JOURNAL OF INFORMATION AND COMMUNICATION TECHNOLOGY
http://e-journal.uum.edu.my/index.php/jict

How to cite this article:
Lucky, H., & Suhartono, D. (2022). Investigation of pre-trained bidirectional encoder representations from transformers checkpoints for Indonesian abstractive text summarization. *Journal of Information and Communication Technology, 21*(1), 71-94. https://doi.org/10.32890/jict2022.21.1.4

Investigation of Pre-Trained Bidirectional Encoder Representations from Transformers Checkpoints for Indonesian Abstractive Text Summarization

*[1]Henry Lucky & [2]Derwin Suhartono
[1,2]Computer Science Department,
Bina Nusantara University, Indonesia

[1]henry.lucky@binus.ac.id
[2]dsuhartono@binus.edu
*Corresponding author

Received:16/3/2021 Revised: 9/6/2021 Accepted: 7/7/2021 Published: 11/11/2021

ABSTRACT

Text summarization aims to reduce text by removing less useful information to obtain information quickly and precisely. In Indonesian abstractive text summarization, the research mostly focuses on multi-document summarization which methods will not work optimally in single-document summarization. As the public summarization datasets and works in English are focusing on single-document summarization, this study emphasized on Indonesian single-document summarization. Abstractive text summarization studies in English frequently use Bidirectional Encoder Representations from Transformers (BERT), and since Indonesian BERT checkpoint is available, it was employed in this study. This study investigated

the use of Indonesian BERT in abstractive text summarization on the IndoSum dataset using the BERTSum model. The investigation proceeded by using various combinations of model encoders, model embedding sizes, and model decoders. Evaluation results showed that models with more embedding size and used Generative Pre-Training (GPT)-like decoder could improve the Recall-Oriented Understudy for Gisting Evaluation (ROUGE) score and BERTScore of the model results.

Keywords: Abstractive text summarization, BERTSum model, BERT Score, GPT-like decoder, ROUGE score.

INTRODUCTION

Text summarization is one of the solutions that has been used to obtain quick and accurate data because it allows information to be gained more quickly and precisely without losing the meaning from the actual document (Widyassari et al., 2019). In its application in technology, text summarization can facilitate several aspects of work on search engines, digital business, and journalistic media (Adelia et al., 2019). In general, there are two approaches to do text summarization, which are extractive and abstractive. In the extractive approach, the system generates summaries by selecting important information in form of sentences or phrases from the source text, which is similar to classification problems. In contrast, the abstractive approach generates summaries by paraphrasing and generating new sentences or phrases while keeping the information related to the source text. Text summarization with extractive approaches is easier to implement and has more straightforward methods; therefore, the research in that area are more developed than research in abstractive approaches. However, the abstractive approach is ideal for summarizing text as it follows how humans generate summaries (Devianti & Khodra, 2019; Nallapati et al., 2016b).

The most used model for abstractive text summarization is sequence-to-sequence models, which consist of encoder and decoder as they give great results. Several works have used this model (Nallapati et al., 2016a; Nallapati et al., 2016b; See et al., 2017; Shi et al., 2021; Zhou et al., 2017), starting from Rush et al. (2015) who successfully applied the model in machine translation tasks. Moreover, with the emergence of the transformer model (Vaswani et al., 2017), which

is a breakthrough in Natural Language Processing (NLP), other breakthrough language models that use contextual representation pre-training have also emerged, such as Embeddings from Language Models (ELMo) (Peters et al., 2018), Generative Pre-trained Transformer-2 (GPT-2) (Radford et al., 2019), Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019), and Bidirectional and Auto-Regressive Transformer (BART) (Lewis et al., 2019). From that point, research on abstractive text summarization have begun to shift using these models as references because they are considered best practices. The most frequently encountered studies are using BERT as the foundation to build their models (Liu & Lapata, 2020; Rothe et al., 2020; Savelieva et al., 2020; Zhang, Kishore, Wu, et al., 2019). BERT's success has influenced other researchers to produce their own BERT version in other languages, such as the Chinese BERT (Cui et al., 2019), French BERT (Martin et al., 2019), German BERT (Rönnqvist et al., 2019), and Indonesian BERT (Koto, Rahimi, Lau, et al., 2020; Wilie et al., 2020).

As this paper was written, there were two well-known large-scale Indonesian BERT checkpoints with the same name, IndoBERT (Koto, Rahimi, Lau, et al., 2020; Wilie et al., 2020), which are used for several Indonesian NLP and Natural Language Understanding (NLU) tasks for benchmarking. Wilie et al. (2020) leveraged their pre-trained IndoBERT model checkpoints for single-sentence classification, sentence-pair classification, single-sentence sequence labeling, and sentence-pair sequence labeling tasks on 12 datasets. Meanwhile, Koto, Rahimi, Lau, et al. (2020) leveraged their model checkpoint for sequence labeling, semantic, and coherency tasks on nine datasets, including IndoSum in an extractive manner. There are no benchmarks for abstractive text summarization tasks from both papers.

However, Indonesian abstractive text summarization is recently gaining attention because the newly released large-scale dataset named Liputan6 (Koto, Lau & Baldwin, 2020) has highly abstractive gold summaries. There is also another summarization dataset, IndoSum (Kurniawan & Louvan, 2018). Both datasets are news document-summary pairs and have the potential to become benchmark datasets in Indonesian text summarization, such as Gigaword corpus (Rush et al., 2015), Newsroom (Grusky et al., 2018), XSum (Narayan et al., 2018), and CNN/Daily Mail (CNNDM) (Hermann et al., 2015) in English. However, the models and methods used for Indonesian abstractive text summarization are considered obsolete as compared

to the English text summarization models. The methods that have been employed include the use of Sentence Fusion (Christie & Khodra, 2016), Abstractive Meaning Representation (Severina & Khodra, 2019), Genetic Semantic Graph (Devianti & Khodra, 2019), and Bidirectional Gated Recurrent Unit (BiGRU) in sequence-to-sequence models (Adelia et al., 2019). The methods utilized are outdated as research in English are using pre-trained language models for this task. There are also some problems regarding the evaluation result as there are hardly any standards for datasets and evaluation methods used in Indonesian text summarization. Since research in English frequently use BERT in their abstractive text summarization models, this paper would like to investigate and leverage two IndoBERT checkpoints (Koto, Rahimi, Lau, et al., 2020; Wilie et al., 2020) for the task in this paper.

This paper aims to investigate two IndoBERT checkpoints for abstractive text summarization tasks using the state-of-the-art model utilizing BERT, following Liu and Lapata (2020). The investigation proceeds by using various combinations of model encoders, model embedding sizes, and model decoders based on the findings while investigating IndoBERT checkpoints. The result of this study is reported with the IndoSum dataset on Recall-Oriented Understudy for Gisting Evaluation (ROUGE) (Lin, 2004) and BERTScore (Zhang, Kishore, Wu, et al., 2019) metrics.

RELATED WORKS

This section reviews the related research works to contextualize the present work. This section is divided into two parts: a review of research on abstractive text summarization in English for the general information of abstractive text summarization, and research on abstractive text summarization in Indonesian to identify the current development in Indonesian research.

English Abstractive Text Summarization

In recent works of English abstractive text summarization, the most used models are transformer-based and sequence-to-sequence models. Hoang et al. (2019) used a pre-trained Generative Pre-Training (GPT) model as a starting point for summarizing abstractive text. Their

research proposed source embedding and domain-adaptive training that could facilitate the use of the GPT model as a text summary. Even though the model used parameters from the pre-trained GPT, there were differences in the type of language between the pre-training dataset and the article summary dataset, which were fictional stories and new. With domain-adaptive training, the model was trained to produce a type of language similar to the training dataset. Next, the model was trained on three datasets, Newsroom (Grusky et al., 2018), XSum (Narayan et al., 2018), and CNNDM (Hermann et al., 2015), to produce a summary of an article. The model scored a significant increase in ROUGE-L on two datasets, Newsroom and XSum. At the same time, the other model achieved higher scores in human evaluation on non-redundancy, coherence, and focus.

There are also some works that incorporate BERT. One of them utilized BERT in a sequence-to-sequence model that had a decoder (Zhang, Kishore, Wu, et al., 2019). The decoder used was a standard transformer decoder. However, there was a difference with this BERT, where it was pre-trained while the decoder was trained from scratch. With this situation, it was afraid that the decoder would not be able to use the context of BERT optimally; therefore, a two-stage decoding process was created to make maximum use of BERT's capabilities. On the CNNDM dataset, compared to previous studies, this study succeeded in improving performance with ROUGE. In another work (Liu & Lapata, 2020), BERT was also used in a sequence-to-sequence manner. This research proposed a new training method where the encoder and decoder had different optimizers. The encoder was configured to learn slower because it had gone through pre-training, while the decoder learned faster to keep up with the encoder. In addition, a two-stage training was carried out where in the first stage, the encoder was trained on summarizing extractive text, and then in the second stage, the model was trained on summarizing abstractive text. They produced excellent scores in extractive and abstractive for minimal parameter models. Afterward, the model in the previous work (Liu & Lapata, 2020) was used by Savelieva et al. (2020) to produce abstractive summarization of written instructions.

Indonesian Abstractive Text Summarization

There are numerous extractive text summarizations in Indonesian (Cai et al., 2019; Christian et al., 2016; Garmastewira & Khodra,

2019; Halim et al., 2020; Hidayat et al., 2015; Najibullah, 2015); however, the abstractive part is not further investigated. Although there are already particular datasets for summarizing text (Koto, Lau & Baldwin. 2020; Kurniawan & Louvan, 2018), these datasets are not widely used. One of the initial research in summarizing abstractive texts (Christie & Khodra, 2016) summarized many documents by using the Sentence Fusion method. Sentence Fusion is a method for generating a sentence from a collection of similar sentences and has been called a semi-extractive method. In implementing this method, machine learning was not required in the process and was more inclined to a clustering method with light pre-processing in the form of Part-of-Speech (POS) tagging and eliminating stopwords. The dataset used was in the form of Indonesian news articles from a previous research (Ilyas, 2015) with additional data taken by the researchers themselves. They used the ROUGE metrics in their research. However, in evaluating the clustering method, they did not mention the ROUGE scores. Oddly, they did not use ROUGE for evaluating the produced summary. Instead, human evaluation was used on grammatical and informativity. Another research (Devianti & Khodra, 2019) adapted the Genetic Semantic Graph method by using extraction of Subject, Verb, Object, and Adverbial (SVOA) from sentences plus some rules, cosine equations based on word embedding to calculate word similarities, and heuristic rules for Natural Language Generation (NLG). The dataset used was in the form of news articles taken from previous research (Christie & Khodra, 2016; Garmastewira & Khodra, 2019). ROUGE-2 recall was used for evaluating the summaries.

The Abstractive Meaning Representation (AMR) method was used by Severina and Khodra (2019) to summarize the text of many documents in an abstractive way. The existing AMR graph was a highly specific tree structure for English because it was based on grammar rules. This study tried to make an AMR graph in Indonesian and used it in summarizing text. Before being made into the AMR graph, the existing documents went through Agglomerative Hierarchical Clustering to select sentences that represented multiple documents. After the AMR graph was created, the graph was re-selected by using Integer Linear Programming (ILP) and supervised learning via the perceptron. With the dataset that was self-gathered by the researchers, this study used ROUGE recall for evaluating the summaries.

The works mentioned are multi-document abstractive summarizations, which depend on clustering (Christie & Khodra, 2016) and graphs

(Devianti & Khodra, 2019; Severina & Khodra, 2019) to pool the documents in the dataset. Such systems will not work optimally in single-document abstractive text summarization because of the difference in the number of the texts. In addition, the methods make the systems very dependent on the limited Indonesian resource available in the summarization dataset. Meanwhile, modern works in English (Hoang et al., 2019; Liu & Lapata, 2020; Savelieva et al., 2020; Zhang, Cai, Xu, et al., 2019) used transfer learning with pre-trained models, which have been pre-trained on other datasets, to achieve better results in single-document abstractive text summarization.

For single-document abstractive text summarization, there is a work that utilized the sequence-to-sequence model (Adelia et al., 2019). This work used BiGRU as an encoder and Gated Recurrent Unit (GRU) with the attentional model as a decoder alongside a dataset in the form of an Indonesian journal document with an abstract as the summary target that was self-gathered by the researchers. The summary results contained repeated words, and the cohesion of the sentence was still not optimal, whereby the language elements in the sentences were used to construct the summary lack a relationship with one another.

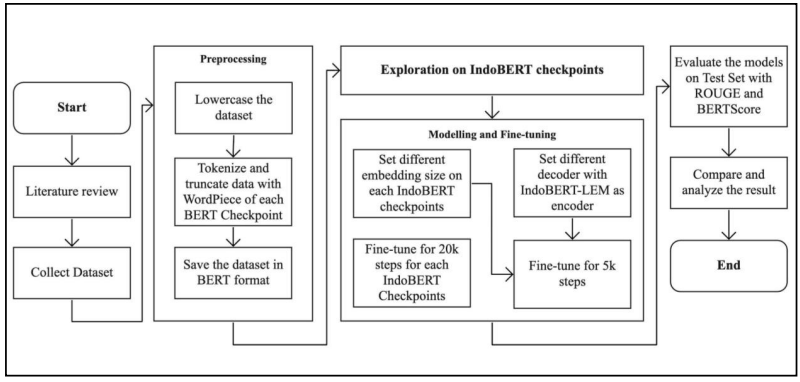
It can be concluded that Indonesian abstractive text summarization methods used in available research are still not optimal for single-document abstractive text summarization. Furthermore, there is another problem with the datasets and evaluation metrics employed. Each research used different datasets and evaluation metrics, which made the methods difficult to compare. As there is a large gap between the research progress in English and Indonesian abstractive text summarizations, this paper's objective is to close this gap. This paper addresses two problems that can be found in Indonesian research. First, to make the result easy to compare with other papers, the Indonesian public dataset IndoSum is used for training and testing the model. This paper also employs ROUGE (Lin, 2004) and BERTScore (Zhang, Kishore, Wu, et al., 2019) as evaluation metrics, following Koto, Lau & Baldwinl. (2020). Second, to reach the results gained in English research, BERT is used as there are currently two Indonesian BERT checkpoints with no benchmark on abstractive text summarization tasks. Experimental research in this paper investigates the use of them in building abstractive text summarization models.

METHODOLOGY

This section explains the methodology used in this paper to reach the research objectives, which starts from literature review, data collection, pre-processing, checkpoints exploration, modeling and fine-tuning, and evaluation as shown in Figure 1. A literature review was conducted to identify research problems in abstractive text summarization, mainly Indonesian. The next step was to collect a dataset that would be used in the training and model evaluation. Then, a pre-processing of the dataset was performed. After conducting an exploration on the IndoBERT checkpoints, the designing of the model using BERT was carried out. The model that was designed would then be fine-tuned and then evaluated with the ROUGE and BERTScore metrics.

Figure 1

The IndoBERT Checkpoints Investigation Method



Models and Exploration on IndoBERT Checkpoints

The model used in this paper for abstractive text summarization followed the model by Liu and Lapata (2020). Their model utilized a pre-trained BERT checkpoint as the encoder and standard transformers for the decoder. There were some variants of the model, namely extractive summarization (BERTSumExt), abstractive summarization (BERTSumAbs), and hybrid summarization that utilized extractive and abstractive methods (BERTSumExtAbs). This paper used the abstractive model, BERTSumAbs, for the experiments. For the

encoder-side, two Indonesian BERT checkpoints were applied. The first was IndoBERT (indobert-base-p2) from Wilie et al. (2020), which was trained in two phases for 1M and 68k steps. In the first phase, it was pre-trained with 128 tokens, while in the second phase, it was pre-trained with 512 tokens. The model was pre-trained on the Indo4B dataset, consisting of 3.6B words from various sources that could be seen as a general dataset. The second was IndoBERT (indobert-base-uncased) from Koto, Rahimi, Lau, et al. (2020), which was trained for 2.4M steps on the dataset, comprising 220M words from three main corpora, Indonesian Wikipedia, news articles, and Indonesian Web Corpus. To avoid misleading as they both have identical names, from this point they will be called IndoBERT-NLU (indobert-base-p2) and IndoBERT-LEM (indobert-base-uncased), following their paper titles.

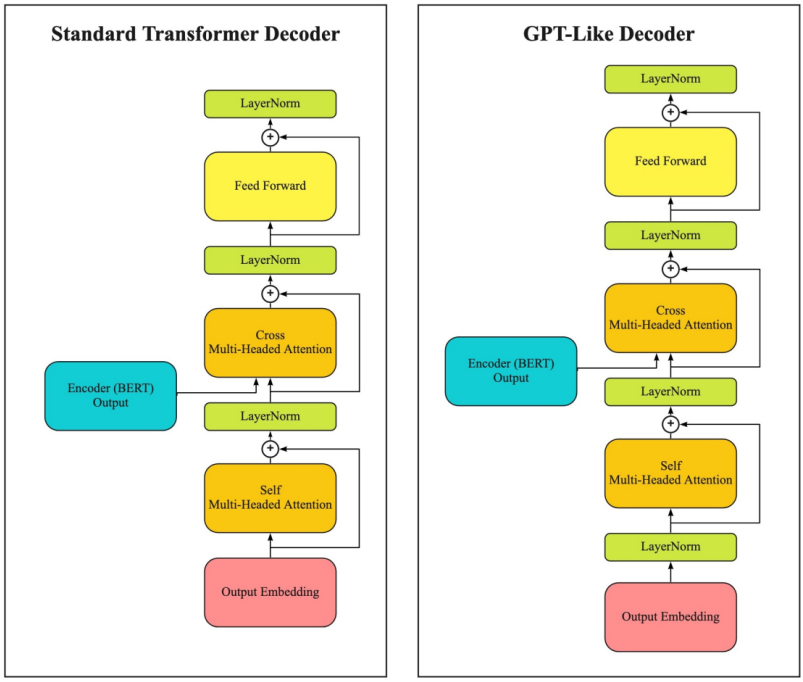
As both checkpoints came from benchmark papers, the papers used the IndoBERT checkpoints for benchmarking in some tasks. For IndoBERT-NLU, there were 12 tasks divided into four categories: single-sentence classification, single-sentence sequence-tagging, sentence-pair classification, and sentence-pair sequence labeling. For IndoBERT-LEM, there were seven tasks divided into three categories: morpho-syntax and sequence labeling, semantic, and discourse coherence. There wasa summarization task in the semantic category; however, they only benchmarked the extractive model. There was no abstractive summarization benchmark with IndoBERT from their respective paper.

Both shared the same parameter numbers. Both had 12 layers, a hidden size of 768, filter size of 3,072, and 12 attention heads. Nevertheless, the vocabulary (vocab) size and embedding layers were different. IndoBERT-NLU claimed it had a vocab size of 30,522; however, it was found that it had a vocab size of 30,521 in the actual checkpoint. In contrast to the vocab size, the embedding size in the model was set to 50,000.

Meanwhile, IndoBERT-LEM had a vocab and embedding size of 31,923. As for the decoder, this paper used six layers of standard transformer decoder with a hidden size of 768, filter size of 2048, and 8 attention heads (the architecture can be seen in Figure 2). Note that this decoder was not pre-trained. The embedding and vocab size of the decoder followed each of the IndoBERT checkpoints.

Figure 2

Architecture Comparison of Standard Transformer Decoder (left) and GPT-Like Decoder (right).



The models were fine-tuned for 20,000 steps in total (~44 epochs) to the IndoSum dataset. The encoder had already been pre-trained while the decoder had been initialized randomly. The fine-tuning might be unstable as the encoder might overfit while the decoder underfit or vice-versa. In order to make the fine-tuning more stable, two Adam optimizers with $\beta_1 = 0.9$ and $\beta_2 = 0.999$ for encoder and decoder were used with different learning rates and warm-up steps as presented in Equation 1.

$$lr_x = \widetilde{lr}_x \cdot \left(step^{-0.5}, step \cdot warmup_x^{-1.5}\right) \quad (1)$$

where x denotes either encoder e or decoder d. For the encoder, it was set as $\widetilde{lr}_e = 2e^{-3}$ and $warmup_e = 8{,}000$ while for the decoder, it was set as $\widetilde{lr}_d = 5e^{-2}$ and $warmup_d = 4{,}000$. This learning schedule would make the pre-trained encoder learn to be slower and the decoder to be faster while keeping the fine-tuning stable as was done for 20,000 steps.

Investigated Model Variants

The main model in this paper consisted of two BERTSumAbs models with different encoders, IndoBERT-NLU and IndoBERT-LEM. This section describes three other variations of the model.

IndoBERT-NLU-30kEmb: Earlier, it was mentioned that Indo BERT-NLU had different sizes of embeddings and vocab configuration so that IndoBERT-NLU was made to have the same size as Indo BERT-LEM. Another BERTSumAbs IndoBERT-NLU was fine-tuned with an embedding size similar to its vocab size of 30,521.

IndoBERT-LEM-50kEmb: Further investigation studied whether increasing the size of the embedding in IndoBERT-LEM to 50,000, as in IndoBERT-NLU, could increase the value of the evaluation. Another BERTSumAbs IndoBERT-LEM with an embedding size of 50,000 was fine-tuned.

IndoBERT-LEM-GPT: BERT was a stack of transformer encoders and GPT-2 was a stack of transformer decoders. Meanwhile, GPT-2 was known for its capability to train data and the parameter contained in the data. Some tinkering was made to the architecture where the layer normalization (Ba et al., 2016) was moved to the input of each sub-block and an additional layer normalization was added after the final attention block as seen in Equation 2 – Equation 5 and Figure 2.

$$\widetilde{h}^{l} = LN\left(h^{l-1}\right), \tag{2}$$

$$\widetilde{h}^{l} = LN\left(\widetilde{h}^{l} + SelfMHAtt(\widetilde{h}^{l})\right), \tag{3}$$

$$\widetilde{h}^{l} = LN\left(\widetilde{h}^{l} + CrossMHAtt(q_e, k_e, v\widetilde{h}^{l})\right), \tag{4}$$

$$\widetilde{h}^{l} = LN\left(\widetilde{h}^{l} + FFN(\widetilde{h}^{l})\right) \tag{5}$$

where h^0 output embedding and $\tilde{h}^0$ indicates temporary value. *LN* is the layer normalization, *MHAtt* is the multi-headed attention, *SelfMHAtt* gets input from y, while *CrossMHAtt* gets input from *SelfMHAtt(value)* and encoder output (query & key), and superscript *l* indicates the number of layers.

It is interesting to observer whether a GPT-like architecture in the decoder model could increase the evaluation scores. The three model

variants were fine-tuned with the same hyperparameters as the main models for 5,000 steps (~11 epochs). This paper also showed the results of the main models' checkpoint at 5,000 steps for comparison.

Table 1 shows the combination of embedding size, encoder, and decoder for all the models mentioned, including the main models and variant models.

Table 1

Details of Experiments on the Models.

No.	Model	Emb. Size		Steps		Encoder		Decoder	
		30k	50k	5k	20k	LEM	NLU	STD	GPT
	Main								
1	BertSumAbs w/ IndoBERT-LEM	✓			✓	✓		✓	
2	BertSumAbs w/ IndoBERT-NLU		✓		✓		✓	✓	
3	BertSumAbs w/ IndoBERT-LEM-5kStep	✓		✓		✓		✓	
4	BertSumAbs w/ IndoBERT-NLU-5kStep		✓	✓			✓	✓	
	Variants								
5	BertSumAbs w/ IndoBERT-NLU-30kEmb	✓		✓			✓	✓	
6	BertSumAbs w/ IndoBERT-LEM-50kEmb		✓	✓		✓		✓	
7	BertSumAbs w/ IndoBERT-LEM-GPT	✓		✓		✓			✓

The 30k embedding size varied from model to model, following the checkpoint's vocab size. For IndoBERT-LEM, the embedding size as 31,923, while for IndoBERT-NLU, it was 30,521. STD in the decoders stood for standard transformer decoder.

RESULTS AND DISCUSSION

The models were evaluated using the IndoSum dataset (Kurniawan & Louvan, 2018). Another summarization dataset, Liputan6 (Koto, Lau & Baldwin, 2020), was actually more abstractive and much more extensive than IndoSum. However, IndoBERT-LEM used the data in pre-training. There might be bias when the dataset was used with the IndoBERT-LEM checkpoint, and to compare the checkpoints fairly, this paper only employed the IndoSum dataset as a benchmark. IndoSum consisted of 19k document-summary pairs with 5-fold cross-validation to make the result more general as it was a low resource dataset. However, only the first fold of the dataset was used to make benchmarking easier for future work. The gold summaries on IndoSum appeared to have a high degree of extraction, signifying that it copied sentences from the source articles most of the time.

The case was lowered and the input documents and gold summaries were truncated to 512 tokens and 128 tokens, respectively, during the fine-tuning. The findings reported the ROUGE *F1* scores (Lin, 2004), particularly R-1 (unigram overlaps) and R-2 (bigram overlaps) for informativeness and R-L (longest common subsequence) for fluency, as well as BERTScore (Zhang, Kishore, Wu, et al., 2019), following Koto, Lau & Baldwin. (2020) as the metrics to count the probability based on BERT's contextual embedding that could capture more similarities between the gold summaries and system summaries. This paper used the ninth layer of cased version of multilingual BERT to compute BERTScore.

Table 2 shows the test F1 scores of R-1, R-2, R-L, and BERTScore (BS) of all models described in the previous section. To the best of the authors' knowledge, there was no other abstractive summarization research using IndoBERT checkpoints (Koto, Rahimi, Lau, et al., 2020; Wilie et al., 2020) with the IndoSum dataset. Therefore, this paper only showed the scores of baseline and extractive models from previous studies. Nevertheless, Koto, Lau and Baldwin (2020) used IndoBERT-LEM in an abstractive summarization task to evaluate their dataset, Liputan6, using the same model as the present study, the BERTSumAbs model. In addition, a BERTSumAbs model with a random encoder and decoder was trained in this paper; however, it generated a sentence with random words for all articles in the test set, thus it was not included in the table. In general, all the models were still underperformed against the Oracle baseline. Nevertheless,

as can be seen, most of the models outperformed the Lead-3 baseline by a large margin. Koto, Rahimi, Lau, et al. (2020) used IndoBERT-LEM for extractive summarization task with the BERTSumExt model and compared it with other BERT checkpoints, such as Multilingual BERT (MBERT) (Devlin et al., 2019) and monolingual Malaysian BERT, MalayBERT. From their experiments, the model built with IndoBERT-LEM had more ROUGE points than the rest. Compared to the BERTSumExt model with IndoBERT-LEM, the proposed abstractive model scores still lagged behind it. It had been predicted as the IndoSum dataset contained more extractive labels so that the extractive models should work better with the dataset.

Table 2

Results for the IndoSum First Fold Test Set.

Model	R-1	R-2	R-L	BS
Baseline				
Oracle*	79.27	72.52	78.82	-
Lead-3*	62.86	54.50	62.10	-
Extractive Models				
NeuralSum* (Cheng & Lapata, 2016)	67.60	61.16	66.86	-
NeuralSum 300 Emb. Size* (Kurniawan & Louvan, 2018)	67.96	61.65	67.24	-
BERTSumExt w/ IndoBERT-LEM* (Koto, Rahimi, Lau, et al., 2020)	69.93	62.86	69.21	-
Investigated Models				
Main				
BERTSumAbs w/ IndoBERT-LEM [1]	**68.80**	**60.86**	**67.97**	**86.01**
BERTSumAbs w/ IndoBERT-NLU [2]	66.32	58.02	65.45	83.44
BERTSumAbs w/ IndoBERT-LEM-5kStep [3]	68.23	60.17	67.40	85.54
BERTSumAbs w/ IndoBERT-NLU-5kStep [4]	64.33	55.98	63.38	82.65
Variants				
BERTSumAbs w/ IndoBERT-NLU-30kEmb [5]	62.70	54.45	61.68	82.59
BERTSumAbs w/ IndoBERT-LEM-50kEmb [6]	68.83	60.79	67.98	85.98
BERTSumAbs w/ IndoBERT-LEM-GPT [7]	**69.20**	**61.35**	**68.36**	**86.22**

R-1, R-2, R-L are ROUGE metrics. BS is BERTScore computed using bert-base-multilingual (layer 9) as suggested in Zhang, Kishore, Wu, et al. (2019). Note that models with * were computed using 5-fold validation of the IndoSum dataset. The bolded scores are the highest in main models and variant models.

For the next part, the two main models were compared using IndoBERT-LEM and IndoBERT-NLU as their encoders as presented in Figure 3. It was pointed out that IndoBERT-LEM outperformed IndoBERT-NLU in all scores. Furthermore, the R-L model with IndoBERT-LEM only improved +0.57 point from 5k steps to 20k steps. Meanwhile, the model with IndoBERT-NLU improved +2.07 point, higher than that of IndoBERT-LEM, indicating that IndoBERT-NLU needed more steps to converge.

Figure 3

Comparison of the Main Models at 5k Steps and 20k Steps.

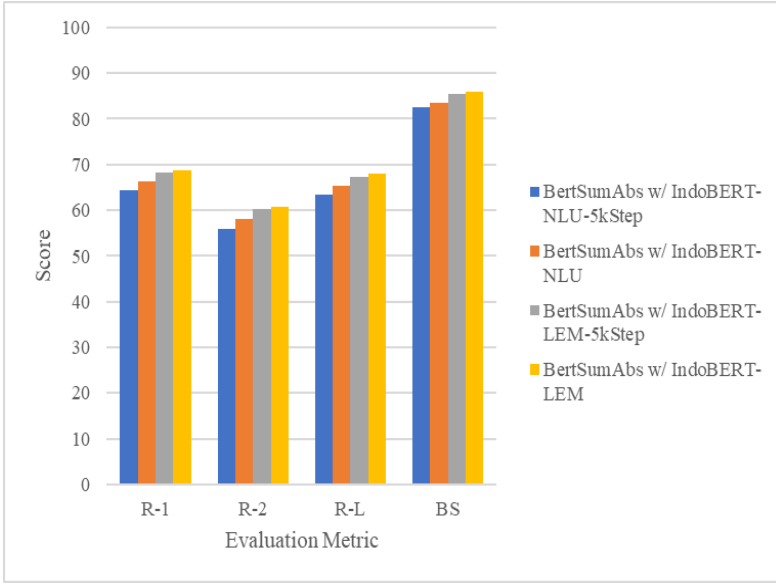


Table 3 shows the generated summary from the main models. An article that had a high abstractive reference summary was chosen. It can be seen that the summary generated was identical to the reference

summary in the beginning and began to differ in the middle to the end of the paragraph. However, the generated summary mostly still followed the facts from the article. The model with IndoBERT-LEM generated "during those three and a half hours" was taken from the article even though it was not supposed to be there. However, it succeeded in referring "Indonesia" to "Its soul". Meanwhile, the model with IndoBERT-NLU generated "spin wind", which as different but still had the same meaning as "windmills" in Indonesian. However, it resulted in more incorrect and repeated words and unneeded random symbols at the end of the summaries. This might be because IndoBERT-NLU had been pre-trained on 128 tokens; therefore, it could hardly handle the text summarization task with 512 tokens dataset.

Table 3

Generated Summary from the Main Models with Highly Abstractive Reference Summary.

Summary Type	Indonesian	English (Translated)
Reference Summary	mohammad hatta, bukan hanya seseorang yang dihormati dan dibanggakan di negaranya sendiri. namanya abadi di negara lain, termasuk di belanda, negara yang pernah menjajah Indonesia. sedari muda, hatta telah mendiami negeri kincir angin untuk meneruskan pendidikannya. hatta menjadi mahasiswa handelshogeschool atau sekolah tinggi bisnis di rotterdam sejak 19 september 1921. hatta tumbuh menjadi pemuda berjiwa nasionalis.	muhammad hatta, is not just someone who is respected and proud in his own country. his name is eternal in other countries, including in the netherlands, a country that once colonized indonesia. since he was young, hatta has inhabited the land of windmills to continue his education. hatta has been a student of handelshogeschool or business college in rotterdam since 19 september 1921. he has grown up to be a young nationalist .
BertSumAbs w/ IndoLEM	mohammad hatta, bukan hanya seseorang yang dihormati dan dibanggakan di negaranya sendiri. namanya abadi di negara lain, termasuk di belanda, negara yang pernah menjajah indonesia. **selama 3 setengah jam itu**, hatta menguliti praktik eksploitasi yang dilakukan rezim	muhammad hatta, is not just someone who is respected and proud in his own country. his name is eternal in other countries, including in the netherlands, a country that once colonized indonesia. **during those 3 and a half hours**, hatta peeled off the

(continued)

Summary Type	Indonesian	English (Translated)
	kolonial di hindia belanda. ia memanfaatkan hasil belajarnya dalam bidang ekonomi - politik untuk melakukan hal tersebut dengan bernas. dalam salah satu bagian pledoi, hatta menulis, " kami percaya atas kekuatan yang ada dalam jiwanya. " (indonesia) .	exploitation practices carried out by the colonial regime in dutch indies. he uses the results of his studies in the field of political economy to do it fully. in one of the pledoi parts, hatta wrote, " we believe in the strength that is in its soul. " (indonesia) .
BertSumAbs w/ IndoBenchmark	mohammad hatta, bukan hanya seseorang yang dihormati dan dibanggakan di negaranya sendiri. namanya abadi di negara lain, termasuk di belanda, negara yang pernah menjajah indonesia. sejak muda, hatta telah mendiami negeri putar angin untuk meneruskan pendidikannya. hatta menjadi mahasiswa handelshoges, belanda**, dan demokrasi kita]**. hingga kini hatta **hatta** kembali melakukan pertemuan, hatta selalu menunjukkan dirinya sebagai sosok yang andal di balik meja perundingan.**). /).**	muhammad hatta, is not just someone who is respected and proud in his own country. his name is eternal in other countries, including in the netherlands, a country that once colonized indonesia. since he was young, hatta has inhabited the country of the spin wind to continue his education. hatta became a student of handelshoges, the netherlands**, and our democracy]**. until now hatta **hatta** has returned to do conferences, hatta has always shown himself as a reliable figure behind the negotiating table.**). /).**

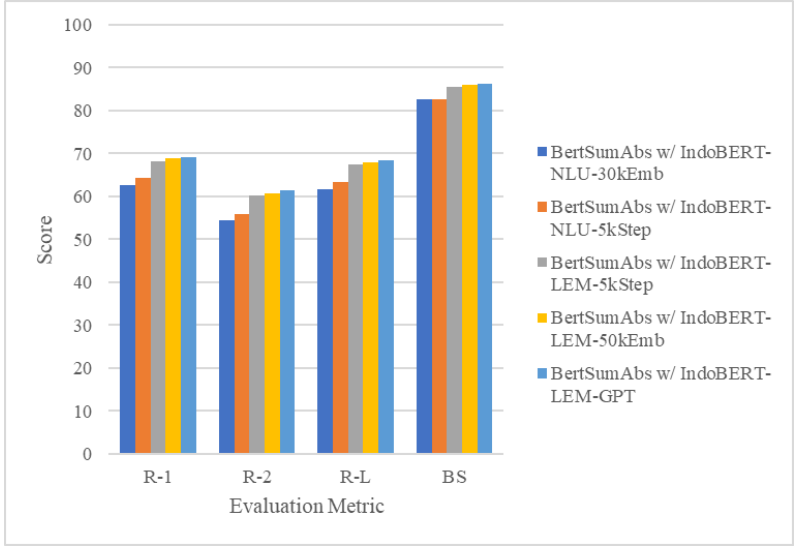
Bear in mind that spaces were given for every symbol, following the generated summaries. Wrong, incomplete, and unneeded words and symbols were highlighted with bold font style. Rephrased words that had the same meaning and served as reference words were highlighted with underlined font style.

Therefore, in the domain of abstractive news summarization, it can be ensured that IndoBERT-LEM was preferred over IndoBERT-NLU as it had been pre-trained longer on a news dataset with 512 tokens. Meanwhile, the latter was pre-trained on a more general dataset with 128 tokens even though with more extensive data than the previous dataset (220M words vs 3.6B words). This finding is consistent with Lewis et al. (2020), whereby the model that was pre-trained specifically on news data performed better in abstractive news summarization than the model that was pre-trained on more general data.

The next experiment was comparing the model variations, whereby different sizes of embedding were set as in Figure 4. The results showed that more embedding could improve the performance of the models. Even the model with IndoBERT-LEM-50kEmb, which was only trained for 5k steps, was on par with the main IndoBERT-LEM model. It was observed that IndoBERT-NLU was pre-trained with 50k embedding size while IndoBERT-LEM was pre-trained with 32k embedding size. However, regardless of that, the models that were fine-tuned with more embedding size outperformed the models with less embedding size. Surprisingly, the last model variation, BERTSumAbs with IndoBERT-LEM-GPT, outperformed all other models even though it was only trained for 5k steps. Nevertheless, it had unstable fine-tuning with the same hyperparameter, whereby the model loss suddenly raised and remained there until the last steps. Therefore, the best checkpoint based on dev set loss was used to compute the scores. It was hypothesized that the learning rate might still be too big for the model. Tinkering with the decoder architecture showed promising results although more research is needed.

Figure 4

Comparison of the Model Variations and Main Models at 5k Steps.



Regarding the use of BERTScore, it revealed a higher score than ROUGE as it computed the similarity between words. However, it was found that the metric as still in line with ROUGE throughout the experiments and provided the same or even less insight than ROUGE. It might be due to the generated summaries that were more extractive. BERTScore should give more insight when the generated summaries were highly abstractive as the words might differ from the reference summaries but still had similar meaning.

CONCLUSION

This paper presented the results of Indonesian abstractive text summarization using the BERTSum and IndoBERT models. Two IndoBERT checkpoints were used, and further findings motivated this research to conduct experimental research on the embedding size and different decoders. The results showed that in the abstractive summarization task, the IndoBERT model, which was trained for more steps with more news data and embedding size, managed to achieve higher ROUGE scores. In addition, the model that used a GPT-like decoder achieved higher scores than the regular model that used a standard transformer decoder. This finding suggests that there are other possibilities for improving the BERTSum model, although more research is needed.

For future studies, research in Indonesian abstractive news summarization may utilize the optimal IndoBERT checkpoint and differentiate the decoder architecture on different datasets to observe another possibility of achieving higher scores. More research is also needed to examine the effectiveness of the BERTScore metric in abstractive text summarization to make better assessment of the text summarization system.

ACKNOWLEDGMENT

This research did not receive a specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

REFERENCES

Adelia, R., Suyanto, S., & Wisesty, U. N. (2019). Indonesian abstractive text summarization using bidirectional gated

recurrent unit. *Procedia Computer Science*, *157*, 581–588. https://doi.org/10.1016/j.procs.2019.09.017

Ba, J. L., Kiros, J. R., & Hinton, G. E. (2016). Layer normalization. *ArXivPreprint ArXiv:1607.06450*. https://arxiv.org/abs/1607.06450

Cai, Z., Lin, N., Ma, C., & Jiang, S. (2019). Indonesian automatic text summarization based on a new clustering method in sentence level. In *Proceedings of the 2019 International Conference on Big Data Engineering* (pp. 30–35). https://doi.org/10.1145/3341620.3341626

Cheng, J., & Lapata, M. (2016). Neural summarization by extracting sentences and words. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 484–494). https://dx.doi.org/10.18653/v1/P16-1046

Christian, H., Agus, M. P., & Suhartono, D. (2016). Single document automatic text summarization using term frequency-inverse document frequency (TF-IDF). *ComTech: Computer, Mathematics and Engineering Applications*, *7*(4), 285–294. https://doi.org/10.21512/comtech.v7i4.3746

Christie, F., & Khodra, M. L. (2016). Multi-document summarization using sentence fusion for Indonesian news articles. In *2016 International Conference On Advanced Informatics: Concepts, Theory and Application (ICAICTA)* (pp. 1–6). IEEE. https://doi.org/10.1109/ICAICTA.2016.7803134

Cui, Y., Che, W., Liu, T., Qin, B., Yang, Z., Wang, S., & Hu, G. (2019). Pre-training with whole word masking for Chinese BERT. *ArXiv Preprint ArXiv:1906.08101*. https://arxiv.org/abs/1906.08101

Devianti, R. S., & Khodra, M. L. (2019). Abstractive summarization using genetic semantic graph for indonesian news articles. In *2019 International Conference of Advanced Informatics: Concepts, Theory and Applications (ICAICTA)* (pp. 1–6). IEEE. https://doi.org/10.1109/ICAICTA.2019.8904361

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *ArXiv Preprint ArXiv:1810.04805*. https://arxiv.org/abs/1810.04805

Garmastewira, G., & Khodra, M. L. (2019). Summarizing Indonesian news articles using Graph Convolutional Network. *Journal of Information and Communication Technology*, *18*(3), 345–365. https://doi.org/10.32890/jict2019.18.3.4675

Grusky, M., Naaman, M., & Artzi, Y. (2018). Newsroom: A Dataset of 1.3 Million Summaries with Diverse Extractive Strategies. *ArXiv Preprint ArXiv:1804.11283*. https://arxiv.org/abs/1804.11283

Halim, K., Novianus Palit, H., & Tjondrowiguno, A. N. (2020). Penerapan *recurrent neural network* untuk pembuatan ringkasan ekstraktif otomatis pada berita berbahasa Indonesia. *Jurnal Infra*, *8*(1), 221–227. http://publication.petra.ac.id/index.php/teknik-informatika/article/view/9797

Hermann, K. M., Kocisky, T., Grefenstette, E., Espeholt, L., Kay, W., Suleyman, M., & Blunsom, P. (2015). Teaching machines to read and comprehend. *Advances in Neural Information Processing Systems*, *28*, 1693–1701. https://ora.ox.ac.uk/objects/uuid:050e7840-1ff3-49db-8d36-e83ed0adf8f7

Hidayat, E. Y., Firdausillah, F., Hastuti, K., Dewi, I. N., & Azhari, A. (2015). Automatic text summarization using latent Drichlet allocation (LDA) for document clustering. *International Journal of Advances in Intelligent Informatics*, *1*(3), 132–139. https://doi.org/10.26555/ijain.v1i3.43

Hoang, A., Bosselut, A., Celikyilmaz, A., & Choi, Y. (2019). Efficient adaptation of pretrained transformers for abstractive summarization. *ArXiv Preprint ArXiv:1906.00138*. https://arxiv.org/abs/1906.00138

Ilyas, R. (2015). Peringkas otomatis dengan ekstraksi informasi untuk kumpulan berita online (Tesis Magister Institut Teknologi Bandung). *Institut Teknologi Bandung, Bandung*.

Koto, F., Lau, J. H., & Baldwin, T. (2020). Liputan6: A large-scale Indonesian dataset for text summarization. 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* (pp. 598–608). https://aclanthology.org/2020.aacl-main.60

Koto, F., Rahimi, A., Lau, J. H., & Baldwin, T. (2020). IndoLEM and IndoBERT: A benchmark dataset and pre-trained language model for Indonesian NLP. In *Proceedings of the 28th International Conference on Computational Linguistics* (pp. 757–770). https://dx.doi.org/10.18653/v1/2020.coling-main.66

Kurniawan, K., & Louvan, S. (2018). Indosum: A new benchmark dataset for Indonesian text summarization. In *2018 International Conference on Asian Language Processing (IALP)* (pp. 215–220). IEEE. https://doi.org/10.1109/IALP.2018.8629109

Lewis, M., Aghajanyan, A., Ghazvininejad, M., Wang, S., Ghosh, G., & Zettlemoyer, L. (2020). Pre-training via paraphrasing. *ArXiv Preprint ArXiv:2006.15020*. https://arxiv.org/abs/2006.15020

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., & Zettlemoyer, L. (2019). BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *ArXiv Preprint ArXiv:1910.13461*. https://doi.org/10.18653/v1/2020.acl-main.703

Lin, C.-Y. (2004). Rouge: A package for automatic evaluation of summaries. *Text Summarization Branches Out*, 74–81. https://www.aclweb.org/anthology/W04-1013.pdf

Liu, Y., & Lapata, M. (2020). Text summarization with pretrained encoders. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (pp. 3721–3731). https://dx.doi.org/10.18653/v1/D19-1387

Martin, L., Muller, B., Suárez, P. J. O., Dupont, Y., Romary, L., de la Clergerie, É. V., Seddah, D., & Sagot, B. (2019). CamemBERT: A tasty French language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 7203–7219). http://dx.doi.org/10.18653/v1/2020.acl-main.645

Najibullah, A. (2015). Indonesian text summarization based on naïve bayes method. In *Proceeding of he International Seminar and Conference on Global Issues* (Vol. 1, No. 1). https://www.publikasiilmiah.unwahas.ac.id/index.php/ISC/article/view/1265/1366

Nallapati, R., Xiang, B., & Zhou, B. (2016a). Sequence-to-sequence RNNs for text summarization. *ICLR 2016*, 4–7.

Nallapati, R., Zhou, B., dos Santos, C., Gulçehre, Ç., & Xiang, B. (2016b). Abstractive text summarization using sequence-to-sequence RNNs and beyond. *ArXiv Preprint ArXiv:1602.06023*. https://arxiv.org/abs/1602.06023

Narayan, S., Cohen, S. B., & Lapata, M. (2018). Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (pp. 1797–1807). https://dx.doi.org/10.18653/v1/D18-1206

Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)* (pp. 2227–2237). https://dx.doi.org/10.18653/v1/N18-1202

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI Blog*, *1(8)*, 9. http://www.persagen.com/files/misc/radford2019language.pdf

Rönnqvist, S., Kanerva, J., Salakoski, T., & Ginter, F. (2019). Is multilingual BERT fluent in language generation? *DL4NLP 2019*, 29. https://ep.liu.se/ecp/163/ecp19163.pdf#page=35

Rothe, S., Narayan, S., & Severyn, A. (2020). Leveraging pre-trained checkpoints for sequence generation tasks. *Transactions of the Association for Computational Linguistics*, *8*, 264–280. https://doi.org/10.1162/tacl_a_00313

Rush, A. M., Chopra, S., & Weston, J. (2015). A neural attention model for abstractive sentence summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing* (pp. 379-389). https://dx.doi.org/10.18653/v1/D15-1044

Savelieva, A., Au-Yeung, B., & Ramani, V. (2020). Abstractive summarization of spoken and written instructions with BERT. *ArXiv Preprint ArXiv:2008.09676.* https://arxiv.org/abs/2008.09676

See, A., Liu, P. J., & Manning, C. D. (2017). Get to the point: Summarization with pointer-generator networks. *ArXiv Preprint ArXiv:1704.04368*. https://arxiv.org/abs/1704.04368

Severina, V., & Khodra, M. L. (2019). Multidocument abstractive summarization using abstract meaning representation for Indonesian language. In *2019 International Conference of Advanced Informatics: Concepts, Theory and Applications (ICAITA)* (pp.1-6). IEEE. https://doi.org/10.1109/ICAICTA.2019.8904449

Shi, T., Keneshloo, Y., Ramakrishnan, N., & Reddy, C. K. (2021). Neural abstractive text summarization with sequence-to-sequence models. *ACM Transactions on Data Science, 2*, *1*(1), 1–37. https://doi.org/10.1145/3419106

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *ArXiv Preprint ArXiv:1706.03762*. https://arxiv.org/abs/1706.03762

Widyassari, A. P., Affandy, A., Noersasongko, E., Fanani, A. Z., Syukur, A., & Basuki, R. S. (2019, July). Literature review of automatic text summarization: Research trend, dataset and method. In *2019 International Conference on Information and Communications Technology (ICOIACT)* (pp. 491–496). IEEE. https://doi.org/10.1109/ICOIACT46704.2019.8938454

Wilie, B., Vincentio, K., Winata, G. I., Cahyawijaya, S., Li, X., Lim, Z. Y., Soleman, S., Mahendra, R., Fung, P., Bahar, S., & Purwarianti, A. (2020). 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* (pp. 843–857). https://www.aclweb.org/anthology/2020.aacl-main.85

Zhang, H., Cai, J., Xu, J., & Wang, J. (2019). Pretraining-based natural language generation for text summarization. In *CoNLL 2019 - 23rd Conference on Computational Natural Language Learning, Proceedings of the Conference* (pp. 789–797). http://dx.doi.org/10.18653/v1/K19-1074

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2019). BERTScore: Evaluating text generation with BERT. *ArXiv Preprint ArXiv:1904.09675*. https://arxiv.org/abs/1904.09675

Zhou, Q., Yang, N., Wei, F., & Zhou, M. (2017). Selective encoding for abstractive sentence summarization. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 1095–1104). https://dx.doi.org/10.18653/v1/P17-1101