Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 1, January 2025
DOI : https://doi.org/10.33395/sinkron.v9i1.14278
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Analyzing PEGASUS Model Performance with ROUGE on Indonesian News Summarization

Fatih Fauzan Kartamanah[1)*], Aldy Rialdy Atmadja[2)], Ichsan Budiman [3)]
[1)2)3)]Universitas Islam Negeri Sunan Gunung Djati Bandung, Indonesia
[1)] fatihfauzan26@gmail.com, [2)] aldyrialdy@uinsgd.ac.id, [3)] ichsanbudiman@uinsgd.ac.id

Submitted :Dec 6, 2024 | **Accepted** : Jan 2, 2025 | **Published** : Jan 8, 2025

Abstract: Text summarization technology has been rapidly advancing, playing a vital role in improving information accessibility and reducing reading time within Natural Language Processing (NLP) research. There are two primary approaches to text summarization: extractive and abstractive. Extractive methods focus on selecting key sentences or phrases directly from the source text, while abstractive summarization generates new sentences that capture the essence of the content. Abstractive summarization, although more flexible, poses greater challenges in maintaining coherence and contextual relevance due to its complexity. This study aims to enhance automated abstractive summarization for Indonesian-language online news articles by employing the PEGASUS (Pre-training with Extracted Gap-sentences Sequences for Abstractive Summarization) model, which leverages an encoder-decoder architecture optimized for summarization tasks. The dataset utilized consists of 193,883 articles from Liputan6, a prominent Indonesian news platform. The model was fine-tuned and evaluated using the Recall-Oriented Understudy for Gisting Evaluation (ROUGE) metric, focusing on F-1 scores for ROUGE-1, ROUGE-2, and ROUGE-L. The results demonstrated the model's ability to generate coherent and informative summaries, achieving ROUGE-1, ROUGE-2, and ROUGE-L scores of 0.439, 0.183, and 0.406, respectively. These findings underscore the potential of the PEGASUS model in addressing the challenges of abstractive summarization for low-resource languages like Indonesian language, offering a significant contribution to summarization quality for online news content.

Keywords: Abstractive summarization, Indonesian language, Natural Language Processing, PEGASUS, ROUGE

INTRODUCTION

Reading is a complex process involving visual, cognitive, and psycholinguistic activities. It extends beyond pronouncing words, promoting understanding, critical thinking, and mental health (Lubis, 2020; Tahmidaten & Krismanto, 2020).

In Society 5.0, technology plays a key role in digital literacy, enabling effective use and engagement with digital environments (Sugiarto & Farid, 2023). According to the Indonesian Internet Service Providers Association (APJII) (Haryanto, 2024), internet penetration in Indonesia has reached 64.8% since 2018. This figure has gradually increased to 73.7% in 2020, 77.01% in 2022, and 78.19% in 2023. By 2024, the number of internet users in Indonesia is projected to reach 221,563,479 out of a total population of 278,696,200 in 2023, resulting in an internet penetration rate of 79.5%. This shift from print to digital formats, such as blogs, which offers quick access to information.

Indonesia ranks low in reading interest, with just 0.001% of the population showing interest (Sulfa Saguni & Rahmayanti Yusuf, 2023). UNESCO reports only 1 out of 1,000 Indonesians enjoy reading (Rahmawati, 2020). Another study by Central Connecticut State University ranked Indonesia 60th out of 61 countries in terms of reading interest (Nursakinah et al., 2023). UNESCO further states that Indonesia ranks 60th out of 61 countries globally in literacy culture, with only 1% of its population interested in reading and 99% showing no interest in this activity (Mulasih & Dwi Hudhana, 2020).

*Fatih Fauzan Kartamanah

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 1, January 2025
DOI : https://doi.org/10.33395/sinkron.v9i1.14278
e-ISSN : 2541-2019
p-ISSN : 2541-044X

The rise of technology and the overwhelming flow of online information contribute to this low interest of reading (Laksana et al., 2022). On the other hand, the effective use of technology, particularly text summarization, can help by offering quick and efficient summaries (Laksana et al., 2022).

Text summarization includes extractive and abstractive methods. Abstractive summarization rewrites the text with new words, while extractive summarization selects parts of the text using sentence ranking techniques (Kirmani et al., 2024). Several studies have explored text summarization for Indonesian, including Koto et al. (2020) who developed the Liputan6 dataset, Hartawan et al. (2024) who used BART for summarization, and Aurelia et al. (2024) who enhanced BART with data augmentation.

This research aims to analyze the PEGASUS model's performance in abstractive summarization of Indonesian news using ROUGE-Scoring. The dataset used is from Liputan6, which can be compared with other models.

LITERATURE REVIEW

Text summarization condenses long documents into shorter versions for easier understanding (A. Joshi et al., 2019). It is a part of Natural Language Processing (NLP), where computers learn to understand and process human language (J. Zhang et al., 2020). NLP enables systems to comprehend the structure and meaning of the source text. There are two summarization approaches: extractive, which selects and rearranges phrases, and abstractive, which generates a new summary based on understanding (T. Zhang et al., 2024).

Several studies have contributed to Indonesian language text summarization. (Koto et al., 2020) created the Liputan6 dataset and evaluated summarization models using canonical and Xtreme test sets, employing ROUGE metrics (R1, R2, RL) and BERTSCORE. They tested models like LEAD-N, pointer-generator models (PTGEN, PTGEN+COV), and BERT-based models (mBERT, IndoBERT). PTGEN with a coverage mechanism reduced repetition, while BERT models used an extractive-abstractive approach. The primary focus was dataset development, with modeling approaches discussed briefly.

(Hartawan et al., 2024) implemented the BART model for summarization, focusing on data pre-processing, including tokenization and truncation, followed by training with AdamW optimizer. The model was fine-tuned with parameters like learning rate and batch size, then evaluated with ROUGE-1, ROUGE-2, and ROUGE-L metrics. Their results showed competitive performance, with minor ROUGE score differences compared to other models.

Another relevant study is (Aurelia et al., 2024), which focuses on the development of the BART model using a multiple dataset approach and data augmentation techniques. This study involves three stages of fine-tuning the BART model with three datasets: Indosum, Liputan6, and an augmented Liputan6 dataset. The process starts with data preprocessing and tokenization, followed by training the model on each dataset. Data augmentation is performed using the ChatGPT API to generate abstractive summaries, which are then combined with the original data to enrich the dataset. The model is then retrained with the augmented dataset. Evaluation is done using ROUGE scores to compare the model-generated summaries with human-created summaries, and manual inference is also conducted to ensure the quality of the summaries is contextually accurate.

METHOD

Research Flow

This research process includes several steps carried out sequentially to achieve the results. The author follows the workflow outlined in the figure 1, starting from the initial development stage to the completion.

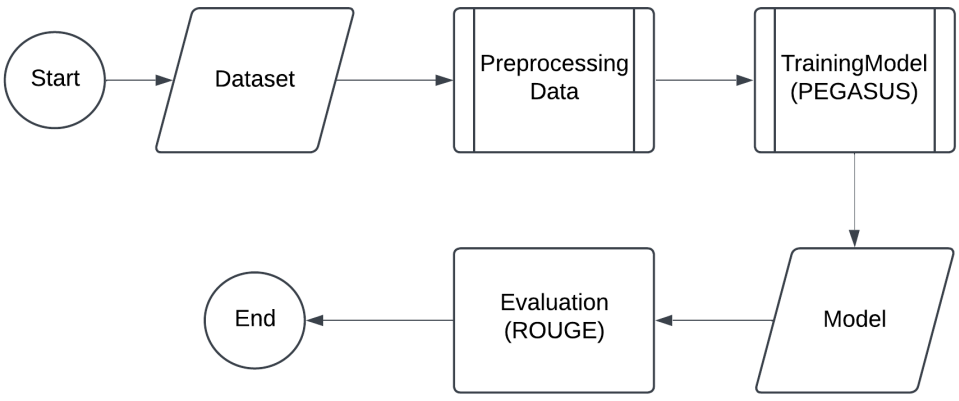

Fig. 1 Research Flow

*Fatih Fauzan Kartamanah

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 1, January 2025
DOI : https://doi.org/10.33395/sinkron.v9i1.14278
e-ISSN : 2541-2019
p-ISSN : 2541-044X

The process begins with selecting the dataset, which is then processed through the preprocessing stage to clean and prepare the data. The processed data is used to train the PEGASUS model during the training phase. Once the model is trained, its performance is evaluated using ROUGE metrics, which compare the model's generated summaries with reference summaries. The process concludes after evaluation and analysis of the results are conducted.

Dataset

The dataset used in this study is Liputan6 (Koto et al., 2020), which contains a large-scale collection of Indonesian-language news article data gathered from an online news portal called Liputan6.com. This dataset includes clean articles and clean summaries, consisting of original news articles and human-generated summaries, with a total of 215,827 document-summary pairs. The data was collected over a span of ten years (2000–2010) and covers topics such as politics, business, technology, health, and entertainment, reflecting key issues during that period. This dataset is divided into two types of tests: the Canonical variant and the Xtreme variant. The Canonical variant contains manually created text summaries, while the Xtreme variant presents summaries generated automatically.

In the Xtreme variant, test and development data are specifically selected. Document-summary pairs that have less than 90% novel 4-grams in their summaries are discarded, so the Xtreme variant contains more abstract data. This means that only summaries that are significantly different from the original text are included in this variant. However, the training data in the Xtreme variant remains the same as in the Canonical variant, without additional filtering based on the number of novel 4-grams.

Table 1 Distribution Table (Koto et al., 2020)

Variant	#Doc			% of Novel n-grams			
	Train	Dev	Test	1	2	3	4
Canonical	193,883	10,972	10,972	16.2	52.5	71.8	82.4
Xtreme	193,883	4,948	3,862	22.2	66.7	87.5	96.6

Table 1 shows the data splitting for the Canonical and Xtreme dataset variants, where the novel 4-gram value for Xtreme is 96.6%. This means that the Xtreme variant is more abstractive than the Canonical variant. Therefore, in this research, the variant model used is the Canonical variant as the training data, and the Xtreme variant (Test) as the test data.

Text Preprocessing

The preprocessing stage in the context of machine learning model training aims to prepare the dataset so it can be efficiently processed by the model. The Liputan6 dataset has already undergone preprocessing steps such as removing HTML entities, converting all words to lowercase, segmenting sentences based on punctuation, and removing unnecessary punctuation marks and special characters (Koto et al., 2020).

Modern Transformer models tend to focus more on developing sophisticated architectures, while the data preprocessing aspect is often overlooked (Siino et al., 2024). However, proper preprocessing can significantly enhance the performance of Transformer models. Essentially, Transformer models like PEGASUS leverage contextual information and subword tokenization, thus eliminating the need for traditional steps like stopword removal, stemming, or lemmatization. PEGASUS utilizes a tokenization method called SentencePiece (J. Zhang et al., 2020), a text-processing tool designed to break words into subwords and reassemble them, tailored specifically for Neural Network-based text systems, especially for tasks such as text generation (Kudo & Richardson, 2018). The SentencePiece algorithm used by PEGASUS is Unigram (J. Zhang et al., 2020). As a result, preprocessing becomes simpler as manual word removal or modification is unnecessary.

During the preprocessing stage, text is extracted from the clean article and clean summary in the dataset and undergoes tokenization to break the text into smaller units or tokens (Zouhar et al., 2023). Tokenization is performed using a specific tokenizer for the PEGASUS model, namely google/pegasus-cnn_dailymail. This tokenizer converts raw text (clean article and clean summary) into a series of indices or tokens (Wolf et al., 2020), where each word or word fragment is assigned a specific numerical index that the model can interpret. The tokenization process also involves truncation to ensure the text does not exceed the maximum (Fiok et al., 2021), determining the maximum input and target lengths, and applying padding to ensure consistent text length across batches.

*Fatih Fauzan Kartamanah

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 1, January 2025
DOI : https://doi.org/10.33395/sinkron.v9i1.14278
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Table 2 Tokenized Article

Original Article Summary
Konflik Ambon telah berlangsung selama tiga tahun. Partai Bulan Bintang wilayah Maluku siap membantu pemerintah menyelesaikan kasus di provinsi tersebut.
Tokenized Article Summary
12962 2073 16277 49333 661 110 144 76467 129 551 19787 45716 7258 46402 110 144 20711 23048 24408 107 3643 8025 29111 1321 88899 838 38114 304 34936 11140 25651 15800 10384 213 15076 63661 891 32405 386 35094 1024 11521 6366 304 42451 9994 116 1579 4218 2717 10555 7639 70792 2819 107 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0

Table 2 displays the text before and after data preprocessing (tokenization) is applied. The Original Summary represents the summary of an original article from the Liputan6 dataset, which is readable and understandable. The Tokenized Article Summary is the summary of the original article after being tokenized into a format that the model can interpret (numbers representing words). The function of padding is visible in the extra zeros at the end of the Processed Article Summary. Padding ensures all processed summaries have the same length by adding zeros, making it easier for the model to handle (Alrasheedi et al., 2023). These zeros do not alter the meaning of the summary but are used to maintain input consistency, enabling the model to process large datasets more quickly and efficiently.

This preprocessing function is applied to both training and testing datasets, ensuring that the function is executed on every batch of the dataset. This process guarantees that the entire dataset is prepared in a consistent and optimized input format for sequence-to-sequence model training, commonly used in transformer models for text summarization tasks.

PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization

The PEGASUS model is a Transformer-based encoder-decoder architecture designed specifically for abstractive text summarization (J. Zhang et al., 2020). Unlike traditional models that rely on large-scale, high-quality document-summary pairs, PEGASUS introduces a novel pre-training objective called Gap Sentences Generation (GSG) (J. Zhang et al., 2020). In this approach, important sentences are removed from the input document, and the model is trained to generate those sentences from the remaining text, mimicking the extractive summarization process.

The PEGASUS model has been tested on 12 diverse public abstractive summarization datasets, covering domains such as news, science, stories, instructions, emails, patents, and legal documents. Measured by ROUGE scores (Lin, 2004), the model achieved state-of-the-art performance across all these tasks, demonstrating its effectiveness across various content types. PEGASUS also excelled in low-resource settings, outperforming previous best results on six datasets with only 1,000 examples.

The success of this model stems from its ability to comprehend and generate text resembling summaries from a large corpus, facilitated by the GSG method. This method encourages the model to understand the overall context and produce coherent summaries. Furthermore, human evaluations confirmed that summaries generated by PEGASUS are comparable to those written by humans, further validating its performance.

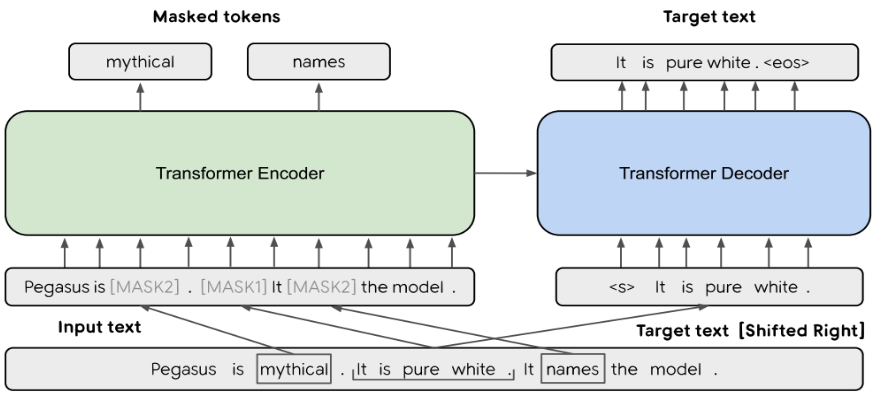

Fig. 2 Base architecture of PEGASUS (J. Zhang et al., 2020)

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 1, January 2025
DOI : https://doi.org/10.33395/sinkron.v9i1.14278
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Figure 2 shows the working mechanism of the basic architecture of the PEGASUS model. The PEGASUS architecture is a Transformer-based encoder-decoder model, consisting of two components: an encoder to understand the input text and a decoder to generate the summarized text. The model is trained using two techniques: Gap Sentences Generation (GSG) and Masked Language Model (MLM).

In GSG, certain sentences in the text are removed (e.g., the second sentence out of three) and replaced with [MASK1]. The model's task is to predict the removed sentences, making them the target output to be generated.

On the other hand, in MLM, some words in the remaining sentences are randomly removed and replaced with [MASK2]. The model is trained to guess these missing words. This dual approach of removing entire sentences and specific tokens from the remaining text helps PEGASUS better understand the context, enabling it to generate more coherent summaries.

Recall-Oriented Understudy for Gisting Evaluation (ROUGE)

The most commonly used evaluation metric for text summarization is ROUGE (Barbella et al., 2021). ROUGE is an automated evaluation package designed for text summarization tasks to assess how well a summary aligns with a reference summary (J. Zhang et al., 2020). This metric measures the overlap between the system-generated summary and the human-written reference summary based on the number of overlapping n-grams, without considering the semantic alignment or syntactic structure of the sentences. ROUGE-1, ROUGE-2, and ROUGE-L are the most commonly used metrics in the literature because they effectively represent the level of detail or granularity of the analyzed texts (Aurelia et al., 2024). A high ROUGE score indicates that the model-generated summary shares more words or phrases with the reference summary, making it more similar to a human-created summary. Conversely, a low score suggests fewer similarities between the two summaries (Tay et al., 2019).

In ROUGE evaluation, the system-generated summary is compared against the reference (gold) summary created by humans, which serves as the quality standard for the summarized text (Ng & Abrecht, 2015). ROUGE is widely adopted for summarization evaluation due to its established standard and extensive use in various studies (Yuliska & Syaliman, 2020). Inspired by research conducted by (Hartawan et al., 2024), the ROUGE variants used are ROUGE-N and ROUGE-L. This research analyzed the performance of a modern transformer model for abstractive summarization called BERT using a similar dataset and evaluation metric, namely Liputan6 and ROUGE.

ROUGE-N is a metric that measures the similarity between the model-generated summary and the reference summary by calculating the overlap of n-grams sequential word units in the text, such as unigrams (single words), bigrams (two-word sequences), or trigrams (three-word sequences) (Barbella & Tortora, 2022). The more overlapping n-grams between the two summaries, the higher the ROUGE-N score, indicating that the model summary is more similar to the reference summary.

$$ROUGE - N = \frac{\sum_{S \in \{ReferenceSummaries\}} \sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{S \in \{ReferenceSummaries\}} \sum_{gram_n \in S} Count \ \ (gram_n)} \quad (1)$$

The equation above shows the formula for calculating ROUGE-N, which is one of the metrics used to assess the quality of text summaries generated by a model. ROUGE-N measures the similarity between the model-generated summary and the reference summary based on the occurrence of n-grams, or sequences of words with a specific length. For example, ROUGE-1 for unigrams (single words), ROUGE-2 for bigrams (two words), and so on. In this formula, S represents the set of reference summaries, while $gram_n$ refers to the n-grams in the summary. The function $Count_{match}(gram_n)$ calculates the number of matching n-grams between the model's summary and the reference, while $Count \ \ (gram_n)$ counts the total n-grams in the reference summary. The higher the ROUGE-N value, the more similar the model's summary is to the reference summary, indicating better summary quality (Lin, 2004).

Meanwhile, ROUGE-L is a text summary evaluation metric that measures the longest common subsequence of words between the machine-generated summary and the gold-standard reference summary (Yuliska & Syaliman, 2020).

$$R_{lcs} = \frac{LCS(X,Y)}{m} \quad (2)$$

$$P_{lcs} = \frac{LCS(X,Y)}{n} \quad (3)$$

*Fatih Fauzan Kartamanah

e-ISSN : 2541-2019
p-ISSN : 2541-044X

$$F_{lcs} = \frac{(1 + \beta^2) R_{lcs} P_{lcs}}{R_{lcs} + \beta^2 P_{lcs}} \quad (4)$$

The figure shows the formula for calculating ROUGE-L, a metric used to evaluate the quality of summaries by considering word sequences using Longest Common Subsequence (LCS). ROUGE-L measures the similarity between the model-generated summary and the reference summary based on the longest matching word sequence between the two, making it more sensitive to sentence structure. The formula for ROUGE-L includes three components: Recall (R_{lcs}), Precision (P_{lcs}), and F-Measure (F_{lcs}).

The formulas in the image represent a Longest Common Subsequence (LCS)-based metric used for evaluating texts, particularly in the field of Natural Language Processing (NLP). Recall (R_{lcs}) measures the ratio of the LCS length ($LCS(X,Y)$) between the reference text (X) and the summary text (Y) to the length of the reference text (m), assessing how much information from the reference text is captured in the summary. Precision (P_{lcs}) calculates the ratio of the LCS length to the length of the summary text (n), evaluating the proportion of the summary that is relevant to the reference text. The F-measure (F_{lcs}) combines Recall and Precision using the parameter β, which allows for weighting between the two. When β = 1, Recall and Precision are weighted equally. Thus, (F_{lcs}) provides a comprehensive measure of the balance between Recall and Precision in assessing the quality of the summary. This metric is useful for evaluating how well a summary accurately and relevantly reflects the information from the reference text (Lin, 2004).

RESULT

Pegasus Modelling

After preprocessing, the data is ready to be used to train the model to generate abstractive summaries. This training process involves splitting the data into two parts: training data and test data. This allows the model to learn from the training data and then be evaluated on the test data to measure its performance on other data. The model used is google/pegasus-cnn_dailymail.

The model setup for training includes several important parameters. One of them is the number of epochs, which refers to the full cycles of training data to be processed by the model, in this case, it is set to 1 epoch. The batch size for both training and evaluation is set to 4, indicating the number of samples processed at each training step. Warmup steps are performed for 500 steps at the beginning of the training, allowing a gradual increase in the learning rate to enhance training stability. Regularization is applied with a weight decay rate of 0.01 to help prevent overfitting, which is a condition where the model performs very well on the training data but poorly on new but similar data because it memorizes specific details of the training data rather than learning more generalizable patterns.

During the training process, several parameters are set. The model periodically logs its performance every 100 steps and saves the training results every 10,000 steps, which will help maintain the training process if any unexpected issues arise during the training. Once the training is completed, the model is tested on the test data, which will give an indication of how well the model can generate accurate and relevant summaries. The details of the parameter settings can be seen in the following table.

Table 3 Fine-tune settings

per-device train batch size	4
per-device evaluation batch size	4
Number of epoch	1
Warm up steps	500
Logging steps	100
Save steps	10000
Weight decay	0,01

With the settings in the table 3 above, the model is expected to learn effectively from the training data and then be tested on the test data to ensure that the model can capture the necessary information and produce high-quality results.

*Fatih Fauzan Kartamanah

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 1, January 2025
DOI : https://doi.org/10.33395/sinkron.v9i1.14278
e-ISSN : 2541-2019
p-ISSN : 2541-044X

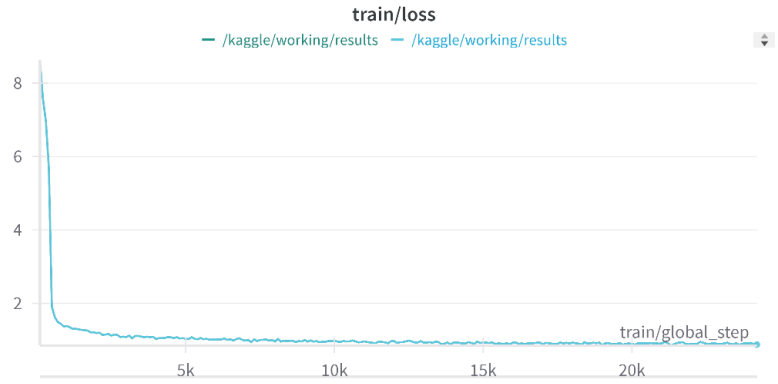

Fig. 3 Train/loss graph

Figure 3 shows the loss value during training. The initial loss value is high (around 8) and decreases rapidly in the early steps, which is normal in model training. After about 5,000 steps, the loss stabilizes at a low value (around 1), indicating that the model has learned well and stabilized with lower errors.

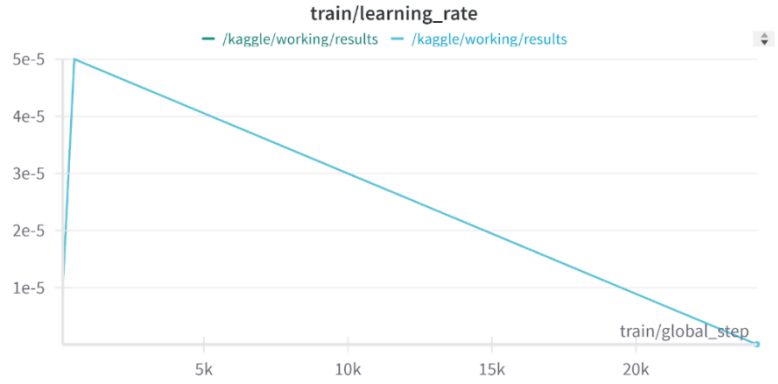

Fig. 4 Train/learning rate graph

Figure 4 shows the learning rate value during training. The learning rate starts at around 5e-5 and decreases linearly until it reaches close to 0 at the end of training. This learning rate decay strategy is used to ensure the model learns quickly at the beginning but does not change too much in the final steps, helping to prevent overfitting.

Summary Results Using the Pegasus Model

After the summary is generated, a decoding process is performed to convert the token representation back into readable text. The final result is a summary produced from the provided article.

Table 4. Summary results using the Pegasus model

Article 1	Kompas.com, Jakarta: Stadion Gelora Bung Tomo, Surabaya, bergemuruh dengan sorak sorai para pendukung Tim Nasional Indonesia U-19 pada Kamis (11/5/2024) sore. Garuda Muda baru saja menyelesaikan pertandingan terakhir Grup A Piala AFF U-19 2024 dengan hasil gemilang, mengalahkan Vietnam dengan skor 2-1. Kemenangan ini mengantarkan Timnas Indonesia U-19 ke babak semifinal, di mana mereka akan berhadapan dengan Thailand, juara grup. Pertandingan melawan Vietnam berlangsung sengit sejak menit awal. Kedua tim saling menyerang dengan determinasi tinggi untuk meraih poin penuh. Timnas Indonesia U-19 membuka skor pada menit ke-35 melalui gol tendangan bebas melengkung dari Arkhan Fikri. Gol ini disambut dengan kegembiraan luar biasa oleh para suporter yang memadati stadion. Vietnam menyamakan kedudukan pada menit ke-73 melalui gol Nguyen Van Toan. Namun, Garuda Muda tidak menyerah.

*Fatih Fauzan Kartamanah

	Pada menit ke-60, Ronaldo Kwateh mencetak gol penentu kemenangan melalui sepakan keras dari dalam kotak penalti. Gol ini memastikan Timnas Indonesia U-19 lolos ke babak semifinal. Perjalanan Timnas Indonesia U-19 di Piala AFF U-19 2024 tidaklah mudah. Garuda Muda tergabung di Grup A yang cukup berat bersama Vietnam, Thailand, Filipina, dan Myanmar. Namun, dengan kerja keras, kekompakan tim, dan dukungan penuh dari para suporter, Timnas Indonesia U-19 berhasil lolos ke babak semifinal. Di babak semifinal, Timnas Indonesia U-19 akan berhadapan dengan Thailand, juara Grup A. Thailand merupakan salah satu tim terkuat di Asia Tenggara dan menjadi favorit juara di turnamen ini. Namun, Garuda Muda tidak gentar. Mereka bertekad untuk memberikan perlawanan sengit dan meraih hasil terbaik di turnamen ini. Dukungan penuh dari para suporter menjadi salah satu faktor penting yang mendorong semangat Timnas Indonesia U-19. Saat bertanding melawan Vietnam, Stadion Gelora Bung Tomo dipenuhi oleh para suporter yang memberikan dukungan tiada henti. Sorak sorai dan chant mereka menggema di seluruh stadion, memberikan semangat bagi para pemain untuk berjuang di lapangan. Timnas Indonesia U-19 telah menunjukkan mentalitas juara di sepanjang turnamen ini. Mereka pantang menyerah dan selalu berjuang hingga akhir. Garuda Muda siap tempur di babak semifinal untuk meraih mimpi mereka, yaitu membawa pulang trofi Piala AFF U-19 2024 ke Indonesia.
Reference Summary	Timnas Indonesia U-19 berhasil lolos ke babak semifinal Piala AFF U-19 2024 setelah mengalahkan Vietnam dengan skor 2-1. Garuda Muda akan berhadapan dengan Thailand di babak semifinal. Pertandingan melawan Vietnam berlangsung sengit, dengan Timnas Indonesia U-19 membuka skor terlebih dahulu melalui gol Arkhan Fikri, kemudian disamakan oleh Vietnam, dan akhirnya Ronaldo Kwateh mencetak gol penentu kemenangan. Timnas Indonesia U-19 menunjukkan mentalitas juara dan pantang menyerah di sepanjang turnamen ini. Dukungan penuh dari para suporter menjadi kekuatan tambahan bagi Garuda Muda. Mari kita dukung Timnas Indonesia U-19 untuk meraih mimpi mereka, yaitu membawa pulang trofi Piala AFF U-19 2024 ke Indonesia!
Generated Summary	Timnas Indonesia U-19 membuka skor pada menit ke-35 melalui gol tendangan bebas melengkung Arkhan Fikri. Gol ini disambut dengan kegembiraan luar biasa dengan para suporter yang memadati stadion. Namun, Garuda Muda pantang menyerah dan selalu berjuang hingga akhir.
Article 2	Liputan6.com, Jakarta: Tentara Nasional Indonesia hendaknya benar-benar profesional. TNI juga harus berada di atas seluruh kekuatan politik yang ada. Demikian permintaan mantan Presiden Partai Keadilan Sejahtera Hidayat Nur Wahid, di sela-sela Munas PKS, Sabtu (19/6), di Jakarta. TNI adalah alat negara yang harus netral dan berada di atas seluruh kekuatan politik yang ada. Ini untuk menjaga keamanan teritorial dan keutuhan Negara Kesatuan Republik Indonesia, kata Hidayat, seperti ditulis Antara. Menurut Hidayat, TNI baru mungkin memiliki hak pilih pada pemilu jika diatur secara konstitusional melalui perundang-undangan. Saya kira wacana TNI memiliki hak pilih dalam pemilu harus melalui pembahasan lebih lanjut di DPR, ujar anggota Komisi I DPR RI itu. Dari pembicaraan dengan pimpinan TNI, kata Hidayat, sampai saat ini TNI masih memilih belum terlibat di pemilu. Hal ini belajar dari pengalaman pada pemilu 1955, di mana TNI terlibat di pemilu sehingga terbelah pada sejumlah kekuatan politik. Kondisi ini membuat sistem keamanan nasional menjadi tidak optimal. Sebelumnya, Presiden Susilo Bambang Yudhoyono mengatakan, suatu saat TNI harus diberikan haknya untuk memberikan hak suara pada pemilu jika sudah tidak ada hambatan yang mengganggu kekompakan. Bisa tidaknya anggota TNI menggunakan hak pilihnya dalam pemilu maupun pilkada, kata Presiden, dapat ditentukan oleh undang-undang yang dibahas oleh pemerintah bersama DPR.
Reference Summary	Dalam sebuah Munas PKS di Jakarta, mantan Presiden PKS, Hidayat Nur Wahid, menyerukan agar Tentara Nasional Indonesia (TNI) tetap profesional dan netral. Dia menegaskan bahwa TNI harus berada di atas segala kekuatan politik untuk menjaga keamanan teritorial dan keutuhan Negara Kesatuan Republik Indonesia (NKRI).
Generated Summary	Mantan Presiden Partai Keadilan Sejahtera Hidayat Nur Wahid mengatakan, tentara Nasional Indonesia hendaknya benar-benar profesional. TNI juga harus berada di atas seluruh kekuatan politik yang ada.

*Fatih Fauzan Kartamanah

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 1, January 2025
DOI : https://doi.org/10.33395/sinkron.v9i1.14278
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Article 3	Liputan6.com, Jakarta: Presiden Joko Widodo (Jokowi) meresmikan Bendungan Kuningan di Kabupaten Kuningan, Jawa Barat, pada hari ini, Rabu (10/5/2024). Bendungan ini merupakan salah satu proyek infrastruktur strategis nasional yang diharapkan dapat membantu mengatasi kekeringan di musim kemarau, khususnya di wilayah Jawa Barat. Bendungan Kuningan memiliki kapasitas tampung air sebesar 15 juta meter kubik, menjadikannya bendungan terbesar di Jawa Barat. Pembangunan bendungan ini menelan biaya sebesar Rp 2,2 triliun dan ditargetkan dapat mengairi sawah seluas 6.000 hektare. Selain itu, bendungan ini juga diharapkan dapat menjadi sumber air baku untuk keperluan domestik dan industri di wilayah sekitar. Dalam sambutannya, Jokowi menyampaikan harapannya agar Bendungan Kuningan dapat memberikan manfaat yang maksimal bagi masyarakat. "Bendungan ini diharapkan dapat membantu mengatasi kekeringan, meningkatkan produktivitas pertanian, dan mendorong pertumbuhan ekonomi di wilayah Kuningan," ujar Jokowi. Jokowi juga meminta kepada pemerintah daerah dan masyarakat untuk bersama-sama menjaga dan memelihara Bendungan Kuningan. "Bendungan ini adalah aset negara yang harus kita jaga bersama. Mari kita gunakan bendungan ini dengan sebaik-baiknya untuk kepentingan bersama," tutur Jokowi.
Reference Summary	Presiden Jokowi meresmikan Bendungan Kuningan di Jawa Barat dengan harapan dapat membantu atasi kekeringan, meningkatkan produktivitas pertanian, dan mendorong pertumbuhan ekonomi. Bendungan ini memiliki kapasitas tampung 15 juta meter kubik, merupakan yang terbesar di Jawa Barat, dan menghabiskan dana Rp 2,2 triliun. Jokowi meminta pemerintah daerah dan masyarakat untuk menjaga dan memelihara bendungan ini.
Generated Summary	Presiden Joko Widodo (Jokowi) meminta kepada pemerintah daerah dan masyarakat untuk bersama-sama menjaga dan memelihara Bendungan Kuningan.

ROUGE Evaluation

After completing the training process, the evaluation of the Pegasus model was conducted using the ROUGE metric to assess the quality of the generated summaries. This evaluation utilized several types of metrics, namely ROUGE-1, ROUGE-2, and ROUGE-L. These metrics measure how closely the summaries generated by the model align with the reference summaries.

The evaluation results provide scores for each type of ROUGE, including recall, precision, and F1 score. Recall indicates how much of the content in the reference summary is captured in the generated summary, while precision measures how much of the generated summary is relevant to the reference summary. The F1 score offers an overall metric by combining recall and precision.

The evaluation results are presented in bar charts, visualizing the comparison between recall, precision, and F1 score for each ROUGE metric. These charts provide a clear representation of the Pegasus model's performance in generating summaries and assist in analyzing the quality of the outputs. This evaluation helps determine whether the model has met the desired standard for performing abstractive summarization.

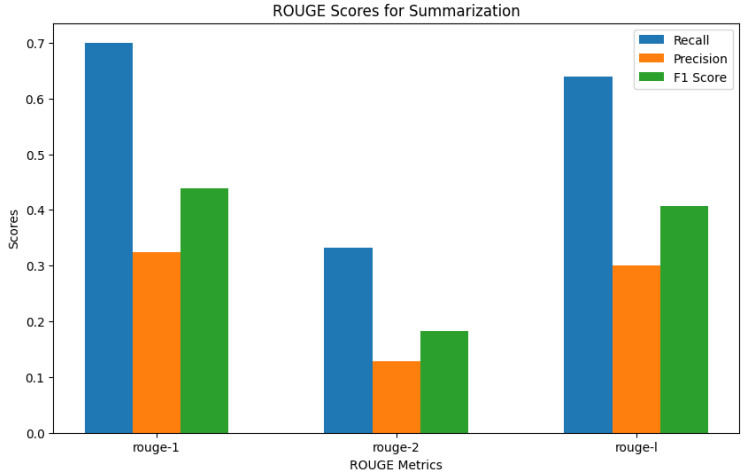

Fig. 5 ROUGE scores bar chart

Figure 5 illustrates the evaluation results of the model's performance in generating summaries using ROUGE metrics. For ROUGE-1, the recall score reached 0.70, indicating that the model successfully captured 70% of the key content from the reference summaries. The precision score was 0.32, meaning 32% of the generated summaries were relevant to the reference summaries. The F1 score for ROUGE-1 was recorded at 0.44, representing the harmonic mean of recall and precision. For ROUGE-2, the recall score was 0.33, with a precision score of 0.13,

*Fatih Fauzan Kartamanah

and an F1 score of 0.18. Meanwhile, ROUGE-L recorded a recall score of 0.64, a precision score of 0.30, and an F1 score of 0.41.

DISCUSSION

The results indicate that the Pegasus model successfully learns to perform abstractive summarization, as evidenced by the loss value decreasing from approximately 8 to 1 during training and stabilizing after around 5,000 steps. This behavior reflects effective learning driven by the chosen hyperparameters, such as the 1 epoch training duration, a batch size of 4, and a 500-step warmup for the learning rate. The ROUGE evaluation highlights the model's ability to capture a significant portion of the reference summaries' key content, as shown by the high recall scores.

Several previous studies have also performed text summarization using the Liputan6 dataset. The table 5 below shows a comparison of the F1 score evaluation results for ROUGE-1, ROUGE-2, and ROUGE-L.

Table *5* Model comparison

Research	Model	Dataset	R1	R2	RL
Koto et al., 2020	LEAD-1	Liputan6 Xtreme	27,27	11,56	23,60
	LEAD-2	Liputan6 Xtreme	31,10	12,78	27,63
	LEAD-3	Liputan6 Xtreme	29.54	12,05	26,68
	ORACLE	Liputan6 Xtreme	43.69	18.57	38.84
	PTGEN	Liputan6 Xtreme	30.41	12.05	27.51
	PTGEN+COV	Liputan6 Xtreme	30.27	11.81	27.26
	BERTEXT (mBERT)	Liputan6 Xtreme	31.83	12.63	28.37
	BERTABS (mBERT)	Liputan6 Xtreme	33.26	13.82	30.12
	BERTEXTABS (mBERT)	Liputan6 Xtreme	33.86	14.13	30.73
	BERTEXT (IndoBERT)	Liputan6 Xtreme	31.95	12.74	28.47
	BERTABS (IndoBERT)	Liputan6 Xtreme	34.59	15.10	31.19
	BERTEXTABS (IndoBERT)	Liputan6 Xtreme	34.84	15.03	31.40
Hartawan et al., 2024	BART	Liputan6 Xtreme	37.19	14.03	33.85
Aurelia et al., 2024	Model II BART-Indosum-Liputan6	Indosum, Liputan6 Canonical	33.51	23.91	32.47
	Model III BART-Indosum-Liputan6-Liputan6Augmented	Indosum, Liputan6 Canonical, Liputan6 Augmented GPT3.5	40.93	**34.09**	40.37
Research result	Pegasus	Liputan6 Xtreme	**43,94**	18,31	**40,64**

As seen in the table 5, the Pegasus model proposed by the author achieved the best F1 scores for ROUGE-1 and ROUGE-L compared to the previous studies listed, with ROUGE-1 and ROUGE-L scores of 43.94 and 40.64, respectively. These results indicate that, overall, the Pegasus model can be a good choice for abstractive text summarization in Indonesian.

Although achieving good overall results, this study has several limitations, including the use of a single epoch and a batch size of 4, which while constrained by memory limitations, may reduce the diversity of training samples and hinder the model's ability to generalize. The reliance solely on ROUGE metrics for evaluation, which do not

*Fatih Fauzan Kartamanah

e-ISSN : 2541-2019
p-ISSN : 2541-044X

fully capture semantic quality. Additionally, computational resource constraints restricted the number of training steps and data used, potentially affecting performance. Future research with improved resources and methodologies, such as manual evaluations by experts, could enhance the quality and reliability of the Pegasus model's abstractive summarization capabilities.

CONCLUSION

Based on the evaluation results using ROUGE metrics with the Liputan6 dataset, the PEGASUS model demonstrates a strong performance in generating text summaries. With a ROUGE-1 F1 score of 0.439, or approximately 43.9%, the model effectively captures key information from the text. ROUGE-2, with an F1 score of 0.183 or 18.3%, shows that the model is able to retain some information in the form of sequential n-grams, although its performance is lower compared to ROUGE-1. Additionally, ROUGE-L, with an F1 score of 0.406 or 40.6%, indicates that the PEGASUS model can generate summaries with a coherent structure that is relevant to the source text.

Overall, the findings of this study confirm that PEGASUS is an effective choice for abstractive text summarization in Indonesian. A comparison with other models previously tested shows that PEGASUS can compete in terms of summary quality, making a significant contribution to the development of automatic summarization technology, particularly for languages with limited resources such as Indonesian. Future research could further explore PEGASUS' potential in a multilingual context, expanding its ability to summarize text in various languages and enhancing its applicability in multilingual communities.

REFERENCES

A. Joshi, E. fidalgo, E. alegre, & L. Fernández-Robles. (2019). SummCoder: An Unsupervised Extractive Text Summarization Framework Based On Deep Auto-encoder. *Expert Syst Appl*, *129*. https://doi.org/10.1016/j.eswa.2019.03.045

Alrasheedi, F., Zhong, X., & Huang, P. C. (2023). Padding Module: Learning the Padding in Deep Neural Networks. *IEEE Access*, *11*, 7348–7357. https://doi.org/10.1109/ACCESS.2023.3238315

Aurelia, M., Monica, S., & Girsang, A. S. (2024). Transformer-based abstractive indonesian text summarization. *International Journal of Informatics and Communication Technology (IJ-ICT)*, *13*(3), 388. https://doi.org/10.11591/ijict.v13i3.pp388-399

Barbella, M., Risi, M., & Tortora, G. (2021). A comparison of methods for the evaluation of text summarization techniques. *Proceedings of the 10th International Conference on Data Science, Technology and Applications, DATA 2021*, 200–207. https://doi.org/10.5220/0010523002000207

Barbella, M., & Tortora, G. (2022). *ROUGE metric evaluation for Text Summarization techniques*. https://doi.org/10.2139/ssrn.4120317

Fiok, K., Karwowski, W., Gutierrez, E., Davahli, M. R., Wilamowski, M., & Ahram, T. (2021). Revisiting text guide, a truncation method for long text classification. *Applied Sciences (Switzerland)*, *11*(18). https://doi.org/10.3390/app11188554

Hartawan, G., Sa'adillah Maylawati, D., & Uriawan, W. (2024). BIDIRECTIONAL AND AUTO-REGRESSIVE TRANSFORMER (BART) FOR INDONESIAN ABSTRACTIVE TEXT SUMMARIZATION. *JIP (Jurnal Informatika Polinema)*, *10*(4), 535–541. https://doi.org/10.33795/jip.v10i4.5242

Haryanto, A. T. (2024, February 7). *APJII Jumlah Pengguna Internet Indonesia Tembus 221 Juta Orang*. Asosiasi Penyelenggara Jasa Internet Indonesia (APJII). https://apjii.or.id/berita/d/apjii-jumlah-pengguna-internet-indonesia-tembus-221-juta-orang#:~:text=Asosiasi%20Penyelenggara%20Jasa%20Internet%20Indonesia%20(APJII)%20mengumumkan%20jumlah%20pengguna%20internet,jiwa%20penduduk%20Indonesia%20tahun%202023

Kirmani, M., Kaur, G., & Mohd, M. (2024). Analysis of Abstractive and Extractive Summarization Methods. *International Journal of Emerging Technologies in Learning (IJET)*, *19*(01), 86–96. https://doi.org/10.3991/ijet.v19i01.46079

Koto, F., Lau, J. H., & Baldwin, T. (2020). Liputan6: A Large-scale Indonesian Dataset for Text Summarization. *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, 598–608. https://doi.org/10.18653/v1/2020.aacl-main.60

Kudo, T., & Richardson, J. (2018). SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 66–71. https://doi.org/10.18653/v1/D18-2012

Laksana, M., Karyawati, A., Putri, L., Santiyasa, I., Sanjaya ER, N., & Kadnyanan, I. (2022). Text Summarization terhadap Berita Bahasa Indonesia menggunakan Dual Encoding. *JELIKU (Jurnal Elektronik Ilmu Komputer Udayana)*, *11*(2), 339–348. https://doi.org/10.24843/JLK.2022.v11.i02.p13

*Fatih Fauzan Kartamanah

Lin, C.-Y. (2004). ROUGE: A Package for Automatic Evaluation of Summaries. *Proceedings of the Workshop on Text Summarization Branches Out*, 74–81.

Lubis, S. S. W. (2020). MEMBANGUN BUDAYA LITERASI MEMBACA DENGAN PEMANFAATAN MEDIA JURNAL BACA HARIAN. *Pionir: Jurnal Pendidikan*, *9*(1). https://doi.org/10.22373/pjp.v9i1.7167

Mulasih, & Dwi Hudhana, W. (2020). URGENSI BUDAYA LITERASI DAN UPAYA MENUMBUHKAN MINAT BACA. *Lingua Rima: Jurnal Pendidikan Bahasa Dan Sastra Indonesia*, *9*(2). https://doi.org/10.31000/lgrm.v9i2.2894

Ng, J.-P., & Abrecht, V. (2015). Better Summarization Evaluation with Word Embeddings for ROUGE. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 1925–1930. https://doi.org/10.18653/v1/D15-1222

Nursakinah, A. L., Sartika, D., Aisyah Humairah, N., Nur Tuada, R., Rahmadhani, A., & Arifin, S. (2023). PERANAN POJOK BACA DALAM MENUMBUHKAN BUDAYA LITERASI SISWA DI SMPN SATAP LENGGO. *PATIKALA: Jurnal Pengabdian Kepada Masyarakat*, *2*(4), 757–763. https://doi.org/10.51574/patikala.v2i4.828

Rahmawati. (2020). Komunitas Baca Rumah Luwu Sebagai Inovasi Sosial Untuk Meningkatkan Minat Baca Di Kabupaten Luwu. *DIKLUS: Jurnal Pendidikan Luar Sekolah*, *4*(2), 158. https://doi.org/10.21831/diklus.v4i2.32593

Siino, M., Tinnirello, I., & La Cascia, M. (2024). Is text preprocessing still worth the time? A comparative survey on the influence of popular preprocessing methods on Transformers and traditional classifiers. *Information Systems*, *121*. https://doi.org/10.1016/j.is.2023.102342

Sugiarto, & Farid, A. (2023). Literasi Digital Sebagai Jalan Penguatan Pendidikan Karakter Di Era Society 5.0. *Cetta: Jurnal Ilmu Pendidikan*, *6*(3), 580–597. https://doi.org/10.37329/cetta.v6i3.2603

Sulfa Saguni, D., & Rahmayanti Yusuf, N. (2023). KINERJA DINAS PERPUSTAKAAN KOTA MAKASSAR DALAM MENINGKATKAN MINAT BACA MASYARAKAT THE PERFORMANCE OF THE LIBRARY OFFICE OF MAKASSAR CITY IN INCREASING PUBLIC READING INTEREST. *Jurnal Administrasi Negara*, *29*(1). https://doi.org/10.33509/jan.v29i1.2243

Tahmidaten, L., & Krismanto, W. (2020). Permasalahan Budaya Membaca di Indonesia (Studi Pustaka Tentang Problematika & Solusinya). *Scholaria: Jurnal Pendidikan Dan Kebudayaan*, *10*(1), 22–33. https://doi.org/10.24246/j.js.2020.v10.i1.p22-33

Tay, W., Joshi, A., Zhang, X., Karimi, S., & Wan, S. (2019). Red-faced ROUGE: Examining the Suitability of ROUGE for Opinion Summary Evaluation. In M. Mistica, P. Massimo, & A. MackinlaY (Eds.), *Proceedings of the 17th Annual Workshop of the Australasian Language Technology Association* (pp. 52–60). Australasian Language Technology Association.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., Von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., … Rush, A. M. (2020). Transformers: State-of-the-Art Natural Language Processing. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 38–45. https://doi.org/10.18653/v1/2020.emnlp-demos.6

Yuliska, Y., & Syaliman, K. U. (2020). Literatur Review Terhadap Metode, Aplikasi dan Dataset Peringkasan Dokumen Teks Otomatis untuk Teks Berbahasa Indonesia. *IT Journal Research and Development*, *5*(1), 19–31. https://doi.org/10.25299/itjrd.2020.vol5(1).4688

Zhang, J., Zhao, Y., Saleh, M., & Liu, P. J. (2020). PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization. *International Conference on Machine Learning*. https://doi.org/10.48550/arXiv.1912.08777

Zhang, T., Ladhak, F., Durmus, E., Liang, P., Mckeown, K., & Hashimoto, T. B. (2024). Benchmarking Large Language Models for News Summarization. *Transactions of the Association for Computational Linguistics*, *12*, 39–57. https://doi.org/10.1162/tacl

Zouhar, V., Meister, C., Gastaldi, J. L., Du, L., Sachan, M., & Cotterell, R. (2023, June 29). Tokenization and the Noiseless Channel. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. https://doi.org/10.18653/v1/2023.acl-long.284

*Fatih Fauzan Kartamanah