



Abstractive Dialogue Summarization using Fine-Tuning Pre-Trained Language Model BART

Desty Rodiah¹, Hanif Syahri Ramadhani²

^{1,2}Departement of Informatic Engineering, Faculty of Computer Science, Universitas Sriwijaya, Inderalaya, Indonesia

*destyrodiah@ilkom.unsri.ac.id

Abstract

The increasing volume of conversational data on digital platforms necessitates effective automatic summarization methods. This research examines the application of a pre-trained BART (Bidirectional and Auto-Regressive Transformers) model for abstractive dialogue summarization on the DialogSum dataset, containing 14,460 English dialogues. The model is fine-tuned using a sequence-to-sequence framework with systematic hyperparameter tuning, including variations in learning rate, beam size, and training epochs. The optimal configuration is achieved using a learning rate of 2×10^{-5} , a beam size of 6, and 5 training epochs. Model performance is evaluated using ROUGE and BERTScore metrics. The experimental findings reveal that the suggested model achieves an average ROUGE-L F1-score of 0.40, suggesting moderate structural resemblance between produced and reference summaries, along with an average BERTScore F1-score of 0.69, demonstrating strong semantic correspondence. These findings suggest that fine-tuned BART is effective in preserving semantic relevance in abstractive dialogue summarization, although further improvements are required to enhance lexical precision and discourse-level coherence.

Keywords: abstractive summarization; BART; dialogue; natural language processing; sequence-to-sequence

How to Cite: D. Rodiah and H. S. Ramadhani, "Abstractive Dialogue Summarization using Fine-Tuning Pre-Trained Language Model BART", *J. RESTI (Rekayasa Sist. Teknol. Inf.)*, vol. 10, no. 3, pp. 745 - 755, Jun. 2026.
Permalink/DOI: <https://doi.org/10.29207/resti.v10i3.7488>

Received: January 13, 2026

Accepted: May 24, 2026

Available Online: June 21, 2026

*This is an open-access article under the [CC BY 4.0 License](#)
Published by [Ikatan Ahli Informatika Indonesia](#)*

1. Introduction

Nowadays, communication through digital platforms has become an integral part of daily activities, leading to a rapid increase in conversational data generated from instant messaging, online forums, and social media. This type of data contains valuable information that can be utilized for various applications, including dialogue analysis and automatic summarization [1]. However, the large volume of conversational data and the diversity of language styles often make it difficult for users to efficiently identify the core information [2]. Consequently, automatic summarization of dialogues has emerged as a significant research area within Natural Language Processing (NLP).

A method for automatic summarization is abstractive summarization, which creates brief summaries by rephrasing the original passage instead of pulling sentences directly [3]. The progression of deep learning methods has greatly enhanced the effectiveness of neural sequence-to-sequence models in abstractive summarization systems. Recent developments in NLP

indicate that pre-trained language models play a significant role in improving abstractive summarization performance [4], [5].

A notable model is BART (Bidirectional and Auto-Regressive Transformers), structured as a sequence-to-sequence framework that integrates a bidirectional encoder with an autoregressive decoder [6]. This architecture enables BART to effectively capture contextual information while generating coherent and fluent summaries, making it suitable for abstractive summarization tasks [7], [8]. Previous studies have shown that BART achieves competitive performance on various summarization tasks, as measured using ROUGE evaluation metrics [9-12]. While ROUGE is effective for measuring lexical overlap, it has limitations in capturing semantic similarity in abstractive summarization. Embedding-based metrics like BERTScore have been developed to assess semantic similarity that goes beyond mere exact word matching in order to tackle this problem [13].

Several previous studies have explored automatic summarization using different neural approaches. Previous studies [3] show that sequence-to-sequence models with attention significantly improve performance compared to traditional statistical methods, although challenges in maintaining coherence remain. The evolution of summarization techniques toward deep learning approaches has also been widely discussed [14], highlighting their effectiveness in capturing contextual information. In addition, it has been reported that abstractive summarization provides more flexible and human-like summaries but requires deeper semantic understanding [15]. Recent surveys [4] show that pre-trained language models like BERT, BART, and T5 greatly improve summarization effectiveness via contextual representation and fine-tuning methods. Furthermore, survey studies on dialogue summarization [2] emphasize that transformer-based models dominate current approaches while addressing challenges such as multi-speaker interactions and informal language. Recent studies have further highlighted the rapid advancement of abstractive summarization techniques. Recent survey papers [16], [17] provide a comprehensive taxonomy of summarization methods, including sequence-to-sequence models, transformer-based architectures, and large language models, while also identifying key challenges such as factual inconsistency and evaluation limitations. In addition, recent journal studies [18] demonstrate that transformer-based models with fine-tuning [19] significantly improve summarization performance, although challenges remain in capturing structural information and maintaining coherence. Furthermore, recent work on dialogue summarization [20] shows that large language models and transformer-based approaches are effective in handling conversational data (UK parliamentary debates), particularly in capturing contextual dependencies and semantic meaning.

In the specific context of dialogue summarization, recent work has proposed specialized models designed to better handle conversational data. For example, DialogLM presents a pre-trained language model that employs a window-based denoising pre-training method and sparse attention techniques to grasp long-range connections in dialogue sequences [1]. This approach demonstrates strong performance on dialogue datasets such as DialogSum and SAMSum. Recent studies such as DialogLM [1] demonstrate that incorporating denoising strategies and sparse attention significantly improves long-range dependency modeling in dialogue summarization. Similarly, transformer-based approaches evaluated on the DialogSum dataset [10] show that contextual representation plays a crucial role in improving summarization quality. Nevertheless, most existing studies primarily focus on structured documents or domain-specific datasets, such as news articles [21], [22] and clinical reports [9] while dialogue summarization remains more challenging due to

informal language, multi-speaker interactions, and implicit contextual dependencies [2], [23].

Furthermore, although several studies have explored the DialogSum dataset using transformer-based architectures [10-12] studies that specifically evaluate the effectiveness of fine-tuned BART models for everyday conversational dialogue summarization remain limited. Most existing approaches either focus on general document summarization or propose specialized architectures without examining the effectiveness of fine-tuned BART models for conversational dialogue summarization.

Based on the analysis of previous studies, several research gaps can be identified. First, many summarization studies focus on structured documents rather than conversational dialogue data, which have unique linguistic characteristics. Second, several works analyze summarization techniques at a survey level without providing empirical evaluation for specific models in dialogue summarization tasks. Third, the application of BART for abstractive summarization on conversational datasets such as DialogSum has not been extensively investigated. This research seeks to systematically assess the BART model's effectiveness for abstractive dialogue summarization on the DialogSum dataset through a fine-tuning strategy, evaluating its performance with both lexical and semantic metrics. The research aims to explore the model's ability to create coherent and informative summaries while retaining the core meaning of conversational data.

To overcome these limitations, this research explores the efficacy of the BART model for summarizing dialogues in an abstractive manner with the DialogSum dataset. The proposed approach applies a fine-tuning strategy based on sequence-to-sequence learning to adapt the pre-trained BART model to the characteristics of conversational data [8], [24]. Unlike prior studies that either focus on general document summarization or propose specialized architectures, this work provides an empirical evaluation of fine-tuned BART specifically for conversational dialogue summarization using a dual-metric evaluation approach (ROUGE and BERTScore). The effectiveness of the suggested model is assessed with both ROUGE and BERTScore metrics to deliver a thorough evaluation of summarization quality. ROUGE is utilized to assess the lexical correspondence between the produced summaries and their related reference summaries, while BERTScore measures semantic similarity by using contextual embeddings that reflect meaning beyond mere word matching. This research seeks to achieve a more dependable and comprehensive assessment of the generated summaries by merging lexical-based and semantic-based evaluation metrics, measuring not only superficial textual resemblance but also the capacity of the model to maintain the essential semantic information and contextual significance found in the reference summaries.

This research has three primary contributions. This research utilizes the BART model for the purpose of abstractive dialogue summarization, employing the DialogSum dataset. Second, the study performs fine-tuning of the pre-trained BART model using a sequence-to-sequence learning approach to better adapt the model to conversational dialogue data. Third, the suggested method is assessed using ROUGE and BERTScore metrics to offer a thorough evaluation of the quality of summarization. The results of this research aim to offer understanding regarding the usefulness of transformer-based models for practical dialogue summarization tasks and aid in creating more efficient summarization systems for conversational data.

2. Methods

This research utilizes a systematic approach made up of multiple sequential stages, as demonstrated in Figure 1. Overall, the suggested framework encompasses data preparation, model training, fine-tuning, summary creation, and assessment. Every phase is intended to guarantee that the process of abstractive dialogue summarization is performed in an organized and repeatable way.

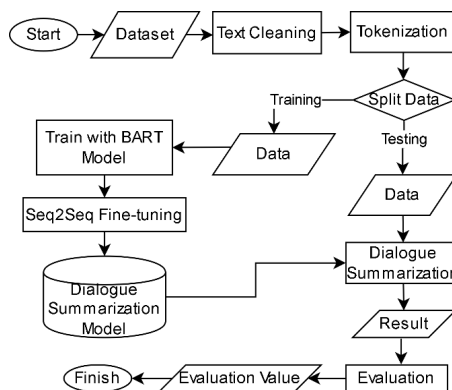


Figure 1. Research Framework

The study starts with dataset preparation, then moves on to text preprocessing to minimize noise and standardize the input data. Once preprocessing is complete, the data is tokenized and split into training, validation, and testing sets with a ratio of 70:15:15. The training data are then used to fine-tune a pre-trained BART model using a sequence-to-sequence learning approach. Subsequently, the trained model is applied to the testing data to generate dialogue summaries. Finally, the quality of the produced summaries is assessed using the ROUGE metric to gauge their resemblance to reference summaries. Figure 1 presents the overall research framework, while the following subsections describe each phase of the methodology in detail.

2.1 Text Summarization

Text summarization is a branch of natural language processing (NLP) that generates automatic summaries encompassing key sentences and pertinent information from a source text. Two primary methods of text

summarization are extractive summarization and abstractive summarization [15]. Extractive summarization is a method in text summarization that functions by pulling out key sentences or phrases from the source document. The architecture of extractive summarization shown in Figure 2. This process involves analyzing the text for features such as sentence position, sentence length, and the presence of keywords, and then sorting these sentences to form a summary [14].

Abstractive summarization is a technique in text summarization that aims to produce summaries of long texts by recreating new sentences that include the main information. This technique works by generating text summaries of the main ideas, not by copying word for word the most important sentences from a text [25].

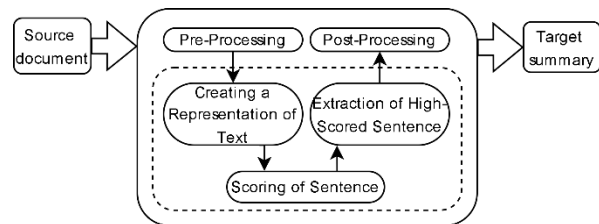


Figure 2. Extractive Summarization Architecture.

Abstractive summarization has three main stages to work perfectly. These stages include information extraction, which involves extracting important information from a document such as nouns and verbs with context, identifying relationships between objects, or abstracting verb-object pairs. The next stage is content selection, which is the selection of phrases for the text summary that can be done optimally using Integer Linear Programming (ILP), which selects all sentences simultaneously, and finally surface realization, which is combining candidate phrases using grammatical and syntactic rules to produce a coherent summary [26]. The architecture of abstractive summarization shown in Figure 3.

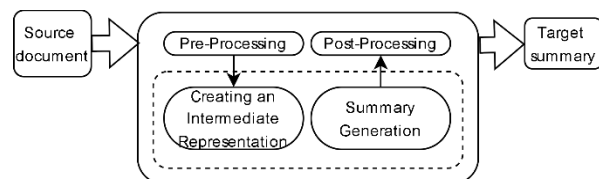


Figure 3. Abstractive Summarization Architecture.

2.2 Dataset

This research uses secondary data in the form of a collection of 14,460 English dialogs. The dataset needed in this research is obtained from DialogSum which is an English conversation dataset. This dataset was updated most recently in 2021. This dataset is utilized in studies related to natural language processing, particularly for summarizing English dialogues. The dataset was collected by downloading the dataset file with a .rar extension through the huggingface platform. The dialog in this dataset was

collected from three public dialog corporations, namely Dailydialog [27], DREAM [28], and MuTual [29].

The downloaded dataset is then extracted to produce a file with a .csv extension. The DialogSum dataset consists of 14,460 dialogue–summary pairs. The dataset is divided into training (70%), validation (15%), and testing (15%) groups, resulting in 10,122 training samples, 2,169 validation samples, and 2,169 testing samples. This split is applied consistently across all experiments to ensure reproducibility.

2.3 Text pre-processing

Text pre-processing is one of the stages in NLP that is useful for preparing text into data that is ready to be used and processed at the next stage [30]. All preprocessing steps are applied consistently across the entire dataset to ensure uniform input representation for the model. The preprocessing workflow includes these steps.

Text Cleaning: The pre-processing used is Text cleaning which is a stage of cleaning text from special characters, symbols, punctuation marks, and other unnecessary characters [31]. This stage aids in making the input text more uniform and appropriate for additional processing.

Tokenization using the BART Tokenizer. After the cleaning process, the text is processed using the BART tokenizer, which is a subword-based tokenizer designed specifically for the BART model. Rather than dividing the text solely by spaces, the tokenizer breaks it down into smaller subword segments utilizing a byte-pair encoding (BPE) approach. This method enables the model to more effectively manage uncommon words and out-of-vocabulary tokens by portraying them as combinations of subword elements [32], [33]. Subword tokenization has become a common technique in modern transformer-based models to improve vocabulary coverage and language representation [34].

Padding: Following tokenization, a padding method is utilized to guarantee that all token sequences maintain uniform length. Padding functions by incorporating special tokens into shorter sequences to ensure they align with the longest sequence length in the batch. This process enables efficient batch processing during model training [33].

Truncation: Truncation is used to ensure that overly long sequences do not surpass the model's maximum input length. In this step, tokens that exceed the predefined maximum sequence length are removed while preserving the most relevant information from the dialogue text. This mechanism ensures that all input sequences remain compatible with the model architecture [33].

By applying these preprocessing steps text cleaning, subword tokenization using the BART tokenizer, padding, and truncation the dialogue data become more structured and compatible with the BART model, which

improves the efficiency and stability of the training process.

2.4. BART Model

BART (Bidirectional and Auto-Regressive Transformers) is a denoising autoencoder created for pretraining sequence-to-sequence architectures [22]. The design of BART illustrated in Figure 4. BART contains 140 million parameters; it operates by distorting text with a random noising function and subsequently learning to rebuild the original text. BART integrates components from different architectures, primarily BERT (featuring a bidirectional encoder) and GPT (which employs a left-to-right decoder), rendering it exceptionally adaptable for diverse natural language processing tasks [33].

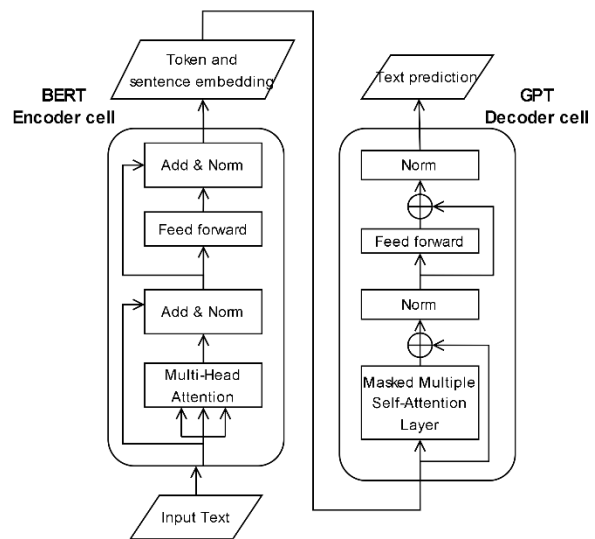


Figure.4 Pre-trained Language Model BART Architecture.

BART combines elements from several transformer-based architectures, particularly BERT, which utilizes a bidirectional encoder, and GPT, which employs an autoregressive left-to-right decoder. This encoder–decoder structure allows BART to capture contextual relationships in the input sequence while generating coherent output sequences [33].

In this study, the BART-base model (facebook/bart-base) is used as the pre-trained model for fine-tuning. This model contains approximately 140 million parameters and is widely used in text generation tasks due to its balance between performance and computational efficiency. The use of the facebook/bart-base model ensures that the experimental setup is transparent and reproducible, allowing other researchers to replicate the training process and evaluation using the same publicly available model configuration.

2.5 Sequence-to-sequence (seq2seq)

Sequence-to-sequence (seq2seq) is a neural network framework crafted to convert an input sequence into an output sequence, finding extensive use in Natural Language Processing (NLP) applications like machine

translation and text summarization. The architecture of seq2seq shown in Figure 5. The structure typically includes an encoder that transforms the input text into a contextual representation, along with a decoder that produces the output sequence incrementally using this representation [35].

The seq2seq architecture is applied through Bidirectional and Auto-Regressive Transformers (BART), a transformer-centric encoder-decoder model crafted for sequence generation activities like summarization. BART integrates a bidirectional encoder with an autoregressive decoder to encompass contextual details and produce cohesive summaries [33].

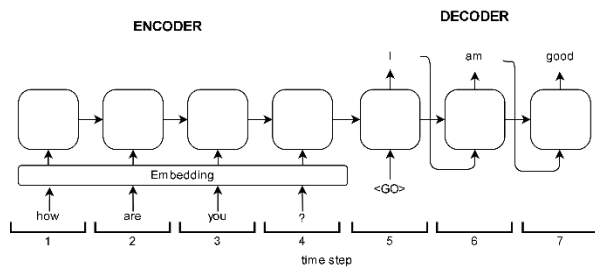


Figure 5. Sequence-to-sequence Architecture.

To evaluate the model's efficacy, various hyperparameters were examined, such as the learning rate (2×10^{-5} and 3×10^{-5}), the count of training epochs (3 and 5), and the beam search setting (num_beams = 4 and 6). The selected learning rates follow commonly used settings for fine-tuning pre-trained transformer models, where relatively small learning rates are recommended to preserve the knowledge acquired during pretraining while allowing the model to adapt to downstream tasks [34]. In the same way, the count of training epochs was adjusted to investigate how training time influences model performance while reducing the chance of overfitting.

Furthermore, the beam search parameter was adjusted to enhance the quality of the produced summaries throughout the decoding phase. Beam search is commonly utilized in sequence generation tasks to investigate several candidate sequences and choose the most likely output. Increasing the number of beams allows the model to consider a larger search space and potentially produce more informative summaries [33]. Therefore, the selected hyperparameter configurations aim to balance training stability, computational efficiency, and generation quality.

2.6 Evaluation

ROUGE is short for Recall-Oriented Understudy for Gist Evaluation. ROUGE is a collection of metrics utilized to assess summaries generated by text summarization algorithms [36]. ROUGE operates by employing precision, recall, and F1-Score to assess the n-gram overlap between the summary produced by the model, often called the candidate summary, and the

reference summary given [37]. According to [38] the ROUGE metric has an equation, namely Equation 1.

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

n represents the length of n-grams, $Count_{match}(gram_n)$ indicates the quantity of n-grams that match between the candidate summary and the reference summary, while $Count(gram_n)$ signifies the total number of n-grams found in the reference summary. In this study, only one form of ROUGE is utilized, specifically ROUGE-N, a metric that assesses the quantity of n-grams shared between the candidate summary and the reference summary.

Two ROUGE-N metrics will be utilized, specifically ROUGE-1, which assesses the similarity of each word in the candidate summary to the reference summary, also known as unigram. Additionally, ROUGE-2 refers to the comparison of a pair of words between the candidate summary and the reference summary, also known as bigram. According to [39], the equation below is utilized in determining the ROUGE-N metric evaluation value through Equations 2-4 [40].

$$Precision = \frac{ngramRef \cap ngramCand}{ngramCand} \quad (2)$$

Precision assesses the degree of relevance of the n-grams produced by the candidate summary in relation to the reference summary.

$$Recall = \frac{ngramRef \cap ngramCand}{ngramRef} \quad (3)$$

Recall assesses the count of n-grams produced by the reference summary that appear in the candidate summary.

$$F1\ Score = \frac{2 \times Recall \times Precision}{(Recall + Precision)} \quad (4)$$

F1-Score measures the performance or quality of the model that generates summaries by summing the precision and recall values into a comprehensive value of how similar the candidate summaries by the system are to the reference summaries.

BERTScore measures the semantic similarity between a generated summary and a reference summary by utilizing contextual token embeddings derived from a pre-trained BERT model. Given a candidate summary $C = \{c_1, \dots, c_m\}$ and a reference summary $R = \{r_1, \dots, r_n\}$, BERTScore calculates the cosine similarity of the token embeddings. Precision is calculated by matching each token in C to the most similar token in R , while recall is computed in the opposite direction. The ultimate F1-score is derived as the harmonic average of precision and recall. This evaluation method based on embeddings allows BERTScore to recognize semantic similarity that goes beyond mere lexical overlap, making it ideal for abstractive summarization [13].

2.7 Training Configuration

The model underwent fine-tuning via the HuggingFace Transformers library utilizing the PyTorch framework. Training was conducted utilizing the Adam optimizer with a batch size of 4.

The training procedure utilized hyperparameters with learning rates (2×10^{-5} and 3×10^{-5}), epoch counts for training (3 and 5), and the beam search setting (`num_beams` = 4 and 6) capping the output length at 128 tokens. The training process was conducted using Python and PyTorch and NLTK.

2.8 Experimental Setup

To ensure reproducibility, the complete experimental configuration is summarized as follows. The study employs the pre-trained BART-base model (facebook/bart-base), which has roughly 140 million parameters. The model is adjusted through the HuggingFace Transformers library utilizing the PyTorch framework. Training is conducted using the Adam optimizer with a batch size of 4. The learning rates evaluated in this study are 2×10^{-5} and 3×10^{-5} , with training epochs set to 3 and 5. During decoding, beam search is applied with `num_beams` values of 4 and 6, and the maximum output sequence length is limited to 128 tokens.

The experiments are conducted using Python, with additional preprocessing support from the NLTK library. All experiments follow the same dataset split (70:15:15) and preprocessing pipeline to ensure consistency and reproducibility.

The evaluation of model performance employs both ROUGE (ROUGE-1, ROUGE-2, and ROUGE-L) and BERTScore metrics. ROUGE assesses lexical overlap, whereas BERTScore analyzes semantic similarity by utilizing contextual embeddings.

3. Results and Discussions

This research employs the DialogSum dataset, a genuine dialogue summarization dataset that includes common English conversations, presented in [10] and obtainable through Github (<https://github.com/cylnlp/DialogSum>). This study uses the DialogSum dataset consisting of 14,460 dialogue instances, each paired with a human-written reference summary. The data is divided into training, validation, and testing groups at a ratio of 70:15:15, yielding 10,122 training, 2,169 validation, and 2,169 testing samples.

Figure 6 presents the word cloud generated from the DialogSum dataset, illustrating the most frequent terms and providing an overview of the dominant themes in the dialogues. Additionally, Figure 7 illustrates the distribution chart of dialogue and summary lengths, emphasizing the span where most dialogue instances and their related summaries are focused.

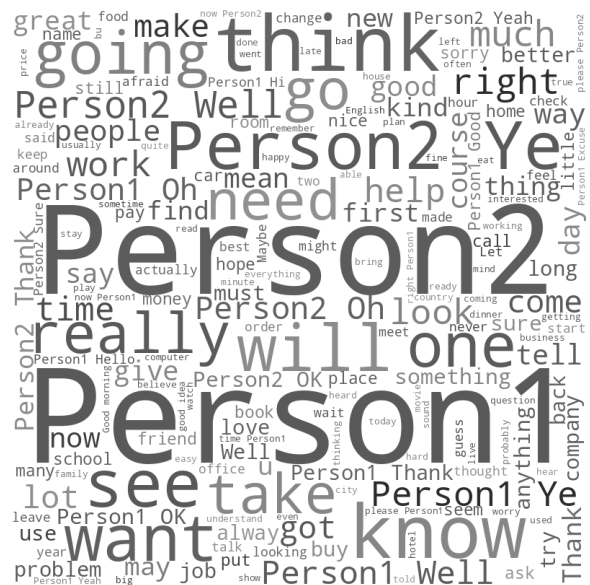


Figure 6. WordCloud from the DialogSum dataset

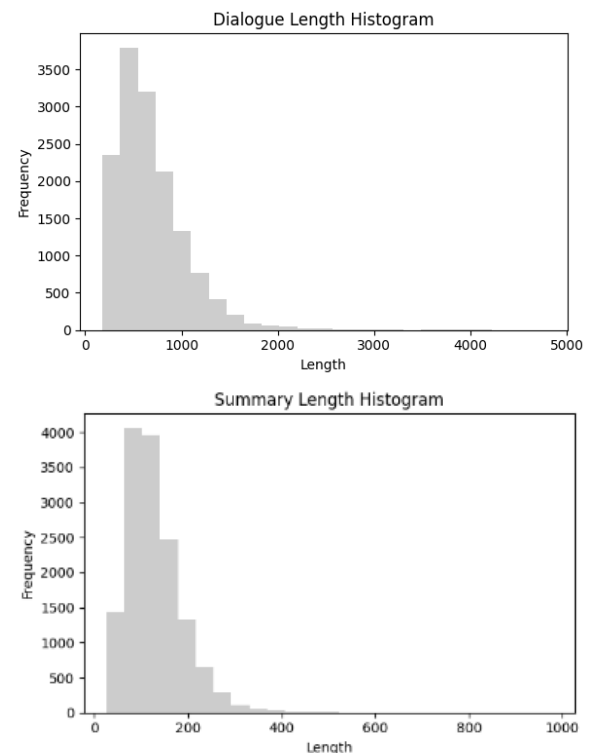


Figure 7. The distribution diagram of dialogue and summary lengths

Prior to using the dataset, text preprocessing was performed, including text cleaning and tokenization using the BART tokenizer to reduce noise and ensure consistent input representation. The following example presents the original dialogue and reference summary alongside their preprocessed versions, demonstrating the effectiveness of the applied preprocessing steps.

Original Dialogue:

#Person1#: Excuse me. Is anyone sitting here?

#Person2#: No, nobody.

#Person1#: You don't mind if I smoke, do you?

#Person2#: Well, to be frank, yes, I do.

#Person1#: Oh, I'm sorry. But this isn't a no-smoker, is it? I mean would you mind if I smoke here?
 #Person2#: Actually it is. Perhaps you haven't noticed the sign.
 #Person1#: Sign? What's sign?
 #Person2#: There, on the window.
 #Person1#: Oh, sorry. I didn't notice it. Sorry.
 #Person2#: That's all right.

Original Reference Summary:

#Person1# wants to smoke but #Person2# reminds #Person1# that smoking is not allowed here.

Preprocessed Dialogue:

Person1 Excuse me Is anyone sitting here Person2 No nobody Person1 You dont mind if I smoke do you Person2 Well to be frank yes I do Person1 Oh Im sorry But this isnt a nosmoker is it I mean would you mind if I smoke here Person2 Actually it is Perhaps you havent noticed the sign Person1 Sign Whats sign Person2 There on the window Person1 Oh sorry I didnt notice it Sorry Person2 Thats all right

Preprocessed Reference Summary:

Person1 wants to smoke but Person2 reminds Person1 that smoking is not allowed here

The BART model is fine-tuned using a sequence-to-sequence approach with variations in learning rate (2×10^{-5} and 3×10^{-5}), number of epochs (3 and 5), and beam size (4 and 6), resulting in eight experimental configurations. Table 1 displays the combinations of hyperparameter settings utilized in the experiments, outlining the differences in learning rate, beam size (num_beams), and the number of training epochs for each experimental setup. The best-performing configuration was selected as the final model setting based on evaluation results.

Table 1. Hyperparameter Combinations for BART Fine-Tuning Experiment

Experiment ID	Learning Rate	Num Beams	Epochs
LR2e5_B4_E3	2×10^{-5}	4	3
LR2e5_B4_E5	2×10^{-5}	4	5
LR2e5_B6_E3	2×10^{-5}	6	3
LR2e5_B6_E5	2×10^{-5}	6	5
LR3e5_B4_E3	3×10^{-5}	4	3
LR3e5_B4_E5	3×10^{-5}	4	5
LR3e5_B6_E3	3×10^{-5}	6	3
LR3e5_B6_E5	3×10^{-5}	6	5

During training, the produced summaries were evaluated against the reference summaries, which acted as the ground truth, utilizing ROUGE metrics. The ROUGE assessment comprises ROUGE-1, ROUGE-2, and ROUGE-L, evaluated based on precision, recall, and F-measure. Furthermore, Table 2 displays the evaluation outcomes for each experimental setup, illustrating the performance comparison based on ROUGE scores and validation loss.

According to Table 2 and Figure 8, the top-performing setup is LR2e5_B6_E5, which secures a ROUGE-1

score of 49.21, ROUGE-2 score of 23.12, and ROUGE-L score of 40.67. This indicates that a lower learning rate combined with sufficient training epochs and moderate beam size provides optimal performance.

Table 2. Training Result of Experimental Configuration

Experiment ID	ROUGE-1	ROUGE-2	ROUGE-L	Loss
LR2e5_B4_E3	47.42	21.45	38.79	1.58
LR2e5_B4_E5	48.94	22.91	40.43	1.54
LR2e5_B6_E3	48.79	22.72	40.30	1.63
LR2e5_B6_E5	49.21	23.12	40.67	1.64
LR3e5_B4_E3	48.97	22.23	40.05	2.00
LR3e5_B4_E5	48.79	22.62	40.26	1.81
LR3e5_B6_E3	48.60	22.15	39.79	2.29
LR3e5_B6_E5	48.81	22.49	40.24	2.03

Figure 8. Comparison of ROUGE-1, ROUGE-2, and ROUGE-L scores across different hyperparameter configurations. The results indicate that lower learning rates combined with sufficient training epochs and moderate beam sizes yield more stable summarization performance.

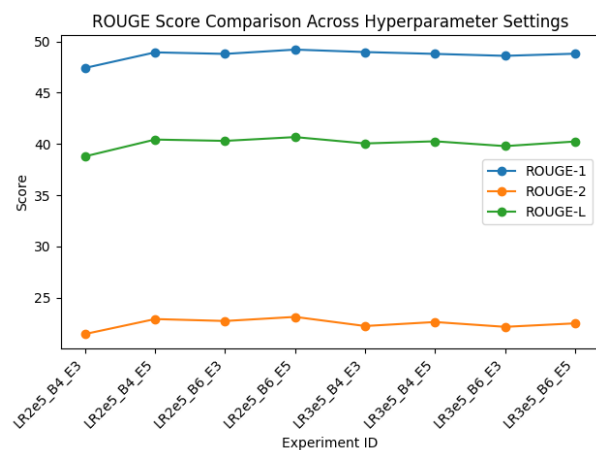


Figure 8. ROUGE Score Comparison Across Hyperparameter Setting

For the testing phase, the best-performing model configuration, LR2e5_B6_E5, was employed to generate dialogue summaries. Table 3 presents five representative examples selected from the 2,169 test dialogues, showing the system-generated summaries alongside their corresponding reference summaries. These examples are provided to qualitatively illustrate the model's ability to capture key information, maintain coherence, and preserve semantic meaning when compared to the ground-truth summaries.

The assessment procedure involves contrasting the reference summary with the summary produced by the system. ROUGE evaluation measures performance using precision (P), recall (R), and F1-score (F1) to assess lexical overlap between the summaries. In addition, BERTScore is employed to evaluate semantic similarity by capturing contextual meaning beyond exact word matching, which is particularly important for abstractive summarization. The evaluation results are presented in Table 4.

Table 3. System Summary Results

ID	Preprocessed Dialogue	Preprocessed Reference Summary	System Summary
1	Person1 Excuse me Is anyone sitting here Person2 No nobody Person1 You dont mind if I smoke do you Person2 Well to be frank yes I do Person1 Oh Im sorry But this isnt a nosmoker is it I mean would you mind if I smoke here Person2 Actually it is Perhaps you havent noticed the sign Person1 Sign Whats sign Person2 There on the window Person1 Oh sorry I didnt notice it Sorry Person2 Thats all right	Person1 wants to smoke but Person2 reminds Person1 that smoking is not allowed here	Person1 asks Person2 if Person1 wants to smoke but Person2 refuses because the sign is on the window
2	Person1 Oh hi Linda Im glad Ive caught you Person2 Hi Chris Person1 Im having a party this Saturday Ive just moved into a new flat Were starting at about five thirty Can you come Person2 Id love to Thank you very much Person1 Heres my new address Person2 Thanks Church Street Randwick Person1 Right The partys at five thirty pm Person2 I wont forget and Ill be there in time Person1 See you then Person2 Bye	Chris invites Linda to the party and gives her the address Linda will be there on time	Chris invites Linda to his new flat party Linda agrees
3	Person1 How can you keep in touch with your aunt in America Person2 Oh I make a phone call once a month Person1 Its very expensive right Person2 Yeah Person1 Why not send Email through the Internet Person2 Is it convenient and cheap Person1 Sure You can send Email to each other every day through the internet In addition you can make a phone call through the internet too Person2 Really Im so glad to hear that But I dont know how to get access to the Internet Could you show me Person1 No problem Ill tell you Person2 Ill appreciate it very much	Person1 tells Person2 Person2 can keep in touch with Person2s aunt in America through the internet and willing to show Person2 how	Person1 teaches Person2 how to keep in touch with Person2s aunt in America via email
4	Person1 I usually get a lot of information on the computer and use Email to send messages to my friends Person2 Me too Recently Im fascinated with netchat Ive made many friends on the net Every day I talk to them and share many interesting things with them I really enjoy it Person1 Dont you think its a sheer waste of both time and money Person2 Im afraid not Every coin has two sides I think using internet has a great deal to do with human interaction	Person1 and Person2 are talking about netchat Person2 thinks it has both advantages and disadvantages	Person1 and Person2 are fascinated with netchat They think using the internet has a great deal to do with human interaction
...	...	...	...
2,169	Person1 I want to buy a black suit Person2 Certainty Do you want double breasted suit or single breasted suit Person1 I prefer single breasted suit Person2 What about this one Its a very nice 3piece suit Person1 That looks great Do you have ties here	Person1 prefers a singlebreasted suit and Person2 recommends one	Person2 recommends a 3piece suit to Person1

Table 4. Rouge Score Evaluation Results

ID	ROUGE-1			ROUGE-2			ROUGE-L			BERT-Score		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
1	0,57	0,44	0,50	0,38	0,29	0,33	0,50	0,39	0,44	0,70	0,74	0,72
2	0,68	0,71	0,70	0,38	0,40	0,39	0,55	0,57	0,56	0,77	0,73	0,75
3	0,55	0,80	0,65	0,38	0,57	0,46	0,45	0,67	0,54	0,87	0,81	0,84
4	0,47	0,33	0,39	0,21	0,15	0,18	0,47	0,33	0,39	0,65	0,71	0,68
...	...	...	...	...	...	...	...	...	...	...	...	...
2,16	0,56	0,71	0,63	0,13	0,17	0,14	0,22	0,29	0,25	0,74	0,69	0,71

Figure 9 presents the final assessment outcomes indicating that the suggested model reaches a ROUGE-1 score of 0.49, a ROUGE-2 score of 0.23, a ROUGE-L score of 0.40, and a BERTScore of 0.69. These results indicate that the model performs well in capturing the main content and semantic meaning of dialogues, although limitations remain in preserving bigram structures and sentence-level coherence.

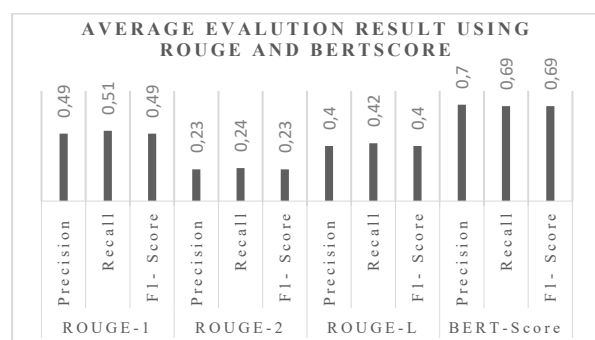


Figure 9. Average Evaluation Result Using ROUGE and BERTScore

To further evaluate the effectiveness of the proposed approach, the results are compared with prior studies on

the DialogSum dataset. Compared to [11], the proposed approach achieves comparable ROUGE-1 performance (49.21 vs. 49.62), indicating similar capability in capturing core information. However, lower performance is observed in ROUGE-2, ROUGE-L, and BERTScore, suggesting limitations in structural coherence and semantic alignment.

In comparison with [12], the proposed model outperforms baseline and standard fine-tuned models, and achieves comparable results to topic-enhanced PEGASUS models. Notably, this performance is achieved without incorporating additional mechanisms such as topic modeling, demonstrating the effectiveness of standard BART fine-tuning.

These findings highlight that dialogue summarization remains a challenging task due to informal language, multi-speaker interactions, and implicit context. The observed performance gap also indicates that further improvements are needed in modeling discourse structure and semantic consistency.

3.1 Key Finding

Overall, these findings indicate that the proposed BART-based model is more effective in preserving semantic meaning than exact lexical and structural similarity in dialogue summarization tasks. This confirms that abstractive summarization models tend to prioritize meaning representation over surface-level text matching. While the model demonstrates strong capability in capturing contextual information, as

reflected by the high BERTScore, its limitations in bigram overlap and structural coherence suggest the need for further improvement in maintaining sentence structure and discourse flow. Therefore, this study highlights that combining lexical-based metrics (ROUGE) and semantic-based metrics (BERTScore) is essential to obtain a more comprehensive evaluation of abstractive summarization performance.

These findings are consistent with the initial hypothesis of this study, which assumes that the BART model is capable of effectively capturing semantic information in dialogue summarization tasks. The results, particularly the high BERTScore, support this hypothesis, although the lower ROUGE-2 score indicates that improvements are still needed in maintaining lexical and structural coherence. These results also address the identified research gap regarding the limited exploration of BART for conversational dialogue summarization on the DialogSum dataset, demonstrating that fine-tuned BART can effectively handle informal and multi-speaker conversational data. This further confirms the study's contribution by offering an empirical assessment of BART through both lexical and semantic evaluation criteria. These results indicate that transformer-based models like BART are better suited for applications needing semantic comprehension of conversational data instead of precise lexical matching, which has significant consequences for practical dialogue summarization systems.

3.2 Limitations and Future Work

Even though the suggested model exhibits strong performance, it still reveals shortcomings in structural consistency and lexical accuracy. Therefore, future work will focus on several directions. First, incorporating structure-aware learning mechanisms, such as discourse-aware modeling or role-based representations, may improve the model's ability to capture multi-speaker interactions and dialogue flow. Second, integrating advanced decoding strategies, including diverse beam search or contrastive decoding, could enhance the fluency and informativeness of generated summaries.

Furthermore, upcoming studies could investigate the application of more extensive pre-trained models and instruction-tuned language models to enhance semantic alignment and contextual comprehension. Another promising direction is the integration of hybrid approaches, combining abstractive and extractive techniques to balance semantic representation and structural accuracy. Finally, extending the evaluation by incorporating human assessment metrics, such as coherence, readability, and factual consistency, as well as testing on multilingual or cross-domain dialogue datasets, would provide a more comprehensive validation of the proposed approach.

4. Conclusions

This study demonstrates that a fine-tuned BART model can achieve competitive performance for abstractive dialogue summarization on the DialogSum dataset, particularly in preserving semantic relevance. The experimental results show an average ROUGE-1 F1-score of 0.49, ROUGE-2 of 0.23, ROUGE-L of 0.40, and BERTScore F1 of 0.69. These findings indicate that the model is effective in capturing the main content and semantic meaning of conversational data, although performance in lexical overlap and structural coherence remains moderate. The relatively low ROUGE-2 score reflects the inherent characteristics of abstractive dialogue summarization, where paraphrasing and informal multi-speaker interactions reduce exact n-gram overlap. In contrast, the higher BERTScore confirms that the generated summaries maintain strong semantic similarity with the reference summaries, highlighting the importance of embedding-based evaluation for this task. From the experimental analysis, it is observed that a lower learning rate combined with sufficient training epochs and moderate beam size provides a stable balance between model convergence and summary quality. This finding is consistent with common practices in fine-tuning transformer-based sequence-to-sequence models. Even with these encouraging outcomes, various constraints persist. The study relies solely on automatic evaluation metrics without human assessment, and it evaluates only a single model without direct comparison to baseline methods. In addition, the use of a single English dataset may limit the generalizability of the findings.

Overall, this study confirms that BART is suitable for dialogue summarization tasks, particularly for preserving semantic information. Future work should focus on incorporating baseline comparisons, human evaluation, and structure-aware learning approaches to further improve summary coherence and robustness.

Acknowledgements

This project received funding from the Faculty of Computer Science, Universitas Sriwijaya.

References

- [1] M. Zhong, Y. Liu, Y. Xu, C. Zhu, and M. Zeng, "DialogLM: Pre-trained Model for Long Dialogue Understanding and Summarization," *Proc. AAAI Conf. Artif. Intell.*, vol. 36, no. 10, pp. 11765–11773, Jun. 2022, doi: [10.1609/aaai.v36i10.21432](https://doi.org/10.1609/aaai.v36i10.21432).
- [2] X. Feng, X. Feng, and B. Qin, "A Survey on Dialogue Summarization: Recent Advances and New Frontiers," in *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*, California: International Joint Conferences on Artificial Intelligence Organization, Jul. 2022, pp. 5453–5460. doi: [10.24963/ijcai.2022/764](https://doi.org/10.24963/ijcai.2022/764).
- [3] S. Gupta and S. K. Gupta, "Abstractive summarization: An overview of the state of the art," *Expert Syst. Appl.*, vol. 121, pp. 49–65, May 2019, doi: [10.1016/j.eswa.2018.12.011](https://doi.org/10.1016/j.eswa.2018.12.011).
- [4] A. A. Syed, F. L. Gaol, A. Boediman, T. Matsuo, and W. Budiharto, "A Survey of Abstractive Text Summarization Utilising Pretrained Language Models," N. T. Nguyen, T. K. Tran, U. Tukayev, T.-P. Hong, B. Trawiński, and E.

- Szecerbicki, Eds., Cham: Springer International Publishing, 2022, pp. 532–544. doi: [10.1007/978-3-031-21743-2_42](https://doi.org/10.1007/978-3-031-21743-2_42).
- [5] L. Xiao, L. Wang, H. He, and Y. Jin, "Modeling Content Importance for Summarization with Pre-trained Language Models," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, B. Webber, T. Cohn, Y. He, and Y. Liu, Eds., Stroudsburg, PA, USA: Association for Computational Linguistics, Nov. 2020, pp. 3606–3611. doi: [10.18653/v1/2020.emnlp-main.293](https://doi.org/10.18653/v1/2020.emnlp-main.293).
 - [6] T. Shi, Y. Keneshloo, N. Ramakrishnan, and C. K. Reddy, "Neural Abstractive Text Summarization with Sequence-to-Sequence Models," *ACM/IMS Trans. Data Sci.*, vol. 2, no. 1, pp. 1–37, Feb. 2021, doi: [10.1145/3419106](https://doi.org/10.1145/3419106).
 - [7] C. Prasad, J. S. Kallimani, D. Harekal, and N. Sharma, "Automatic Text Summarization Model using Seq2Seq Technique," in *2020 Fourth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)*, IEEE, Oct. 2020, pp. 599–604. doi: [10.1109/I-SMAC49090.2020.9243572](https://doi.org/10.1109/I-SMAC49090.2020.9243572).
 - [8] C. Zheng, K. Zhang, H. J. Wang, L. Fan, and Z. Wang, "Enhanced Seq2Seq Autoencoder via Contrastive Learning for Abstractive Text Summarization," in *2021 IEEE International Conference on Big Data (Big Data)*, IEEE, Dec. 2021, pp. 1764–1771. doi: [10.1109/BigData52589.2021.9671819](https://doi.org/10.1109/BigData52589.2021.9671819).
 - [9] S. Macharla, A. Madamanchi, and N. Kancharla, "nav-nlp at RadSum23: Abstractive Summarization of Radiology Reports using BART Finetuning," in *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2023, pp. 541–544. doi: [10.18653/v1/2023.bionlp-1.55](https://doi.org/10.18653/v1/2023.bionlp-1.55).
 - [10] Y. Chen, Y. Liu, L. Chen, and Y. Zhang, "DialogSum: A Real-Life Scenario Dialogue Summarization Dataset," in *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2021, pp. 5062–5074. doi: [10.18653/v1/2021.findings-acl.449](https://doi.org/10.18653/v1/2021.findings-acl.449).
 - [11] C. Lundberg, L. Sánchez Viñuela, and S. Biales, "Dialogue Summarization using BART," in *Proceedings of the 15th International Conference on Natural Language Generation: Generation Challenges*, Association for Computational Linguistics, Jul. 2022, pp. 121–125. [Online]. Available: <https://aclanthology.org/2022.inlg-genchal.17/>
 - [12] V. Chauhan, P. Roy, L. Dey, and T. Goel, "TCS_WITM_2022 @ DialogSum: Topic oriented Summarization using Transformer based Encoder Decoder Model," in *Proceedings of the 15th International Conference on Natural Language Generation: Generation Challenges*, Association for Computational Linguistics, Jul. 2022, pp. 104–109. [Online]. Available: <https://aclanthology.org/2022.inlg-genchal.15/>
 - [13] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "BERTScore: Evaluating Text Generation with BERT," *CoRR*, vol. abs/1904.0, Feb. 2020, doi: [10.48550/arXiv.1904.09675](https://doi.org/10.48550/arXiv.1904.09675).
 - [14] P. Verma and A. Verma, "A Review on Text Summarization Techniques," *J. Sci. Res.*, vol. 64, no. 01, pp. 251–257, 2020, doi: [10.37398/JSR.2020.640148](https://doi.org/10.37398/JSR.2020.640148).
 - [15] 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](https://doi.org/10.1016/j.jksuci.2020.05.006).
 - [16] N. Almohaimed and A. M. Azmi, "Abstractive text summarization: A comprehensive survey of techniques, systems, and challenges," *Comput. Sci. Rev.*, vol. 57, p. 100762, Aug. 2025, doi: [10.1016/j.cosrev.2025.100762](https://doi.org/10.1016/j.cosrev.2025.100762).
 - [17] H. Shakil, A. Farooq, and J. Kalita, "Abstractive text summarization: State of the art, challenges, and improvements," *Neurocomputing*, vol. 603, p. 128255, Oct. 2024, doi: [10.1016/j.neucom.2024.128255](https://doi.org/10.1016/j.neucom.2024.128255).
 - [18] T. Bao, H. Zhang, and C. Zhang, "Enhancing abstractive summarization of scientific papers using structure information," *Expert Syst. Appl.*, vol. 261, p. 125529, Feb. 2025, doi: [10.1016/j.eswa.2024.125529](https://doi.org/10.1016/j.eswa.2024.125529).
 - [19] O. Büyüç, "A Comprehensive Evaluation of Large Language Models for Turkish Abstractive Dialogue Summarization," *IEEE Access*, vol. 12, pp. 124391–124401, Sep. 2024, doi: [10.1109/ACCESS.2024.3454342](https://doi.org/10.1109/ACCESS.2024.3454342).
 - [20] Y. Altameemi, M. Altamimi, A. Alkhalil, D. Uliyan, and R. F. Mansour, "Text summarization method of argumentative discourse by combining the BERT-transformer model," *Front. Artif. Intell.*, vol. 8, Nov. 2025, doi: [10.3389/frai.2025.1654496](https://doi.org/10.3389/frai.2025.1654496).
 - [21] Z. Nie et al., "Domain-Specific Multi-Document Political News Summarization Using BART and ACT-GAN," *IEEE Access*, vol. 13, pp. 143737–143750, 2025, doi: [10.1109/ACCESS.2025.3598058](https://doi.org/10.1109/ACCESS.2025.3598058).
 - [22] C. Adhik, S. Sri Lakshmi, and C. Muralidharan, "Text summarization using BART," 2024, p. 020038. doi: [10.1063/5.0217004](https://doi.org/10.1063/5.0217004).
 - [23] K. Jin, Y. Wang, L. Santos, H. G. Oliveira, X. Yang, and S. K. Im, "Role-aware adapters for dialogue summarization in Seq2Seq models," *Appl. Soft Comput.*, vol. 187, p. 114293, Feb. 2026, doi: [10.1016/j.asoc.2025.114293](https://doi.org/10.1016/j.asoc.2025.114293).
 - [24] A. R. Fabbri et al., "Improving Zero and Few-Shot Abstractive Summarization with Intermediate Fine-tuning and Data Augmentation," Apr. 2021, doi: [10.48550/arXiv.2010.12836](https://doi.org/10.48550/arXiv.2010.12836).
 - [25] 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](https://doi.org/10.1016/j.eswa.2020.113679).
 - [26] H. Lin and V. Ng, "Abstractive Summarization: A Survey of the State of the Art," *Proc. AAAI Conf. Artif. Intell.*, vol. 33, no. 01, pp. 9815–9822, Jul. 2019, doi: [10.1609/aaai.v33i01.33019815](https://doi.org/10.1609/aaai.v33i01.33019815).
 - [27] Y. Li, H. Su, X. Shen, W. Li, Z. Cao, and S. Niu, "DailyDialog: A Manually Labelled Multi-turn Dialogue Dataset," Oct. 2017, doi: [10.48550/arXiv.1710.03957](https://doi.org/10.48550/arXiv.1710.03957).
 - [28] K. Sun, D. Yu, J. Chen, D. Yu, Y. Choi, and C. Cardie, "DREAM: A Challenge Data Set and Models for Dialogue-Based Reading Comprehension," *Trans. Assoc. Comput. Linguist.*, vol. 7, pp. 217–231, Nov. 2019, doi: [10.1162/tac1_a_00264](https://doi.org/10.1162/tac1_a_00264).
 - [29] L. Cui, Y. Wu, S. Liu, Y. Zhang, and M. Zhou, "MuTual: A Dataset for Multi-Turn Dialogue Reasoning," Apr. 2020, doi: [10.48550/arXiv.2004.04494](https://doi.org/10.48550/arXiv.2004.04494).
 - [30] L. Hickman, S. Thapa, L. Tay, M. Cao, and P. Srinivasan, "Text Preprocessing for Text Mining in Organizational Research: Review and Recommendations," *Organ. Res. Methods*, vol. 25, no. 1, pp. 114–146, Jan. 2022, doi: [10.1177/1094428120971683](https://doi.org/10.1177/1094428120971683).
 - [31] N. Garg and K. Sharma, "Text pre-processing of multilingual for sentiment analysis based on social network data," *Int. J. Electr. Comput. Eng.*, vol. 12, no. 1, p. 776, Feb. 2022, doi: [10.11591/ijece.v12i1.pp776-784](https://doi.org/10.11591/ijece.v12i1.pp776-784).
 - [32] R. Sennrich, B. Haddow, and A. Birch, "Neural Machine Translation of Rare Words with Subword Units," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, K. Erk and N. A. Smith, Eds., Stroudsburg, PA, USA: Association for Computational Linguistics, Aug. 2016, pp. 1715–1725. doi: [10.18653/v1/P16-1162](https://doi.org/10.18653/v1/P16-1162).
 - [33] M. Lewis et al., "BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension," Oct. 2019, doi: <https://doi.org/10.48550/arXiv.1910.13461>.
 - [34] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "{BERT}: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of the 2019 Conference of the North, J. Burstein, C. Doran, and T. Solorio, Eds., Stroudsburg, PA, USA: Association for Computational Linguistics*, Jun. 2019, pp. 4171–4186. doi: [10.18653/v1/N19-1423](https://doi.org/10.18653/v1/N19-1423).
 - [35] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to Sequence Learning with Neural Networks," *Adv. Neural Inf. Process. Syst.*, vol. 27, Dec. 2014, doi: <https://doi.org/10.48550/arXiv.1409.3215>.
 - [36] A. Auriemma Citarella, M. Barbella, M. G. Ciobanu, F. De Marco, L. Di Biase, and G. Tortora, "Assessing the effectiveness of ROUGE as unbiased metric in Extractive vs. Abstractive summarization techniques," *J. Comput. Sci.*, vol. 87, no. November 2024, p. 102571, May 2025, doi: [10.1016/j.jocs.2025.102571](https://doi.org/10.1016/j.jocs.2025.102571).
 - [37] M. Barbella and G. Tortora, "Rouge Metric Evaluation for Text Summarization Techniques," *SSRN Electron. J.*, 2022, doi:

- [10.2139/ssrn.4120317](https://doi.org/10.2139/ssrn.4120317).
- [38] C.-Y. Lin, "ROUGE: A Package for Automatic Evaluation of Summaries," in *Association for Computational Linguistics*, Association for Computational Linguistics, Jul. 2004, pp. 74–81. [Online]. Available: <https://aclanthology.org/W04-1013/>
- [39] Halimah, Surya Agustian, and Siti Ramadhani, "Peringkasan teks otomatis (automated text summarization) pada artikel berbahasa indonesia menggunakan algoritma lexicrank," *J. CoSciTech (Computer Sci. Inf. Technol.*, vol. 3, no. 3, pp. 371–381, Dec. 2022, doi: [10.37859/coscitech.v3i3.4300](https://doi.org/10.37859/coscitech.v3i3.4300).
- [40] E. Davoodijam and M. Alambardar Meybodi, "Evaluation metrics on text summarization: comprehensive survey," *Knowl. Inf. Syst.*, vol. 66, no. 12, pp. 7717–7738, Dec. 2024, doi: [10.1007/s10115-024-02217-0](https://doi.org/10.1007/s10115-024-02217-0).