# DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

# DELHI TECHNOLOGICAL UNIVERSITY

## CO301: Software Engineering

## Project Report

## Courier Management System

**Submitted To:**

**Mr. Rohit Beniwal**
**Assistant Professor**

**Submitted By:**

**Harshit Chopra**
**(2K21/CO/184)**

**Himanshu**
**(2K21/CO/200)**

# ABSTRACTIVE SUMMARIZATION OF PODCAST TRANSCRIPTS

## ABSTRACT

This project aims to revolutionize podcast consumption by developing an innovative abstractive summarization system using the Spotify Podcast Dataset. Leveraging over 100,000 transcribed episodes, the solution employs an extractive module and a BART model for generating concise, human-readable summaries. The absence of ground truth summaries is addressed by utilizing creator-provided descriptions for training supervised models.

## INTRODUCTION

Podcasts have emerged as a popular medium for knowledge dissemination, but the overwhelming volume of content poses challenges for users in efficient discovery. This project focuses on enhancing the user experience by providing accurate and concise abstractive summaries, enabling users to quickly decide which podcasts to engage with. Leveraging the rich Spotify Podcast Dataset, the goal is to develop a system that aids users in making informed choices about the content they consume.

## PROBLEM STATEMENT

The challenge lies in the sheer volume of podcast content available, making it difficult for users to identify episodes that align with their interests. Traditional show notes and metadata often fall short in providing detailed and informative summaries, necessitating the development of an advanced summarization system.

## DATASET DESCRIPTION

The Spotify Podcast Dataset is a groundbreaking collection that encompasses more than 100,000 transcribed podcast episodes, offering a diverse range of content. This dataset includes not only raw audio files but also meticulously generated transcripts and accompanying metadata. The transcripts are obtained through the Google Cloud Platform's Speech-to-Text API, ensuring a high level of accuracy in capturing spoken content.

### TRANSCRIPTS

The transcriptions serve as a valuable resource for understanding the spoken content within each episode. These transcripts are not mere verbatim representations but are the result of advanced speech recognition technology, providing a textual representation of the podcast content.

### AUDIO FILES

In addition to transcripts, the dataset includes the original raw audio files of each podcast episode. This multimodal aspect introduces the possibility of exploring integrative approaches that consider both textual and audio features for a more comprehensive summarization model.

### METADATA

Accompanying each episode are metadata details that include information such as episode title, duration, publication date, and potentially relevant tags or categories. This metadata enriches the dataset, offering contextual information that can be leveraged during the summarization process.

### EPISODE DESCRIPTIONS

While no ground truth summaries are provided, the episode descriptions authored by the podcast creators serve as proxies for summaries. These descriptions, written with the intention of attracting listeners, provide a valuable source of training data for supervised models.

## USE CASE DIAGRAM

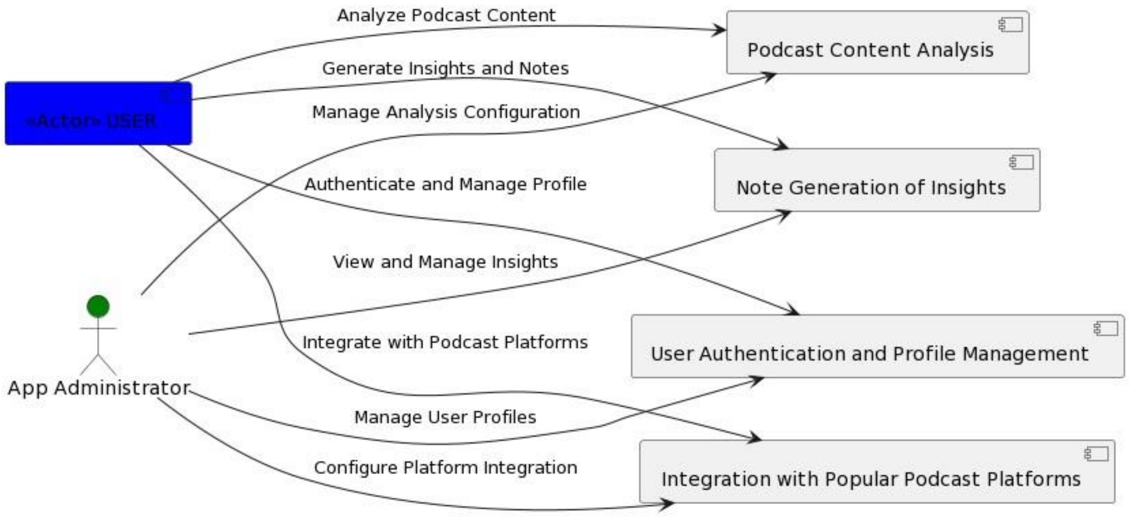
Analyze Podcast Content
Podcast Content Analysis
Generate Insights and Notes
«Actor» USER
Manage Analysis Configuration
Note Generation of Insights
Authenticate and Manage Profile
View and Manage Insights
Integrate with Podcast Platforms
User Authentication and Profile Management
App Administrator
Manage User Profiles
Configure Platform Integration
Integration with Popular Podcast Platforms

*Fig 1: Use case diagram of project*

# PROPOSED METHODOLOGY

The proposed methodology for abstractive summarization of podcast transcripts involves a two-step process: an Extractive Module to identify key segments from the transcript and an Abstractive Summarizer using a BART model for generating concise, human-readable summaries.

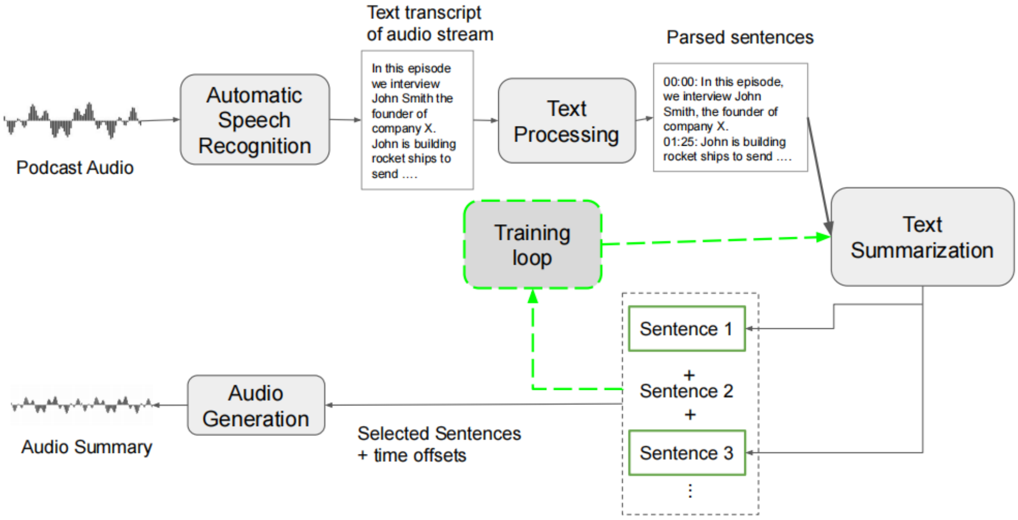
Text transcript of audio stream
Parsed sentences
Automatic Speech Recognition
In this episode we interview John Smith the founder of company X. John is building rocket ships to send ....
Text Processing
00:00: In this episode, we interview John Smith, the founder of company X.
01:25: John is building rocket ships to send ....
Podcast Audio
Training loop
Text Summarization
Sentence 1
+
Sentence 2
+
Sentence 3
⋮
Audio Generation
Audio Summary
Selected Sentences
+ time offsets

*Fig 2: Level 1-DFD*

## I. Extractive Module

The Extractive Module aims to select salient chunks from podcast transcripts, serving as input to the subsequent abstractive summarization step. This module involves the following steps:

### 1. **Pre-processing**

- Tokenization: The transcript is tokenized into individual words or sub-word units, creating a suitable input for further analysis.

- Stop-word Removal: Common stop-words are removed to focus on meaningful content.

### 2. **Feature Extraction**

- TF-IDF (Term Frequency-Inverse Document Frequency): Key terms are identified based on their importance in the context of the entire transcript.

- Sentence Embeddings: Utilizing embeddings, the module captures semantic information to identify essential sentences.

3. **Salient Chunk Selection**

- Using the extracted features, the module selects salient chunks, which represent important segments within the transcript.

- Sentence importance scores contribute to the decision-making process.

## II. Abstractive Summarizer (BART Model)

The Abstractive Summarizer employs a BART (Bidirectional and Auto-Regressive Transformers) model with an encoder-decoder architecture. This model has proven effective in various natural language generation tasks. The steps involved in the Abstractive Summarizer are as follows:

1. **Data Preparation**

- The salient chunks identified by the Extractive Module, along with additional context from the transcript, are prepared as input sequences for training the BART model.

2. **Model Architecture**

- Encoder-Decoder Structure: The BART model consists of an encoder to process the input sequence and a decoder to generate the abstractive summary.

- Attention Mechanism: Attention mechanisms allow the model to focus on relevant parts of the input during both encoding and decoding.

3. **Training**

- Supervised Learning: The model is trained using a dataset where the input is the selected salient chunks, and the target is the corresponding abstractive summary.

- Fine-tuning: Hyperparameters are fine-tuned to optimize the model's performance on podcast summarization.

4. **Inference**

- During inference, the trained model takes new transcript data and generates abstractive summaries.

- Beam Search: The decoding process incorporates beam search to explore multiple potential summaries and select the most suitable one.

## Model Integration

The Extractive Module and Abstractive Summarizer are seamlessly integrated, creating a comprehensive system for podcast summarization. The selected salient chunks act as input to the Abstractive Summarizer, ensuring that the generated summaries are based on key content identified in the extraction phase.

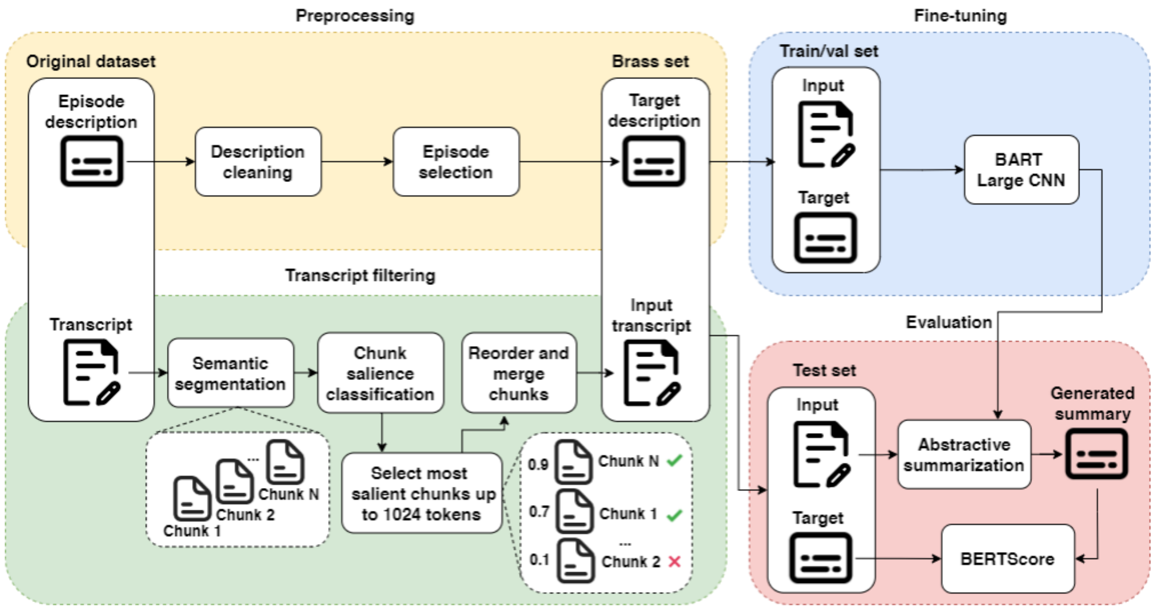

*Fig 3: Proposed Methodology*

# ADVANTAGES

### 1. Content Accuracy

The combination of an extractive module and an abstractive summarizer enhances the accuracy of content representation in the generated summaries. The extractive module identifies salient chunks, ensuring that the abstractive summarizer focuses on key content. This two-step process contributes to the precision and relevance of the summaries, providing users with an accurate reflection of the podcast content.

### 2. Human-Readability

The abstractive summarization process using a BART model is designed to produce concise and easily understandable summaries. This human-readable output is crucial for users who seek quick insights into podcast episodes without delving into lengthy transcripts. The summarizer's ability to capture the essence of the content in a coherent and succinct manner contributes to an enhanced user experience.

### 3. Dataset Utilization

The project leverages the extensive Spotify Podcast Dataset, a resource with over 100,000 transcribed episodes, providing a diverse set of podcasts for training and testing. The utilization of this large-scale dataset enhances the model's ability to generalize across various podcast genres, styles, and topics, ensuring robust performance in real-world scenarios.

## CHALLENGES

### 1. Lack of Ground Truth Summaries

One major challenge is the absence of ground truth summaries in the dataset. To overcome this, the project relies on episode descriptions provided by podcast creators as proxies for summaries. However, this introduces potential biases as creators may emphasize certain aspects in their descriptions, impacting the model's training and generalization.

### 2. Model Complexity

Fine-tuning the BART model for optimal performance is a non-trivial task. The complexity lies in determining the appropriate hyperparameters, training strategies, and balancing the trade-off between model size and computational efficiency. Rigorous experimentation and tuning are essential to ensure the model's effectiveness in generating high-quality summaries.

### 3. Multimodal Data Handling

The integration of raw audio files, transcripts, and metadata introduces challenges in data preprocessing and model integration. Handling multimodal data requires careful consideration of alignment between audio and textual features. Ensuring a seamless integration of these diverse data types is crucial for the overall success of the abstractive summarization system.

## IMPACT ON BUSINESS OPERATIONS

### 1. Enhanced User Experience

Implementing abstractive summarization significantly improves the efficiency of podcast discovery, providing users with concise summaries that aid in decision-making. This enhanced user experience can attract and retain users, fostering increased engagement with the podcast platform.

### 2. Content Monetization:

Precise summarization contributes to targeted advertising and content recommendation strategies. By understanding the content at a deeper level, the platform can deliver more relevant advertisements and recommendations to users. This targeted approach enhances revenue streams through increased ad effectiveness and user engagement.

### 3. Business Intelligence:

The generated summaries can be utilized for business intelligence purposes, extracting valuable insights into user preferences and popular podcast topics. This data can inform content creators, advertisers, and platform administrators, enabling strategic decision-making and content optimization.