Improving Indonesian Text Classification Using Multilingual Language Model

Ilham Firdausi Putra
Sekolah Teknik Elektro dan Informatika
Institut Teknologi Bandung
Bandung, Indonesia
ilhamfputra31@gmail.com

Ayu Purwarianti
Sekolah Teknik Elektro dan Informatika
Institut Teknologi Bandung
Bandung, Indonesia
ayu@stei.itb.ac.id

Abstract— **Compared to English, the amount of labeled data for Indonesian text classification tasks is very small. Recently developed language models called multilingual language model have shown its ability to create multilingual representations. In this paper, we investigate the effect of combining English and Indonesian data on building Indonesian text classification (e.g. sentiment analysis and hate-speech) using multilingual language models. Using feature-based approach, we observe its performance on various data sizes and total added English data. The experiment showed that the addition of English data, especially if the amount of Indonesian data is small, improves the performance. Using fine-tuning approach, we further showed its effectiveness at utilizing English language on building Indonesian text classification models.**

Keywords—multilingual language model, text classification, sentiment analysis, hate-speech classification, Indonesian text

I. Introduction

With the development of Indonesian people's access to the internet, more and more text data is available digitally. This data is full of information and very useful if processed. For business owners, for example, citizen comments on the internet can be analyzed for sentiment to find out their reaction to something. Then for those who have a website, detecting violations in online conversations such as hate speech or abuse automatically can be very helpful.

Building such good automated text classifiers needs a lot of labeled data, something that Indonesian language lacks compared to other languages such as English. For example, the aforementioned task of sentiment analysis and hate-speech detection have English dataset like Yelp Review[1] (598K) and Jigsaw Toxic Comment[2] (1902K) respectively. In contrast with Indonesian dataset from recent publications such as [2] (12K) & [3] (11K) for sentiment analysis, and [4] (12K) for hate-speech detection.

Recent deep, contextualized language models trained on corpora of multiple languages provide powerful, general-purpose linguistic representation across language [1, 5]. These language models were trained on a corpora of multiple languages without any multilingual objective. Surprisingly, an empirical investigation has shown that its hidden representation does share a common subspace that represents useful linguistic information in a language-agnostic way [6]. Based on these findings, we seek to improve Indonesian language's lack of data with English data by utilizing multilingual language models.

In this paper, we investigate the effect of combining English and Indonesian data on building Indonesian sentiment analysis and hate-speech detection using multilingual language models. Using feature-based approach, our findings show that the addition of English dataset, especially if Indonesian language data is small, improves the performance of Indonesian text classification. But there are cases where the addition of excessive English data decreases classification performance. Using a fine-tuning approach, we further improve the result of previous research on sentiment analysis [2, 3] and hate-speech detection [4] on Indonesian language.

II. Related Works

Previous works have conducted Indonesian sentiment analysis using various text representation and model approaches. Farhan and Khodra [2] conducted experiments on various text representation with support vector machine, random forest, and multilayer perceptron as model. An evaluation on its own dataset yields 0.8521 F1-score using a neural network with TF-IDF as its feature. Crisdayanti and Purwarianti [3] conducted experiments on various text representations and neural network topology models. An evaluation on its own dataset yields 0.9369 F1-score using Bi-LSTM with word embedding enhanced with paragraph vector. Ibrohim and Budi [4] conducted experiments on multi-label hate-speech and abusive language in Indonesian twitter. An evaluation on its own dataset yields 77.36% average accuracy using random forest with unigram as its feature. Previous works on text classification using multilingual language models have focused on its zero-shot capability across 15 XNLI languages [5]. Using its largest multilingual model, [5] showed that their model is very competitive and even outperforming monolingual language model on various languages. There is no research in Indonesian text classification using multilingual language models, especially in sentiment analysis and hate speech classification.

III. Proposed Method

The latest Indonesian text classification model have been developed using representation from word embeddings and sequential model such as recurrent neural networks (RNN), long short-term memory (LSTM) [13] or gated recurrent unit (GRU) [14] neural networks. While the combination of word embedding with RNN, LSTM, or GRU as the model have been very successful, it still possesses some limitation [1, 8]. First, the nature of sequential processing restricts the model since it can only attend to previous token. Second, pretrained word representation failed to capture deeper meaning on a

---

[1] https://www.yelp.com/dataset

[2] https://www.kaggle.com/c/jigsaw-unintended-bias-in-toxicity-classification/data

document of text. Both problems were solved using recently developed language model such as BERT [1].

BERT use Transformer architecture and pretrained on millions of texts with masked language model (MLM) objective. MLM randomly masked a token on its input and ask the model to predict the masked token. This novel objective paired with Transformer and its self-attention mechanism allow the model to learn from the context of the whole document. This method proven massively successful, topping various benchmark and significantly improve the performance from previously state-of-the-art sequential based language model such as ELMo [15].

On the other hand, labeled Indonesian data is scarce in comparison to English. To utilize English data on Indonesian task, we need a cross-lingual representation of English and Indonesian language on shared vector space. There are two main methods on producing cross-lingual representation on shared vector space: *alignment* and *joint optimization* [9]. Earlier works on producing cross-lingual representation have relied on the former. Mikolov, Le, & Sutskever [10] trained an embedding of two different languages independently and aligned the two using small bilingual data. While it had great success, [9] argues that there are critical downsides such as it rely on having good parallel data and a key assumption that the embedding spaces of each language is isomorphic, which [11, 12] proves does not hold for many language pairs

Recent development on joint training methods have shown great results in obtaining cross-lingual representation without any parallel data. The method simply trains Transformer based language models on corpora of multiple languages, without any multilingual objective. The resulting models have shown good generalization ability across language and language-agnostic representation inside its hidden state [6]. Using this cross-lingual and strong pretrained representation, we add English text data to train our Indonesian text classification model.

One such multilingual language model that we used on the experiments is Multilingual BERT [1] and XLM-RoBERTa [5] (henceforth, mBERT and XLM-R). Denoting the number of Transformer blocks as L, the hidden size as H, and the number of self-attention heads as A, Multilingual BERT is using the BERT-base architecture (L=12, H=768, A=12, Total Parameters=110M) and was trained on concatenation of Wikipedia corpora from 104 languages. The XLM-R variant that we use in the experiment is XLM-R Large (L = 24, H = 1024, A = 16, 550M params). Following the improvement in BERT pretraining approach [7] and adding more balanced data across all languages, XLM-R Large have substantially more parameters, better pre training technique, and trained on the CommonCrawl corpus containing 100 languages.

IV. EXPERIMENTS

In this section, we describe the building blocks of our experiment. Overall, it consists of three training data scenarios, two training approaches, and five datasets.

A. Training Data Scenarios

We investigate the model performance on three different scenarios, each differ by the combination of language used in its training data: monolingual, zero-shot, and multilingual. In the monolingual scenario, we use Indonesian text to train and validate the model. In the zero-shot scenario, we use English text to train the model while being validated on Indonesian

text. Lastly, we use a combination of Indonesian and English text to train the model while being validated on Indonesian text in the multilingual scenario. Using these scenarios, we observe the improvement of the added English text.

B. Training Approaches

There are two approaches on applying large pre-trained language representation to downstream tasks: *feature-based* and *fine-tuning* [1]. On a feature-based approach, we extract fixed features from the pretrained model. In this experiment, we use the last hidden state, which is 768 for mBERT and 1024 for XLM-R Large, as the feature. This extracted feature is then fed into a single dense layer, the only layer we trained on feature-based approach, connected with dropout before finally end on a sigmoid function. On the contrast, fine-tuning approach trains all the language model parameters, 110M for mBERT and 550M for XLM-R Large, including the last dense layer, on the training data binary cross entropy loss.

Using the feature-based scenario, we run many experiments as the expensive and multilingual representation have been precomputed on all the data. In all training data scenarios, we vary the total data used. More specifically, we train the model using [500, 1000, 2500, 5000, 7500, Max] text data. Specific to multilingual training data scenario, we vary the amount of added English data by [0.25, 0.5, 0.75, 1, 1.5, 2, 3, 4, 5, 6, 7, 8, 9, 10] times the amount of Indonesian text data. We refer to a multilingual experiment with added English data N times the amount of Indonesian text data as multilingual(N).

In contrast to feature-based scenarios, fine-tuning the full language model is expensive and resource intensive. But as shown on [1], fully fine-tuning the full language model will result in a better text classifier. We fine-tuned the best performing model on feature-based scenarios. The experiment was reduced to only using maximum total data and added English data multiplier up to 3.

C. Datasets

For Indonesian language sentiment analysis dataset, we used data originated from previous related work by [2, 3]. Farhan and Khodra [2] used reviews crawled from TripAdvisor. Crisdayanti and Purwarianti [2] used text from Twitter, Zomato, TripAdvisor, Facebook, Instagram, and Qraved. For English language we use data from Yelp that they have open sourced. The details on the train, test, and label distribution for each data can be seen at Table I.

TABLE I.   SENTIMENT DATASET DETAILS

Data		Train	Test
Source	*Label*		
Farhan & Khodra [2]	Positive	6281	1125
	Negative	6108	1304
Crisdayanti & Purwarianti [3]	Positive	7151	208
	Negative	3830	204
Yelp Review	Positive	299000	N/A
	Negative	299000	N/A

For Indonesian language hate-speech dataset , we used data originated from previous related work by [4]. Ibrohim and Budi [4] crawled tweets from Twitter and annotated the
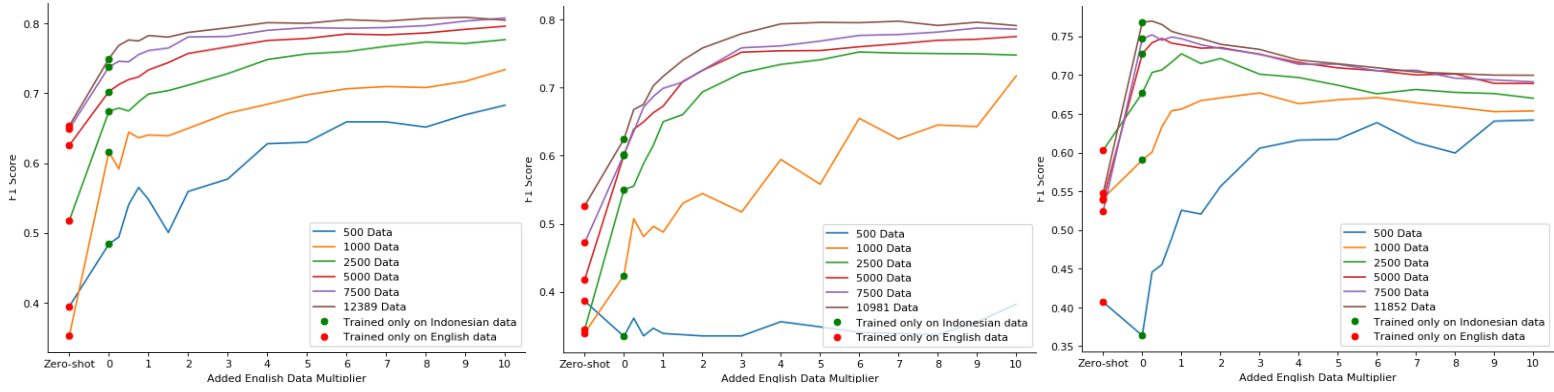
Fig. 1. Feature-based experiment result with XLM-R on [2] (left), [3] (middle), and [4] (right)
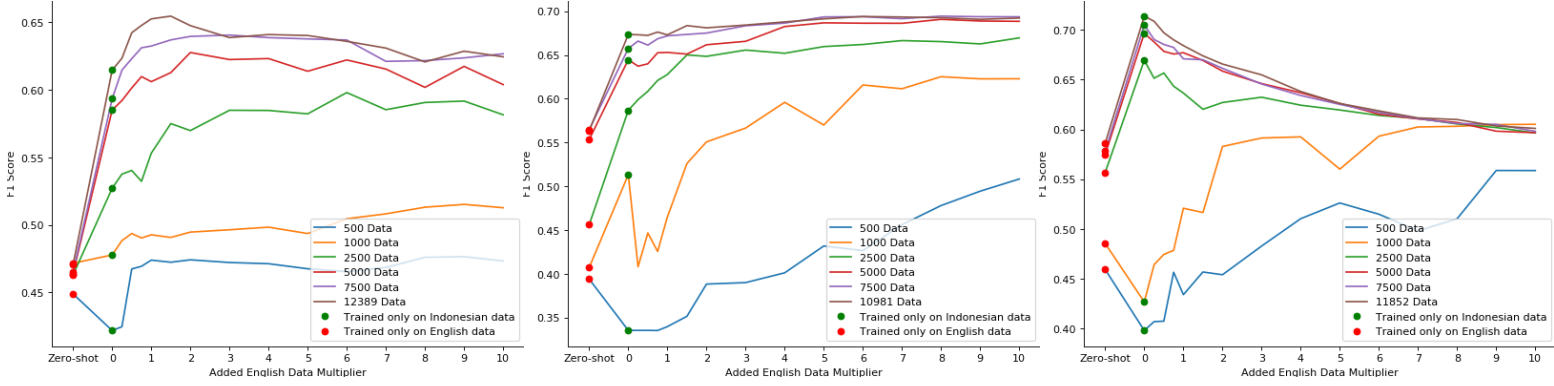

Fig. 2. Feature-based experiment result with mBERT on [2] (left), [3] (middle), and [4] (right)

text with the help of 30 diverse annotators. The annotation not only includes whether or not the text falls into hate-speech, abusive, or normal, but also its target, category, and level. For English language, we use data from Jigsaw. The data comes from various conversations over the internet annotated by dozens and up to thousands of annotators per text. The data also has rich and detailed labels up to its level and category. But the fine-grained labels don't exactly match. More importantly, the two dataset differs in its main label. Ibrohim and Budi [4] annotate the text into normal, abusive, or hate-speech meanwhile Jigsaw annotate the text into normal or toxic. We simplify the label of [4] into a simple binary label indicating the text is normal or hate-speech/abusive. The hate-speech dataset details can be seen at Table II.

TABLE II.   HATE-SPEECH DATASET DETAILS

Data		Train	Test
Source	Label		
Ibrohim & Budi [4]	Hate-speech	6578	731
	Normal	5274	1586
Jigsaw Toxic Comment	Hate-speech	152111	N/A
	Normal	1750083	N/A

D. Training Reproducibility and Hyperparameters

The data is split into training, validation, and test sets. Each experiment will train the model using the training set and

validate it to the validation set on each epoch. After each epoch, we will evaluate whether we will continue, reduce the learning rate, or stop the training process based on validation set performance and the hyperparameter set on each condition. In the end, we use the model from the best performing epoch based on its validation performance to predict the test set.

On feature-based experiment, we set the final layer dropout probability to 0.2, the learning rate reducer patience to 5, and the early stopping patience to 12. On full fine-tune experiment, we set the final layer dropout probability to 0.2, the learning rate reducer patience to 0, and the early stopping patience to 4. Every validation and prediction use 0.5 as its label threshold.

To ensure reproducibility, we set every random seed possible on each experiment. On the feature-based experiment, we average the result of 6 different runs by varying the seed from 1-6. Running the same experiment on feature-based approach results in the same final score. On the full fine-tune experiment, we only run one experiment. While the result should not differ substantially, the exact reproducibility cannot be guaranteed as the training was done on a TPU[3].

V. RESULTS

We build the system as mentioned in the previous section. The next subsection talk in detail about the result by each training approach.

---

[3] https://suneeta-mall.github.io/2019/12/22/Reproducible-ml-tensorflow.html

A. Feature-based experiment

The result of feature-based experiment with XLM-R model on all datasets can be seen on Fig 1. Through this result, we can see that adding English data can help the performance of the model. On [2] & [3] dataset, adding English data consistently improves the performance. But on [4], there's a point where the added English data results in worse performance. We hypothesize this is due to the large difference in what constitutes hate-speech (or toxic by Jigsaw dataset) between the datasets used.

The result of feature-based experiment with mBERT model on all datasets can be seen on Fig 2. The same phenomena is observed on mBERT based experiment, although the performance is substantially lower. This is expected as XLM-R is designed to improve mBERT on various design choices.

Defining the gain as the difference between monolingual and its highest multilingual performance, Table III shows the gains averaged on all dataset across total data and model. The highest gain can be seen on the lowest amount of total data used, 500, with F1-score gain of 0.176 using XLM-R model and 0.129 using mBERT model. The results suggest that the lower the amount of data used; the more gains yield by adding English data to the training set.

TABLE III.   AVERAGE F1-SCORE GAINS

Total Data	Average Gain	
	XLM-R	*mBERT*
500	0.176221	0.129394
1000	0.165718	0.109215
2500	0.118456	0.051226
5000	0.095780	0.029564
7500	0.086930	0.028043
MAX	0.077875	0.020184

B. Full fine-tune experiment

The result of fully fine-tuning all parameters, in addition to utilizing English data, proved to be effective in building a better Indonesian text classification model. On [2] dataset, the highest performance achieved on zero-shot scenario where it yielded 0.893 F1-score, improving the previous works of 0.834. On [3] dataset, the highest performance achieved on multilingual(1.5) scenario where it yielded perfect F1-score, improving the previous works of 0,9369. On [4] dataset, the highest performance achieved on multilingual(3) scenario where it yielded 0.898 F1-score and 89.9% accuracy. To provide a fair comparison with the previous work by Ibrohim & Budi [4], we also ran the experiment using the original label and monolingual scenario. The experiment yielded 89.52% average accuracy, improving the previous works of 77.36%.

VI. CONCLUSION

In this work, we investigate the use of Indonesian and English text data with multilingual language models to improve Indonesian language's lack of data. Through experiment on sentiment analysis and hate-speech detection task, we show that the addition of English text data with the utilization of multilingual language models can improve model performance on Indonesian tasks. The less Indonesian text data used, the greater improvement yielded from adding English text data. And finally, the full utilization of a pre-trained language model combined with added data by multilingual representation have successfully improved the result of Indonesian sentiment analysis and hate-speech detection.

REFERENCES

[1] J. Devlin et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: arXiv:1810.04805 [cs] (2019). arXiv: 1810.04805. URL: http://arxiv.org/abs/1810.04805.

[2] A. N. Farhan & M. L. Khodra. "Sentiment-specific word embedding for Indonesian sentiment analysis". In: 2017 International Conference on Advanced Informatics, Concepts, Theory, and Applications (ICAICTA). 2017, 1–5. DOI: 10.1109/ICAICTA.2017.8090964.

[3] I. A.P. A. Crisdayanti & A. Purwarianti. "Improving Bi-LSTM Performance for Indonesian Sentiment Analysis Using Paragraph Vector". In: (2019).

[4] M. O. Ibrohim & I. Budi. "Multi-label Hate Speech and Abusive Language Detection in Indonesian Twitter". In: Proceedings of the Third Workshop on Abusive Language Online. Association for Computational Linguistics, 2019, 46–57. DOI: 10.18653/v1/W19-3506. URL: https://www.aclweb.org/anthology/W19-3506.

[5] A. Conneau et al. "Unsupervised Cross-lingual Representation Learning at Scale". In: arXiv:1911.02116 [cs] (2020). arXiv: 1911.02116. URL: http://arxiv.org/abs/1911.02116.

[6] T. Pires, E. Schlinger & D. Garrette. "How multilingual is Multilingual BERT?" In: arXiv:1906.01502 [cs] (2019). arXiv: 1906.01502. URL: http://arxiv.org/abs/1906.01502.

[7] Y. Liu et al. "RoBERTa: A Robustly Optimized BERT Pretraining Approach". In: arXiv:1907.11692 [cs] (2019). arXiv: 1907.11692. URL: http://arxiv.org/abs/1907.11692.

[8] A. Vaswani et al. "Attention Is All You Need". In: arXiv:1706.03762 [cs] (2017). arXiv: 1706.03762. URL: http://arxiv.org/abs/1706.03762.

[9] Z. Wang et al. "Cross-lingual Alignment vs Joint Training: A Comparative Study and A Simple Unified Framework". In: arXiv:1910.04708 [cs] (2019). arXiv: 1910.04708. URL: http://arxiv.org/abs/1910.04708.

[10] T. Mikolov, Q. V. Le & I. Sutskever. "Exploiting Similarities among Languages for Machine Translation". In: arXiv:1309.4168 [cs] (2013). arXiv: 1309.4168. URL: http://arxiv.org/abs/1309.4168.

[11] Anders Søgaard, Sebastian Ruder, and Ivan Vulic. On the limitations of unsupervised bilingual dictionary induction. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 778–788, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1072. URL https://www.aclweb.org/anthology/P18-1072.

[12] Barun Patra, Joel Ruben Antony Moniz, Sarthak Garg, Matthew R. Gormley, and Graham Neubig. Bilingual lexicon induction with semi-supervision in non-isometric embedding spaces. In The 57th Annual Meeting of the Association for Computational Linguistics (ACL), Florence, Italy, July 2019. URL https://www.aclweb.org/anthology/P19-1018.

[13] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. Neural computation, 9(8):1735–1780, 1997

[14] Junyoung Chung, Çaglar Gülçehre, Kyunghyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. CoRR, abs/1412.3555, 2014.

[15] Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018a. Deep contextualized word representations. In NAACL.