JOIN (Jurnal Online Informatika)
p-ISSN: 2528-1682, e-ISSN: 2527-9165
Volume 10 Number 1 | June 2025: 196-204
DOI: 10.15575/join.v10i1.1506

Performance of Machine Learning Algorithms on Automatic Summarization of Indonesian Language Texts

Galih Wiratmoko[1], Husni Thamrin[2], Endang Wahyu Pamungkas[3]
[1]Department of Informatics, Universitas Muhammadiyah Madiun, Indonesia
[2,3]Department of Informatics, Universitas Muhammadiyah Surakarta, Indonesia

Article Info

Article history:

Received December 3, 2024
Revised April 1, 2025
Accepted April 16, 2025
Published May 15, 2025

Keywords:

Abstractive algorithms
Bahasa Indonesia
Hybrid model
T5-model
Text summarization

ABSTRACT

Automatic text summarization (ATS) has become an essential task for processing huge amounts of information efficiently. ATS has been extensively studied in resource-rich languages like English, but research on summarization for under-resourced languages, such as Bahasa Indonesia, is still limited. Indonesian presents unique linguistic challenges, including its agglutinative structure, borrowed vocabulary, and limited availability of high-quality training data. This study conducts a comparative evaluation of extractive, abstractive, and hybrid models for Indonesian text summarization, utilizing the IndoSum dataset which contains 20,000 text-summary pairs. We tested several models including LSA (Latent Semantic Analysis), LexRank, T5, and BART, to assess their effectiveness in generating summaries. The results show that the LexRank+BERT hybrid model outperforms traditional extractive methods, achieving better ROUGE precision, recall, and F-measure scores. Among the abstractive methods, the T5-Large model demonstrated the best performance, producing more coherent and semantically rich summaries compared to other models. These findings suggest that hybrid and abstractive approaches are better suited for Indonesian text summarization, especially when leveraging large-scale pre-trained language models.

Corresponding Author:

Husni Thamrin
Dept. of Informatics, Fac. of Communications and Informatics, Universitas Muhammadiyah Surakarta
Jl. A. Yani, Pabelan, Kartasura, Sukoharjo, Indonesia. 57162
Email: husni.thamrin@ums.ac.id

1. INTRODUCTION

Text summarization plays a crucial role in information retrieval and management because the methods enable users to quickly grasp the essence of large volumes of text while saving time and enhancing comprehension. Without summarization, individuals may struggle to extract key information from large volumes of text, leading to wasted time, decreased comprehension, and the potential for them to overlook key details [1], [2]. As effective text summarization techniques evolve, they improve information retrieval and pave the way for more advanced applications in various fields, such as news reporting, academic research, legal documentation, and social media. News article summarization, for example, allows readers to stay informed on current events without having to read lengthy articles, providing concise overviews that highlight the main points [3]. In academic settings, summarization aids students and researchers in distilling complex information into manageable insights, facilitating better understanding and retention of key concepts. Social media summarization provides users with quick insights into trending topics, allowing them to grasp the key discussions and opinions without sifting through countless posts [4].

The problem with text summarization is selecting which algorithm performs best to tackle the underlying tasks. The common approaches to automatic text summarization are extractive, abstractive, and hybrid [5][6]. Determining the most effective approach for text summarization is not a trivial issue because each approach has different characteristics. The extractive approach is based on selecting essential sentences from the original text, maintaining coherence but lacking flexibility [7]. In contrast, the abstractive approach focuses on a deep understanding of the document and re-state the ideas in new sentences that may have different structures [8]. The hybrid approach combines both extractive and abstractive methods, aiming to leverage the strengths of each [6][9].

In the context of Bahasa Indonesia, abstractive and extractive text summarization methods have been explored, with varying results. Conventional extractive techniques such as LexRank, Latent Semantic Analysis (LSA), and PageRank have been observed to produce ROUGE scores ranging from 0.38 to 0.64 [10][11]. Although abstractive techniques have not been extensively studied in-depth, there are works using the T5 and GPT-2 models that results in ROUGE scores of 0.61 and 0.51. Researchers have proposed improvements to these models, suggesting bias reduction in the generated summaries, the implementation of pre-processing steps, and the use of more powerful hardware for training on large datasets [12][13].

Despite these advances, there remain challenges in finding the most effective combination of summarization techniques, particularly for Indonesian-language texts. Continuous research is needed to evaluate and compare various new methods to find the optimal solution. Summarizing texts in Bahasa Indonesia presents unique challenges due to linguistic complexities, such as its agglutinative structure, flexible word order, and the use of borrowed words from various languages (e.g., Dutch, Javanese, and Arabic). Furthermore, observation of texts in Bahasa Indonesia poses specific challenges, particularly due to the limited availability of resources compared to widely spoken languages such as English. Recently, the IndoSum dataset, which contains 20,000 text-summary pairs, has emerged as the largest resource for Indonesian text summarization research. Several studies have utilized this dataset such as those conducted by [14][15], but further efforts are needed both to expand the dataset and to evaluate new methods to achieve better summarization performance.

The preceding paragraphs emphasize the needs to explore automatic summarization of Indonesian language texts. Investigations is required into unexplored abstractive summarization methods and technical improvements to explored methods such as in the finetuning process following suggestions of previous studies [12]. This study builds on the success of models like BART and T5-Large, which have demonstrated superior performance in English summarization tasks [8][15][16], but remain under-explored for Bahasa Indonesia. This paper presents a study that addresses this research gap by examining extractive and abstractive methods on the IndoSum dataset. To further complete the study, observations were also made on a hybrid model that combines the both extractive and abstractive methods. Hence, this paper contributes in determining which algorithms that work best to conduct summarization of Indonesian texts by investigating various models, i.e. abstractive, extractive and hybrid. Several methods have not been tested on texts in Bahasa Indonesia, so this work is novel in its scope.

The rest of the paper is organized as follows: In Section 2, we describe the methodology of the study, detailing the dataset used, pre-processing steps, and the configurations of the extractive, abstractive, and hybrid models employed for summarization. Section 3 presents the results and discussion, divided into two subsections: the first discusses the performance of extractive and hybrid models, while the second focuses on the results of abstractive models. Finally, Section 4 provides the conclusion, summarizing the key findings of the research and offering insights for future work in improving the performance and application of text summarization for Indonesian-language texts.

2. METHOD

This research was conducted in four main stages as illustrated in the methodological framework in Figure 1. The first stage involved selecting the dataset, which used the IndoSum dataset comprising 20,000 articles with corresponding gold-standard summaries [16]. The second stage was pre-processing to ensure the text is clean, consistent, valid, and ready for analysis [17][18]. The effectiveness of the pre-processing steps is expected to improve the performance of text summarization [19]. Pre-processing activities include removing unwanted characters, tokenization, text normalization, and stop word removal.

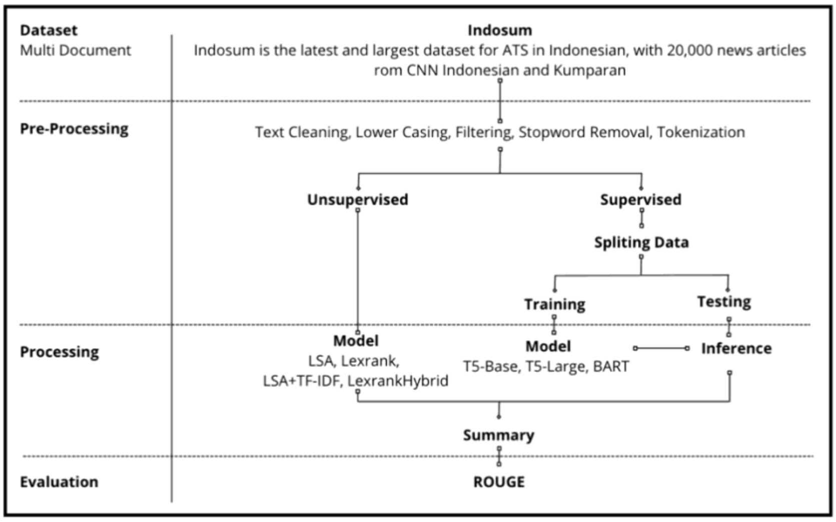

Dataset
Multi Document
Indosum
Indosum is the latest and largest dataset for ATS in Indonesian, with 20,000 news articles rom CNN Indonesian and Kumparan
Pre-Processing
Text Cleaning, Lower Casing, Filtering, Stopword Removal, Tokenization
Unsupervised
Supervised
Spliting Data
Training
Testing
Processing
Model
LSA, Lexrank, LSA+TF-IDF, LexrankHybrid
Model
T5-Base, T5-Large, BART
Inference
Summary
Evaluation
ROUGE

Figure 1. Research Method

The third stage focused on testing various extractive, abstractive, and hybrid models. Testing is carried out on identical datasets, which aims to ensure that the experimental conditions for each method were consistent. During the testing process, information is collected about the effectiveness, accuracy, and relevance of the summarization produced by each model. The methods tested are LSA (latent semantic analysis) and LexRank for extractive summarization models, T5 (Text-To-Text Transfer Transformer) and BART (Bidirectional and Auto-Regressive Transformers) for abstractive summarization models, and a combination of LSA+TF-IDF (Term Frequency-Index Document Frequency) and Lexrank+BERT (Bidirectional Encoder Representation from Transformer) methods for hybrid models.

In addition, this study tested the T5-Large model, which had not previously been used for Indonesian text summarization. The parameters used in the tests included 6 epochs with a batch size of 4, an input length of 512 tokens, a summary length of 128 tokens, and optimization using the AdamW method with a learning rate of 1e-4. Although T5-base is lighter and faster in training T5-Large was chosen due to its ability to capture complex data structures, which can improve performance despite higher computing resource requirements. The T5-large method has a number of parameters of 737 million and a model estimation size of 2.9 gigabytes (GB) in training in this study.

This study tested the BART model that can be used to generate new text summaries by feeding input text into a pre-defined summarization pipeline. The objective function describes the modeling training process, where many hyperparameters are randomly selected within a certain range. In this study, a learning rate in the interval 1e-5 to 1e-4 was used. While the per_device_train_batch_size parameter was selected one from the values {8, 16, 32} and gradient_cumulation_steps was selected one from the set of values {2, 4, 8}. Hyperparameter optimization was performed with Optuna to determine at which values the hyperparameters were set, either for the learning rate, per device train batch size, or 32, and gradient_cumulation_steps. In addition, other parameters use fixed values and are not applied to the optimization process, for example epochs at a value of 3, per device eval batch size of 8, weight decay of 0.01, save total limit of 2. Furthermore, the predict with generate value is set to active (True) and FP16 is set to active (True) to speed up training using 16-bit floating point. After determining the optimal hyperparameters with a series of iterations, Optuna continues with the final training of the model using these hyperparameters.

This study tested the hybrid LSA+TF-IDF and Lexrank+BERT models to correct errors and improve the quality of the summary by utilizing the strengths of each model. The LSA method consists of three main steps, namely creating an input matrix, using singular value decomposition (SVD), and selecting sentences. The input document is represented by a matrix as a column representing a sentence, a row representing a word, and each cell indicating the importance of the word in the sentence. The improvement of the LSA model is done by combining TF-IDF. Lexrank+BERT improves the weaknesses of the traditional Lexrank model by adding key extraction and combination scores to improve the relevance and coverage of the summary. After tokenization, keywords are extracted using keyBERT with the aim of helping to identify important words. The formation of a combination score is done by combining the LexRank score with the sentence position score and the presence of keywords previously done using keyBERT.

Evaluation of summary results from various techniques can use various criteria such as clarity, summary length, and accuracy. In this study, the evaluation was carried out using the ROUGE (Recall-Oriented Understudy for Disposition Evaluation) metric which is widely used and accepted as a measure of text summary quality [20]. ROUGE-N assesses the similarity between the automatic summary and the reference summary using n-grams, with "N" indicating the size of the n-grams used. Quite popular n values are 1 and 2 which produce the ROUGE-1 and ROUGE-2 metrics, which respectively evaluate the number of unigrams (individual words) and bigrams (adjacent word pairs) that match between the automatic summary results and the reference summary. The formula for determining ROUGE-N is stated in equation (1) where gramn refers to n-grams and and countMatch(gramn) is the largest number of n-grams that appear together in the generated summary and the reference summary [21][22].

$$ROUGE-N=\frac{\sum_{s\epsilon\,Refsum}\sum_{gramn\in s}countMatch(gramn)}{\sum_{s\epsilon\,\{Refsum\}}\sum_{gramn\in s}count(gramn)} \tag{1}$$

The performance evaluation of the summarization method is also carried out using the ROUGE-L method, where L indicates the use of the Longest Common Sequence (LCS) technique, namely the number of consecutive word sequences in the summary. This metric offers a more refined assessment of the structural similarity between two texts. The calculation of ROUGE-L can be done using the formula in equation (2), where the LCS term (S, S') indicates how much the longest word sequence is the same between the automatic summary results and the reference summary [21][22].

$$ROUGE-L=\frac{\sum_{s\epsilon\,Refsum}LCS(S,S')}{\sum_{s\epsilon\{\,Refsum\}}length(S)} \tag{2}$$

Both ROUGE metrics (namely ROUGE-N and ROUGE-L) calculate precision, recall, and F-measure, which are the main components in evaluating model performance. Equations (3) to (5) show the formulas in determining these three metrics.

$$Precision=\frac{Correct}{Correct+Wrong} \tag{3}$$
$$Recall=\frac{Correct}{Correct+Missed} \tag{4}$$
$$F-measure=2\cdot\frac{Precision\cdot Recall}{Precision+Recall} \tag{5}$$

3. RESULT AND DISCUSSION

This section discusses the comparative analysis of the performance of several proposed approaches. The analysis is mainly conducted on the performance of the hybrid model and the abstractive model. Both methods have not been widely explored for the process of summarizing Indonesian text.

3.1. Extractive and Hybrid Models

The first experiment was conducted on the extractive and hybrid models using a sample of 20,000 articles evaluated using the ROUGE-1, ROUGE-2, and ROUGE-L metrics. The results are shown in Table 1 and Table 2. The table shows that the hybrid LSA+TF-IDF model outperforms the single LSA method. The application of the hybrid model gives higher scores in all ROUGE metrics. The value of this metric indicates that the resulting summary covers most of the words in the original document. The superiority of the hybrid model is reinforced by the results of observations on the hybrid Lexrank+BERT method which shows better summarization results than the Lexrank method alone. The hybrid model gives higher metric values for all precision, recall, and F-Measure values.

Table 1. ROUGE-1 and ROUGE-2 for hybrid models

	ROUGE-1			ROUGE-2		
	P	R	F	P	R	F
LSA	0.3648	0.6193	0.4524	0.2771	0.4789	0.3460
LSA+TF-IDF	0.6445	**0.8055**	0.6296	0.5096	**0.6616**	0.5418
Lexrank	0.3695	0.5992	0.4493	0.2734	0.4530	0.3353
Lexrank+BERT	**0.6943**	0.7084	**0.6935**	**0.5473**	0.5649	**0.5560**

Table 2. ROUGE-L for hybrid models

	ROUGE-L		
	P	R	F
LSA	0.3181	0.5428	0.3953
LSA+TF-IDF	0.5583	**0.6855**	0.5945
Lexrank	0.2793	0.4582	0.3411
Lexrank+BERT	**0.5665**	0.5849	**0.6260**

The results in Table 1 and Table 2 show that the LSA method yielded modest results in text summarization, with a ROUGE precision in the range from 0.2771 to 0.3648, a recall in the range of 0.4789 to 0.6193, and an F-measure in the range of 0.3460 to 0.4524. This indicates that LSA captures some key features but struggles to maintain a balance between precision and recall, especially in generating concise summaries. When combining LSA with TF-IDF, the performance significantly improved. The ROUGE precision rose to the range between 0.5096 and 0.6445, and the precision rose to the range between 0.6616 and 0.8055, and the F-measure in the range between 0.5418 and 0.6296. This improvement suggests that incorporating TF-IDF helps enhance the extraction of relevant content, leading to better summarization performance.

LexRank showed moderate performance, similar to LSA. The ROUGE precision for Lexrank was in the range from 0.2734 to 0.3695, the recall was in the range from 0.4530 to 0.5992, and the F-measure was in the range from 0.3353 to 0.4493. Lexrank's performance was close to LSA, indicating that it effectively identifies important sentences but with limited improvement over the basic LSA model. The combination of Lexrank and BERT yielded the best results among all the methods. The ROUGE precision was in the range of 0.5473 to 0.6943, with a recall in the range of 0.5473 to 0.7084, and an F-measure in the range of 0.5560 to 0.6935. These results indicate that combining LexRank with BERT leads to better summarization by improving both precision and recall in identifying key content.

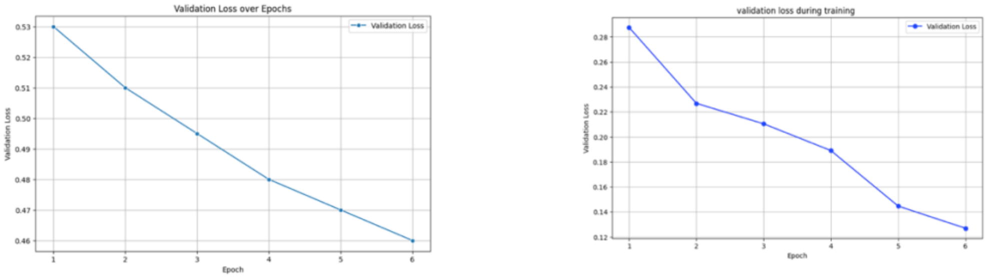
3.2. Abstractive Models

The second experiment evaluated abstractive models using the T5-Base, T5-Large, and BART models. The experiment used 20,000 data points, of which 75% were used for training, and the remaining for validation and testing.

The training process was carried out using PyTorch Lightning which facilitates the determination of specifications for various functions such as training steps, validation steps, and testing steps. All steps are required in the training activity in a structured manner. The training process utilizes the ModelCheckPoint function which stores the most optimal model during training so that the best model can be obtained. The model with the best parameters has minimum validation loss. In addition, TensorBoardLogger is used to display a visual depiction of the training process to assist the process of model analysis and improvement.

The use of models in PyTorch Lightning facilitates the specification of certain functions such as training steps (training_step), validation (validation_step), and testing (test_step), all of which play a role in an efficient and structured training procedure. The use of the ModelCheckPoint callback guarantees the most optimal model storage considering the minimum validation loss value, while TensorBoardLogger offers a visual representation of the training process, helping in the analysis and improvement of the model. After the model is trained, the inference process is carried out on the validation data to generate text summaries.

The training of the BART and T5 models is carried out in 6 epochs. Figure 2 shows how the validation loss value changes in each epoch for the training of the T5-Large model (figure 2.a) and the training of the BART model (figure 2.b). It can be seen that the models improve as the training process progresses. Training is stopped at 6 epochs because the models obtained with larger epochs no longer produce significantly better summarization performance.



(a)

(b)

Figure 2. Model training (a) T5-large (b) BART

After the training process is complete, the model is tested using 25% of the data from the dataset. Table 3 shows the evaluation metric values of ROUGE-1 and ROUGE-2 against the automatic summarization results compared to the reference summary. While Table 4 shows the metric values for ROUGE-L. For the three metrics, the Precision, Recall and F-Measure values are presented.

Table 3. R-1 and R-2 for abstractive summarization

	ROUGE-1			ROUGE-2		
	P	R	F	P	R	F
T5-Base	0.7246	0.7277	0.7231	0.6573	0.**6597**	0.**6557**
T5-Large	0.**7647**	**0.7876**	**0.7786**	0.**6864**	0.6145	0.6153
BART	0.6208	0.5886	0.5686	0.5852	0.5413	0.5268

Table 4. ROUGE-L for abstractive summarization

	ROUGE-L		
	P	R	F
T5-Base	0.6958	0.6979	0.6940
T5-Large	0.**7278**	**0.74450**	**0.7507**
BART	0.5722	0.5754	0.5576

Table 3 and Table 4 show that the T5-Base model exhibited strong performance across the metrics, especially in terms of precision and recall. It achieved a precision in the range of 0.6573 to 0.7246, a recall in the range of 0.6579 to 0.7277, and an F-measure in the range between 0.6557 and 0.7231. These results suggest that T5-Base is generally effective in both generating relevant and comprehensive summaries, hitting a good balance between precision and recall.

The T5-Large model, a more advanced version of T5-Base, outperformed T5-Base in most metrics. The model achieved a precision in the range of 0.6864 to 0.7647, a recall in the range of 0.6145 to 0.7876, and an F-measure in the range from 0.6153 to 0.7786. The results show significant improvement in precision and recall over the base model. However, in ROUGE-2, while the precision was still high at 0.6864, the recall and F-Measure dropped to about 0.61. This indicates that T5-Large generates more accurate summaries, but may struggle to cover all relevant content, especially for longer or more complex texts.

The BART model performed slightly below the T5 models. The results suggest that BART may be capable of generating decent summaries. However, it is less effective than both versions of T5, particularly in achieving precision and recall. BART showed lower overall performance suggesting that it may struggle to produce summaries that are as comprehensive or accurate as the T5 models.

3.3. Discussion

The results show that the combination of Lexrank and BERT yields the best results among extractive and hybrid models. This phenomenon is due to the complementary strengths of both models. Lexrank is good at extractive summarization. It is a graph-based algorithm that effectively ranks sentences by measuring their lexical similarity [23], drawing inspiration from Google's PageRank. While it excels in identifying salient sentences for text summarization, its reliance on lexical similarity can limit its ability to capture deeper semantic meanings, potentially leading to less precise sentence selection[24]. On the other hand, BERT is a deep learning model that excels in understanding the semantic context of words within sentences. Unlike traditional approaches that only consider word-level relationships, BERT captures the meaning of words in context by analyzing entire sentence structures. The attention mechanisms in BERT enable it to discern relationships between words and their roles in sentences, contributing to its ability to understand nuanced meanings [25], [26]. By combining the graph-based approach of Lexrank with BERT's powerful semantic understanding, the summarization becomes both more comprehensive and accurate. BERT's contextual embeddings refine Lexrank's sentence rankings, ensuring that the selected sentences are not only important but also semantically rich and relevant to the summary. This statement agrees with [27] that stated that the combination of LexRank and BERT outperforms traditional methods, as BERT enhances the representation of sentences, leading to more coherent and comprehensive summaries.

Among abstractive models under investigations, the performance of the T5 models (T5-Base and T5-Large) is superior over BART. This may be explained by differences in their architectures, training strategies, and objectives, which impact how effectively they handle text summarization tasks. The T5 (Text-to-Text Transfer Transformer) model exemplifies a unified approach to natural language processing (NLP) by framing all tasks as text-to-text problems [28]. This design allows T5 to generalize effectively across various NLP tasks, including summarization, question answering, and text classification [29]. BART (Bidirectional and Auto-Regressive Transformers) employs a denoising autoencoder objective, which effectively reconstructs corrupted text sequences [30]. While this approach excels in text generation tasks, it does not explicitly emphasize span prediction, which is crucial for certain applications [31] where understanding and predicting spans is critical such as question answering and abstractive summarization. The T5 models' text-to-text formulation and focus on directly generating sequences make them better suited for abstractive summarization compared to BART's emphasis on reconstructing corrupted inputs.

This study provides better metric values than several previous studies that have been observed. The model with larger parameters applied in summarizing Indonesian text shows better results with rouge-1 and ROUGE-2 metric values of 0.7647 and 0.7876 respectively, while previous research by [12] using T5-base with fewer parameters and smaller training data produced summarization with ROUGE-1 of 0.61 and ROUGE-2 of 0.51. These results emphasize the importance of the number of model parameters and dataset size in influencing the performance of the summarization model. This statement is in line with findings in other studies that show an increase in summary quality with the use of larger datasets as shown in research with the BBC News dataset [32][33]. In this study, the ROUGE-1 and ROUGE-2 scores, which were initially 0.45 and 0.26, could be increased to 0.69 and 0.59 after using a larger dataset.

4. CONCLUSION

This study presents a comparative evaluation of extractive, abstractive, and hybrid models for automatic text summarization. The presentation is an effort to solve the problem of selecting best algorithms or models to conduct summarization of texts in Bahasa Indonesia using the IndoSum dataset as samples. The results show that the hybrid models, particularly the LexRank+BERT model, outperform traditional extractive methods in terms of ROUGE precision, recall, and F-measure. In contrast, the abstractive models, especially T5-Large, showcased superior results, generating more coherent and semantically rich summaries compared to other approaches. The comparison of T5 and BART models highlighted that the T5 framework, with its text-to-text formulation, was better suited for summarization tasks.

The findings show the importance of combining extractive and abstractive techniques to optimize summary quality. Moreover, the results underline the effect of model size and training data volume in improving summarization performance for supervised models. Future work could focus on further optimization of hybrid and abstractive models, as well as using larger datasets and exploring new architectures. Potential future works include fine-tuning pre-trained models like T5 and BART specifically for Indonesian texts which may yield better results in abstractive summarization tasks.

ACKNOWLEDGEMENTS

The authors thank the Directorate of Research, Technology and Community Development (Direktorat Riset, Teknologi dan Pengabdian Masyarakat, DRTPM), the Republic of Indonesia, that has funded the work through the scheme of Thesis Research Grant contract number: 007/LL6/PB/AL.04/2024, 196.98/A.3-III/LRI/VI/2024.

REFERENCES

[1] L. Abualigah, M. Q. Bashabsheh, H. Alabool, and M. Shehab, "Text Summarization: A Brief Review," 2020, pp. 1–15. doi: 10.1007/978-3-030-34614-0_1.

[2] J. Patel, N. Chauhan, and K. Patel, "Text Summarization Using Natural Language Processing," *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, pp. 16–22, May 2023, doi: 10.32628/10.32628/CSEIT2390298.

[3] J. Deny, S. Kamisetty, H. V. R. Thalakola, J. Vallamreddy, and V. K. Uppari, "Inshort Text Summarization of News Article," in *2023 7th International Conference on Intelligent Computing and Control Systems (ICICCS)*, IEEE, May 2023, pp. 1104–1108. doi: 10.1109/ICICCS56967.2023.10142549.

[4] V. Sireesha, N. P. Hegde, and M. Yashwanth, "Automatic Text Summarization Using Intelligent Systems," 2022, pp. 253–262. doi: 10.1007/978-981-16-9705-0_25.

[5] Á. Hernández-Castañeda, R. A. García-Hernández, Y. Ledeneva, and C. E. Millán-Hernández, "Extractive Automatic Text Summarization Based on Lexical-Semantic Keywords," *IEEE Access*, vol. 8, pp. 49896–49907, 2020, doi: 10.1109/ACCESS.2020.2980226.

[6] B. Khan, Z. A. Shah, M. Usman, I. Khan, and B. Niazi, "Exploring the Landscape of Automatic Text Summarization: A Comprehensive Survey," *IEEE Access*, vol. 11, pp. 109819–109840, 2023, doi: 10.1109/ACCESS.2023.3322188.

[7] N. Giarelis, C. Mastrokostas, and N. Karacapilidis, "Abstractive vs . Extractive Summarization : An Experimental Review," 2023.

[8] Á. Hernández-Castañeda, R. A. García-Hernández, Y. Ledeneva, and C. E. Millán-Hernández, "Extractive Automatic Text Summarization Based on Lexical-Semantic Keywords," *IEEE Access*, vol. 8, pp. 49896–49907, 2020, doi: 10.1109/ACCESS.2020.2980226.

[9] B. M. Gurusamy, P. K. Rengarajan, and P. Srinivasan, "A hybrid approach for text summarization using semantic latent Dirichlet allocation and sentence concept mapping with transformer," vol. 13, no. 6, pp. 6663–6672, 2023, doi: 10.11591/ijece.v13i6.pp6663-6672.

[10] N. Khotimah and A. S. Girsang, "Indonesian News Articles Summarization Using Genetic Algorithm," *Engineering Letters*, vol. 30, no. 1, pp. 152 – 160, 2022.

[11] Y. M. Sari and N. S. Fatonah, "Peringkasan Teks Otomatis pada Modul Pembelajaran Berbahasa Indonesia Menggunakan," vol. 7, no. 2, pp. 153–159, 2021.

[12] A. Bahari and K. E. Dewi, "peringkasan teks otomatis abtraktif menggunakan transformer pada teks bahasa indonesia," *KOMPUTA : Jurnal Ilmiah Komputer dan Informatika*, vol. 13, no. 1, 2024.

[13] A. nur Khasanah and M. Hayati, "Abtsractive-Based Automatic Text Summarization On Indonesian News Using GPT2," vol. X, no. 1, pp. 9–18, 2023.

[14] A. S. Girsang and F. J. Amadeus, "Extractive Text Summarization for Indonesian News Article Using Ant System Algorithm," vol. 14, no. 2, pp. 295–301, 2023, doi: 10.12720/jait.14.2.295-301.

[15] K. E. Dewi and N. I. Widiastuti, "Automatic Summarization of Indonesian Texts Using a Hybrid Approach," vol. 15, no. 1, 2022.

[16] K. Kurniawan and S. Louvan, "I NDO S UM : A New Benchmark Dataset for Indonesian Text Summarization," *2018 International Conference on Asian Language Processing (IALP)*, pp. 215–220, 2018.

[17] N. Babanejad, H. Davoudi, A. Agrawal, A. An, and M. Papagelis, "The Role of Preprocessing for Word Representation Learning in Affective Tasks," *IEEE Trans Affect Comput*, vol. 15, no. 1, pp. 254–272, 2024, doi: 10.1109/TAFFC.2023.3270115.

[18] E. Qais and V. M. N., "TxtPrePro: Text Data Preprocessing Using Streamlit Technique for Text Analytics Process," in *2023 International Conference on Network, Multimedia and Information Technology (NMITCON)*, 2023, pp. 1–6. doi: 10.1109/NMITCON58196.2023.10275887.

[19] G. Khekare, C. Masudi, Y. K. Chukka, and D. P. Koyyada, "Text Normalization and Summarization Using Advanced Natural Language Processing," in *2024 International Conference on Integrated Circuits and Communication Systems (ICICACS)*, 2024, pp. 1–6. doi: 10.1109/ICICACS60521.2024.10498983.

[20] M. A. Dursun and S. Serttaş, "A Multi-Metric Model for analyzing and comparing extractive text summarization approaches and algorithms on scientific papers," vol. 1, pp. 31–48, 2024, doi: 10.24012/dumf.1376978.

[21] A. R. Lubis, H. R. Safitri, and M. Lubis, "improving text summarization quality by combining T5-based models and convulutional seq2seq models," vol. 5, no. 1, pp. 451–459, 2023.

[22] A. Al Foysal and R. Böck, "Who Needs External References ?— Text Summarization Evaluation Using Original Documents," pp. 970–995, 2023.

[23] G. Erkan and D. R. Radev, "LexRank: Graph-based Lexical Centrality as Salience in Text Summarization," *Journal of Artificial Intelligence Research*, vol. 22, pp. 457–479, Dec. 2004, doi: 10.1613/jair.1523.

[24] R. Mihalcea and D. Radey, "Graph-based natural language processing and information retrieval.," *Cambridge University Press*, 2011.

[25] K. Clark, U. Khandelwal, O. Levy, and C. D. Manning, "What does BERT look at? an analysis of BERT's attention," *arXiv preprint arXiv:1906.04341*, 2019, doi: 10.18653/v1/W19-4828.

[26] S. MacAvaney, A. Yates, A. Cohan, and N. Goharian, "CEDR," in *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, New York, NY, USA: ACM, Jul. 2019, pp. 1101–1104. doi: 10.1145/3331184.3331317.

[27] S. Narayan, S. B. Cohen, and M. Lapata, "Ranking Sentences for Extractive Summarization with Reinforcement Learning," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2018, pp. 1747–1759. doi: 10.18653/v1/N18-1158.

[28] C. Raffel *et al.*, "Exploring the limits of transfer learning with a unified text-to-text transformer," *Journal of machine learning research*, vol. 21, no. 140, pp. 1–67, 2020.

[29] L. Xue *et al.*, "mt5: A massively multilingual pre-trained text-to-text transformer," *arXiv preprint arXiv:2010.11934*, 2020, doi: 10.48550/arXiv.2010.11934.

[30] M. Lewis *et al.*, "Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," *arXiv preprint arXiv:1910.13461*, 2019, doi: 10.48550/arXiv.1910.13461.

[31] M. Joshi, D. Chen, Y. Liu, D. S. Weld, L. Zettlemoyer, and O. Levy, "Spanbert: Improving pre-training by representing and predicting spans," *Trans Assoc Comput Linguist*, vol. 8, pp. 64–77, 2020.

[32] J. Ranganathan and G. Abuka, "Text Summarization using Transformer Model," in *2022 Ninth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, 2022, pp. 1–5. doi: 10.1109/SNAMS58071.2022.10062698.

[33] A. M. A. Zeyad and A. Biradar, "Advancements in the Efficacy of Flan-T5 for Abstractive Text Summarization: A Multi-Dataset Evaluation Using ROUGE and BERTScore," in *2024 International Conference on Advancements in Power, Communication and Intelligent Systems (APCI)*, 2024, pp. 1–5. doi: 10.1109/APCI61480.2024.10616418.