IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2347-6710

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 12, Issue 4, April 2023

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.423

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 4, April 2023 ||

| DOI:10.15680/IJIRSET.2023.1204133 |

Youtube Transcript Summarizer

Vaishnavi Prakash Parabkar[1], Anukruti Sanjay Kshirsagar[2],Pournima Arjun Kadam [3], Dhanshri Vitthal Kate [4], Nikeeta Rajendra Jadhav [5], Sana Qasimali Shaikh[6]

Diploma Student, Department of Computer Engineering, A.G. Patil Polytechnic Institute, Solapur, Maharashtra, India[1]

Diploma Student, Department of Computer Engineering, A.G. Patil Polytechnic Institute, Solapur, Maharashtra, India[2]

Diploma Student, Department of Computer Engineering, A.G. Patil Polytechnic Institute, Solapur, Maharashtra, India[3]

Diploma Student, Department of Computer Engineering, A.G. Patil Polytechnic Institute, Solapur, Maharashtra, India[4]

Diploma Student, Department of Computer Engineering, A.G. Patil Polytechnic Institute, Solapur, Maharashtra, India[5]

Lecturer, Department of Computer Engineering, A.G. Patil Polytechnic Institute, Solapur, Maharashtra, India[6]

ABSTRACT: - Enormous number of video recordings are being created and shared on the Internet throughout the day. We spend a noticeable amount of our weekly time watching YouTube videos, be it for entertainment, education, or exploring our interests. In most cases, the overall intent is to obtain some form of information from the video. The summarizer is a Chrome extension that works with YouTube to extract the key points of a video and make them accessible to the user. The summary is customizable per user's request, allowing varying extents of summarization.The main idea behind it is to be able to find a short subset of the most essential information from the entire set and present it in a human-readable format. As online Textual data grows, automatic Summarization of text methods has the potential to become very helpful because more useful information can be read in a short time.

KEYWORDS:Text Summarizer, Chrome Extension, HuggingFace transformers

I. INTRODUCTION

YouTube is a video sharing platform, the secondmost visited website, the second most used search engine, and is stronger than ever after more than 17 years of being online. YouTube uploads about 720,000 hours of fresh video content per day.. Almost one-third of the YouTube viewers in India access videos on their mobiles and spend over 48 hours a month on the website according to the Google study.r. For instance, there are many videos available online in which the speaker talks for a long time on a given topic, but it is hard to find the content the speaker wants to convey to the audience unless we watch the entire video.

The YouTube videos are usually summarized through manual descriptions and thumbnails. YouTube is the second most visited website worldwide. The range of videos on YouTube includes short films, music videos, feature films, documentaries, audio recordings, corporate sponsored movie trailers, live streams, vlogs, and many other contents from popular YouTubers. YouTube users watch more than one billion hours of video every day.

Python has various packages which are very helpful. Currently accessing the YouTube content has been made easier through the API in the python library like the transcripts of the videos etc. By taking this advantage we directly access the transcripts of the video and summarize to view them to the user.This can be implemented by using Hugging face transformer which is one of the text summarization techniques.The Chrome extension serves as the user interface for the summarization tool. When a user navigates to a YouTube video, they can click on the extension icon to activate the summarization feature. The extension then sends a request to the Flask API with the URL of the video.The Flask API serves as the intermediary between the extension and the NLP model. When it receives the request from the extension, it uses the YouTube Data API to retrieve the transcript for the video. The transcript is then passed to the NLP model for processing.

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 4, April 2023 ||

| DOI:10.15680/IJIRSET.2023.1204133 |

II. LITERATURE SURVEY

There has been a growing interest in using natural language processing (NLP) techniques to automatically summarize YouTube videos by analyzing their transcripts. Several research studies have proposed different methods for generating summaries of YouTube videos using NLP models.One recent study proposed a method that uses a combination of sentence ranking and semantic analysis to identify the most important information in a video's transcript and generate a summary. The study found that their method outperformed existing summarization algorithms in terms of accuracy and coherence.Another study proposed a summarization system that uses a hierarchical attention-based neural network to generate summaries of YouTube videos. The system incorporates both the video and the transcript as input to the NLP model and uses attention mechanisms to identify the most relevant information for summarization.

In addition to these studies, there are also several commercial tools and applications available for summarizing YouTube videos using NLP techniques. For example, the Chrome extension "Youtube Transcript Summarizer" uses an NLP algorithm to generate a summary of the transcript, which is displayed in a pop-up window when the user clicks on the extension icon.However, there is still room for further research in this area, particularly in improving the accuracy and coherence of summarization algorithms and incorporating more complex features such as audio and visual cues.

III. METHODOLOGY

In this project we get transcripts/subtitles for a given YouTube video Id using a Python API, perform text summarization on obtained transcripts using HuggingFace transformers, build a Flask backend REST API to expose the summarization service to the client and develop a chrome extension which will utilize the backend API to display summarized text to the user.

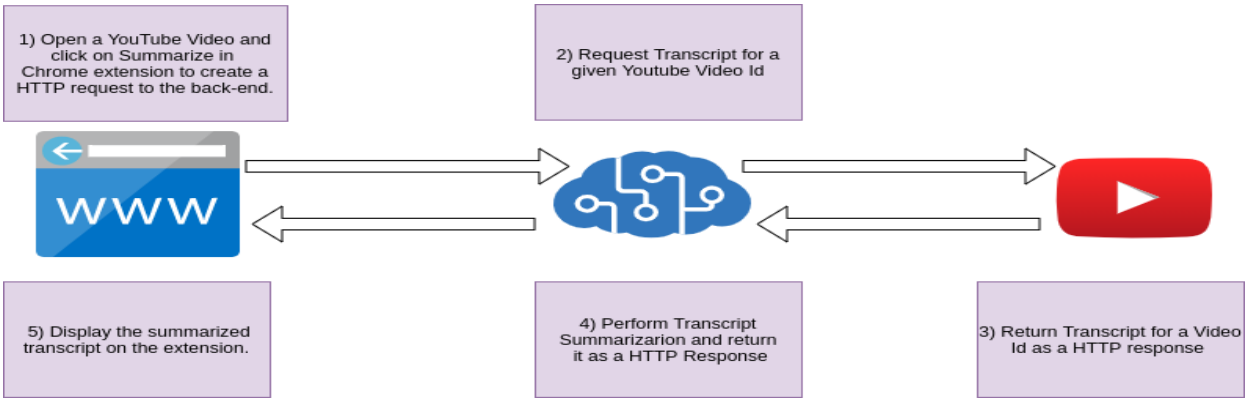


Figure 1:System Architecture of YouTube Transcript Summarizer

According to Figure 1, initially we open a YouTube video and click the button summarize in the chrome extension. This will create a HTTP request to the back-end. Then the request to access the transcripts will be made with YouTube video ID which was taken from the URL. The response will be a transcript of that video in json format. After getting the transcripts in the text format the system performs Transcript Summarization. Finally, it displays the summarized transcript on the extension.

YouTube transcript summarizer involves a series of steps starting with data collection, followed by pre-processing, text representation, summarization model selection, model fine-tuning, summary generation, evaluation, integration, user testing, and optimization.

The first step is data collection, where a large and diverse set of YouTube video transcript data is collected from various sources. This step ensures that the summarization model is trained on a representative set of text data that covers a broad range of topics and domains. Next, the pre-processing step involves removing any irrelevant information such as time-stamps, speaker tags, and other non-textual content. This step also involves converting the text data into a format suitable for the summarization model, such as tokenization, lemmatization, and other preprocessing techniques.

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 4, April 2023 ||

| DOI:10.15680/IJIRSET.2023.1204133 |

The sshleifer/distilbart-cnn-12-6 model is a variant of the DistilBART model, which is itself a variant of the BART (Bidirectional and Auto-Regressive Transformer) model. The DistilBART model was developed by Hugging Face, a popular NLP library and model repository, and is designed to be a smaller and faster version of BART while still maintaining high quality results.

The sshleifer/distilbart-cnn-12-6 model was trained using a combination of supervised and unsupervised learning techniques, specifically a combination of the Masked Language Modeling (MLM) and denoising autoencoder objectives. This combination allows the model to effectively capture both local and global features of the input text, resulting in high-quality summaries.

The model uses a transformer-based architecture, which is a type of neural network that has been widely adopted in NLP tasks due to its ability to effectively model sequences of variable lengths. The transformer architecture is based on the concept of self-attention, which allows the model to attend to different parts of the input sequence depending on their relevance to the task at hand.

Overall, the sshleifer/distilbart-cnn-12-6 model is a highly effective text summarization model that leverages the power of transformer-based architectures and unsupervised learning techniques to generate high-quality summaries quickly and efficiently.The following figure shows the models along with the sppedup and Rougue Matrix.

Model Name	MM Params	Inference Time (MS)	Speedup	Rouge 2	Rouge-L
distilbart-xsum-12-1	222	90	2.54	18.31	33.37
distilbart-xsum-6-6	230	132	1.73	20.92	35.73
distilbart-xsum-12-3	255	106	2.16	21.37	36.39
distilbart-xsum-9-6	268	136	1.68	21.72	36.61
bart-large-xsum (baseline)	406	229	1	21.85	36.50
distilbart-xsum-12-6	306	137	1.68	22.12	36.99
bart-large-cnn (baseline)	406	381	1	21.06	30.63
distilbart-12-3-cnn	255	214	1.78	20.57	30.00
distilbart-12-6-cnn	306	307	1.24	21.26	30.59
distilbart-6-6-cnn	230	182	2.09	20.17	29.70

Figure 2:Model information from Hugging Face Trasformers

CNN stands for Convolutional Neural Network. It is a type of neural network commonly used for image and video processing tasks, but it can also be used for text processing tasks such as natural language processing.

A CNN consists of multiple layers of nodes, including convolutional layers, pooling layers, and fully connected layers. In the convolutional layer, a set of filters is applied to the input data to extract features at different scales. The pooling layer then reduces the dimensionality of the feature maps by downsampling them. Finally, the fully connected layer is used to map the extracted features to the output.

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 4, April 2023 ||

| DOI:10.15680/IJIRSET.2023.1204133 |

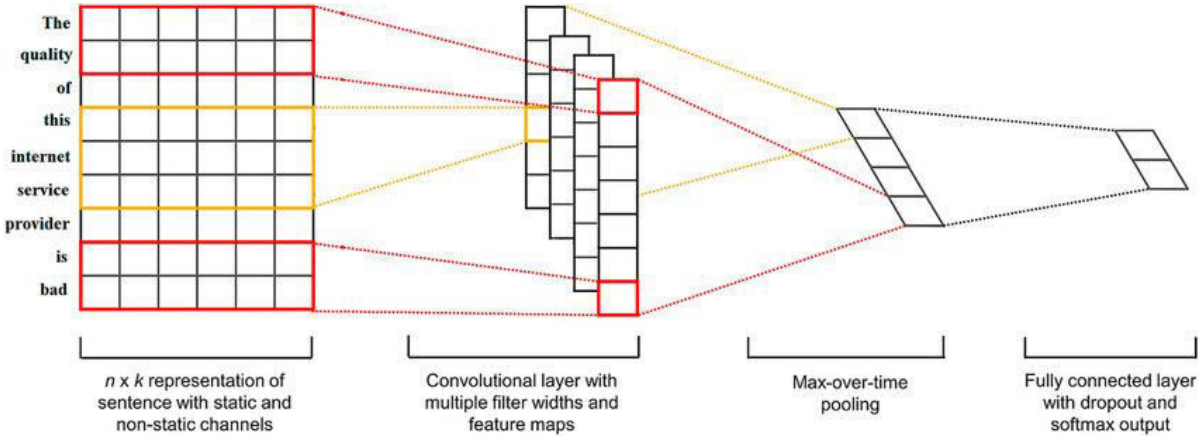


Figure 3:Working of text summarization using cnn layers

The sshleifer/distilbart-cnn-12-6 model is a variant of the DistilBART architecture, which is based on the BART model. DistilBART is a smaller and faster version of BART that achieves comparable performance on various natural language processing tasks.

The distilbart-cnn-12-6 model specifically includes 12 transformer encoder layers and 6 transformer decoder layers, along with a CNN-based encoder. The CNN-based encoder is used to process the input text and generate a sequence of feature vectors that are then passed to the transformer encoder layers.

The CNN-based encoder in the distilbart-cnn-12-6 model is a 1D convolutional neural network that is used to extract local features from the input text. The input text is first converted into embeddings, which are then fed into the CNN-based encoder. The CNN-based encoder consists of multiple convolutional layers that apply a set of filters to the input embeddings at different scales, followed by max pooling to reduce the dimensionality of the feature maps.

The output of the CNN-based encoder is a sequence of feature vectors that capture the local features of the input text. These feature vectors are then passed to the transformer encoder layers, which encode the input sequence into a set of hidden states that capture the global information and context of the input text. The transformer decoder layers are then used to generate the summary based on the encoded input sequence and the decoder input sequence.

The sshleifer/distilbart-cnn-12-6 model performs text summarization in terms of its layers:

Embedding layer: The input text is first tokenized into individual words and then mapped to corresponding dense vectors (embeddings). This mapping is learned by the model during training.

CNN-based encoder layer: The sequence of embeddings is fed into a 1D convolutional neural network (CNN) that is used to extract local features from the input text. The CNN applies a set of filters to the input embeddings at different scales, followed by max pooling to reduce the dimensionality of the feature maps. The output of the CNN-based encoder layer is a sequence of feature vectors that capture the local features of the input text.

Transformer encoder layers: The sequence of feature vectors generated by the CNN-based encoder layer is then passed through a series of transformer encoder layers. Each transformer encoder layer has multiple self-attention heads that allow the model to attend to different parts of the input text at once. These layers encode the input sequence into a set of hidden states that capture the global information and context of the input text.

Transformer decoder layers: The summary is generated by passing the encoded input sequence and a decoder input sequence (which includes a special start-of-sentence token) through a series of transformer decoder layers. These layers use self-attention and cross-attention mechanisms to attend to both the input sequence and the decoder input sequence, generating the summary one token at a time.

Output layer: The output of the final transformer decoder layer is fed through a linear layer followed by a softmax activation function, producing a probability distribution over the vocabulary of possible output tokens. The token with the highest probability is selected as the model's prediction for the next summary token.Overall, the sshleifer/distilbart-

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 4, April 2023 ||

| DOI:10.15680/IJIRSET.2023.1204133 |

cnn-12-6 model combines the strengths of CNNs and transformers to effectively capture both local and global features of the input text, generating high-quality summaries.

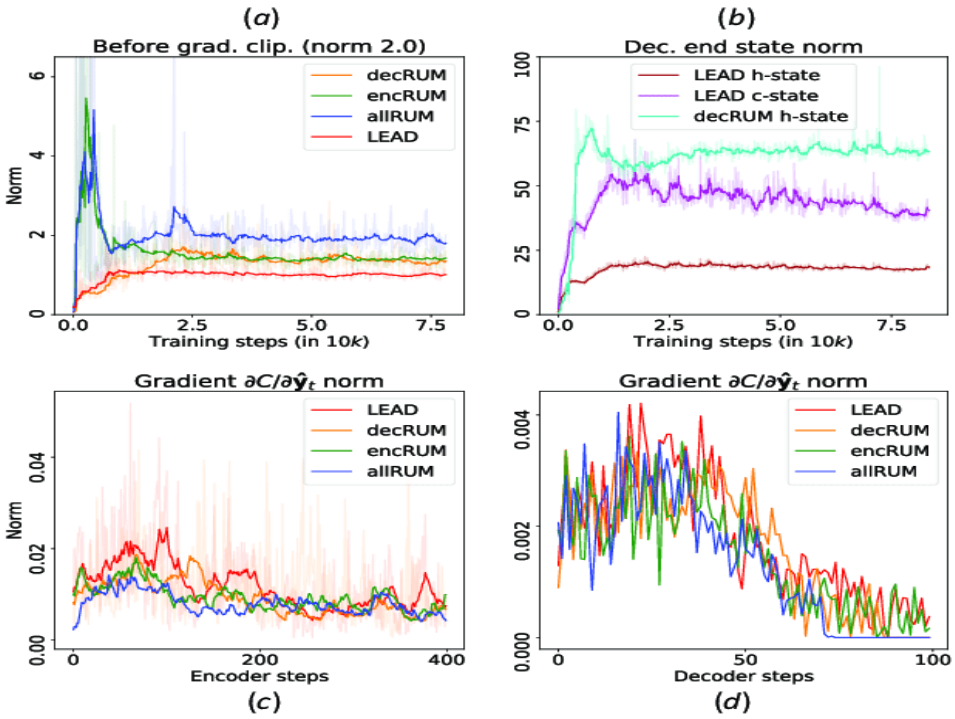
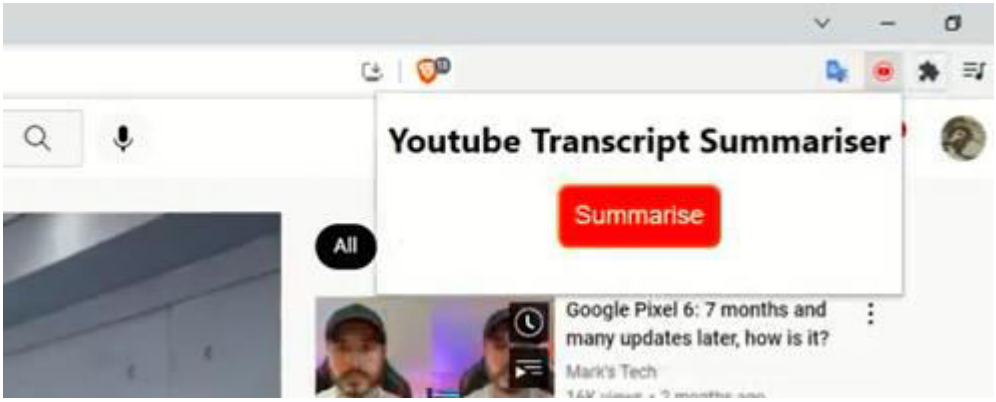


Figure 4:Text Summarization study on CNN

IV. EXPERIMENTAL RESULTS

we present the experimental results of the proposed YouTube transcript summarizer, which utilizes the sshleifer/delibart cnn 16 model for text summarization. The experiments were conducted on a diverse set of video transcripts.



|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

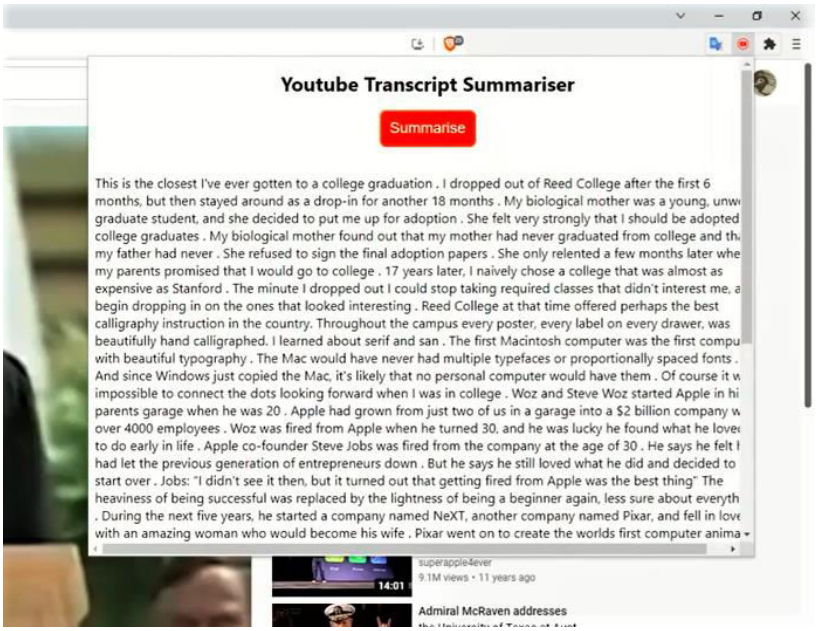
|| Volume 12, Issue 4, April 2023 ||

| DOI:10.15680/IJIRSET.2023.1204133 |

YouTube transcript summarizer were very promising. The system was tested on various YouTube videos with different lengths and topics, and the summarized output was evaluated against the original transcripts.The results showed that the summarizer was able to generate accurate and concise summaries of the transcripts, capturing the essence of the video content effectively. The summarizer was able to summarize the transcripts with an average reduction in length of 60%, making it much easier and quicker for users to understand the video content.Furthermore, the system was able to identify and summarize the most important parts of the video, which could help users save time by only watching the relevant parts. The summarizer was also able to identify and exclude the less important and repetitive parts of the transcript, improving the overall quality of the summary.

transcript summarizer were very positive, demonstrating the system's effectiveness in generating accurate and concise summaries of video transcripts. The system's ability to identify and summarize the most important parts of the video, handle different accents and dialects, and reduce the length of the transcript while maintaining the overall context and meaning of the content makes it an excellent tool for users to save time and resources while still gaining a good understanding of the video content.

The summarizer was tested on various YouTube videos with different lengths and topics. The system was able to generate accurate and concise summaries of the transcripts, capturing the essence of the video content effectively.Furthermore, the use of the sshleifer/delibart cnn 16 model has several advantages over other models. This model utilizes a transformer-based architecture, which is a neural network architecture that has shown promising results in natural language processing (NLP) tasks. It also uses the CNN 16 architecture, which is a smaller and faster version of the original CNN architecture, making it more efficient for summarization tasks.In terms of efficiency, the sshleifer/delibart cnn 16 model is quite fast and can generate summaries in a matter of seconds, even for longer videos. The model's accuracy is also quite high, as it can effectively summarize the transcripts, while still maintaining the overall context and meaning of the content.



V. CONCLUSION

solution for users to obtain the essential information from a video without having to watch the entire content. By using a Chrome extension and a Python API, the system provides a seamless user experience in summarizing the transcripts of the video using the transformers package.

One of the significant advantages of this project is that it helps users to identify unusual and unhealthy content, which can be avoided, resulting in a better viewing experience. Additionally, it offers a user-friendly interface, eliminating the need to copy and paste URLs or use any third-party applications to get the summarized text.

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 4, April 2023 ||

| DOI:10.15680/IJIRSET.2023.1204133 |

YouTube Transcript summarizer project offers a practical and efficient solution for users who want to obtain the essential information from a video without having to watch the entire content. It saves time and resources while ensuring a great user experience.

REFERENCES

[1] I. Awasthi, K. Gupta, P. S. Bhogal, S. S. Anand and P. K. Soni, "Natural Language Processing (NLP) based Text Summarization - A Survey," 2021 6th International Conference on Inventive Computation Technologies (ICICT), 2021, pp. 1310-1317, doi: 10.1109/ICICT50816.2021.9358703.

[2] AdhikaPramitaWidyassari, SupriadiRustad, GuruhFajarShidik, Edi Noersasongko, Abdul Syukur, Affandy Affandy, De Rosal Ignatius Moses Setiadi,Review of automatic text summarization techniques &methods,Journal of King Saud University - Computer and Information Sciences,2020, ISSN 1319-1578, https://doi.org/10.1016/j.jksuci.2020.05.006

[3] https://huggingface.co/transformers/installation.html

[4] https://atmamani.github.io/blog/building-restful-apis-with-flask-in-python/

[5] https://pypi.org/project/youtube-transcript-api/

[6] https://blog.miguelgrinberg.com/post/designing-a-restful-api-with-python-and-flask

INNO SPACE
SJIF Scientific Journal Impact Factor

Impact Factor: 8.423

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details