iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 12 **Issue:** XI **Month of publication:** November 2024

DOI: https://doi.org/10.22214/ijraset.2024.65452

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 XI Nov 2024- Available at www.ijraset.com

BhashaBlend: Enabling Multilingual understanding of videos through NLP for Deaf and Hearing Impaired users

Milind Kamble[1], Kashish Goel[2], Aditya Kadgi[3], Kanika Gorka[4]
Department of Electronics and Telecommunication, Vishwakarma Institute of Technology, Pune, Maharashtra, India-411037

Abstract: ***In India's diverse educational landscape, linguistic barriers and hearing impairments present significant challenges to effective learning. The research introduces an innovative NLP-driven solution aimed at overcoming these hurdles by offering a comprehensive educational platform. This platform includes features such as subtitle generation, text summarization, Q&A generation for content recall, multilingual translation, and Indian Sign Language (ISL) support. Leveraging a fine-tuned Whisper model for Speech-to-Text Translation (STT) and conducting a comparative study of BERT, T5, and TextRank for subtitle summarization, this approach ensures optimized language understanding and content generation across various NLP applications. Through the utilization of ConceptNet and BERT for question answering tasks, the system demonstrates impressive performance metrics, including a 95.12% accuracy rate in subtitle generation and an average WER of 3.43%. While manual evaluation was necessary due to the absence of specific datasets, this innovative approach represents an advancement in addressing educational accessibility gaps in India, fostering the Right to Education to all individual users in an educational environment.***

Keywords: Text, audio, video, subtitle, Summarization, Translation.

I. INTRODUCTION

A. Subtitle Generation and Translation

Recent research shows that 69% of viewers prefer silent video viewing in public spaces, while only 25% prefer private settings. Additionally, 92% prefer sound-off viewing on mobile devices. The global prevalence of online video content is significant, with an audience reach of 92.3% in Q2 2023. Understanding context and environment is crucial in video consumption habits. The percentage of users who would rather watch Facebook videos with subtitles is shown in Fig 1.[1]

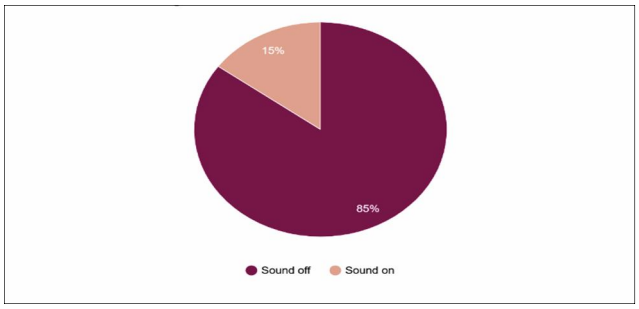


Fig 1: Percentage of videos on Facebook viewed with sound on/off.

For example, Discovery Digital Network, a YouTube channel with over 100 million videos, experimented with adding subtitles to their videos to observe the impact on video views. They found that YouTube videos with subtitles had a 7.32% higher total view count. This rise confirms that videos are a preferred online format, covering everything from educational content to product reviews. Up to one-third of all internet users watch videos on YouTube every day. This is a substantial chunk of the user base. When videos have closed captioning or any other type of subtitles, users are more engaged and spend more time on the platform. As we delve into the nuances of video consumption, these statistics become essential, revealing the interaction between viewer behavior, context, and the widespread popularity of online videos.

©IJRASET: All Rights are Reserved

According to a survey conducted on YouTube, closed captioning can increase user engagement time on a particular video by up to 40%. A study by Verizon Media and Publicis Media found that up to 80% of participants prefer watching videos with subtitles. Subtitles enhance video accessibility, comprehension, and engagement, especially for those with hearing loss. With over 5% of the world's population having some form of hearing loss, closed captioning could benefit at least 360 million people worldwide. The limitations of previous works include potential challenges in accurately identifying speakers and gaps in speech, as well as the need to ensure precise time synchronization. The traditional subtitle generation systems saw a dip in the accuracy for noisy videos.[2] In contrast, despite being trained on huge data, many systems face constraints related to its reliance on Deep-Speech's high memory usage and limited language support, which may hinder its suitability for resource-constrained environments and diverse linguistic contexts. Additionally, Deepspeech saw an increase in WER (word error rate) for longer length videos.[3]

B. Multilingual Language Translation

India's linguistic diversity highlights the importance of translation in bridging languages and preserving cultural heritage. Despite English's global dominance, translation remains crucial for effective communication and cultural exchange, especially online. Translation plays a crucial role in ensuring social harmony and peace through effective intercultural communication. The Language translation NLP market is projected to reach US$5.92 billion by 2024, with an anticipated annual growth rate of 12.40% leading to a market volume of US$11.94 billion by 2030. The United States is expected to be the largest market, reaching US$2,122.00 million by 2024. These statistics highlight the growing recognition of translation's indispensable role in fostering connectivity and understanding globally. As the market expands, translation remains vital for cultural preservation, communication, and societal cohesion. [4] The major limitations in the previous systems related to multilingual language translation, especially in Indian languages was lack in accuracy, moderate accuracy and the contextual understanding was lot during the translation.[5]

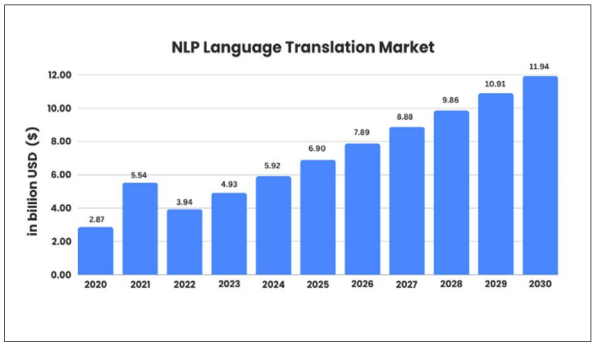


Fig 2: NLP Language Translation Market in USD ($) bn by 2030

The Fig 2 talks about the projection of the Multilingual Language Translation market by 2030.This also states the dire need of translation of different languages as a way of providing inclusive educational platforms to all.

C. Subtitle Summarization

Subtitles in videos serve to help viewers quickly grasp the content's essence, making a summary strategy essential to save time. As data continues to explode, with over 180 zettabytes predicted by 2025, clear and concise summaries are crucial for condensing vast amounts of information into easily digestible formats for efficient use and evaluation.

D. Question-Answer Generation

Questions and answers in education promote engagement, comprehension, and feedback. They encourage critical thinking and knowledge retention while offering valuable insights for instructors to enhance teaching methods. Overall, they play a crucial role in assessing understanding and fostering active learning. The potential challenges previously included accurately identifying verbs and matching them with tokens, which may affect the precision of the generated answers. Additionally, the restricted domain for answer extraction may limit the system's applicability to broader contexts. [6] Moreover, the problem lies in the datasets on which it is being trained. For example, if a system is trained on the basis of medical data, then its reliance on POS and NER for text labeling will affect drastically, which may result in inaccuracies, particularly in complex sentence structures or ambiguous contexts. Furthermore, while achieving an 80% accuracy rate is commendable, there may still be room for improvement to address cases of ambiguous or nuanced questions that require deeper semantic understanding.[6]

E. Indian Sign Language(ISL) Translation

The Americans with Disabilities Act (2016) mandates auxiliary aids like closed captions and subtitles for video accessibility, ensuring fundamental rights. Fig 3 depicts the most popular accessibility elements utilized by viewers[7].

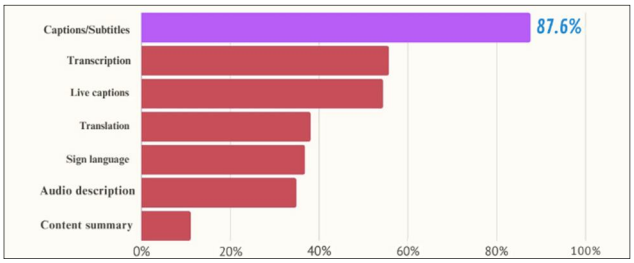


Fig 3: Video Accessibility Analysis for viewers over the internet

The following Fig 4 illustrates the distribution of video genres over the Internet [8].

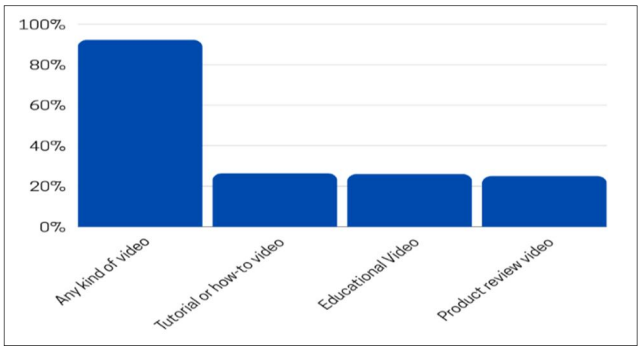


Fig 4: Percentage of genre of videos over the Internet

The introduction of a universal knowledge assimilation tool is driven by the rapidly expanding education industry, with the global e-learning market expected to reach around $400 billion by 2026, doubling its size from approximately $200 billion in 2019. Notably, the learning management system (LMS) market alone generated over $18 billion in revenue in 2019.[9]

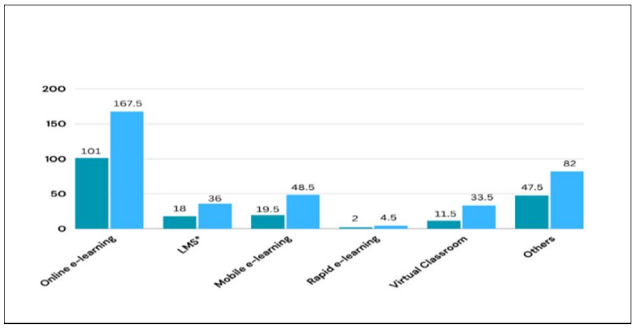


Fig 5: E-learning market size 2019-2026 in US ($) billion dollars

The Fig 5 states the projected market size of the E-learning industry, also stating a need of why Online e-learning has become an important part in the after the COVID-19 pandemic. Previously, there were limited domain-specific solutions in education, hindering comprehensive e-learning capabilities. Additionally, language accessibility poses a significant challenge in India, additionally for individuals with auditory or hearing impairments, as existing learning tools often lacked provisions for their needs. While efforts were made to provide education to all, achieving this goal on a cross-cultural platform presented difficulties. The research aimed to address these issues by facilitating education in an interactive way, regardless of language barriers. Moreover, the system aims to cater to individual learning paces, offering solutions to aid concept retention. Recognizing the efficiency of learning in one's mother tongue, the objective was to leverage new technologies effectively in education. Through the development of a web-based platform integrating natural language processing (NLP) and deep learning (DL) capabilities, the research successfully overcame these challenges.

II. RELATED WORKS

The paper proposes an automatic subtitle generation and semantic video summarization technique leveraging speech recognition and NLP-based text summarization algorithms. It introduces five text summarization methods, including Luhn's, LSA, Text Rank, Lex Rank, and Edmundson's algorithms, and demonstrates their efficacy through experimental results. The approach shows promise for enhancing video navigation and content accessibility on platforms like YouTube and Dailymotion in the era of big data. [10] The objective was to develop an LDA-based subtitles summarization model for educational videos. Methodology involved dataset expansion through human summaries from subtitles, followed by LDA model application for extractive summarization. Performance comparison with TF-IDF and LSA models showed LDA's superiority in ROUGE scores and human evaluation. Future work may address challenges with topics rich in numbers, Booleans, and verbs, focusing on length control and punctuation improvements. [11] The paper examines subtitle segmentation parameters like BLEUbr and a new metric, Sigma, showcasing different sensitivities to classification noises. The limit projection method estimates system results, highlighting Sigma's consistency with other metrics. Limitations include challenges in distinguishing homogeneous production and biases from alignment algorithms. Future work involves developing balanced metrics and validating human decision links, offering insights for improved subtitle segmentation assessment. [12] This research introduces a Python-based approach for efficient YouTube video summarization, utilizing the YouTube Transcript API and advanced BERT algorithms. The method includes extractive summarization, video-to-text extraction, URL validation, and language translation, offering benefits for teaching, patent research, and information extraction. [13] This work employs a variety of datasets and evaluation criteria to conduct a thorough comparison between AI-based translation approaches and conventional machine translation (MT) procedures. Supported by experimental and control group evaluations, it shows that AI translation performs faster, more affordably, more reliably, and more accurately than traditional methods. It also looks at the incorporation of AI translation in the classroom and how it might help students become more proficient and self-assured. Overall, the study highlights how artificial intelligence (AI) translation technology is revolutionizing both translation operations and education.[14]

The researchers proposed an Android-based application programming interface (API) powered by Firebase, providing a comprehensive solution to the ongoing problem of language barriers. It solves the problem that users and travelers have when they come across messages written in different languages. The suggested method enables users to translate scanned text images into the languages of their choice even in the absence of an internet connection by combining text recognition and translation functionalities. By leveraging technologies such as OCR, ML Kit, and Firebase, the solution efficiently addresses challenges in language translation by streamlining the process and enhancing user engagement. [15] The study aimed to generate computer-based exam questions aligned with learning goals, utilizing NLP tools like SpaCy, NLTK, and Sense2vec. NLTK managed sentence segmentation, while Sense2vec provided choices and accurate answers, with future plans focusing on automating answer validation and enhancing assessment efficiency and accuracy. [16] The authors proposed a method where NLP based methods were used to extract features like POS tags, synonyms, hypernyms etc. Apache Solar server was used to store all the information (Named Entity Recognition) was performed on the given corpus and questions based on keywords Who, When, and where were formed. A parse tree was formed to find the possible answers to a particular question and the best viable sentence was chosen based on search done by the Solar server. The advantages of this system were that it was useful for single word answers. The limitations were finding answers for similar but differently named proper nouns [17]

The research involved creating a Q&A system based on specific domain to support a cohesive learning environment. A semantic relation was extracted from the words, and synonyms of the extracted keywords were extracted using WordNet. The similarity of these words was calculated using cosine similarity, and the word with the higher cosine was selected as the exact answer. The proposed method was not able to handle all types of questions. Furthermore, it could not provide answers to multiple questions where the user's query combines two or more questions. [18]

The paper delves into sign language processing, examining datasets, recognition, translation, and modeling through interdisciplinary collaboration. With insights from a workshop involving 39 experts, the study aims to tackle challenges and foster inclusive, efficient computer-based solutions for sign language communication. [19] The paper presents EDUZONE, an educational tool tackling online learning challenges via video summarization, indexing, and digital human technology, boosting learner engagement and efficiency. While praised for its accuracy and usability, it faces latency in video summarization and relies on speech-to-text accuracy. Future efforts may center on improving summarization accuracy, integrating handwritten content, and refining the digital human assistant's training. [20]

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

III. PROPOSED SYSTEM

Unlike traditional methods reliant on ".srt" files, our innovative dynamic subtitle creation minimizes storage, seamlessly embedding subtitles into videos while offering translation, summarization, Q&A sessions, and ISL transcription via this Universal Knowledge Assimilation Tool. The system architecture for the system is shown in Fig 6.

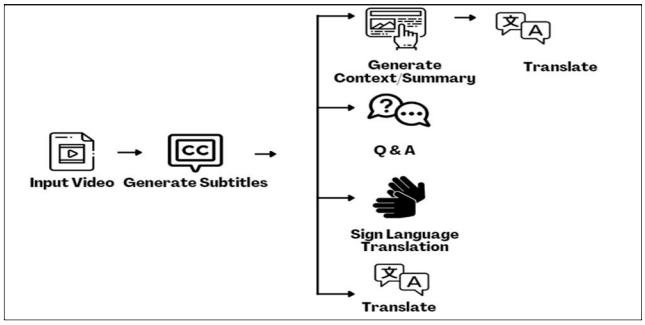


Fig 6: Proposed System Architecture

The flowchart of the propounded system is as follows as represented in Fig 7.

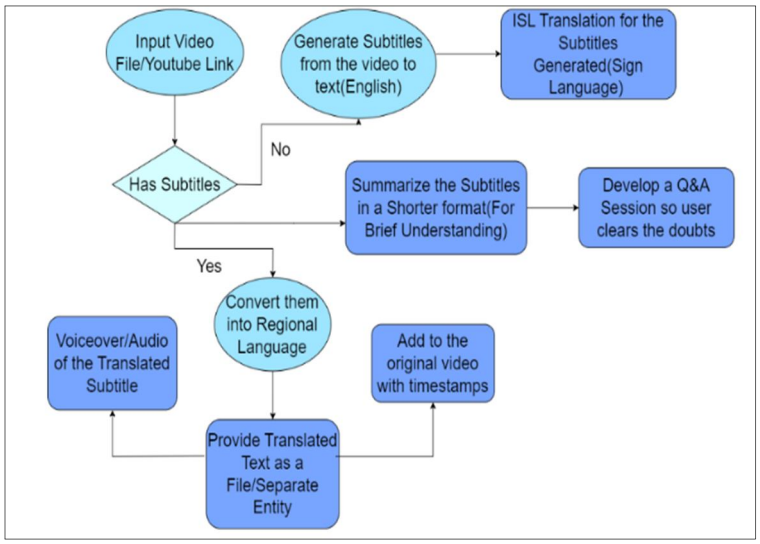


Fig 7: Flowchart of Proposed System

A. Subtitle Generation and Translation

Traditionally, generating subtitles entailed extracting audio, creating SRT transcripts, and synchronizing them with videos, a cumbersome and accuracy-intensive process. However, our system transforms this by leveraging OpenAI's "Whisper" mode. We forego SRT files, instead overlaying word-level timestamps directly onto videos, ensuring a seamless viewing experience. This eliminates time-consuming synchronization, making the process more efficient and accessible, particularly for videos with unfamiliar accents or languages.

B. Video Summarization

Before TextRank, BERT, and T5 integration, conventional summarization methods, reliant on word frequency and rule-based approaches, struggled with contextual comprehension and semantic nuances, often yielding incoherent summaries.

C. Question-Answer Generation

Thus, a Question and Answering system was developed to clear any specific doubts that the user has in the video. This feature was implemented keeping in mind the need of QA systems for educational videos, where the student can use this feature to clear his/her doubts or might revise thus providing him/her a one-stop solution to test the concepts.

D. ISL(Indian Sign Language) Translation

With over 70 million deaf individuals worldwide and 200+ sign languages, effective communication methods are crucial. Sign language serves as a proficient means for deaf individuals to connect with the world. Our prototype addresses this need, targeting primarily those with hearing disabilities, along with their families, professionals, and content creators. The first step in the subtitle generating process is to accept a.mp4 file as input. Users can upload directly from their device or submit a YouTube link. For YouTube links, Pytube is utilized, and ffmpeg is used to extract audio and create.mp3 files. Equipped with 680,000 hours of training data, the whisper library functions as an automatic speech recognition (ASR) system that facilitates multilingual transcription and English translation. The Whisper architecture which has implemented as a simple encoder-decoder transformer is shown in Fig 8.

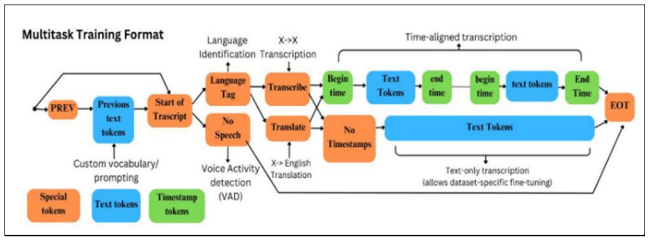


Fig 8: The encoder-decoder transformer of Whisper (a STT) model

The whisper ASR system generates subtitles paired with accurate timestamps, while MoviePy facilitates the addition of burned-in subtitles to the video. For translating subtitles, the googletrans library was selected due to its extensive language support. NMT models are employed for translating English text to various Indic languages. The simplified equation representing the translation process is as follows:

$$y^{\wedge} = \arg max_y P(Y|x) \quad \square\square\square$$

The y^ is the translated text in the target language, along with X being the input text in English, y represents possible translation in the targeted language calculated for the conditional probability of generating the target translation. Subtitle summarization utilizes BART extractive summarization, T5, and TextRank algorithms. BART and T5 involve tokenization, encoding, and decoding mechanisms, with BART emphasizing abstractive summary generation. Post-processing ensures a coherent summary from the generated text. Fig 9 depicts the tokenization of words in all the mentioned methods.

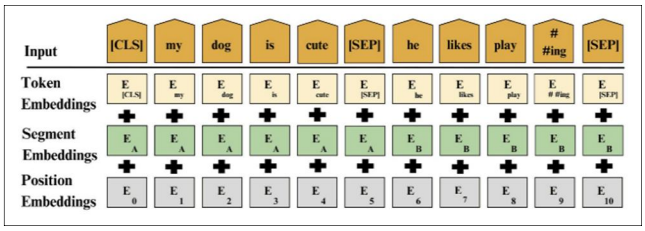


Fig 9: Tokenization of words in a sentence

The model is pre-trained on tokens "t," which look backward to tokens "k" from the past, in order to compute the current token.

$$L_1(T) = \sum_i^k \log P(t_i|t_{i-k}, \ldots, t_{i-1}; \theta) \quad \square\square\square$$

To make the model more robust on certain tasks, in our instance summarization, it is been modified in a supervised manner to maximize the probability of label "y" given feature vectors x1–xn.

$$L_2(C) = \sum_x^y \log P(y|x_1, \ldots . x_n) \quad \square\square\square$$

Now combining (2) and (3), we get an equation where the symbol lambda (λ) represents a weight parameter that has been learned to regulate the impact of language modeling.

$$L_3(C) = L_1(C) + L_3(C) \quad \square 4\square$$

The equations for summarizing text for the BERT model and T5 are as follows:

$$BARTSumm(X) = argmax_Y \sum_{i=1}^n P(y_i|y_{<I}, X) \quad \square\square\square$$

$$T5Summarize(X) = argmax_Y \sum_{i=1}^n P(y_i|y_{<I}, X) \square\square\square$$

In the above equations (5) & (6), *X* represents the input text, *Y* represents the summary and *P* represents the conditional probability of generating the i-th token in the summary of the previously generated tokens $y_{<i}$.. The equations represent the process of maximizing the conditional probability to find the summary Y from input text X. The TextRank algorithm, on the other hand, initializes by tokenizing raw text, preprocesses by removing stop words and noise, constructs a sentence graph based on similarity, and applies the PageRank algorithm iteratively to select top-ranked sentences for the extractive summary. The initial approach for generating MCQs involved utilizing Sense2vec to create distractors, leveraging neural network-generated word embeddings called "senses" capturing semantic data based on labels like Entity Name. WordNet's hypernyms provided abstract concepts, like countries in Asia for the keyword "India." These strategies, along with fine-tuning a BERT model on subtitles, facilitated diverse and contextually relevant distractor generation, enhancing the robustness and semantic richness of the MCQ options. In our context we have used "lava" as the keyword for generating the distractors. The vector embeddings for similar words were ('molten_rock|NOUN', 0.8014), ('lava_pit|NOUN', 0.7955), ('volcano NOUN', 0.7906), ('molten_lava|NOUN', 0.7889). Various methods like Sense2vec, WordNet, and Concept-net were used to generate incorrect options. WordNet provided abstract concepts, while Concept-net gave related terms through API calls. Additionally, fine-tuning a BERT model on subtitles allowed handling out-of-vocabulary words effectively by breaking text into sub-word units using Word-Piece tokenization. The tokenization can be mathematically represented as:

$$S_i = WordPTokenize(t_i) = \{s_{i1}, s_{i2}, \ldots . s_{im}\} \quad \square\square\square$$

Where *m* is the number of sub-word units in token t_i after tokenization. The overall tokenization of the input sequence involves the Word-Piece tokenizer to each token t_i in the sequence, resulting in a sequence of sub-word units $S=(S_1,S_2,\ldots S_n)$.The equation of every Word-Piece tokenizer applied on every token is as follows:

$$S_i = WordPTokenize(t_1)..WordPTokenize(t_n) \quad \square 8 \square$$

In BERT-based question answering, token scores for answer start and end positions are stored. The highest-scoring token indices are determined using 'torch.argmax (start_scores)' and 'torch.argmax (end_scores)'. These indices identify the answer span, which is then sliced from the 'tokens' list. Finally, these tokens are joined to form the answer string. This approach efficiently generates a coherent textual representation of the answer. The pseudo code for the two approaches followed for the same are given in Fig 10 and Fig 11. Fig 10 depicts the pseudo code for generating the MCQs and Fig 11 represents the One-word Answers.

```
function generate_mcq_and_answers(keyword):
    # Approach 1: Using Sense2vec
    # Generate distractor images using Sense2vec package
    sense2vec_embeddings = sense2vec.generate_embeddings(keyword)
    similar_words = sense2vec.find_similar_words(sense2vec_embeddings)
    mcq_options_1 = similar_words[:4]

    # Approach 2: Using WordNet
    wordnet_hypernyms = wordnet.get_hypernyms(keyword)
    mcq_options_2 = wordnet_hypernyms[:4]

    # Approach 3: Using ConceptNet
    conceptnet_related_terms = conceptnet.get_related_terms(keyword)
    mcq_options_3 = conceptnet_related_terms[:4]

    return mcq_options_1, mcq_options_2, mcq_options_3
```

Fig 10: Pseudo code for generating MCQ options(distractors)

```
function extract_answer_span(context_tokens, question_tokens, start_scores, end_scores):
    # Tokenize input context and question
    context_word_pieces = tokenize(context_tokens)
    question_word_pieces = tokenize(question_tokens)

    # Calculate total context tokens
    total_tokens = len(context_word_pieces)

    # Find start and end positions
    start_position = torch.argmax(start_scores)
    end_position = torch.argmax(end_scores)

    # Ensure start position is within context
    start_position = min(start_position, total_tokens - 1)

    # Ensure end position is within context and greater than start position
    end_position = min(max(end_position, start_position), total_tokens - 1)

    # Retrieve answer span tokens
    answer_tokens = context_word_pieces[start_position:end_position + 1]

    # Convert answer tokens to string
    answer_string = ''.join(answer_tokens)

    return answer_string
```

Fig 11: Pseudo code for One-word Question and Answers

E. ISL Translation

In the initial project phase, text data from Indian Government sources and sign language GIFs, alongside individual letter images, were collected and standardized. Text underwent preprocessing, while the translation algorithm iteratively processed words to display corresponding sign language GIFs or letter images through Tkinter, incorporating user-friendly timing parameters for seamless content display and relevance scores based on keyword frequency and text similarity. Let's consider the relevance score for a GIF Gi as R(Gi). The equation for calculating the relevance score can be represented as:

$$R(G_i) = \sum_{j=1}^{n} w_j . f_j(G_i)$$ □□□

Where *n* is the total number of factors considered(e.g. frequency of keywords, presence of tags, similarity of descriptors),where w_j is the weight assigned to factor *j*(normalized to 1).$f_i(G_i)$ is the value of factor *j* for GIF.The evaluation process involved meticulous comparison with the original text to assess translation precision and scrutinized the algorithm's efficacy in selecting the appropriate mode of translation, alongside gathering user feedback for overall assessment.

IV. RESULTS AND DISCUSSION

The following Table 1 represents a textual format of the preferred technologies used for the numerous features of the multimodal platform.

TABLE I.

Task	Technology Experimented	Reason Used	Feature
Subtitle Generation	Whisper Wav2vec	Whisper-Works in noisy environme	Scrolling Subtitles
Translation	DeepL googletrans	Googletrans-100 Languages DeepL-Only 35 Languages (Non Indic)	Translation of Subtitles
Content Summarizati	T5, BART and TextRank	BART for Conte based summariz	Pipeline ba ed content summariza on
Question & Answering Model	T5, ConceptN WordNet and BERT	Context (Subtitl based Q-A)	MCQ Single blan answers
ISL-Indian Sign Langua Translation	ML and NLI	Easy understand for students with auditory impairm nts.	GIF based video of phrases/let ers

Table 1: Textual representation of the preferred technologies used for the system

The research achieved several goals, including the development of a deep-learning pipeline for accurate STT conversion to generate subtitles for videos. NLP algorithms were implemented for context generation, multilingual translations, and question-answering. Integration of ISL transcription enhances accessibility for users with hearing impairments.

A. Subtitle Generation

The verification dataset includes .mp4 videos with durations ranging from 60 to 200 seconds, featuring predominantly English-language audio content. It offers a diverse mix of educational and entertainment-focused videos, suitable for various applications such as subtitle generation and content analysis.



Fig 12: Official Video of the CIPAM (Ministry of Commerce and Industries) of India

Fig 12 shows the examples of the types of videos included in the dataset used for the purpose of subtitle generation. The speech-to-text output is overlaid onto the original video, synchronized with timestamps and keyword highlighting for clarity. Table 2 represents a comprehensive and comparative analysis of the WER (Word Error Rate) while STT conversion using existing technologies. As there was an unviability of a domain-related data set, some examples from YouTube videos were taken as a way to analyze the performance of the fine-tuned model.

TABLE II.

COMPARISON OF WORD ERROR RATE(WER)				
Video Title	Whisper (WER)		Wave2Vec	Auto Generated by You tube
	Small	Medium		
Pre-Chewed Food Delivery | Explainer Video	8.51%	5.32%	38.54%	30.93%
Animated Explainer Video Demo	25.43%	14.09%	36.86%	23.08%
Explainer Cafe's Animated	1.9%	0.6%	55.13%	34.62%
PPC Adwords Animated Explainer Video	4.05%	0.00%	42.07%	25.09%
Infographics Animation | Explainer video production	5%	0.00%	45.00%	20%
How to hot-swap the smart heads and their positions Atlas Copco	1%	0.00%	36.00%	16.50%
Atlas Copco - Home of Industrial Ideas	7.64%	0.69%	48.61%	22.92%

Table 2: Comparative Analysis of WERs for different methodologies

The Fig 13 and Fig 14 shows the input video from the user and the output video respectively:

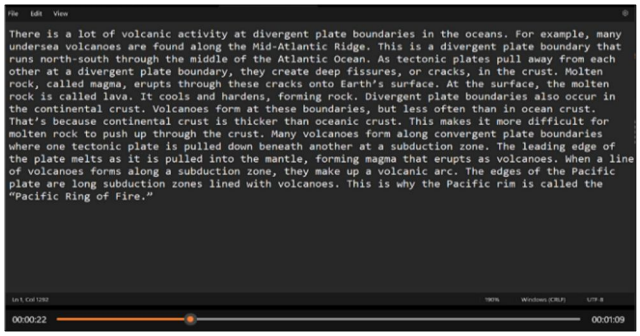


Fig 13: Input video(From the local device or Youtube Link)

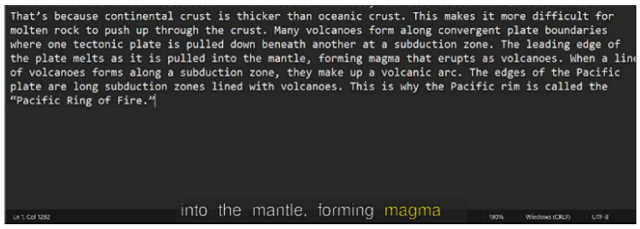


Fig 14: Output video

B. Subtitle(Text) Summarization

The outputs for the three methodologies implemented for the summarization of the subtitles were BART Extractive summarization, T5 summarization and Text-Rank algorithm. For displaying the extractive summarization of the subtitles using BART, a GUI-based interface was applied to display the summarized text. This GUI serves as a way to provide to summarize any text apart from the subtitles. The Fig 15 shows the output thus provided by the Extractive BART summarizer model which is trained on the input context. The following Table 3 represents the Cosine similarity of the original(manual) subtitles to our model for STT conversion.

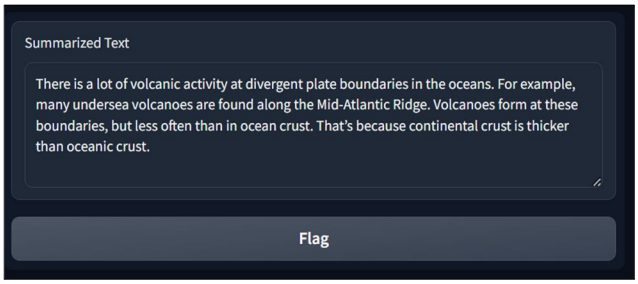


Fig 15: Output-Box for BART based Summarization.

TABLE III.

Text Similarity of Actual Text to Generated text		
Cosine/Spacy Similarity	*BART*	*T5*
T5	96.59%	-
TextRank	95.9%	93.6%
BART	-	88.78%

Table 3: Cosine similarity of Original Subtitle to Generated Text

C. Subtitle Translation

For using the Translation feature for the system, a Flask page was created to enter the input text and a choice of Indian regional languages as this feature tries to bridge the barrier between accessibility of educational content from English to other Indic languages. Fig 16 is the initial page where there is pre-defined text extracted as soon as the subtitles are made.

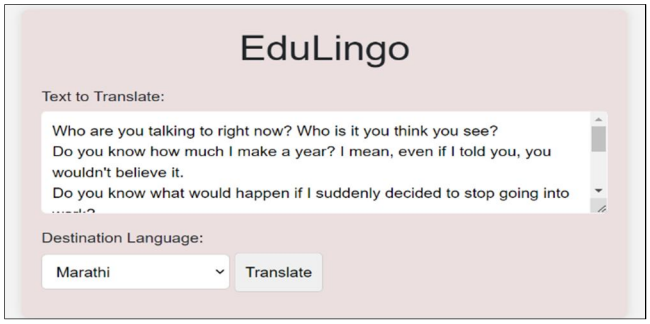


Fig 16: Input Text-Box for Multilingual Text Translation

The translated text obtained in the selected language is then dynamically updated on the page. Fig 17 shows the translated text obtained in Marathi.

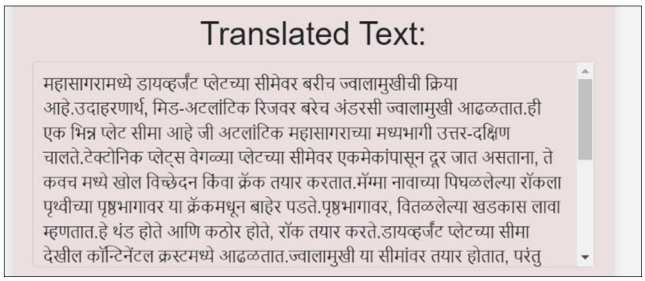


Fig 17: Output-Box for Multilingual Text Translation

D. Question and Answer Generation

Several methodologies were curated for creating a Question-and-Answer Generation model based on the subtitles namely MCQs and one-word type question and answers. Fig 18 and Fig 19 are MCQ based questions built on each and every line of the context(subtitles).

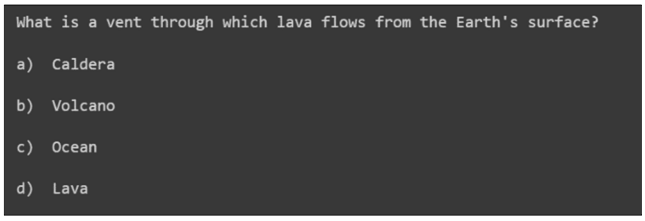


Fig 18: MCQ based on distractors used by ConceptNet

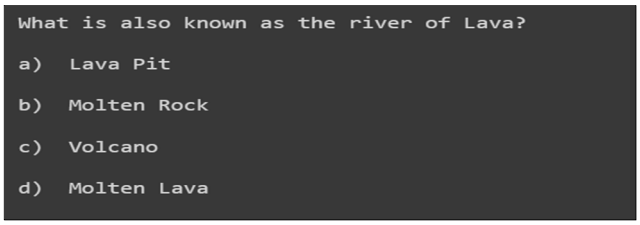


Fig 19: MCQ based on distractors used by Word-Net

E. Indian Sign Language(ISL) Transcription

An essential feature is converting subtitles to sign language, catering to auditory impairments for a clear understanding at the user's pace.

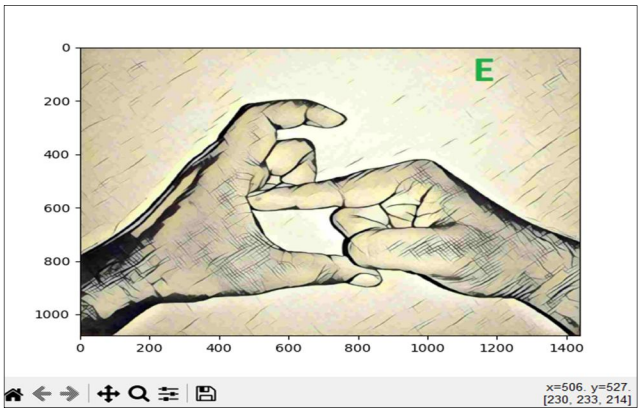


Fig 20: Letter based GIF representation using Tkinter

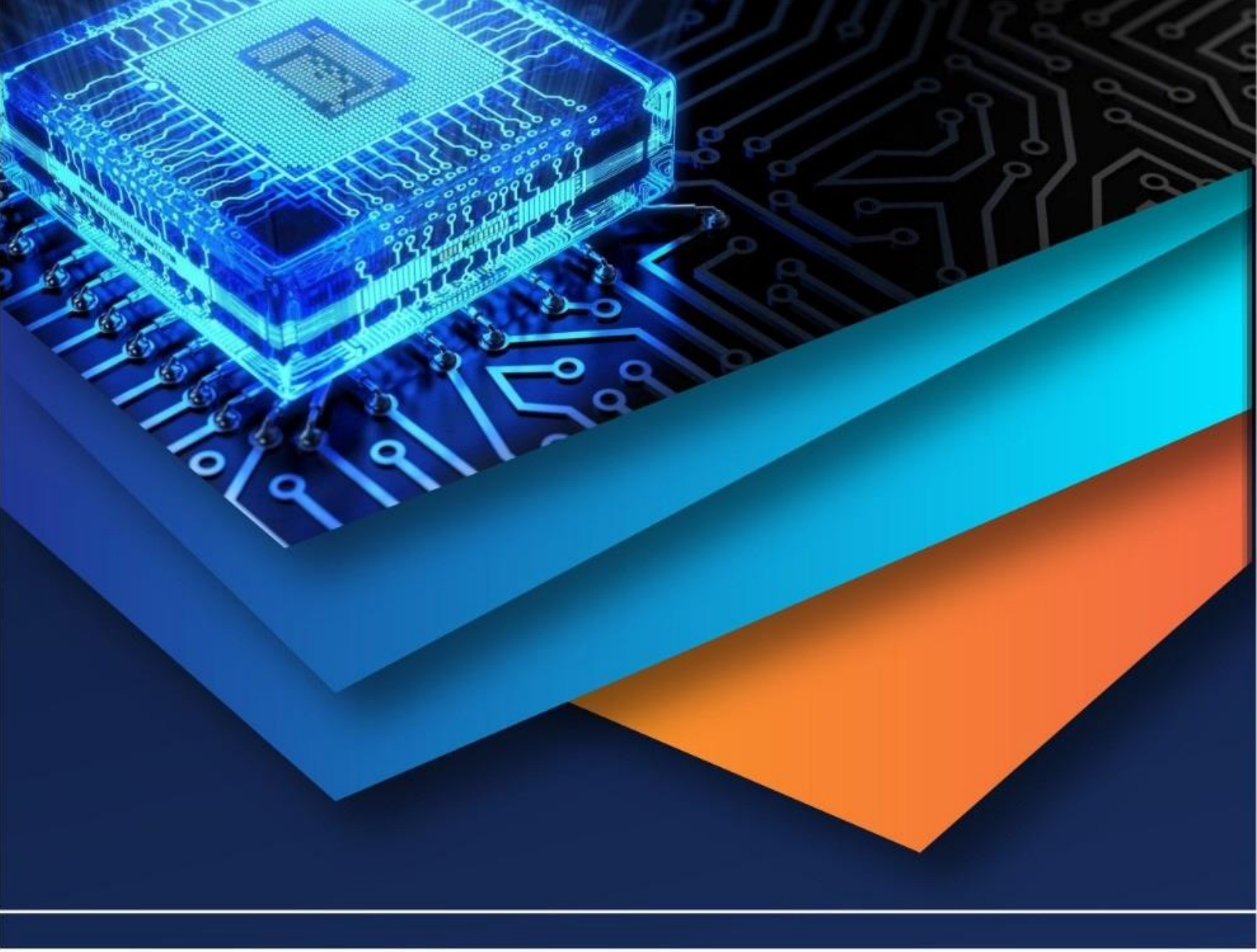
Fig 20 shows the letter-based representation of the word provided it is not found in the dataset. Letter based GIFs will be displayed if the specific word is not found in the dataset. The study found that integrating Whisper for speech-to-text (STT) improved the effectiveness of videos recorded in noisy environments, enhancing learning through multilingual support and real-time subtitles. However, optimizing speech recognition algorithms and addressing video quality issues are critical for long videos, especially regarding subtitle translation. Despite limitations in providing subtitles in other Indic languages, the research positively impacted students, particularly those for whom English is a third language. Further work is needed for Indian Sign Language (ISL) translation and indexing GIFs of phrases from a government-approved database.

V. CONCLUSION

In this project, there was a successful attempt to create a multimodal and multilingual based platform to provide accessibility of education to all students despite language constraints. Also, the system helped pave to way to provide educational content to users having hearing impairments. Support of English language to the videos and ISL transcription for certain phrases were some of the limitations in the research. Future features include adding multiple language support to subtitles especially in Indic languages namely Devnagari and Dravidian languages based in India. Additional features to be worked on involve real-time human voice cloning for the audio, and a 3-D-based model for ISL translation.

REFERENCES

[1] 80% of People Prefer Video Subtitles. Here's How they Affect Engagement. https://www.kapwing.com/resources/subtitle-statistics/.
[2] Akhil Kanade, Sourabh Gune, Shubham Dharamkar, Rohan Gokhale,"Automatic Subtitle Generation for Videos",International Journal of Engineering Research and General Science,vol. 3, Issue 6, 2015
[3] Prachi Sharma, Manasi Raj, Pooja Jangam, Sana Bhati, Prof. Neelam Phadnis, "Automatic Generation of Subtitle in Videos," SSRG International Journal of Computer Science and Engineering , vol. 6, no. 4, pp. 11-15, 2019, https://doi.org/10.14445/23488387/IJCSE-V6I4P103
[4] Language translation NLP - Worldwide, https://www.statista.com/outlook/tmo/artificial-intelligence/natural-language-processing/language-translation-nlp/worldwide.
[5] M. S. B. C., “Machine Translation Using Open NLP and Rules Based System ‘English to Marathi Translator’”, IJRITCC, vol. 2, no. 6, pp. 1730–1733, Jun. 2014.
[6] Shivani Singh,Nishtha Das,Rachel Michael,Dr. Poonam Tanwar,"The Question Answering System Using NLP and AI",International Journal of Scientific & Engineering Research vol. 7, Issue 12, December-2016.
[7] Priti Gumaste, Shreya Joshi, Srushtee Khadpekar, Shubhangi Mali,"AUTOMATED QUESTION GENERATOR SYSTEM USING NLP LIBRARIES",International Research Journal of Engineering and Technology (IRJET),vol. 07,Issue: 06,June 2020.
[8] Most popular video content type worldwide in 3rd quarter 2023, by weekly usage reach, https://www.statista.com/statistics/1254810/top-video-content-type-by-global-reach/.
[9] Size of the global e-learning market in 2019 and 2026, by segment(in billion U.S. dollars), https://www.statista.com/outlook/emo/online-education/worldwide.
[10] Aswin, VB & Javed, Dr. Mohammed & Parihar, Parag & Aswanth, K & Druval, CR & Dagar, Anpam & C V, Aravinda. “NLP Driven Ensemble Based Automatic Subtitle Generation and Semantic Video Summarization Technique”,2021.
[11] S. S. Alrumiah and A. A. Al-Shargabi, "Educational videos subtitles’ summarization using latent dirichlet allocation and length enhancement," Computers, Materials & Continua, vol. 70, no.3, pp. 6205–6221, 2022. https://doi.org/10.32604/cmc.2022.021780
[12] Alina Karakanta, François Buet, Mauro Cettolo, and François Yvon, 2 “Evaluating Subtitle Segmentation for End-to-end Generation Systems”, In Proceedings of the Thirteenth Language Resources and Evaluation Conference, pages 3069–3078, Marseille, France. European Language Resources Association.,2022.
[13] Ilampiray P, Naveen Raju D, Thilagavathy A, Mohamed Tharik M, Madhan Kishore S, Nithin A.S, Infant Raj ,"Video Transcript Summarizer",E3S Web Conf. 399 04015,2023.
[14] Yu Yuxiu, "Application of translation technology based on AI in translation teaching," Systems and Soft Computing, vol. 6, pp. 200072, Jan. 2024, https://doi.org/10.1016/j.sasc.2024.200072.
[15] M. Vaishnavi, H. R. Dhanush Datta, V. Vemuri, and L. Jahnavi, "Language Translator Application," International Journal for Research in Applied Science & Engineering Technology (IJRASET), vol. 10, no. 7, pp. 45.98, Jul. 2022.
[16] Puneeth Thotad , Shanta Kallur , Sukanya Amminabhavi,“Automatic Question Generator Using Natural Language Processing”, Journal of Pharmaceutical Negative Results, pp. 2759–2764, Dec. 2022.
[17] A. M. Agrawal, A. Atri, A. Chowdhury, R. Koneru, K. A. Batchu, and S. C. R. Mallavaram, “Question Answering System Using Natural Language Processing”, IJRESM, vol. 4, no. 12, pp. 16–19, Dec. 2021.
[18] Maram Almotairi,Fethi Fkih,"Developing a Semantic Question Answering System for E-learning Environments using Linguistic Resources",Journal of Education and e-Learning Research,vol. 9, no. 4,pp. 224-232, 2022.
[19] Sarah S. Alrumiah, Amal A. Al-Shargabi,"Educational Videos Subtitles’ Summarization Using Latent Dirichlet Allocation and Length Enhancement",Computers, Materials & Continua, vol. 70(3), pp. 6205-6221,2022,https://doi.org/10.32604/cmc.2022.021780.
[20] T. Wangchen, P. N. Tharindi, K. C. C. Chaveena De Silva, W. D. Thushan Sandeepa, N. Kodagoda and K. Suriyawansa, "EDUZONE – A Educational Video Summarizer and Digital Human Assistant for Effective Learning," 2022 7th International Conference on Information Technology Research (ICITR), Moratuwa, Sri Lanka, pp. 1-6,2022.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)