Journal of Artificial Intelligence and Engineering Applications

Website: https://ioinformatic.org/

15th October 2025. Vol. 5. No. 1; e-ISSN: 2808-4519

Whisper Model and Lexrank Algorithm as an Efficient Solution for YouTube Video Transcription and Summarization

Labid Muwaffaq Wicaksana[1], Mufti Ari Bianto[2]*, Ian Pasha[3], Adam Azzamul Husni Huda[4]

[1,2,3]*Computer Engineering, Faculty of Science, Technology, and Education, Universitas Muhammadiyah Lamongan, Indonesia*
wicaksanaabid@gmail.com[1], muftiari10@gmail.com[2*], pashanetwork3@gmail.com[3]

Abstract

This study developed an automated system for transcribing and summarizing YouTube video content to help users efficiently access important information from long-duration videos. The system only requires a video link as input, then automatically extracts the audio, transcribes it using the Whisper Small model, and generates a summary using the LexRank algorithm, which selects key sentences based on graph centrality. Transcription quality is evaluated using the Word Error Rate (WER) metric, with an average score of 0.3703 or approximately 37%, indicating a fairly good level of accuracy. Meanwhile, the summarization evaluation using ROUGE metrics yielded average F1-Scores of 33% for ROUGE-1, 10% for ROUGE-2, and 19% for ROUGE-L, reflecting the relevance of the generated summaries to manual references. The average transcription processing time is around 0.17 seconds per word, while the summarization process takes less than 1 second. All results—including transcriptions, summaries, and evaluation metrics—are automatically saved in CSV format. This system demonstrates stable performance and holds strong potential for various video-based knowledge extraction applications, such as in education, journalism, research, and digital documentation.

Keywords: *YouTube, Whisper, transcription, summarization, automation*

1. Introduction

In the digital era dominated by multimedia content, video has become one of the primary media for delivering information, particularly through platforms like YouTube. Educational videos, tutorials, seminars, and online lectures are now increasingly accessible to the general public [1]. However, as the duration and volume of available videos continue to grow, users often face challenges in efficiently absorbing information, especially when time is a limiting factor. This issue highlights the need for a system capable of quickly and concisely extracting and presenting important information from videos [2].The advancement of artificial intelligence (AI) and machine learning technologies has introduced innovative solutions to complex problems, including the processing and presentation of information from long-duration videos. One highly effective approach is video summarization, which involves condensing video content into its core information within a short time frame [3]. Such summaries help users understand the video content more easily and reduce data load [4].

Research interest in video summarization continues to grow and has become a widely discussed topic. This aligns with the increasing demand for fast and efficient access to information, particularly in the context of video-based learning, which is becoming more popular. Long videos often make it difficult for users to find truly relevant information, resulting in an ineffective search process. This condition drives the continued development of video summarization research as a solution to deliver concise, accessible information while maintaining the essence of the original video. Therefore, video summarization plays a significant role in improving efficiency and facilitating users in obtaining the information they need [5].Whisper by OpenAI is a key component in this system due to its capabilities as an accurate and efficient automatic speech recognition (ASR) model, supported by built-in filtering features to improve transcript quality [6]. For summarization, the LexRank algorithm is employed as a graph-based extractive method that selects the most representative sentences from the transcript based on sentence centrality within a graph network [7][8]. Recent studies show that LexRank remains relevant and competitive for summarizing Indonesian-language documents, even when compared with latent semantic analysis (LSA) and neural models [7]. Furthermore, the integration of LexRank with other machine learning approaches has shown promising results in multi-document scenarios using ROUGE as the evaluation metric [8].

Building on previous research, this study aims to develop an intelligent web-based system capable of performing automatic transcription and summarization of YouTube videos using only the video link as input. The system is designed to simplify the information extraction process from long-duration videos efficiently and accurately, without requiring complex manual intervention. The system implements the Whisper model for audio transcription and the LexRank algorithm to summarize the transcribed text. The transcription quality is evaluated using the Word Error Rate (WER) metric by comparing the output to official subtitles, while the summarization quality is assessed using ROUGE metrics to measure the similarity between the generated summaries and manual references. The entire process is designed to run on standard hardware without the need for GPU acceleration, with each process logged into a CSV file for documentation and evaluation

purposes.Through this approach, the study not only delivers a practical technical solution for general users but also opens opportunities for developing video-based knowledge extraction systems applicable in various domains such as education, research, journalism, and digital documentation.digital.

2. Research Methodology

This study employs a quantitative experimental approach aimed at developing and evaluating the performance of a web-based system for automating the transcription and summarization of YouTube videos. The system uses OpenAI's Whisper Small model to convert video audio into text, along with the LexRank algorithm to perform extractive text summarization based on sentence ranking. Evaluation is carried out using two approaches: transcription accuracy is measured with the Word Error Rate (WER) metric by comparing the transcription output to the official YouTube subtitles, while summarization effectiveness is assessed using ROUGE metrics by comparing the automatically generated summaries with manually created reference summaries.

2.1. System Architecture

The system architecture developed in this study is designed to perform fully automated transcription and summarization of YouTube videos from start to finish. Users only need to enter the YouTube video link as input, after which the system automatically downloads and extracts the audio using the yt_dlp library.

The extracted audio is then transcribed into text using the Whisper model. Once transcription is complete, the resulting text is summarized using the LexRank algorithm, which selects the most important sentences from the entire transcript.

The summary produced by the LexRank process is then evaluated using ROUGE metrics by comparing the automatically generated summaries with manually created reference summaries. All process data, including duration, transcription results, summary output, system specifications, and execution time, are automatically recorded in a CSV file for documentation and performance analysis. The detailed system flow is illustrated in Figure 1.

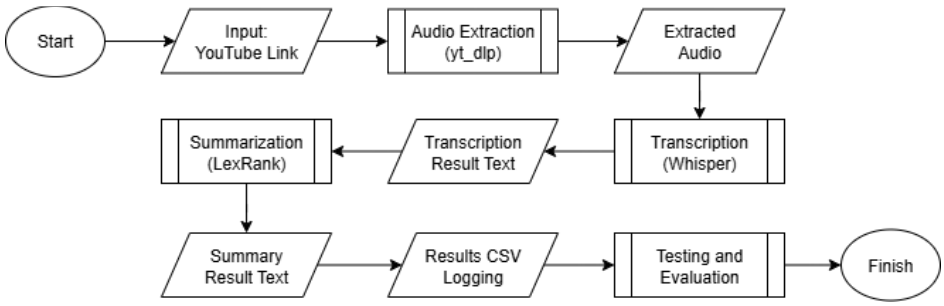

Fig. 1: Research System Flowchart

2.2. Audio Extraction Using the yt_dlp Library

The research objects used in this experiment consist of a collection of educational videos sourced from the YouTube platform. These videos were randomly selected with varying durations to ensure that system testing could cover a range of time scenarios and content complexity levels. The choice of educational videos was based on the need to evaluate the system's effectiveness in filtering important information from content with learning value.

Data collection was carried out through an audio extraction method using the yt_dlp library, which is used to directly download audio from YouTube videos [9]. Once the audio was successfully extracted, the resulting files served as input for the automatic transcription and summarization processes. This approach allowed the system to be tested directly on real-world data available on an open platform, with the goal of reflecting the system's performance in actual usage scenarios.

2.3. Whisper Model Text Transcription

Whisper is a model developed using a Large-Scale Weak Supervision approach, where audio recognition is trained on a massive dataset sourced from diverse environments, speaker variations, and multiple languages [6]. The model employs an encoder-decoder architecture with parameter sizes ranging from a small model with 39 million parameters to a large model with up to 1.55 billion parameters. One of Whisper's main advantages is its multilingual capability, supporting more than 98 languages. OpenAI specifically designed Whisper's data format to ensure the model can handle various tasks flexibly and robustly under different conditions. In the Whisper system, the transformer architecture used in both the encoder and decoder enables efficient and accurate transcription of text from audio [10].

This study uses the Whisper small variant, primarily considering its faster processing speed and better transcription accuracy compared to lower variants, even though it is less accurate than the larger models. The small model is considered sufficiently efficient, offering a balance between accuracy and processing time that remains within reasonable limits [5]. It is also freely available and open source, making it easily accessible and usable without additional costs. Transcription results are evaluated using the Word Error Rate (WER) metric to measure the quality and accuracy of the generated text [11].

2.4. LexRank Summarization Algorithm

LexRank is an extractive automatic text summarization method that uses a graph-based approach to determine the importance of each sentence in a document. In this algorithm, each sentence is represented as a node, and the connections between sentences are calculated based on cosine similarity. The concept of centrality is then used to identify the highest-weighted sentences considered most representative of the overall document content. One of LexRank's advantages is its unsupervised nature, which does not require training data and is flexible for use in multiple languages, including Indonesian. Previous research has shown that LexRank can produce concise and relevant summaries while preserving the original sentence structure [7].

In this study, the LexRank algorithm is applied to summarize the transcribed text of YouTube videos processed using the Whisper model. LexRank was chosen for its simplicity, processing speed, and ability to extract important sentences from long transcription texts. Its implementation enables the system to generate concise summaries that still effectively capture the core information of the videos. This aligns with the primary objective of the study, which is to develop a web-based system for automating the transcription and summarization of YouTube video content.

2.5. Word Error Rate (WER) Evaluation

Word Error Rate (WER) testing is used to evaluate the transcription error level by comparing the system-generated transcript with the reference text. WER is calculated as the ratio of the number of edit errors (substitutions, insertions, and deletions) to the total number of words in the reference. A lower WER value indicates better and more accurate transcription quality [11]. The formula for word error rate (WER) is defined as shown in Equation (1).

$$WER = \frac{S+D+I}{N} \tag{1}$$

Where S is the number of substituted words, D is the number of deleted words, I is the number of inserted words, and **N** is the total number of words in the reference. Word Error Rate (WER) is used in this study to measure the accuracy of the transcription output produced by the Whisper model compared to the reference transcript from YouTube's official subtitles. The WER value provides an indication of how well the transcription system can reproduce text that matches the original spoken content in the video. By comparing the automatic transcription results with the reference text, WER helps evaluate the model's performance in accurately recognizing and writing spoken language. A lower WER value indicates better transcription performance with fewer errors [11].

2.6. ROUGE Evaluation Metric (Recall-Oriented Understudy for Gisting Evaluation)

ROUGE (Recall-Oriented Understudy for Gisting Evaluation) is an evaluation method consisting of several indicators for measuring the quality of automatically generated text. This approach is widely used for assessing text summarization algorithms and has become a recognized evaluation standard in numerous studies [12]. In this research, two ROUGE variants are used: ROUGE-N, which measures recall based on n-grams, and ROUGE-L, which assesses similarity based on the longest common subsequence found between the generated summary and the reference text [13].

ROUGE-N is a metric used to measure recall by comparing the occurrence of n-grams between the reference summary and the text produced by the automatic summarization system. Typically, the value of n ranges from 1 to 4, but ROUGE-1 and ROUGE-2 are the most commonly used in evaluations. In the calculation of ROUGE-N, p denotes the number of matching n-grams between the reference and the machine-generated summary, while q represents the total number of n-grams in the reference summary [14]. The formula for ROUGE-N is defined as shown in Equation (2).

$$ROUGE - N = \frac{p}{q} \tag{2}$$

ROUGE-L evaluates the quality of text summaries by comparing the longest common subsequence (LCS), which is the longest sequence of words appearing in both the machine-generated summary and the reference summary. In the calculation, m represents the number of words in the reference summary [13]. The formula for ROUGE-L is shown in Equation (3) below.

$$ROUGE - L = \frac{LCS}{m} \tag{3}$$

In this study, the ROUGE scores are measured from the text summaries generated by the LexRank algorithm, which are derived from transcriptions produced using the Whisper Small model. These ROUGE scores are used to assess how well the combination of the Whisper model and the LexRank algorithm can produce relevant and accurate summaries.

2.7. Data Collection

Data collection in this study is carried out by recording all process results into a file named logs.csv. This file stores important information such as the URL of the processed video, video duration, the number of words in the transcription, summary length, processing time at each stage (download, transcription, and summarization), device specifications (RAM, CPU, and OS), and the final process status indicating success or failure. This data serves as the basis for system performance analysis and evaluation.

2.8. Experimental Environment

The experimental environment in this study uses a standard laptop without GPU support. The device specifications include the Windows 10 operating system, 8GB of RAM, and an Intel Core i5 processor. Testing is conducted under these conditions to evaluate the system's performance on hardware with moderate computing capabilities.

3. Results and Discussion

3.1. Audio Extraction Using the yt_dlp Library

In this first stage, data in the form of audio is collected from videos randomly selected from the YouTube platform. For the experiment, five test videos with different durations were used to ensure that the system was tested across various processing times and content complexities. Audio from each video was downloaded using the yt_dlp library, which allows direct audio extraction from YouTube sources. Using videos of varying durations aims to comprehensively evaluate the system's performance in terms of processing speed, transcription accuracy, and the quality of the generated summaries for inputs of different lengths. The details of the test results are presented in Table 1 below.

Table 1: Audio Extraction Using the yt_dlp Library

Content ID	Content Title	Video Duration	Download Time	Internet Speed
v=zcr3A8Z_MaA	When you wake up in the morning (motivational video) | spoken word | Merry Riana	02:37 Minutes	7.86 Seconds	3.05 Mbps
v=uqGf4PWDOUw	Tips for Discipline in Building Maudy Ayunda's Booklist	08:22 Minutes	16.02 Seconds	1.5 Mbps
v=H-DeO-hnyTc	How to Improve IQ and Brain Intelligence	16:19 Minutes	22.73 Seconds	1.06 Mbps
v=DbkjezHz3ys	Change Your Life in 2025 in Under 25 Minutes	28:28 Minutes	54.74 Seconds	0.44 Mbps

Based on the testing results for four educational YouTube videos of varying durations, it can be concluded that the audio extraction process using the yt_dlp library demonstrates high efficiency. Although video duration influences download time, the main factor determining download speed is the stability and speed of the internet connection during the process. This is evident in longer videos that required more time to download due to slower internet speeds. Overall, yt_dlp was able to complete the audio extraction process in under one minute per video, making it a practical and reliable solution for an automated YouTube video transcription system.

3.2. Text Transcription Using the Whisper Model

The results of this test demonstrate the performance of the Whisper Small variant in automatically transcribing four YouTube videos of varying durations. Transcription was performed without any manual intervention, and the evaluation focused on transcription processing time, the number of words produced, and efficiency in terms of transcription time per word (in seconds). The collected data provides insights into the model's performance and consistency in handling videos of different lengths automatically. The details of the test results are presented in Table 2 below.

Table 2: Whisper Small Results

Content ID	Video Duration	Whisper Model	Word Count	Transcription Time	Time Per Word
v=zcr3A8Z_MaA	02:37 Minutes	Small	234 Words	37.02 Seconds	0.158 Seconds
v=uqGf4PWDOUw	08:22 Minutes	Small	1165 Words	176.06 Seconds	0.151 Seconds
v=H-DeO-hnyTc	16:19 Minutes	Small	2354 Words	390.43 Seconds	0.166 Seconds
v=DbkjezHz3ys	28:28 Minutes	Small	4134 Words	775.64 Seconds	0.188 Seconds

Transcription testing on four YouTube videos with varying durations shows that the Whisper Small model is capable of performing transcription with fairly good time efficiency. The average time required to transcribe one word ranged from approximately 0.151 to 0.188 seconds per word. Although the video duration increased, the transcription time per word remained relatively stable, indicating that the model's performance is consistent across different word counts and video lengths.

Whisper Small also proved to be efficient and practical for use on devices without a GPU, making it a suitable choice for web-based or standard desktop systems. With a balanced level of accuracy and processing time still within a reasonable range, this model is well-suited for use in automatic transcription applications related to education, documentation, or video content analysis.

3.3. Summarization Using the LexRank Algorithm

The system developed in this study applies the LexRank algorithm to automatically summarize the transcription results of YouTube videos. The summarization process begins by parsing the text using an English parser and tokenizer to split the document into individual sentences. After the segmentation is completed, the system calculates the total number of sentences in the document to determine the length of the generated summary.

The summary length is determined adaptively based on the number of sentences contained in the original text. This strategy aims to maintain proportionality and content relevance without sacrificing the core information. If the text contains only a few sentences, the system will retain most of them in the summary. For longer documents, the system limits the maximum number of sentences to ensure the summary remains concise and efficient. Table 3 below summarizes the adaptation scheme for summary length based on the total number of sentences in the source text.

Table 3: Summary Length Adaptation Scheme

Original Number of Sentences	Number of Sentences in Summary
≤ 5 sentences	All Sentences
6 – 15 sentences	4 sentences
16 – 30 sentences	6 sentences
31 – 50 sentences	8 sentences
51 – 100 sentences	10 sentences
> 100 sentences	Minimum 15 or 12.5% of Total Sentences

After the number of sentences is determined, LexRank selects the sentences with the highest centrality scores within the graph structure formed from inter-sentence similarity. These sentences are then reordered into a coherent summary that represents the main content of the document. In addition to generating the summary, the system also records processing time to measure the efficiency of the LexRank algorithm on documents of varying lengths.

To evaluate the effectiveness and efficiency of the LexRank algorithm in automatically summarizing the transcribed text of YouTube videos, tests were conducted on four videos with varying durations and word counts. The analyzed parameters include the length of the original text, the length of the generated summary, summarization processing time, and the total time from download to evaluation. Table 4 below presents the complete results of the summarization process using the LexRank algorithm.

Table 4: LexRank Algorithm Summarization Results

Content ID	Video Duration	Word Count	Summary Length	Transcription Time	LexRank Time	Total Time
v=zcr3A8Z_MaA	02:37 Minutes	234 Words	74 Words	37.02 Seconds	0.3 Seconds	48.55 Seconds
v=uqGf4PWDOUw	08:22 Minutes	1165 Words	182 Words	176.06 Seconds	0.17 Seconds	196.14 Seconds
v=H-DeO-hnyTc	16:19 Minutes	2354 Words	262 Words	390.43 Seconds	0.52 Seconds	418.27 Seconds
v=DbkjezHz3ys	28:28 Minutes	4134 Words	1450 Words	775.64 Seconds	0.15 Seconds	836.08 Seconds

Based on the data in Table 4: LexRank Algorithm Summarization Results, it can be concluded that the LexRank algorithm demonstrates fast and efficient performance in summarizing text derived from video transcriptions with varying durations and word counts. The LexRank processing time to generate a summary is relatively very short (averaging less than 1 second), even for texts exceeding 4,000 words. This indicates that LexRank is able to maintain time efficiency even as input complexity increases.

In terms of results, the length of the generated summaries correlates with the length of the original text, where longer videos with more words produce longer summaries. This shows that the system automatically adjusts the summary length based on content complexity.

Overall, the combination of fast processing time and proportionally consistent summary quality makes LexRank an effective extractive method for presenting the core information of long-duration videos in a concise and informative manner.

Table 5 below presents the transcription results generated by the Whisper Small model from several YouTube videos used as research subjects. It also shows the summaries obtained using the LexRank algorithm. This presentation aims to provide a comparative overview of the system's outputs at the transcription and automatic summarization stages, which then serve as the basis for evaluating quality and accuracy in this study.

Table 5: Whisper Transcription Results and LexRank Algorithm Summaries

Transcribed Text ***(Whisper Small Model)***	Hi curious people, Maudia Yunda here. Welcome to my book list again, where I share interesting grids of books I read. Of all the content, I will only share 3 insights that are personally interesting to me. Today we are going to distinguish the book Atomic Habits by James Clear. Let's talk about the author. James Clear has been resetting habits and decision making for years. Initially, James was known through a newsletter that waited to become 100,000 subscribers in just 2 years. To this day, his work frequently appears in The New York Times, Forbes, and Business Insider. The first insight, forget the goals and try to build a system. This insight emphasizes that goals and systems are two different things. Goals or objectives are the results to be achieved. But the system is the process that leads to those outcomes. James Clear reminds us to focus on building systems that can get us closer to our desired goals...
Summary Text ***(LexRank Algorithm)***	James Clear reminds us to focus on building systems that can bring us closer to our desired goals. Too many people focus on the wrong thing. So we have to focus on the system and the exhaustion that we do every day. First, there are 4 surefire ways to be able to build habits. If you keep the guitar in the closet, you will definitely never practice. But if you put the guitar in a place that is easy to see, like in the middle of the room, for example, you will definitely be reminded to practice more often. So for example, you want to have a running habit on a treadmill. Now there are tips and tricks to make difficult habits more habitable. You will definitely be a mugger for other TV and you can use that time to do other more productive things. Another way to make habits more potent is to change your mindset. We have to make it more satisfying to be able to build long-term habits.

The transcription of an 8-minute 22-second video using the Whisper Small model produced text containing 1,165 words, while the summary generated by the LexRank algorithm was 182 words long. The transcription process took approximately 176 seconds, whereas summarization was much faster at only 0.17 seconds, resulting in a total processing time of around 196 seconds. The summary produced by LexRank successfully captured the main ideas in the transcription text, emphasizing the importance of building systems to achieve goals and practical strategies for forming habits. This demonstrates that the combination of Whisper Small and LexRank can efficiently generate shorter yet still informative summaries, supporting the research objective of making long video content easier to understand.

3.4. Word Error Rate (WER) Evaluation

This study employed a programming-based approach to evaluate the transcription accuracy of the Whisper model on a set of YouTube videos. The evaluation was performed by comparing the transcription results to the official subtitles available on the platform. The process began with extracting the ID from each video link, followed by retrieving the reference subtitles using the youtube_transcript_api library, which allows direct access to YouTube subtitle text via API.

The transcription error rate was measured using the Word Error Rate (WER) metric, which accounts for errors in the form of substitutions, deletions, and insertions relative to the total number of words in the reference. The evaluation results for each video were documented in an Excel file for structured record-keeping. Table 6 below summarizes the comparison scheme between the automatic transcription results and the reference subtitles, along with the WER values for each tested video.

Table 6: Word Error Rate (WER) Evaluation

Content ID	Video Duration	Whisper Model	Word Count	World Error Rate (WER)
v=zcr3A8Z_MaA	02:37 Minutes	Small	234 Words	0.3843
v=uqGf4PWDOUw	08:22 Minutes	Small	1165 Words	0.4094
v=H-DeO-hnyTc	16:19 Minutes	Small	2354 Words	0.3181
v=DbkjezHz3ys	28:28 Minutes	Small	4134 Words	0.3697

Based on the data in Table 6, it can be concluded that the Whisper Small variant is capable of producing automatic transcriptions with reasonably good accuracy, even when run on standard CPU hardware without post-processing refinements. The Word Error Rate (WER) values for the four videos range from 0.31 to 0.41, which is still within an acceptable range for an open-source automatic transcription model. It is notable that even as word counts increase with video duration, the WER does not show significant spikes, indicating the model's consistent performance on longer inputs.

3.5. Evaluation Using the Recall-Oriented Understudy for Gisting (ROUGE) Metric

The quality of the summaries generated by the system was evaluated using the ROUGE metric (Recall-Oriented Understudy for Gisting Evaluation), which is the standard for automatic summarization tasks [12]. ROUGE measures the overlap of n-grams, phrases, and word sequences between the system-generated summary and a manually written reference summary. This study used three ROUGE metric variants: ROUGE-1 (unigram overlap), ROUGE-2 (bigram overlap), and ROUGE-L (Longest Common Subsequence). Each metric was evaluated in terms of recall, precision, and F1-Score to provide a comprehensive picture of system performance. The ROUGE measurement results for five sample video contents are shown in Table 7 below.

Table 7: Evaluasi Ringkasan

Content ID	Score	ROUGE-1	ROUGE-2	ROUGE-L
v=zcr3A8Z_MaA	*Recall*	0.656	0.258	0.593
	Precision	0.283	0.109	0.256
	F1-Score	0.396	0.153	0.358
v=uqGf4PWDOUw	*Recall*	0.528	0.096	0.433
	Precision	0.153	0.027	0.126
	F1-Score	0.238	0.042	0.195
v=H-DeO-hnyTc	*Recall*	0.623	0.248	0.525
	Precision	0.454	0.181	0.383
	F1-Score	0.526	0.209	0.443
v=DbkjezHz3ys	*Recall*	0.643	0.099	0.273
	Precision	0.096	0.014	0.040
	F1-Score	0.167	0.025	0.071

Based on the evaluation results shown in Table 7, in general, the recall values for the ROUGE-1 metric tend to be higher than the precision values. This indicates that the summarization system has a relatively good ability to capture important information from the reference summaries, although its precision still needs improvement. The lower ROUGE-2 scores suggest that matching at the bigram (consecutive word pair) level is more challenging, indicating difficulties in maintaining local cohesion within the summaries. Meanwhile, ROUGE-L, which accounts for word sequence order, shows intermediate results with considerable variation across the different content samples. These findings suggest that while the system is capable of generating informative summaries, there remains room for further development to improve syntactic and semantic alignment with the reference summaries.

4. Conclusion

This study presents an innovative automated system for transcribing and summarizing YouTube video content using only a video link as input—eliminating the need for complex manual steps. The main novelty of the system lies in its ability to perform the entire process automatically on a standard CPU device, enabling broader and more inclusive adoption.

Testing results demonstrate competitive performance: the Whisper Small model achieved Word Error Rates (WER) ranging from 31% to 41%, which is reasonably good for long and varied video content. The LexRank algorithm successfully summarized transcripts exceeding 4,000 words in under one second, producing concise yet informative summaries. Evaluation using ROUGE metrics yielded average F1-Scores of 33% for ROUGE-1, 10% for ROUGE-2, and 19% for ROUGE-L—indicating that the automatic summaries closely approximate human-written references. Another advantage of the system is its automatic logging of transcription results, summaries, and evaluation metrics in CSV format, supporting structured digital documentation.

This system offers a practical, fast, and accessible approach to transforming lengthy video content into more structured information, with strong potential for implementation in education, journalism, research, and digital archiving. Nonetheless, there remains room for future improvement—such as enhancing summary coherence, reducing reliance on official subtitles for evaluation, expanding multilingual content support, and integrating generative summarization models for more natural and contextually rich outputs..

References

[1] M. A. M. Ardiansyah and M. L. Nugraha, "Analisis pemanfaatan media pembelajaran YouTube dalam meningkatkan pemahaman konsep matematika peserta didik," in *Proc. Seminar Nasional Riset dan Inovasi Teknologi (SEMNAS RISTEK)*, vol. 6, no. 1, Jan. 2022. doi: 10.30998/semnasristek.v6i1.5828.

[2] D. Ramadhina and I. Rohman, "Problematika guru dalam penggunaan video YouTube sebagai media pembelajaran di sekolah dasar," *Mimbar Ilmu*, vol. 27, no. 1, pp. 117–123, 2022. doi: 10.23887/mi.v27i1.45598.

[3] D. Wong, "Effectiveness of learning through video clips and video learning improvements between business related postgraduate and undergraduate students," Int. J. Mod. Educ., vol. 2, no. 7, pp. 119–127, 2020, doi: 10.35631/IJMOE.27009.

[4] H. B. U. Haq, M. Asif, and M. Bin Ahmad, "Video summarization techniques: a review," *Int. J. Sci. Technol. Res*., vol. 9, no. 11, pp. 146–153, 2020.

[5] M. Fadlilah, A. Atmadja, and M. Firdaus, "Pemanfaatan Transformer untuk peringkasan teks: Studi kasus pada transkripsi video pembelajaran," Building of Informatics, Technology and Science (BITS), vol. 6, no. 3, pp. 2111–2119, 2024, doi: 10.47065/bits.v6i3.6342.

[6] R. F. Khoiroh, E. Julianto, S. A. Ardiyansa, H. A. Fajri, A. A. R. Yasa, and B. Sangapta, "Implementasi speech recognition Whisper pada debat calon wakil presiden Republik Indonesia," Explore, vol. 14, no. 2, pp. 67–74, 2024, doi: 10.35200/ex.v14i2.115.

[7] Wiratmoko, G. (2025). Evaluating the Effectiveness of the LexRank and LSA Algorithm in Automatic Text Summarization for Indonesian Language. Eduvest - Journal of Universal Studies, Vol. 5 No. 2, 3407–3415. DOI: 10.59188/eduvest.v5i2.1663

[8] Mustansiriyah University. (2021). An Approach for Multi-Document Text Summarization Using Extreme Learning Machine and LexRank. International Journal of Engineering Research and Advanced Technology, 7(5), 19–28. DOI: 10.31695/IJERAT.2021.3704.

[9] A. Rawat, V. Rawat, N. Singh, N. Kuchhal, J. Barmola, and H. S. Negi, "An enhance version of YouTube video downloader using Python," in Proc. 2023 International Conference on Computer Science and Emerging Technologies (CSET), Oct. 2023, pp. 1–6. doi: 10.1109/CSET58993.2023.10346693.

[10] D. Ferdiansyah and C. S. K. Aditya, "Implementasi automatic speech recognition bacaan Al-Qur'an menggunakan metode Wav2Vec 2.0 dan OpenAI-Whisper," in Proc. Jurnal Teknik Elektro dan Komputer TRIAC, vol. 11, no. 1, pp. 11–16, 2024. doi: 10.21107/triac.v11i1.24332.

[11] Z. Dyarbirru and S. Hidayat, "Metode Wavelet-MFCC dan Korelasi dalam Pengenalan Suara Digit", *jtim*, vol. 2, no. 2, pp. 100–108, Aug. 2020. DOI: 10.35746/jtim.v2i2.99

[12] M. Barbella and G. Tortora, "Rouge metric evaluation for text summarization techniques," in *Proc. SSRN Electronic Journal*, 2022. doi: 10.2139/ssrn.4120317.

[13] G. Hartawan, D. S. Maylawati, and W. Uriawan, "Bidirectional and Auto-Regressive Transformer (BART) for Indonesian Abstractive Text Summarization," in Jurnal Informatika Polinema, vol. 10, no. 4, pp. 535–542, 2024. doi: 10.33795/jip.v10i4.5242

[14] Y. Yuliska and K. U. Syaliman, "Literatur review terhadap metode, aplikasi dan dataset peringkasan dokumen teks otomatis untuk teks berbahasa Indonesia," in *Proc. IT J. Res. Dev.,* vol. 5, no. 1, pp. 19–31, 2020. doi: 10.25299/itjrd.2020.vol5(1).4688.