



Hybrid Video Transcription Summarization with a BERT-Based Clustering and BART

Fathul Agit Darmawan¹, Muhammad Bima Mauludin², Christian Sri Kusuma Aditya^{3*}

^{1,2,3}Department of Informatics, Faculty of Engineering, University of Muhammadiyah Malang, Indonesia

¹fathulagitd169@webmail.umm.ac.id, ²bimamauludin14@webmail.umm.ac.id, ³christianskadiyaa@umm.ac.id

Abstract

The use of video as a medium for information and education is rapidly increasing across online platforms. However, long durations and unstructured delivery often hinder audiences from grasping the core message, presenting challenges for the development of automatic summarization methods for monologues, interviews, and podcasts. Extractive methods often yield less coherent summaries, while abstractive methods may overlook important details. To address this issue, this study proposes a hybrid approach combining extractive and abstractive techniques. In the extractive stage, sentences are represented using BERT embeddings and clustered using two methods, namely K-Means Clustering and Hierarchical Clustering (agglomerative). The abstractive stage then employs the BART model to generate summaries that are more coherent and informative. Experimental evaluations on 20 Human Metapneumovirus (HMPV) videos indicate the strongest performance on monologues, with ROUGE-1 of 57%, ROUGE-2 of 30%, and ROUGE-L of 32%. Although lower performance was observed for interviews and podcasts due to dynamic interactions and frequent speaker shifts, the hybrid approach consistently surpassed extractive-only and abstractive-only baselines. These results highlight the effectiveness of the hybrid approach and its potential for developing more adaptive video summarization in the future.

Keywords: BART; BERT; clustering; hybrid approach; video summarization

How to Cite: F. A. Darmawan, M. B. Mauludin, and C. S. K. Aditya, "Hybrid Video Transcription Summarization with a BERT-Based Clustering and BART", *J. RESTI (Rekayasa Sist. Teknol. Inf.)*, vol. 9, no. 6, pp. 1439 - 1450, Jan. 2026.

Permalink/DOI: <https://doi.org/10.29207/resti.v9i6.7066>

Received: August 10, 2025

Accepted: December 25, 2025

Available Online: January 11, 2025

This is an open-access article under the CC BY 4.0 License

Published by Ikatan Ahli Informatika Indonesia

1. Introduction

The use of video as a medium for delivering public information continues to increase, especially through online platforms such as YouTube [1]. Behind this easy access, videos have limitations in terms of the speed of information consumption due to their long duration and content structure that is not always systematic [2]. This condition can make it difficult for people to quickly absorb important information, particularly in the context of educational videos, policy speeches, or lengthy public discussions. As a result, the potential for misinformation, misconceptions, or even missing crucial information becomes higher, especially for audiences with limited time or access. Efforts to extract the core information from videos in text form face particular challenges, especially when the speech-to-text transcriptions are long, repetitive, and unorganized. Technologies like Whisper have enabled automatic conversion of speech to text with high accuracy [3], [4],

but the resulting text still requires further processing to be utilized in a concise and informative manner.

Text summarization has become a relevant approach to simplify the content of video transcriptions. In natural language processing, two main summarization approaches are known: extractive and abstractive. The extractive approach focuses on selecting important sentences from the original text, while the abstractive approach rearranges information into a new, more concise form. The BERT model has been widely used in extractive summarization due to its ability to build context-based sentence representations [5], [6], whereas the BART model demonstrates superior performance in abstractive summarization through an encoder-decoder architecture specifically designed to understand and generate new text [7], [8]. However, the integrated application of both methods to handle video transcriptions remains very limited in the literature.

This becomes increasingly important because video transcriptions have unique characteristics that

differentiate them from written documents. The conversational structure, pauses, repetitions, and rapid, complex topic shifts make video transcripts more difficult to summarize effectively. Moreover, since transcriptions stem from oral communication that is dynamic and spontaneous, the text tends to lack narrative coherence [9]. Previous research has largely focused on transcribing written documents or summarizing structured oral communication, leaving a gap in addressing the specific challenges posed by the unpredictable nature of video transcriptions [10]. Additionally, previous studies also suggest the use of hybrid methods for text summarization, which combine both extractive and abstractive techniques, as a promising solution for more effective summaries [11]. Therefore, a summarization approach capable of handling the linguistic diversity in video transcripts and producing summaries that remain relevant and easily understood is required.

To address these challenges, this study proposes a hybrid approach that leverages the strengths of both extractive and abstractive methods. The extractive method alone tends to produce less natural summaries as it merely selects raw sentences from the original text [12], while the abstractive approach struggles with excessively long and unstructured texts [13]. Therefore, an initial sentence selection is performed using the BERT model, after which the sentences are grouped based on semantic similarity using clustering algorithms. Two clustering methods are employed: *K-Means Clustering* and *Hierarchical Clustering (agglomerative)*. *K-Means* is utilized to efficiently identify representative sentences from various topic groups while reducing redundancy [14]. whereas *Hierarchical Clustering* constructs a dendrogram structure that represents the relationships among sentences, thus enabling the identification of representative sentences within a hierarchical context

[15]. In the final stage, the selected sentences from each cluster are processed using the BART model to generate summaries that are more coherent, narratively flowing, and concise.

This method is applied to a dataset consisting of transcriptions from twenty public YouTube videos, covering various types of real-life communication such as monologues, interviews, and podcasts, thus allowing analysis of the method in a broader and more realistic context. Summary quality evaluation is performed by comparing model outputs with manual summaries using the ROUGE metric. Therefore, this study aims to produce an effective video transcription summarization method that presents information in a concise, coherent, and easily digestible manner. The main contributions lie in the integration of two complementary summarization approaches, the application to the underexplored domain of video transcriptions, and the thorough evaluation of the quality of generated summaries.

2. Methods

This section outlines the research design, data acquisition process, methodology for transcription and summarization, and evaluation of the summary results.

2.1 Research Design

This study analyzes the transcription and summarization of videos on Human Metapneumovirus (HMPV), a topic widely discussed in early 2025 due to rising infections and public health concerns. Selected videos include monologues, interviews, and podcasts covering diverse medical and scientific perspectives. Figure 1 illustrates the overall workflow of data collection, transcription, summarization, and evaluation.

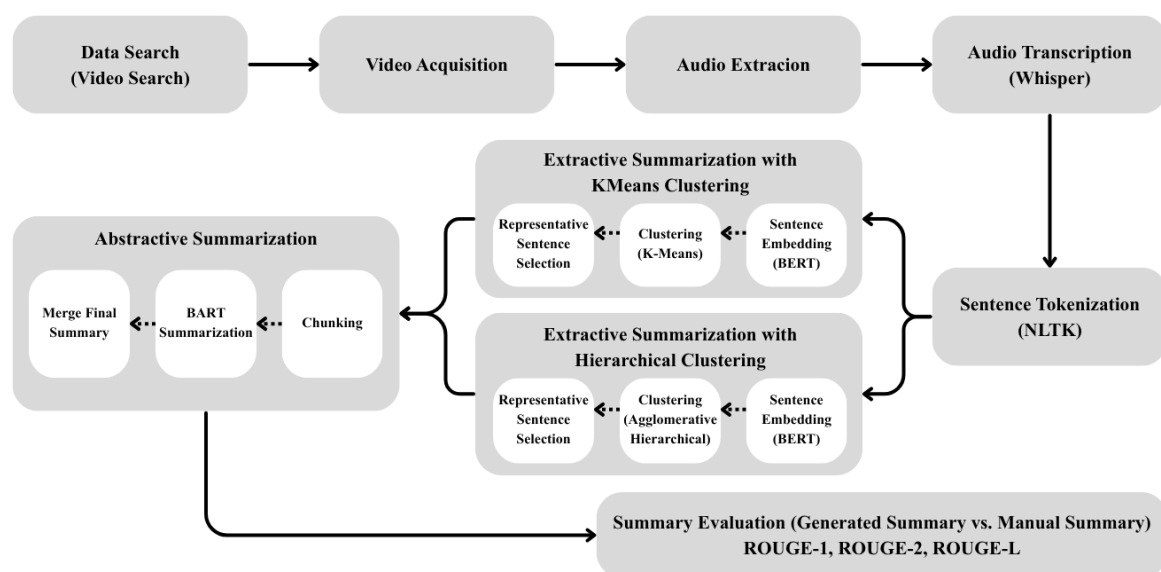


Figure 1. Research flow for automatic summarization of Human Metapneumovirus (HMPV) videos.

2.2 Data Source and Acquisition

The dataset in this study consists of twenty videos manually collected from the YouTube platform, representing four different types of communication:

monologue, interview, podcast, and short interview, with five videos for each type. The videos were directly downloaded using the *yt_dlp* library. Table 1 summarizes the details of the videos used in this study.

Table 1. Details of Videos Used

Video Title	Type of Video	Duration (s)	Channel YouTube
What you need to know about HMPV	Monologue 1	08:25	TEN HEALTH
HMPV Outbreak: The Rapidly Spreading Virus You Haven't Heard Of	Monologue 2	08:32	TEN HEALTH
HMPV Outbreak: What is HMPV Virus, Symptoms, spread, and how to prevent	Monologue 3	08:10	Peak Hive
What is the HMPV virus?	Monologue 4	02:41	The Hindu
HMPV virus in China: What is HMPV? Know Symptoms, Treatment, Prevention	Monologue 5	03:46	midday india
Doctor: Human Metapneumovirus an influenza-like illness, not as severe as COVID or even flu ANC	Interview 1	08:18	ANC 24/7
HMPV Virus India Dr. K. Srinath Reddy Rules Out COVID-Like Threat Exclusive Interview N18V	Interview 2	17:51	CNN-News18
Understanding HMPV: Risks, Precautions, and Insights Exclusive Interview with Dr. MC Misra	Interview 3	14:57	Oneindia News
HMPV Virus Explained: Dr. Ravi Godse Talks Symptoms & Impact Health China Pandemic Vaccine	Interview 4	31:53	The Free Press Journal
HMPV Virus: Doctors Explain Everything You Need To Know 3 Cases Found in India NewsX	Interview 5	22:49	NewsX Live
Human Metapneumovirus (HMPV) #podcast	Podcast 1	20:37	Chronicles Home Podcasts
Understanding HMPV: Unveiling the Flu-Like Threat	Podcast 2	12:03	Eleven Minute Podcast
The Medical Minute Podcast- 2025 Ep1 Mysterious Metapneumovirus, Measles & Depo-Provera Controversy	Podcast 3	21:34	The Medical Minute with Dr Anne & Dr Moz
HEALTH LIVE PODCAST: What is Metapneumovirus and how can we avoid it.	Podcast 4	15:33	SALUD_EN_VIVO
Audio Podcast Human Metapneumovirus HMPV. Is there any chance of another Covid-like pandemic?	Podcast 5	16:16	Medical Microbiology and Infections PODCAST
Human metapneumovirus infections up 36%, CDC says	Short Interview 1	03:23	CBC New
What to know about human metapneumovirus GMA	Short Interview 2	02:27	Good Morning America
Human metapneumovirus	Short Interview 3	03:00	Amesh Adalja
What is HMPV? The signs, symptoms and spread	Short Interview 4	02:34	The Australian
HMPV Virus Precautions One Needs To Take To Avoid HMPV Infection: Ex WHO Chief Scientist Explains	Short Interview 5	02:31	NDTV

Table 1 summarizes the 20 videos used in this study, consisting of five videos for each communication type: monologue, interview, short interview, and podcast. Monologues involve a single speaker delivering information without interaction, while interviews feature question-and-answer exchanges between an interviewer and a guest in a more in-depth format. Short interviews present similar interactions in a more concise manner, and podcasts typically consist of relaxed yet informative discussions that explore topics in greater depth.

2.3 Research Procedure

The research consists of several main stages: video transcription, extractive summarization, abstractive summarization, and summary evaluation.

The initial step converts video content into text using OpenAI's Whisper model, a state-of-the-art NLP system for accurate and efficient speech-to-text transcription [16]. Videos are downloaded from YouTube using the *yt_dlp* library and converted into .wav format with *moviepy.editor.VideoFileClip* for transcription. The audio is transcribed using the Whisper *small* model, and the resulting text is

segmented into sentences with *nlk.sent_tokenize* to support subsequent processing.

The objective of extractive summarization is to identify the most relevant and representative sentences from a transcription. The process starts with sentence tokenization using *nlk.sent_tokenize*, followed by semantic feature extraction with BERT, where each sentence is converted into an embedding using the *bert-base-nli-mean-tokens* model from Sentence Transformers. BERT captures deep contextual semantics by considering bidirectional word context, making it effective for representing sentence meaning [17], [18]. The embedding of a sentence is computed by averaging the embeddings of its tokens, as shown in Equation 1:

$$Emb_{sentence} = \frac{1}{n} \sum_{i=1}^n Emb_{token_i} \quad (1)$$

Each token in a sentence is converted into a numerical vector (embedding), and these vectors are then summed and averaged to produce a single representative vector for the entire sentence. Although this method does not account for word order or contextual dependencies between tokens, it is widely used due to its simplicity

and reasonable effectiveness in capturing the overall semantic representation of a sentence [19], [20].

Once the sentence embeddings are obtained, sentence clustering using K-Means is performed. K-Means is an unsupervised machine learning algorithm that partitions the data into a predefined number of clusters by minimizing the variance within each cluster. It iteratively assigns data points to the nearest cluster centroid and updates the centroids based on the mean of the assigned points [21]. The distance between a sentence and a cluster centroid is computed using the Euclidean distance, as defined in Equation 2:

$$d(x, c) = \sqrt{\sum_{i=1}^n (x_i - c_i)^2} \quad (2)$$

Equation (2) calculates the Euclidean distance between two vectors x and c in an n -dimensional space, measuring how close a data point is to a cluster center. This metric forms the foundation of the K-Means clustering algorithm, which is widely applied in various research areas including customer segmentation, text clustering, and multidimensional data analysis [22], [23]. In the final stage, one sentence closest to the centroid in each cluster is selected as the most representative sentence to be included in the summary.

In addition to K-Means, this study also compares the results with the Hierarchical Clustering method. Unlike K-Means, which requires the number of clusters to be defined in advance, Hierarchical Clustering constructs a hierarchical structure (dendrogram) that illustrates the similarity relationships among sentences [15]. Broadly, there are two main approaches to Hierarchical Clustering: Agglomerative (bottom-up) and Divisive (top-down). In the agglomerative approach, which is the most widely used, each sentence starts as an individual cluster, and pairs of clusters are iteratively merged based on a distance metric until all sentences are grouped into a single cluster. In contrast, the divisive approach begins with all sentences in one large cluster and recursively splits them into smaller clusters [24].

In this study, we employ the agglomerative approach, as it is computationally more practical and widely supported in existing libraries. The distance between clusters can be determined using various linkage strategies such as *single linkage*, *complete linkage*, or *average linkage*. For instance, in *average linkage*, the distance between two clusters C_a and C_b is calculated by averaging the pairwise distances of all sentences between the two clusters, as defined in Equation 3:

$$d(C_a, C_b) = \frac{1}{|C_a| \cdot |C_b|} \sum_{x \in C_a} \sum_{y \in C_b} d(x, y) \quad (3)$$

where $d(x, y)$ denotes the Euclidean distance between two sentence vectors x and y . The strength of Hierarchical Clustering lies in its ability to reveal the natural structure of the data without requiring an assumption of the number of clusters, thereby providing a more flexible view of sentence relationships [25].

In summary, the extractive process enables the selection of the most representative sentences from the transcript based on semantic mapping using BERT and clustering techniques. While K-Means offers a simple and efficient approach to generating summaries, Hierarchical Clustering provides an alternative perspective by emphasizing the hierarchical relationships among sentences.

Abstractive summarization is performed using Facebook's BART model via the transformers library. BART (Bidirectional and Auto-Regressive Transformers) combines the strengths of both encoder-decoder architectures, allowing it to understand input context deeply and generate fluent, human-like summaries by predicting text sequences in an autoregressive manner [26].

To prevent context loss, the extractive summary is split into chunks of up to 500 characters. This size balances staying within BART's token limit while preserving sufficient context. Larger chunks risk truncation, while smaller ones may reduce coherence. Each chunk is summarized using BART with controlled parameters (max_length=80, min_length=20, do_sample=True, temperature=0.4, top_p=0.7). The resulting summaries are then combined into a single coherent summary representing the full transcript.

To evaluate summary quality, generated summaries are compared with manually written references using ROUGE (Recall-Oriented Understudy for Gisting Evaluation). ROUGE is a standard evaluation metric in text summarization that measures n-gram and word sequence overlap to assess the informativeness and relevance of generated summaries [27]. Its importance lies in providing an objective and quantitative benchmark, enabling consistent performance comparison across different summarization models and datasets. Each ROUGE metric consists of three main components, as summarized in Table 2.

Table 2. ROUGE metrics explanation

ROUGE Metric	Compared Unit	Similarity Measure
ROUGE-1	unigram (single word)	Precision, Recall, F1
ROUGE-2	bigram (consecutive word pairs)	Precision, Recall, F1
ROUGE-L	LCS (Longest Common Subsequence)	Precision, Recall, F1

The F1 score in ROUGE is the harmonic mean of Precision and Recall for a specific n-gram, as shown in Equation 4.

$$F1 = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (4)$$

The F1 score represents a balance between precision and recall. If one metric is high while the other is low, the F1 score will be moderate, whereas balanced values result in a higher F1 score. Thus, it provides a more comprehensive measure of summary quality by considering both accuracy and coverage.

Equation (4) defines the F1-score as a performance metric that balances precision and recall, particularly useful for imbalanced data. A high F1-score indicates strong overall model performance and is widely used in evaluation tasks. [28].

3. Results and Discussions

This section presents the results of the transcription and summarization process of videos on the topic of Human Metapneumovirus (HMPV), along with a comprehensive discussion of the findings. The results are presented in tables and visualizations to facilitate understanding. Evaluation was conducted using the ROUGE metric to compare the automatic summaries with manual summaries.

3.1 Transcription Results

After the transcription process using the Whisper model was carried out, transcription results were obtained for each video. Table 3. shows the video duration, the number of words in the transcription, and the number of sentences.

Table 3. Video Transcription Results

Type of Video	Duration (s)	Word Count	Sentence Count
Monologue 1	08:25	1163	89
Monologue 2	08:32	1141	73
Monologue 3	08:10	1317	81
Monologue 4	02:41	353	23
Monologue 5	03:46	480	26
Interview 1	08:18	1088	76
Interview 2	17:51	2504	147
Interview 3	14:57	2027	136
Interview 4	31:53	4832	425
Interview 5	22:49	3425	179
Podcast 1	20:37	3755	383
Podcast 2	12:03	2293	186
Podcast 3	21:34	3826	256
Podcast 4	15:33	2834	279
Podcast 5	16:16	2903	238
Short Interview 1	03:23	593	42
Short Interview 2	02:27	453	47
Short Interview 3	03:00	624	45
Short Interview 4	02:34	302	16
Short Interview 5	02:31	424	20

Table 3 summarizes the transcription results of 20 HMPV-related videos in terms of duration, word count, and sentence count. The videos span different formats—monologue, interview, podcast, and short interview—resulting in clear variations in length and content density. Word counts range from 302 to 4,832 and sentence counts from 16 to 425, with longer formats such as podcasts and interviews providing more detailed discussions, particularly *Interview 4*. In contrast, monologues and short interviews are more concise and focus on key points, reflecting distinct information delivery styles across video types.

3.2 Extractive Summary Results with K-Means

The extractive summary was obtained by selecting the most representative sentences based on clustering results using BERT embeddings and the K-Means

algorithm. To determine an appropriate summary length, a proportional rule was applied based on the total number of sentences in the original transcript. Table 4 outlines this rule, which generally corresponds to around 20–40% of the full content, depending on the transcript's length.

Table 4. Summary Length Guidelines Based on Total Sentence Count

Sentence Range	Count	Summary Sentence Count	Estimated Percentage (%)
$21 \leq N \leq 50$		10 sentences	$\pm 20\%$
$51 \leq N \leq 100$		25 sentences	$\pm 25\%$
$101 \leq N \leq 200$		50 sentences	$\pm 25\text{--}30\%$
$201 \leq N \leq 300$		80 sentences	$\pm 26\text{--}40\%$
$301 \leq N \leq 400$		100 sentences	$\pm 25\text{--}33\%$
$N > 400$		150 sentences	$\pm 35\%$

The number of sentences in each summary (Table 4) is determined by transcript length and complexity. Short transcripts (≤ 50 sentences) are summarized into 10 sentences, as they usually cover limited topics. Medium-length transcripts (51–100 sentences) use 25 sentences to maintain coverage, while more complex transcripts (101–200 sentences) require around 50 sentences for balance. Longer transcripts (over 200 sentences) use 80–150 sentences to represent topic diversity, making this approach effective for long-format content such as interviews and podcasts.

These sentence counts were determined through multiple experiments using different summary lengths. The ranges presented in Table 4 produced relatively better Silhouette scores than other configurations. Although the Elbow Method was tested, it often suggested too few clusters, resulting in summaries that were overly short and less informative, likely because most transcripts focus on a single main topic (HMPV), causing embedding similarity.

The number of sentences in the original transcript not only determines summary length but also guides the number of clusters used in the K-Means algorithm. By adjusting the number of clusters proportionally to document length, each cluster can represent a distinct information segment, from which representative sentences are selected to form the final summary.

To further illustrate this process, K-Means visualizations were generated for each video type, including monologues, interviews, podcasts, and short interviews. These visualizations show how sentence embeddings are grouped, providing insights into content distribution and communication styles across video categories.

The K-Means visualization of the monologue video, as shown in Figure 2, presents an example of a dispersed cluster distribution without a consistent dominant pattern. This visualization illustrates varying centroid arrangements, reflecting the complexity of the content and topics within the monologue. In general, monologues with more dynamic idea variation tend to produce denser clusters, whereas monologues with

simpler or more focused narratives exhibit fewer and more widely separated clusters.

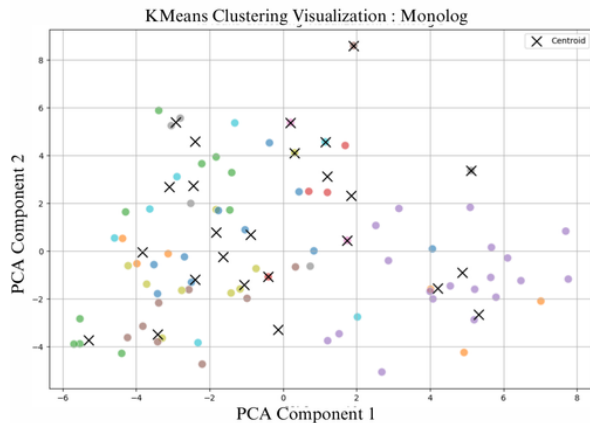


Figure 2. Sample K-Means Clustering Visualization for Monolog

The K-Means clustering of the five interview videos, as shown in Figure 3, reveals distinct distribution patterns for each session, reflecting differences in content and interaction styles. Interviews 1 and 4 display wider and more separated clusters, indicating greater topic diversity, whereas Interviews 2 and 3 exhibit denser and overlapping clusters, suggesting more homogeneous dialogue. Interview 5 presents a balanced and well-structured cluster distribution.

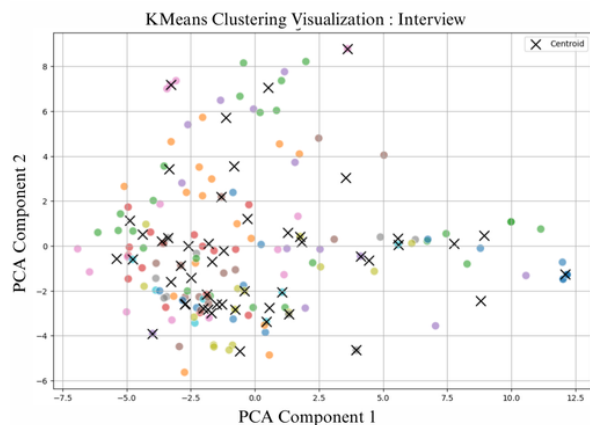


Figure 3. Sample K-Means Clustering Visualization for Interview

The K-Means visualizations of the five podcast videos, as shown in Figure 4, exhibit more evenly distributed and loosely formed clusters compared to interview sessions. This pattern reflects broader topic variation and more dynamic theme transitions. The scattered centroids across the PCA space indicate that podcast content tends to be more narrative and multi-thematic, characterized by higher semantic diversity and a more flexible conversational style.

The K-Means clustering of the five short interview videos demonstrates diverse and evenly distributed clusters, indicating balanced variation in topics and communication styles. As illustrated in Figure 5, many data points are located close to their respective centroids, reflecting strong thematic cohesion, while several outliers suggest the presence of uniquely

characterized segments. Overall, short interview content appears varied yet well-structured, allowing for effective thematic grouping.

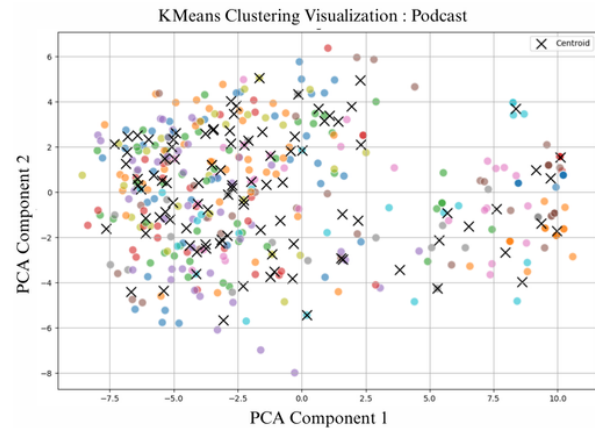


Figure 4. Sample K-Means Clustering Visualization for Podcast

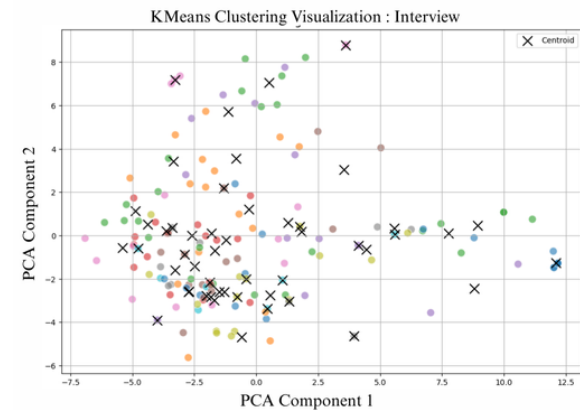


Figure 5. Sample K-Means Clustering Visualization for Short Interview

3.3 Extractive Summary Results with Hierarchical Clustering

This study also uses Hierarchical Clustering with an agglomerative approach and ward linkage. BERT sentence embeddings are grouped based on semantic similarity, with the number of clusters set to about 25% of the total sentences to balance coverage and conciseness. Cluster quality is evaluated using the Silhouette Score, and one representative sentence per cluster is selected based on proximity to the centroid. The final summary is limited to 40 sentences, with PCA visualization used to illustrate semantic distribution, resulting in concise and coherent summaries that capture topic diversity.

The hierarchical clustering visualization for monologue content, as shown in Figure 6, exhibits a more evenly distributed structure, although distinct group separations remain visible. This indicates that, despite monologues typically involving a single speaker and a central theme, there are still notable variations in idea delivery. Small and closely positioned clusters represent subtle differences in emphasis or rhetorical

style, while remaining within a more centralized narrative framework compared to other content types.

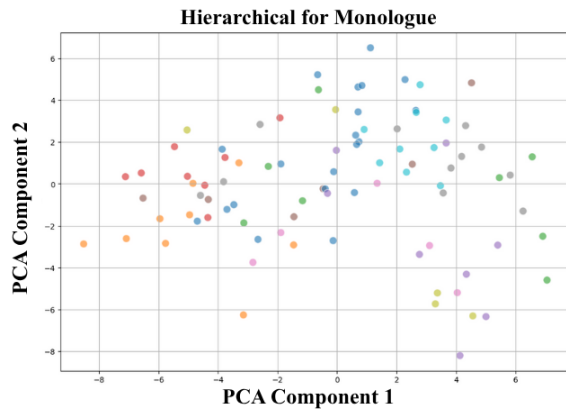


Figure 6. Sample Hierarchical Clustering Visualization for Monolog

The PCA visualization of interview data using Hierarchical Clustering shows scattered clusters, indicating that interviews cover multiple distinct topics with limited overlap. As illustrated in Figure 7, the separated clusters reflect variations in topics, question–answer dynamics, and dialogue structures, which are consistent with the interactive nature of interviews.

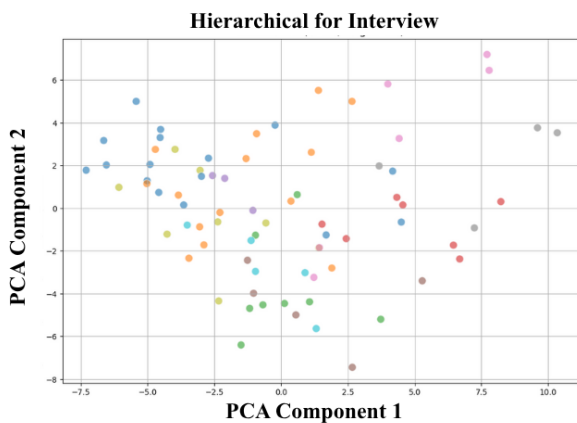


Figure 7. Sample Hierarchical Clustering Visualization for Monolog

The hierarchical clustering results for podcast content reveal dense distributions with many closely spaced small clusters, indicating high idea diversity and conversational complexity. The overlapping clusters reflect multi-topic discussions that are typical of long-form podcasts, which require careful summarization to maintain coherence while capturing topic diversity. This clustering pattern is illustrated in Figure 8.

Figure 9. illustrates the hierarchical clustering visualization of short interview content, which shows sparsely distributed and clearly separated clusters, indicating distinct topical divisions despite the limited number of sentences. The presence of small clusters reflects the concise and focused nature of short interviews, which typically address only one or two main issues.

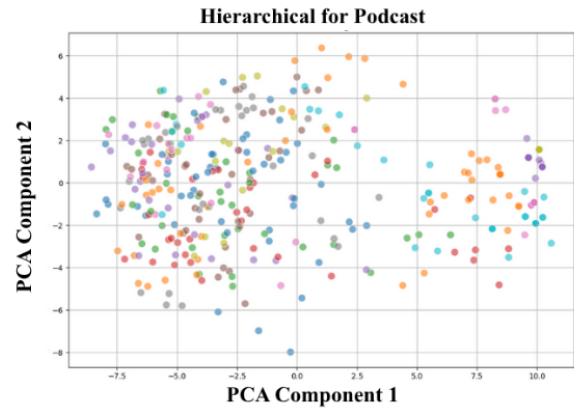


Figure 8. Sample Hierarchical Clustering Visualization for Monolog

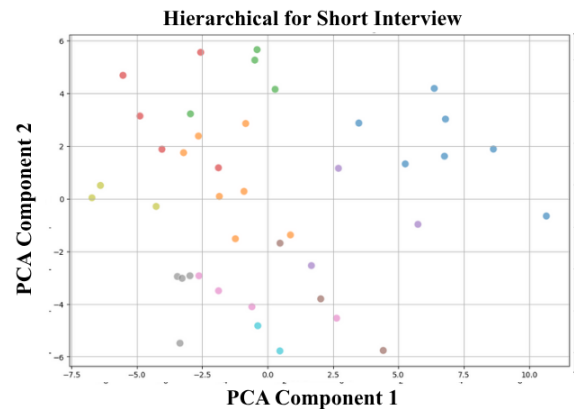


Figure 9. Sample Hierarchical Clustering Visualization for Monolog

3.4 Selection of Representative Sentences via KMeans Clustering and Hierarchical Clustering

In the final extractive summarization stage, one representative sentence was selected from each cluster to capture the core semantics of the transcript. For K-Means, the sentence closest to the cluster centroid was chosen, while for Hierarchical Clustering, selection was based on proximity to the cluster's central point. BERT-based sentence embeddings and Euclidean distance were used to ensure semantic centrality and minimize redundancy. The selected sentences were reordered to preserve coherence, and the evaluation results across transcript types are summarized in Table 5. Table 5 summarizes the clustering evaluation results for KMeans and Hierarchical Clustering across all transcripts, reporting sentence counts, silhouette scores, and the number of representative sentences. Overall silhouette scores range from 0.05 to 0.27, indicating moderate semantic cohesion typical of text embeddings. Podcast 4 achieves the highest scores in both methods, reflecting clearer topic separation. Monologues and short interviews show fewer clusters and lower scores, suggesting more focused content. Hierarchical Clustering generally produces fewer clusters than KMeans while achieving higher silhouette scores, indicating more compact and semantically cohesive groupings. Overall, both methods effectively segment transcripts into coherent topical units for summarization.

Table 5. Number of Clusters and Evaluation Metrics for KMeans and Hierarchical Clustering

Type of Video	Silhouette Score		Repr. Sentences (Cluster)	
	K-Means	Hierarchical :	K-Means	Hierarchical :
Monologue 1	0.0685	0.0929	25	22
Monologue 2	0.0786	0.1052	25	18
Monologue 3	0.0486	0.0763	25	20
Monologue 4	0.0598	0.1143	10	5
Monologue 5	0.1233	0.1185	10	6
Interview 1	0.0566	0.0979	25	19
Interview 2	0.1048	0.1237	50	36
Interview 3	0.1086	0.1489	50	34
Interview 4	0.0985	0.1183	150	106
Interview 5	0.0894	0.1147	50	44
Podcast 1	0.1515	0.1713	100	95
Podcast 2	0.0651	0.0977	50	46
Podcast 3	0.1089	0.1297	80	64
Podcast 4	0.2466	0.2744	80	69
Podcast 5	0.1057	0.1307	80	59
Short Interview 1	0.0844	0.1220	10	10
Short Interview 2	0.0617	0.0902	10	11
Short Interview 3	0.0827	0.0869	10	11
Short Interview 4	0.0788	0.1030	3	4
Short Interview 5	0.0846	0.1197	10	5

3.5 Abstractive Summary Results

After the extractive summarization process, each summary was further condensed using the BART model to produce an abstractive summary. This model generates sentences that are more natural and linguistically coherent [29]. The process consisted of three main stages: chunking, BART summarization, and merging the final summary.

The chunking stage was necessary to divide the extractive summaries into smaller segments that would fit within the input length limit of the BART model (maximum 1024 tokens). Extractive summaries, especially for podcasts and interview videos, often exceeded this limit. Therefore, they were divided into several chunks, each containing a few sentences and limited to a maximum of 500 characters per chunk. This character-based segmentation was used to maintain a safe margin below the token limit, ensuring that each chunk could be processed by the model without truncation. Importantly, at this abstractive stage, the same process was applied to summaries produced by both KMeans and Hierarchical clustering, ensuring consistency in the abstractive summarization results. Table 6 shows the number of chunks generated per video type. The number of chunks mainly depends on the length and complexity of the extractive summaries. Longer content such as interviews and podcasts required more chunks because their summaries often exceeded BART's input limit, while shorter videos needed fewer chunks. For example, Interview 4 produced the most chunks under K-Means (22), reflecting its lengthy and rich content, whereas Short Interview 4 required only one chunk in both methods due to its concise summary. Overall, the Hierarchical method generated fewer chunks than K-Means for longer videos, indicating more compact summaries.

Table 6. Number of Chunks Generated per Video

Type of Video	Number of Chunks	
	K-Means :	Hierarchical :
Monologue 1	4	4
Monologue 2	6	4
Monologue 3	5	4
Monologue 4	2	2
Monologue 5	3	2
Interview 1	4	4
Interview 2	12	10
Interview 3	10	8
Interview 4	22	6
Interview 5	14	12
Podcast 1	15	6
Podcast 2	9	8
Podcast 3	14	8
Podcast 4	12	6
Podcast 5	15	7
Short Interview 1	2	2
Short Interview 2	2	2
Short Interview 3	2	2
Short Interview 4	1	1
Short Interview 5	3	2

The chunking process ensured that long summaries could be processed effectively without losing key information. Chunk boundaries were chosen carefully typically at sentence ends to preserve semantic integrity within each chunk. This way, each segment remained coherent and meaningful when passed through the BART model. Below is a sample chunk from one monologue video:

Chunk 1 of Monologue 4

Cough, fever, nasal congestion, sore throat and shortness of breath are common symptoms.. Avoid close contact with people who are sick.. The Indian Council of Medical Research has detected two cases of human metanomevirus or HMPV in Karnataka.. Wash your hands often with soap and water for at least 20 seconds.. The human metanomevirus is a virus that usually causes symptoms similar to the common cold..

Chunk 2 of Monologue 4

HMPV spreads through direct contact with someone who has it or from touching things contaminated with the virus.. But what is HMPV?. HMPV infections are more common in the winter and early spring.. It often causes upper respiratory infections in people of all ages but it can sometimes cause lower respiratory infections like pneumonia or a toe.. Currently, there is no specific antiviral therapy to treat HMPV and no vaccine to prevent HMPV..

Each chunk was then processed using the BART model. The model paraphrased and condensed the content into a more fluent and natural form, typically producing fewer sentences while preserving the main ideas of each chunk.

After the BART summarization stage, all abstractive outputs from the chunks were combined into one final summary for each video. Careful attention was given to maintaining a logical flow and coherence across segments. The merging process was generally effective in producing a unified summary without significant

redundancy. The following is an example of a final abstractive summary from a monologue video :

Final Abstractive Summary (Monologue 4)

Human metanavirus or HMPV is a virus that usually causes symptoms similar to the common cold. Cough, fever, nasal congestion, sore throat and shortness of breath are common symptoms. Avoid close contact with people who are sick. HMPV infections are more common in the winter and early spring. It often causes upper respiratory infections in people of all ages. HMPV spreads through direct contact with someone who has it or from touching things contaminated with the virus.

After completing the entire process, final abstractive summaries were generated for each video. Table 7 presents the word and sentence counts of these hybrid summaries, which represent the final outputs for each video.

Table 7. Word and Sentence Counts of Final Abstractive Summaries

Type of Video	Original Summary Content (Extractive)		Hybrid Summaries Result	
	Word	Sentence	Word	Sentence
K-Means :				
Monologue 1	302	25	117	12
Monologue 2	402	25	217	14
Monologue 3	354	25	156	11
Monologue 4	139	10	77	6
Monologue 5	159	10	78	7
Interview 1	325	25	116	9
Interview 2	889	50	436	24
Interview 3	740	50	345	25
Interview 4	781	150	781	52
Interview 5	947	50	458	31
Podcast 1	1120	100	403	33
Podcast 2	654	50	274	21
Podcast 3	1115	80	449	31
Podcast 4	938	80	366	26
Podcast 5	1024	80	466	28
Short Interview 1	150	10	70	4
Short Interview 2	93	10	51	5
Short Interview 3	108	10	64	4
Short Interview 4	74	3	28	2
Short Interview 5	189	10	97	6
Hierarchical :				
Monologue 1	291	14	159	14
Monologue 2	290	12	147	13
Monologue 3	289	13	131	10
Monologue 4	103	4	56	5
Monologue 5	109	4	56	4
Interview 1	259	14	127	9
Interview 2	636	21	314	16
Interview 3	569	25	258	16
Interview 4	488	25	218	19
Interview 5	816	21	398	25
Podcast 1	468	28	168	13
Podcast 2	539	21	294	16
Podcast 3	667	22	339	23
Podcast 4	477	19	220	18
Podcast 5	522	26	281	17
Short Interview 1	157	7	49	4
Short Interview 2	94	7	40	4
Short Interview 3	132	8	70	5
Short Interview 4	77	3	27	2
Short Interview 5	102	4	69	5

Table 7 summarizes the word and sentence counts of the final abstractive summaries generated through the hybrid approach. For monologues, the summaries range

from 56 to 217 words with 4 to 14 sentences, indicating concise yet sufficiently detailed representations. Interview summaries tend to be longer, particularly Interview 4 and Interview 5, which reached up to 398–458 words and more than 25 sentences, reflecting their higher complexity and richer content. Podcasts also produced relatively extensive summaries, ranging from 168 to 466 words with 13 to 33 sentences, due to the broader thematic scope of discussion. In contrast, short interviews consistently resulted in the briefest summaries, containing only 27 to 97 words and 2 to 6 sentences, aligning with their inherently compact content.

3.6 Summary Quality Evaluation Using ROUGE

The quality evaluation was conducted by comparing the automatic abstractive summaries with the manual summaries. The ROUGE metrics used include ROUGE-1, ROUGE-2, and ROUGE-L, calculated based on the F1 score. The evaluation results are presented separately for each video. Table 8 presents the ROUGE F1 scores for the hybrid summaries from 20 videos analyzed in this study.

Table 8. ROUGE F1 Scores for Hybrid Summaries

Type of Video	Result		
	R-1	R-2	R-L
K-Means :			
Monologue 1	0.406130	0.115830	0.222222
Monologue 2	0.441687	0.079800	0.133995
Monologue 3	0.535593	0.238908	0.244068
Monologue 4	0.567308	0.300971	0.317308
Monologue 5	0.501931	0.140078	0.239382
Interview 1	0.376623	0.065359	0.142857
Interview 2	0.422977	0.131234	0.234987
Interview 3	0.372807	0.105727	0.175439
Interview 4	0.172648	0.046647	0.095053
Interview 5	0.285714	0.065421	0.105590
Podcast 1	0.316667	0.093645	0.133333
Podcast 2	0.270742	0.048246	0.122271
Podcast 3	0.266453	0.041868	0.112360
Podcast 4	0.252252	0.018083	0.104505
Podcast 5	0.293839	0.050713	0.110585
Short Interview 1	0.577778	0.227273	0.488889
Short Interview 2	0.217143	0.034682	0.114286
Short Interview 3	0.308370	0.106667	0.149780
Short Interview 4	0.239316	0.017391	0.136752
Short Interview 5	0.308300	0.031873	0.134387
Hierarchical :			
Monologue 1	0.346405	0.111842	0.183007
Monologue 2	0.396341	0.085890	0.146341
Monologue 3	0.511111	0.156716	0.170370
Monologue 4	0.401961	0.099010	0.235294
Monologue 5	0.328889	0.098655	0.160000
Interview 1	0.364179	0.072072	0.149254
Interview 2	0.265263	0.063425	0.117895
Interview 3	0.333333	0.082609	0.138528
Interview 4	0.276056	0.028329	0.129577
Interview 5	0.266041	0.050235	0.109546
Podcast 1	0.358839	0.063660	0.131926
Podcast 2	0.297436	0.025773	0.117949
Podcast 3	0.331839	0.072072	0.143498
Podcast 4	0.292011	0.016620	0.137741
Podcast 5	0.340528	0.053012	0.119904
Short Interview 1	0.227642	0.033058	0.097561
Short Interview 2	0.250000	0.080460	0.170455
Short Interview 3	0.292237	0.082949	0.136986
Short Interview 4	0.321429	0.127273	0.232143
Short Interview 5	0.178010	0.031746	0.083770

Based on Table 8, the abstractive summarization model performs differently across video types. Monologues achieve the highest and most consistent ROUGE scores due to their structured, single-speaker nature. Interviews show more variability as a result of multiple speakers and conversational complexity, while podcasts generally obtain lower scores because of topic shifts and spontaneous discussions. Short interviews present mixed results, largely influenced by transcript length and clarity.

The consistently higher scores in monologues can be explained by their relatively simple structure, they involve a single speaker, fewer interruptions, and more organized content flow, making it easier for the model to extract and rephrase central ideas. On the other hand, interviews and podcasts generally have lower scores because of overlapping speech, informal expressions, and frequent topic transitions, which reduce coherence and hinder the model's ability to preserve semantic continuity. Short interviews vary because their brevity sometimes aligns well with the reference summary, but in other cases the limited content provides fewer opportunities for accurate matches.

In addition to hybrid results, this study compares the performance of three summarization approaches: extractive-only, abstractive-only, and hybrid. The goal is to understand the strengths of each strategy.

The results in Table 9 show that the hybrid summarization method generally performs better than extractive-only and abstractive-only approaches, particularly in more complex transcript types such as interviews and podcasts. For interviews, the hybrid (K-Means) approach achieved ROUGE-2 of 0.08 and ROUGE-L of 0.15, compared to extractive-only (0.06, 0.11) and abstractive-only (0.06, 0.09). Similarly, for podcasts, hybrid methods yielded R-2 of 0.05 and R-L of 0.12–0.13, outperforming extractive-only (0.04, 0.08–0.10) and abstractive-only (0.04, 0.08). These improvements indicate that the hybrid strategy effectively balances semantic accuracy and coherence in longer, multi-speaker transcripts.

For monologues and short interviews, the hybrid method produced results that were comparable to or slightly better than other approaches. In monologues, hybrid (K-Means) achieved R-2 of 0.18 and R-L of 0.23, which is close to extractive-only (0.19, 0.21) and slightly above abstractive-only (0.15, 0.24). In short interviews, ROUGE-1 scores across all methods were similar (0.33–0.36), but hybrid approaches maintained a balance between factual coverage and sentence fluency, with R-L around 0.20. Overall, the hybrid approach leverages the strengths of both extractive (factual preservation) and abstractive methods (improved readability), producing summaries that are both informative and coherent.

Table 9. Comparison of ROUGE F1 Scores

Type of Video	Method	R1-Mean	R2-Mean	RL-Mean
Monologue	Extractive-only (K-Means)	0.50	0.19	0.21
	Extractive-only (Hierarchical)	0.42	0.12	0.18
	Abstractive-only	0.41	0.15	0.24
	Hybrid (K-Means)	0.49	0.18	0.23
	Hybrid (Hierarchical)	0.30	0.06	0.13
	Extractive-only (K-Means)	0.23	0.06	0.11
Interview	Extractive-only (Hierarchical)	0.25	0.05	0.11
	Abstractive-only	0.19	0.06	0.09
	Hybrid (K-Means)	0.33	0.08	0.15
	Hybrid (Hierarchical)	0.30	0.06	0.13
	Extractive-only (K-Means)	0.18	0.04	0.08
	Extractive-only (Hierarchical)	0.24	0.04	0.10
Podcast	Abstractive-only	0.16	0.04	0.08
	Hybrid (K-Means)	0.28	0.05	0.12
	Hybrid (Hierarchical)	0.32	0.05	0.13
	Extractive-only (K-Means)	0.36	0.09	0.20
	Extractive-only (Hierarchical)	0.33	0.07	0.13
	Abstractive-only	0.35	0.08	0.19
Short Interview	Hybrid (K-Means)	0.33	0.08	0.20
	Hybrid (Hierarchical)	0.25	0.07	0.14

3.7 Comparison with Previous Studies

The previous study by Gurusamy et al. (2023) proposed a hybrid approach that combined *semantic LDA* and *sentence concept mapping* for extractive summarization, followed by a T5 Transformer model for abstractive summarization. Their evaluation on the DUC2002 dataset (news articles) reported competitive performance, achieving a ROUGE-1 score of 48.35% [30].

In contrast, the present study adopts a hybrid framework that leverages BERT embeddings with clustering methods (KMeans and Hierarchical) for extractive summarization, followed by BART for abstractive summarization. The evaluation was conducted on video transcript data, which presents more complex challenges compared to news text, such as spoken language characteristics, frequent repetitions, and sudden topic shifts. As summarized in Table 10, the results show that the proposed method achieved ROUGE-1 scores of 57% using KMeans and 51% using Hierarchical clustering for Monologue videos.

These results demonstrate that, despite the significant differences in datasets (news articles vs. video transcripts), the proposed hybrid approach remains competitive and even slightly surpasses the previous study in terms of keyword capture (ROUGE-1). The focus on ROUGE-1 is justified because it provides a more stable measure of essential information

representation, whereas ROUGE-2 and ROUGE-L scores in video transcripts tend to be lower due to the unstructured nature of spoken language, frequent repetitions, and topic shifts.

Table 10. Comparison of Hybrid Summarization Results with Gurusamy et al. (2023)

Study	Hybrid Method	Dataset	Data Characteristics	ROUGE-1
Gurusamy et al. (2023)	Semantic LDA + Sentence Concept Mapping + T5	DUC2002 (news articles)	Formal text, well-structured topics, clean sentences	48.35%
This study (KMeans)	BERT Embeddings + KMeans + BART	Video transcripts	Spoken language, frequent repetitions, abrupt topic shifts	57%
This study (Hierarchical)	BERT Embeddings + Hierarchical + BART	Video transcripts	Spoken language, frequent repetitions, abrupt topic shifts	51%

3.8 Future Work Suggestions

To enhance summarization quality, especially for dialog-based content, BART can be combined with extractive methods such as LexRank to improve structure and coherence. Low ROUGE-2 scores indicate the need for better bigram capture through further fine-tuning or more context-aware training data. Human evaluation is recommended to complement ROUGE metrics and reduce stylistic bias, while speaker segmentation may improve results for multi-speaker content.

Future research should also address the limitation of topic uniformity in the dataset, which mainly focuses on HMPV. This lack of thematic diversity may limit semantic variation and hinder model performance. Using datasets with more diverse topics could improve evaluation robustness and overall summarization effectiveness.

4. Conclusions

This study proposed a hybrid automatic summarization approach combining extractive techniques using BERT-based sentence embeddings and clustering with abstractive summarization using the BART model to effectively simplify long and unstructured video transcriptions. Experiments on various video types such as monologues, interviews, and podcasts showed competitive ROUGE F1 scores, especially for content with a narrative structure. Summary quality varied depending on the video type and transcription style, with narrative formats being easier to summarize accurately compared to free-form conversations. These conversations showed lower ROUGE-2 and ROUGE-L scores, likely due to stylistic differences between manual and model-generated summaries. While the model demonstrated promising results, particularly for monologue content, there remains significant room for improvement, especially in handling complex dialogues and conversational formats.

One important limitation of this study lies in the lack of topic diversity in the dataset, since most of the transcriptions revolve around a single theme, which is HMPV. This uniformity may have reduced the variation in semantic representations, making it difficult for clustering and summarization models to detect subtopic boundaries. Future work may consider integrating

speaker segmentation or applying summarization techniques that are more sensitive to conversational context, especially for multi-speaker content. It is also recommended to expand the dataset with more diverse topics to enable more robust evaluations and improve generalizability. Overall, this method shows strong potential for automatic summarization systems, with further development suggested in the form of more adaptive extractive methods, the use of more advanced language models, and the inclusion of human evaluations to improve both summary quality and interpretability.

References

- [1] J. P. Suharsono and D. Nurahman, "Pemanfaatan Youtube Sebagai Media Peningkatan Pelayanan Dan Informasi," *Ganaya J. Ilmu Sos. Dan Hum.*, vol. 7, no. 1, pp. 298–304, Mar. 2024, doi: 10.37329/ganaya.v7i1.3157.
- [2] L. Parabhoi, R. R. Sahu, R. S. Dewey, M. K. Verma, A. Kumar Seth, and D. Parabhoi, "YouTube as a source of information during the Covid-19 pandemic: a content analysis of YouTube videos published during January to March 2020," *BMC Med. Inform. Decis. Mak.*, vol. 21, no. 1, p. 255, Dec. 2021, doi: 10.1186/s12911-021-01613-8.
- [3] J. K. Höhne, T. Lenzner, and J. Claassen, "Automatic speech-to-text transcription: evidence from a smartphone survey with voice answers," *Int. J. Soc. Res. Methodol.*, pp. 1–8, Jan. 2025, doi: 10.1080/13645579.2024.2443633.
- [4] A. Koencke, A. S. G. Choi, K. X. Mei, H. Schellmann, and M. Sloane, "Careless Whisper: Speech-to-Text Hallucination Harms," 2024, doi: 10.48550/ARXIV.2402.08021.
- [5] Le Ngoc Thang, Nguyen Minh Tien, Do Nhat Minh, Nguyen Chi Thanh, and Le Quang Minh, "A method to utilize prior knowledge for extractive summarization based on pre-trained language models," *Vietnam J. Sci. Technol.*, Dec. 2024, doi: 10.15625/2525-2518/20241.
- [6] Wicaksono, G. W., Azhar, Y., Hidayah, N. P., & Andreawana, A., "Automatic summarization of court decision documents over narcotic cases using bert", *JOIV: International Journal on Informatics Visualization*, 7(2), 416-422. 2023. doi: https://dx.doi.org/10.30630/joiv.7.2.1811
- [7] G. Hartawan, D. S. Maylawati, and W. Uriawan, "Bidirectional and Auto-Regressive Transformer (BART) for Indonesian Abstractive Text Summarization," *J. Inform. Polinema*, vol. 10, no. 4, pp. 535–542, Aug. 2024, doi: 10.33795/jip.v10i4.5242.
- [8] M. Ülker and A. B. Özer, "The Bart-based Model for Scientific Articles Summarization," *JUCS - J. Univers. Comput. Sci.*, vol. 30, no. 13, pp. 1807–1828, Dec. 2024, doi: 10.3897/jucs.115121.
- [9] N. B. Raut, A. S. Pranesh, B. Nagulan, S. Pranesh, and R. Vasantharajan, "An Extensive Survey on Audio-to-Text and Text Summarization for Video Content," in *2023 3rd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA)*, Bengaluru, India: IEEE, Dec. 2023, pp. 1251–1257. doi: 10.1109/ICIMIA60377.2023.10426376.

- [10] M. Y. Jaradeh, M. Stocker, and S. Auer, "MORTY: Structured Summarization for Targeted Information Extraction from Scholarly Articles," in *From Born-Physical to Born-Virtual: Augmenting Intelligence in Digital Libraries*, vol. 13636, Y.-H. Tseng, M. Katsurai, and H. N. Nguyen, Eds., in Lecture Notes in Computer Science, vol. 13636, Cham: Springer International Publishing, 2022, pp. 290–300. doi: 10.1007/978-3-031-21756-2_23.
- [11] A. I. Jony, A. T. Rithin, and S. I. Edrish, "A Comparative Study and Analysis of Text Summarization Methods," *Malays. J. Sci. Adv. Technol.*, pp. 118–129, Mar. 2024, doi: 10.56532/mjsat.v4i2.231.
- [12] Z. Edress and Y. Ortakci, "Optimizing Text Summarization with Sentence Clustering and Natural Language Processing," *Int. J. Adv. Comput. Sci. Appl.*, vol. 15, no. 10, 2024, doi: 10.14569/IJACSA.2024.01510115.
- [13] J. Yang, F. Wei, N. Huber-Fliflet, A. Dabrowski, Q. Mao, and H. Qin, "An Empirical Analysis of Text Segmentation for BERT Classification in Extended Documents," in *2023 IEEE International Conference on Big Data (BigData)*, Sorrento, Italy: IEEE, Dec. 2023, pp. 2793–2797. doi: 10.1109/BigData59044.2023.10386783.
- [14] A. Firdaus, N. Yusliani, and D. Rodiah, "Text Summarization with K-Means Method," *Sriwij. J. Inform. Appl.*, vol. 2, no. 2, July 2021, doi: 10.36706/sjia.v2i2.25.
- [15] Lviv Polytechnic National University, Knyaz Roman Str. 5 Lviv, 79013, Ukraine, N. I. Boyko, O. A. Tkachyk, and Lviv Polytechnic National University, Knyaz Roman Str. 5 Lviv, 79013; Ukraine, "Hierarchical clustering algorithm for dendrogram construction and cluster counting," *Inform. Math. Methods Simul.*, vol. 13, no. 1–2, pp. 5–15, Apr. 2023, doi: 10.15276/imms.v13.no1-2.5.
- [16] A. L. Haz, E. D. Fajrianti, N. Funabiki, and S. Sukaridhoto, "A Study of Audio-to-Text Conversion Software Using Whispers Model," in *2023 Sixth International Conference on Vocational Education and Electrical Engineering (ICVEE)*, Surabaya, Indonesia: IEEE, Oct. 2023, pp. 268–273. doi: 10.1109/ICVEE59738.2023.10348186.
- [17] M. Salıcı and Ü. E. Ölçer, "Impact of Transformer-Based Models in NLP: An In-Depth Study on BERT and GPT," in *2024 8th International Artificial Intelligence and Data Processing Symposium (IDAP)*, Malatya, Türkiye: IEEE, Sept. 2024, pp. 1–6. doi: 10.1109/IDAP64064.2024.10710796.
- [18] M. S. Islam and L. Zhang, "A Review on BERT: Language Understanding for Different Types of NLP Task," Jan. 26, 2024, *Computer Science and Mathematics*. doi: 10.20944/preprints202401.1857.v1.
- [19] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of the 2019 Conference of the North, Minneapolis, Minnesota: Association for Computational Linguistics*, 2019, pp. 4171–4186. doi: 10.18653/v1/N19-1423.
- [20] J. Wang et al., "Utilizing BERT for Information Retrieval: Survey, Applications, Resources, and Challenges," *ACM Comput. Surv.*, vol. 56, no. 7, pp. 1–33, July 2024, doi: 10.1145/3648471.
- [21] Suharso, W., Arifianto, S., Wibowo, H., Chandranegara, D.R. and Syaifuddin, S., "Music Features Pada Bidang Ilmu Komputer Menggunakan Modularity Clustering", *Just IT: Jurnal Sistem Informasi, Teknologi Informasi, dan Komputer*, vol 12, no 1, pp. 39-46, December 2022, doi: https://doi.org/10.24853/justit.13.1.%25p
- [22] Mrs. J. Sirisha, V. Lakshmi Prathyusha, P. Naga Anupriya, M. Suma Sri, and P. Naga Hema, "Customer Segmentation using K-Means Clustering," *Int. J. Adv. Res. Sci. Commun. Technol.*, pp. 170–175, Nov. 2022, doi: 10.48175/IJARST-7618.
- [23] Devaghi. P and S. Sudha, "Exploratory Data Analysis and Data Segmentation using K means Clustering," in *2023 3rd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)*, Greater Noida, India: IEEE, May 2023, pp. 145–147. doi: 10.1109/ICACITE57410.2023.10183143.
- [24] R. R. Yager, "Intelligent control of the hierarchical agglomerative clustering process," *IEEE Trans. Syst. Man Cybern. Part B Cybern.*, vol. 30, no. 6, pp. 835–845, Dec. 2000, doi: 10.1109/3477.891145.
- [25] J.-T. Chien, "Hierarchical Theme and Topic Modeling," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 3, pp. 565–578, Mar. 2016, doi: 10.1109/TNNLS.2015.2414658.
- [26] D. Jiang, S. Cao, and S. Yang, "Abstractive summarization of long texts based on BERT and sequence-to-sequence model," in *2021 2nd International Conference on Information Science and Education (ICISE-IE)*, Chongqing, China: IEEE, Nov. 2021, pp. 460–466. doi: 10.1109/ICISE-IE53922.2021.00112.
- [27] C.-Y. Lin, "ROUGE: A Package for Automatic Evaluation of Summaries," *Assoc. Comput. Linguist.*, vol. Text Summarization Branches Out, pp. 74–81, July 2004.
- [28] T. A. Mir, D. Banerjee, D. Upadhyay, and R. S. Rawat, "Binary Classification of Lung Diseases Using Convolutional Neural Networks and Support Vector Machines," in *2024 4th Asian Conference on Innovation in Technology (ASIANCON)*, Pimari Chinchwad, India: IEEE, Aug. 2024, pp. 1–6. doi: 10.1109/ASIANCON62057.2024.10837890.
- [29] J. Chen, S. Wang, Z. Qi, Z. Zhang, C. Wang, and H. Zheng, "A Combined Encoder and Transformer Approach for Coherent and High-Quality Text Generation," 2024, *arXiv*. doi: 10.48550/ARXIV.2411.12157.
- [30] B. M. Gurusamy, P. K. Rengarajan, and P. Srinivasan, "A hybrid approach for text summarization using semantic latent Dirichlet allocation and sentence concept mapping with transformer," *Int. J. Electr. Comput. Eng. IJECE*, vol. 13, no. 6, p. 6663, Dec. 2023, doi: 10.11591/ijece.v13i6.pp6663-6672.