iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 12 **Issue:** III **Month of publication:** March 2024

DOI: https://doi.org/10.22214/ijraset.2024.59365

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 12 Issue III Mar 2024- Available at www.ijraset.com

Video Transcripts Summarization using OpenAI Whisper and GPT Model

A D Bhargavi[1], Prof. Ch. D V Subba Rao[2]
Department of Computer Science and Engineering, Sri Venkateswara University College of Engineering, Tirupati-517501

Abstract: ***In today's digital age, a vast amount of video content is generated and shared on the internet every minute. However, extracting relevant information from these videos can be time-consuming and challenging. This is where video transcript summarization comes in, providing a concise summary of video content without the need to watch the entire video. The video transcript summarization system aims to streamline the process of extracting key insights and information from video content by generating concise and informative summaries from their transcripts. In the dynamic landscape of video content, existing approaches to transcript summarization have encountered challenges that limit their effectiveness. Issues such as compromised accuracy, prolonged processing times and a restricted focus on YouTube's captioned videos have prompted the exploration of alternative solutions. This paper introduces a novel approach that tackles these limitations head-on, employing innovative technologies to redefine the landscape of video transcript summarization. To address the identified limitations, we propose an alternative strategy that hinges on the utilization of advanced technologies like OpenAI Whisper Model Automatic Speech Recognition(ASR) for video's Transcription and Generative Pre-trained Transformer (GPT) model for Summarization. Through the integration of the OpenAI Whisper model for transcription and GPT models for summarization, our proposed approach strives to redefine standards in accuracy, processing efficiency, multilingual support, and exhibit the capability to generate either extractive or abstractive summaries. In the context of natural language processing (NLP) and text summarization ROUGE (Recall-Oriented Understudy for Gisting Evaluation) metrics will be used to evaluate the quality of generated summaries compared to human referenced summaries.***

Keywords*: Transcript summarization, OpenAI whisper ASR, Generative Pre-trained transformer, Extractive Summarization, Abstractive Summarization, Natural Language Processing, ROUGE*

I. INTRODUCTION

Summarizing longer videos through video transcripts summarization, a key application of NLP, aims to create brief and cohesive summaries that effectively capture the essential information from the original content. This technology proves beneficial in time-constrained scenarios or when a rapid overview of the video content is required. Additionally, as videos stand as a primary source for acquiring information and knowledge, individuals often prefer them over reading lengthy documents. However, the abundance of videos on a single topic may lead to a challenge of choosing the most relevant one. This is where video transcript summarization becomes crucial, offering concise summaries to aid in the selection of desired videos. By providing summaries, users can make informed decisions on which videos align with their specific learning objectives, enhancing the overall efficiency of knowledge acquisition. Utilizing a blend of techniques such as automatic speech recognition for text extraction and text summarization to generate summary, video summarization seeks to distil the core essence of the video into a concise textual format.

A. Automatic Speech Recognition

Automatic Speech Recognition (ASR) is a transformative technology that enables machines to convert spoken language into written text automatically. This innovative field within the broader realm of Natural Language Processing (NLP) has gained prominence for its applications in various industries, from transcription services to voice-activated virtual assistants.ASR systems leverage advanced algorithms and machine learning techniques to process and interpret spoken language, transforming audio signals into transcriptions that can be easily understood by computers. The primary goal of Automatic Speech Recognition is to bridge the gap between human communication and machine understanding, allowing for seamless interactions in voice-driven applications.

B. Text Summarization

Text Summarization is a fascinating area within Natural Language Processing (NLP) that focuses on automatically condensing lengthy textual content into shorter, coherent versions while retaining the essential information and meaning.

The primary objective of text summarization is to distil the key insights from a document, making it more accessible and time-efficient for users. Extractive and Abstractive Summarization are two approaches to perform text summarization. In extractive summarization, the summary is generated by selecting and extracting specific sentences or passages directly from the original text. The selected sentences are considered representative of the main ideas and key information in the document. This approach doesn't involve rephrasing or generating new sentences; it relies on extracting existing content. In abstractive summarization, the system generates a summary by paraphrasing and rephrasing the content in a way that captures the core meaning. It involves understanding the input text and creatively producing a condensed version that may not necessarily include verbatim sentences from the original.

II. RELATED WORK

"Learning to Summarize YouTube Videos with Transformers: A Multi-Task Approach" published by R. Sudhan, D.R. Vedhaviyassh, G. Saranya in August 2023. In this paper, they presented a system for summarizing YouTube videos that use NLP and machine learning techniques to retain the important details without any data loss. The proposed approach involves downloading the video' audio, converting it to WAV format, performing speech-to-text conversion using the Hugging Face Automatic Speech Recognition model, and then using transformers and pipeline for summarization.

"Youtube Transcript Summarizer Using Flask" published by Surabhi Bandabe1, Janhavi Zambre, Pooja Gosavi, Roshni Gupta, Prof. J. A. Gaikwad in April 2023. This paper Utilized a Python API, find the transcripts and subtitles for a particular YouTube video ID. If transcripts are available then perform text summarization on obtained transcripts using HuggingFace transformers else extract audio from the video by using speech to text conversion and summarize the converted text.

"Summarization of Video Clips using Subtitles by Eleesa Anil, Sherine Sebastian, Janice Johnson, Janhavi Rane, K.Priya Karunakaran in March 2023.This paper proposed a method to create a video summary in a way that it contains only the necessary and important information in a concise format by using various NLP algorithms such as Textrank, LexRank and LSA(Latent Semantic Analysis).

"Text Summarization using Transformer Model" published by Jaishree Ranganathan, Gloria Abuka, in November 2022. This paper proposed a text summarization method based on the Text-to- Text Transfer Transformer (T5) model. They used the University of California, Irvine's (UCI) drug reviews dataset and manually created human summaries for the ten most useful reviews of a particular drug for 500 different drugs from the dataset. They fine-tuned the Text-to- Text Transfer Transformer (T5) model to perform abstractive text summarization. The model's effectiveness was evaluated using the ROUGE metrics, and model achieved an average of ROUGE1, ROUGE2, and ROUGEL scores of 45.62, 25.58, and 36.53, respectively.

In addition to above literature survey, many more papers were published for video transcript summarization. The primary objective of all existing papers and the proposed approach is to generate a concise summary. The difference lies in the usage of techniques and approaches, which have encountered challenges such as lower ROUGE F1 score and increased processing time. In addition most existing models have predominantly focused on either extractive summarization or abstractive summarization but not both.

III. DESIGN AND METHODOLOGY

A. System Design

The proposed system design as shown in Figure 1 comprises into five stages which together will make up the final product. Also the flow chart of system design is as shown in Figure 2 which explains step by step working flow at user interface level.

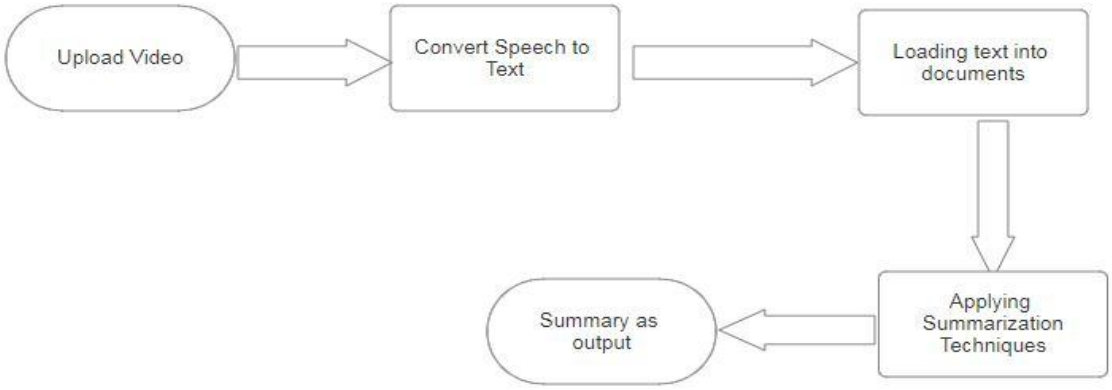


Figure 1.Proposed System Design

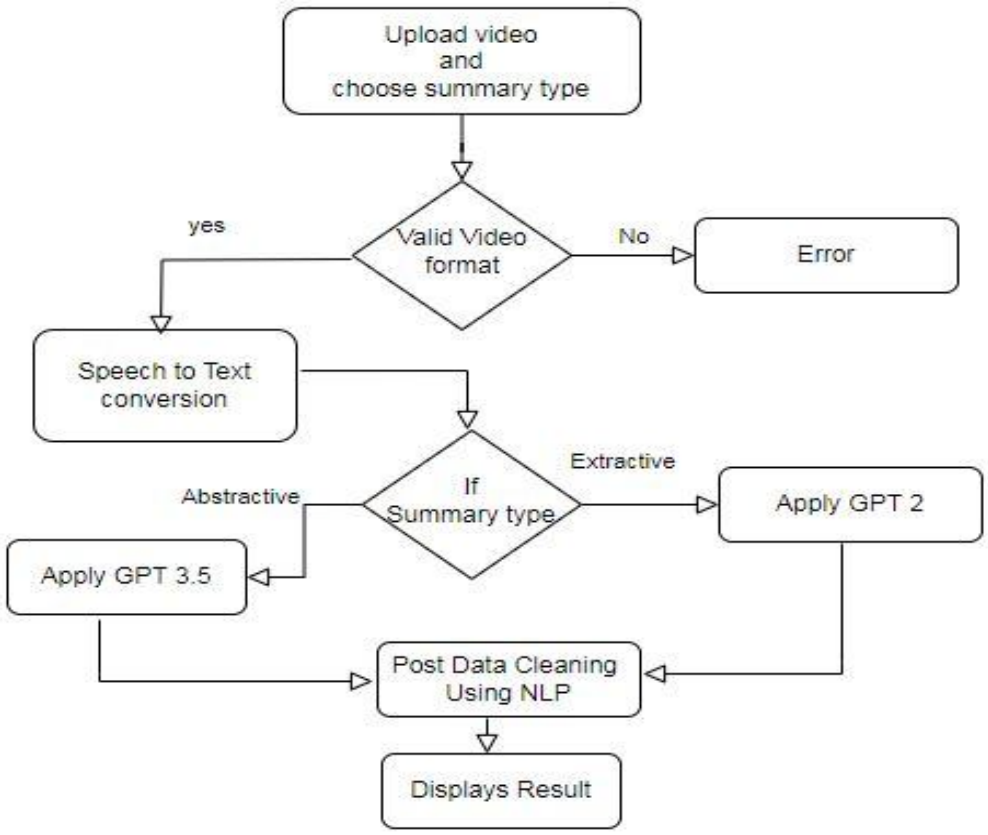


Figure 2: Flow chart of System Design User Interface

As shown in the Figure 2, the system flow starts with uploaded video and choosen summary type. For valid video format, action starts with generating a transcript from the video and undergo some text preprocessing using Lang chain .Once transcript generated, summarization is performed based on user's choice using GPT models. At last we will apply some post data cleaning technique for generated summary to achieve more accurate format of summary to be displayed.Else for invalid video format, system simply results in error.

B. Methodology

As shown in Figure 1, our proposed system entails different recent approches in NLP, automatic speech recognition and summarization models to get conise summary for user uploaded video.

1) Speech to Text Conversion

Speech-to-text conversion, also known as Automatic Speech Recognition (ASR), is the technology that transforms spoken language into written text. This process involves the use of algorithms and models to recognize and transcribe spoken words into a textual format. In our design, OpenAI Whisper ASR model will be used for speech to text conversion.So uploaded video will be auto translated and transcripted to English language using OpenAI Whisper ASR.

2) Text Preprocessing

Text preprocessing is a crucial step in natural language processing (NLP) and machine learning tasks. It involves cleaning and transforming raw text data into a format that can be easily understood and utilized by algorithms. Langchain used as text pre processor as it converts unstructured data into structured documents for easy analysis.

3) Transcript Summarization

Transcript summarization involves condensing the content of a transcript from video into a shorter and more concise form while retaining the key information.To achieve this,GPT Text summarization models like GPT 2 will be used to generate extractive summary and GPT 3.5 used for abstractive summary.

4) Post Text Cleaning using NLP

Once summary got generated, we perform post processing cleaning technique to remove unwanted characters, extra spaces from summary by using Natural Language Processing to make it more accurate.

5) User Interface using Flask

Video Transcript Summarization is a web application that allows users to upload video and get their summary .This application is developed by using Flask framework.

IV. METRICS ANALYSIS AND RESULTS

A. Metrics Analysis

Evaluation metrics for automatic speech recognition is Word Error Rate (WER) and for text summarization is ROUGE metrics.

1) Word Error Rate

WER is the ratio of errors in a transcript to the total words spoken. A lower WER in speech-to-text means better accuracy in recognizing speech. For example, a 20% WER means the transcript is 80% accurate. We calculated the WER for a few transcriptions of the proposed OpenAI whisper ASR and the Google ASR . The results say that OpenAI whisper model has high accuracy than Google ASR. WER between Google and OpenAI Whisper ASR for few samples is shown in Figure 3.

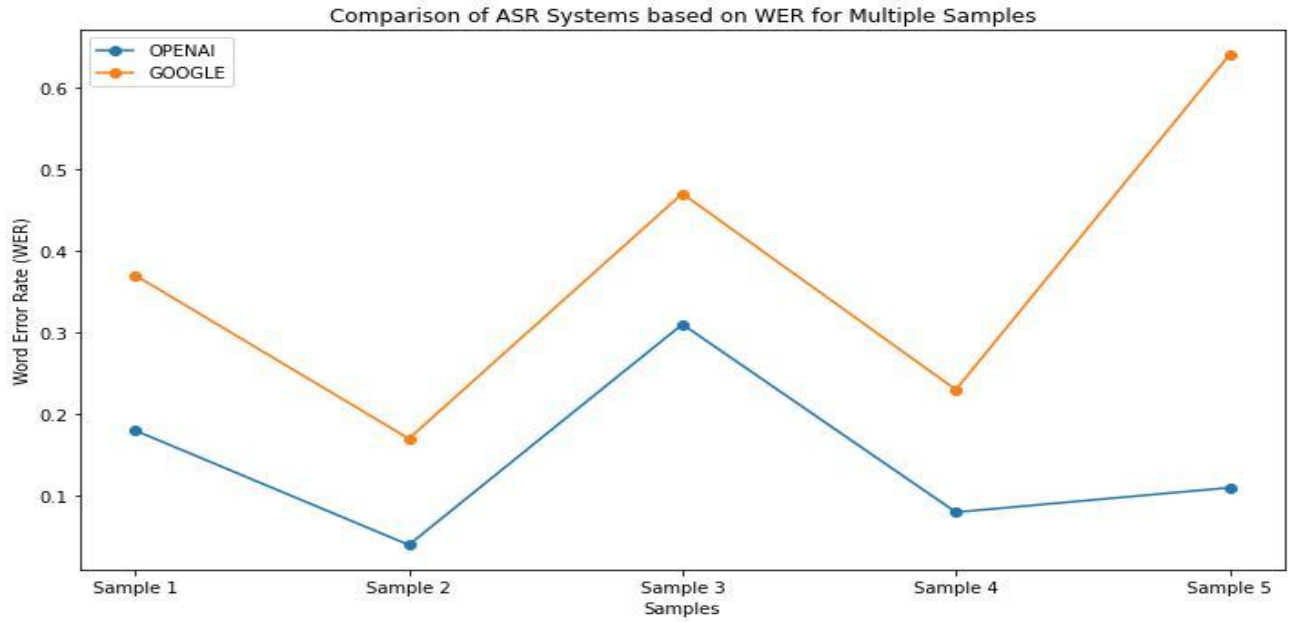


Figure 3 WER comparison between OpenAI Whisper and Google ASR

2) Rouge

ROUGE, or Recall-Oriented Understudy for Gisting Evaluation, is a set of metrics and a software package used for evaluating automatic summarization in natural language processing. The metrics compare an automatically produced summary against a set of references (human-produced) summary. Both precision and recall play a very important role in determining the quality of the summary generated by a model. Moreover, the F1 — score helps in maintaining this balance by combining both precision and recall into a single value.

The three main variants of Rouge scores are Rouge — 1 measures the overlap of unigrams (single words) between the generated summary and the reference summary.Rouge — 2 evaluate the overlap of bigrams (pairs of adjacent words) between the generated summary and reference summary.Rouge — L measures the longest subsequence between the generated summary and reference summary. A subsequence is a sequence of words that appear in the same order, but not necessarily consecutively.
In this paper, we evaluated ROUGE metrics average F1 score between few samples of human generated and system generated summary for different approaches like proposed system and existing system models in both type of summarizations.

For Abstractive Summarization approaches:
ROUGE metrics for existing system models like T5, BART and proposed system model GPT 3.5 are tabulated in Table 1.Their metrics visualization is shown in Figure 4. From these visualization we can clearly see that GPT 3.5 model has high F1 score than other models.

Model	ROUGE -1 F1	ROUGE-2 F1	ROUGE-L F1
GPT 3.5	0.71	0.58	0.65
BART	0.56	0.32	0.39
T5	0.35	0.16	0.29

Table 1 ROUGE metrics between GPT 3.5, BART and T5 abstractive summarization models

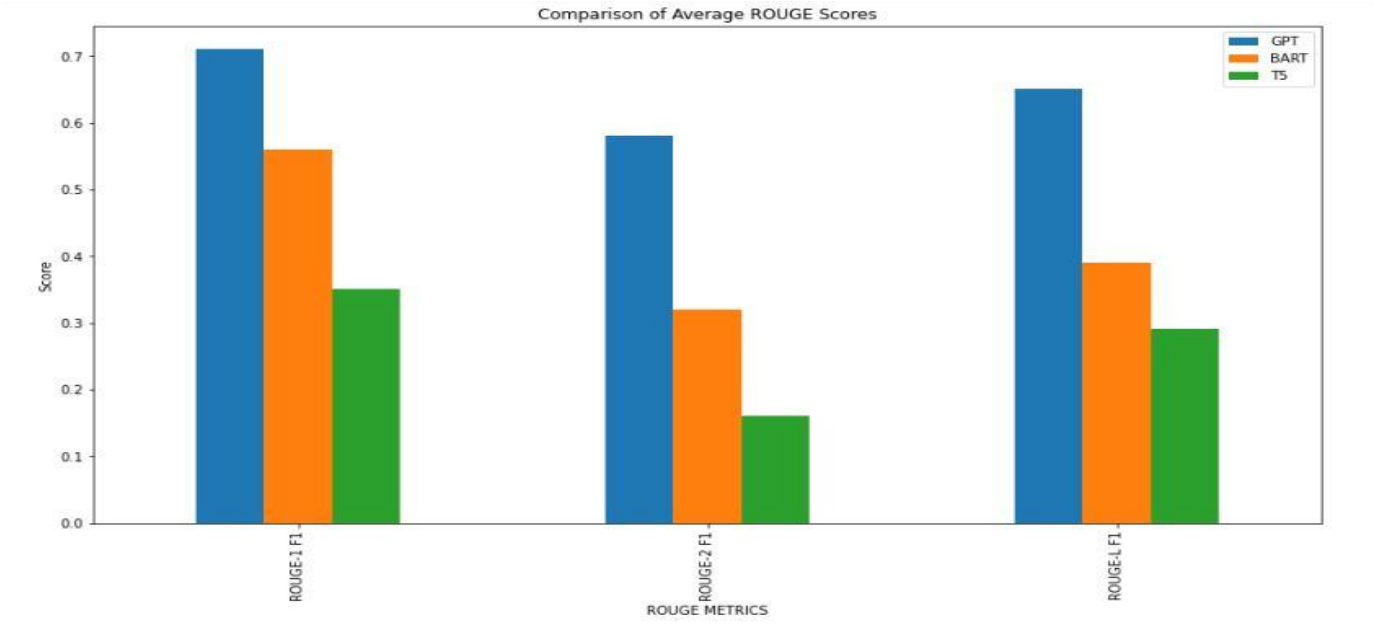


Figure 4 Visualization of F1 score between GPT 3.5, BART and T5 abstractive summarization models

For Extractive Summarization approaches:
ROUGE metrics for existing system models like XLNET, BERT and proposed system GPT 2 model are tabulated in Table 2. Their visualization is shown in Figure 5 which shows that GPT 2 model has high F1 score than other models.

MODEL	ROUGE -1 F1	ROUGE-2 F1	ROUGE-L F1
GPT 2	0.60	0.44	0.50
XLNET	0.35	0.30	0.25
BERT	0.11	0.01	0.06

Table 2 ROUGE metrics between GPT 2, XLNET and BERT extractive summarization models

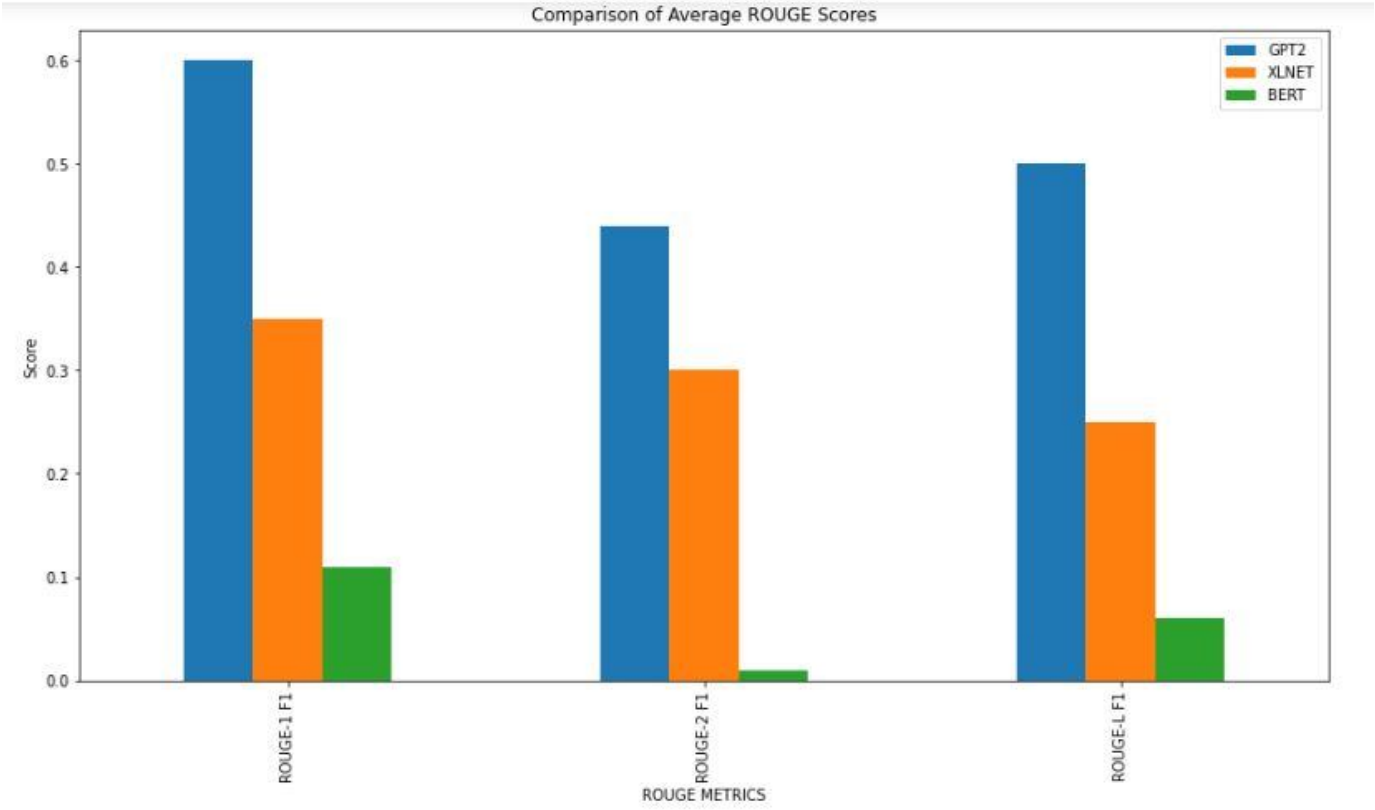


Figure 5 Visualization of F1 score between GPT 2, XLNET and BERT extractive summarization models

B. Results

The developed Home Page is shown in Figure 6 which allows user to upload video by choosing file, choose summary from drop down list and click on Generate button. After successful execution, summary will be displayed in text area. For example the abstractive and extractive summary for same sample video is shown in Figure 7 and Figure 8. Download button is used to download the displayed summary into a text file for future reference when user has to choose video based on multiple summaries. Downloaded text file sample is shown in Figure 9.

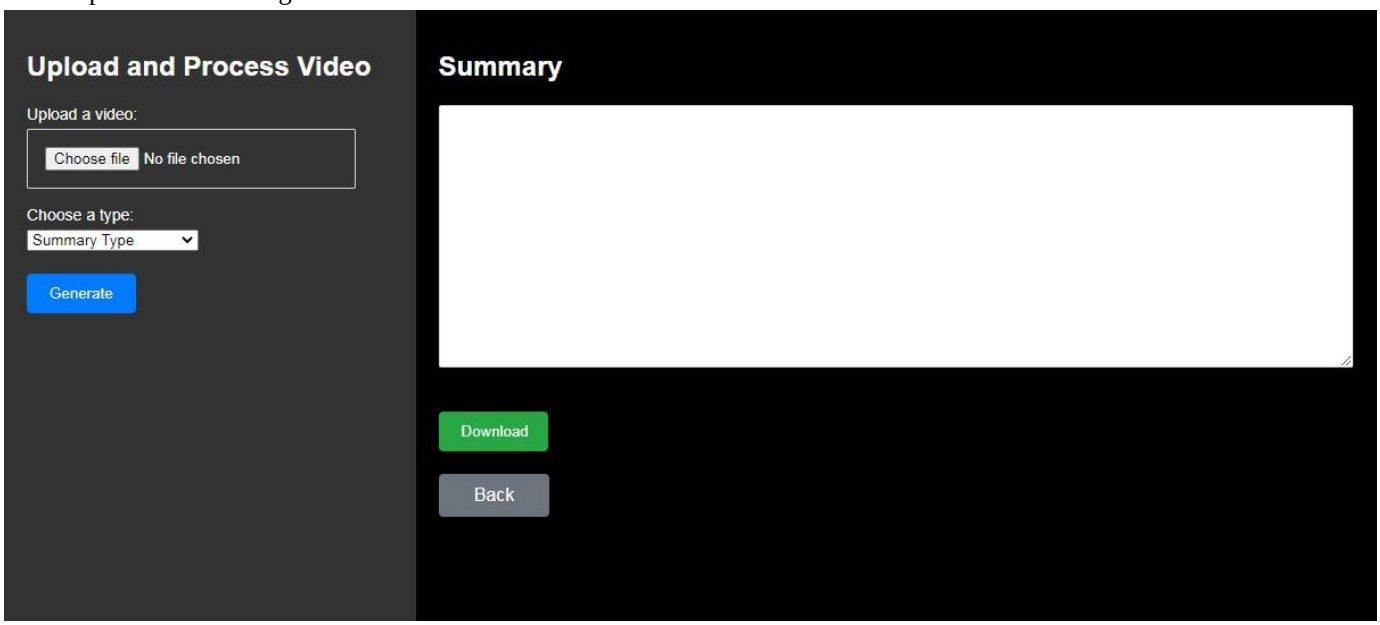


Figure 6 Screenshot of Home page and summary display

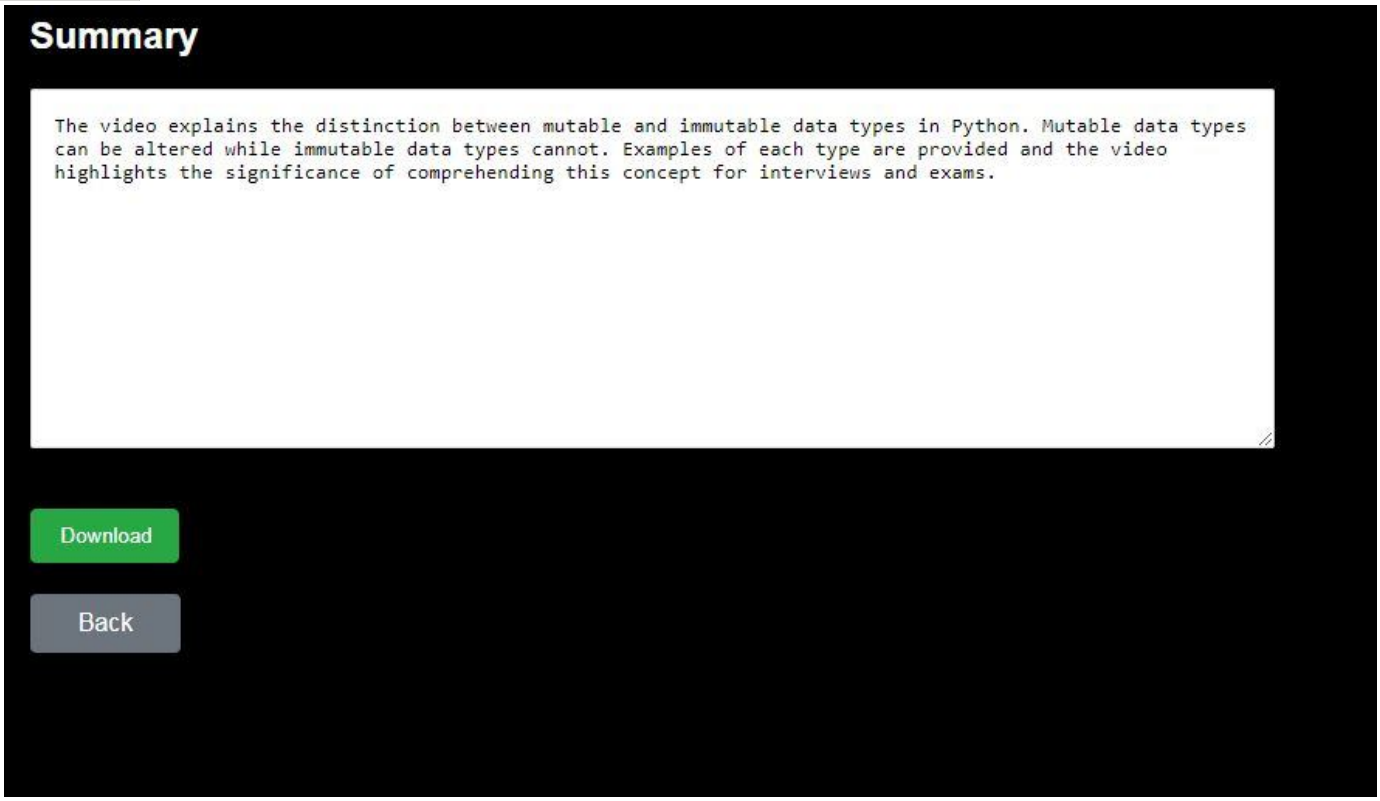


Figure 7 Abstractive summary for sample video

Figure 8 Extractive summary for sample video

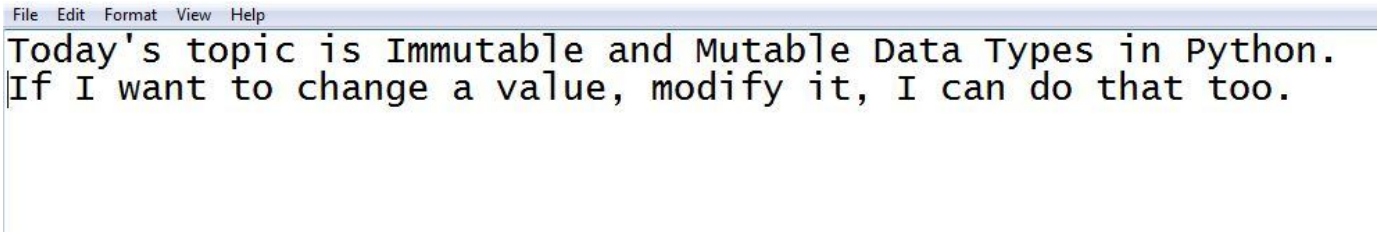


Figure 9 Downloaded file

For uploaded video file which has only music without speech the output summary is shown in Figure 10.

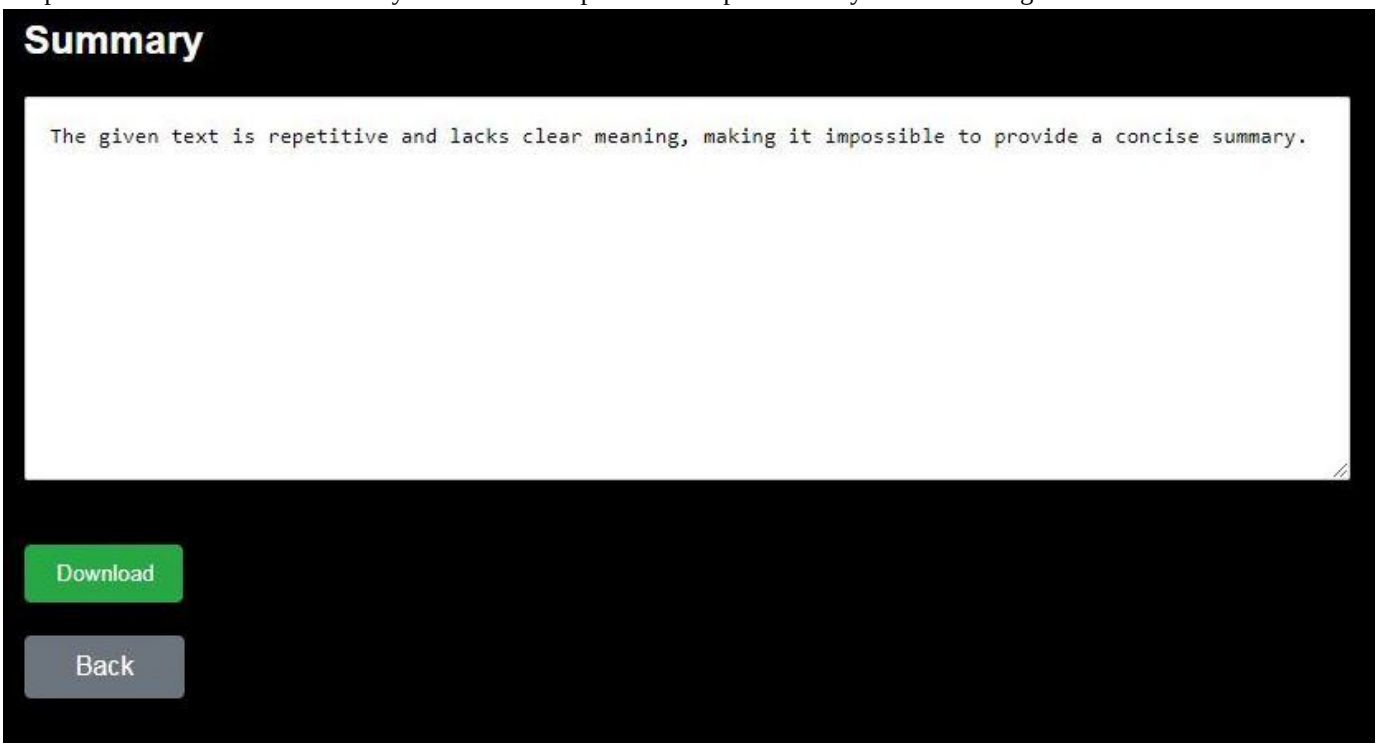


Figure 10 Abstractive summary for sample music video

V. CONCLUSION

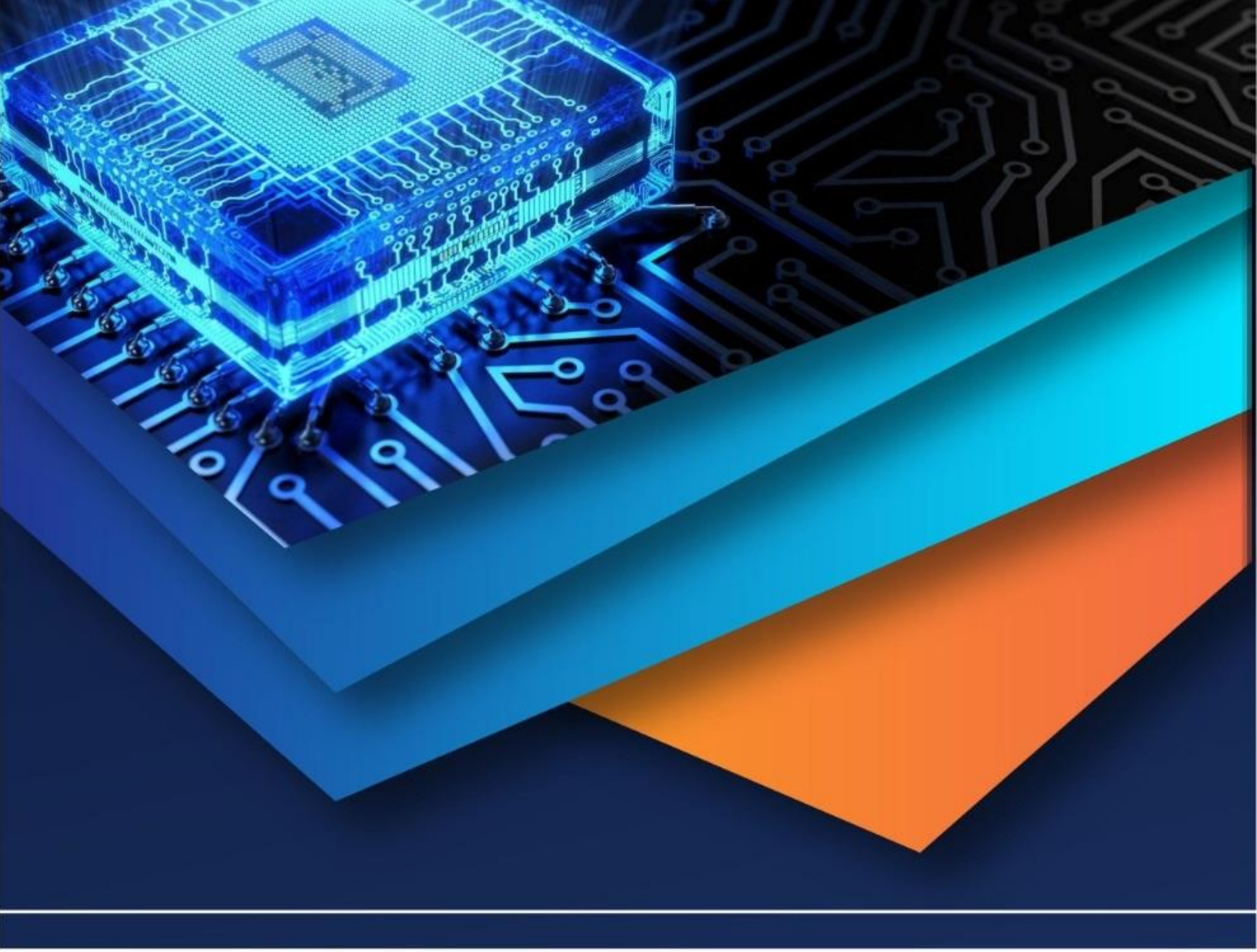
The proposed Video Transcript Summarization system takes input as multi-language video and the user chooses summary type then clicks on the generate button on the web page. Now the system converts the video's speech into the transcripts of that video with the help of openAI whisper ASR and the converted transcripts are then summarized with the GPT model transformer package based on user's summary choice. The summarized text is displayed to the user on the web page. The users of this initiative benefit greatly from the savings of their time and money. This enables us to comprehend the salient points of the video without watching the entire video. Also, it assist user in identifying and avoiding harmful content so that it won't interfere with their viewing experience. In conclusion, the proposed system outperforms the existing system in text processing as evidenced by Rouge metrics. For proposed system, abstractive summary F1 scores are 0.71, 0.58, 0.65 and extractive summary F1 scores are 0.60, 0.44, 0.50 respectively surpass those of the existing system, abstractive summary F1 scores 0.56,0.32,0.39 and extractive summary F1 scores 0.35,0.30,0.25.

VI. FUTURE SCOPE

Providing user customization in the future scope of a video transcript summarization project could involve allowing users to personalize the summarization process according to their preferences or specific needs..This might include options to prioritize certain topics, adjust summary length. Customization enhances user experience by tailoring the summarization output to better meet the user's unique requirements.

REFERENCES

[1] R. Sudhan, D.R. Vedhaviyassh, G. Saranya,"Learning to Summarize YouTube Videos with Transformers: A Multi-Task Approach"(IEEE August 2023)
[2] Tirath Tyagi; Lakshaya Dhari; Yash Nigam; Renuka Nagpal ,"Video Summarization using Speech Recognition and Text Summarization" (IEEE,July 2023)
[3] Ilampiray , Naveen Raju , Thilagavathy , Mohamed Tharik , Madhan Kishore , Nithin A , Infant Raj ,"VIDEO TRANSCRIPT SUMMARIZER"(E3S Web of Conferences 2023)
[4] Siddhartha, Prashu Pandey, Ansh Saxena, Anupam kumar Sharma," Youtube Transcript Summarizer" (IJRES May 2023)
[5] Youtube Transcript Summarizer Using Flask (IJRASET ,April 2023)
[6] Summarization of Video Clips using Subtitles(IEEE,MARCH 2023)
[7] P Nagaraj; V Muneeswaran.; B Rohith; B Sai Vasanth; G Veda Varshith Reddy; A Koushik Teja," Automated Youtube Video Transcription To Summarized Text Using Natural Language Processing" (IEEE May 2023)
[8] Atluri Naga Sai Sri Vybhavi; Laggisetti Valli Saroja; Jahnavi Duvvuru; Jayanag Bayana,"Video Transcript Summarizer", (IEEE April 2022)
[9] Gousiya Begum , N. Musrat Sultana , Dharma Ashritha," YOUTUBE TRANSCRIPT SUMMARIZER" (IJCRT June 2022)
[10] Porwal, Khushi and Srivastava, Harshit and Gupta, Ritik and Pratap Mall, Shivesh and Gupta, Nidhi, Video Transcription and Summarization using NLP (July 14, 2022).
[11] Text Summarization using Transformer Model (IEEE,November 2022)
[12] Krishna Kulkarni; Rushikesh Padaki," Video Based Transcript Summarizer for Online Courses using Natural Language Processing" (IEEE December 2021)
[13] R Sanjana et al., "Video Summarization using NLP", International Research Journal of Engineering and Technology (IRJET), 2021.
[14] https://www.width.ai/post/4-long-text-summarization-methods
[15] https://pypi.org/project/bert-extractive-summarizer
[16] https://dev.to/eteimz/understanding-langchains-recursivecharactertextsplitter



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY