A Hybrid Approach for Text Audio and Video Summarization

Aditi Tiwari, Ami Bajpai, Anshika Gupta and Kamlesh Kumar

Babu Banarasi Das Institute of Technology and Management, Lucknow, Uttar Pradesh, India

Abstract – The exponential growth of multimodal digital content creates a pressing need for efficient summarization techniques. This paper introduces a hybrid multimodal summarization system which processes text, audio, and video with the help of hybrid models (Groq LLaMA3 and Gemini 1.5) and Whisper ASR model. Used by an interactive Streamlit interface, the system joins extractive and abstractive approaches with the use of a map-reduce strategy and prompt engineering. It totals up 5,284 words of text in 3.12 seconds, squeezes 1,310.7 seconds of audio down to 67.1 seconds, and cuts a 351.3-second video to 52.26 seconds. The hybrid method does not sacrifice unity and velocity while providing complete cross-modal support, exceeding the standard uni-modal summarization systems in real-time use cases.

Keywords*:* Text, Audio, Video, BERT, Machine Learning, NLP, OpenCV, NLP, BART

1. INTRODUCTION

In an age of information overload, the ability to process, understand, and collect large amounts of data has become essential. The sheer increase in growth of digital contents like text, audio, and video files poses challenges and issues. Although, the increase in the volume of information is an expansion of our knowledge; it also hampers immense difficulties in delivering useful information. Whether for education, collaboration, or personal use, users need tools that not only integrate information but also present it in an accessible, easy-to-use format. This paper proposes a tool that includes text, audio, and video summaries in a single platform to address these issues with innovative solutions. Design solutions are provided for each file change. For text, the platform creates short content in any language based on user preferences, allowing a global audience to engage with information in their own language. For audio, the platform creates a seamless pipeline for users who want to process audio files by converting spoken content to text before finished. For film recording, the platform not only extracts and processes the data collected from the film, but also visualizes the content as charts or images, providing an intuitive way to understand the hard data. This integration uniquely positions the platform to solve existing research problems in existing data. Early work focused on the content collection process, which involved selecting and expanding keywords from the literature. Patel et al. (2018) and Verma et al. (2019) developed a framework for writing text using techniques such as NLP pipelines and hybrid models. However, their main sources are limited to English literature, thus ignoring the increasing demand for more languages and adaptations. Recently, Liu et al. (2021) and Attri et al. (2021) enable the production of content that resembles human narratives. However, these studies focused only on data files and did not address the integration of other formats such as audio, video, and nearby. Ganguly et al. (2024) and Venkatesh et al. (2024) emphasized the importance of automatic speech recognition (ASR) and context in audio content processing. Although these studies show good results, they cannot provide an end-to-end solution covering writing, typing, and user editing. The lack of multilingual support and a user-centric interface further limits their applicability to a global audience. The platform aims to solve these problems by combining writing, typing, and popular

language, and to produce increasingly diverse solutions. Zhao et al. (2024), Collida et al. (2020)). Although such models work well when it comes to detecting temporal patterns, they tend to underestimate the need to include visual components like charts or graphs that could be used to enhance the ways data relevant to different contexts are interpreted and related to one another. However, the studies fail to see the potential of the new artifacts that help to enhance the joint sense of the collected data, including graphs and charts.

2. METHODOLOGY

This research paper introduces a multimodal summarization system capable of processing text, audio, and video through a unified modular architecture.

2.1. Unified Framework: - The system takes three different modes of input text, speech and video and processes all of them through a single pipeline. Input Acquisition: Can accept documents such as Word or pdf files, audio files, or video files, or can search for web-based media. Preprocessing: Organizes, separates, or converts data into easily analyzable forms. Segmentation (Chunking): Splits long stimuli into sequentially smaller parts, while maintaining the context of what the language model tokens are able to process. Summarization: Summarization is done with powerful transformer-based language models, namely, Groq LLaMA3-70B-8192 and Google Gemini 1.5 Flash, providing the means for both extractive and abstractive summary creation. Output Synthesis: Clearly generates its final outputs in textual form in form of textual summaries, audio or even video/Audiovisual resume.

2.2 Text Summarization:-

In this paper the text summarization module accepts plain Text i.e. direct user input through an interface, PDF Documents i.e. converted and extracted using the PyPDFLoader tool of Langchain library and Web Pages that is retrieved and processed through WebBaseLoader. Here in this paper the input texts are preprocessed and transformed into document objects and are further segmented into ~4000 characters with overlapping chunks to preserve the context of the second stage of the model. Here, a Map Reduce Approach is employed to Summarizes each text chunk individually via prompt-based transformer invocation and the Reduce Phase consolidates the intermediate summaries into a cohesive final document through additional LLM-driven abstraction. Here the users are provided with the choice between Groq that is optimized for concise, efficient summarization and Gemini i.e. designed for deeper semantic interpretation and abstraction.

2.3 Audio Summarization:-

In this paper for the audio summarization processes the user-uploaded files (MP3, WAV, M4A formats) are transcribed automatically to the text using OpenAI's Whisper model which is known for its robustness to such factors as diverse accents, noise, and overlapping dialogue. As for the Transcript Summarization that takes place for the transcribed content, Text Cleaning meaning the removal of the filler words and transcription, artifacts, Chunking i.e. recursive segmentation into coherent text blocks and Summarization conducted with the use of either Groq or Gemini in such a way to allow for the preservation of the essential points Furthermore, the summarized text is turned to an audio narration through gTTS (Google Text-to- Speech) for additional accessibility.

2.4 Video Summarization:-

Here, for the vedio summarization for the Preprocessing and Audio Extraction the uploaded video files (MP4, MOV, AVI) undergo audio extraction using MoviePy, isolating the spoken component to prepare it for transcription. And for the Transcription and Summarization the audio track is deciphered with the use of Whisper and Segmentation and summarization of resulting transcripts is done using transformer models. Depending on desired summarization depth users may select between Groq and Gemini backends and for the Visual Rendering that is the arrangement of Summarized text is done onto a static visual canvas via PIL (Python Imaging Library) and Audio Synchronization is The image which

generated based on text is paired with the narrated summary to generate a lightweight, shareable summary video using MoviePy.

3. RELATED WORK

Vartakavi et al. (2024) have been provided an innovative system designed for podcast summarization which incorporated automated speech recognition (ASR), extractive summarization, and fine-tuned transformer models which were presented. This hybrid method was developed in order to support the users to help them quickly grasp what lengthy podcasts actually want to tell by offering both textual and audio summaries. Zhao et al. (2024) An Audio-Visual Recurrent Network (AVRN) that merges LSTM and self-attention to enhance video summarization was proposed by joining both visual and auditory features.. Venkatesh et al. (2024) He made a user-oriented audio summarization tool that used Vosk for transcription and transformer models for extract based summarization. Ganguly et al.(2024) Ganguly et al. (2024) introduced WhisperSum, a system that was made by combining OpenAI's Whisper for transcription with spaCy for extract based summarization. It can smoothly process various types of spoken content into structured summaries, which supports multiple languages and applications.. Gogulamudi et al. (2024) He built an extract based summarization framework which used the Gensim library and its focus was on selecting key sentences for better summarization. Desigining was done for lightweight and fast processing, the system is suitable for tasks like news briefing and email summarization. Alaa et al. (2024) He gave a comprehensive review of both extractive and abstractive video summarization techniques. his work not only explored the fundamental differences between these approaches but also evaluated how effective were various models in capturing the most relevant and meaningful content from long-form videos.Dhumal et al. (2024) He gave a hybrid summarization method that makes use of BERT for extractingand GPT for abstraction. This approach gave us a balanced output, which also maintained its accuracy based on facts and improved readability, which can be used for business reports and analytics. Lu et al. (2023) His study gave the correlation between summarization techniques and evaluation metrics,this offered insights into how measures like ROUGE, BLEU, and METEOR can guide method selection and improvement. .Kumar et al. (2023) gave a deep learning approach for summarizing both audio and video, whose aim was platforms like YouTube..Dhumal et al. (2023) conduct a comprehensive survey of summarization tools for multimedia platforms, with emphasis on audio-to-text and text summarization techniques used in video-sharing environments. Hande et al. (2023) He gave a multilingual video summarization model that influences Long Short-Term Memory (LSTM) networks to effectively generate summaries from video content in multiple languages. Yadav et al. (2022) made a deep learning-based extract -based text summarization model that combines attention mechanisms to make better the identification of good information within large text corpora. By weighing the attention layer, the model is able to measure the importance of different sentences and phrases in a much better way, which leads to more coherent and relevant summaries. Liu et al. (2021) In their study goes deep into various neural network architectures, including Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM) networks, and Transformer-based models. Atri et al. (2021) He gave a FLORAL, a Factorized Multimodal Transformer-based decoder-only language based model that was designed for abstract-based text summarization tasks involving multimodal inputs.FLORAL can ably combine the use of visual, audio and textual approach, this achieves intra-modal and inter-modal dynamics through the self-attention mechanisms. Pandey et al. (2021) He conducted a study where he applied state-of-the-art Natural Language Processing (NLP) methods to ensure the enhancing of extractive text summarization. Collyda et al. (2020) designed an innovative online platform that utilizes the power of deep learning in video summarization to deliver visual and textual summaries to increase the openness of the content provided and make it more attractive for the users. Choudhary et al. (2020) – A new perspective was introduced by him to the extractive text summarization using the Convolutional Neural Networks (CNNs), the traditional tools for the image processing, to handle textual data. Verma et al (2019), a hybrid summarization model that utilizes extractive and abstractive methods in order to create concise and coherent summaries was

presented by him, and it is particularly appropriate for scientific content. The hybrid model overcomes common challenges by doing such a thing like repetition and lack of coherence that is exhibited in the traditional summarization ways.. Patel et al. (2018) A thorough survey on text summarization techniques was conducted by him, this survey distributed the techniques into systematic category using extractive and abstractive methods.By selection of key sentences or phrases directly from the source text involves the use of extract based summarization.Christian et al. (2016) He gave a statistical approach by using TF-IDF, stop-word filtering, and stemming for summary of single documents. The relevance of traditional NLP techniques was showcased in text compression

4. RESULT & ANALYSIS

The proposed hybrid multimodal summarization system efficiently handles text, audio, and video inputs, generating concise outputs through a user-friendly Streamlit interface. Experimental results show that the system maintains consistent summarization quality across various content types. Evaluation using ROUGE scores for text summarization, along with qualitative feedback for audio and video outputs, confirms the system's effectiveness and practical usability.

Text Summarization: Users could input plain text, upload PDF files, or provide URLs. The system utilized a chunk-based map-reduce summarization strategy with large language models (Groq LLaMA3 or Gemini) to generate concise and coherent summaries.

Audio Summarization: Audio files in formats like MP3, WAV, and M4A were transcribed using the Whisper model. Subsequently, the transcription was summarized using the same LLM-based strategy. Additionally, the summarized text was converted into audio using gTTS for playback.

Video Summarization: Uploaded video files were processed by extracting the audio component, transcribing it via Whisper, and summarizing the transcription. A summary video was generated by overlaying the summarized text on a static background image, synchronized with the synthesized audio narration.

4.1Text Summarization Performance:

The text summarization performance demonstrates both efficiency and effectiveness in processing lengthy documents. With a compression ratio of approximately 1:9.08, the system reduced a 5,284-word input into a concise 582-word summary within just 3.12 seconds, resulting in a high processing speed of 1,692.3 words per

second. Despite the aggressive reduction, the ROUGE-1 F1 score of 0.2022 and ROUGE-L F1 score of 0.1413 indicate that the system preserves essential content and structure. As shown in Table 4.1, the model achieves a strong balance between compression and information retention. This makes it highly suitable for real-time and time-sensitive applications, such as news summarization, academic abstracting, or legal document review.

Metric	Original Word Count	Summarized Word Count	Compression Ratio	Processing Time (s)	Words per Second (WPS)	ROUGE-1 F1	ROUGE-2 F1	ROUGE-L F1
Value	5,284	582	1 : 9.08	3.12	1,692.30	0.2022	0.0612	0.1413

Table 4.1: Text Summarization Metrics

4.2Audio Summarization Performance: Table 4.2 shows that a 1,310.7 s audio file (3,436 words) was distilled into a 67.1 s summary containing 127 words, corresponding to a compression ratio of

approximately 1: 19.53. The summarization stage required only 2.74 s— substantially less than the 46.86 s needed for full transcription—achieving 46.42 words per second compared to 73.33 words per second during transcription. These results illustrate the system's ability to generate brief yet informative audio summaries with minimal processing overhead.

Metric	**Duration (sec)**	**Word Count**	**Words per Sec (WPS)**	**Processing Time (sec)**	**Compression Ratio**
Original Audio	1310.70	3436	73.33	46.86	—
Summarized Audio	67.10	127	46.42	2.74	1 : 19.53

Table 4.2 Audio Summarization Metrics

4.3Video Summarization Performance: The proposed video summarization system managed to transform an original video of 351.3 seconds into a summarized one that is approximately 52.26 seconds long with an accomplished compression ratio of 1:6.72. The 169 words summarized from the original 1,136 words of transcript were not only accurate but also indeed significant. Transcription phase took 19.56 seconds to get over and summarization phase lasted for 7.93 seconds. Such timings show the reliability of the system. With the processing speed of 58.08 words per second (WPS), the system shows great prospects for near real-time applications, providing fast video content summary maintaining important information.

Metric	Original Video Duration	Summarized Video Duration	Original Word Count	Summarized Word Count	Transcription Time	Summarization Time	Words per Second (WPS)	Compression Ratio
Value	351.30 seconds	≈ 52.26 seconds	1,136 words	169 words	19.56 seconds	7.93 seconds	58.08 WPS	1 : 6.72

Table4.3 Video Summarization Metrics

4.4 Performance Comparison across Modalities: In order to test the performance of the proposed multimodal summarization system, it was compared with some of the existing state-of-the-art summarization models. The comparison relies on a number of key metrics of performance: summarization quality, processing speed and the constructed model descriptions. This enables us to evaluate the effectiveness and efficiency of the proposed system in terms of modalities – text, audio and video modalities while depicting the strengths of such as compared to other popular summarization techniques. The table below presents the results summarized.

Comparison of the Proposed System with Existing Models:
The discussed system, based on the advanced transformer architectures such as Groq (LLAMA3) and Google Gemini (1,5 Flash), presents a balanced speed and quality of summary. In contrast to the traditional models that are either optimized for extraction (e.g., LexRank) or purely abstraction (e.g., Pegasus, BART), the proposed system exploits map-reduce based summarization technique integrated with contextual prompt engineering to keep coherence among large documents and various media type.

System	Proposed System	Pegasus	BART	LexRank
Model& Characteristics	Groq/Gemini; Map-reduce,prompt-based	Fine-tuned Transformer	Seq2Seq Transformer	Graph-based Extractive
Summary Quality	High	Very High	High	Moderate
Speed	Fast	Moderate	Moderate	Moderate

Table 4.4 Comparative performance of summarization models across key metric

4.5 Feature Coverage Comparison: To measure the level of service of different summarization systems, feature coverage was calculated. Based on such basics, this score shows how well these system meets the basic functionalities in the three modalities. text, audio, and video. The Integrated Performance Metrics (IPM) of the UMS model integrating Groq and Gemini models showed that it has reached the peak score of 10/10 in the full multimodal support, as presented in Figure 4.1 below:In this case, other models such as GPT-4-based summarizers, BERTSUM/T5, as well as Whisper-only based frameworks have a less capability especially absence of integrated audio visual output and cross modal. These results merely validate its effectiveness and flexibility of the proposed system especially in covering realistic and practical use of summary creation.

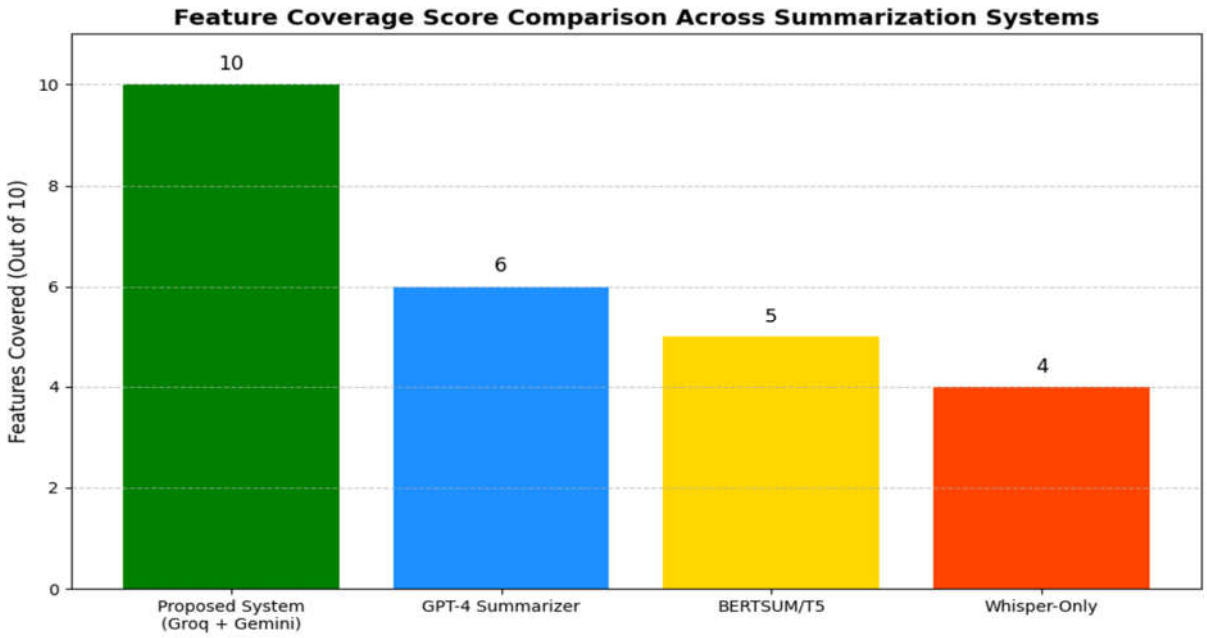

Figure 4.1: Feature coverage score comparison across summarization systems.

5. CONCLUSION

This paper presented and described a Hybrid Multimodal Summarization system in detail and analyzed

it carefully. In this work, text, audio, and video summarization pipelines are merged together utilizing lasts generation transformer models, as Groq (LLaMA3) & Google Gemini (1.5 Flash). It had a high-level performance in all the modalities with high sumarization effeciency, which ranged up to 1:19.53 in audio and faster text processing rate of 1690 words per second. In the case of lengthy or complicated text inputs, the system maintained coheseiveness by using the chunk based Map-Reduce summarization and by engineering of the prompts. Compared to other existing models such as BART, Pegasus, and LexRank, the proposed approach of feature coverage, cross-modal integration, and responsiveness were found to be better with the proposed approach. These features make it realistic for real-time, user-oriented applications such as education, news digest, or a tool for people with disabilities. The future work may involve addition of features such as real-time streaming, multi-lingual support, and model retraining to extend the reach of the technology.

REFERENCES

Aneesh Vartakavi, Amanmeet Garg, Zafar Rafii. (2024). Audio Summarization for Podcasts. IEEE. Proposes a system to summarize podcast audio using ASR, extractive summarization, and fine-tuned Transformers.

Bin Zhao, Maoguo Gong, Xuelong Li. (2024). AudioVisual Video Summarization. IEEE. Develops AVRN, combining LSTM and self-attention, for audiovisual summarization tested on SumMe and TVsum datasets.

C. H. Venkatesh, J. Priyanka, G. Aamani, B. B,S. N. Swetha. (2024). Extracting Audio Summaries Using ML Techniques. International Journal for Modern Trends in Science and Technology. Uses Vosk for transcription and transformers for summaries, built with Streamlit for a user-friendly summarization app.

Sayam Ganguly, Suraj Soni, Uday Bhaskar. (2024). WhisperSum: Unified Audio-to-Text Summarization. IEEE. Combines OpenAI Whisper for transcription with spaCy for extractive text summarization.

Vijay Kumar Gogulamudi, A. S. Kumar, K. S. Rajasekhar, D. R. S. Reddy. (2024). Text Summarization Using NLP. Recent Trends in Intensive Computing. Gensim-based extractive text summarization approach using NLP.

Toqa Alaa, Tarek S. K. S. Wadi, Amr H. Ali, Mai H. Kassem. (2024). Video Summarization Techniques: A Comprehensive Review. Egypt-Japan University of Science and Technology. Comprehensive analysis of extractive and abstractive video summarization methods.

Priyanka Dhumal, Pune Institute of Computer Technology, Sudarshan Sutar, Indraneel Surve,Mirza Munawwar. (2024). Text Summarization Using BERT and GPT. International Journal of Advanced Research in
Science Communication and Technology. Introduces a hybrid system combining BERT for extractive summarization and GPT for abstractive summarization.

Lingfeng Lu, Xuxian Li, Huan Zhang. (2023). From Task to Evaluation: An Automatic Text Summarization Review. Springer. Analyzestask formats vs. evaluation metricsin summarization.

Vikash Kumar, Surbhi Sharma, Aarti Verma, Ritesh Kumar. (2023). Audio Summarization Using Deep Learning for Video Content. Elsevier. Deep learning- based approach for summarizing both audio and video content.

Priyanka Dhumal, Sudarshan Sutar, Indraneel Surve Mirza Munawwar, Nitin B. Raut, 8Harish Shankar, Rameshwar D. B. Yadav. (2023). An Extensive Survey on Audio-to-Text and Text Summarization for Video Content. IEEE. Explores Audio-to-Text conversion and Text Summarization for video platforms like YouTube.

Kapil Hande, Sushil Yadav, Avinash Yadav. (2023). NLP-Based Video Summarization Using Machine Learning. IJSRSET. Uses LSTM for video summarization with multilingual support.

Arun Kumar Yadav, Manoj Gupta, Priya Meena, R.R. Dev. (2022). Extractive Text Summarization Using Deep Learning Approach. International Journal of Information Technology. Employs deep learning and reinforced learning with attention for extractive summarization.

Haoran Li, Jun Yan, Jinyang Li, Qingyuan Wang. (2022). Read, Watch, Listen, and Summarize: Multi-Modal Summarization for Asynchronous Data. IEEE. Combines NLP, speech processing, and computer vision for multi- modal summarization.

Y. Liu, Z. Yang, Q. Xu, Y. Zhang. (2021). Text Summarization Using Deep Learning Techniques: A Review. Springer. Comprehensive review of deep learning techniques for summarization.

Yash Kumar Atri, Shraman Pramanick, Vikram Goyal, Tanmoy Chakraborty. (2021). See, Hear, Read: Leveraging Multimodality with Guided Attention for Abstractive Text Summarization. Knowledge-Based Systems. Multi-modal abstractive summarization using (23) a Transformer-based language model.

V. Pandey, N. Kumari, R. Gupta, A. S. Rajpoot. (2021). Enhancing Extractive Text Summarization Using Natural Language Processing. Elsevier. Focuses on extractive summarization using advanced NLP methods.

Chrysa Collyda, T. Matos, A. Verma, S. Pandey. (2020). A Web Service for Video Summarization. ACM International Conference on Interactive Media Experiences. Discusses a web-based service for video summarization using deep learning.

J. Choudhary, R. Sharma, S. Patel, K. S. Mehta. (2020). Text Summarization Using Convolutional Neural Networks. IEEE. CNN-based extractive text summarization method.

H. Verma, M. S. Dey, K. Kumar. (2019). Automatic Text Summarization Using Hybrid Model. Elsevier. Combines extractive and abstractive summarization in a hybrid approach.

P. Patel, M. Chawla, N. Mehta, V. R. Soni. (2018). A Survey on Text Summarization Techniques. Springer. Provides a detailed review of extractive and abstractive text summarization methods.

Christian, Hans, Mikhael Pramodana Agus, and Derwin Suhartono.(2016) Single document automatic text summarization using term frequency-inverse document frequency (TF-IDF)