus

In this paper, the models are pre-trained on massive data collected from three sources, namely Oscar corpus, Wikipedia corpus and news corpus. The whole corpus used for pre-training has a size of 1.3G, which is four times more than the corpus used by the existing Tagalog pre-trained models. We use 99% of the corpus as the training set and 1% as the validation set. The corpus statistics for pre-training are shown in Table 1.

Table 1. Statistics of the pre-training corpus.

Source	Size of File	Num. of Document	Num. of Line	Num. of Tokens
Oscar	417M	–	3580299	78236499
Wiki	283M	174444	2004429	5173224
News	744M	396941	6341530	140186315

4.1 Oscar

Oscar corpus is a large-scale unlabeled corpus constructed by Ortiz et al. [18]. In order to create the multilingual OSCAR corpus, Ortiz et al. (2019) reproduce the pipeline proposed by Grave et al. [19] to process, filter and classify Common Crawl, which is a non-profit organization that produces and maintains an open, freely available repository of crawled data from the web. The filtering step used to create OSCAR involves keeping only the lines containing at least 100 UTF-8 encoded characters. Finally, as in Grave et al. [19], the OSCAR corpus is deduplicated, i.e. for each language, only one occurrence of a given line is included. In this paper, we only use Tagalog corpus from the OSCAR corpus.

4.2 Wiki

Wikipedia is a multilingual, free encyclopedia that contains a lot of text information. We use the Tagalog Wikipedia corpus "WikiText-TL-39" [8] as one of the training corpora. "TL" stands for Tagalog and "39" refers to the dataset having 39 million tokens in the training set. In this paper, we use the training set, validation set and test set of this corpus for pre-training.

4.3 News

We crawl massive news articles from 13 Tagalog news websites to construct a large-scale news corpus for Tagalog. The corpus is comprised of around 400,000 news articles as shown in Table 2.

Table 2. Statistics of news websites.

Website	Num. of Document
https://www.pna.gov.ph	85655
http://balita.net.ph/	37051
http://bandera.inquirer.net	73525
http://cnnphilippines.com	96
http://eaglenews.com	9802
https://www.bworldonline.com	7728
https://tonite.abante.com.ph	28045
https://www.topgear.com.ph	355
https://philnews.ph	181
https://kickerdaily.com	11153
https://www.hataw tabloid.com	39486
https://www.remate.ph	93908
https://www.pinoyparazzi.com	9956
Total	396941

5 Experiment

5.1 Downstream Tasks

POS Tagging. The existing Tagalog POS tagging datasets cannot meet the development of deep learning technology in terms of scale and quality and all of them are not publicly available. Therefore, we build a dataset containing 14438 samples (totally 286706 words) within 39 tags based on the Tagalog news articles crawled from Bailta[4]. In the annotating process, each sample is labeled by two annotators. Then samples with the same labeling results are added to the dataset. Instead, samples with different annotation results will be annotated again by the third Annotators. If the annotation results are the same as one of the first two persons, They will also be added to the dataset. A split of (70%, 15%, 15%) of the dataset is respectively for (training, test, validation). Statistics of the POS tagging dataset and POS Tagset are represented in Table 3 and Table 4.

Table 3. Data distribution of the POS tagging dataset.

Data	Num. of Sentence	Num. of Token
Train	10108	195468
Dev	2165	46971
Test	2165	44267
Total	14438	286706

Table 4. Statistics of the POS tagging dataset.

Tag	Proportion (%)	Explanation
CN	13.7018	Common noun
AD	2.2776	Auxiliary verb
P	4.9263	Particle
CP	3.6675	Completed
PREP	13.4961	Preposition
A	3.1935	Adjective
ART	5.5548	Article
PN	6.0414	Proper noun
Z	10.8941	Punctuation
INF	3.1307	Infinitive

(*continued*)

[4] http://balita.net.ph/.

Table 4. (*continued*)

Tag	Proportion (%)	Explanation
CS	4.2158	Connection structure
INTP	0.1531	Interrogative pronoun
PP	4.8457	Personal pronoun
CC	1.986	Coordinating conjunction
SC	2.5556	Subordinating conjunction
NP	0.3007	Negative pronoun
F	10.013	Foreign Word
INTADV	0.1779	Indefinite adverb
NADV	0.8273	Negative adverb
CT	1.3251	Contemplated
DP	1.1943	Demonstrative pronoun
INC	1.8667	Incompleted
JOD	0.6446	Ordinal number of adjective
CD	1.1224	Cardinal number
X	1.1451	Unknown
INDP	0.1221	Indefinite pronoun
INT	0.0743	Interjection
AS	0.1102	Adjective, superlative degree
DADV	0.3101	Demonstrative adverb
VOD	0.0345	Ordinal number of adverb
INDADV	0.0244	Indefinite adverb
DD	0.0593	Demonstrative determiner
HADV	0.0014	Adverb of the same class
NUM	0.0007	Numeral
ADJ	0.0028	Adjective
ADV	0.0003	Adverb
SADV	0.001	Adverb, superlative degree
QD	0.0007	Quantitative determiner
V	0.001	Verb

Natural Language Inference. Natural Language Inference (NLI) is a sentence-pair classification for inference of the relationship between two sentences, such as a sentence with a premise and a sentence with a hypothesis. Their relationship can be entailment, neutrality and contradiction. NewsPH-NLI [10] is an NLI benchmark dataset in Tagalog comprised of multiple news articles from all major Tagalog news sites online. The dataset is divided into (420000, 90000, 9000) documents for (training, test, validation) sets.

Hate Speech Classification. Hate Speech dataset [9] is a collection of tweets mined in real-time during the 2016 Philippine Presidential Election debates, and from tweets related to the 2016 election hashtags. The dataset is introduced as a binary classification task benchmark in Tagalog, with each tweet labeled as 0 (non-hate) or 1 (hate). The training set has 10,000 labeled examples with 5340 and 4660 non-hate and hate tweets respectively. An even split of 4232 validation and 4232 test samples are included for evaluation.

Dengue Classification. Dengue dataset [9], a multiclass classification dataset, is composed of tweets collected from Twitter in the Tagalog language. There are five labels for each tweet in the dataset: absent, dengue, health, mosquito, and sick. The dataset is represented as a low-data dataset, with only 4015 training examples and an even split of 500 validation and 500 test examples. More importantly, the classes are highly imbalanced with a distribution of (905, 49, 1804, 528, 1035) samples in five labels (absent, dengue, health, mosquito, sick).

5.2 Pre-training

Two improvements are made in our pre-trained models compared with the existing pre-trained models [8]: (1) **more data**; and (2) **a larger vocabulary size**. The vocabulary in the models pre-trained by Cruz and Cheng [8] is 32K, while the size of our pre-trained dictionary is 52K. In addition, in the pre-training stage in [8], the model pre-training epoch exceeds 50 batches, while we only conduct 5 batches. The BERT and ELECTRA models we build are both uncased models, because in general, the performance of uncased models is better than the cased models [8]. When training BERT and ELECTRA, we use the word piece [20] segmentation method, and when training RoBERTa, we use the BPE [21] segmentation method. The hyperparameters in the pre-training stage are shown in Table 5.

Table 5. Hyperparameters for pre-training.

Parameter	BERT	ELECTRA	RoBERTa
Layer Num	12	12	12
Hidden Size	768	768	768
FFN inner hidden size	3072	3072	3072
Attention heads	12	12	12
Vocab Size	52000	52000	52000
Tokenizer Type	Word Piece	Word Piece	BPE
Adam β_1	0.9	0.9	0.9
Adam β_2	0.999	0.999	0.98
Adam ϵ	1e−6	1e−6	1e-6
Learning Rate Decay	Linear	Linear	Linear
Weight Decay	0.01	0.01	0.01
Batch Size	128	128	128
Peak Learning Rate	1e-4	1e-4	5e-4
Dropout	0.1	0.1	0.1
Attention Dropout	0.1	0.1	0.1
Epoch	5	5	5
Warmup Steps	5K	5K	5K
Max Length	512	512	512

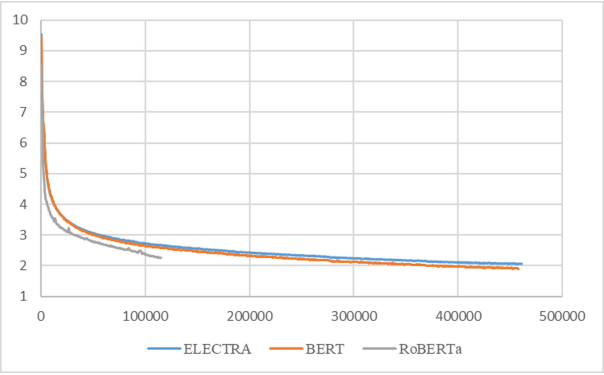

Fig. 3. Pre-training losses for all models over the steps.

5.3 Fine-Tuning

We compare our pre-trained models with six existing pre-trained models. For classification tasks, we report the accuracy of the test set as in [9, 10], and report the accuracy,

precision, recall and F1-score for POS tagging task. We use the same classification fine-tune code[5] from Cruz et al. [8–10]. In the classification task, we uniformly set the maximum length to 128, while in the POS tagging task, the maximum length is set to 200. For two small ELECTRA models, we fine-tune for 3 epochs with a learning rate of 2e-4. For three base BERT model and base ELECTRA model, we fine-tune for 3 epochs with a learning rate of 5e-5. For RoBERTa model, we fine-tune for 5 epochs with a learning rate of 3e-5 for NewsPH-NLI task because it needs a longer training time and a smaller learning rate to converge. And for other task, we fine-tune RoBERTa for 3 epochs with a learning rate of 5e-5. For XLM model, we fine-tune for 5 epochs with a learning rate of 1e-6 for hate speech classification task and NIL task, and fine-tune for 3 epochs with a learning rate of 5e-5 for POS tagging task and dengue classification task. The GPU we use for model fine-tuning is TIAN RTX.

5.4 Experiment Results and Analysis

As shown in Fig. 3, in the warm-up phase, the loss values of BERT, ELECTRA and RoBERTa drop significantly, while in the subsequent phases, the loss values slowly decrease. Among them, the loss value of ELECTRA dropped from 9.5212 to 2.0620, the loss value of BERT dropped from 9.4844 to 1.9041, and the loss value of RoBERTa dropped from 8.5030 to 2.2657. It seems that BERT model fits the best.

Table 6. The result of NewsPH-NLI.

Model	Test Loss	Test Accuracy
BERT (base, cased)	0.3169	0.8870
BERT (base, uncased)	0.3114	0.8884
ELECTRA (base, cased)	0.2572	0.9113
ELECTRA (base, uncased)	0.2528	0.9168
ELECTRA (small, cased)	0.1896	0.9279
ELECTRA (small, uncased)	0.1948	0.9249
XLM	0.2116	0.9115
BERT (base, uncased, our)	0.1872	0.9466
ELECTRA (base, uncased, our)	**0.1858**	**0.9489**
RoBERTa (base, uncased, our)	0.3678	0.9128

[5] https://github.com/jcblaisecruz02/Filipino-Text-Benchmarks.

Table 7. The result of hate speech classification.

Model	Test Loss	Test Accuracy
BERT (base, cased)	0.6172	0.7695
BERT (base, uncased)	0.5849	0.7862
ELECTRA (base, cased)	0.6264	0.7648
ELECTRA (base, uncased)	0.5925	0.7608
ELECTRA (small, cased)	**0.4725**	0.7883
ELECTRA (small, uncased)	0.5009	0.7683
XLM	0.5508	0.7110
BERT (base, uncased, our)	0.5578	0.8193
ELECTRA (base, uncased, our)	0.5588	**0.8264**
RoBERTa (base, uncased, our)	0.5497	0.7930

Table 8. The result of dengue classification.

Model	Test Loss	Test Accuracy
BERT (base, cased)	0.1886	0.9318
BERT (base, uncased)	0.1708	0.9405
ELECTRA (base, cased)	0.1953	0.9288
ELECTRA (base, uncased)	0.1750	0.9330
ELECTRA (small, cased)	0.1833	0.9316
ELECTRA (small, uncased)	0.1754	0.9296
XLM	0.2014	0.9133
BERT (base, uncased, our)	**0.1395**	**0.9541**
ELECTRA (base, uncased, our)	0.1454	0.9525
RoBERTa (base, uncased, our)	0.1572	0.9425

Table 6–9 present the results of different pre-trained models for 4 downstream tasks. Our models outperform the existing models in three text classification tasks: (1) Our pre-trained ELECTRA model works best in NLI task and hate speech classification task which achieve an accuracy of 0.9489 and 0.8264; (2) for dengue classification task, our pre-trained BERT model reaches state-of-the-art performance with an accuracy of 0.9541. It is worthwhile to note that in POS tagging task, our pre-trained BERT model and ELECTRA model have the same F1-score, and BERT reaches state-of-the-art performance with an accuracy of 0.9532.

Table 9. The result of POS tagging.

Model	Accuracy	Precision	Recall	F1
BERT (base, cased)	0.9441	0.9259	0.9222	0.9241
BERT (base, uncased)	0.9429	0.9264	0.9222	0.9243
ELECTRA (base, cased)	0.9358	0.9149	0.9134	0.9142
ELECTRA (base, uncased)	0.9368	0.9164	0.9140	0.9152
ELECTRA (small, cased)	0.9412	0.9231	0.9190	0.9211
ELECTRA (small, uncased)	0.9387	0.9200	0.9162	0.9181
XLM	0.9517	0.9352	0.9328	0.9340
BERT (base, uncased, our)	**0.9532**	**0.9381**	0.9351	**0.9366**
ELECTRA (base, uncased, our)	0.9531	0.9379	**0.9353**	**0.9366**
RoBERTa (base, uncased, our)	0.9473	0.9301	0.9271	0.9286

6 Conclusion

In this paper, we present three monolingual language models for Tagalog pre-trained in a much larger corpus. Additionally, we construct a part-of-speech (POS) tagging dataset to relieve the insufficient sequence-labeled resources in Tagalog. Experimental results demonstrate the effectiveness of our pre-trained models in various Tagalog natural language processing (NLP) tasks of POS tagging, hate speech classification, dengue classification and natural language inference (NLI). By publicly releasing the pre-trained models, we hope that they can have implications for future research for Tagalog NLP.

Acknowledgement. This work was supported by the National Natural Science Foundation of China (No. 61572145), the Major Projects of Guangdong Education Department for Foundation Research and Applied Research (No. 2017KZDXM031) and National Social Science Foundation of China (No. 17CTQ045). The authors would like to thank the anonymous reviewers for their valuable comments and suggestions.

References

1. Devlin, J., Chang, M.W., Lee K., Toutanova K.: BERT: pre-training of deep bidirectional transformers for language understanding, In: Proceedings of NAACLHLT 2019, pp. 4171–4186 (2019)
2. Cui, Y., Che, W., Liu, T., Qin, B., Yang, Z., Wang, S., Hu, G.: Pre-training with whole word masking for Chinese BERT. CORR (2019)
3. de Vries, W., van Cranenburgh, A., Bisazza, A., Caselli, T., van Noord, G., Nissim, M.: BERTje: a Dutch BERT model. CORR (2019)
4. Vu, X.S., Vu, T., Tran, S.N., Jiang, L.: ETNLP: a visual-aided systematic approach to select pre-trained embeddings for a downstream task. In: Proceedings of the International Conference on Recent Advances in Natural Language Processing, pp. 1285–1294 (2019)

5. Martin, L., et al.: CamemBERT: a tasty French language model. In: Annual Meeting of the Association for Computational Linguistics, pp. 7203–7219 (2020)
6. Lample, G., Conneau A.: Cross-lingual language model pretraining. CORR (2019)
7. Conneau, A., et al.: Unsupervised cross-lingual representation learning at scale. In: Annual Meeting of the Association for Computational Linguistics, pp. 8440–8451 (2020)
8. Cruz, J.C.B., Cheng, C.: Evaluating language model finetuning techniques for low-resource languages. CORR (2019)
9. Cruz, J.C.B., Cheng, C.: Establishing baselines for text classification in low-resource languages. CORR (2020)
10. Cruz, J.C.B., Resabal, J.K., Lin, J., Velasco, D. J., Cheng, C.: Investigating the true performance of transformers in low-resource languages: A case study in automatic corpus creation. CORR (2020)
11. Xue, L., et al.: mT5: a massively multilingual pre-trained text-to-text transformer. CORR (2021)
12. Liu, Y., et al.: RoBERTa: a robustly optimized BERT pretraining approach. CORR (2019)
13. Cheng, C., Rabo, S.: TPOST: a template-based, n-gram part-of-speech tagger for Tagalog. J. Res. Sci. Comput. Eng. **3**(1) (2004)
14. Reyes, C.D.E., Suba, K.R.S., Razon, A.R., Naval Jr., P.C.: SVPOSTA part-of-speech tagger for Tagalog using support vector machines. In: Proceedings of the 11th Philippine Computing Science Congress (2011)
15. Olivo, J.F.T., Hari, P.J.T., dela Fuente, M.B.: CRFPOST: part-of-speech tagger for Filipino texts using conditional random fields. In: Proceedings of the 2nd International Conference on Algorithms, Computing and Artificial Intelligence, pp. 444–449 (2019)
16. Clark, K., Luong, M.T., Le, Q.V., Manning, C.D.: ELECTRA: pre-training text encoders as discriminators rather than generator. In: Proceedings of International Conference on Learning Representations (2020)
17. Vaswani, A., et al. Attention is all you need. CORR (2017)
18. Suárez, P.O., Romary, L., Sagot, B.: A monolingual approach to contextualized word embeddings for mid-resource languages. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (2020)
19. Grave, E., Bojanowski, P., Gupta, P., Joulin, A., ikolov, T.: Learning word vectors for 157 languages. In: Proceedings of the 11th Language Resources and Evaluation Conference, European Language Resource Association (2018)
20. Wu, Y., Schuster, M., Chen, Z., Le, Q.V., Norouzi, M., Macherey, W., et al.: Google's neural machine translation system: bridging the gap between human and machine translation. CORR (2016)
21. Sennrich, R., Haddow, B., Birch, A.: Neural machine translation of rare words with subword units. CORR (2015)