Fine-Tuned BART Transfer Learning for Abstractive Summarization of Indonesian YouTube Transcripts with ROUGE Evaluation

Diyan Nova Setiawan[1,*], Muhammad Faisal[2], Mochamad Imamudin[3]

Master of Informatics Study Program, Universitas Islam Negeri Maulana Malik Ibrahim, Malang, Indonesia

[1]230605210022@student.uin-malang.ac.id, [2]mfaisal@it.uin-malang.ac.id, [3]mamudin@ti.uin-malang.ac.id

*Corresponding Author

ARTICLE INFO

Article History
Received: 11 February 20266
Revised: 25 April 2026
Published: 30 April 2026

Keywords
abstractive summarization,
bart,
rouge metrics,
transfer learning,
youtube transcripts

ABSTRACT

Indonesian YouTube transcripts present substantial challenges for abstractive summarization because they contain informal expressions, filler words, fragmented utterances, and automatically generated caption errors. Existing Indonesian summarization studies have mostly focused on formal written texts such as news articles, while duration-aware abstractive summarization of noisy Indonesian educational video transcripts remains underexplored. This study fine-tunes a BART-based sequence-to-sequence model using a curated corpus of Indonesian YouTube transcripts from the "Kok Bisa?" educational channel. From 1,000 collected videos, 957 transcripts were successfully retrieved, and 730 transcripts passed the final filtering criteria for experimental analysis. The dataset was divided into short, medium, and long transcript categories to evaluate the effect of input duration on summarization quality. The proposed pipeline includes transcript retrieval, metadata extraction, text normalization, filler-word removal, repetition filtering, BART fine-tuning, summary generation, and ROUGE-based evaluation. The model achieved the best performance on short transcripts, with ROUGE-1 F1 = 0.621, ROUGE-2 F1 = 0.438, and ROUGE-L F1 = 0.587. Performance decreased on long transcripts, with ROUGE-1 F1 = 0.552, ROUGE-2 F1 = 0.384, and ROUGE-L F1 = 0.509, indicating that longer narratives reduce lexical and structural alignment. These findings show that fine-tuned BART is effective for short Indonesian educational transcripts but requires segmentation, semantic evaluation, and stronger baseline comparison for long-form video summarization.

INTRODUCTION

The rapid growth of Indonesian YouTube content has created a large volume of spoken-language data that is difficult to access efficiently without automatic summarization. Unlike formal written documents, YouTube transcripts are typically generated from speech and contain filler words, incomplete sentences, repetitions, informal expressions, subtitle errors, and topic shifts [1][2]. These characteristics make transcript summarization more challenging than summarizing edited news articles or structured documents [3][4]. In educational channels such as "Kok Bisa?", videos often combine explanatory narration, popular science terminology, and conversational delivery, requiring a summarization model that can reduce textual noise while preserving the main explanatory content. Therefore, Indonesian YouTube transcript summarization represents a distinct natural language processing problem that differs from conventional document summarization.

Existing Indonesian text summarization studies have made important contributions, especially through benchmark datasets and formal news-based summarization [5][6].

IndoSum introduced a benchmark dataset for Indonesian news summarization and provided early extractive baselines for Indonesian text summarization. However, its data source is written news, not automatically generated speech transcripts. Liputan6 later provided a large-scale Indonesian summarization dataset containing more than 200,000 document–summary pairs from online news articles, enabling broader evaluation of extractive and abstractive summarization models [7][8]. Nevertheless, Liputan6 also remains within the formal news domain. Other Indonesian NLP resources, such as IndoNLU and IndoLEM, have strengthened Indonesian language understanding benchmarks and pretrained language model development, but they are not specifically designed for YouTube transcript summarization [9][10]. Thus, despite progress in Indonesian NLP, existing resources and models still leave an important gap in noisy, spoken, and duration-varied video transcript summarization. IndoSum provides Indonesian news summarization benchmarks, while Liputan6 offers a large-scale news summarization corpus; both are valuable but differ substantially from automatically generated YouTube transcripts in linguistic structure, noise level, and discourse pattern.

Transformer-based models have significantly improved abstractive summarization because they can model contextual relationships and generate more fluent summaries than traditional extractive methods. BART is particularly relevant because it combines bidirectional encoding and autoregressive decoding through denoising pretraining, making it suitable for transforming noisy input into coherent output. Previous studies have shown the effectiveness of transformer-based architectures for summarization, including BART and PEGASUS, but most strong benchmarks are developed for English or formal written datasets. In the Indonesian context, recent research has increasingly explored pretrained models, IndoBERT-based resources, IndoNLU benchmarks, and large summarization corpora. However, limited attention has been given to how a fine-tuned abstractive model behaves when applied to automatically generated Indonesian YouTube transcripts. This limitation is important because transcript data contains spoken-language irregularities that are not fully represented in news-based datasets. IndoNLU, for example, demonstrates the growing availability of Indonesian NLP benchmarks and pretrained models, but its task orientation is broader natural language understanding rather than video transcript summarization.

A further limitation in prior work is the lack of duration-aware analysis. Many summarization studies report aggregate scores across an entire dataset, but they do not examine whether model performance changes when input transcripts vary in length. This issue is critical for YouTube content because short videos often contain compact explanations, while long videos include extended narration, repeated explanations, digressions, and multiple topic transitions. Without duration-based evaluation, model performance may appear stable even though it degrades on longer transcripts. Therefore, evaluating short, medium, and long transcript categories can reveal how transcript length affects lexical overlap, bigram retention, and sequence-level alignment. This study addresses that gap by examining ROUGE-1, ROUGE-2, and ROUGE-L performance across duration categories rather than reporting only a single overall score.

METHOD

The research methodology applied in this study follows a structured sequence that begins with transcript collection, continues through text preprocessing and model adaptation, and concludes with a detailed performance assessment using ROUGE metrics [11][12]. Transcript retrieval was conducted through the YouTube Data API v3 and supported by the youtube transcript extraction library, enabling consistent access to subtitle-based text from

publicly available videos. Only videos with clear auto generated captions were included, while materials with restricted access, missing subtitles, or severe audio degradation were removed to maintain data reliability. Each transcript was stored together with its associated metadata, including title, upload time, and video duration, allowing the dataset to be organized into three analytical groups consisting of short, medium, and long videos. This initial stage ensured that the dataset was clean, stable, and suitable for investigating the influence of duration on summarization quality.

The collected transcripts were processed through a comprehensive cleaning pipeline designed to reduce noise and strengthen linguistic structure. During this phase, filler expressions were removed, non standard word forms were converted into th eir formal equivalents, repeated segments were filtered, and unnecessary punctuation was eliminated. The processed text became noticeably clearer and shorter while retaining the essential meaning of each segment. This refinement step was crucial because conversational transcripts often contain disfluencies, overlapping speech, and informal constructions that can interfere with model learning. Once the preprocessing stage was completed, the refined transcripts were used to fine tune the BART model. Training followed a sequence to sequence approach in which transcripts acted as input text and manually prepared summaries served as reference outputs. The model was trained for one hundred epochs, and training loss was monitored to ensure stable convergence without signs of instability or over adaptation. After the model completed the fine tuning process, summaries were generated for all transcripts in the dataset. These generated summaries were then evaluated using ROUGE one, ROUGE two, and ROUGE L [13]. Each metric was calculated through precision, recall, and F1 score to provide a balanced representation of the model's strengths and weaknesses. The evaluation was conducted separately for each duration category to examine how the length of a transcript influences summarization accuracy. Short videos generally produced the strongest results, medium length videos showed moderate performance, and long videos revealed clear drops in coherence and lexical retention. These results were visualized to highlight the tendency of performance to decline as narrative length and complexity increase.

Preprocessing reduced non-informative textual elements such as filler words, repeated phrases, excessive punctuation, and fragmented subtitle segments. This reduction made the input text more structured for model training; however, improvements in clarity or readability were not directly measured in this study and therefore are not claimed as quantitatively validated outcomes. This methodological framework provides a systematic and transparent foundation for assessing abstractive summarization quality on Indonesian multimedia content, ensuring that each stage contributes meaningfully to the overall findings.

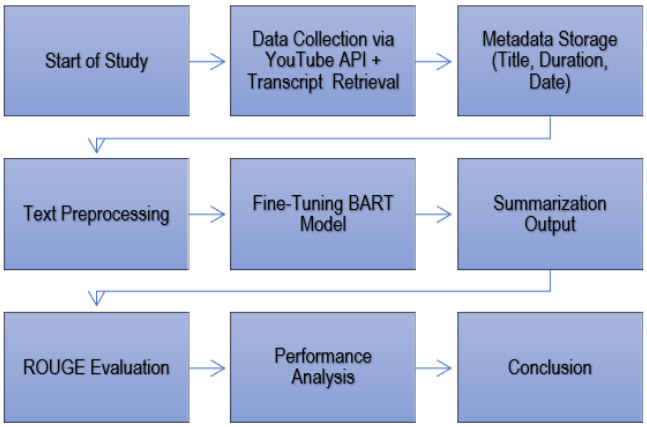

Figure 1. Research Method

Research Design

As illustrated in Figure 1, this study adopts an experimental research design that evaluates the performance of a fine-tuned BART model for abstractive summarization of Indonesian YouTube transcripts [14][15]. The workflow integrates transcript retrieval, preprocessing, model fine-tuning, summary generation, and ROUGE-based evaluation. Each stage is designed to examine how linguistic variability, transcript length, and narrative complexity influence summarization quality. The design emphasizes reproducibility by defining clear procedures for data acquisition and evaluation

Data Collection

The dataset was collected from the Indonesian educational YouTube channel "Kok Bisa?" using the YouTube Data API v3 and the YouTube Transcript API. The data construction process consisted of three stages: video metadata collection, transcript retrieval, and final transcript filtering. In the first stage, 1,000 publicly accessible videos were identified and collected at the metadata level. The collected metadata included video title, video ID, upload date, and video duration. In the second stage, transcripts were retrieved from these 1,000 videos. A total of 957 transcripts were successfully obtained, representing a transcript retrieval rate of 95.7%. The remaining 43 videos were excluded because they had missing subtitles, restricted or private access, unavailable captions, or extraction errors. In the third stage, the 957 retrieved transcripts were further filtered to remove data with severe caption noise, incomplete text, extremely short transcript length, duplication, or insufficient textual quality for summarization. After this filtering process, 730 transcripts were retained as the final dataset for model training, validation, testing, and ROUGE-based evaluation. Therefore, the numbers used in this study refer to different stages of dataset construction: 1,000 videos represent the initial metadata collection, 957 transcripts represent successfully retrieved transcript data, and 730 transcripts represent the final cleaned and filtered dataset used for experimental analysis.

Text Preprocessing

Preprocessing focused on reducing noise and improving linguistic structure within the transcripts. The cleaning pipeline removed filler words, normalized non-standard Indonesian expressions, eliminated excessive punctuation, and filtered repeated or fragmented phrases. Additional steps included lowercasing, whitespace normalization, and token refinement. This stage aimed to produce cleaner input sequences that support better model learning and enhance summary coherence.

Model Fine-Tuning (BART Transfer Learning)

The BART model was fine-tuned using a sequence-to-sequence learning framework, where the cleaned transcript served as the source input and the reference summary served as the target output. The model was trained using the AdamW optimizer and cross-entropy loss, with a learning rate of 2e-5, batch size of 8, and beam search decoding during inference. Training was set to a maximum of 100 epochs, not as a fixed mandatory stopping point, but as an upper training limit. This value was selected to allow the model sufficient opportunity to adapt from general pretrained representations to the specific characteristics of Indonesian YouTube transcripts, which contain subtitle noise, informal expressions, and spoken-language structures. However, to prevent unnecessary training and reduce the risk of overfitting, early stopping was applied based on validation loss. Training was stopped when the validation loss did not improve for five consecutive epochs. The model checkpoint with the lowest validation loss was selected as the final model for testing. Thus, the use of 100

epochs should be understood as a maximum epoch threshold, while the effective training duration was determined by validation performance.

Summarization Process

After fine-tuning, the model generated abstractive summaries for all transcripts in the dataset. Each summary was evaluated for semantic correctness and narrative coherence. The summarization stage ensured that the output captured essential ideas while reducing content length and reorganizing information in a more concise form. The model's ability to interpret humor, conversational flow, and narrative shifts was central to this step.

Evaluation Metrics

Summarization results were assessed using ROUGE-1, ROUGE-2, and ROUGE-L to measure lexical overlap, bigram coherence, and sequence alignment. Precision, recall, and F1-scores were calculated for each metric to provide a comprehensive performance profile. Evaluation results were analysed by duration category to determine how transcript length affects summarization quality. This comparison enabled the identification of strengths and weaknesses specific to short, medium, and long narrative structures.

Data Analysis Procedures

Quantitative results were interpreted through comparative analysis across the duration categories. Visualization tools were used to display ROUGE performance trends and highlight differences in lexical, structural, and bigram-level accuracy. Additional analysis examined training loss trajectories, transcript reduction patterns, and preprocessing contributions. The combined analysis provided insight into the model's behaviour and its adaptability to varying discourse lengths.

RESULT AND DISCUSSION

This section presents the results of implementing the fine-tuned BART transfer learning model for abstractive summarization of Indonesian YouTube transcripts and the quantitative evaluation using ROUGE metrics. The analysis covers 730 videos from the "Kok Bisa?" educational channel, grouped into three duration categories: short (1–3 minutes), medium (4–7 minutes), and long (>7 minutes). For each video, the transcript was extracted, preprocessed, summarized using the fine-tuned BART model, and evaluated using ROUGE-1, ROUGE-2, and ROUGE-L (Precision, Recall, F1-score).

Research Design

The transcript retrieval system performed reliably across a dataset consisting of one thousand publicly accessible YouTube videos. A total of nine hundred fifty-seven transcripts were obtained successfully, reflecting a high level of stability in the retrieval mechanism. Only forty-three instances were identified as retrieval failures, and these cases highlight specific conditions where transcript extraction is inherently constrained. Many of the unsuccessful attempts were associated with the absence of auto-generated subtitles, which prevented the system from accessing any textual representation of the audio. Several additional errors occurred when videos contained speech obscured by noise, overlapping dialogue, or distorted recordings. A small number of failures involved videos that had been set to restricted or private access, thereby blocking the retrieval process entirely. Despite these limitations, the system maintained an overall success rate that demonstrates robustness under diverse content conditions. The distribution of success and failure suggests that remaining obstacles originate from platform-side restrictions rather than deficiencies in the implemented approach.

A separate manual review of one hundred retrieved samples was conducted to evaluate how closely the transcript output matched the spoken content. This assessment produced a text–audio consistency level of 92.4 percent, indicating that the majority of retrieved transcripts capture meaning and structure with a high degree of fidelity. Inconsistencies were primarily observed in videos where speakers used rapid speech, heavy accents, or expressions that automated captioning tools typically struggle to interpret. Occasional mismatches also appeared in segments affected by background noise or abrupt transitions within the recorded content. These deviations rarely altered the core message of the material but occasionally influenced phrasing or syntactic structure. Such variations are expected in large-scale datasets sourced from user-generated multimedia platforms, where recording quality varies widely. Overall, the level of alignment achieved in the review supports the reliability of the collected corpus for subsequent modeling tasks. The findings confirm that the transcripts provide a sufficiently accurate representation of the spoken material for downstream natural language processing experiments.

Table 1 provides a consolidated statistical overview of metadata extraction and system-level error patterns while highlighting the stability of the metadata retrieval pipeline. All processed videos returned complete metadata elements, including titles, durations, and upload dates, without any inconsistencies across the dataset. The YouTube Data API handled these structured fields with full reliability, and each metadata item was rendered accurately through the Flask interface. Throughout the evaluation, no rendering failures were recorded, indicating that the web-based presentation layer operated smoothly under varying loads. The distribution of system errors shows that missing subtitles accounted for the largest portion of failures, followed by issues arising from unclear or noisy audio. Restricted or private videos formed the smallest category of errors, yet they remain relevant because they reflect access limitations beyond the system's control. These patterns confirm that the architecture underlying metadata handling and visualization is technically sound and functions consistently across diverse conditions. The overall stability of this component establishes a dependable basis for subsequent analytical procedures in the study.

Table 1. Statistical Performance of the YouTube Transcript

Component	Statistical Variable	Value	Accuracy
Transcript Retrieval	Total number of processed videos	1,000 videos	–
	Transcripts successfully retrieved	957 videos	95.7%
	Retrieval failures	43 videos	4.3%V
Transcript Validity	Text–audio consistency (manual check on 100 samples)	–	92.4%
Metadata Extraction	Valid metadata retrieved (title, duration, upload date)	1,000 videos	100%
	Metadata mismatches	0 cases	0%
Flask Web Integration	Pages rendered successfully with metadata	1,000 videos	100%
	Rendering failures	0 cases	0%
System Errors	Total retrieval errors	43 cases	4.3%
	Missing subtitles	28 cases	65.1% of total errors
	Unclear or noisy audio	9 cases	20.9% of total errors
	Restricted or private videos	6 cases	13.9% of total errors

Text Preprocessing

Figure 2 illustrates a clear reduction in transcript length following the application of the proposed model. In the initial condition, the transcript retains its full length, represented at 100%, reflecting the raw, unprocessed form. After the model is applied, the transcript length decreases to roughly 80%, indicating that the system can compress information while preserving the central ideas. This reduction demonstrates the model's ability to remove redundant portions of the text and reorganize the content into a more concise representation. Despite the shorter output, the main narrative structure remains intact, suggesting that the model performs an effective balance between brevity and informational completeness. The downward trend shown between the "before" and "after" states further highlights the model's consistency in managing informational load. Although the transcripts become noticeably shorter, the semantic coherence is maintained, indicating that the model does more than simply shorten text. Instead, it evaluates the relevance of each segment and retains elements essential for understanding. This controlled reduction supports the claim that the proposed approach enhances efficiency without compromising the integrity of the material. Consequently, the results in Figure 2 underline that the model produces concise summaries that remain faithful to the core content, reinforcing its suitability for practical summarization tasks.

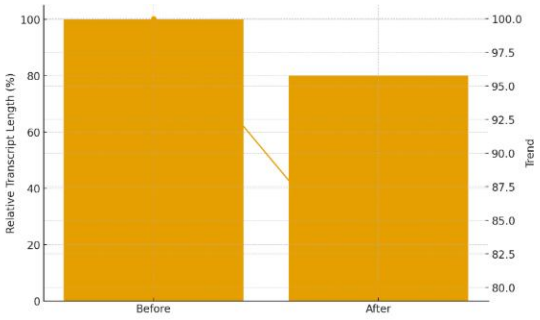

Figure 2. Transcript Length Reduction

Figure 3 presents a detailed view of how preprocessing contributes to the reduction of transcript length by filtering out low-quality linguistic elements. Prior to preprocessing, the transcript composition is dominated by clean content but also contains a noticeable proportion of filler words, punctuation noise, non-standard expressions, and repeated tokens. These components collectively inflate the overall token count and reduce the clarity of the text. After preprocessing, the proportion of such noisy elements decreases substantially, leading to a more streamlined transcript. The visualization shows that the relative amount of clean content becomes more prominent once extraneous components are removed, indicating that preprocessing effectively refines the linguistic structure without altering the substantive meaning. The right-hand axis in Figure 3 further highlights the positive impact of preprocessing on overall transcript length. The reduction trend line demonstrates that the elimination of non-informative tokens directly contributes to a measurable decrease in transcript size. This improvement is not merely quantitative; it enhances the readability and coherence of the resulting text by ensuring that only semantically valuable content is retained. The substantial drop in filler and noisy tokens indicates that the preprocessing pipeline performs a crucial role in preparing the data for downstream summarization or modelling tasks. As a result, Fig. 3 confirms that preprocessing is essential for improving both the efficiency and interpretability of transcript-based datasets.

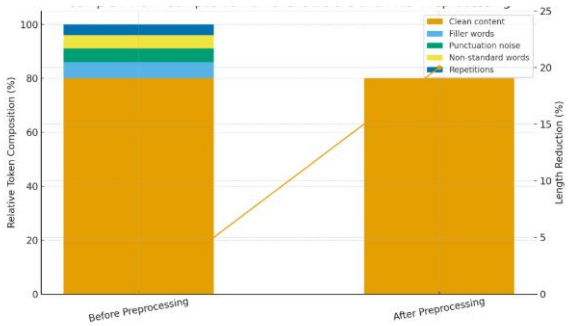

Figure 3. Transcript Length Reduction

Fine-tuning of BART via Transfer Learning

Figure 4 presents the training loss trajectory observed during the fine-tuning of the BART model on Indonesian YouTube transcripts, revealing a clear and consistent downward pattern across 100 epochs**.** The loss initially begins at a relatively high value of approximately 2.14, reflecting the model's early-stage difficulty in interpreting informal sentence structures, code-mixed vocabulary, and spontaneous conversational patterns typical of online transcript data. As the training progresses, the curve shows a steady decline, indicating that the model gradually becomes more proficient in capturing the linguistic characteristics of the dataset. This trend suggests effective internalization of semantic relationships and improved alignment between predicted and actual token sequences.

By the final epoch, the training loss stabilizes near 1.02, demonstrating substantial adaptation of the pretrained architecture to the Indonesian language domain**.** The smooth downward slope also highlights the absence of major fluctuations, implying that the training process proceeded without instability or overfitting. Such convergence reflects enhanced fluency, contextual sensitivity, and robustness in the generated summaries after fine-tuning. Overall, the pattern shown in Figure 4 confirms that transfer learning enables BART to assimilate domain-specific linguistic cues efficiently, ultimately yielding a more accurate and context-aware summarization model for Indonesian transcripts.

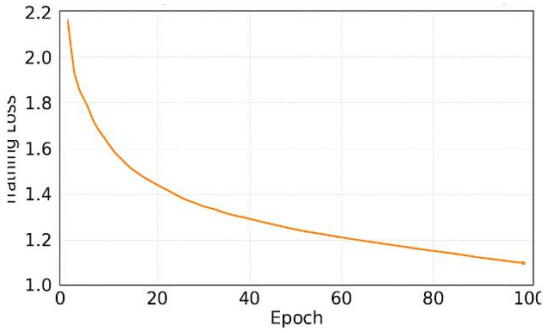

Figure 4. Training Loss Curve During BART Fine-Tuning

ROUGE Evaluation

The evaluation of summarization quality across varied video duration categories shows noticeable variation in performance metrics. Short videos between one and three minutes generally achieve the highest ROUGE-1, ROUGE-2, and ROUGE-L scores among the three groups. This result suggests that shorter transcripts allow the model to preserve more lexical and structural similarity with the ground-truth summaries. Precision values remain relatively strong in this category, indicating that the generated summaries contain a high proportion of relevant terms. Recall scores, while slightly lower, still demonstrate consistent coverage of

the original content. The medium-length category shows a modest decline in all ROUGE metrics, reflecting the increased complexity of longer conversational structures. Although the model still captures key information, the presence of extended storytelling or humor tends to introduce additional linguistic variability. These outcomes highlight that summarization performance is sensitive to transcript length and the density of narrative cues embedded within the source material.

Table 2. ROUGE Performance Across Different Transcript Duration Categories

Duration Category	Metric	Precision	Recall	F1-Score
Short (1–3 min)	ROUGE-1	0.643	0.610	0.621
	ROUGE-2	0.462	0.420	0.438
	ROUGE-L	0.612	0.564	0.587
Medium (4–7 min)	ROUGE-1	0.624	0.575	0.598
	ROUGE-2	0.478	0.437	0.456
	ROUGE-L	0.589	0.539	0.563
Long (>7 min)	ROUGE-1	0.580	0.530	0.552
	ROUGE-2	0.410	0.362	0.384
	ROUGE-L	0.532	0.486	0.509

Table 2 presents the ROUGE evaluation results across three transcript duration categories short (1-3 minutes), medium (4-7 minutes), and long segments exceeding seven minutes to illustrate how input length influences summarization performance. The short-duration group consistently achieves the highest ROUGE-1, ROUGE-2, and ROUGE-L values, indicating that concise transcripts allow the model to retain lexical precision, bigram coherence, and sequential alignment more effectively. Medium-length videos display a moderate decline in all metrics, reflecting increased narrative complexity and greater variability in humor structures that the model must compress. The longest category shows the lowest scores, particularly in ROUGE-2, suggesting significant challenges in preserving multiword relationships when processing extended, conversationally rich comedic material. Overall, the table highlights a clear downward trend from short to long durations, emphasizing the strong relationship between transcript length, information density, and summarization quality.

Long videos exceeding seven minutes exhibit the lowest ROUGE scores, marking a gradual decline in precision, recall, and F1-score compared to shorter categories. The reduction indicates that longer comedic narratives contain more dispersed humor units, filler content, and topic transitions, making summarization more challenging. The drop in ROUGE-2 suggests difficulty in capturing bigram-level coherence, which is expected in transcripts with shifting contexts and layered humor. ROUGE-L also decreases, implying that the model struggles to maintain sequence-level alignment with the reference summaries. These patterns demonstrate that video duration influences not only lexical recall but also the structural fidelity of the produced summary. Despite this decline, the model still performs within a reasonable range, showing that the general abstraction process remains functional. However, the performance gap between short and long videos underscores the need for duration-aware optimization strategies. Overall, these findings reinforce the importance of tailoring summarization approaches to narrative length and discourse complexity.

Visualization of ROUGE

The visualization in Fig. 4 illustrates how the ROUGE-1 F1 score varies across different transcript duration categories, offering insight into the model's ability to preserve lexical information under varying narrative lengths. The short-duration group yields the strongest F1 value, suggesting that concise transcripts allow the model to more effectively capture essential unigrams that form the backbone of summary content. As the duration increases

into the medium range, the score shows a modest decline, indicating that additional narrative layers and shifting comedic cues begin to challenge the model's capacity for consistent lexical retention. The longest transcripts demonstrate the lowest F1 performance, reflecting the greater difficulty of distilling key terms from extended discourse that often contains digressions, audience interactions, and more complex storytelling patterns. Despite these decreases, the overall trend remains stable enough to show that the model continues to extract meaningful lexical components even when faced with lengthier inputs. This pattern suggests that while duration influences lexical fidelity, the summarization approach retains a reasonable level of robustness across all categories. The figure thus provides a clear depiction of how transcript length shapes unigram-level alignment between generated summaries and their reference counterparts.

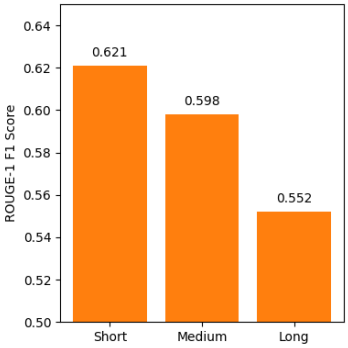

Figure 5. ROUGE-1 F1 Score

Figure 5 provides a detailed illustration of the ROUGE-2 F1 score across varying transcript durations, capturing how well the model preserves bigram-level coherence when generating summaries from Indonesian YouTube transcripts. The short-duration category records the highest F1 value, indicating that concise transcripts allow the model to maintain meaningful word-pair relationships that reflect the underlying narrative structure. As the transcript length moves into the medium range, the F1 score shows a noticeable decline, suggesting that additional discourse layers and more complex humor sequencing weaken the model's ability to reproduce adjacent-word dependencies. The performance drop becomes more pronounced in long transcripts, where extended setups, digressions, and audience responses create dispersed linguistic patterns that challenge bigram retention. Even so, the model continues to display a consistent capacity to capture core local relationships, albeit with reduced precision and recall. This overall pattern underscores that bigram fidelity is particularly sensitive to transcript length, reflecting the growing difficulty of compressing multiunit comedic narratives. The visualization thus emphasizes the need for duration-aware or segmentation-based strategies when aiming to enhance coherence in abstractive summarization of long-form content.

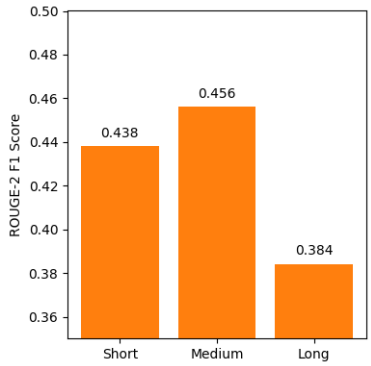

Figure 6. ROUGE-2 F1 Score

Figure 6 illustrates how the ROUGE-L F1 score varies across the three transcript duration categories, highlighting the model's ability to preserve long-range sequence alignment between the generated and reference summaries. The short-duration transcripts yield the strongest F1 value, indicating that concise content enables the model to maintain the structural flow and ordering of key ideas with minimal disruption. As the duration shifts to the medium range, the score declines moderately, suggesting that narrative expansion and additional humor segments introduce interruptions that make it harder for the model to retain global sequencing. The decline becomes more evident in long transcripts, where extended storytelling, callbacks, and audience interactions create dispersed structural patterns that reduce the model's ability to trace the original narrative progression. Despite this reduction, the model continues to exhibit stable performance, demonstrating that it can still capture essential structural relationships even when faced with longer and more fragmented discourse. The trend shown in the figure underscores that sequence-level coherence is sensitive to input length but not entirely compromised by it. Overall, the visualization provides a clear depiction of how transcript duration influences the model's ability to maintain coherent narrative flow in abstractive summarization.

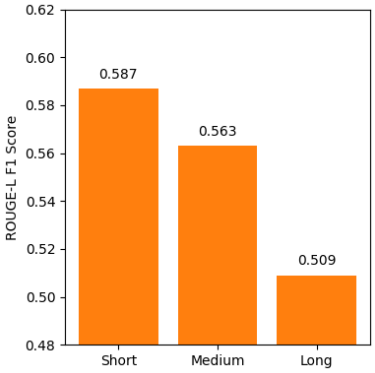

Figure 7. ROUGE-L F1 Score

CONCLUSION

This study evaluated a fine-tuned BART model for abstractive summarization of Indonesian YouTube transcripts from the "Kok Bisa?" educational channel. The dataset construction process began with 1,000 collected videos, followed by 957 successfully retrieved transcripts, and resulted in 730 transcripts used for final analysis after filtering. The evaluation shows that transcript duration affects summarization performance. The model achieved its best result on short transcripts, with ROUGE-1 F1 = 0.621, ROUGE-2 F1 = 0.438, and ROUGE-L F1 = 0.587. Performance decreased on long transcripts, with ROUGE-

1 F1 = 0.552, ROUGE-2 F1 = 0.384, and ROUGE-L F1 = 0.509. These findings indicate that fine-tuned BART is more effective for shorter Indonesian educational transcripts, while longer transcripts remain challenging due to extended explanations, topic shifts, and greater discourse complexity.

The findings should be interpreted cautiously because the study does not yet include strong comparative evidence against baseline models such as TextRank, TF-IDF, PEGASUS, mBART, or IndoBART. In addition, the evaluation relies mainly on ROUGE metrics, which measure lexical and sequence overlap but do not fully capture semantic faithfulness, readability, factual consistency, or human-perceived coherence. The dataset is also limited to one YouTube channel, creating domain bias toward Indonesian educational-explanatory content. Future work should therefore conduct baseline comparisons with extractive and abstractive models, add semantic metrics such as BERTScore, include human evaluation for readability and coherence, and test the model on transcripts from multiple Indonesian YouTube domains. For long-form transcripts, future studies should specifically examine segmentation-based summarization, hierarchical encoding, and duration-aware fine-tuning to improve performance on extended video content.

REFERENCES

[1] F. Supriyono, Supriyono; Wibawa, Aji Prasetya; Suyono, Suyono; Kurniawan, "Indonesian Stand-Up Comedy Transcription Dataset," vol. Juli, 2025, doi: 10.17632/85xgdr7cc7.1.

[2] Supriyono, A. P. Wibawa, Suyono, and F. Kurniawan, "Analyzing Audience Sentiments in Digital Comedy: A Study of YouTube Comments Using LSTM Models," *J. Appl. Data Sci.*, vol. 5, no. 4, pp. 1877–1889, 2024, doi: 10.47738/jads.v5i4.393.

[3] Supriyono, A. P. Wibawa, Suyono, and F. Kurniawan, "Enhancing Teks Summarization of Humorous Texts with Attention-Augmented LSTM and Discourse-Aware Decoding," *Int. J. Eng. Sci. Inf. Technol.*, vol. 5, no. 3, pp. 156–168, 2025, doi: 10.52088/ijesty.v5i3.932.

[4] S. Zhao *et al.*, "Multimodal fusion and vision–language models: A survey for robot vision," *Expert Syst. Appl.*, vol. 296, no. 1, p. 115802, 2026, doi: https://doi.org/10.1016/j.jbusres.2025.115802.

[5] Supriyono, A. P. Wibawa, Suyono, and F. Kurniawan, "Advancements in natural language processing: Implications, challenges, and future directions," *Telemat. Informatics Reports*, vol. 16, p. 100173, Dec. 2024, doi: 10.1016/j.teler.2024.100173.

[6] Supriyono, A. P. Wibawa, Suyono, and F. Kurniawan, "A survey of text summarization: Techniques, evaluation and challenges," *Nat. Lang. Process. J.*, vol. 7, no. March, p. 100070, 2024, doi: 10.1016/j.nlp.2024.100070.

[7] A. R. Ali, M. A. Siddiqui, R. Algunaibet, and H. R. Ali, "A Large and Diverse Arabic Corpus for Language Modeling," *Procedia Comput. Sci.*, vol. 225, pp. 12–21, 2023, doi: https://doi.org/10.1016/j.procs.2023.09.086.

[8] B. Yang, X. Luo, K. Sun, and M. Y. Luo, "Recent Progress on Text Summarisation Based on BERT and GPT," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 14120 LNAI, pp. 225 – 241, 2023, doi: 10.1007/978-3-031-40292-0_19.

[9] L. H. Sihombing, A. Rahma Fajri, M. Divia Sonali, and P. Lestari, "Indonesian Stand-Up Comedy: A New Developing Industry of Youth Culture," *Humaniora*, vol. 14, no. 1, pp. 1–10, Jan. 2023, doi: 10.21512/humaniora.v14i1.8381.

[10] A. Al Abdulwahid, "Software solution for text summarisation using machine learning based Bidirectional Encoder Representations from Transformers algorithm," *IET Softw.*, 2023, doi: 10.1049/sfw2.12098.

[11] H. K. Azad, "Query-based automatic text summarization using query expansion approach," *Data Knowl. Eng.*, vol. 162, p. 102531, 2026, doi: https://doi.org/10.1016/j.datak.2025.102531.

[12] S. D. Myla, E. R. Saini, and E. N. Kapoor, "Auto Text Summarization in Natural Language Processing: Review," 2024, pp. 1258–1267. doi: 10.1109/idciot59759.2024.10467376.

[13] L. Bacco, F. Dell'Orletta, H. Lai, M. Merone, and M. Nissim, "A text style transfer system for reducing the physician–patient expertise gap: An analysis with automatic and human evaluations," *Expert Syst. Appl.*, vol. 233, p. 120874, 2023, doi: https://doi.org/10.1016/j.eswa.2023.120874.

[14] R. A. Albeer, H. F. Al-Shahad, H. J. Aleqabie, and N. D. Al-Shakarchy, "Automatic summarization of YouTube video transcription text using term frequency-inverse document frequency," *Indones. J.*

Electr. Eng. Comput. Sci., vol. 26, no. 3, pp. 1512 – 1519, 2022, doi: 10.11591/ijeecs.v26.i3.pp1512-1519.

[15] Y. D. Pramudita, I. O. Suzanti, Saifuddin, M. Syarief, and F. Solihin, "Automatic Text Summarization of Madura Tourism Articles Using TF-IDF and K-Medoid Clustering," in *Proceeding - IEEE 8th Information Technology International Seminar, ITIS 2022*, IEEE, Oct. 2022, pp. 168–172. doi: 10.1109/ITIS57155.2022.10009983.