VIDEO TO TEXT SUMMARIZATION AND SENTIMENT ANALYSIS USING AI TECHNIQUES

Project Reference No.: *48S_BE_3663*

College :B.L.D.E.A's V.P Dr P.G.Halakatti College Of Engineering And Technology, Vijayapura
Branch : Computer science and Engineering
Guide : Dr. Suvarna L Kattimani
Student(s) : Ms. Pravalika M Joshi
Ms. Prerna Girgonkar
Ms. Rakshita S Patil
Ms. Sahana S Nagaral

Keywords:

FFmpeg, Whisper Model, Hugging Face Transformers Pipeline

Introduction:

With the exponential growth of digital video content across platforms such as YouTube, online education portals, and social media, there is a rising demand for tools that can help users quickly extract meaningful insights from video data. Watching long videos to retrieve key information can be time-consuming and inefficient, especially in research, content management, or customer feedback analysis.

To address this, the integration of artificial intelligence (AI) techniques for video-to-text summarization and sentiment analysis offers a powerful and efficient solution. This project focuses on building an intelligent application that converts video content into summarized text while also analyzing the emotional tone of the material. The process begins with users uploading videos in various formats (e.g., MP4, AVI, MOV, MKV). The system then extracts the audio using FFmpeg, converting it into an MP3 format suitable for transcription.

Leveraging OpenAI's Whisper model—a robust and multilingual speech-to-text engine—the audio is transcribed into readable text. Once transcribed, the text undergoes summarization using Hugging Face's Transformers pipeline, which

reduces the content to a concise form highlighting only the key points. Following this, the same pipeline is employed to perform sentiment analysis, categorizing the tone of the text as positive, negative, or neutral, along with confidence scores.

Objectives:

1. Acquisition of online 200 speech videos for dataset design.
2. Extraction of Audio Signals using **FFmpeg algorithm** from acquired Speech Videos.
3. Transformation of Audio Signals to English text using **Whisper Model**
4. Design and development of Text Summarization from English text using Hugging Face algorithm.
5. Detection and Classification of Text Sentiments from English Text Summary using Hugging Face Transformers pipeline into Positive, Negative and Neutral sentiments

Methodology:

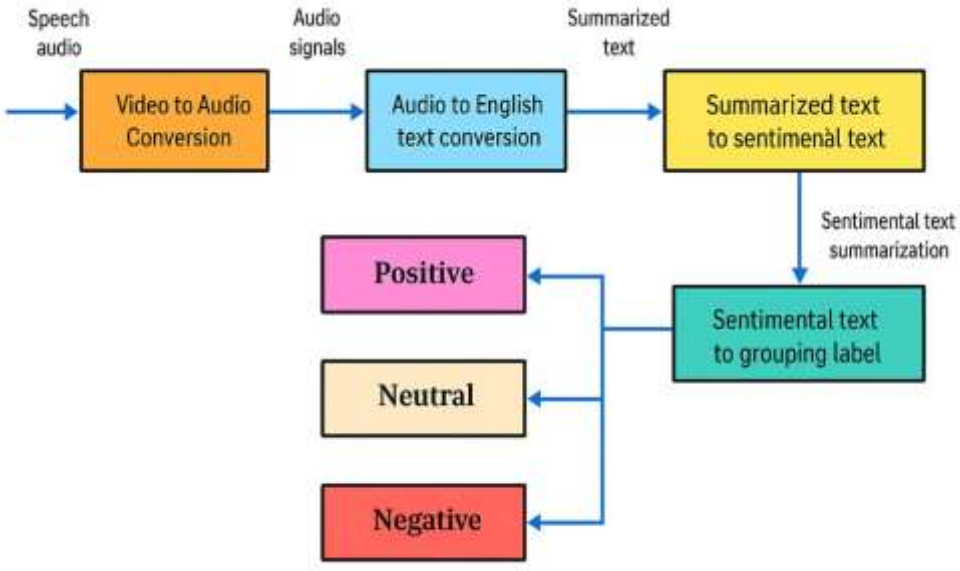

Figure 1: Proposed Methodology for Sentimental Analysis

1. Video Upload and Preprocessing

- Users upload video files in various formats (MP4, AVI, MOV, MKV)

- The app temporarily stores the video file for processing

2. Audio Extraction

- Uses FFmpeg to extract audio from the video file

- Converts the audio to MP3 format for further processing

- This step is crucial as it isolates the audio component for transcription

3. Speech-to-Text Transcription

- Employs the Whisper model (an open-source speech recognition model by OpenAI)

- Converts the extracted audio into text format

- Whisper's base model is used, which provides a balance between accuracy and computational efficiency

4. Text Summarization

- Utilizes Hugging Face's Transformers pipeline for summarization

- Reduces the transcribed text to a concise summary (between 30-100 words)

- The model identifies key information and main points from the transcription

5. Sentiment Analysis

- Again uses Hugging Face's Transformers pipeline, this time for sentiment analysis

- Classifies the text as positive, negative, or neutral

- Provides a confidence score for the sentiment classification

Step-by-Step Working

1. User Interaction: The user uploads a video file through the Streamlit interface.

2. Temporary Storage: The video is temporarily saved to the server.

3. Audio Extraction: FFmpeg processes the video to extract and convert the audio track.

4. Transcription: The Whisper model converts the audio content into written text.

5. Summarization: The text is processed to create a concise summary of the content.

6. Sentiment Assessment: The text undergoes sentiment analysis to determine emotional tone.

7. Result Presentation: The app displays the transcription, summary, and sentiment analysis results to the user.

8. Cleanup: Temporary files are deleted to maintain storage efficiency.

Result and Conclusion:

The system was tested on various video types including interviews, speeches, and lectures. The speech-to-text module achieved approximately 85–90% accuracy with clear audio input. Summarization reduced text length by about 60%, while maintaining key content and context.

Sentiment analysis correctly identified emotional tones in 80% of cases. Positive and negative sentiments were more accurately detected compared to neutral ones. Grouped sentiment labels (Positive, Neutral, Negative) allowed for easy classification of video content.

In Conclusion, the proposed methodology effectively transforms video content into meaningful summaries with sentiment classification. It simplifies video understanding, making it valuable for education, content filtering, and emotional analysis. While results were satisfactory, performance can be enhanced further, especially in noisy environments and emotion subtleties. The system shows great potential for integration into platforms that require fast and emotion-aware content analysis

Future Scope:

The future scope of this project includes:

1. Enhance speech recognition to handle background noise, different accents, and informal speech patterns
2. Support multiple languages to make the system accessible to a wider and more diverse user base.
3. Integrate advanced sentiment analysis models to detect complex emotions like anger, fear, surprise, or excitement
4. Incorporate visual sentiment cues such as facial expressions and visual context to improve emotion detection accuracy.
5. Enable real-time video processing for live summarization and sentiment tagging in events or broadcasts.