International Journal of Engineering Trends and Technology
ISSN: 2231–5381 / https://doi.org/10.14445/22315381/IJETT-V71I11P220

Volume 71 Issue 11, 190-200, November 2023

Original Article

Augmented-Based Indonesian Abstractive Text Summarization using Pre-Trained Model mT5

Andre Setiawan Wijaya[1], Abba Suganda Girsang[2]

[1,2]Computer Science Department, BINUS Graduate Program - Master of Computer Science, Bina Nusantara University, Jakarta, Indonesia.

[1]Corresponding Author : andre.wijaya002@binus.ac.id

Received: 10 July 2023 Revised: 07 September 2023 Accepted: 09 October 2023 Published: 04 November 2023

Abstract *- Nowadays, up-to-date information is endlessly generated by online users; however, sometimes, information on the internet often needs a lot of time to read by readers. Therefore, tools like automatic text summarization are especially important today. Although Indonesian is one of the most used languages in the world, its research in abstractive automatic text summarization is very limited compared to other languages like English and Mandarin. Recently, many pre-trained models for NLP have been developed and are able to generate abstractive automatic text in the English language. Recently, using data augmentation in NLP has also gained a lot of interest; according to research, applying data augmentation in the training set can improve the performance of NLP downstream tasks such as aspect-based sentiment analysis and machine translation. Thus, this research tries to augment the Indonesian news dataset, Liputan6, using the backtranslation method, which will be used to train and fine-tune mT5, mBART, and IndoBART model to generate Indonesian abstractive automatic text summarization task, then compare the result of the summarization and ROUGE score with the models trained and fine-tuned using non-augmented Liputan6 datasets. The result shows that models that are trained with the augmented Liputan6 dataset gained an increase in ROUGE-1, ROUGE-2, and ROUGE-L scores.*

Keywords *- Backtranslation, Data augmentation, Fine-tune pre-trained model, Indonesian abstractive text summarization, Indonesian automatic text summarization.*

1. Introduction

Nowadays, information is scattered very quickly through the internet. The latest information is continuously being generated and uploaded by online users via online media, like online news portals and social media. However, this information is often more than being needed by online users. Sometimes, this information takes a lot of time to read, so online users must select and sort the best data for the specific information they need in the fastest time possible. However, the user has limited time to digest all the available information. Therefore, automatic text summarization is very important in this era, which produces information endlessly [1][2][3].

Automatic text summarization aims to make a shorter summary from the original text version [4]. The summary produced by automatic text summarization must capture the main idea of the document want to summarize so that the user can get the main points without having to read the entire document [5]. There are two types of automatic text summarization: extractive text summarization and abstractive text summarization. Extractive text summarization is where the entire result of the summary consists of words or main idea sentences extracted from the original document. Meanwhile, abstractive automatic text summarization produces summaries that could have new sentences, such as paraphrasing the original document, so that the results of the summaries look like summaries that humans write [6][7].

Research on abstractive automatic text summarization was considered difficult; researchers focused more on conducting research on extractive text summaries [8]. Still, there are abstractive text summarization research can be found, such as Chopra [9] using an encoder-decoder architecture with a recurrent neural network layer inside the encoder and the decoder, and the researcher stated that the proposed method achieved satisfactory ROUGE-1, ROUGE-2, ROUGE-L results on the GIGAWORD, and DUC-2004 dataset.

Chen et al. [10] compared the use of an encoder-decoder architecture with unidirectional GRU and bidirectional GRU in the DUC-2004 and LCSTS (*A Large Scale Chinese Short Text Summarization Dataset*) dataset; researchers found that bidirectional GRU ROUGE results were better than unidirectional GRU.

Besides using neural networks, research using pre-trained models on abstractive text summarization tasks can be found as follows. Saito et al. [11] use the encoder–decoder based on the BERT (Bidirectional Encoder Representation from Transformer) model on the CNN/Daily Mail Dataset to gain a satisfying ROUGE result. Kieuvongngam et al. [12] adopted BERT and GPT-2 models to generate abstractive text summarization on the COVID-19 Open Research Dataset and stated that the generated text summary showed a good level of readability.

Numerous research on abstractive text summarization with foreign language datasets has been carried out. However, interest in research on abstractive text summarization using the Indonesian dataset still gets less attention compared to other foreign languages. Research on Spotted Hyena Optimization with Deep Learning based Automatic Text Summarization can be found by A. Leoraj, and M. Jeyakarthic [13] by using a ABiGRU encoder and unidirectional GRU decoder, and this study shows the result of ROUGE-1 and ROUGE-2, 0.011975, and 0.01199 respectively, witch a machine summary that produces repeated words.

In recent NLP studies, there has been increasing interest in incorporating data augmentation because of its efficient results for adding new data in a limited number of datasets without directly collecting more data, which can result in better performance [14]. Data augmentation research in the field of NLP can be seen in the research of Sayali Hande, and M.A. Potey [15]; in this study, researchers combined the Easy Data Augmentation (EDA) method and backtranslation methods on small Indonesian language sentiment analysis datasets and used the LSTM, Bi- LSTM, and CNN to perform aspect-based sentiment analysis tasks, the results show that EDA combined with backtranslation can improve model performance. Li et al. [15] used the backtranslation and self-learning methods to carry out data augmentation in machine translation tasks, and the results show that these methods can improve performance in IWSLT German-English translation tasks, and IWSLT English-Turkish translation tasks.

Based on the analysis of previous research, it was found that encoder-decoder models such as mT5 and mBART generally outperform model architectures that only use decoders or encoders only [17]. Models with an encoder-decoder architecture, such as mT5, have also succeeded in achieving state-of-the-art results on NLP tasks such as paraphrase identification, natural language inference, and question answering on the XTREME multilingual benchmark [18]. Previous research also found that using the Data Augmentation technique on training datasets can improve the model’s performance. Hence, this study aims to propose an abstractive automatic summarization method on Indonesian language datasets using pre-trained models mT5, mBART, and IndoBART; then, each of these models will be trained with the Liputan6 dataset will be augmented, and the Liputan6 dataset will not be augmented, then the results of each model will be compared using ROUGE evaluation metric.

2. Literature Review

2.1. Natural Language Processing

Natural language processing is a branch of artificial intelligence science and linguistics that deals with text and speech. Natural language processing aims to study the interaction between computers and human language. Techniques in natural language processing are used to perform useful tasks that use human language as input, such as automatic summarization, machine translation, question answering, sentiment analysis, topic segmentation, and speech recognition [19][20][21].

2.2. Automatic Text Summarization

Automatic text summarization is a well-known task in the field of Natural Language Processing (NLP) where the task is to shorten a large amount of information from one or more documents into a concise form by selecting important information and discarding unimportant information from the original document using certain algorithms/method used by machines [22][23]. Automatic text summary can be divided into two according to its techniques, namely, extractive text summarization, where summary results will be obtained by extracting words from the document, and abstractive text summarization, which generates a summary that contains new phrases or can produce words that are not present in the original document [24].

2.3. Data Augmentation

Data augmentation is a method that has been frequently implemented in the field of computer vision, where the augmented data is in the form of images. The augmented image undergoes a transformation process, such as cropping, rotation, flipping, and noise injection, to produce a different image [25]. In the field of Natural Language Processing (NLP), data augmentation adopts a similar concept but is implemented on data in the form of text; text data augmentation can be interpreted as a process of transforming text into new sentences.

Data Augmentation in the field of NLP has recently received increasing interest because it can efficiently increase the amount of dataset, especially data with low language resources [26]. In its application, using too little training dataset can cause model overfitting [27]. Meanwhile, collecting large amounts of datasets requires a great deal of effort and requires a lot of time [28]. One of the data augmentation methods implemented in this research is backtranslation. Backtranslation is a method whereby a text is translated from one language to another and then translated back into the original language; for example, an English text is translated into Chinese and then translated again into English. In general, by using backtranslation, there are new words that do not exist in the original text and still maintain the semantics of the original sentence so that data can be reproduced quickly [29][25].

2.4. mT5

T5 (Text-to-Text Transfer Transformer) is a pre-trained language model whose main function is using a unified 'text-to-text' format for all text-based NLP problems. The T5 approach is particularly common for generative tasks, such as machine translation or abstractive summarization, where the task format requires the model to generate text dependent on the given input [30]. T5 uses an encoder-decoder transformer architecture based on the proposed method by Vaswani et al.[31]. T5 was previously trained on the "span-corruption" objective in masked language modeling, where successive ranges of input tokens are replaced with mask tokens and the model is trained to reconstruct the masked token. In their research, Xue et al. [30] wanted to create a massive multilingual model that follows the T5 model as closely as possible, called mT5. The mT5 model is trained using the mC4 dataset, and the mC4 dataset is a common crawl web scrape dataset consisting of 101 languages (including Indonesian) as a form of multilingual dataset from the C4 dataset, which only contains the English language.

2.5. mBART

mBART is a multilingual language model that uses a denoising autoencoder to train sequence-to-sequence tasks; mBART itself is based on the BART language model [32]. BART is trained by destructing the text using an arbitrary noising function, and the model learns to reconstruct the original text. BART uses a Transformer-based neural machine translation architecture; this model can be seen as a generalization of BERT (using a two-way encoder) and GPT (using a left-to-right decoder) [33]. The mBART language model works by being trained once for multiple languages, thus providing a set of parameters that can be fine-tuned for various language pairs on a supervised or unsupervised basis without having to concentrate on a particular task.

2.6. IndoBART

IndoBART is a language model that follows an encoder-decoder architecture; the IndoBART model is based on and adopts the implementation of the mBART language model. IndoBART is trained in 3 languages: Indonesian, Javanese, and Sundanese. IndoBART consists of 6 encoder layers 6 decoder layers, with 12 heads, 768 embedding sizes, and 132M parameter sizes [34].

2.7. ROUGE (Recall-Oriented Understudy for Gisting Evaluation)

Lin [35] stated that ROUGE (Recall-Oriented Understudy for Gisting Evaluation) serves to measure how well the quality of a summary is produced by a machine or computer (computer-generated). ROUGE works to measure the quality of a summary by counting the words that overlap between the words in the summary produced by machines and the words in the summary produced by humans (ground truth / gold summary). For example, if the result of the machine-generated summary is "the cat was found under the bed" and the human summary is "the cat was under the bed", ROUGE will count overlapping words to measure the quality of the machine-generated summary. Equation 1 shows the calculation of ROUGE values in general.

$$ROUGE = \frac{TotalOverlappingWords}{TotalWordsGeneratedInMachineSummary} \tag{1}$$

2.8. Related Works

In 2016, Chopra et al. [9] proposed an encoder with an attention mechanism and convolution network to compute the value of the sentence input and a decoder with a recurrent network to capture the important words used in the generated summary. The datasets used in this research are Gigaword and DUC-2004 datasets. This study obtained the results of ROUGE-1, ROUGE-2, and ROUGE-L, respectively, 33.78, 15.97, and 31.15 on the Gigaword dataset, and the results of ROUGE-1, ROUGE-2, ROUGE-L, respectively 28.97, 8.26, and 24.06 in the DUC-2004 dataset.

Wang [36] conducted research in abstractive automatic text summarization on the LCSTS dataset by using an encoder-decoder architecture with a recurrent neural network and adding an attention mechanism to the encoder layer. This study uses the ROUGE evaluation model to calculate how good the proposed method is. This study's ROUGE-1, ROUGE-2, and ROUGE-L values were 38.70%, 26.50% and 37.65%.

Aksenov et al. [37] utilize an encoder and decoder based on a BERT (Bidirectional Encoder Representation from Transformer) pre-trained model to produce abstractive text summarization on the CNN/DailyMail and SwissText dataset; this proposed method shows that the results of ROUGE-1, ROUGE-2, ROUGE-L were 31.51, 14.1, and 29.77 respectively. Baykara & Güngör [18] use various pre-trained models such as BERT, mBART, and mT5 to perform abstractive automatic text summarization on Turkish language datasets, TR-News, MLSum (TR). The results obtained by researchers shows that mT5 has higher ROUGE-1 and ROUGE-2 score in TR-News and MLSum (TR) datasets compared to other pre-trained models.

In 2021, research conducted by Hasan et al. [38] proposed a large-scale multilingual dataset for abstractive summarization research, which includes 44 languages. In addition to proposing the datasets, the researcher also performs automatic abstractive summarization using a pre-trained model; the pre-trained model used is mT5. The top 10 languages that received the best ROGUE were then obtained from this research, and these languages are English, Chinese, Hindi, Spanish, French, Arabic, Bengali, Russian, Portugese, and Indonesian. Indonesian's ROUGE-1, ROUGE-2, and ROUGE-L values are 36.17, 16.70 and 30.50, respectively.

Adelia et al. [13] conducted research on Indonesian abstractive text summarization using Bidirectional Gated Recurrent Units (BiGRU) as encoder and unidirectional Gated Recurrent Units (GRU) with attention model as its decoder. The dataset used in this research was a private Indonesian language journal dataset, which the researchers themselves collected. This study obtained ROUGE-1 and ROUGE-2 scores of 0.11975 and 0.01199, respectively. The researcher stated that the machine summary still had repetitive words and a lack of semantics. Saputra et al. [39] also conducted a study on Indonesian abstractive text summarization using 3 layers of long short-term memory (LSTM) as the encoder and decoder model. The research obtains F1 ROUGE-1, ROUGE-1 precision, and ROUGE-1 recall scores of 0.13826, 0.14211, and 0.13595, respectively.

3. Methodology

The research methodology in this study can be divided into 5 sub-sections: dataset, data construction, data augmentation, fine-tuning pre-trained model, and Evaluation. The workflow of this research methodology can be seen in Figure 1.

3.1. Dataset

The dataset will be used in this study is the Liputan6 dataset [41]. The liputan6 dataset is a dataset sourced from Liputan6.com, an online news portal; within a span of 10 years, the news content covers various topics and events that occurred, especially in Indonesia, from October 2000 to October 2010, the liputan6 dataset contains approximately 215 thousand pairs news and gold summaries.

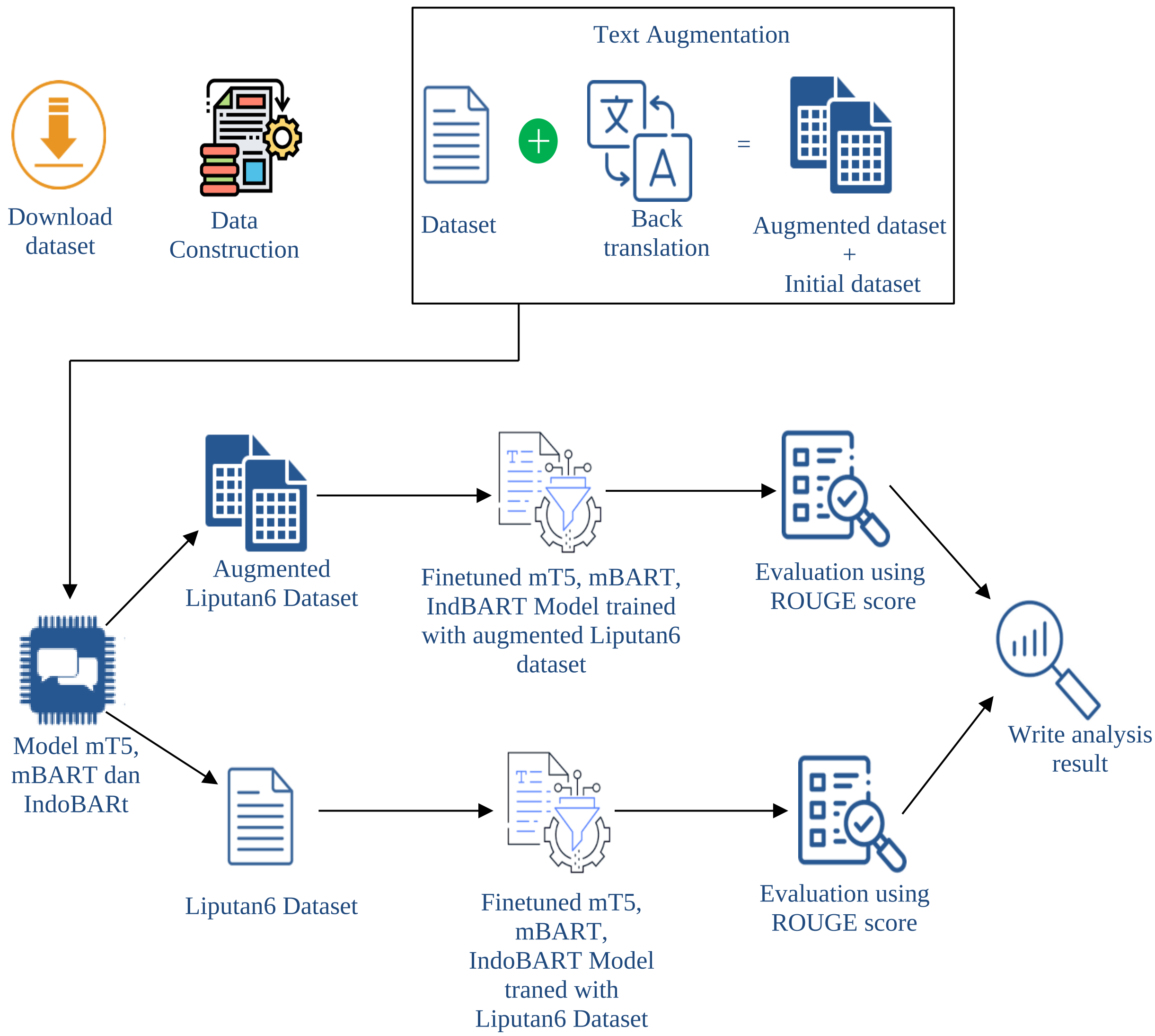

Fig. 1 Research methodology

The Liputan6 dataset is an open-source dataset, but this dataset can only be used for non-commercial activities such as academic research. Due to limited computational resources in this study, the Liputan6 dataset that is going to be used will be cut into 10,000 pairs of articles and gold summaries for the training set, 1,250 pairs of articles and gold summaries for the validation set, and 1,250 pairs of articles and gold summaries for the test. Set, this truncated Liputan6 dataset also serves the purpose of becoming a small dataset (low resource dataset) for later comparison with the performance of models trained using the augmented Liputan6 dataset. The augmented Liputan6 dataset contains 20,000 pairs of articles and gold summaries (an additional 10,000 pairs using the back translation method) with the same amount of validation and testing set.

3.2. Data Construction

The data used is a dataset from Liputan6 [41]. The Liputan6 dataset contains news articles from a period of 10 years. The number of documents in the Liputan6 dataset is 215,827 pairs of documents, along with its gold summary. The original form of the dataset is in the form of words cut into pieces, so the dataset needs to be constructed into complete sentences. Therefore, it can be tokenized properly by the model, for example, as:

Liputan6 sample
[["Liputan6", ".", "com", ",", "Kuta", ":", "Masyarakat", "Bali", "baru-baru", "ini", "menggelar", "lomba", "mendirikan", "telur", "di", "atas", "piring", "di", "Pantai", "Kuta", "."], ["Susah", "memang", ",", "tapi", "lomba", "ini", "membuat", "banyak", "orang", "penasaran", "."], ["Warga", "setempat", "dan", "turis", "lokal", "maupun", "mancanegara", "banyak", "mencobanya", "."], ["Akhirnya", ",", "ada", "wisatawan", "asing", "yang", "berhasil", ".", "Lomba", "mendirikan", "telur", "ini", "adalah", "menjadi", "patokan", "untuk", "memulai", "pelaksanaan", "ritual", "tahunan", "umat", "Budha", "yang", "disebut", "Bak", "Cang", "."], ["Ritual", "ini", "digelar", "setiap", "tanggal", "lima", "bulan", "lima", "pada", "tahun", "Cina", "."], ["Tujuannya", "untuk", "menghormati", "negarawan", "sekaligus", "sastrawan", "Cina", "Ciek", "Yen", "yang", "hidup", "pada", "Dinasti", "Chi", "pada", "278", "sebelum", "masehi", "."], ["Cang", "juga", "bertujuan", "untuk", "memohon", "keselamatan", "dan", "pembersihan", "diri", "dengan", "cara", "mandi", "di", "laut", ".", "(", "BOG/Putu", "Setiawan", ")", "."]]
Liputan6 sample after data construction
Liputan6.Com Kuta: Masyarakat Bali baru-baru ini menggelar lomba mendirikan telur di atas piring di Pantai Kuta. Susah memang, tapi lomba ini membuat banyak orang penasaran. Warga setempat dan turis lokal maupun mancanegara banyak mencobanya. Akhirnya, ada wisatawan asing yang berhasil. Lomba mendirikan telur ini adalah menjadi patokan untuk memulai pelaksanaan ritual tahunan umat Budha yang disebut Bak Cang. Ritual ini digelar setiap tanggal lima bulan lima pada tahun Cina. Tujuannya untuk menghormati negarawan sekaligu sastrawan Cina Ciek Yen yang hidup pada Dinasti Chi pada 278 sebelum masehi. Cang juga bertujuan unutk memohon keselamatan dan pembersihan diri dengan cara mandi di laut. (BOG/Puitu Setiawan,).
Liputan6 sample (English translation)
Liputan6.Com Kuta: Balinese people recently held a competition to place eggs on a plate on Kuta Beach. It is difficult, but this competition makes a lot of people curious. Many local residents and foreign tourists try to do it. Finally, some foreign tourists are successful. This egg standing competition is a benchmark for starting the practice of the annual Buddhist ritual called Bak Cang. This ritual is held every fifth day of the fifth month of the Chinese year. The aim is to honor the Chinese statesman and writer Ciek Yen, who lived during the Chi Dynasty in 278 BC. Cang also aims to seek safety and cleanse oneself by bathing in the sea. (BOG/Puitu Setiawan,).

Furthermore, the original Liputan6 datasets are in the form of JSON files (1 JSON file containing 1 pair of articles and a gold summary). Thus, the Liputan6 dataset in this study will be loaded and converted into a CSV file format to make processing easier.

3.3. Data Augmentation

The Liputan6 dataset is going to be augmented using the back-translation method, where sentences will be translated from the source language to a foreign language and translated back to the original language.

After carrying out the back-translation process, there will generally be differences between the resulting and initial sentences. In this study, the dataset will be translated from Indonesia to English and English to Indonesia. (Bolded words are new words generated from the back-translation method.)

Dataset sample before backtranslation
Liputan6.Com Kuta: Masyarakat Bali baru-baru ini menggelar lomba mendirikan telur di atas piring di Pantai Kuta. Susah memang, tapi lomba ini membuat banyak orang penasaran. Warga setempat dan turis lokal maupun mancanegara banyak mencobanya. Akhirnya, ada wisatawan asing yang berhasil. Lomba mendirikan telur ini adalah menjadi patokan untuk memulai pelaksanaan ritual tahunan umat Budha yang disebut Bak Cang. Ritual ini digelar setiap tanggal lima bulan lima pada tahun Cina. Tujuannya untuk menghormati negarawan sekaligus sastrawan Cina Ciek Yen yang hidup pada Dinasti Chi pada 278 sebelum masehi. Cang juga bertujuan unutk memohon keselamatan dan pembersihan diri dengan cara mandi di laut. (BOG/Puitu Setiawan,).

Dataset sample after back-translation
LIPUTAN6.com Kuta: **Orang - orang** Bali baru -baru ini **mengadakan kompetisi** untuk **membangun** telur di atas piring di pantai Kuta. **Sulit** memang, **tetapi kompetisi** ini membuat banyak orang penasaran. **Penduduk** setempat dan **wisatawan** lokal **dan asing** banyak mencoba. Akhirnya, ada wisatawan asing yang berhasil. **Kompetisi** yang **membangun** telur ini **merupakan tolok ukur** untuk memulai ritual tahunan **orang -orang** Buddha yang disebut Bak Cang. Ritual ini **diadakan** setiap tanggal kelima bulan kelima **Tiongkok**. Tujuannya **adalah** untuk menghormati negarawan sastra Cina Ciek Yen yang **tinggal** di Dinasti Chi pada 278 SM. CANG juga bertujuan untuk **meminta** keselamatan dan diri sendiri dengan mandi di laut. (Bog/Puitu Setiawan,).

3.4. Finetuning mT5, mBART, and IndoBART

The pre-trained models that will be used are IndoBART, mT5, and IndoBART, which can be loaded using the NLP library provided by HuggingFace. Fine-tuning that will be carried out in this study begins with adjusting the hyperparameter values used in IndoBART, mT5, and mBART. These hyperparameter values include maximum sequence length, learning rate, batch size, epoch, and optimizer. After setting the hyperparameter values, the next step is to carry out the training, validation, and testing processes on the IndoBART, mT5, and mBART models using the Liputan6 datasets.

3.5. Evaluation

The performance from each pre-trained model will be evaluated using the ROUGE metric. ROUGE-N evaluates the summary by calculating the recall based on the n-gram comparison overlapping between the gold and system summaries. If n = 1, then ROUGE-1 calculates word matches unigram. If n=2, then ROUGE-2 calculates bigram match words. Equation 2 shows how to calculate ROUGE-N, where x is the number of n-grams that overlap between the gold summary and the system summary, and y is the number of words generated by the system summary. ROUGE-L evaluates the summary based on the longest common subsequence or the longest series of words that are the same between the gold and system summaries. ROUGE-L evaluation can be calculated using equation 3, where LCS is the longest series of words overlapping the gold summary and system summary, while y is the number of words produced by the system summary. ROUGE-1, ROUGE-2 and ROUGE-L will be used in this study.

$$ROUGE - (N) = \frac{X}{y} \tag{2}$$

$$ROUGE - L = \frac{LCS}{y} \tag{3}$$

4. Results & Discussions

Several hyperparameter fine-tuning settings have been used for each pre-trained language model (IndoBART, mT5-Base, and mBART-Large-50). The best ROUGE model results are obtained with the hyperparameter configurations listed in Table 2.

Table 2 shows that the mT5-base language model that has been fine-tuned and trained with an augmented dataset shows the best ROUGE-1 and ROUGE-2 performance, with a ROUGE-1 value of 0.3856 and ROUGE-2 value of 0.2329, while the best ROUGE-L performance is obtained by the mBART-Large-50 model trained using augmented liputan6 dataset with a value of 0.3331. Results also show that the model trained with the augmented liputan6 dataset has increased ROUGE-1, ROUGE-2, and ROUGE-L scores.

A sample of the abstractive summary generated of each model will be taken and analyzed. An abstractive summary of each model will be shown in Table 4,5,6,7.

Table 1. Total Liputan6 dataset pairs used

	Training Set	Validation Set	Testing Set
Liputan6	10.000	1.250	1.250
Liputan6-Augmented	20.000	1.250	1.250

Table 2. Hyperparameter finetuning of each model

Model	Dataset	Learning Rate	Epoch	Batch Size	Duration	Optimization	Scheduler
mT5-Base	Liputan6	5.00e-4	5	8	1h 50m	Adafactor	Linear
	Liputan6- Augmented	5.00e-4	5	8	2h 45m	Adafactor	Linear
mBART-Large-50	Liputan6	4.00e-5	5	8	2h 00m	AdamW	Linear
	Liputan6-Augmented	4.00e-5	5	8	3h	AdamW	Linear
IndoBART	Liputan6	3.00e-5	12	8	55m	AdamW	Polynomial
	Liputan6-Augmented	3.00e-5	10	8	1h 20m	AdamW	Polynomial

Table 3. ROUGE score of each Pre-Trained model

Model	Dataset	ROUGE-1	ROUGE-2	ROUGE-L
mT5-Base	Liputan6	0.3430	0.1720	0.2730
	Liputan6-Augmented	**0.3856**	**0.2329**	0.3263
mBART-Large-50	Liputan6	0.3481	0.1711	0.3243
	Liputan6-Augmented	0.3814	0.2151	**0.3331**
IndoBART	Liputan6	0.3153	0.1549	0.2723
	Liputan6-Augmented	0.3325	0.1688	0.2821

Table 4. Input article and its gold summary

Input sample article from Liputan6 dataset	
Input Article	Liputan6. Com, Jakarta: Indonesia Corruption Watch (ICW) mengadakan syukuran atas pemberhentian Marzuki Darusman dari jabatan Jaksa Agung , Kamis (7/6) . Soalnya, menurut Ketua ICW Teten Masduki, Marzuki hanya memanfaatkan Kejaksaan Agung sebagai alat politik demi kepentingan kekuasaan. Hal itu diutarakan Teten di sela-sela acara syukuran ICW tersebut di Jakarta. Teten menjelaskan, banyak kasus-kasus besar di Kejagung yang tidak berhasil dituntaskan dalam kepemimpinan Marzuki. Bahkan, Marzuki lebih mementingkan kekuasaan daripada penegakan hukum. Teten mencontohkan, kasus penghentian penyidikan terhadap Presiden Gus Dur dalam Buloggate karena tidak ada bukti kuat. Karena itulah, ICW sangat gembira dengan pergantian jabatan Jaksa Agung dari Marzuki kepada Baharudin Lopa. (HFS/Lita Hariyani dan Ari Trisna).
Input Article (English Translation)	Liputan6.Com, Jakarta: Indonesia Corruption Watch (ICW) held thanksgiving for the dismissal of Marzuki Darusman from the position of Attorney General Thursday (7/6). The problem is, according to ICW Chairman Teten Masduki, Marzuki only uses the Attorney General's Office as a political tool for the sake of power. Teten stated this on the sidelines of the ICW Thanksgiving event in Jakarta. Teten explained that many major cases at the Attorney General's Office were not successfully resolved under Marzuki's leadership. In fact, Marzuki is more concerned with power than law enforcement. Teten gave an example, the case of stopping the investigation into President Gus Dur in Buloggate because there was no strong evidence. For this reason, ICW is very happy with the change of position of Attorney General from Marzuki to Baharudin Lopa. (HFS/Lita Hariyani and Ari Trisna).
Gold Summary	ICW mengadakan syukuran atas pemberhentian Marzuki Darusman. Sebab, banyak kasus-kasus besar di Kejagung yang tidak berhasil dituntaskan dalam kepemimpinan Marzuki.
Gold Summary (English Translation)	ICW held thanksgiving for the dismissal of Marzuki Darusman. This is because many major cases at the Attorney General's Office were not successfully resolved under Marzuki's leadership.

4.1. Summary Generated by mT5-Base Model

Based on the summary generated by the model in Table 5 (Bolded words are newly generated words, and the underlined phrase is the model's attempt to paraphrase), an mT5-base model trained using the liputan6 dataset can change vocabulary from "*pemberhentian*" to "*pergantian*", the model can also shorten "*kejaksaan agung*" into "*kejagung*". The model can also generate new words like "*mengatakan*" not present in the original article.

Meanwhile, an mT5-base model trained using an augmented liputan6 dataset can refer to "*Indonesian Watch Corruption*" as an "*ICW*" model can also produce new words such as "*menggelar*" and "*sejumlah*", which is not present in the original article, model can also replace word such as "*pemberhentian*" to "*pergantian*". Model is also able to input the last word from the original article, "*kepada Baharudin Lopa,*" into the first sentence of the generated summary. The model can also put phrases such as "*menurut Teten Masduki*" into the last sentence of the summary, which the phrase can be seen in the middle of the original article.

4.2. Summary Generated by mBART-Large-50 Model

Based on the summary generated by the mBART in Table 6 (Bolded words are newly generated words, and the underlined phrase is the model's attempt to paraphrase), the mBART-Large-50 model trained using Liputan6 can change the word "*Menurut Ketua ICW Teten Masduki*" to "*Menurut Ketua Indonesian Corruption Watch Teten Masduki*", the model can also generate new word such as the word "*terkesan*" which is not in the original article.

Table 5. Summary generated by mT5-Base model

Input Article	Liputan6. Com, Jakarta: Indonesia Corruption Watch (ICW) mengadakan syukuran atas pemberhentian Marzuki Darusman dari jabatan Jaksa Agung , Kamis (7/6) . Soalnya, menurut Ketua ICW Teten Masduki, Marzuki hanya memanfaatkan Kejaksaan Agung sebagai alat politik demi kepentingan kekuasaan. Hal itu diutarakan Teten di sela-sela acara syukuran ICW tersebut di Jakarta. Teten menjelaskan, banyak kasus-kasus besar di Kejagung yang tidak berhasil dituntaskan dalam kepemimpinan Marzuki. Bahkan, Marzuki lebih mementingkan kekuasaan daripada penegakan hukum. Teten mencontohkan, kasus penghentian penyidikan terhadap Presiden Gus Dur dalam Buloggate karena tidak ada bukti kuat. Karena itulah, ICW sangat gembira dengan pergantian jabatan Jaksa Agung dari Marzuki kepada Baharudin Lopa. (HFS/Lita Hariyani dan Ari Trisna).
Generated summary by mT5-Base Model	
mT5-Base (Liputan6)	Indonesia Corruption Watch mengadakan syukuran atas pergantian Marzuki Darusman dari jabatan Jaksa Agung. Marzuki hanya memanfaatkan Kejagung sebagai alat politik. Ketua ICW Teten Masduki mengatakan, kasus-kasus besar di Kejagung yang tidak berhasil dituntaskan dalam kepemimpinan Marzuki.
mT5-Base (Liputan6) (English translation)	Indonesia Corruption Watch held a thanksgiving for replacing Marzuki Darusman from the position of Attorney General. Marzuki only uses the Attorney General's Office as a political tool. ICW Chair Teten Masduki said that major cases at the Attorney General's Office had not been resolved under Marzuki's leadership.
mT5-Base (Liputan6-Augmented)	ICW menggelar syukuran atas pergantian Marzuki Darusman dari jabatan Jaksa Agung kepada Baharudin Lopa. Marzuki Darusman hanya memanfaatkan Kejaksaan Agung sebagai alat politik demi kepentingan kekuasaan. Sejumlah kasus-kasus besar di Kejagung yang tidak berhasil dituntaskan dalam kepemimpinan Marzuki menurut Teten Masduki.
mT5-Base (Liputan6-Augmented) (English Translation)	ICW held a thanksgiving for replacing Marzuki Darusman from the position of Attorney General to Baharudin Lopa. Marzuki Darusman only uses the Attorney General's Office as a political tool for the sake of power. According to Teten Masduki, a number of major cases at the Attorney General's Office were not resolved under Marzuki's leadership.

Table 6. Summary generated by mBART-Large-50

Input Article	Liputan6. Com, Jakarta: Indonesia Corruption Watch (ICW) mengadakan syukuran atas pemberhentian Marzuki Darusman dari jabatan Jaksa Agung , Kamis (7/6) . Soalnya, menurut Ketua ICW Teten Masduki, Marzuki hanya memanfaatkan Kejaksaan Agung sebagai alat politik demi kepentingan kekuasaan. Hal itu diutarakan Teten di sela-sela acara syukuran ICW tersebut di Jakarta. Teten menjelaskan, banyak kasus-kasus besar di Kejagung yang tidak berhasil dituntaskan dalam kepemimpinan Marzuki. Bahkan, Marzuki lebih mementingkan kekuasaan daripada penegakan hukum. Teten mencontohkan, kasus penghentian penyidikan terhadap Presiden Gus Dur dalam Buloggate karena tidak ada bukti kuat. Karena itulah, ICW sangat gembira dengan pergantian jabatan Jaksa Agung dari Marzuki kepada Baharudin Lopa. (HFS/Lita Hariyani dan Ari Trisna).
Generated model by mBART-Large-50 model	
mBART-Large-50 (Liputan6)	Menurut Ketua Indonesia Corruption Watch Teten Masduki, banyak kasus-kasus besar di Kejagung yang tidak berhasil dituntaskan dalam kepemimpinan Marzuki sehingga dia terkesan lebih mementingkan kekuasaan demi penegakan hukum
mBART-Large-50 (Liputan6) (English Translation)	According to the Chairman of Indonesia Corruption Watch, Teten Masduki, many major cases at the Attorney General's Office were not successfully resolved under Marzuki's leadership, so he gave the impression that he was more concerned with power for the sake of law enforcement.
mBART-Large-50 (Liputan6-Augmented)	Ketua ICW Teten Masduki menilai Marzuki hanya memanfaatkan Kejaksaan Agung sebagai alat politik demi kepentingan kekuasaan. Banyak kasus besar di Kejagung yang tidak berhasil dituntaskan dalam kepemimpinannya karena dia lebih mementingkan kekuasaan daripada penegakan hukum.
mBART-Large-50 (Liputan6-Augmented) (English Translation)	ICW Chairman Teten Masduki believes that Marzuki is only using the Attorney General's Office as a political tool for the sake of power. Many major cases at the Attorney General's Office were not successfully resolved under his leadership because he was more concerned with power than law enforcement.

Table 7. Summary generated by IndoBART

Input Article	Liputan6. Com, Jakarta: Indonesia Corruption Watch (ICW) mengadakan syukuran atas pemberhentian Marzuki Darusman dari jabatan Jaksa Agung , Kamis (7/6) . Soalnya, menurut Ketua ICW Teten Masduki, Marzuki hanya memanfaatkan Kejaksaan Agung sebagai alat politik demi kepentingan kekuasaan. Hal itu diutarakan Teten di sela-sela acara syukuran ICW tersebut di Jakarta. Teten menjelaskan, banyak kasus-kasus besar di Kejagung yang tidak berhasil dituntaskan dalam kepemimpinan Marzuki. Bahkan, Marzuki lebih mementingkan kekuasaan daripada penegakan hukum. Teten mencontohkan, kasus penghentian penyidikan terhadap Presiden Gus Dur dalam Buloggate karena tidak ada bukti kuat. Karena itulah, ICW sangat gembira dengan pergantian jabatan Jaksa Agung dari Marzuki kepada Baharudin Lopa. (HFS/Lita Hariyani dan Ari Trisna).
Generated model by IndoBART model	
IndoBART (Liputan6)	icw teten masduki mengadakan syukuran atas pemberhentian Marzuki Darusman.
IndoBART (Liputan6) (English Translation)	ICW Teten Masduki held thanksgiving for the dismissal of Marzuki Darusman.
IndoBART (Liputan6-Augmented)	indonesia corruption watch (icw) mengadakan syukuran atas pemberhentian Marzuki Darusman.
IndoBART (Liputan6-Augmented) (English Translation)	Indonesia Corruption Watch (ICW) held a thanksgiving for the dismissal of Marzuki Darusman.

Finetuned mBART-Large-50 model trained using augmented liputan6 dataset. The model can generate new words that are not found in the original article, such as “menilai”. In addition, this model can also make words that refer, such as the word “kepemimpinannya” which refers to “kepemimpinan Marzuki Darusman”. This model can also produce causal conjunctions, the model can produce causal sentences “karena dia lebih mementingkan kekuasaan daripada penegakan hukum” because of the phrase “Banyak kasus besar di Kejagung yang tidak berhasil ditutaskan”. However, both models failed to capture the main idea from the sentence of “Indonesia Corruption Watch mengadakan syukuran atas pemberhentian Marzuki Darusman dari jabatan Jaksa Agung”.

4.3. Summary Generated IndoBART Model

According to the summary produced by the IndoBART model in table 7, IndoBART trained using Liputan6 and augmented Liputan6 produces summary that is extractive, where these two models take the main sentence from the input article “Indonesia Corruption Watch (ICW) mengadakan syukuran atas pemberhentian Marzuki Darusman dari Jabatan Jaksa Agung”, both models fail to capture the main idea as why Marzuki Darusman was dismissed from the position of Attorney General which is contained in the sentence of “Teten menjelaskan, banyak kasus-kasus besar di Kejagung yang tidak berhasi dituntaskan dalam kepemimpinan Marzuki.

Bahkan, Marzuki lebih mementingkan kekuasaan daripada penegakan hukum”. The difference between these two models is IndoBART (Liputan6) model can write down the name “Teten Masduki” but cannot explain that Teten Masduki is the chairman of Indonesian Corruption Watch; the model immediately generates the sentence “*icw teten masduki*.” Meanwhile, the IndoBART model trained using the liputan6 augmented dataset does not write down the name of the head of ICW, but it makes the generated summary not too awkward to read. This lack of summarization performance by the IndoBART pre-trained model might be because of the relatively small dataset used to train the model.

5. Conclusion

This paper proposed the Indonesian abstractive text summarization using fine-tuned pre-trained models and applying a data augmentation method, namely back translation in the small-sized liputan6 training dataset; the aim is to increase the dataset efficiently and effectively, thus improving the ROUGE summarization score.

Based on the results obtained in this experiment, models trained using the augmented Liputan6 dataset can provide better ROUGE-1, ROUGE-2, and ROUGE-L scores, thus increasing the accuracy of generated summary, looking at the results of the summary produced by each model, models that are trained using the augmented Liputan6 dataset can also generate more words, the models also have the ability to generate new words which could not be found in the original article, models could also paraphrase the generated summary, and still able to capture the main idea from the original article.

Future Works

Suggestions that can be given for further research such as fine-tuning various pre-trained models and the Indonesian dataset to get a varied evaluation and summary results, performing hyperparameter search tuning by using combinations of many hyperparameters in one training run so that the fine-tuning process is more effective, resulting in maximum evaluation results, researchers can also categorize liputan6 dataset to compare the performance by its category, for example comparing results between sports news category and entertainment news category. In this conducted study, IndoBART models seem to need more training datasets to improve their generated summary; therefore, feeding more training datasets for the IndoBART model can result in higher ROUGE scores.

References

[1] Abirami Rajasekaran, and R. Varalakshmi, "Review on Automatic Text Summarization," *International Journal of Engineering and Technology*, vol. 7, no. 2.33, pp. 456-460, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[2] Annapurna P. Patil et al., "Automatic Text Summarizer," *2014 International Conference on Advances in Computing, Communications, and Informatics (ICACCI)*, pp. 1530-1534, 2014. [CrossRef] [Google Scholar] [Publisher Link]

[3] Aarti Patil et al., "Automatic Text Summarization," *International Journal of Computer Applications*, vol. 109, no. 17, pp. 1-2, 2015. [CrossRef] [Google Scholar] [Publisher Link]

[4] Yogan Jaya Kumar et al., "A Review on Automatic Text Summarization Approaches," *Journal of Computer Science*, vol. 12, no. 4, pp. 178-190, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[5] Wafaa S. El-Kassas et al., "Automatic Text Summarization: A Comprehensive Survey," *Expert Systems with Applications*, vol. 165, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[6] M. Firoz Mridha et al., "A Survey of Automatic Text Summarization: Progress, Process and Challenges," *IEEE Access*, vol. 9, pp. 156043-156070, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[7] Ahmad T. Al-Taani, "Automatic Text Summarization Approaches," *2017 International Conference on Infocom Technologies and Unmanned Systems (Trends and Future Directions) (ICTUS)*, pp. 93-94, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[8] Mahak Gambhir, and Vishal Gupta, "Recent Automatic Text Summarization Techniques: A Survey," *Artificial Intelligence Review*, vol. 47, no. 1, pp. 1-66, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[9] Sumit Chopra, Michael Auli, and Alexander M. Rush, "Abstractive Sentence Summarization with Attentive Recurrent Neural Networks," *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 93-98, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[10] Qian Chen et al., "Distraction-Based Neural Networks for Document Summarization," *arXiv*, pp. 1-8, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[11] Itsumi Saito et al., "Length-Controllable Abstractive Summarization by Guiding with Summary Prototype," *arXiv*, pp. 1-8, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[12] Virapat Kieuvongngam, Bowen Tan, and Yiming Niu, "Automatic Text Summarization of Covid-19 Medical Research Articles Using Bert and GPT-2," *arXiv*, pp. 1-13, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[13] A. Leoraj, and M. Jeyakarthic, "Spotted Hyena Optimization with Deep Learning-Based Automatic Text Document Summarization Model," *SSRG International Journal of Electrical and Electronics Engineering*, vol. 10, no. 5, pp. 153-164, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[14] Steven Y. Feng et al., "A Survey of Data Augmentation Approaches for NLP," *arXiv*, pp. 1-21, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[15] Sayali Hande, and M.A. Potey, "Temporal Text Summarization of TV Serial Excerpts Using Lingo Clustering and Lucene Summarizer," *SSRG International Journal of Computer Science and Engineering*, vol. 2, no. 5, pp. 16-22, 2015. [Google Scholar] [Publisher Link]

[16] Yu Li et al., "A Diverse Data Augmentation Strategy for Low-Resource Neural Machine Translation," *Information*, vol. 11, no. 5, pp. 1-12, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[17] Colin Raffel et al., "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer," *Journal of Machine Learning Research*, vol. 21, no. 1, pp. 1-67, 2020. [Google Scholar] [Publisher Link]

[18] Batuhan Baykara, and Tunga Güngör, "Turkish Abstractive Text Summarization Using Pretrained Sequence-to-Sequence Models," *Natural Language Engineering*, vol. 29, no. 5, pp. 1-30, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[19] Abhimanyu Chopra, Abhinav Prashar, and Chandresh Sain, "Natural Language Processing," *International Journal of Technology Enhancements and Emerging Engineering Research*, vol. 1, no. 4, pp. 131-134, 2013. [Google Scholar] [Publisher Link]

[20] Alpa Reshamwala, Dhirendra S. Mishra, and Prajakta Ganesh Pawar, "Review on Natural Language Processing," *IRACST Engineering Science and Technology*, vol. 3, no. 1, pp. 113-116, 2013. [Google Scholar]

[21] Mansi Agarwal, and Abhishek Saxena, "An Overview of Natural Language Processing," *International Journal for Research in Applied Science and Engineering Technology*, vol. 7, no. 5, pp. 2811-2813. 2019. [CrossRef] [Google Scholar] [Publisher Link]

[22] C.S. Saranyamol, and L. Sindhu, "A Survey on Automatic Text Summarization," *International Journal of Computer Science and Information Technologies*, vol. 5, no. 6, pp. 7889-7893, 2014. [Google Scholar] [Publisher Link]

[23] Dani Gunawan et al., "Automatic Text Summarization for Indonesian Language Using TextTeaser," *IOP Conference Series: Materials Science and Engineering*, vol. 190, no. 1, pp. 1-7, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[24] Adhika Pramita Widyassari et al., "Literature Review of Automatic Text Summarization: Research Trend, Dataset, and Method," *2019 International Conference on Information and Communications Technology (ICOIACT)*, pp. 491-496, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[25] Connor Shorten, and Taghi M. Khoshgoftaar, "A Survey on Image Data Augmentation for Deep Learning," *Journal of Big Data*, vol. 6, no. 60, pp. 1-48, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[26] Steven Y. Feng et al., "A Survey of Data Augmentation Approaches for NLP, *arXiv*, pp. 1-21, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[27] Jiaao Chen et al., "An Empirical Survey of Data Augmentation for Limited Data Learning in NLP," *Transactions of the Association for Computational Linguistics*, vol. 11, pp. 191-211, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[28] Huu-Thanh Duong, and Tram-Anh Nguyen-Thi, "A Review: Preprocessing Techniques and Data Augmentation for Sentiment Analysis, *Computational Social Networks*, vol. 8, no. 1, pp. 1-16, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[29] Pei Liu et al., "A Survey of Text Data Augmentation," *2020 International Conference on Computer Communication and Network Security (CCNS)*, pp. 191-195, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[30] Linting Xue et al., "mT5: A Massively Multilingual Pre-Trained Text-to-Text Transformer," *arXiv*, pp. 1-17, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[31] Ashish Vaswani et al., "Attention is all you Need," *31st Conference on Neural Information Processing Systems (NIPS 2017)*, pp. 1-11, 2017. [Google Scholar] [Publisher Link]

[32] Yinhan Liu et al., "Multilingual Denoising Pre-Training for Neural Machine Translation," *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 726-742, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[33] Mike Lewis et al., "Bart: Denoising Sequence-to-Sequence Pre-Training for Natural Language Generation, Translation, and Comprehension, *arXiv*, pp. 1-10, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[34] Samuel Cahyawijaya et al., "IndoNLG: Benchmark and Resources for Evaluating Indonesian Natural Language Generation, *arXiv*, pp. 1-24, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[35] Chin-Yew Lin, "Rouge: A Package for Automatic Evaluation of Summaries," *Association for Computational Linguistics*, pp. 74-81, 2004. [Google Scholar] [Publisher Link]

[36] Zhe Wang, "An Automatic Abstractive Text Summarization Model based on Hybrid Attention Mechanism," *Journal of Physics: Conference Series*, vol. 1848, no. 1, pp. 1-7, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[37] Dmitrii Aksenov et al., "Abstractive Text Summarization based on Language Model Conditioning and Locality Modeling," *arXiv*, pp. 1-10, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[38] Tahmid Hasan et al., "XL-Sum: Large-Scale Multilingual Abstractive Summarization for 44 Languages," *arXiv*, pp. 1-11, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[39] Rike Adelia, Suyanto Suyanto, and Untari Novia Wisesty "Indonesian Abstractive Text Summarization Using Bidirectional Gated Recurrent Unit," *Procedia Computer Science*, vol. 157, pp. 581-588, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[40] Muhammad Alfhi Saputra, Wikky Fawwaz Al Maki, and Nur Andini, "Abstractively Automatic Indonesian Text Summarizer Using Long Short-term Memory Method," *eProceedings of Engineering*, vol. 8, no. 2, pp. 1-15, 2021. [Google Scholar] [Publisher Link]

[41] Fajri Koto, Jey Han Lau, and Timothy Baldwin, "Liputan6: A Large-Scale Indonesian Dataset for Text Summarization," *arXiv*, 2020. [CrossRef] [Google Scholar] [Publisher Link]