 ISSN:2302-8149
e-ISSN:2540-9719

Comparative Analysis of T5 Model Performance for Indonesian Abstractive Text Summarization

[1]Mohammad Wahyu Bagus Dwi Satya*, [2]Ardytha Luthfiarta, [3]Mohammad Noval Althoff
[1,2,3]Informatics Engineering, Faculty of Computer Science, Universitas Dian Nuswantoro,
[1,2,3]Jl. Imam Bonjol No.207, Pendrikan Kidul, Kec. Semarang Tengah, Semarang, Jawa Tengah, Indonesia
*e-mail: bagus8545@gmail.com

(***received*:** 6 December 2024, ***revised*:** 7 March 2025, ***accepted*:** 10 March 2025)

Abstract

The rapid growth of digital content has created significant challenges in information processing, particularly in languages like Indonesian, where automatic summarization remains complex. This study evaluates the performance of different T5 (Text-to-Text Transfer Transformer) model variants in generating abstractive summaries for Indonesian texts. The research aims to identify the most effective model variant for Indonesian language summarization by comparing T5-Base, FLAN-T5 Base, and mT5-Base models. Using the INDOSUM dataset containing 19,000 Indonesian news article-summary pairs, we implemented a 5-Fold Cross-Validation approach and applied ROUGE metrics for evaluation. Results show that T5-Base achieves the highest ROUGE-1, ROUGE-2, and ROUGE-L scores of 73.52%, 64.50%, and 69.55%, respectively, followed by FLAN-T5, while mT5-Base performs the worst. However, qualitative analysis reveals various summarization errors: T5-Base exhibits redundancy and inconsistent formatting, FLAN-T5 suffers from truncation issues, and mT5 often generates factually incorrect summaries due to misinterpretation of context. Additionally, we assessed computational performance through training time, inference speed, and resource consumption. The results indicate that mT5-Base has the shortest training time and fastest inference speed but at the cost of lower summarization accuracy. Conversely, T5-Base, while achieving the highest accuracy, requires significantly longer training time and greater computational resources. These findings highlight the trade-offs between accuracy, error tendencies, and computational efficiency, providing valuable insights for developing more effective Indonesian language summarization systems and emphasizing the importance of model selection for specific language tasks.

Keywords: natural language processing, text summarization, transformers, T5, ROUGE

1 Introduction

The ability to quickly absorb and understand information in the digital era, marked by technological advancements, is becoming increasingly crucial[1]. Every day, the internet is filled with millions of documents, articles, and other content, creating new challenges for users in filtering and extracting relevant information. This challenge is particularly felt by organizations and public institutions that must manage large volumes of text data, often resulting in decreased efficiency and loss of focus[2]. Manually processing large amounts of text data is becoming increasingly difficult because of the sheer scale of the available information. Automatic text summarization technology has emerged as a promising solution to address this issue[3]. This process aims to produce a shortened version of a longer text while retaining the core and main meanings of the information. Thus, users can quickly obtain important information without having to read the entire document. The demand for automatic summarization tools is increasing, especially for languages such as Indonesian, where automatic summarization remains a complex challenge that is yet to be fully resolved[4]. This solution offers efficiency in information management, saving the time and resources required to process text.

Automatic text summarization involves two main approaches, extractive and abstractive. The extractive approach is simpler, as the system selects only the most relevant sentences or phrases from the source text without creating new sentences [5]. Extractive summarization typically involves techniques such as sentence ranking based on relevance, which allows the model to identify the main

 ISSN:2302-8149
e-ISSN:2540-9719

components of the document. While it can produce quick and easily generated summaries, extractive summarization sometimes lacks in flow and readability because the resulting summary still retains the structure of the original sentences[6].

On the other hand, the abstractive approach is different in that it does not simply replicate the content from the original text. Instead, this approach generates new sentences by restructuring the information to capture the essential meaning in a coherent and natural form[5]. This method requires a system to analyze the source text, identify the most relevant information, and then synthesize a summary that accurately conveys the main concepts, resulting in a summary that is natural and easy to read. Although more complex compared to extractive summarization, abstractive summarization offers higher quality by producing summaries that are contextually accurate and coherent, making it suitable for applications in which coherence and readability are highly important[7]. Therefore, the abstractive text summarization approach is becoming increasingly important, especially in contexts where the quality and understanding of information are critical.

In the context of developing automatic summarization technology, the T5 model (Text-To-Text Transfer Transformer) has emerged as one of the latest innovations, offering unique capabilities in abstractive text summarization[8]. This model transforms various natural language processing tasks into a text-to-text format, allowing greater flexibility in handling different types of input and output[9]. Compared to previous models such as BERT, which is predominantly used for extractive summarization, T5 excels in generating more natural and coherent summaries by restructuring information into a more concise and readable format[10]. Previous research has shown that T5-based models achieve higher ROUGE scores than traditional extractive methods, particularly in news and scientific document summarization tasks[9]. With this flexibility, T5 has the potential to be adapted for Indonesian text summarization, providing an innovative approach to address existing challenges in this field. Although several previous studies have explored automatic text summarization in Indonesian using extractive, there is still a significant gap in the development of effective abstractive summarization methods for Indonesian. In particular, there is no comprehensive study comparing the performance of different variants of modern transformer models such as T5, FLAN-T5, and mT5 in the context of Indonesian text summarization. This study fills the gap by conducting a systematic analysis of the capabilities of the three model variants, with a particular focus on the aspects of semantic accuracy, text coherence, and summarization effectiveness in the Indonesian linguistic context. Through this comparison, we aim to evaluate how well these models handle the complexity of abstractive summarization in Indonesian and to determine which variant offers the best balance in terms of accuracy, fluency, and coherence in generating high-quality summaries. The significance of this research lies in its contribution to the development of high-quality abstractive summarization for the Indonesian language, an area that remains underexplored. By systematically comparing the performance of T5, FLAN-T5, and mT5 models, this study provides valuable insights into their effectiveness in handling Indonesian text summarization.

The findings help bridge the gap in existing research by identifying the strengths and weaknesses of each model in terms of semantic accuracy, coherence, and summarization efficiency. Furthermore, this study has practical implications for organizations, media platforms, and digital information management systems that require accurate and efficient summarization tools for large-scale Indonesian text processing. The results can serve as a reference for future improvements in transformer-based summarization, offering guidelines for model selection and fine-tuning strategies to optimize performance in low-resource language settings. By advancing the capabilities of abstractive summarization in Indonesian, this research contributes to the broader field of natural language processing and supports the development of more intelligent, automated information-processing systems.

2 Literature Review

The literature review contains a discussion of the research that highlights the key points of this study. The sources of the literature review are taken from various references, including journals, books, theses, and other scientific works.

 ISSN:2302-8149
e-ISSN:2540-9719

2.1 K-Fold Cross-Validation

K-Fold Cross-Validation is a validation technique used to measure model performance by dividing the dataset into multiple balanced subsets. In this process, the model is trained using $k-1$ subsets and tested on the remaining subset. This process is repeated k times, so each subset serves as test data exactly once. Previous research that applied K-Fold Cross-Validation for automatic text summarization in Hindi has shown that using K-Fold Cross-Validation can improve accuracy, especially with limited-sized datasets[11]. This approach helps to evaluate model performance more accurately and reduces the risk of overfitting, providing a more consistent estimate of model performance across different data.

2.2 Related Research

Various studies have examined automatic text summarization, both through extractive and abstractive approaches, on news articles and documents. In the extractive approach, pretrained encoder methods such as Bidirectional Encoder Representation Transformer (BERT) have been applied to the 'CNN/DailyMail' dataset, achieving ROUGE-1, ROUGE-2, and ROUGE-L scores of 43.23%, 20.24%, and 39.63%, respectively[12]. These results demonstrate the model's effectiveness in extracting key information from lengthy texts. On the other hand, studies using the abstractive approach also leverage transformer-based models. In one such study, the use of BERT resulted in ROUGE-1, ROUGE-2, and ROUGE-L scores of 41.72%, 19.39%, and 38.76%, respectively[13]. Although both approaches are promising for text summarization, the results obtained still indicate room for improvement, particularly in terms of accuracy and coherence.

Our observations show that the transformer architecture, with its self-attention mechanism, exhibits superior performance in summarizing large-sized texts[14]. Therefore, we plan to apply transformer architectures with pretrained language models (PTLMs) on large-scale data. Recent research has focused on using T5 (Text to Text Transfer Transformer) derivatives for text summarization, demonstrating significant improvements in summary quality. For instance, one study utilized a T5 model further fine-tuned with a news dataset, achieving a ROUGE-1 score of 43.02%, ROUGE-2 score of 14.50%, and ROUGE-L score of 37.43%[15]. Another study also reported that T5 derivatives were capable of generating more informative and coherent summaries, with higher ROUGE scores than previous models.

Another study using T5 derivatives on scientific documents showed promising results, with a ROUGE-1 score of 45.00%, ROUGE-2 score of 20.00%, and ROUGE-L score of 40.00%[16], indicating its effectiveness in capturing the core information from more complex texts. Furthermore, another research by showed that T5 derivative models can be well adapted for languages other than English, producing high-quality summaries on Persian news datasets[17]. These findings suggest that the use of T5 models on non-English datasets holds great potential for improving the quality of generated summaries. Further research is needed to compare the performance of various T5 model variants in the context of the Indonesian language, so as to identify the most effective model for generating informative and coherent summaries.

3 Research Method

The following is the research methodology to be applied for automatic text summarization, shown in Figure 1.

 ISSN:2302-8149
e-ISSN:2540-9719

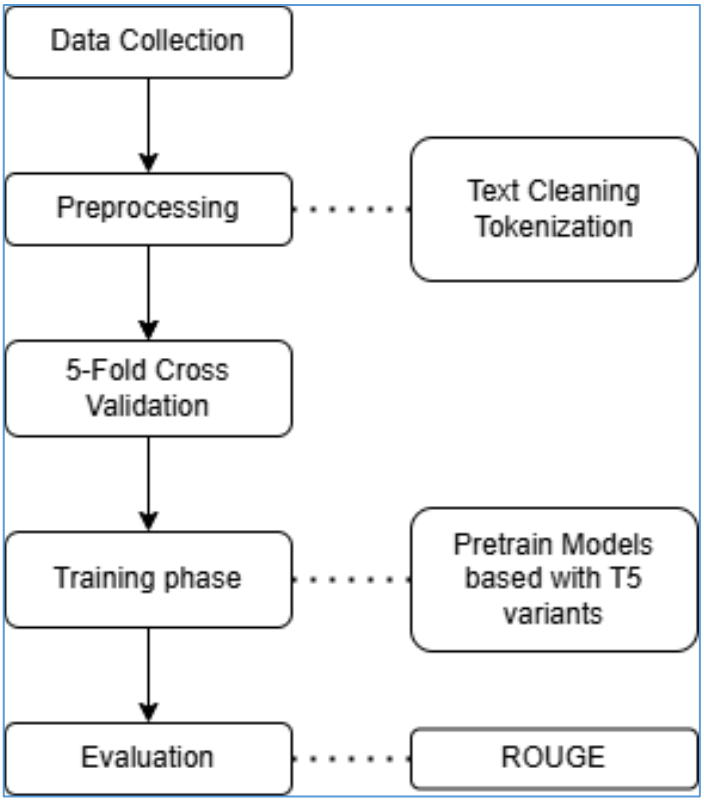

Figure 1. Research framework diagram

3.1 Data Collection

This study utilizes the INDOSUM dataset, an Indonesian-language dataset specifically designed for automatic text summarization[18]. The dataset was obtained from various Indonesian online news portals and is available in a public repository on *GitHub - kata-ai/indosum*. It contains 19,000 news article-summary pairs, where each article is accompanied by a human-annotated gold summary.

The dataset includes articles ranging from 100 to 1,500 words, with summaries between 20 to 100 words. Written in formal Indonesian, it captures variations in sentence complexity, vocabulary, and journalistic tone. The summaries, created by human annotators, ensure accuracy and coherence, making them a reliable benchmark for evaluating summarization models.

INDOSUM is categorized into sports, entertainment, inspiration, technology, headlines, and the entertainment industry, each presenting unique challenges for summarization. Sports involves event reporting, entertainment includes informal language, inspiration focuses on sentiment and storytelling, technology requires precise handling of technical terms, headlines cover diverse factual topics, and the entertainment industry deals with media business trends. These categories represent different domains of news coverage; however, this study does not differentiate between them during training and evaluation. Instead, all news categories are treated equally to assess the general effectiveness of abstractive summarization models.

3.2 Preprocessing

In this study, the preprocessing method applied includes several key steps to ensure data quality and consistency before being used in the model. The steps taken are as follows:

1) Text Cleaning: This process aims to clean the text of unwanted characters. The steps include:
    a. Lower Casing: Converting all letters to lowercase to standardize the text format. This is important to avoid discrepancies between identical words with different letter cases.
    b. Filtering with Regex: Using specific regex patterns, such as ([a-z0-9- +,.\/\\]*[a-z0-9][a-z0-9-+]*), to remove irrelevant characters or those that don't match the desired pattern. Using regex allows for more efficient and accurate text cleaning.

 ISSN:2302-8149
e-ISSN:2540-9719

2) Tokenization: After text cleaning, the next step is tokenization, where the text is divided into smaller units, such as words or phrases. In the context of the T5 model and its variants, tokenization using the SentencePiece method is particularly relevant.
    a. SentencePiece: This is a subword-based tokenization method that enables the model to handle a larger vocabulary by converting words into smaller segments, thus reducing out-of-vocabulary (OOV) issues[19]. With SentencePiece, the T5 model can work better with various languages and dialects and capture more complex language structures.
    b. Prefix for Tasks: In the context of T5, this approach also involves using a prefix that defines the task. For example, before the input sentence, a prefix like "summarize: " is added to inform the model of the type of output expected[9]. This approach provides additional context to the model, thereby improving the accuracy and relevance of the results generated.

3.3 5-Fold Cross Validation

After the preprocessing process is complete, the next step is 5-Fold Cross-Validation. This involves dividing the dataset into five folds. In each iteration, four folds are used to train the model, while one-fold is used as a validation set to test the model's performance during training. This process is repeated five times, with each fold serving as the validation set once and as part of the training folds four times. Once all five iterations are complete, the results from all folds are combined to obtain the model's average performance.

3.4 Model T5 Variants

The T5 model operates using a "text-to-text transfer learning" approach, where each NLP task is viewed as a transformation from input text to output text. By leveraging the self-attention mechanism, as shown in Figure 2, T5 can capture complex relationships between words within a sentence, allowing the model to apply varying degrees of attention to each word depending on the context. This self-attention mechanism is central to the Transformer architecture, enabling the model to focus on specific parts of the input while processing information. This enhances the model's ability to understand context and meaning.

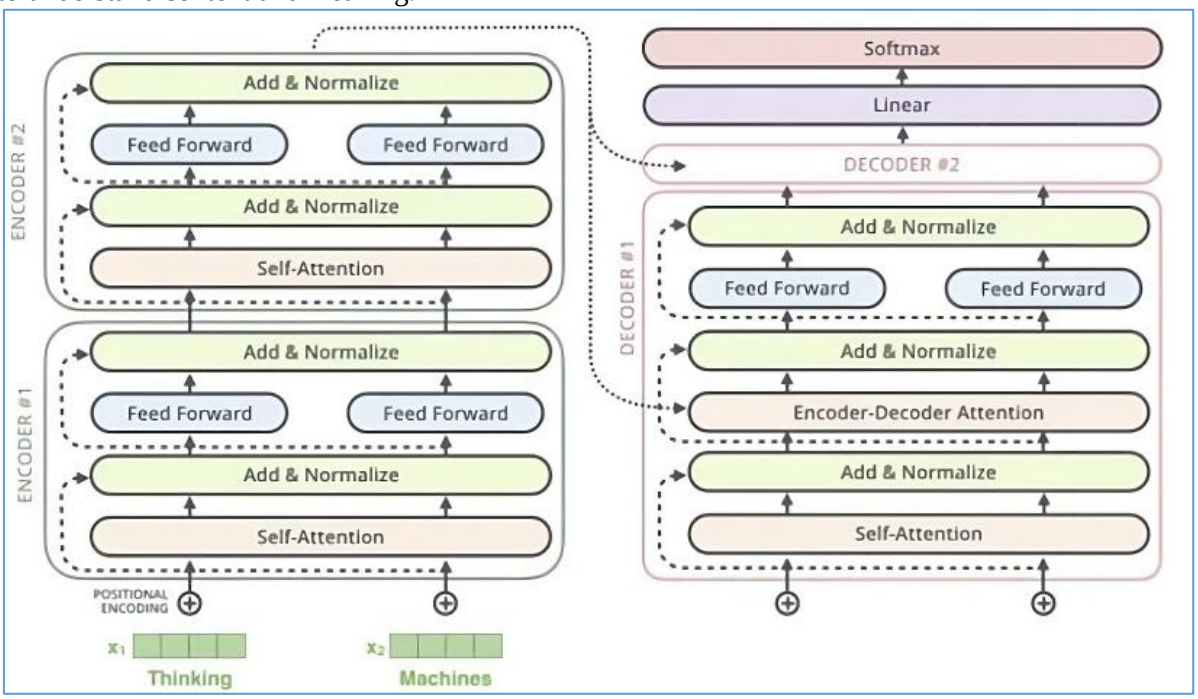

Figure 2. Transformer architecture

In this study, we used several T5 model variants to analyze their performance on summarization tasks. The models applied include T5-Base, FLAN-T5 Base, and mT5-Base. Below is a summary of each model along with its parameter count:

 ISSN:2302-8149
e-ISSN:2540-9719

1. T5-Base: This model has approximately 220 million parameters. T5-Base is designed to optimize various NLP tasks with a text-based approach.
2. FLAN-T5 Base: FLAN-T5 is a variant of T5 trained using a task instruction tuning approach and also has around 220 million parameters.
3. mT5-Base: This model is the multilingual version of T5 with the same parameter count, totaling 220 million. mT5 is designed to support multiple languages.

3.5 Experiment Setup

For this study, we experimented with three transformer-based models: T5, Flan-T5, and mT5. Several scenarios were designed to systematically evaluate their performance:

1. The performance of T5, Flan-T5, and mT5 was evaluated under identical training setups to identify the most effective model for the task.
2. Learning rates (0.0001 and 0.0003) were tested to optimize performance on the development set, ensuring each model was tuned for maximum effectiveness.

The selection of these hyperparameters was guided by their impact on model convergence and performance [20]. A learning rate of 0.0001 was chosen as a standard starting point for fine-tuning pretrained models, providing stability during gradient updates. Conversely, 0.0003 was selected to explore faster convergence, balancing the risk of overshooting optima. Batch size was set at 8 to accommodate computational constraints while maintaining sufficient gradient variability for effective learning.

Table 1 provides an overview of the hyperparameter configurations employed during training, which include the number of training steps, learning rate, batch size, and additional parameters. These hyperparameters were fine-tuned iteratively based on observed performance on the development subset, ensuring that each model was trained to achieve its maximum potential

Table 1. Hyperparameter value

Hyperparameter	Experiment 1 Value	Experiment 2 Value
Batch size	8	8
Learning rate	0.0001	0.0003
Optimizer	AdamW	AdamW
Max sequence length	512 tokens	512 tokens
Total training steps	10000	10000

These experiments aimed to understand how architectural differences and hyperparameter adjustments influence summarization performance. The training process is performed with early stopping to prevent overfitting. The model that performs best on dev is kept for further evaluation on the test subset.

3.6 Evaluation

The Recall-Oriented Understudy for Gisting Evaluation (ROUGE) is an evaluation method used to assess summarization systems automatically by comparing the generated summaries with reference summaries or human-crafted summaries[21]. ROUGE evaluation is reported in terms of precision, recall, and F1-score. Precision measures the proportion of n-grams present in the generated summary that also appear in the reference summary. The precision score can be calculated with the following formula:

$$Precision = \frac{number\ of\ overlapped\ words}{number\ of\ words\ in\ the\ model\ summary} \quad (1)$$

Recall measures the proportion of n-grams in the reference summary that are included in the resulting summary. The following formula calculates the recall value:

$$Recall = \frac{number\ of\ overlapped\ words}{word\ count\ in\ human\ summary} \quad (2)$$

F1-score is the average between precision and recall which provides a comprehensive evaluation measure[21]. The f-measure value is obtained from the following formula:

 ISSN:2302-8149
e-ISSN:2540-9719

$$F1 - score = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (3)$$

ROUGE uses several metrics that are used to standardize whether a summary is good or not, namely, ROUGE-N and ROUGE-L. ROUGE-N is used to compare n-grams between the generated summary and the reference summary [21]. ROUGE-1 focuses on unigrams (individual words) in this comparison, while ROUGE-2 compares bigrams (two consecutive words or a combination of two words). The formula to calculate ROUGE-N is as follows:

$$ROUGE - N = \frac{\sum_{gram_n \in referencesummar} Count_{match}(gram_n)}{\sum_{gram_n \in referencesummar} Count(gram_n)} \quad (4)$$

ROUGE-L measures the Longest Common Subsequence (LCS) based comparison between the generated summary and the reference summary. This metric assesses how well the generated summary preserves the order and structure of the reference summary[21]. The formulas for calculating ROUGE-L precision, recall, and f1-score are as follows:

$$Precision_{lcs} = \frac{LCS(X,Y)}{m} \quad (5)$$

$$Recall_{lcs} = \frac{LCS(X,Y)}{n} \quad (6)$$

LCS (X, Y) is the length of the longest common subsequence between X and Y, m is the number of tokens in the reference summary, and n is the number of tokens in the generated summary.

$$F_{lcs} = \frac{(1+\beta^2) \cdot R_{lcs} \cdot P_{lcs}}{R_{lcs} + \beta^2 \cdot P_{lcs}} \quad (7)$$

Where β (beta) is a parameter used to balance the contribution between precision and recall in the calculation of f1-score.

4 Results and Analysis

The results of this study demonstrate the good performance of the T5 model and its variants in generating automatic summaries from various text sources. The study also discusses broader implications for text analysis applications in digital contexts, providing a solid foundation for future research and development in summarization methodologies. Using evaluation metrics such as ROUGE, this study highlights the effectiveness of the transformer-based model approach in generating informative and relevant summaries, as well as its ability to handle text in multiple languages.

4.1 Data Selection

Table 2. Data row information

Column	Non-Null Count	Dtype
category	14263 non-null	object
gold_labels	14263 non-null	object
id	14263 non-null	object
paragraphs	14263 non-null	object
source	14263 non-null	object
source_url	14263 non-null	object
summary	14263 non-null	object

The data used in this research initially has several columns as seen in Table 2. These columns include 'category', 'gold_labels', 'id', 'paragraphs', 'source', 'source_url', and 'summary'. However, to focus on the text summary task, we only selected the 'paragraphs' and 'summary' columns. The 'paragraphs' column contains the main text that needs to be summarized, while the 'summary' column provides the target summary as a reference for model performance evaluation.

 ISSN:2302-8149
e-ISSN:2540-9719

```
Train Data - Paragraphs and Summary:
                                                  paragraphs  \
0  [[[Jakarta, ,, CNN, Indonesia, -, -, Timnas, A...
1  [[[Kebakaran, melanda, kawasan, pemukiman, di,...
2  [[[Jakarta, ,, CNN, Indonesia, -, -, Bek, Juve...
3  [[[Jakarta, ,, CNN, Indonesia, -, -, Predikat,...
4  [[[Jakarta, ,, CNN, Indonesia, -, -, Macet, me...

                                                 summary
0  [[Timnas, Argentina, ditahan, imbang, ketika, ...
1  [[Kebakaran, melanda, kawasan, pemukiman, di, ...
2  [[Bek, Juventus, ,, Giorgio, Chiellini, ,, mem...
3  [[Tahun, ini, pendaftar, kategori, Film, Anima...
4  [[Benjamin, David, (, 40, ), ,, warga, Munich,...
```

Figure 3. Selection row data

After the data selection process, only the 'paragraphs' and 'summary' columns are used, as shown in Figure 4. By simplifying the data into these two columns, we can reduce irrelevant information and speed up the data processing process.

4.2 Data Pre-Processing

After the data selection process, the next step is pre-processing which includes several important steps to prepare the data before it is used in the summary model. As shown in Figure 5, we apply lower casing to standardize all text to lowercase, which aims to reduce processing complexity and avoid differences in recognition of uppercase and lowercase letters by the model.

	paragraphs	summary
0	summarize: jakarta cnn indonesia timnas argent...	timnas argentina ditahan imbang ketika berhada...
1	summarize: kebakaran melanda kawasan pemukiman...	kebakaran melanda kawasan pemukiman di jalan k...
2	summarize: jakarta cnn indonesia bek juventus ...	bek juventus giorgio chiellini memuji gonzalo ...
3	summarize: jakarta cnn indonesia predikat film...	tahun ini pendaftar kategori film animasi terb...
4	summarize: jakarta cnn indonesia macet menjadi...	benjamin david 40 warga munich jerman selama b...

Figure 4. Preprocessed row

In addition, we added the prefix “summarize:” to the 'paragraphs' field to signal to the model that the task to be performed is a text summary.

4.3 5-Folding

To ensure a robust evaluation of model performance, we implemented a 5-Fold Cross-Validation strategy. This approach divides the dataset into five subsets, where in each iteration, four subsets are used for training, and one subset is reserved for validation (dev) or testing. This method enhances model generalization and minimizes overfitting.

In our setup, the dataset is split into 80% training, 5% development (dev), and 15% test data for each fold. Table 3 presents the distribution of data across the five folds.

 ISSN:2302-8149
e-ISSN:2540-9719

Table 3. Article per fold

	train	dev	test
Fold 1	14262	750	3762
Fold 2	14263	749	3762
Fold 3	14290	747	3737
Fold 4	14272	750	3752
Fold 5	14266	747	3761

By systematically rotating the validation and test sets across different iterations, we ensure that every data point is used for both training and evaluation, leading to a more reliable performance assessment.

4.4 Training Process

To ensure robust evaluation and comprehensive dataset coverage, we employed a fold-based evaluation methodology. This approach systematically divides the dataset into multiple folds, each containing three distinct subsets: training, validation, and testing. The training subset is used for primary model development, allowing the model to learn patterns and representations from the data. The development subset is allocated for fine-tuning hyperparameters and preventing overfitting by providing performance feedback during training. Lastly, the testing subset is strictly reserved for assessing the model's generalization capability on unseen data. By independently training models on each fold, this methodology ensures that every data point is used for both training and testing, yielding an exhaustive evaluation of model performance across the entire dataset.

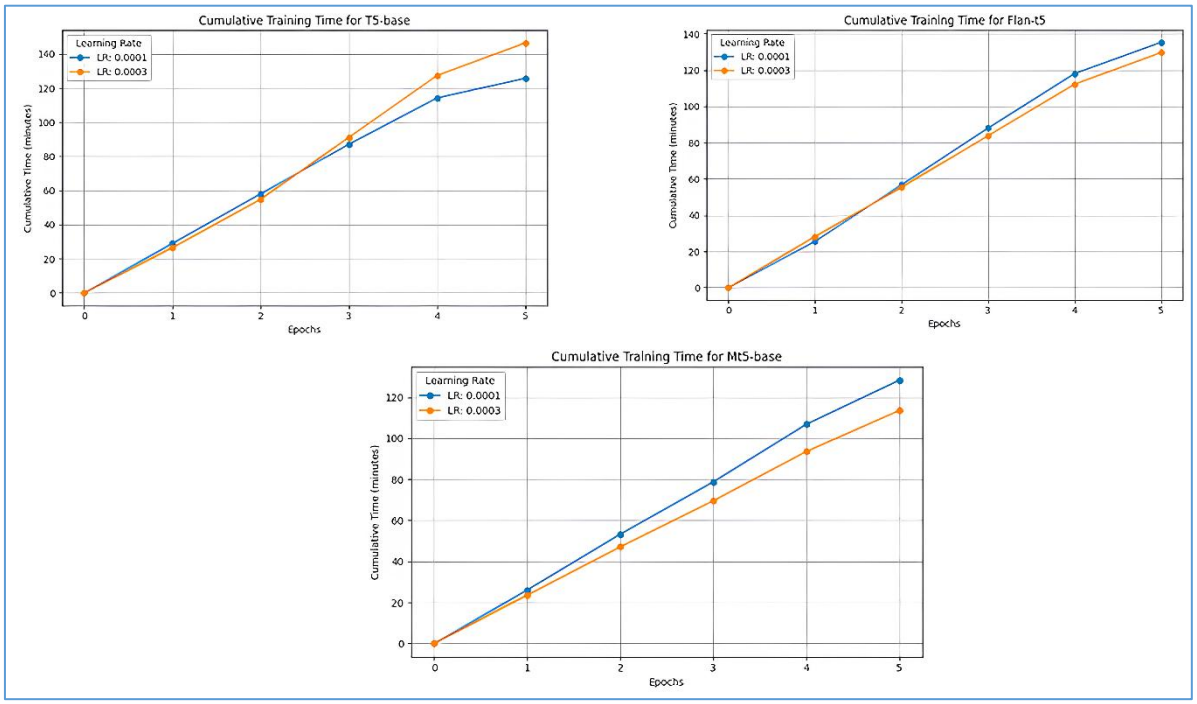

Figure 5 Training Time each Model for 1-fold

The training efficiency of the models—T5-base, Flan-T5, and mT5-base—is illustrated in the cumulative training time plots (Figure 6), showing the impact of different learning rates (0.0001 and 0.0003) over five epochs for one-fold of data. The T5-base model exhibits the longest cumulative training time, requiring approximately 140 minutes by the fifth epoch when trained with a learning rate of 0.0001. Similarly, Flan-T5 also reaches around 140 minutes under the same conditions, demonstrating comparable computational demands between these two models. In contrast, mT5-base consistently requires less training time, with cumulative durations staying under 120 minutes by the fifth epoch. Interestingly, mT5 exhibits faster convergence at the higher learning rate of 0.0003, completing training more quickly than with 0.0001. This behavior suggests that mT5 may benefit

 ISSN:2302-8149
e-ISSN:2540-9719

from higher learning rates, leveraging its multilingual pretraining to accelerate optimization. While T5 and Flan-T5 demand longer training times, their architectural robustness and fine-tuning mechanisms ensure superior summarization performance, albeit at the cost of higher computational resources. On the other hand, mT5's faster training at 0.0003 highlights its potential for tasks requiring rapid adaptation, but its summarization accuracy remains lower in comparison. This trade-off between computational efficiency and performance underscores the importance of selecting models and hyperparameters based on specific task requirements and resource constraints.

4.5 Summarization Analysis

Analysis of model performance was conducted quantitatively and qualitatively to evaluate the ability of the T5, Flan-T5, and mT5 models to summarize text in Indonesian. Each model was tested to understand its strengths and limitations in producing concise, accurate, and relevant summaries.

4.5.1 Quantitative Analysis

Table 4. Comparison of average ROUGE values

Model	Metrics Evaluation (Average 5-Fold)		
	ROUGE 1	ROUGE 2	ROUGE L
NEURALSUM[18]	46	48	47
ALBERT[22]	45.28	40.77	44.39
BERT2GPT[23]	62	56	60
Our Research			
$T5_{(lr: 0.0003)}$	72.43	63.05	68.48
$Flan\text{-}T5_{(lr: 0.0003)}$	66.98	54.88	62.39
$mT5_{(lr: 0.0003)}$	58.61	46.11	52.80
$T5_{(lr: 0.0001)}$	**73.51**	**64.49**	**69.54**
$Flan\text{-}T5_{(lr: 0.0001)}$	72.00	61.94	67.68
$mT5_{(lr: 0.0001)}$	58.13	45.09	52.38

As shown in Table 4, the performance of the T5, Flan-T5, and mT5 models is evaluated using ROUGE-1, ROUGE-2, and ROUGE-L metrics, averaged over a 5-fold cross-validation. The T5 model with a learning rate of 0.0001 achieves the highest scores, with ROUGE-1, ROUGE-2, and ROUGE-L values of 73.51%, 64.49%, and 69.54%, respectively, demonstrating its superior ability to generate coherent and accurate summaries.

Flan-T5, trained with the same learning rate, follows closely with scores of 72.00%, 61.94%, and 67.68% for ROUGE-1, ROUGE-2, and ROUGE-L, respectively. Although slightly lower than T5, these results indicate strong summarization capabilities, albeit with minor limitations in linguistic fluency and detail.

In contrast, the multilingual mT5 model performs significantly worse at a learning rate of 0.0001, with ROUGE-1, ROUGE-2, and ROUGE-L scores of 58.13%, 45.09%, and 52.38%, respectively. However, when trained with a higher learning rate of 0.0003, mT5 achieves better scores of 58.61%, 46.12%, and 52.80% for ROUGE-1, ROUGE-2, and ROUGE-L, respectively, indicating its potential to benefit from optimized hyperparameter settings. Despite this improvement, mT5 still lags significantly behind T5 and Flan-T5, suggesting challenges in summarizing Indonesian texts due to its generalized multilingual training, which may dilute its focus on specific linguistic features.

Furthermore, the training efficiency of these models is illustrated in the cumulative training time plots (Figure 6). The graphs show that both T5 and Flan-T5 require slightly longer training times compared to mT5, regardless of the learning rate. For instance, at the 0.0001 learning rate, T5 and Flan-T5 reach approximately 140 minutes by the fifth epoch, while mT5 requires less than 120 minutes. This indicates a trade-off between training time and summarization performance, where T5 and Flan-T5 exhibit superior accuracy at the cost of marginally higher computational demands.

4.5.2 Qualitative Analysis

To provide a deeper understanding of the summarization performance across models, we conducted a qualitative analysis of the summaries generated by T5, Flan-T5, and mT5 for an Indonesian-language text. Table 5 and Table 6 presents example outputs of each model, along with a summary of references for comparison.

 ISSN:2302-8149
e-ISSN:2540-9719

Table 5. Output of summarized result

Model	Summary Result
Article	Jerussalem (ANTARA News)-Menteri Energi Israel, Yuval Steinitz, Jumat (Sabtu WIB), mengatakan, Amerika Serikat telah meninggalkan Israel menyusul sikap abstain Amerika Serikat dalam Resolusi Dewan Keamanan PBB yang menyatakan pembangunan pemukiman Israel di wilayah yang mereka duduki adalah ilegal. Akibat keputusan Amerika Serikat yang langka itu karena tidak menggunakan hak veto, Dewan Keamanan PBB pun secara bulat meminta agar Israel menghentikan pembangunan pemukiman di wilayah Palestina yang mereka duduki. " Ini jelas bukan suatu resolusi menentang pemukiman, tapi resolusi anti-Israel, menentang negara Yahudi dan masyarakat Yahudi. Amerika Serikat malam ini telah meninggalkan satu-satu sekutunya di Timur Tengah, " kata Steinitz yang dikenal dekat dengan Perdana Menteri Israel, Benjamin Netanyahu.
Reference Summary	Menteri Energi Israel Yuval Steinitz mengatakan, Amerika Serikat telah meninggalkan Israel menyusul sikap abstain Amerika Serikat dalam Resolusi Dewan Keamanan PBB. Sikap tersebut berkaitan dengan pembangunan pemukiman Israel di wilayah yang mereka duduki adalah ilegal. Perdana Menteri Israel, Benjamin Netanyahu mengatakan bahwa itu bukan suatu resolusi menentang pemukiman, tapi resolusi anti-Israel, menentang negara Yahudi dan masyarakat Yahudi.
T5 Model Summary	METERI Energi Israel, Yuval Steinitz, Jumat (Sabtu WIB), mengatakan, Amerika Serikat telah meninggalkan Israel menyusul sikap abstain Amerika Serikat dalam Resolusi Dewan Keamanan PBB yang menyatakan pembangunan pemukiman Israel di wilayah yang mereka duduki adalah ilegal.
Flan-T5 Model Summary	Amerika Serikat telah meninggalkan Israel menyusul sikap abstain Amerika Serikat dalam Resolusi Dewan Keamanan PBB yang menyatakan pembangunan pemukiman Israel di wilayah yang mereka duduki adalah ilegal yang menghentikan pembangunan pemukiman Israel di wilayah Palestina yang mereka duduki adalah
mT5 Model Summary	menurut sekjen israel Benjamin Netanyahu sikap abstain amerika Serikat dalam resolusi dewan keamanan pBB yang menyatakan pembangunan pemukiman di wilayah yang mereka duduki adalah ilegal Steinitz mengatakan bahwa keputusan abstain itu tidak menggunakan hak veto dan meminta agar Israel menghentikan pembangunan pemukiman di Palestina yang mereka duduki

The T5 model captures the main ideas and correctly highlights details such as the stance of Israel's Energy Minister, Yuval Steinitz, and the United Nations Security Council's resolution. However, there are issues in formatting and redundancy. For instance, the T5 model output includes the word "MENTERI" in all capital letters, which is inconsistent with standard capitalization conventions in Indonesian and may indicate an error in tokenization or preprocessing. This inconsistency, combined with redundant phrasing, makes the T5 summary less polished and slightly verbose. Despite these drawbacks, T5 achieves a higher ROUGE score than Flan-T5. However, its failure to capture more nuanced aspects—like Netanyahu's specific stance—indicates that T5 may struggle with contextual depth.

Flan-T5, tuned for following structured instructions, demonstrates an improved ability to summarize complex, context-rich texts but encounters a significant truncation issue, stopping at the word "adalah" (meaning "is"), as shown in Table 5. This truncation suggests the summary was cut off prematurely, impacting its coherence and completeness. The truncation likely stems from an insufficient token limit, which restricts the model's ability to complete the summary. Consequently, Flan-T5 scores lower on ROUGE compared to T5, reflecting the lack of comprehensive lexical overlap with the reference summary due to this cutoff. Adjusting Flan-T5's token limits could mitigate such issues in similar tasks.

 ISSN:2302-8149
e-ISSN:2540-9719

The mT5 model, designed for multilingual tasks, performs relatively well, capturing the central topic on Israeli settlements. It references Benjamin Netanyahu but inaccurately labels him as the “sekjen israel” (Secretary of Israel) instead of Prime Minister, which shifts the factual accuracy of the summary. This inconsistency implies that mT5 may face challenges with nuanced details in lower-resource languages like Indonesian, potentially due to limited high-quality data for fine-tuning.

Table 6. Another output of summarized result

Model	Summary Result
Article	Jakarta, CNN Indonesia-- Kematian vokalis Linkin Park, Chester Bennington masih menyisakan duka bagi para penggemarnya di seluruh dunia. Para fan bersimpati dengan melakukan berbagai kegiatan untuk memberikan penghormatan kepada sang idola. Linkin Park merangkum doa dan dukungan para penggemar itu dengan mengunggah sebuah video di YouTube tepat di hari ke-50 kematian Bennington. Dari ratusan peringatan di dunia, peringatan di Semarang, Indonesia menjadi sorotan dalam video Linkin Park bertajuk Chester Bennington-Memorials Around the World. Mengenang Bennington di Semarang dilakukan oleh sekelompok pemuda yang kompak mengenakan pakaian berwarna hitam. Di sebuah taman, mereka menyalakan lilin sambil melantunkan nyanyian untuk Bennington, dengan diiringi petikan gitar. Indonesia menjadi sedikit negara yang masuk dalam video Linkin Park. Nama Indonesia bersanding dengan Peru, Brasil, Meksiko, Jerman, Amerika Serikat, Yunani, Belarusia, Filipina, Kazakhstan, Rusia, China, Chille, Prancis dan Belanda. Di Indonesia sendiri sebenarnya perayaan tak hanya di lakukan di Semarang, tapi juga Jakarta, Bandung dan Surabaya. Tak diketahui apa alasan Linkin Park memilih Semarang dan beberapa kota lainnya di dunia dalam video itu. " Terima kasih untuk para penggemar di seluruh dunia atas curahan cinta dan dukungan kalian semua, " tulis Linkin Park di penghujung video. Cuplikan video itu juga diunggah ke akun resmi Linkin Park di Instagram. Chester Bennington memutuskan untuk mengakhiri hidupnya dengan bunuh diri di rumahnya di Los Angeles, Amerika Serikat pada 20 Juli lalu. Selain video ini, Linkin Park pun berencana membuat acara khusus mengenang rekan sejawatnya itu. Mereka sedang menggarap sebuah proyek di Los Angeles. Informasi rinci tentang acara itu bakal diumumkan dalam waktu dekat.
Reference Summary	Linkin Park merangkum doa dan dukungan para penggemar itu dengan mengunggah sebuah video di YouTube tepat di hari ke-50 kematian Bennington. Dari ratusan peringatan di dunia, peringatan di Semarang menjadi sorotan dalam video yang bertajuk Chester Bennington - Memorials Around the World. Mengenang Bennington di Semarang dilakukan sekelompok pemuda yang mengenakan pakaian hitam dengan berkumpul di sebuah taman, mereka menyalakan lilin.
T5 Model Summary	Kematian vokalis Linkin Park, Chester Bennington masih menyisakan duka bagi para penggemarnya di seluruh dunia. Para fan bersimpati dengan melakukan berbagai kegiatan untuk memberikan penghormatan kepada sang idola.
Flan-T5 Model Summary	Kematian vokalis Linkin Park, Chester Bennington masih menyisakan duka bagi para penggemarnya di seluruh dunia. Para fan bersimpati dengan melakukan berbagai kegiatan untuk memberikan penghormatan kepada sang idola Linkin Park merangkum doa dan dukungan para penggemar itu dengan mengunggah sebuah video di YouTube
mT5 Model Summary	kematian vokalis Linkin Park, Chester Bennington masih menyisakan duka bagi para penggemarnya di seluruh dunia dengan mengunggah sebuah video di YouTube tepat di hari ke-50 kematian Bennington. Dari ratusan peringatan di dunia, peringatan di Semarang menjadi sorotan dalam video Linkin Park bertajuk Chester Bennington Memorials Around the Worldsyngton Around the World

 ISSN:2302-8149
e-ISSN:2540-9719

In the additional Table 6 provided, the T5, Flan-T5, and mT5 models demonstrate an improved alignment with the reference summary, though minor errors remain. The T5 model effectively captures the primary content and context, accurately highlighting the public memorial events for Chester Bennington around the world. However, it slightly misses the mention of the memorial in Semarang. Despite these minor issues, T5 achieves a high ROUGE score, reflecting strong alignment with the reference summary. The Flan-T5 model also performs well in terms of coherence and structure, summarizing the key aspects of the memorial events with a clear and concise approach. Unlike the previous example, the truncation issue has been addressed, resulting in a completer and more cohesive summary.

The mT5 model, designed for multilingual applications, demonstrates strong summarization abilities and correctly captures the central theme and relevant details. However, it makes a translation error, inaccurately rendering "Memorials Around the World" as "Around the Worldsyngton Around the World".

Overall, this study highlights each model's ability to generate high-quality summaries with minor lexical or translation errors. The T5 model excels in content coverage but requires fine-tuning to reduce verbosity. Flan-T5, although generally cohesive, could benefit from adjustments to minimize redundancy. Finally, mT5 shows promising multilingual capability but requires additional language-specific tuning to handle idiomatic phrases and contextually nuanced terms accurately. Although T5 achieves the highest ROUGE scores, its tendency toward verbosity and redundancy aligns with previous studies that emphasize the trade-off between fluency and brevity in transformer-based summarization models. This observation is consistent with earlier research on English-language summarization tasks using T5, where high recall scores were often accompanied by excessive length in generated summaries. In contrast, Flan-T5, designed for improved instruction-following, demonstrates a structured summarization style but encounters truncation issues similar to findings in prior work on instruction-tuned models, where strict adherence to training patterns occasionally limits flexibility. Meanwhile, mT5's performance indicates a persistent challenge in adapting multilingual models for low-resource languages like Indonesian, echoing previous research that noted a decline in performance when these models are not fine-tuned on sufficient domain-specific data.

5 Conclusion

This study provides a comprehensive evaluation of T5 model variants for Indonesian text summarization, highlighting key findings, practical implications, and areas for improvement. T5-Base outperforms FLAN-T5 and mT5 across all ROUGE metrics, achieving the highest scores of 73.52%, 64.50%, and 69.55% for ROUGE-1, ROUGE-2, and ROUGE-L, respectively. While T5-Base excels in capturing key ideas, it struggles with redundancy and formatting. FLAN-T5 generates structured summaries but suffers from truncation issues, while mT5 demonstrates multilingual capability but lacks context-specific accuracy in Indonesian.

These findings have practical implications for automated summarization applications, particularly in news and information retrieval. However, the study has limitations, including potential dataset bias toward news articles and reliance on ROUGE scores, which may not fully capture semantic coherence. Additionally, training strategies such as reinforcement learning and prompt tuning remain unexplored.

Future research should address these gaps by incorporating more diverse datasets, human evaluations, and advanced fine-tuning techniques. Hybrid approaches combining extractive and abstractive methods could further enhance summarization quality. Additionally, optimizing multilingual models like mT5 for Indonesian remains a crucial avenue for improvement. This study contributes to advancing high-quality automated summarization in Indonesian, paving the way for more effective language models.

Reference

[1] D. Qiu and B. Yang, "Text summarization based on multi-head self-attention mechanism and pointer network," *Complex Intell. Syst.*, vol. 8, no. 1, pp. 555–567, Feb. 2022, doi: 10.1007/s40747-021-00527-2.

 ISSN:2302-8149
e-ISSN:2540-9719

[2] D. Bawden and L. Robinson, "Information Overload: An Introduction," in *Oxford Research Encyclopedia of Politics*, Oxford University Press, 2020. doi: 10.1093/acrefore/9780190228637.013.1360.
[3] C. Setyawan, N. Benarkah, and V. R. Prasetyo, "Automatic Text Summarization Berdasarkan Pendekatan Statistika pada Dokumen Berbahasa Indonesia," *Keluwih J. Sains Dan Teknol.*, vol. 2, no. 1, Art. no. 1, Feb. 2021, doi: 10.24123/saintek.v2i1.4045.
[4] W. Widodo, M. Nugraheni, and I. P. Sari, "A comparative review of extractive text summarization in Indonesian language," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 1098, no. 3, p. 032041, Mar. 2021, doi: 10.1088/1757-899X/1098/3/032041.
[5] W. S. El-Kassas, C. R. Salama, A. A. Rafea, and H. K. Mohamed, "Automatic text summarization: A comprehensive survey," *Expert Syst. Appl.*, vol. 165, p. 113679, Mar. 2021, doi: 10.1016/j.eswa.2020.113679.
[6] N. Giarelis, C. Mastrokostas, and N. Karacapilidis, "Abstractive vs. Extractive Summarization: An Experimental Review," *Appl. Sci.*, vol. 13, no. 13, p. 7620, Jun. 2023, doi: 10.3390/app13137620.
[7] A. P. Widyassari *et al.*, "Review of automatic text summarization techniques & methods," *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 34, no. 4, pp. 1029–1046, Apr. 2022, doi: 10.1016/j.jksuci.2020.05.006.
[8] C. Raffel *et al.*, "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer," Sep. 19, 2023, *arXiv*: arXiv:1910.10683. Accessed: Oct. 31, 2024. [Online]. Available: http://arxiv.org/abs/1910.10683
[9] G. S. Ramesh, V. Manyam, V. Mandula, P. Myana, S. Macha, and S. Reddy, "Abstractive Text Summarization Using T5 Architecture," in *Proceedings of Second International Conference on Advances in Computer Engineering and Communication Systems*, A. B. Reddy, B. V. Kiranmayee, R. R. Mukkamala, and K. Srujan Raju, Eds., in Algorithms for Intelligent Systems. , Singapore: Springer Nature Singapore, 2022, pp. 535–543. doi: 10.1007/978-981-16-7389-4_52.
[10] A. Garg *et al.*, "NEWS Article Summarization with Pretrained Transformer," in *Advanced Computing*, vol. 1367, D. Garg, K. Wong, J. Sarangapani, and S. K. Gupta, Eds., in Communications in Computer and Information Science, vol. 1367. , Singapore: Springer Singapore, 2021, pp. 203–211. doi: 10.1007/978-981-16-0401-0_15.
[11] A. Urlana, S. M. Bhatt, N. Surange, and M. Shrivastava, "Indian Language Summarization using Pretrained Sequence-to-Sequence Models," Mar. 25, 2023, *arXiv*: arXiv:2303.14461. Accessed: Oct. 31, 2024. [Online]. Available: http://arxiv.org/abs/2303.14461
[12] Y. Liu and M. Lapata, "Text Summarization with Pretrained Encoders," Sep. 05, 2019, *arXiv*: arXiv:1908.08345. Accessed: Aug. 19, 2024. [Online]. Available: http://arxiv.org/abs/1908.08345
[13] M. Ramina, N. Darnay, C. Ludbe, and A. Dhruv, "Topic level summary generation using BERT induced Abstractive Summarization Model," in *2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS)*, Madurai, India: IEEE, May 2020, pp. 747–752. doi: 10.1109/ICICCS48265.2020.9120997.
[14] Q. A. Itsnaini, M. Hayaty, A. D. Putra, and N. A. M. Jabari, "Abstractive Text Summarization using Pre-Trained Language Model 'Text-to-Text Transfer Transformer (T5),'" *Ilk. J. Ilm.*, vol. 15, no. 1, pp. 124–131, Apr. 2023, doi: 10.33096/ilkom.v15i1.1532.124-131.
[15] G. E. Abdul, I. A. Ali, and C. Megha, "Fine-Tuned T5 for Abstractive Summarization," *Int. J. Perform. Eng.*, vol. 17, no. 10, p. 900, 2021, doi: 10.23940/ijpe.21.10.p8.900906.
[16] Y. Ding, Y. Qin, Q. Liu, and M.-Y. Kan, "CocoSciSum: A Scientific Summarization Toolkit with Compositional Controllability," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Singapore: Association for Computational Linguistics, 2023, pp. 518–526. doi: 10.18653/v1/2023.emnlp-demo.47.
[17] V. N. Mahmoodabadi and F. Ghasemian, "Persian Text Summarization via Fine Tuning mT5 Transformer," 2023.
[18] K. Kurniawan and S. Louvan, "Indosum: A New Benchmark Dataset for Indonesian Text Summarization," *Int. Conf. Asian Lang. Process. IALP*, pp. 215–220, 2018, doi: 10.1109/IALP.2018.8629109.

 ISSN:2302-8149
e-ISSN:2540-9719

[19] S. Mehta, D. Shah, R. Kulkarni, and C. Caragea, "Semantic Tokenizer for Enhanced Natural Language Processing," Apr. 24, 2023, *arXiv*: arXiv:2304.12404. Accessed: Oct. 31, 2024. [Online]. Available: http://arxiv.org/abs/2304.12404

[20] X. Wang, W. Tian, and Z. Liao, "Framework for Hyperparameter Impact Analysis and Selection for Water Resources Feedforward Neural Network," *Water Resour. Manag.*, vol. 36, no. 11, pp. 4201–4217, Sep. 2022, doi: 10.1007/s11269-022-03248-4.

[21] R. Yacouby and D. Axman, "Probabilistic Extension of Precision, Recall, and F1 Score for More Thorough Evaluation of Classification Models," in *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems*, Online: Association for Computational Linguistics, 2020, pp. 79–91. doi: 10.18653/v1/2020.eval4nlp-1.9.

[22] A. L. Putra, Sanjaya, M. R. Fachruradzi, and A. Y. Zakiyyah, "Resource Efficient Abstractive Text Summarization in Indonesian with ALBERT," in *2024 International Conference on Smart Computing, IoT and Machine Learning (SIML)*, Surakarta, Indonesia: IEEE, Jun. 2024, pp. 81–85. doi: 10.1109/SIML61815.2024.10578175.

[23] M. Nasari, A. Maulina, and A. S. Girsang, "Abstractive Indonesian News Summarization Using BERT2GPT," in *2023 IEEE 7th International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE)*, Purwokerto, Indonesia: IEEE, Nov. 2023, pp. 369–375. doi: 10.1109/ICITISEE58992.2023.10405359.