KNOWTUBE

YouTube Video Transcript Summarizer

PROBLEM STATEMENT

To provide abstractive and extractive methods to summarize youtube videos on same platform.

TEAM MEMBERS

Prajakta Ravindra Deokule
C22019221332

Chinmayi Sujit Adsul
C22019221304

Sonali Pramod Ingle
C22019221351

IDEA BEHIND THE PROJECT

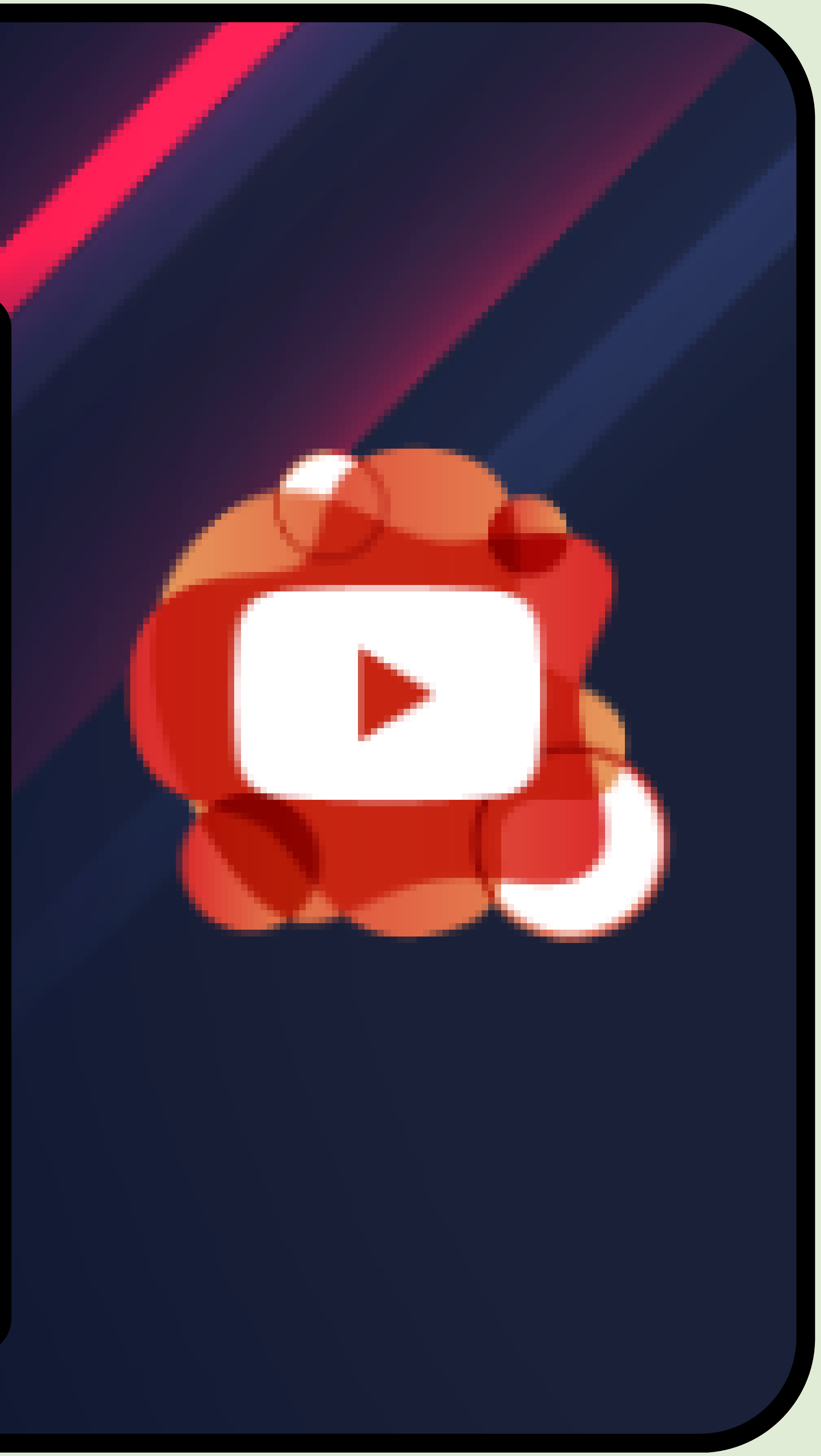
- Due to the pandemic, people's internet usage increased a lot. An enormous number of video recordings are created and shared on the internet everyday. It is tedious to spend time watching the entire video which may have a longer duration than expected and sometimes our efforts may become futile if we don't find relevant information in it.

- Summarizing transcripts of YouTube videos allows us to quickly look out for the important patterns in the video. It helps us to save time and effort to go through the whole content of the video.

OBJECTIVES

To build a model to summarize a youtube video transcript

- The user has to enter the video link he/she wants to summarize. The user is provided options to choose from the various summarization techniques.
- The Summarization techniques included are : BART(sshleifer/distilbart-cnn-12-6 , facebook/bart-large-cnn) for abstractive summarisation and NLTK, Gensim models for extractive summarisation.



Summarization Techniques?

Abstractive

Extractive

ABSTRACTIVE

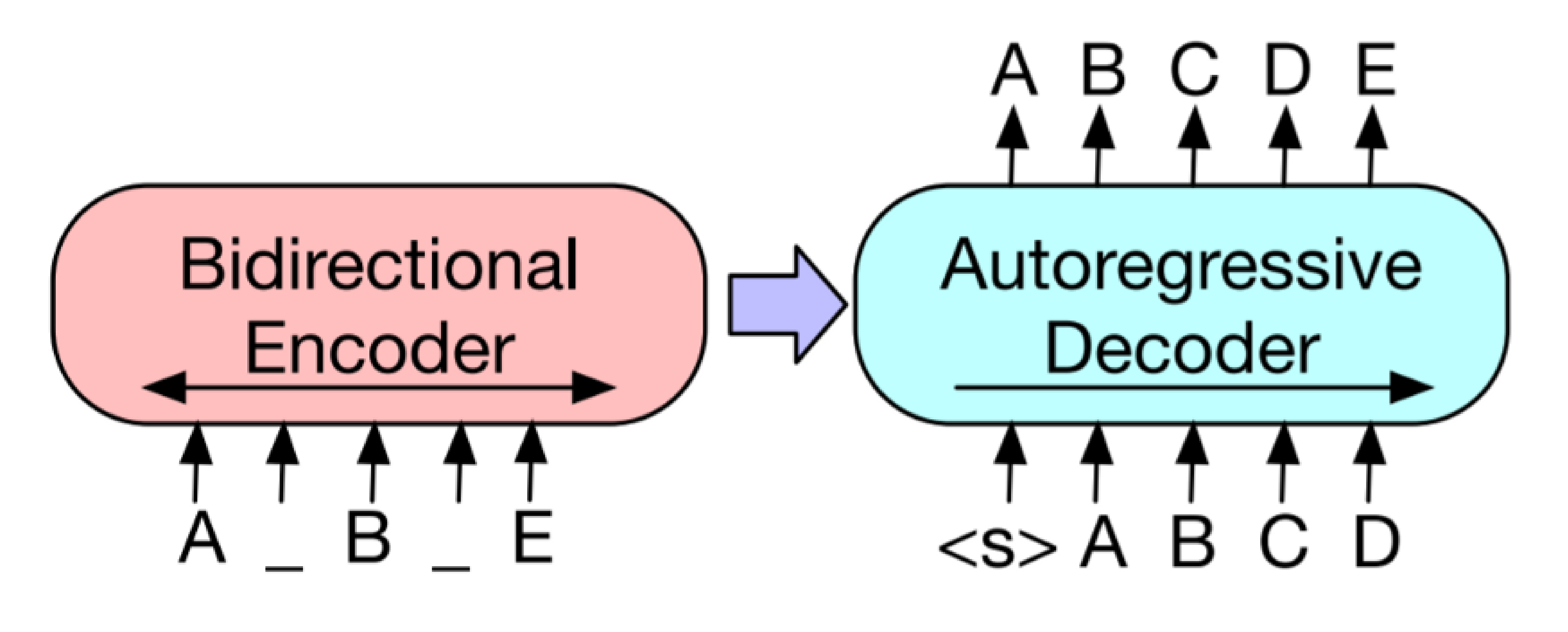
Abstractive summarization in Natural Language Processing generates a concise summary of the source text with its own words. This is more complex than extractive summarization as it involves paraphrasing.

Modules Used :

- BART
  - sshleifer/distilbart-cnn-12-6
  - facebook/bart-large-cnn

BART SEQUENCE-TO-SEQUENCE

- BART stands for Bidirectional Auto Regressive Transformers. It has both an encoder (like BERT) and a decoder (like GPT), essentially getting the best of both worlds.



Transformer's pipeline BART summarisation model

- Applying NLP operations from scratch becomes tedious since it requires various steps to be performed.
- HuggingFace is a startup in the Natural Language Processing (NLP) domain, which offers its library of models for use to a large number of organizations including Microsoft and AWS. It provides access to several pre-trained Transformer models with BART being just one of them.
- For the text summarization task, you can choose fine-tuned BART models from the HuggingFace model explorer website. You can find the configuration and training description of every model uploaded there.

PIPELINE

- The pipelines are a great and easy way to use models for inference.

- These pipelines are objects that abstract most of the complex code from the library, offering a simple API dedicated to several tasks, including Summarization, Named Entity Recognition, Masked Language Modeling, Sentiment Analysis, Feature Extraction and Question Answering.

Important Parameters

- **task (str) –Task defining which pipeline will be returned.** Currently accepted tasks are:

  "summarization": will return a SummarizationPipeline

  - "feature-extraction": will return a FeatureExtractionPipeline
  - "sentiment-analysis": will return a TextClassificationPipeline
  - "ner": will return a TokenClassificationPipeline
  - "question-answering": will return a QuestionAnsweringPipeline
  - "fill-mask": will return a FillMaskPipeline
  - "translation_xx_to_yy": will return a TranslationPipeline
  - "text-generation": will return a TextGenerationPipeline
- model (str or PreTrainedModel or TFPreTrainedModel, optional, defaults to None)
- tokenizer(str or PretrainedTokenizer,optional_defaults to None)

EXTRACTIVE

- The system will extract the important paragraphs and contents from the given passage and combine these extracted paragraphs to create the summarized text.

Modules Used :

- NLTK
- Gensim

- **NLTK**

  - NLTK is a leading platform for building Python programs to work with human language data.

  - NLTK has been called "a wonderful tool for teaching and working in, computational linguistics using Python," and "an amazing library to play with natural language."

  - **nltk.corpus** : The modules in this package provide functions that can be used to read corpus files in a variety of formats.

  - **nltk.tokenize** : NLTK tokenizers to tokenize the transcript

- **Gensim**

- Gensim package provides a method for text summarization.

- It is based on Text Rank algorithm and can handle large text collections

\-gensim.summarization.summarizer(text, ratio, word_count, split)

- gensim.summarization.textcleaner : The clean the text.

FUTURE SCOPE

- Improve user interface
- Deploy the model as application
- Compare summaries from all models
- Provide user best reliable summary by studying all types of summaries

THANK YOU