A Fine-Tuned BART Pre-trained Language Model for the Indonesian Question-Answering Task

Alfonso Darren Vincentio
Informatics Department, Universitas Multimedia Nusantara, Tangerang, Indonesia
alfonso.vincentio@student.umn.ac.id

Seng Hansun
Informatics Department, Universitas Multimedia Nusantara, Tangerang, Indonesia
seng.hansun@lecturer.umn.ac.id (corresponding author)

Received: 4 December 2024 | Revised: 30 December 2024 and 16 January 2025 | Accepted: 19 January 2025

 DOI: https://doi.org/10.48084/etasr.9828

ABSTRACT

The information extraction process from a given context can be time consuming and a Pre-trained Language Model (PLM) based on the transformer architecture could reduce the time needed to obtain the information. Moreover, PLM is easily fine-tuned to accomplish certain tasks, one of which is the Question-Answering (QA) task. In literature, QA tasks are generally fine-tuned using encoder-based PLMs, such as the Bidirectional Encoder Representations from Transformers (BERT), where the generated answers come from the extraction process of the context. In order to be able to return more abstract answers, a PLM with Natural Language Generation (NLG) capability, such as the Bidirectional and Auto-Regressive Transformer (BART), is needed. In this study, we aim to fine-tune the NLG PLM using BART to build a more abstractive generative QA task. Based on the experimental results, the fine-tuned BART model performs well with an 85.84 F1 score and a 59.42 Exact Match (EM) score.

Keywords-BART; BERT; NLG; PLM; QA task

I. INTRODUCTION

In the last few decades, the use of Web technology to store data and make them available to the public has increased. This use is also supported by information storage technologies which are bigger and cheaper. However, due to the massive available data, the process of exploring and finding related information becomes a complex and expensive task and motivates the development of new research areas, such as the Question-Answering Systems (QASs) [1]. A QAS allows the user to ask questions and returns answers directly in a more Natural Language (NL) way rather than a set of relevant documents as commonly found in engine research.

Recent works have shown that transformer-based language models that leverage self-attention mechanism can have a deeper sense of language context and flow by utilizing a novel technique named masked language modeling, which allows bidirectional training and is known for BERT (Bidirectional Encoder Representations from Transformers) [2]. This language model then can be fine-tuned for a wide variety of NL understanding tasks, such as question answering, classification, named entity recognition, etc. As a result, the BERT technique could achieve state-of-the-art results on a wide variety of challenging NL understanding tasks. Natural Language Processing (NLP) is divided into two components, one is known for NL understanding and the other is Natural Language Generation (NLG).

Generally, an encoder-only model such as BERT which performs quite well on many NL understanding tasks is not very good on NLG tasks because it is not trained in a typical auto-regressive style in the first place. For a language model to generate NL, generative pre-training leverages what is known as transformer auto-regressive decoder-only architecture as it is pre-trained to predict the next word of the input sequence, hence the abbreviation is GPT (Generative Pre-trained Transformer). The model in [3] was pre-trained on 40 GB of text and resulted with 1.5 billion parameters as the largest version of it, which was able to achieve state-of-the-art results across many NLG tasks. In order to increase the applicability of a language model across more NL processing tasks, the language model requires both the encoder and decoder layers that reside in the architecture. BART (Bidirectional and Auto-Regressive Transformer) is a pre-trained language model that leverages a bidirectional encoder from BERT and an auto-regressive decoder from GPT research [4]. BART was pre-trained with a text infilling and sentence shuffling approach as it highlights the best score compared to other several document corruption (adding noise) techniques [5]. Therefore, this research aims to fine-tune the BART pre-trained language

model for question-answering tasks using a typologically diverse question-answering dataset.

In addition to using the BART model as a baseline model of QA tasks which has never been used in similar studies, we explored different hyperparameter settings and fine-tuned the BART model. The novelty of the current study is the acquisition of an accurate model for QA tasks in the Indonesian language.

II. METHODS

A. Transformer Model

A language model that is based on transformer architecture generally composes encoder and decoder layers. The transformer architecture that was proposed in [6] consists of an array of six encoders that stack on top of each other and another array of six decoders. All the encoders have the same structure but do not share the same weights, whereas each of them is subdivided into two sub-layers which are a self-attention layer and a feed-forward neural network. The self-attention layer will first be fed with the input sequence, which allows the encoder to look at other words in the input sequence while encoding a certain word. The output of the layer will then be used as an input for the feed-forward neural network. Both those layers can be found in the decoder component. An attention layer that helps the decoder to focus on relevant components of the input sequence is also added to this component. As a result, the transformer architecture was able to break previous performance records for machine translation tasks.

B. Bidirectional and Auto-Regressive Transformer (BART)

The recently popular massively pre-trained language models like BERT (Bidirectional Encoder Representations from Transformers) are able to reach previous state-of-the-art Natural Language Understanding (NLU) tasks in a novel way by utilizing bidirectional autoencoders. They are pre-trained by masking some of the tokens as well as using next sentence prediction approach [2]. On the other side of the NL processing realm which is NLG, GPT is a decoder transformer architecture language model that is well known to achieve state-of-the-art performance by highly leveraging the novel auto-regressive decoder approach that reads context from left-to-right and is pre-trained on a very large corpus which results in a large number of parameters [3, 7]. Generally, the larger parameter count can achieve better results, but it also comes at the higher cost in training computation. Following those studies, BART (Bidirectional and Auto-Regressive Transformer) is based on encoder-decoder (seq2seq) transformer architecture that is proposed to combine BERT for encoding the noised source text and GPT for reconstructing the original text by predicting the masked tokens which achieved state-of-the-art result across various downstream NLG and comprehension tasks [5, 8, 9].

BART was pre-trained by using some document corruption techniques, such as token masking (a randomly sampled subset of the input sequence were masked with [MASK] tokens), token deletion (a randomly sampled subset of the input sequence was deleted, then the model inserted new token in their position), text infilling (several text spans - a group of contiguous tokens that vary in length - was drawn based on Poisson's distribution and each span was replaced by a [MASK] token), sentence permutation (the input sequence was split based on periods or full stop punctuation (.) and the sentences were then shuffled), and document rotation (a single token was uniformly chosen at random and the document was then rotated around that token so that the document would begin with that specific token). The whole pre-training goal was for the model to be able to reconstruct the original text before being corrupted. BART is best trained using text infilling since it consistently performs well across a wide range of tasks.

C. Typologically Diverse Question Answering (TyDi QA) Dataset

TyDi QA (Typologically Diverse Question Answering) is a multilingual dataset that consists of around 200 thousand question-answer pairs in 11 typologically diverse languages, which are Bahasa Indonesia, Arabic, Bengali, Finnish, Japanese, Kiswahili, Korean, Russian, Telugu, Thai, and English [10]. The collection of linguistic properties that each language conveys the languages in TyDi QA is diverse, so the models are expected to perform well on this dataset to generalize to a wide number of languages throughout the world. A quantitative data quality study as well as example-level qualitative linguistic assessments of language phenomena not seen in English-only corpora is also given. In order to eliminate priming effects and give a genuine information-seeking task, questions are prepared by individuals who want to know the answer but do not know it yet, and data is gathered directly in each language without translation.

III. RESULTS AND DISCUSSION

The experiment first begun with loading and pre-processing the TyDi QA dataset from the HuggingFace library [11]. In this research, TyDi QA gold passage is used and is defined as the secondary task in the HuggingFace library. The pre-processing step consists of splitting the TyDi QA training set by 85% for training and the rest by 15% for validation. The TyDi QA validation set that has been provided directly from HuggingFace will be used for testing. As a result, there were a total of 4,847 question-answer pairs for the training set, 855 for the validation set, and 565 for the testing set. Every question-answer pair will also include its passage or context for question answering. After splitting the dataset, the BART pre-trained language model was loaded from the HuggingFace library along with its tokenizer to encode and decode the input sequence based on its available vocabulary. The fine-tuning phase consists of the definition of the PyTorch dataloader to load the data for every defined batch size. It also consists of hyperparameter tuning using a grid search method to locate the best hyperparameter configuration for the model training or fitting. Model evaluation was conducted with the use of F1 score, exact match, and other machine-automated metrics, such as the Bi-lingual Evaluation Understudy (BLEU) [12] and the Recall-Oriented Understudy for Gisting Evaluation (ROUGE-1, ROUGE-2, and ROUGE-L) [13]. Figure 1 shows the research workflow adopted in this study.

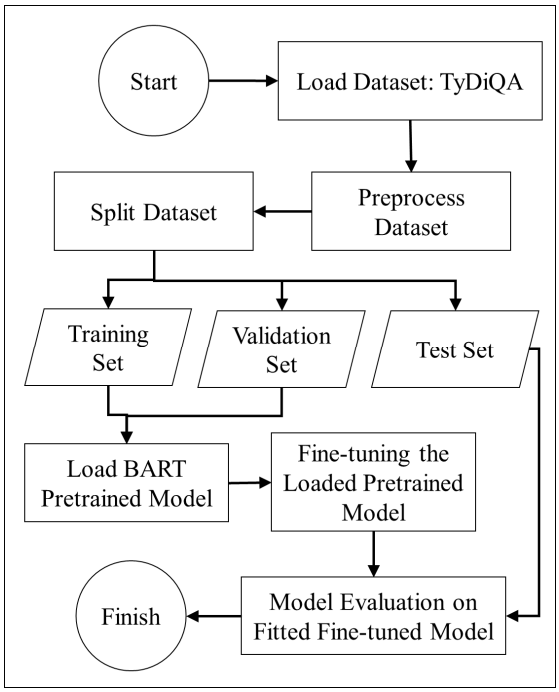

Fig. 1. Research workflow.

A. Hyperparameter Optimization

The hyperparameter optimization process consists of 36 trials across different hyperparameter configurations of the learning rate, gamma value, and maximum sequence length. Table I shows the hyperparameter optimization settings used in this study. Table II displays the training result for three epochs.

TABLE I. HYPERPARAMETER SETTINGS

Hyperparameter	Range
Epoch	[3, 4]
Gamma	[0.8, 0.9, 1.0]
Learning rate	[0.000005, 0.0001]
Max_seq_len	[128, 256, 512]

TABLE II. HYPERPARAMETER OPTIMIZATION TRIALS FOR THREE EPOCHS

Learning rate	Gamma	Max. sequence length	F1	Loss
0.000005	0.8	128	82.444	0.3997
0.000005	0.8	256	81.526	0.3982
0.000005	0.8	512	80.791	0.4141
0.0001	0.8	128	80.487	0.5123
0.0001	0.8	256	81.963	0.5100
0.0001	0.8	512	80.073	0.5846
0.000005	0.9	128	82.483	0.3927
0.000005	0.9	256	82.068	0.3904
0.000005	0.9	512	81.691	0.3953
0.0001	0.9	128	78.105	0.6092
0.0001	0.9	256	79.358	0.6182
0.0001	0.9	512	80.935	0.5716
0.000005	1	128	81.594	0.3735
0.000005	1	256	81.650	0.3826
0.000005	1	512	82.308	0.3796
0.0001	1	128	78.512	0.6764
0.0001	1	256	78.375	0.7072
0.0001	1	512	78.550	0.6112

The hyperparameter optimization trials that on four epochs can be seen in Table IV. The highest obtained scores, across all 36 trials of different hyperparameter configurations, can be seen in bold fonts.

TABLE III. HYPERPARAMETER OPTIMIZATION TRIALS FOR FOUR EPOCHS

Learning Rate	Gamma	Max. sequence length	F1	Loss
0.000005	0.8	128	82.617	0.3732
0.000005	0.8	256	82.309	0.3730
0.000005	0.8	512	82.617	0.3735
0.0001	0.8	128	81.355	0.5609
0.0001	0.8	256	80.890	0.5459
0.0001	0.8	512	80.831	0.5221
0.000005	0.9	128	82.895	0.3576
0.000005	0.9	256	83.249	0.3670
0.000005	0.9	512	82.310	0.3683
0.0001	0.9	128	82.201	0.5494
0.0001	0.9	256	80.699	0.5980
0.0001	0.9	512	80.392	0.5600
0.000005	1	128	82.639	0.3629
0.000005	1	256	82.995	0.3709
0.000005	1	512	82.504	0.3654
0.0001	1	128	78.771	0.6855
0.0001	1	256	77.795	0.7989
0.0001	1	512	78.747	0.7548

B. Analysis

After determining the best hyperparameter using a grid search across all those sets of combinations, the best hyperparameters will be used to train the model for 10 epochs. The visualization of the hyperparameter optimization phase can be seen in Figure 2.

Two testing scenarios were conducted in this study. The first one is the testing of BART pre-trained model performance before fine-tuneing (or baseline model) and the second one is the testing of the pre-trained model performance after fine-tuning (using the best hyperparameters found in the previous step). Fine-tuning plays a key role in transfer learning. It can be seen from the experimental results that the fine-tuned model performed significantly better than the baseline model, as shown in Table IV for the Indonesian language QA task.

Table V shows performance comparison of the proposed model with similar studies in the domain of NLG for the QA task in the Indonesian language. Authors in [14] introduced a new Indonesian Machine Reading Comprehension (MRC) dataset called IDK-MRC, and three different MRC models, IndoBERT, m-BERT, and XLM-R were used to validate it and compare its performance to the standard TyDiQA dataset. Authors in [15] proposed the use of the multilingual Text-to-Text Transfer Transformer (mT5) for automatic question generation in the Indonesian language. Authors in [16] compared IndoBERT-lite and RoBERTa models on TyDiQA and SQuAD datasets. Authors in [17] built an automatic question generation for Bahasa Indonesia using CopyNet. As can be seen, the proposed fine-tuned model performs on par or slightly better than other models.

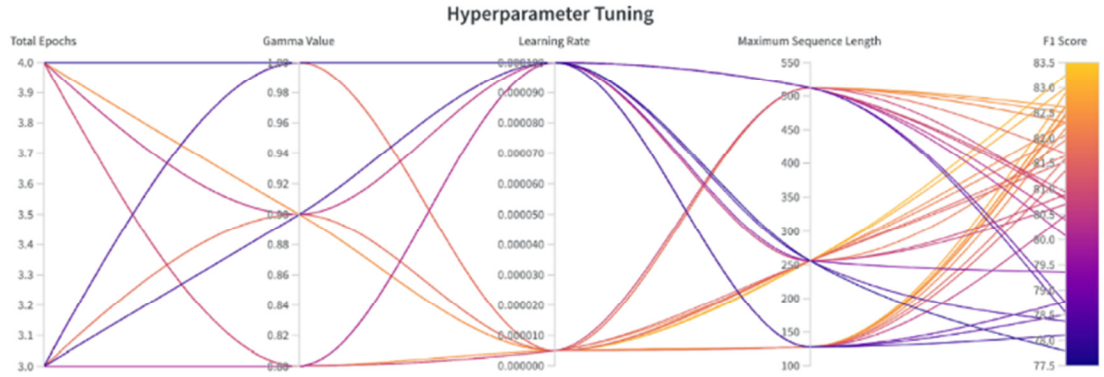

Fig. 2. Hyperparameter tuning result.

TABLE IV. FINE-TUNED AND BASELINE MODELS' RESULTS

Model	BLEU	ROUGE-1	ROUGE-2	ROUGE-L	F1	EM
Fine-Tuned	92.28	85.89	70.44	85.80	85.84	59.42
Baseline	36.84	20.19	14.43	20.07	20.21	0.00

Notes: BLEU (Bi-lingual Evaluation Understudy), EM (Exact Match), ROUGE (Recall-Oriented Understudy for Gisting Evaluation)

TABLE V. COMPARISON OF THE PROPOSED MODEL WITH OTHER STUDIES

Model	BLEU (%)	ROUGE-L (%)	F1 (%)	EM (%)
IndoBERT [14]	-	-	37.08	31.00
m-BERT [14]	-	-	41.23	36.35
XLM-R [14]	-	-	39.74	33.01
mT5 [15]	-	39.57	-	-
Roberta-1.5gb [16]	-	-	87.00	-
CopyNet GRU [17]	16.00	32.00	-	-
CopyNet Bi-GRU [17]	15.00	30.00	-	-
Fine-tuned BART (proposed)	92.28	85.80	85.84	59.42

```
PRED : 2011
LABEL: 2011

PRED : amerika serikat
LABEL: amerika serikat

PRED : gereja kelahiran
LABEL: gereja kelahiran

PRED : stock ef field, south yor ks hire
LABEL: sh ef field, south yor ks hire

PRED : tvri dan sctv
LABEL: tvri dan sctv

PRED : megawati soekarnoputri
LABEL: megawati soekarnoputri

PRED : kelurahan gelora, kecamatan tanah abang, jakarta pusat
LABEL: kelurahan gelora, kecamatan tanah abang, jakarta pusat

PRED : bandar udara internasional h. a. s. han and jo ed din
LABEL: bandar udara internasional h. a. s. han and jo ed din

PRED : 18 onia
LABEL: franc swiss

PRED : 12. 94 6 meter persegi
LABEL: 12. 94 6 meter persegi
```

Fig. 3. Testing samples of the fine-tuned BART model.

Figure 3 shows some prediction results along with their labels of the fine-tuned BART pre-trained language model for the Indonesian QA task. As can be seen from the Figure, the proposed fine-tuned model performed very well with one overall false prediction result and one word false prediction result. The other examples were predicted correctly by the proposed model.

IV. CONCLUSIONS

Question-answering (QA) task is one of the most popular tasks in Natural Language Generation (NLG) gaining traction lately. However, few studies are available focusing on QA tasks in the Indonesian language. Therefore, in this study, we explored, fine-tuned, and optimized the Bidirectional and Auto-Regressive Transformer (BART) model for this task.

The best implementation of the fine-tuned BART pre-trained language model for QA tasks using a typologically diverse QA dataset, based on the results of the conducted research, was when the gamma hyperparameter was set to 0.9, the learning rate value to 0.000005, and the maximum sequence length to 256 with 10 epochs. With these values we could achieve as high as 92.28 BLEU, 85.89 ROUGE-1, 70.44 ROUGE-2, 85.8 ROUGE-L, 85.84 F1, and 59.42 EM.

There many different aspects that can be explored regarding future research,. The first one is fine-tuning the proposed model using different datasets, such as SQuAD [18-20]. Another approach is to fine-tune different kinds of language models

rather than the BART pre-trained language model used in this study [21-24]. Moreover, the use of advanced recurrent neural networks, such as the long short-term memory and the gated recurrent unit, that could handle the long-term dependency of the dataset can be explored [25-27] and the the graph mining approach could be considered [28]. Lastly, we could leverage the use of natural language processing benchmarks to evaluate the proposed model's performance [29-32].

ACKNOWLEDGMENT

The authors would like to acknowledge the support and facility given by the Research and Community Outreach Department (LPPM), Universitas Multimedia Nusantara.

REFERENCES

[1] A. Bouziane, D. Bouchiha, N. Doumi, and M. Malki, "Question Answering Systems: Survey and Trends," *Procedia Computer Science*, vol. 73, pp. 366–375, Jan. 2015, https://doi.org/10.1016/j.procs.2015.12.005.

[2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Minneapolis, MN, USA, Mar. 2019, pp. 4171–4186, https://doi.org/10.18653/v1/N19-1423.

[3] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language Models are Unsupervised Multitask Learners," 2019. Available: https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf

[4] S. Giddaluru, S. M. Maturi, O. Ooruchintala, and M. Munirathnam, "Stance Detection in Hinglish Data using the BART-large-MNLI Integration Model," *Engineering, Technology & Applied Science Research*, vol. 14, no. 4, pp. 15477–15481, Aug. 2024, https://doi.org/10.48084/etasr.7741.

[5] M. Lewis *et al.*, "BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension," in *58th Annual Meeting of the Association for Computational Linguistics*, Jul. 2020, pp. 7871–7880, https://doi.org/10.18653/v1/2020.acl-main.703.

[6] A. Vaswani *et al.*, "Attention is all you need," in *31st International Conference on Neural Information Processing Systems*, Long Beach, CA, USA, Dec. 2017, pp. 6000–6010.

[7] T. B. Brown *et al.*, "Language models are few-shot learners," in *34th International Conference on Neural Information Processing Systems*, Vancouver, BC, Canada, Dec. 2020, pp. 1877–1901.

[8] Y. Liu *et al.*, "Multilingual Denoising Pre-training for Neural Machine Translation," *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 726–742, Dec. 2020, https://doi.org/10.1162/tacl_a_00343.

[9] Y. Tang *et al.*, "Multilingual Translation from Denoising Pre-Training," in *Findings of the Association for Computational Linguistics: ACL-IJCNLP*, Aug. 2021, pp. 3450–3466, https://doi.org/10.18653/v1/2021.findings-acl.304.

[10] J. H. Clark *et al.*, "TyDi QA: A Benchmark for Information-Seeking Question Answering in Typologically Diverse Languages," *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 454–470, Jul. 2020, https://doi.org/10.1162/tacl_a_00317.

[11] T. Wolf *et al.*, "Transformers: State-of-the-Art Natural Language Processing," in *Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Jul. 2020, pp. 38–45, https://doi.org/10.18653/v1/2020.emnlp-demos.6.

[12] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: a method for automatic evaluation of machine translation," in *40th Annual Meeting on Association for Computational Linguistics*, Philadelphia, PA, USA, Jul. 2012, pp. 311–318, https://doi.org/10.3115/1073083.1073135.

[13] C.-Y. Lin, "ROUGE: A Package for Automatic Evaluation of Summaries," in *Text Summarization Branches Out*, Barcelona, Spain, Apr. 2004, pp. 74–81.

[14] R. A. Putri and A. Oh, "IDK-MRC: Unanswerable Questions for Indonesian Machine Reading Comprehension," in *Conference on Empirical Methods in Natural Language Processing*, Abu Dhabi, United Arab Emirates, Dec. 2022, pp. 6918–6933, https://doi.org/10.18653/v1/2022.emnlp-main.465.

[15] M. Fuadi and A. D. Wibawa, "Automatic Question Generation from Indonesian Texts Using Text-to-Text Transformers," in *International Conference on Electrical and Information Technology*, Malang, Indonesia, Sep. 2022, pp. 84–89, https://doi.org/10.1109/IEIT56384.2022.9967858.

[16] B. Richardson and A. Wicaksana, "Comparison of indobert-lite and roberta in text mining for Indonesian language question answering application," *International Journal of Innovative Computing, Information and Control*, vol. 18, no. 6, pp. 1719–1734, Dec. 2022, https://doi.org/10.24507/ijicic.18.06.1719.

[17] M. M. Henry, G. N. Elwirehardja, and B. Pardamean, "Automatic question generation for bahasa indonesia examination using copynet," *Procedia Computer Science*, vol. 245, pp. 953–962, Jan. 2024, https://doi.org/10.1016/j.procs.2024.10.323.

[18] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, "SQuAD: 100,000+ Questions for Machine Comprehension of Text," in *Conference on Empirical Methods in Natural Language Processing*, Austin, TX, USA, Nov. 2016, pp. 2383–2392, https://doi.org/10.18653/v1/D16-1264.

[19] P. Rajpurkar, R. Jia, and P. Liang, "Know What You Don`t Know: Unanswerable Questions for SQuAD," in *56th Annual Meeting of the Association for Computational Linguistics*, Melbourne, Australia, Jul. 2018, pp. 784–789, https://doi.org/10.18653/v1/P18-2124.

[20] M. Artetxe, S. Ruder, and D. Yogatama, "On the Cross-lingual Transferability of Monolingual Representations," in *58th Annual Meeting of the Association for Computational Linguistics*, Jul. 2020, pp. 4623–4637, https://doi.org/10.18653/v1/2020.acl-main.421.

[21] L. Xue *et al.*, "mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer," in *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Jun. 2021, pp. 483–498, https://doi.org/10.18653/v1/2021.naacl-main.41.

[22] Y. Liu *et al.*, "RoBERTa: A Robustly Optimized BERT Pretraining Approach." arXiv, Jul. 26, 2019, https://doi.org/10.48550/arXiv.1907.11692.

[23] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter," in *33rd Conference on Neural Information Processing Systems*, Vancouver, BC, Canada, Dec. 2019, pp. 1–5.

[24] C. Raffel *et al.*, "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer," *Journal of Machine Learning Research*, vol. 21, no. 140, pp. 1–67, 2020.

[25] S. Hansun, A. Suryadibrata, R. Nurhasanah, and J. Fitra, "Tweets sentiment on ppkm policy as a covid-19 response in indonesia," *Indian Journal of Computer Science and Engineering*, vol. 13, no. 1, pp. 51–58, Feb. 2022, https://doi.org/10.21817/indjcse/2022/v13i1/221301302.

[26] S. Hansun, A. Wicaksana, and A. Q. M. Khaliq, "Multivariate cryptocurrency prediction: comparative analysis of three recurrent neural networks approaches," *Journal of Big Data*, vol. 9, no. 1, Apr. 2022, Art. no. 50, https://doi.org/10.1186/s40537-022-00601-7.

[27] T. A. Wotaifi and B. N. Dhannoon, "An Effective Hybrid Deep Neural Network for Arabic Fake News Detection," *Baghdad Science Journal*, vol. 20, no. 4, pp. 1392–1392, Aug. 2023, https://doi.org/10.21123/bsj.2023.7427.

[28] A. K. Jassim, M. J. Hamzah, and A. H. Aliwy, "Using Graph Mining Method in Analyzing Turkish Loanwords Derived from Arabic Language," *Baghdad Science Journal*, vol. 19, no. 6, pp. 1369–1369, Dec. 2022, https://doi.org/10.21123/bsj.2022.6008.

[29] B. Wilie *et al.*, "IndoNLU: Benchmark and Resources for Evaluating Indonesian Natural Language Understanding," in *1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*

and the 10th International Joint Conference on Natural Language Processing, Suzhou, China, Dec. 2020, pp. 843–857, https://doi.org/10.18653/v1/2020.aacl-main.85.

[30] S. Cahyawijaya *et al.*, "IndoNLG: Benchmark and Resources for Evaluating Indonesian Natural Language Generation," in *Conference on Empirical Methods in Natural Language Processing*, Nov. 2021, pp. 8875–8898, https://doi.org/10.18653/v1/2021.emnlp-main.699.

[31] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. Bowman, "GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding," in *EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, Brussels, Belgium, Nov. 2018, pp. 353–355, https://doi.org/10.18653/v1/W18-5446.

[32] S. Gehrmann *et al.*, "The GEM Benchmark: Natural Language Generation, its Evaluation and Metrics," in *1st Workshop on Natural Language Generation, Evaluation, and Metrics*, Bangkok, Thailand, Aug. 2021, pp. 96–120, https://doi.org/10.18653/v1/2021.gem-1.10.