Journal of Computer Engineering, Electronics and Information Technology (COELITE)

Journal homepage: https://ejournal.upi.edu/index.php/COELITE

Improving Transformer Performance for Text Summarization in Video Transcription

Rizky Dwi Putra[1], Aldy Rialdy Atmadja[2], Yana Aditia Gerhana[3]*

[1,2,3]Faculty of Science and Technology, UIN Sunan Gunung Djati Bandung, Indonesia

Correspondence E-mail: 1217050127@student.uinsgd.ac.id

ABSTRACT

In conveying information today, it can take the form of online video content. The rapid growth of online video content has created a strong need for automatic text summarization to improve information efficiency. Summarization is important because it allows audiences to quickly capture the essence of lengthy materials, reduces information overload, and ensures that key points can be accessed without going through the entire content. This study explores the use of Whisper Turbo for transcription and mT5 for summarizing Indonesian-language YouTube videos. Whisper Turbo produces accurate transcriptions, although the results vary depending on audio quality and topic complexity. The transcribed text is then summarized using mT5, which achieves strong performance with a ROUGE-1 F1 score of 54.13% and a ROUGE-L score of 49.39%. These findings indicate that mT5 outperforms the standard T5 model despite using less training data. Overall, the combination of Whisper Turbo and mT5 offers an effective solution for generating concise and reliable summaries of video content, with broad potential applications in education, journalism, and digital documentation.

© 2025 Universitas Pendidikan Indonesia

ARTICLE INFO

Article History:
Submitted/Received 20 Aug 2025
First Revised 16 Sep 2025
Accepted 20 Sep 2025
First Available online 01 Oct 2025
Publication Date 01 Oct 2025

Keyword:
Machine Learning,
Deep Learning,
Text Summarization,
Natural Language Processing (NLP)

1. INTRODUCTION

The development of digital technology has brought about significant changes in various aspects of human life, including how information is produced, disseminated, and consumed. Along with the industrial revolution and digital transformation, society is faced with new challenges in managing the rapid and massive flow of information. One tangible impact of this progress is the increased use of digital media, such as video, as the primary means of conveying information [1]. Video-based media has become popular because it can present content in a visual and engaging way [2]. However, to ensure the effectiveness of information delivery, it is crucial to structure content concisely and relevantly, and avoid unnecessary information so that the main message can be clearly received by the audience [3].

Artificial intelligence (AI) and machine learning technologies offer innovative new approaches to addressing this challenge. One application is video summarization, which can efficiently filter and present the core information from long videos [4]. By generating a summary in the form of a short text, this technique not only saves users time in understanding the video content but also helps reduce data size, thereby benefiting both users and computing systems [5].

As the growth of video content on various digital platforms accelerates, the need for technology capable of quickly filtering relevant information becomes increasingly urgent. Video summarization not only offers efficiency in obtaining information but also opens up significant opportunities across various sectors, such as education, media, corporations, and public services [6]. This technology enables users to grasp the core message of a video without having to watch its entire duration. Therefore, various studies have focused on developing intelligent models capable of understanding the context of video and audio, ensuring that the summarization process is not only fast but also maintains the meaning and accuracy of the information conveyed [7].

To achieve optimal video summarization, the use of the Whisper model is one of the key components in this research. Whisper is an audio transcription model developed using the Large-Scale Weak Supervision approach, trained using large-scale data from various sources, recording backgrounds, and languages. This model uses a transformer-based encoder-decoder architecture, with parameter sizes ranging from 39 million to 1.55 billion, depending on the model version. With support for over 98 languages, Whisper offers extensive multilingual capabilities [8]. The data format used is designed to enable the model to perform various tasks flexibly and reliably. Both the encoder and decoder in this model work efficiently to support accurate audio-to-text transcription [9]. Developed by OpenAI, Whisper known as an advanced Automatic Speech Recognition (ASR) system equipped with various automatic filtering mechanisms to improve transcription accuracy. With these capabilities, Whisper can generate text from speech more accurately, which is crucial in the initial stages of summarization [10].

Subsequently, text summarization is performed using the mT5 (Multilingual Text-to-Text Transfer Transformer) model, which has proven superior through deep training and evaluated using the ROUGE metric. The mT5 model leverages strong language representations through neural network architecture, enabling it to generate high-quality summaries. By training this model on the Liputan6 dataset, the system is expected to adapt more effectively to the characteristics of the Indonesian language [11]. Testing transformer models in summarization commonly uses Rouge because Rouge calculates the overlap between the summary

DOI: https://doi.org/10.17509/coelite.v4i2.89353
p- ISSN 2829-4157 e- ISSN 2829-4149

generated by the model and the reference summary through unigrams (ROUGE-1), bigrams (ROUGE-2), and longer sequences (ROUGE-L for longest common subsequence) [12].

Research on text summarization from video transcripts using transformer architectures has been widely conducted on both English and Indonesian videos. First, this study highlights how the use of Natural Language Processing (NLP) plays a crucial role in accelerating, simplifying, and enhancing the accuracy of content navigation on the YouTube platform. This study demonstrates that NLP can optimize the user experience in consuming digital content. Additionally, the study underscores the challenges and opportunities for future development of NLP-based summarization methods [13]. Additionally, the abstractive summarization approach was also tested using the BART (Bidirectional and Auto-Regressive Transformer) model with training based on the Liputan6 dataset, yielding evaluation scores of ROUGE-1 of 37.19, ROUGE-2 of 14.03, and ROUGE-L of 33.85 [14]. Previous studies that serve as references in this study utilized the T5 (Text-to-Text Transfer Transformer) model and the Whisper model for text summarization in video transcripts. This study utilized a dataset of 50,000 liputan6 transcripts, and the model was able to generate satisfactory summaries, with training evaluation scores of F1-score on the ROUGE metric at 39.23 (ROUGE-1), 13.17 (ROUGE-2), and 23.84 (ROUGEL) [15]. In that study, there was a limitation in token length to minimize chunking or segmentation of the text input, which led to suboptimal summarization results and model performance.

Referring to previous studies that have examined transformer-based automatic summarization technology and automatic speech recognition, this study improves a larger model and proposes an approach that enhances transformer performance while optimizing input token length to reduce text chunking. In addition, this study summarizes videos with a wider range of topics, thereby expanding the scope of evaluation and increasing the relevance of model results in the context of real-world applications.

2. METHODS

This research design began with the data collection stage, which involved collecting videos from YouTube as the main source. The videos were then transcribed using the Whisper model to produce transcript texts as the basis for the next process. The transcription texts will be evaluated for accuracy using the word error rate. The transcripts obtained were not used directly, but went through a preprocessing stage to organize the data into a more structured form suitable for model processing. The preprocessing used was tokenization, which divides the text into small parts such as words or sub-words.

After preprocessing, the transcript text was entered into the mT5 model, which had undergone a learning stage with an Indonesian language dataset taken from liputan6. The mT5 model is an advancement of the T5 model, where mT5 has a larger model size and broader language coverage. In Indonesian text summarization studies, mT5 excels because it is trained on a large dataset that includes Indonesian text [16]. This process produced a summary that was then tested to assess its quality and relevance. The evaluation is carried out using the ROUGE metric, which measures the level of similarity between the generated summary and the reference summary. The evaluation is carried out using the ROUGE metric, which measures the level of similarity between the generated summary and the reference summary. The ROUGE metric is commonly used to evaluate text summarization because it compares the generated summary with the reference summary, producing ROUGE-n and ROUGE-L scores [17]. The final stage of this research is to conclude the effectiveness of the

DOI: https://doi.org/10.17509/coelite.v4i2.89353
p- ISSN 2829-4157 e- ISSN 2829-4149

combination of Whisper and mT5 in generating informative and accurate text summaries from YouTube videos. The method used is described in **Figure 1**.

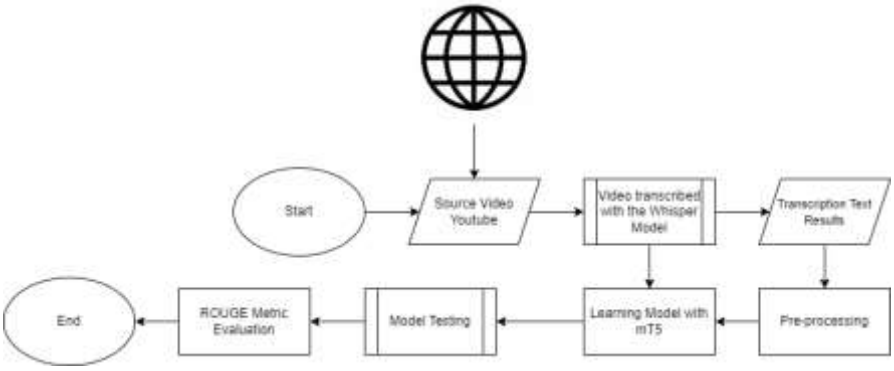

Figure 1. The method of preprocessing.

3. RESULTS AND DISCUSSION

3.1. Text Transcription with the Whisper Turbo Model

Transcription testing in this study was conducted using several YouTube videos with varying durations and topics. All videos were selectively chosen to represent variations in language structure and content length. The transcription process was carried out using the Whisper Turbo model, which is known for its high performance in terms of accuracy and processing speed [18]. Some of the videos used were already equipped with built-in transcripts from YouTube, allowing the calculation of the Word Error Rate (WER) by comparing the Whisper Turbo transcription results with the original transcripts. The WER test results are presented in **Table 1** as a basis for evaluating the effectiveness of Whisper Turbo in automatically generating text transcripts from online videos.

Table 1. Example Sentences Before Tokenization and Tokenization Results

Video Title	Duration (Minute)	Transcription Time (Seconds)	*Word Error Rate* (WER)
Different Tribes, Cultures, And Religions	5	25,09	28,37%
BPUPKI - History of the Investigating Body for Preparatory Efforts for Indonesian Independence	8	46,33	41,44%
History of the First Computer l The Discovery and Development of Computers	20	77,96	30,96%

Table 1 shows the transcription results of YouTube video content, including transcription time and word error rate (WER), which varied. The best WER result was shown in the first video because it was shorter than the other videos, and the sound in the video was clear. The lowest results are shown by the second video compared to the third video, where the second video has a shorter duration than the third video, but the third video has better

DOI: https://doi.org/10.17509/coelite.v4i2.89353
p- ISSN 2829-4157 e- ISSN 2829-4149

sound and information delivery than the second video. This indicates that WER is strongly influenced by the clarity of the audio in the video [19].

3.2. mT5 Model

3.2.1. Fine-Tuning Using the Liputan6 Dataset

The mT5 model was implemented through fine-tuning on the Liputan6 dataset. This dataset was selected because it consists of Indonesian online news articles covering diverse topics and written in a factual style, making it suitable for training summarization models on long input texts. The Liputan6 dataset contains 215,827 data points categorized into canonical and extreme sets. For the fine-tuning process, 21,000 canonical data points were used. The fine-tuning scenario applied an 80%–20% split, with 80% of the data used for training and 20% for testing [20]. The configuration used for fine-tuning is the mT5 model with 3 epochs, a batch size of 1, and a learning rate of 0.0001. This configuration was chosen to maintain a balance between model accuracy and efficiency and to ensure that the model can understand and learn data patterns optimally in the division scenario used.

3.2.2. Evaluation of ROUGE Metrics on mT5 and T5 Models

In this study, the mT5 model will be compared with the T5 model to determine which performs best in summarization. The mT5 model uses the Liputan6 dataset with 21,000 data points, while the T5 model uses 25,000 data points, both with the same configuration. After the fine-tuning process is done, the model will be evaluated using the ROUGE metric to see how well it works before it's used to summarize videos. The results are shown in **Table 2**.

Table 2. Comparison of mT5 and T5 model metric evaluations with Liputan6 dataset fine-tuning

Scenario	Score	ROUGE-1	ROUGE-2	ROUGE-L
mT5 model 21.000 dataset	*Recall*	**64,45**	**29,85**	**58,99**
	Precision	**47,42**	**20,66**	**43,22**
	F1-Score	**54,13**	**24,04**	**49,39**
T5 model 25.000 dataset	*Recall*	51,05	15,08	44,53
	Precision	30,64	8,58	25,79
	F1-Score	37,78	10,85	32,25

The evaluation using the ROUGE metric, as shown in **Table 2**, presents the recall, precision, and F1-score for ROUGE-1, ROUGE-2, and ROUGE-L. The highest values are observed in ROUGE-1, indicating that the model is quite effective in extracting important words from the video text transcription. On the ROUGE-1 metric, mT5 achieved an F1 score of 54.13%, with a recall of 64.45% and a precision of 47.42%. Similar results were also seen on the ROUGE-2 metric, where mT5 achieved an F1 score of 24.04%, much higher than the T5 model. On the ROUGE-L metric, mT5 again showed its superiority with an F1 score of 49.39%, compared to the T5 model. This comparison shows that even though it uses less training data, the mT5 model is capable of producing more accurate and relevant summaries, and has better efficiency than the T5 model in Indonesian text summarization tasks. This is because the mT5 model itself is a development of the T5 model, where the mT5 model has a larger model size and more language coverage. In processing Indonesian text, mT5 is superior because it is trained with a large dataset that includes Indonesian text.

DOI: https://doi.org/10.17509/coelite.v4i2.89353
p- ISSN 2829-4157 e- ISSN 2829-4149

3.2.3. Final Evaluation of Text Summarization Results from Whisper Model Transcription Using ROUGE Metrics

The final evaluation in this study was conducted on the text summarization results of 5-minute, 8-minute, and 20-minute videos that had been transcribed using the Whisper Turbo model. The transcription texts from each video were then summarized using the mT5 model, which had previously been fine-tuned with 21,000 articles from the Liputan6 dataset. In this process, the input length of each transcription was limited to 700 tokens to align with the maximum capacity that the mT5 model could process without causing context loss or the omission of important information. To assess the quality of the generated summaries, the ROUGE metric was used, which consists of ROUGE-1, ROUGE-2, and ROUGE-L. These three metrics are used to measure the extent of alignment between the model-generated summaries and the reference summaries based on the occurrence of single words, consecutive word pairs, and the longest common word sequences. The evaluation results of the text summarization from the Whisper Turbo model's transcriptions are shown in **Table 3**.

Table 3. Final Evaluation of Transcription Results Summary for Whisper Turbo Model and mT5 Model

Video Title	Duration (Minute)	Score	ROUGE-1	ROUGE-2	ROUGE-L
Different Tribes, Cultures, And Religions	5	*Recall*	44,28	10,14	27,14
		Precision	25,61	05,83	15,70
		F1-Score	32,46	07,40	19,89
BPUPKI - History of the Investigating Body for Preparatory Efforts for Indonesian Independence	8	*Recall*	**54,95**	**20,90**	**32,43**
		Precision	**37,88**	**14,37**	**22,36**
		F1-Score	**44,85**	**17,03**	**26,47**
History of the First Computer l The Discovery and Development of Computers	20	*Recall*	50,67	13,63	30,76
		Precision	22,35	06,00	13,57
		F1-Score	31,02	08,33	18,83

With the three videos tested, the second video showed the best performance in terms of summarization, discussing “BPUPKI - History of the Investigation Agency for the Preparation of Indonesian Independence.” This can be seen from the ROUGE-1 F1-Score value, which reached 44.85%, higher than the other two videos. In addition, the ROUGE-2 and ROUGE-L scores for this video were also the highest, at 17.03% and 26.47% respectively, indicating that the summary was not only relevant in terms of content but also more structured and linguistically coherent. Compared to the first and third videos, the second video has a better balance between Recall and Precision, resulting in a summary that not only covers important information but also remains concise and relevant. In this study, the good summarization results can be seen in the second video. This is because the second video is not too long (8 minutes), which minimizes the chunking process and helps preserve the information without losing it. In contrast, the third video, which has a much longer duration, produced less accurate summarization results due to excessive chunking, which caused a significant amount of information loss.

DOI: https://doi.org/10.17509/coelite.v4i2.89353
p- ISSN 2829-4157 e- ISSN 2829-4149

4. CONCLUSION

This study demonstrates that combining the Whisper Turbo model for transcription with mT5 for extractive summarization yields strong results in processing Indonesian-language video content. Whisper Turbo efficiently converts audio to text, while mT5 produces relevant and well-structured summaries, particularly for longer videos. Refinement with the Liputan6 dataset consistently outperforms T5, showing that multilingual models adapt better to Indonesian linguistic structures. The approach has broad applications in education, advertising, and other sectors where video is a key medium for information delivery. Future improvements could focus on expanding the model's ability to handle a wider variety of content, thereby enhancing its generalization capability. In addition, exploring advanced models such as LongT5 and applying hybrid summarization approaches, combined with user evaluation, holds strong potential for achieving near human-level summarization quality. Furthermore, this research can be further developed by incorporating datasets from Indonesian regional languages.

5. ACKNOWLEDGMENT

This study was funded by the UIN Sunan Gunung Djati Bandung. The research resources and facilities provided by the computer science department were instrumental in this study.

6. AUTHORS' NOTE

The authors declare that there is no conflict of interest regarding the publication of this article. Authors confirmed that the paper was free of plagiarism.

7. REFERENCES

[1] A. Faidlatul Habibah and I. Irwansyah, "Era Masyarakat Informasi sebagai Dampak Media Baru," *Jurnal Teknologi Dan Sistem Informasi Bisnis*, vol. 3, no. 2, pp. 350–363, Jul. 2021, doi: 10.47233/jteksis.v3i2.255.

[2] B. Rahmat and D. Darmiati, "Pengembangan Media Pembelajaran dengan Video Based Learning di Akademi Kebidanan Pelamonia," *Lectura: Jurnal Pendidikan*, vol. 12, no. 2, pp. 149–165, Aug. 2021, doi: 10.31849/lectura.v12i2.7268.

[3] H. Haerawan, W. Cale, and U. Barroso, "The Effectiveness of Interactive Videos in Increasing Student Engagement in Online Learning," *Journal of Computer Science Advancements*, vol. 2, no. 5, pp. 244–258, Oct. 2024, doi: 10.70177/jsca.v2i5.1322.

[4] H. Burhan Ul Haq, M. Asif, M. Asif, and M. Bin Ahmad, "Video Summarization Techniques: A Review Article in," *International Journal of Scientific & Technology Research*, 2021, [Online]. Available: www.ijstr.org

[5] H. Lai and X. Yan, "Multimodal Sentiment Analysis with Asymmetric Window Multi-Attentions," *Multimed Tools Appl*, vol. 81, no. 14, pp. 19415–19428, Jun. 2022, doi: 10.1007/s11042-021-11234-y.

[6] P. Saini, K. Kumar, S. Kashid, A. Saini, and A. Negi, "Video Summarization Using Deep Learning Techniques: A Detailed Analysis and Investigation," *Artif Intell Rev*, vol. 56, no. 11, pp. 12347–12385, Nov. 2023, doi: 10.1007/s10462-023-10444-0.

[7] J. Xie, X. Chen, S. Zhao, and S.-P. Lu, "Video Summarization via Knowledge-Aware Multimodal Deep Networks," *Knowl Based Syst*, vol. 293, p. 111670, Jun. 2024, doi: 10.1016/j.knosys.2024.111670.

[8] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust Speech Recognition via Large-Scale Weak Supervision," Dec. 2022, [Online]. Available: http://arxiv.org/abs/2212.04356

[9] R. S. A. Pratama and A. Amrullah, "Analysis Of Whisper Automatic Speech Recognition Performance on Low Resource Language," *Jurnal Pilar Nusa Mandiri*, vol. 20, no. 1, pp. 1–8, Mar. 2024, doi: 10.33480/pilar.v20i1.4633.

[10] Roissyah Fernanda Khoiroh, Eric Julianto, Safrizal Ardana Ardiyansa, H. A. Fajri, Aryaguna Abi Rafdi Yasa, and Brian Sangapta, "Implementasi Speech Recognition Whisper pada Debat Calon Wakil Presiden Republik Indonesia," *Explore*, vol. 14, no. 2, pp. 67–74, Jul. 2024, doi: 10.35200/ex.v14i2.115.

[11] S. Masri, Y. Raddad, F. Khandaqji, H. I. Ashqar, and M. Elhe-Nawy, "Transformer Models in Education: Summarizing Science Textbooks with AraBART, MT5, AraT5, and mBART."

[12] K. F. H. Holle, D. N. Munna, and E. W. Ekaputri, "Performance Evaluation of Transformer Models: Scratch, Bart, and Bert for News Document Summarization," *Jurnal Teknik Informatika (Jutif)*, vol. 6, no. 2, pp. 787–802, Apr. 2025, doi: 10.52436/1.jutif.2025.6.2.2534.

[13] Y. Singh, R. Kumar, S. Kabdal, and P. Upadhyay, "YouTube Video Summarizer Using NLP: A Review," *International Journal of Performability Engineering*, vol. 19, no. 12, pp. 817–823, Dec. 2023, doi: 10.23940/ijpe.23.12.p6.817823.

[14] G. Hartawan, D. Sa'adillah Maylawati, and W. Uriawan, "JIP (Jurnal Informatika Polinema) Halaman| 535 Bidirectional and Auto-Regressive Transformer (Bart) For Indonesian Abstractive Text Summarization".

[15] M. F. Fadlilah, A. R. Atmadja, and M. D. Firdaus, "Pemanfaatan Transformer untuk Peringkasan Teks: Studi Kasus pada Transkripsi Video Pembelajaran," *Technology and Science (BITS)*, vol. 6, no. 3, 2024, doi: 10.47065/bits.v6i3.6342.

[16] S. Nasution, R. Ferdiana, and R. Hartanto, "Towards Two-Step Fine-Tuned Abstractive Summarization for Low-Resource Language Using Transformer T5," *International Journal of Advanced Computer Science and Applications*, vol. 16, no. 2, 2025, doi: 10.14569/IJACSA.2025.01602120.

[17] A. Auriemma Citarella, M. Barbella, M. G. Ciobanu, F. De Marco, L. Di Biasi, and G. Tortora, "Assessing the effectiveness of ROUGE as unbiased metric in Extractive vs. Abstractive summarization techniques," *J Comput Sci*, vol. 87, May 2025, doi: 10.1016/j.jocs.2025.102571.

[18] K. K. R. Nareddy, S. Ternus, and J. Niebling, "Analyzing and Fine-Tuning Whisper Models for Multilingual Pilot Speech Transcription in the Cockpit," Jun. 2025, [Online]. Available: http://arxiv.org/abs/2506.21990

[19] S. Katkov, A. Liotta, and A. Vietti, "Benchmarking Whisper Under Diverse Audio Transformations and Real-Time Constraints," 2025, pp. 82–91. doi: 10.1007/978-3-031-77961-9_6.

[20] I. Gusti *et al.*, "Abstractive Text Summarization to Generate Indonesian News Highlight Using Transformers Model," *Journal of Information Systems and Informatics*, vol. 7, no. 2, 2025, doi: 10.51519/journalisi.v7i2.1082.

DOI: https://doi.org/10.17509/coelite.v4i2.89353
p- ISSN 2829-4157 e- ISSN 2829-4149