© APR 2026 | IRE Journals | Volume 9 Issue 10 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I10-1716256

Transplant Track: Video Content Summarization with Whisper and GPT Transformers

KAMSANI ROHITH REDDY[1], VASAM AKHILESH[2], KUNDANAPALLY RAKESH[3], HALAVATH BALAJI[4]

[1]*Dept of CSE, UG Student, Sreenidhi Institute of Science and Technology, Hyderabad.*
[4]*Asst.Professor, Sreenidhi Institute of Science and Technology, Dept of CSE, Hyderabad.*

Abstract- ***The scalding development of the procedure of creation of digital material has led to bursting of the demand of the short form of the video contents of the reels, shorts and clips that may keep the audience entertained during a finite duration. The long form videos may however be time and labour-intensive since the content creators will have to choose interesting highlights manually. The paper introduces a Video-to-Reel Conversion Platform which is an intelligent platform to convert long videos to small and entertaining short reels. It is referred to as a speech recognizer, sentiment analysis, multimedia and artificial intelligence that is able to extract a valuable content of video information of a long length. The interactive interface of the site may be created with the assistance of the Streamlit and allows individuals to visit the site with the help of offering videos or links on YouTube to be processed. The videos uploaded are stored in a PostgreSQL database such that they are safe and easily accessed. FFmpeg is used to remove the audio track and Whisper, a speech recognition model that uses the approaches of the deep-learning model, is used to produce relevant text transcripts. The analysis is further done with the help of GPT and TextBlob to find out the intensity of emotion, relatability of the subject and interest of the reader. Highlight selection algorithm ranks most interesting contents of the videos according to the following characteristics of transcription sentiment polarity, speech pacing and keyword density which rank the videos. The selected parts will be automatically cut and submitted to FFmpeg and MoviePy to create quality reels in small format. The last reels are made available to the users via the platform by the option of downloading or sharing via the digital platforms. Using the analysis of the experiment, we may observe that the proposed system can contribute to the significant drop in the number of manual laborers to be employed to create the short video highlights and be rather topical and interesting. The automated pipeline is offering a valuable and scalable solution to content creators, marketers, and media employees who would prefer to use long-form video content to produce entertaining short-form media.***

Index Terms- ***Video Highlight Detection, Automated Reel Generation, Speech Recognition, Whisper Model, Sentiment Analysis, FFmpeg, Moviepy, Artificial intelligence, Streamlit, Video content processing.***

I. INTRODUCTION

The emergence of the short-form video platforms, such as Instagram Reels, YouTube Shorts, and Tik Tok, has radically altered the way digital content is viewed. Short videos are very engaging, viral and they work in attracting the attention of the user within a short duration of time. To this, the content creators and digital advertisement agencies are all converging on the shorter versions of the longer videos that include the podcasts, lectures and interviews to the already existing webinars.

Such brief clips are also a typical method of preparation that is tiresome and time-consuming. The manufacturers have to view lengthy videos, choose certain interesting places, cut and share them. When there is a lot of content, it is a tedious undertaking. This has resulted in the high demand of intelligent systems which are able to automatically identify and extract meaningful highlights in long form videos.

The latest superiority in the Artificial Intelligence (AI), Natural Language Processing (NLP), and speech recognition technology has made the automated multimedia analysis possible. The speech-to-text applications, such as whisper, could enable a person to transcribe the speech to the most precise accuracy and the sentiment analysis and topic identification could be used to find the emotional parts of the speech. These technologies might be utilized through automated video editing devices and,

© APR 2026 | IRE Journals | Volume 9 Issue 10 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I10-1716256

hence, one might create great short films automatically.

Automated Video-to-Reel Conversion Platform, which is an artificial intelligence machine, that converts a long video content into short and shareable video reels is the solution offered in this paper. It is a speech recognition system, sentiment and key word extraction system with video editing in an interactive system that was created with Streamlit.

II. METHODOLOGY

The provided system architecture will be founded on a scalable and modular architecture that will include some number of functional units. The modules accomplish a particular task in the processing line which facilitates the processing of the video material between the production of the input and output. The most significant part of the system architecture is the user interface part, video input and storage part, audio extraction part, speech-to-text transcription part, sentiment and topic analysis part, highlight-detection algorithm part, reel generation part and the output delivery part.

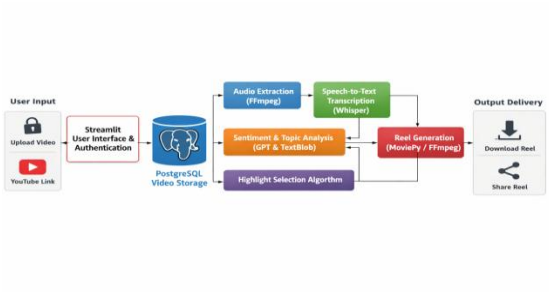

Fig 1: System Architecture

The most significant interface of the system and the users is the user interface layer. According to this interface, one can post the videos, insert the links of the YouTube, trace the course of the processing process and get the reels created.

The video input and storage module is one of the modules that are involved in ingestion and storage of video files. A relational database system is used to store videos uploaded with metadata so as to manage data effectively.

The audio extraction packet will divide the audio track of the video file in a manner that will make it analyze it depending on the speech. As verbal message in the majority of situations is made up of contextual messages, used to mark important spots in a video, analysis of audio tracks is critical in highlighting.

The speech-to-text transcription engine takes the audio that was extracted as a text transcript to convert the audio to text with the help of deep learning models. Such transcripts are also done under natural language processing.

Sentiment and topic analysis module are used to analyze the textual transcript in order to determine the passages that contain emotional data, and the areas that are relevant in the video to discuss.

The detection algorithm of video highlighting is applied in the process of calculating the various values like intensity of the sentiment, density of the key words and pacing of the speech to watch the most fascinating part of the video.

The selected pieces of video works are published with the assistance of reel generation module and reassembled into short and high quality reels.

Lastly, the output delivery module offers the reels that the user created to the user to download or distribute the clips.

The modular design ensures that each of its parts may act independently but possess a coherent processing pipeline, which implies that the system is both flexible and can be extended in the future.

B. User Interface and Authentication.

The platform is an interactive web based platform which is developed using Streamlit a python based data-driven application framework. This interface is made easy enough to be user friendly to the extent that the user can operate the system without necessarily involving a technical aspect.

The users are enabled to upload video files or they can give the URL of other video sharing sites such as YouTube. Once one has uploaded a video, the interface will display real-time processing status,

thereby enabling him/her to see the transcription, analysis and reel generation status. After such processing has been done the resulting reels are then displayed on the interface with a preview and download button which allows you to preview the clip and download.

C. Video Storage and Database Management

Video related information storage and management should be efficient to ensure that the performance and scaleability of the platform is not compromised. The system is founded on PostgreSQL, which is an unshaken relational database management system to store and manipulate the user data, video and processing results metadata.

An entry in the database with the relevant metadata that would contain the user ID, video name, date uploaded, file location and processing status shall be created using a system whereby the user would type the video or a You Tube link. The video is forwarded through the various levels of the pipeline to the database of the data to the present stage of processing.

D. Audio Extraction

The second would be to delete the audio content of the video file after uploading the video and enrolling it to the system. It is a serious procedure as verbal content can contain very crucial information that can give some of the significant events that took place in the video.

The system employs the FFmpeg that is a popular multimedia processing system to extract audio. FFmpeg has a broad video and audio compatibility as well as provides an efficient processing group of multimedia manipulation.

E. Speech-to-Text Transcription

The audio text Whisper model is an automatic speech recognition model that does the speech-to-text transcription by using the audio to text Whisper automatic speech recognition model, which is a complex deep-learning-based automatic speech recognition model. Whisper can also translate spoken words in very high degree of accuracy even in a challenging audio environment that contains noise in the background, speech and different accents overlap. The model may also be applied in the transcription and analyze the audio waveform to produce a written transcript of the audio in the video. The information on the timestamps is also offered in the output because it includes the correlation of all the texts and the video timeline.

F. Sentiment and Topic Analysis

The system is based on natural language processing in the development of a transcript to process the textual information, and generate meaningful insights. It is at this point that the level of feelings and relevance of various parts of the speech are identified.

Two primary instruments are used to achieve this through GPT-based analysis and TextBlob sentiment analysis.

The GPT analysis interprets the text and establishes the most significant issues of discussion, background meaning of the words, and keywords, which are relevant and pertinent to the most significant points of the video. The analysis of the semantic structure of the conversation allows defining what can interest the viewers in the conversation with the assistance of GPT.

G. Reel Generation

The highlight detection module is very important in establishing the most interesting parts of the video. The algorithm compares several characteristics based on the transcript and gives a score to each part depending on its possible relevance and interest.

Some of the major parameters are taken into consideration under the algorithm and they are the intensity of sentiments, repetitions of the keywords, the rate of speech and the relevance of the topic. Sentiment intensity is employed to describe the fact that the presence of emotions exists in which the frequency of key words may be employed to express the presence of meaningful words or phrases. One of the clues towards the topicality of the material to the overall background of the video is the fact that the speaker accentuates the points and the topicality of the subject matter at the specific moment.

© APR 2026 | IRE Journals | Volume 9 Issue 10 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I10-1716256

After the highlight segments have been identified, the system is then used to make the final short-form video reels. It is doneutilizing Moviepy and FFmpeg that offers a great degree of video editing and processing.

The reel generation process involves the extraction of the chosen video pieces having time stamps that were chosen during the highlight detection process. The system also eliminates the redundant frames and flattens all the clips to get started and stopped. The extraction of the video and the end product do not affect the quality of the video and the end product is a professional work and can be distributed on the social media.

III. RESULTS

In order to measure the performance of the proposed Automated Video-to-Reel Conversion Platform, various experimental studies were performed using various forms of long-form video content such as podcasts, lectures, interviews, and webinars. The aim of the assessment was to examine the performance of the system with regard to processing efficiency, highlight detection accuracy, reel generation quality, and the performance of the system.

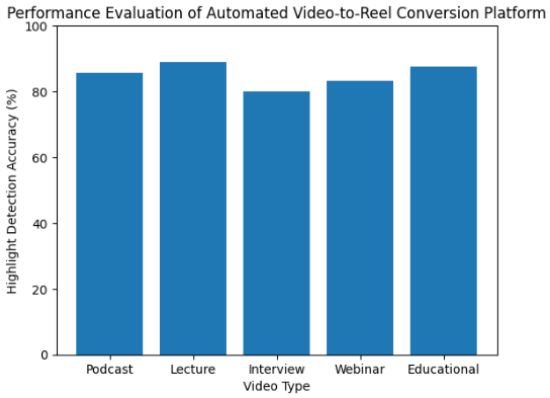

Fig 3: Performance Evaluation Graph

The assessment process was carried out by uploading the videos of different lengths between 10 minutes and 90 minutes and monitoring the work of each processing stage. The system automatically handled the videos by audio extraction, transcription, sentiment analysis, highlight detection, and reel generation steps. The results were compared to manually selected highlights in order to establish the accuracy of the system.

Some of the key metrics that were used in measuring the performance of the system included processing time, the number of highlights detected, accuracy of highlight identification, and output reel duration. The results obtained from the experimental evaluation are summarized in Table I

TABLE.1 Overall Performance Evaluation of the Automated Video-to-Reel Conversion Platform

Video Type	Video Duration (min)	Processing Time (min)	Manual Highlights	System Highlights	Matching Highlights	Accuracy (%)	Generated Reel Duration
Podcast	60	7.5	8	7	6	85.7	45 sec
Lecture	45	6.2	10	9	8	88.9	50 sec
Interview	30	5.1	6	5	4	80.0	35 sec
Webinar	90	8.4	7	6	5	83.3	55 sec
Educational Video	40	5.8	9	8	7	87.5	48 sec
Average	53	6.6	8	7	6	85.1 %	46 sec

IV. DISCUSSION

The experimental findings confirm that the offered system is an efficient solution to the automation of the process of transforming long-form videos into small and interactive reels. Speech recognition, sentiment analysis and highlight detection methods allow the system to detect meaningful parts of long videos without human intervention.

The advanced speech recognition technology to transcribe spoken information accurately is one of the greatest strengths of the system. The text generated by the Whisper model makes it possible to gain a better insight into the narrative form of the video, as

the system would identify significant events, following linguistic and emotional indicators.

The integration of topic detection and sentiment analysis methods is also another major benefit. The intensity of emotion tends to be associated with viewer attentiveness, and the system manages to detect areas where speakers are passionate, excited, or making significant announcements. These are the most appropriate segments that are usually suitable in short-form content.

The automated video editing software like FFmpeg and MoviePy also helps boost the functionality of the system through the efficient extraction and compilation of the chosen parts to high quality reels. Automation of this process is a great way of saving time and effort, which are needed to manually edit the video.

Although it is effective, there are limitations of the system. As an example, videos of extremely low audio quality or containing much background noise can influence the accuracy of transcription. Also, visual only highlights like changes in the dramatic scenes or a visual humorous element might not be identified since the existing system mostly examines the speech content.

However, the system has a high potential as a useful resource to content developers, digital marketers, educators, and media professionals who require repurposing long videos into captivating short-form videos.

V. CONCLUSION

The paper introduced a Automated Video-to-Reel Conversion Platform which will be aimed at simplifying the process of converting long-form video content to small and catchy reels. It is a combination of different high-tech capabilities that include speech recognition, natural language processing, sentiment analysis, and automatic video editing and create a smart highlight detection pipeline.

The suggested system allows the user to add video or even a link to YouTube with the assistance of an interactive interface of Streamlit. The video uploaded is subject to a series of modules including audio extraction, speech transcription using Whisper, sentiment and topic analysis using GPT and TextBlob, highlight extraction and automatic reel creation using FFmpeg and MoviePy.

During experiment basis, whether the system was capable of working with long videos or not and whether the overall accuracy of the highlight detection process is about 85 percent was established. The produced reels were short, informative, and suited the short-form video apps, such as Instagram Reels, and YouTube Shorts.

On the whole, the proposed platform will be an automated and scalable video summarization and repurposing information platform. The system assists content creators to create more and quality content, as the number of manual labor required to extract highlights and edit videos will be diminished, which increases the rates of engagement with the audience.

VI. FUTURE SCOPE

The proposed system has shown good outcomes, but there are some improvements that can be made to enhance its abilities and operations.

Another possible solution is the incorporation of computer vision methods to examine visual features like facial expressions, gestures, and changes of scenes. This would enable the system to identify moments that are visually stimulating and might not be pronounced on analysis of the speech alone.

The other future improvement is the use of multilingual transcription and analysis. The option to support multiple languages would allow the platform to reach a broader global audience and can process videos in different regional languages.

The system might also be expanded to include automatic generation of subtitles to generated reels. Subtitles enhance accessibility and enhance viewing enjoyment particularly when using social media networks where videos are commonly viewed mutely.

Moreover, the approach of including machine learning models based on the data of social media engagement would enhance the predictability of highlights further by pinpointing the parts that are most likely to receive views, likes, and shares.

Last but not least, the next iterations of the platform may feature the ability to integrate directly with social media, whereby users may be able to automatically post generated reels to other platforms such as Instagram, TikTok, and YouTube.ent.

REFERENCES

[1] A. Radford et al., "Robust Speech Recognition via Large-Scale Weak Supervision," *OpenAI Research*, 2022.

[2] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed Representations of Words and Phrases and their Compositionality," *Advances in Neural Information Processing Systems*, 2013.

[3] S. Bird, E. Klein, and E. Loper, *Natural Language Processing with Python*, O'Reilly Media, 2009.

[4] M. Lutz, *Learning Python*, 5th ed., O'Reilly Media, 2013.

[5] D. Jurafsky and J. H. Martin, *Speech and Language Processing*, 3rd ed., Pearson, 2020.

[6] F. Chollet, *Deep Learning with Python*, Manning Publications, 2017.

[7] A. Rosebrock, "Video Processing with OpenCV and Python," *PyImageSearch*, 2018.

[8] R. Reinders, "FFmpeg Basics: Multimedia Handling with a Fast Audio and Video Encoder," *Linux Journal*, 2019.

[9] M. McKinney, *Python for Data Analysis*, O'Reilly Media, 2018.

[10] S. Loria, "TextBlob: Simplified Text Processing," *Python Library Documentation*, 2018.

[11] T. Kluyver et al., "Jupyter Notebooks – A Publishing Format for Reproducible Computational Workflows," *IEEE Conference Proceedings*, 2016.

[12] J. Brownlee, "Deep Learning for Natural Language Processing," *Machine Learning Mastery*, 2019.

Dean and G. Hinton, "Deep Learning," *Nature*, vol. 521, pp. 436–444, 2015.

[13] T. O'Shea and J. Nash, "An Introduction to Convolutional Neural Networks," *arXiv preprint*, 2015.

[14] R. Szeliski, *Computer Vision: Algorithms and Applications*, Springer, 2010.