INTELLIGENT VIDEO CONTENT ANALYSIS WITH NLP

Summarization, Search and Segmentation

By

Elife Kocabey 150120054
İrem Kıranmezar 150121013
İrem Aydın 150120013

CSE4197 Engineering Project 1

Project Specification Document

Supervised by:
Dr. Öğr. Üyesi BETÜL BOZ

Marmara University, Faculty of Engineering

Computer Engineering Department

15.11.2024

Contents

1 PROBLEM STATEMENT

Nowadays, academic researchers and students frequently utilize videos in their work; however, quickly accessing important information within these videos can often be challenging. Dividing videos into specific topics, enabling keyword searches within the content, and generating rapid summaries will reduce time loss and improve efficiency. This project has been developed to analyze speech based videos more effectively and to help researchers extract information from videos more quickly.

2 PROBLEM DESCRIPTION AND MOTIVATION

In the age of digital transformation, video content has become an indispensable resource for both researchers and students in their academic pursuits. Audiovisual materials, including recorded lectures, conferences, workshops, interviews, and documentaries, provide dynamic and engaging sources of information that facilitate a richer understanding of complex subjects. As academic resources evolve beyond traditional texts, the volume of video-based content available for research has dramatically increased. However, the accessibility of this wealth of information is often hampered by the challenges of locating specific insights within lengthy video files. Quickly accessing the desired information from videos is especially important in academic research and projects where time is very valuable.

The main motivation for this project is the need for more efficient analysis of speech based video content. The fact that short videos have become very popular, especially in recent times, and therefore people's focus times have decreased, makes it difficult to comprehend long videos efficiently. Hence this project will offer features such as analyzing uploaded videos by topic, searching for specific words to locate relevant scenes, and summarizing videos. This way, users will be able to access video content more quickly and in a more focused manner. In other words, the platform to be developed will save time for researchers and students by accelerating their work and making the information access process more efficient.

3 MAIN GOAL AND OBJECTIVES

The main goal of this project is to develop a video analysis platform for researchers and students to efficiently access information in academic videos. The platform will categorize videos into chapters, support keyword search to locate specific scenes, and provide concise summaries of content. Additionally, a user-friendly interface will enable easy video uploads and quick access to analysis results.

Objective 1: To design and implement a system that categorizes uploaded videos into chapters based on their content.

Objective 2: To enable keyword search functionality that allows users to locate specific scenes where certain words are mentioned within the videos.

Objective 3: To develop an algorithm for summarizing videos, providing users with a concise overview of the content.

Objective 4: To create a user-friendly interface that facilitates easy video uploads and quick access to the analysis results.

4 RELATED WORK

4.1 Sonix: General Features and Limitations

In recent years, the need for video content analysis in media management, education, and research has grown significantly, leading to the development of various platforms to meet this demand. One of the prominent platforms in this field is Sonix [1], which offers users a range of analysis tools, such as video transcription, topic summarization, and content segmentation. However, Sonix has notable limitations when processing long video content. For instance, uploading and processing an 8-minute and 45-second video on Sonix takes approximately 5 minutes and 22 seconds. This processing time causes a substantial delay, particularly for analyzing longer videos, resulting in a considerable time loss. Our project, with a focus on performance over extensive features, aims to reduce this processing time.

Additionally, Sonix's pricing model, based on analysis time, can lead to high costs for users. The platform offers a 30-minute free trial initially; for continued usage, however, it charges around $10 per hour of video analysis. This pricing structure poses a significant financial burden, especially for academic research. In our project, we aim to support researchers by providing analysis services free of charge, thereby increasing accessibility.

Another limitation of Sonix is the low accuracy rate in features such as topic segmentation and speaker recognition. For example, it has a limited success rate in identifying and distinguishing between different speakers. In this context, our project aims to achieve a minimum accuracy rate of 65% for functions like topic segmentation, summarization, and speaker recognition. These improvements intend to provide more reliable and effective results in video analysis, ultimately enhancing user experience.

4.2 Other Video Summarization Applications

Apart from Sonix, there are various other video summarization and analysis platforms available in the market, such as summarize.tech [2], Monica [3] and ScreenApp [4]. However, most of these platforms primarily support video uploads via URL and offer limited functionality. Many of these applications focus solely on basic speech transcription and superficial summarization, lacking comprehensive analysis features such as detailed content segmentation and keyword search. These limited features make it challenging to analyze long video content and negatively affect user experience.

Moreover, while many platforms offer a free trial initially, they tend to impose high subscription fees after the trial period. This pricing model restricts sustainability, particularly for academic and long-term use cases, making it difficult for users needing lengthy analysis processes.

Our project aims to overcome these limitations by providing users with an extensive solution for long video analysis without cost concerns. In addition to offering a free service, we aim to be a robust analysis tool with advanced functionality, including topic segmentation, keyword search, and customizable summarization. This approach will allow users to analyze the overall context of video content beyond just specific keywords, creating a more appealing and accessible alternative for users requiring prolonged analysis.

4.3 Algorithms and Methods

The study [5], which combines the Whisper model planned to be used for the Speech-to-Text phase of the project to be developed with a web application, focused on the system performance of the application and the comparison of the Transcription Accuracy rates of different Whisper models. The performance of this web application developed using the Streamlit package was measured in different browsers with Google Lighthouse, a speed measurement tool, and according to the data obtained from the experimental results, each browser took less than 10 seconds to load the content completely. In the continuation of the study, measurements were made in order to evaluate the performance of the Whisper model and to observe its potential use in web applications. In these measurements, it was aimed to obtain an acceptable accuracy rate by preserving the transcription speed using WER% (Word Error Rate), which is used to measure the accuracy of ASR systems. As a result of the experiment conducted on three different audio files, it was seen that the small.en model achieved a WER% value below 10%, but higher than the medium.en model, and worked twice as fast as the medium.en model. For this reason, the small.en model was preferred in the web application. In the project, we aim to work with the whisper-large-v3 model, which is not evaluated in this study and uses more parameters than the medium.en model, and to obtain results at maximum speed while keeping the accuracy rate high.

The study [6], which uses the review evaluations of a skin care product as a dataset and provides informative summaries of these reviews for consumers, compared three different models for the summary process. These models are BART, BERT and T5 models. Although all three models can be used for content summary with their own approaches, the study compared the performances of the models using the ROUGE Score evaluation method, which measures the similarities between the model summary and the human summary, and preferred the model that gave the best result to be used in the project. According to the evaluation results, the BART model achieved a higher ROUGE score than the BERT and T5 models by receiving an average score of 0.87718 in ROUGE-1, 0.80689 in ROUGE-2 and 0.87688 in

ROUGE-L. Therefore, it was deemed appropriate to use the BART model in order to provide more useful product information to consumers.

Another study on the BART model to be used in the Summarization part of our project [7] fine-tuned the model and retrained the model to summarize medical text. The aim of the study is to compare the evaluation results of the trained model with other medical text summarization approaches in the literature. The obtained results were evaluated with the ROUGE metric. With the results of 0.7216 in ROUGE-1, 0.5757 in ROUGE-2, and 0.7075 in ROUGE-L, it was seen that this trained version of the BART model generally exhibited higher performance than other similar studies in the literature. As a result of the study, it was revealed that fine-tuning the BART model according to the desired solution had positive effects.

As a result of the information obtained from the two articles [6] [7] and the researches, it was planned to use the BART model in the project to be developed, as it will achieve a higher accuracy rate compared to other models and will be able to achieve good results if the model is later fine-tuned with academic education contents.

The paper[8] presents a topic modeling approach that combines BERT and LDA models to uncover latent themes in financial news. It highlights the shortcomings of traditional LDA-based methods in handling short and context-independent texts, leading to loss of contextual meaning. To address this, a model is developed that integrates BERT's powerful semantic understanding with LDA's thematic coverage, taking both context and topic structure into account.

In the study, UMAP was used for dimensionality reduction, HDBSCAN for clustering, and the c-TF-IDF method for topic representation. Experimental results show that the BERT-LDA model produces consistent and meaningful topics, outperforming traditional methods. This model demonstrates superior performance, especially with short and context-dependent data. Based on the results obtained in this paper, we will use these two models in the most efficient way for our project.

5 SCOPE

5.1 Included in The Project Scope

The project aims to develop a video analysis website. It includes automatically detecting and displaying scenes where a specific word appears when searched, generating a summarized transcript of the video, and dividing the video into sections based on distinct topic headings. The project supports videos that are one hour or shorter and is designed specifically for videos containing speech, as it focuses on speech and audio analysis. Users can only perform operations on videos they have uploaded to the system.

5.2 Excluded from The Project Scope

The project will focus solely on video analysis and will not include editing, trimming, or assembling videos. User experience will be the primary priority, and performance optimizations for heavy user loads, such as thousands of users accessing the system simultaneously, are outside the project scope. Additionally, the project will support only videos that are one hour or shorter; videos longer than one hour are not included within the scope. As it focuses on speech and audio analysis, the project is unsuitable for silent films or videos containing only visual content

5.3 Constraints

- **Hardware Constraints:** The performance of the project depends on server resources and the processing power of the user devices. Processing long videos will significantly utilize system resources.
- **Software Constraints:** The project will rely on existing video processing and natural language processing libraries (e.g., NLP tools). Additionally, the project's functions will be shaped by the features and limitations of the libraries used.
- **Data Constraints:** The quality and accuracy of the datasets used in video analysis will directly affect the accuracy of the results. The application's performance will depend on the breadth and diversity of the training data used.

5.4 Assumptions

- The project assumes that users will upload videos of reasonable length. Processing and analyzing videos longer than one hour will not be within the scope of this project.

- It is assumed that video files can be uploaded in certain formats (e.g., MP4, AVI) during the project.

- It is assumed that users will possess the necessary technical knowledge to use the platform.

- The infrastructure of the system is assumed to provide sufficient bandwidth for users with average internet connections. Additional optimizations may be required to maintain system performance under heavy user load.

- The reliability of third-party video and language processing libraries used in the project is assumed. Compatibility issues may arise with future updates of these libraries.

6 METHODOLOGY AND TECHNICAL APPROACH

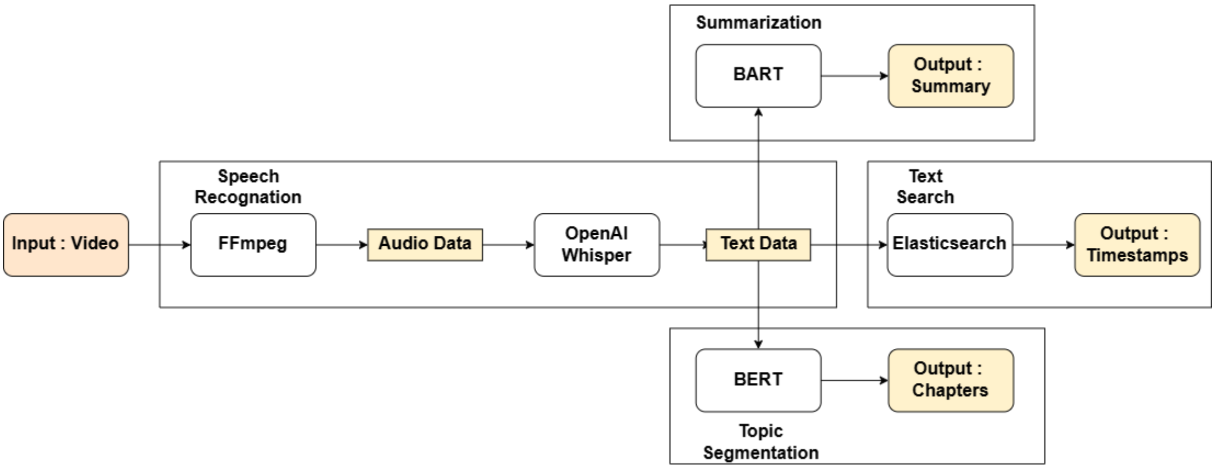


Figure 1 : Project Flow Chart

The system design of the project planned to be developed consists of four main sections, as shown in Figure 1. These sections, created to obtain the desired outputs from the video input to be received from the User Interface, are listed in detail below as Speech Recognition, Text Search , Summarization, and Topic Segmentation , with User Interface.

6.1 Speech Recognition

The first and most important step for the word search, video summarization and topic segmentation stages of the project is to transcribe the speech in the video with as little error rate as possible. In the Speech-To-Text and tagging each text output with timestamps stages of the project, it is planned to use the open source Whisper ASR (Automatic Speech Recognition) model developed by OpenAI, which converts the audio file to text using deep learning techniques. FFmpeg (Fast Forward MPEG), an open source command line tool, will be used to separate audio data from videos and create input for the Whisper model.

Whisper, first proposed in 2022 [9], was developed using the Transformer architecture [10], as in OpenAI's GPT models. The self-attention mechanism offered by Transformer allows Whisper to discover meaningful patterns in speech data. Whisper takes advantage of Transformer’s attention mechanism to take into account the context of each word when converting audio data to text and determines how much importance should be given to each part of the audio content. This provides the ability to cope with challenges such as different accents, word variations or noisy environments.

The attention mechanism at the core of Whisper performs calculations in parallel by assigning weights to each element of the input data (word, syllable or vocal region) in context. In this way, much faster and more efficient results are obtained compared to sequential models such as traditional RNN or LSTM. Whisper takes audio data in the form of sound waves and converts this wave into the Transformer model’s attention-based calculations, ensuring that audio conversations are transcribed correctly.

Whisper will be preferred as the Speech To Text model in this project due to its high accuracy rate, multi-language support and the possibility of fine tuning since it is open source. In this project, it is planned to use the whisper-large-v3 version of the Whisper model family, which has 7.44 Word Error Rate Percentage (WER%), or 92.56% accuracy, according to OpenAI Speech Recognition Leaderboard [11] data. Python's PyTorch [12] library will also be included in the project for the integration of the Whisper model.

6.2 Text Search

The keyword search function enables users to access specific segments within video content quickly and efficiently. Since video content often consists of large and complex datasets, a robust infrastructure is required for fast and accurate searches. To achieve this in the project, Elasticsearch will be used to perform keyword searches within video transcripts [13]. Elasticsearch is an open-source, distributed search engine capable of processing large text datasets effectively. By indexing video transcripts, this system allows users to perform fast and accurate keyword-based queries within the text.

In the initial step, the transcripts generated by Whisper are then transferred to Elasticsearch after being processed into meaningful keywords. This process involves removing unnecessary words and performing tokenization, thereby enhancing the search accuracy and enabling a fast and precise keyword search within the video content.

Subsequently, the transcripts are divided into specific time segments, each labeled with start and end timestamps. This structure allows users to be directed to the particular video section where the searched keyword appears. Elasticsearch indexes the terms within each segment, enabling users to quickly access video segments containing specific keywords. When performing a search on video transcripts, Elasticsearch uses an algorithm called BM25 to rank results. This algorithm evaluates the significance of each word within the video transcript, presenting the most relevant results. This ranking determines the relationship between the searched keyword and the video text, providing users with the most relevant results.

The search results are presented in a user-friendly interface, displaying each result with timestamps and text excerpts. This allows users to quickly navigate to the video segment where the keyword appears and better understand its context. The accurate transcripts provided by Whisper, combined with the fast and precise search capability of Elasticsearch, allow for easy access to information within video content.

6.3 Summarization

The summarization phase is crucial for distilling key information from each video segment, enabling users to access shorter and more comprehensible content. For this purpose, we chose BART (Bidirectional and Auto-Regressive Transformers), as it produces more consistent and higher-quality summaries for complex and lengthy texts compared to other models [6]. BART's bidirectional and auto-regressive features allow it to fully understand the context of the text and logically select the most relevant information.

We will use Hugging Face's Transformers Library to implement BART [14]. Hugging Face provides a robust suite of tools and models that simplify working with language models, offering utilities essential for fine-tuning BART on our project-specific data. For training and model optimization, we rely on PyTorch (Torch) [12]. PyTorch provides a powerful framework for building and training deep learning models, particularly beneficial for handling large datasets in this project.

To prepare text data for summarization, we use BARTTokenizer, a specialized tool designed to work seamlessly with the BART model. BARTTokenizer processes text by breaking it down into meaningful tokens, removing irrelevant parts, and structuring the input according to the model's requirements. This includes splitting text into tokens, managing text length through truncation or padding, and encoding it into a format the model understands, optimizing the data for summarization and ensuring accurate, efficient results.

To enhance this process, SentencePiece is integrated within BARTTokenizer. SentencePiece performs subword tokenization, which breaks words into smaller subword units, allowing the model to handle rare or unknown words more effectively. This subword tokenization is crucial for preserving context, especially in complex or lengthy texts. Together, BARTTokenizer and SentencePiece improve the accuracy and efficiency of text analysis during the summarization phase, making them essential tools in this project.

6.4 Topic Segmentation

The primary objective is to analyze each video transcript and divide the content into meaningful thematic segments. This approach allows users to identify main topics and access relevant information more efficiently, making it easier to explore video content. To achieve this segmentation, we will use a hybrid method that combines both BERT (Bidirectional Encoder Representations from Transformers) and Latent Dirichlet Allocation (LDA) models.

BERT captures contextual information effectively with its bidirectional attention mechanism, allowing for deep analysis of relationships between words and phrases within sentences. This helps us interpret the complex structure of video transcripts.

LDA, on the other hand, identifies hidden topics by clustering frequently co-occurring words. While BERT provides in-depth sentence-level analysis, LDA detects broader patterns. Together, LDA's thematic analysis enhances BERT's contextual understanding, allowing for better identification of main topics in the content [8].

To integrate BERT into our project, we will use the Hugging Face Transformers Library. Hugging Face provides tools for training, fine-tuning, and applying BERT, which will help us tailor the model to the specific needs of our project, improving performance and efficiency. To enhance topic modeling performance, we will use Sentence Transformers, as these embeddings refined by BERT capture semantic relationships more accurately, improving LDA's ability to differentiate meaningful topics.

For improved clustering, we will integrate HDBSCAN and K-Means. HDBSCAN, a density-based clustering algorithm, will automatically determine the optimal number of clusters, ensuring that contextually similar sentences are grouped together. In contrast, K-Means will allow us to test for more defined, fixed clusters when needed, offering flexibility in organizing the data [8].

In the preprocessing stage, we will use NLTK or SpaCy to handle tasks such as tokenization, stop-word removal, and lemmatization. These steps will clean and streamline the data, ensuring that the input for both BERT and LDA is well-structured and relevant, thus improving the overall performance of the topic segmentation process.

6.5 User Interface

All video analysis processes to be implemented will be presented to users via a website. ReactJs, a JavaScript library, and HTML - CSS will be used in the development of this web-based project. In addition, Bootstrap framework will be used for user-friendly designs. The system will be tested for performance using metrics such as processing time and responsiveness to ensure that the user experience is optimized for large video files. The combination of these elements will provide an effective and user-friendly platform for video analysis.

7 PROFESSIONAL CONSIDERATIONS

7.1 Methodological Considerations / Engineering Standards

Table 1 : Basic Tools Used In The Project

Tool	Explanation
	Git: Version Control and Source Code Management
	Github: To Store Source Codes in a Remote Repository
python	Python Programming Language: To Implement ML Models and Algorithms
PyTorch	PyTorch: Machine Learning Library for Python
🤗	HuggingFace: To NLP and Speech Processing Models and Datasets
React	ReactJs: JavaScript Library for User Interface
	Elasticsearch: A Search Engine Based on Apache Lucene for Text Search
	Draw.io: To Sketching tables, Figures and Diagrams

Table 1 Continued

IEEE Xplore	IEEE Xplore : To Literature Survey
zoom	Zoom : To Communication and Project Management

Table 1 summarizes the tools to be used throughout the project.

7.2 Realistic Constraints

7.2.1 Economical

The memberships of similar projects in the market are calculated in terms of the dollar exchange rate as quite expensive. The fact that our project is local and national will provide budget-friendly use for users. The cost of our project is primarily limited to a server and storage space with sufficient performance and capacity since it operates on digital media. However, additional costs such as hardware or software licenses may arise when the project is implemented in real life.

7.2.2 Environmental

Our project does not pose any environmental issues since it only operates on digital media. However, if the project is implemented, there may be indirect environmental impacts, such as energy consumption by the server.

7.2.3 Ethical

Our project carries high ethical responsibilities in terms of privacy and data security. Strict security measures are taken to protect user data, explicit consent is obtained before each analysis, and video data is not stored in the system after analysis. By respecting copyrights, no changes are made to the original content during the analysis, and users' content ownership is protected.

7.2.4 Health and Safety

Since the project offers a software-based solution, it does not have a direct impact on health and safety. However, techniques such as data encryption and user authentication will be used to ensure application security.

7.2.5 Sustainability

The system will be designed to be sustainable and scalable. The server infrastructure will be expandable when necessary, considering the expansion of the user base and increasing data processing demands.

7.2.6 Social

Our project has been carefully designed to avoid causing social discrimination when analyzing video content. Equal opportunities will be provided regarding users' access rights and data usage. The project aims to offer a beneficial and inclusive experience for all users.

7.3 Legal Considerations

The legal aspects to be considered during the project development process are as follows:

Licenses: The software and libraries used in the project will be compliant with open-source licenses.

Data Privacy and Security: Since the project will conduct analyses on videos uploaded by users, full compliance with data protection laws such as KVKK (Personal Data Protection Law) will be ensured. Users' videos will be kept confidential, and no video or analysis results will be shared externally.

Ethical Permissions: In the event that the project collects or analyzes user data, users' consent will be obtained. In this context, ethical permissions and user agreements will be established.

8 MANAGEMENT PLAN

8.1 Description of Task Phases

Phase 1 - Determination of Project Topics and Problem Definition: Identification of the fundamental issues related to the selected topics and determination of possible strategies to achieve the final goal. Additionally, clarification of the project scope and identification of necessary resources.

Phase 2 - Review of Existing Literature: Conducting a comprehensive review of existing studies on video analysis and natural language processing.

Phase 3 - Preparation of Project Documents: Gathering necessary information related to the project and creating a Project Specification Document (PSD).

Phase 4 - Research on Video Analysis Methods: Conducting research on various video analysis techniques and identifying applicable methods. Evaluating the advantages and disadvantages of the selected methods.

Phase 5 - Data Collection and Preprocessing: Collecting video data to be used in the project, performing data cleaning, and executing preprocessing steps. Making the data suitable for analysis.

Phase 6 - Consolidation of Research and Preparation of Analysis Report: Combining tests and research to create an Analysis and Design Document (ADD). Preparing for project presentation.

Phase 7 - Algorithm Development and Optimization: Developing the selected algorithms, optimizing their performance, and detailing the application processes. Testing the developed algorithms in different scenarios.

Phase 8 - Model Validation and Test Scenarios: Creating scenarios to test the accuracy of the developed model and applying them. Evaluating how the model performs on real data.

Phase 9 - User Interface Design and Coding: Conducting design work to ensure the website has a user-friendly interface. After the design phase, developing the interface and performing the necessary coding to facilitate user interaction.

Phase 10 - Evaluation of Project Results: Analyzing the results after the application is completed and evaluating the findings obtained. Documenting the project results and preparing for the presentation.

8.2 Division of Responsibilities

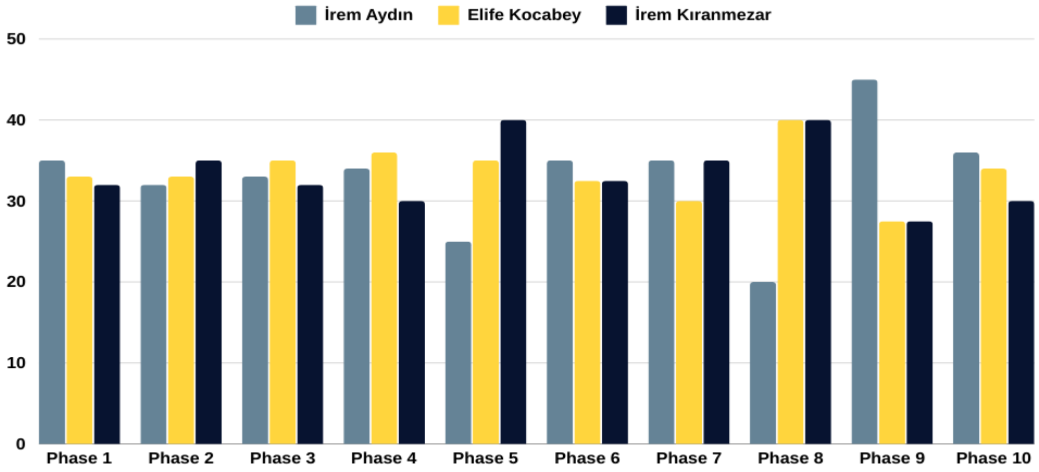
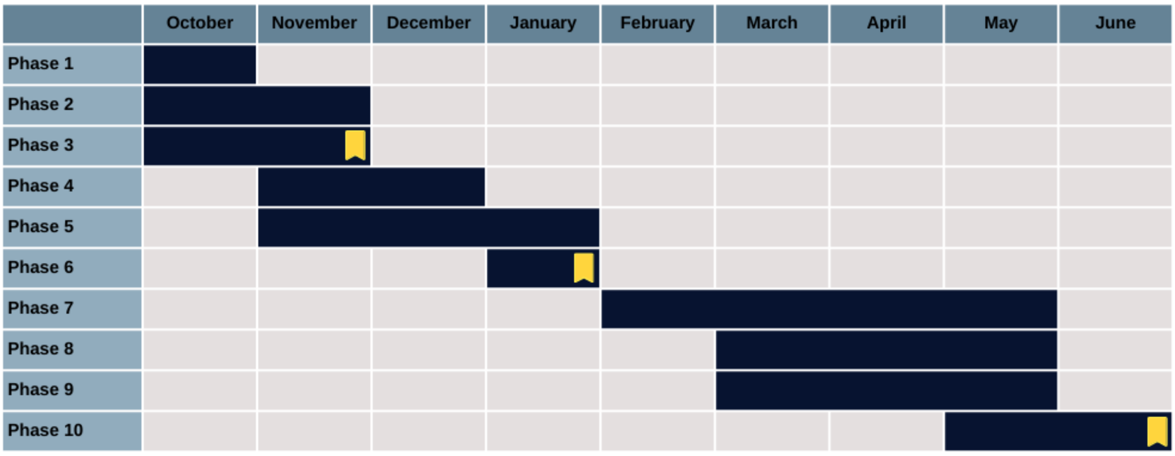


Figure 2 : Division of Responsibilities Among Team Members

Division of responsibilities among team members is given in Figure 2.

8.3 Time Line with Milestones



	October	November	December	January	February	March	April	May	June
Phase 1	■								
Phase 2	■	■							
Phase 3	■	■ 🔖							
Phase 4		■	■						
Phase 5		■	■	■					
Phase 6				■ 🔖					
Phase 7					■	■	■	■	
Phase 8						■	■	■	
Phase 9						■	■	■	
Phase 10								■	■ 🔖

Figure 3 : Timeline with Milestones

Timeline with milestones is shown in Figure 3. There are 3 milestones which refer to:

1) The project topic is determined and the PSD document is completed.

2) Research is completed, the ADD document and presentation are prepared.

3) The software is completed, tests are conducted, results are evaluated with the final report and presentation prepared.

9 SUCCESS FACTORS AND RISK MANAGEMENT

9.1 Measurability / Measuring Success

- **Objective 1:** To design and implement a system that categorizes uploaded videos into chapters based on their content.

  Success Factor: The system is expected to categorize videos with at least 85% accuracy using appropriate titles. This performance will be evaluated using the "F1 Score" given in Equation 3.

- **Objective 2:** To enable keyword search functionality that allows users to locate specific scenes where certain words are mentioned within the videos.

  Success Factor: The keyword search function is expected to find the correct scenes with at least 90% accuracy. Success will be measured by user satisfaction with the search results and the system's ability to locate the scene related to the keyword. This performance will be evaluated using the "F1 Score".

- **Objective 3:** To develop an algorithm for summarizing videos, providing users with a concise overview of the content.

  Success Factor: The developed summarization algorithm is expected to accurately summarize the content of the original video. Success will be measured using ROUGE metrics such as ROUGE-1, ROUGE-2, and ROUGE-L. ROUGE-1 measures the overlap of individual words (unigrams) between the summary and the reference text, while ROUGE-2 evaluates sequential word pairs (bigrams). ROUGE-L calculates the longest common word sequence (Longest Common Subsequence) between the summary and the reference, assessing structural similarity. ROUGE metrics will be calculated using Equations 1, 2, and 3. Success will be achieved if accuracy in these metrics reaches 70% or higher.

- **Objective 4:** To create a user-friendly interface that facilitates easy video uploads and quick access to the analysis results.

Success Factor: User satisfaction will be measured through surveys or evaluation forms regarding the interface. The aim is to receive at least 75% "satisfied" or "very satisfied" feedback from users.

Precision measures the extent to which words in the system summary match those in the reference summaries. The Precision value is calculated using Equation 1.

$$Precision = \frac{TP\ (number\ of\ similar\ words)}{TP + FP\ (total\ words\ in\ system\ summary)}\ (1)$$

Recall measures the extent to which words in the reference summary are included in the system summary. The Recall value is calculated using Equation 2.

$$Recall = \frac{TP\ (number\ of\ similar\ words)}{TP + FN\ (total\ words\ in\ reference\ summary)}\ (2)$$

The F1-score combines precision and recall to assess the balance between the quality of the system-generated summary and its ability to capture relevant information from the reference summaries. The formula for calculating the F1-score is shown in Equation 3.

$$F1\ Score = \frac{2 * Precision * Recall}{Precision + Recall}\ (3)$$

9.2 Risk Management

- If sufficient hardware power cannot be provided and sufficient performance speed cannot be achieved during the speech recognition phase of the project, whisper-medium.en or Mozilla Deepspeech, which have less accuracy but require less processing power, will be used instead of whisper-large-v3 as the ASR model.

- If Elasticsearch fails to provide the desired speed with large datasets or encounters consistency issues in search results, alternative search infrastructures, such as Apache Solr, will be considered.

- If the BART model fails to provide the desired speed and accuracy in summarizing long videos, different models such as T5 will be considered.
- If BERT and LDA models prove insufficient in identifying key topic sections within video content or require longer processing times than desired, alternatives such as using only BERT or other topic modeling algorithms, such as BERTopic, will be considered.

10 BENEFITS AND IMPACT OF THE PROJECT

The project will provide fast and efficient analysis capabilities for academics, researchers, and students working with video content. By enabling quick access to desired information through features like summarization and keyword search, it will save time in academic research without the need to examine long videos in detail.

Scientific Impact: The project aims to provide an innovative approach in the fields of video analysis and natural language processing, inspiring other researchers and contributing to the scientific literature.

Economic/Commercial/Social Impact: The project will serve as a useful tool for students and researchers conducting academic studies, accelerating video analysis processes and reducing costs. This can lead to more effective use of natural resources and minimize environmental impacts.

Potential Impact on New Projects: Our project could serve as a reference source for new studies in the fields of video analysis and natural language processing.

Impact on National Security: Our project is not directly related to national security; however, an effective video analysis system has the potential to contribute indirectly in areas such as emergency management.

*** This document was prepared by the team members, with ChatGPT utilized to enhance the quality of the English language used in the sentences.

REFERENCES

[1] Sonix.ai: https://sonix.ai/ Date accessed: 13.11.2024.

[2] summarize.tech: https://www.summarize.tech Date accessed: 13.11.2024.

[3] monica.im: https://monica.im/features/youtube-summary-with-chatgpt Date accessed: 13.11.2024.

[4] screenapp.io: https://screenapp.io/features/ai-summarizer Date accessed: 13.11.2024

[5] A. L. Haz, E. D. Fajrianti, N. Funabiki & S. Sukaridhoto. 2023 "A Study of Audio-to-Text Conversion Software Using Whispers Model," Sixth ICVEE, Surabaya, Indonesia, 268-273.

[6] Maghfiroh, N. A., Bachtiar, A. B., & Lailil, M. 2023. "Comparative analysis of summarization methods for skin care product reviews: A study on BERT, BART, and T5 models." In Proceedings of the 2023 International Conference on Advanced Mechatronics, Intelligent Manufacture and Industrial Automation, 593-598, IEEE.

[7] Altundogan, T. G., et al. 2023. "BART fine-tuning based abstractive summarization of patients' medical questions texts." In Proceedings of the 2023 4th ICDABI, 174-178, IEEE.

[8] Zhou, M., Kong, Y., & Lin, J. 2022. "Financial topic modeling based on the BERT-LDA embedding. " Proceedings of the 20th IEEE International Conference on Industrial Informatics (INDIN 2022), 495-500.

[9] Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. 2022. “Robust Speech Recognition via Large-Scale Weak Supervision”, arXiv.

[10] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, et al. 2017, "Attention is all you need", Advances in neural information processing systems, vol. 30.

[11] HuggingFace, Open ASR Speech Recognition Models LeaderBoard**:** https://huggingface.co/spaces/hf-audio/open_asr_leaderboard Date accessed: 13.11.2024

[12] Pytorch Library: https://pytorch.org/ Date accessed: 13.11.2024.

[13] Elasticsearch: https://www.elastic.co/elasticsearch Date accessed: 13.11.2024.

[14] HuggingFace, Models: https://huggingface.co/models Date accessed: 13.11.2024.