Leveraging BERT and T5 for Comprehensive Text Summarization on Indonesian Articles

Mohammad Wahyu Bagus Dwi Satya
Informatics Engineering
Faculty of Computer Science
Universitas Dian Nuswantoro
Semarang, Indonesia
111202113441@mhs.dinus.ac.id

Ardytha Luthfiarta
Informatics Engineering
Faculty of Computer Science
Universitas Dian Nuswantoro
Semarang, Indonesia
ardytha.luthfiarta@dsn.dinus.ac.id

Abstract— One of the main challenges in the field of Natural Language Processing (NLP) is developing systems for automatic text summarization. These systems typically fall into two categories: extractive and abstractive. Extractive techniques generate summaries by selecting important sentences or phrases directly from the original text, whereas abstractive techniques focus on rephrasing or paraphrasing the content, producing summaries that resemble human-written ones. In this research, models based on Transformer architectures, including BERT and T5, were used, which have been shown to effectively summarize texts in various languages, including Indonesian. The dataset used was INDOSUM, consisting of Indonesian news articles. The best results were achieved with the T5 model, using the abstractive approach, recorded ROUGE-1, ROUGE-2, and ROUGE-L scores of 69.36%, 61.27%, and 66.17%, respectively. On the other hand, the extractive BERT model achieved ROUGE-1, ROUGE-2, and ROUGE-L scores of 70.82%, 63.99%, and 58.40%.

Keywords—Text Summarization, Bahasa Indonesia, Indosum, Pretrained Model

I. INTRODUCTION

In the digital era marked by technological advancements, the ability to quickly absorb and understand information is becoming increasingly crucial. Every day, the internet is filled with millions of documents, articles, and other content, creating new challenges for users in filtering and extracting relevant and meaningful information. This problem can reduce the efficiency of processing information, often leading to boredom and loss of focus[1].

Automatic text summarization has emerged as a promising solution to address these challenges[2]. This approach aims to create a condensed version of a more extensive text, while still preserving its key ideas and meaning, allowing users to quickly understand the essential information without going through the whole document.

Automatic text summarization in natural language processing fall into two principal categories. The first approach, known as extractive summarization, operates by identifying and preserving crucial elements from the source document. This method creates a distilled version of the original text by compiling its most salient components. The second category adopts a more generative approach. This abstractive method entails synthesizing novel content that encapsulates the core information of the source material. In contrast to its extractive counterpart, this technique may generate summaries comprising linguistic constructions and phraseology not explicitly present in the original document[3].

Despite advancements in Automatic Text Summarization, several challenges persist, particularly in the context of the Indonesian language. One significant challenge is the morphological complexity of Indonesian. Unlike English, which has a relatively fixed word structure, Indonesian employs extensive affixation, including prefixes, suffixes, infixes, and circumfixes, to modify word meanings [4]. This complexity increases the difficulty of accurate text segmentation and summarization, as different morphological variations of the same root word can have distinct syntactic roles and meanings.

Additionally, Indonesian syntax is more flexible compared to English. While English predominantly follows a subject-verb-object (SVO) structure, Indonesian allows variations such as subject-object-verb (SOV) and verb-subject-object (VSO), depending on context, emphasis, and formality. This syntactic flexibility presents a challenge for summarization models, as it can affect coherence and readability when reconstructing sentences [5].

Another challenge is the high level of redundancy in Indonesian discourse. Indonesian often employs synonymous phrases or repetitive structures for emphasis, which can lead to unnecessarily verbose summaries if not properly handled [6]. Unlike English, which tends to be more concise, Indonesian frequently reiterates key information to reinforce meaning. Summarization models must effectively filter out redundant expressions while preserving the main ideas.

Furthermore, Indonesian lacks strict grammatical markers for tense, unlike English, which uses verb conjugations to indicate past, present, or future actions [7]. Instead, temporal information in Indonesian is conveyed through adverbs or contextual cues, making it more difficult for summarization models to accurately infer chronological sequences and maintain temporal coherence within a summary.

While both summarization methodologies aim to condense text, they differ significantly in their treatment of source material. Extractive techniques preserve the original wording, whereas abstractive methods prioritize conveying the essence of the text, potentially sacrificing verbatim reproduction for conciseness and clarity.

Various studies have investigated automatic text summarization using both extractive and abstractive methods on news articles and documents. In extractive summarization, approaches utilizing pretrained encoders, like BERT (Bidirectional Encoder Representations from Transformers), have been applied to datasets such as 'Cnn/DailyMail', achieving scores of 43.23%, 20.24%, and 39.63% on the ROUGE-1, ROUGE-2, and ROUGE-L metrics, respectively[8]. On the other hand, abstractive summarization research using transformer models, including BERT, on datasets such as those from Wikipedia and The Hindu news outlet, reported corresponding scores of 41.72%, 19.39%, and 38%.[9].

Although both approaches are promising in text summarization, the results can still be improved, especially in terms of accuracy and coherence. Based on our observations, the transformer architecture, with its self-attention mechanism, shows superior performance in summarizing large texts[10]. Therefore, we will apply the transformer architecture with pre-trained language models (PTLMs) on large-scale data. As for the extractive approach, we try to use BERT on Indonesian documents, while for the abstractive approach, we try to use T5 on Indonesian documents. This research seeks to demonstrate the effectiveness of these models in performing automatic text summarization (ATS) for Indonesian online news, with the goal of generating more accurate and coherent summaries.

II. METHODOLOGY

A. Automatic Text Summarization (ATS)

Advancements in the field of automatic text summarization (ATS) continue to make significant strides, as demonstrated by the growing body of research in this area [2], [11]. ATS can be divided into two main approaches: extractive and abstractive. Extractive methods typically employ various natural language processing techniques. For instance, one study using an Indonesian news dataset applied the Term Frequency-Inverse Document Frequency (TF-IDF) method in conjunction with Clustering[12]. The study found that this combination produced more cohesive summaries, achieving ROUGE-1, ROUGE-2, and ROUGE-L scores of 49.37%, 38.18%, and 46.87%, respectively.

For the abstractive approach, another study employed the Transformer architecture to summarize English news articles. Specifically, the researchers utilized a pre-trained T5

(Text to Text Transfer Transformer) model, which was further fine-tuned using a news dataset to enhance its summarization capabilities. The evaluation results demonstrated a notable improvement in summary quality, achieving ROUGE-1, ROUGE-2, and ROUGE-L scores of 43.02%, 14.50%, and 37.43%, respectively, representing a significant enhancement compared to the baseline model without fine-tuning[13].

B. Experiment

The following is an overview of the research methodology that will be applied for automatic text summarization, shown in Figure 1.

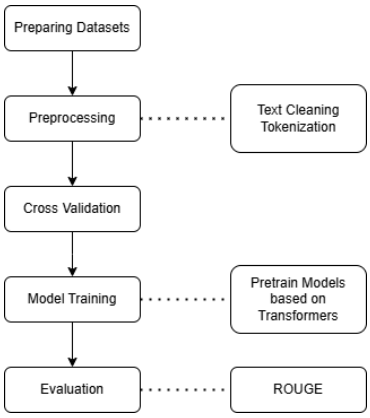

Figure 1 Research Flowchart

1) Preparing Datasets

This study utilizes the INDOSUM dataset [14], which is an Indonesian language dataset specifically designed for automatic text summarization. This dataset contains a collection of news texts accompanied by gold summaries. Overall, the dataset includes 19 thousand pairs of news articles and their summaries. The articles in this dataset are divided into six categories, such as sports, entertainment, inspiration, technology, headlines, and the entertainment industry. However, this study does not take into account the differences between these categories, so all news categories are treated equally.

2) Preprocessing

The preprocessing methods applied in this study include: Text Cleaning: This process involves cleaning text from unwanted characters. Text cleaning steps include changing all letters to lower case to match the text format, as well as filtering text using certain regex patterns, such as ([a-z0-9- +,.\/\\]*[a-z0-9][a-z0-9-+]*), to remove characters that are irrelevant or do not match the desired pattern. After that, Tokenization is a crucial process in text processing for machine learning models such as BERT and T5 that will be used for this study.

In the BERT model, tokenization is done through the WordPiece method [15], which breaks text into subword tokens based on their frequency of occurrence in the training corpus. The model also implements special tokens such as [CLS] for classification representation and [SEP] to separate sentences or paragraphs. For example, the sentence "The quick brown fox" will be broken into tokens such as [CLS], The, qu, ick, brown, fox, and [SEP].

T5 uses Sentencepiece with the Unigram Tokenization method, divides the text into subword tokens based on the probabilistic model and add prefix such as "summarize:". Special tokens including </s> for end of document and <unk> for unrecognized tokens [16]. For example, the sentence "The quick brown fox" can be tokenized as "The", "qu", "ick", "brown", "fox", followed by a final token and a prefix.

3) K-Fold Cross Validation

Following data preparation and refinement, our methodology advances to the implementation of 5-Fold Cross-Validation. This technique divides the complete dataset into five equal segments. Each iteration utilizes four segments to train the model, while the remaining segment acts as a validation set, assessing the model's performance during the training process. Upon completion of all iterations, we aggregate the performance metrics from each fold, calculating their average. This approach yields a more robust and reliable evaluation of the model's summarization capabilities on new, unseen text. By employing this method, we mitigate result variability and obtain a more dependable measure of the model's efficacy[17].

4) Transformer

a) Extractive Summarization with Transformers.

In extractive summarization, Transformer architecture is used to select important sentences or phrases directly from the source text to form a summary. The process starts with an encoder that receives news tokens as input and applies a self-attention mechanism to calculate the inter-relationships between tokens in the document[18]. Transformer models, like BERT, calculate an importance score for each token or sentence based on the attention generated by multi-head attention[19]. The sentence or phrase with the highest score is considered the most representative of the entire document and is selected for inclusion in the final summary.

During the training process, the model is trained to judge the importance of a text section based on the given summary references. In the inference stage, the extractive method produces a list of sentences or phrases selected to form a summary, without changing the original structure or sentences of the document.

b) Abstractive Summarization with Transformers

In contrast, in abstractive summarization, the Transformer architecture is used to generate new summaries that reorder information from the original document. The encoder in Transformer receives news tokens as input and understands the full context of the text. During training, the decoder receives label tokens (summary references) with the initial token "<s>" in front, and the order of these label tokens is shifted right to ensure alignment between the input and output[20]. The learning model maps a sequence of input tokens to a sequence of summary tokens.

During inference, the decoder produces a summary by predicting tokens one by one, starting with the start token "<s>" and continuing until it reaches the end token "</s>" or a specified iteration limit.

5) Evaluation

The Recall-Oriented Understudy for Gisting Evaluation (ROUGE) is a commonly used technique for assessing automatic summarization systems by comparing summaries produced by machines against human-generated reference summaries[21]. ROUGE metrics typically employ three key evaluation criteria: precision, recall, and F1-score. The precision metric quantifies the percentage of n-grams in the system-produced output that are present in the reference human-crafted summary. The precision value can be calculated using the following formula:

$$Precision = \frac{overlapping\ number\ of\ n-grams}{number\ of\ n-grams\ in\ system\ summary} \quad (1)$$

Recall quantifies the fraction of n-grams from the gold-standard summary that are successfully captured in the system-produced output. The recall value is calculated by the following formula:

$$Recall = \frac{overlapping\ number\ of\ n-grams}{number\ of\ n-grams\ in\ reference\ summary} \quad (2)$$

F-measure is the average between precision and recall which provides a comprehensive evaluation measure [22]. The f-measure value is obtained from the following formula:

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

ROUGE uses several metrics that are used as standards for whether a summary result 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 ROUGE-1 focuses on unigrams (each word) in this comparison, while ROUGE-2 compares bigrams (two consecutive words or a combination of two words). The formula for calculating ROUGE-N is as follows [21]:

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

ROUGE-L evaluates the comparison by analyzing the Longest Common Subsequence (LCS) between the generated and reference summaries. This metric assesses how well the generated summary preserves the order and structure of the reference summary[21]. The formula for calculating ROUGE-L precision, recall, and f-measure is as follows:

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

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

In this formula, the function LCS(X,Y) calculates the maximum length of a continuous sequence shared between X and Y. The variable m denotes the total count of linguistic units in the human-crafted synopsis, while n represents the quantity of elements in the machine-produced condensation.

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

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

III. RESULTS AND DISCUSSION

A. Training Process

To ensure a robust evaluation and comprehensive dataset coverage, we employed a 5-Fold Cross-Validation methodology. This technique systematically partitions the dataset into five equal subsets, where each iteration assigns four subsets for training and the remaining subset for validation. This iterative approach ensures that every data point is used for both training and validation, thereby mitigating variability and enhancing the reliability of performance assessment. The training subsets facilitate primary model development by allowing the model to learn patterns and representations from the data. Meanwhile, the validation subset is utilized for fine-tuning hyperparameters and monitoring performance to prevent overfitting. Upon completion of all iterations, the final performance metrics are aggregated and averaged, providing a more comprehensive evaluation of the model's generalization capability across the entire dataset.

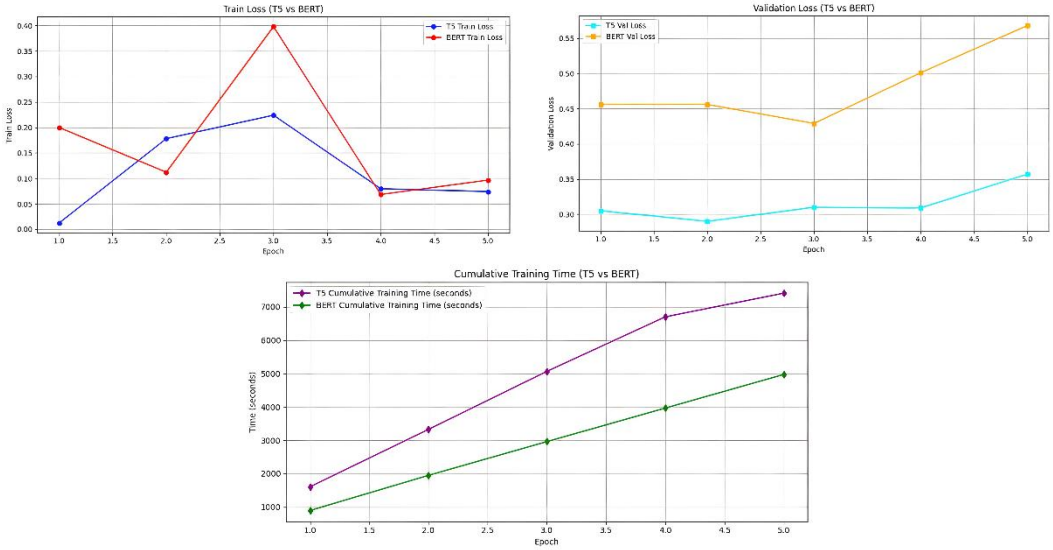

Figure 2 Training and Validation Loss & Cumulative Training Time

Based on the training efficiency illustrated in Figure 2, the cumulative training time comparison between T5 and BERT highlights the trade-offs between performance and computational efficiency. T5 exhibits a significantly higher cumulative training time across all epochs, requiring approximately 7000 seconds by the fifth epoch, whereas BERT completes training in roughly 4500 seconds. This indicates that T5's encoder-decoder architecture demands more computational resources than BERT's single-transformer encoder.

The training loss trends reveal distinct optimization characteristics. T5 maintains a relatively stable downward trend, while BERT experiences fluctuations, particularly at the third epoch, suggesting instability in convergence. On the validation side, T5 consistently demonstrates lower validation loss across all epochs, indicating better generalization to unseen data. In contrast, BERT's validation loss increases over time, which may suggest overfitting. This result highlights T5's ability to generate more fluent summaries, whereas BERT, being extractive, may struggle with generalization as training progresses.

Overall, these findings emphasize the need to balance model selection based on available hardware and efficiency requirements. While T5 demands higher computational resources, it generalizes better. BERT, on the other hand, trains faster but may require additional regularization to maintain performance on unseen data.

B. Qualitative Analysis

Our summarization approach involved adapting advanced neural network architectures sourced from the Hugging Face repository. These models, which utilize transformer technology, underwent further training to specialize in summary generation. For this process, we leveraged the Indosum corpus, a dataset comprising full-length texts paired with their corresponding human-crafted summarization.

In this analysis, we leverage human-generated summaries to conduct a thorough evaluation of the performance of various summarization models. Table 1 provides a comparative overview of these models' effectiveness against human reference summaries. For instance, instead of including the full text of a news article, we can simply provide the title or link for summarization. An example of this approach is demonstrated with the title "*Tiket Konser Band Korea 'Hyukoh' Dijual Mulai Rp 850 Ribu*," which can be accessed at *https://kumparan.com/niken-nurani/tiket-konser-band-korea-hyukoh-dijual-mulai-rp-850-ribu*.

Table 1. Example Summary Results Comparison

Model	Summary Results
Human Reference Summary	*"Band indie asal Korea Selatan , Hyukoh akan berkunjung ke Jakarta dan menggelar konser pertamanya pada 27 November mendatang di The Establishment - Lot 6 SCBD , Jakarta . Penjualan tiket untuk konser Hyukoh di Jakarta dibandrol dengan 2 harga . Yaitu Presale seharga Rp 850.000 dan harga normal seharga Rp 1.100.000. Untuk pembelian tiket dan informasi lebih lanjut dapat ditemukan di situs www.juniconcert.id dan www.kiostix.com"*
BERT (Extractive)	*"band indie asal korea selatan , hyukoh akan berkunjung ke jakarta .grup yang berada di bawah naungan anak label yg entertainment , highgrnd , ini akan menggelar konser pertamanya pada november mendatang .grup debutan 2014 ini diboyong oleh promotor musik , juni concert , yang juga mengkonfirmasi kedatangan oh hyuk , lim hyun - jae , im dong - gun dan lee in - woo ke jakarta ."*
T5 (Abstractive)	*"hyukoh akan berkunjung ke jakarta. grup yang berada di bawah naungan anak label yg entertainment, highgrnd, ini akan menggelar konser pertamanya pada november mendatang. kedatangan hyukoh ini diumumkan melalui akun twitter resmi mereka hyukohh pada 16 september kemarin, grup debutan 2014 ini diboyong oleh promotor musik juni concert yang juga mengkonfirmasi kedatangan oh hyuk lim hyun jae im dong gun dan lee in woo ke jakarta"*

The summary results generated by the BERT model tend to produce content that is more directly extracted from the original text. However, the summarization shows some issues with coherence and structure. For example, the text is presented as one continuous paragraph without proper punctuation, making it difficult to distinguish between different pieces of information. The model retains most of the main information about Hyukoh's visit to Jakarta, their label, and the concert details, but it doesn't effectively organize this information for easy comprehension.

The T5 model demonstrates a more coherent summarization ability that appears closer to what a human might produce. The output is structured into proper sentences with appropriate punctuation, improving readability. This model seems to focus on key information and presents it in a more logical flow. For instance, it groups the information about the band's visit, their label, and the concert announcement more cohesively.

However, while the T5 model maintains better coherence, it appears to have overlooked some details present in the BERT output. For example, the BERT version mentions that this will be Hyukoh's first concert in Jakarta, a detail absents from the T5 summary. This omission slightly reduces the completeness of the T5 summary.

On the other hand, the T5 model includes additional information not present in the BERT output, such as the specific date of the announcement (September 16) and the mention of the band's official Twitter account. This suggests that the T5 model might have access to or be incorporating information from a broader context, which could be seen as both a strength and a potential risk depending on the accuracy of this additional information.

Another summary example from both model that highlights differences in information selection is shown in Table 2. This table provides a comparative analysis of the generated summaries, illustrating potential biases in how each model selects and prioritizes information. Further error analysis reveals that the model occasionally omits key details or overemphasizes less relevant content, indicating possible biases in the summarization process.

Table 2. Another Example Summary Results Comparison

Model	Summary Results
Human Reference Summary	*"Andrea Dovizioso jadi pebalap tercepat di sesi latihan bebas kedua GP Austria . Sementara Valentino Rossi masih kesulitan dan duduk di posisi ke - 12 . Dovizioso membukukan catatan waktu terbaik dengan 1 menit 24,046 detik . Ia unggul 0,234 detik dari Maverick Vinales yang ada di posisi kedua . Zona lima besar dilengkapi oleh Dani Pedrosa (Honda) , Johann Zarco (YZR M1) , dan Jorge Lorenzo (Ducati) ."*
BERT (Extractive)	*"jakarta , cnn indonesia - - andrea dovizioso jadi pebalap tercepat di sesi latihan bebas kedua gp austria . sementara itu valentino rossi masih kesulitan dan kembali duduk di posisi ke - 12 . dovizioso membukukan catatan waktu terbaik dengan 1 menit 24,046 detik."*
T5 (Abstractive)	*"andrea dovizioso jadi pebalap tercepat di sesi latihan bebas kedua gp austria sementara itu valentino rossi masih kesulitan dan kembali duduk di posisi ke 12 dovizioso membukukan catatan waktu terbaik dengan 1 menit 24046 detik ia unggul 0234 detik dari maverick vinales yang ada di posisi kedua zona lima besar dilengkapi oleh dani pedrosa honda johann zarco yzr m1 dan jorge lorenzo ducati"*

The summary results generated by the BERT model tend to be highly extractive and overly selective, focusing primarily on Andrea Dovizioso while omitting crucial details about other riders. The summary only includes Dovizioso's lap time and a brief mention of Valentino Rossi's struggles, leaving out Maverick Vinales, Dani Pedrosa, Johann Zarco, and Jorge Lorenzo. This selective summarization reduces the overall informativeness, making it seem as though the competition was only between Dovizioso and Rossi. Additionally, the model truncates information, stopping abruptly after mentioning Dovizioso's fastest lap time without providing context about his lead over other competitors. Another issue with the BERT summary is its lack of coherence and poor transitions. Since it extracts information in fragments, the result appears disjointed rather than forming a well-structured narrative. The phrase *"sementara itu Valentino Rossi masih kesulitan dan kembali duduk di posisi ke - 12"* introduces slight redundancy with *"kembali"*, which suggests repetition that was not explicitly stated in the original text. Moreover, punctuation inconsistencies make the summary harder to read, as sentence boundaries are not always clearly defined.

On the other hand, the T5 model generates a more comprehensive summary but suffers from punctuation errors and readability issues. Unlike BERT, it successfully includes details about the top five riders, making it more informative. However, it presents the entire summary as a single long sentence without punctuation, making it difficult to distinguish between different pieces of information. The lack of commas, periods, and conjunctions makes the text feel like a run-on sentence, reducing clarity. For example, the phrase *"andrea dovizioso jadi pebalap tercepat di sesi latihan bebas kedua gp austria sementara itu valentino rossi masih kesulitan dan kembali duduk di posisi ke 12"* should have proper punctuation to separate different ideas. Additionally, the absence of capitalization at the beginning of the summary further contributes to its unpolished appearance.

Both models also exhibit bias in content selection. The BERT model prioritizes Andrea Dovizioso and Valentino Rossi, almost completely ignoring other riders who were in the top five. This suggests that the model tends to focus on the most well-known figures rather than providing a balanced summary. The T5 model, while more inclusive, still emphasizes Dovizioso and Rossi more than other competitors, indicating a similar preference for well-known riders.

Overall, the BERT model struggles with completeness, coherence, and content selection, producing an overly selective summary that lacks crucial details. The T5 model, while more informative, suffers from severe punctuation and formatting issues, making it difficult to read. To achieve an ideal summarization, a model should balance factual accuracy, inclusiveness, coherence, and proper formatting while avoiding biased content selection.

C. Quantitative Analysis

Here is Table 2, which compares the evaluation metrics for extractive and abstractive summarization approaches, averaged over 5 folds.

Table 2 Evaluation and Comparison of Average ROUGE Values

Approach	Model	Evaluation Metrics (Average 5 fold)		
		ROUGE 1	ROUGE 2	ROUGE L
Extractive	BERT	70.82	63.99	58.40
Abstractive	T5	69.36	61.27	66.17

The table shows that in the extractive approach, the BERT model achieves a ROUGE-1 score of 70.82, a ROUGE-2 score of 63.99, and a ROUGE-L score of 58.40. In contrast, the T5 model used in the abstractive approach attains a ROUGE-1 score of 69.36 and a ROUGE-2 score of 61.27 but surpasses BERT in the ROUGE-L metric with a score of 66.17. These results suggest that while BERT is more effective at capturing important phrases and bigrams (as indicated by higher ROUGE-1 and ROUGE-2 scores), T5 is better at maintaining the overall structure and coherence of the summary (as reflected in the higher ROUGE-L score).

Next, we attempt to compare the evaluation results with findings from previous studies that employed different models or methods. This comparison is conducted to assess how the models utilized in this research, specifically BERT and T5, perform in relation to other models that have been tested earlier.

Table 3. Comparison of ROUGE Value with previous studies

Approach	Model	Evaluation Metrics (Average 5 fold)		
		ROUGE 1	ROUGE 2	ROUGE L
Extractive	NEURALSUM[14]	67.96	61.65	**67.24**
	LEAD-3[23]	67.66	60.66	64.66
	BERT	**70.82**	**63.99**	58.40
Abstractive	GPT-2[24]	57.33	47.33	53.33
	BERT2GPT[25]	62	56	60
	T5	**69.36**	**61.27**	**66.17**

Referring to Table 3, the evaluation results comparison using the ROUGE metric reveals that among the extractive methods, the BERT model achieves the highest scores on the ROUGE-1 (70.82%) and ROUGE-2 (63.99%) metrics, but falls short compared to NEURALSUM on ROUGE-L. This indicates that while BERT is proficient at capturing relevant n-grams, it is less effective in preserving textual coherence during summarization. On the other hand, NEURALSUM outperforms the others on the ROUGE-L metric with a score of 67.24%, highlighting its superior ability to maintain sequence flow. LEAD-3 delivers consistently stable results across all metrics, albeit slightly behind the other models.

On the other hand, in the abstractive approach, T5 performed the best on all three evaluation metrics, with scores of 69.36% for ROUGE-1, 61.27% for ROUGE-2, and 66.17% for ROUGE-L. This indicates that the T5 model is superior in producing more

abstract and coherent summaries compared to other models. BERT2GPT performed fairly well, but still below T5 in all metrics. GPT-2, despite being popular in various NLP tasks, performed the lowest in this evaluation with scores of 57.33% for ROUGE-1, 47.33% for ROUGE-2, and 53.33% for ROUGE-L, indicating limitations in abstractive summarization compared to newer approaches such as T5.

IV. CONCLUSION

Based on the results and analysis that have been carried out in this study, it can be concluded that the BERT and T5 models provide high accuracy for the task of automatic text summarization on Indonesian language news articles. From several experiments conducted, the best model was achieved using the T5 architecture with an abstractive summarization approach which produced an average score of ROUGE-1 of 69.36%, ROUGE-2 of 61.27%, and ROUGE-L of 66.17%. Meanwhile, for the extractive summarization approach, the BERT model demonstrated superior performance, achieving the highest average ROUGE scores across all metrics. Specifically, it attained 70.82% for ROUGE-1, 63.99% for ROUGE-2, and 58.40% for ROUGE-L. These results strongly suggest that both BERT and T5 models can be effectively applied to text summarization tasks, producing coherent and accurate summaries. This research has great potential for further development. Future studies can explore other transformer-based models or combine several models to improve the quality of the summarization results. In addition, the application of more diverse fine-tuning techniques and exploration of various new datasets can contribute to improving model performance. Given that the selection of model architecture and hyperparameter adjustments such as learning rate and optimizer type significantly affect the results, further research is expected to consider these factors to achieve optimal performance. Furthermore, the use of deeper preprocessing techniques such as stemming, lemmatization, or text normalization is expected to help the model handle linguistic variations better, thereby improving the quality of the resulting summary.

REFERENCES

[1] 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.

[2] W. S. El-Kassas, C. R. Salama, A. A. Rafea, and H. K. Mohamed, “Automatic text summarization: A comprehensive survey,” *Expert Systems with Applications*, vol. 165, p. 113679, Mar. 2021, doi: 10.1016/j.eswa.2020.113679.

[3] A. Bahari and K. E. Dewi, “Peringkasan Teks Otomatis Abstraktif Menggunakan Transformer Pada Teks Bahasa Indonesia,” *KOMPUTA*, vol. 13, no. 1, pp. 83–91, Apr. 2024, doi: 10.34010/komputa.v13i1.11197.

[4] L. G. Andovita, A. Rahmat, and H. Pujiati, “MACRO COHERENCE LEVEL ON STUDENTS’ SCIENTIFIC PAPER,” *J HUM SOC STUD*, vol. 3, no. 2, pp. 107–112, Sep. 2019, doi: 10.33751/jhss.v3i2.1478.

[5] R. Adelia, S. Suyanto, and U. N. Wisesty, “Indonesian Abstractive Text Summarization Using Bidirectional Gated Recurrent Unit,” *Procedia Computer Science*, vol. 157, pp. 581–588, 2019, doi: 10.1016/j.procs.2019.09.017.

[6] A. R. Puspita and H. Rosyidiana, “Eksistensi Kebakuan Bahasa Indonesia dalam Karya Tulis Mahasiswa,” *n.a. bhs.*, vol. 5, no. 2, pp. 161–174, Sep. 2020, doi: 10.32528/bb.v5i2.3521.

[7] A. R. Ilmy and M. L. Khodra, “Parsing Indonesian Sentence into Abstract Meaning Representation using Machine Learning Approach,” Mar. 05, 2021, *arXiv*: arXiv:2103.03730. doi: 10.48550/arXiv.2103.03730.

[8] 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

[9] 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.

[10] 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.

[11] M. F. Mridha, A. A. Lima, K. Nur, S. C. Das, M. Hasan, and M. M. Kabir, "A Survey of Automatic Text Summarization: Progress, Process and Challenges," *IEEE Access*, vol. 9, pp. 156043–156070, 2021, doi: 10.1109/ACCESS.2021.3129786.

[12] D. P. Ismi and F. Ardianto, "Peringkasan Ekstraktif Teks Bahasa Indonesia dengan Pendekatan Unsupervised Menggunakan Metode Clustering," *CBN*, vol. 3, no. 02, p. 90, Oct. 2020, doi: 10.29406/cbn.v3i02.2290.

[13] G. E. Abdul, I. A. Ali, and C. Megha, "Fine-Tuned T5 for Abstractive Summarization," *Int J Performability Eng*, vol. 17, no. 10, p. 900, 2021, doi: 10.23940/ijpe.21.10.p8.900906.

[14] K. Kurniawan and S. Louvan, "Indosum: A New Benchmark Dataset for Indonesian Text Summarization," *International Conference on Asian Language Processing (IALP)*, pp. 215–220, 2018, doi: 10.1109/IALP.2018.8629109.

[15] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," May 24, 2019, *arXiv*: arXiv:1810.04805. Accessed: Aug. 21, 2024. [Online]. Available: http://arxiv.org/abs/1810.04805

[16] 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: Aug. 21, 2024. [Online]. Available: http://arxiv.org/abs/1910.10683

[17] S. Esmaeilzadeh, G. X. Peh, and A. Xu, "Neural Abstractive Text Summarization and Fake News Detection," Dec. 12, 2019, *arXiv*: arXiv:1904.00788. Accessed: Aug. 22, 2024. [Online]. Available: http://arxiv.org/abs/1904.00788

[18] A. Vaswani *et al.*, "Attention Is All You Need," Aug. 01, 2023, *arXiv*: arXiv:1706.03762. Accessed: Aug. 22, 2024. [Online]. Available: http://arxiv.org/abs/1706.03762

[19] S. F. N. Azizah, H. D. Cahyono, S. W. Sihwi, and W. Widiarto, "Performance Analysis of Transformer Based Models (BERT, ALBERT, and RoBERTa) in Fake News Detection," in *2023 6th International Conference on Information and Communications Technology (ICOIACT)*, Yogyakarta, Indonesia: IEEE, Nov. 2023, pp. 425–430. doi: 10.1109/ICOIACT59844.2023.10455849.

[20] A. N. S. Rahayu, T. I. Hermanto, and I. M. Nugroho, "Sentiment Analysis Using K-Nearest Neighbor Based on Particle Swarm Optimization According to Sunscreen's Reviews," *J. Tek. Inform. (JUTIF)*, vol. 3, no. 6, pp. 1639–1646, Dec. 2022, doi: 10.20884/1.jutif.2022.3.6.425.

[21] C.-Y. Lin, "ROUGE: A Package for Automatic Evaluation of Summaries".

[22] 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.

[23] School of Electrical Engineering and Informatics Institut Teknologi Bandung Bandung, Indonesia, R. Wijayanti, M. L. Khodra, School of Electrical Engineering and Informatics Institut Teknologi Bandung Bandung, Indonesia, D. H. Widyantoro, and School of Electrical Engineering and Informatics Institut Teknologi Bandung Bandung, Indonesia, "Single Document Summarization Using BertSum and Pointer Generator Network," *ijeei*, vol. 13, no. 4, pp. 916–930, Dec. 2021, doi: 10.15676/ijeei.2021.13.4.10.

[24] A. N. Khasanah and M. Hayaty, "Abstractive-Based Automatic Text Summarization on Indonesian News Using Gpt-2," *JURTEKSI*, vol. 10, no. 1, pp. 9–18, Dec. 2023, doi: 10.33330/jurteksi.v10i1.2492.

[25] 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.