BERT based summarizer for Audio Podcast transcripts, combined with Topic modeling for user recommendation.

Team 8

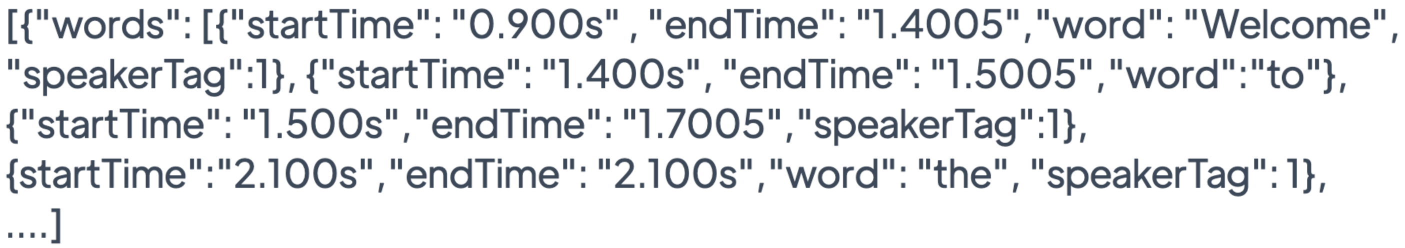
Introduction

- Podcasts are one of the **fastest growing content** available on the internet for consumption.
- Audio podcasts, while entertaining and informative, are usually ignored as text mining content because of the **conversational nature**.
- Our aim is to **summarize the transcripts and model the topics** covered in the transcript.
- This data could be used in recommendation systems, consolidate podcast content from multiple podcasts on the same topic, and other such applications.

Data

Transcript(as text)

- Available in JSON format
- Generated using Google Cloud Platform’s Cloud Speech-to-Text API3(GCP-ASR)
- On average, over 6000 words per episode
- Word Error rate of 18.1%
- Format:



```
[{"words": [{"startTime": "0.900s" , "endTime": "1.4005","word": "Welcome",
"speakerTag":1}, {"startTime": "1.400s", "endTime": "1.5005","word":"to"},
{"startTime": "1.500s","endTime": "1.7005","speakerTag":1},
{startTime":"2.100s","endTime": "2.100s","word": "the", "speakerTag": 1},
....]
```

Data

Audio

- Spotify Podcast Dataset

  Over 100,000 episodes of Audio podcasts data

  Size of the dataset: 2TB

  Median Length of episodes: 31.6 minutes

- Listennotes API

  Over 1,000,000 podcasts available to scrape

  Podcast audio as mp3 format

Speech to Text

- Open-source Speech-to-text APIs such as:
  - **Whisper** – by OpenAI
  - **GCP Speech-to-text** – by Google
  - **DeepSpeech** – by Mozilla
  - **SpeechBrain** – by PyTorch
- We can use these to build transcripts for audio with incoherent transcripts or no transcript at all.

Summarizing Podcast Episodes

Summarizing the content of podcast episodes can be very useful for listeners in several ways:

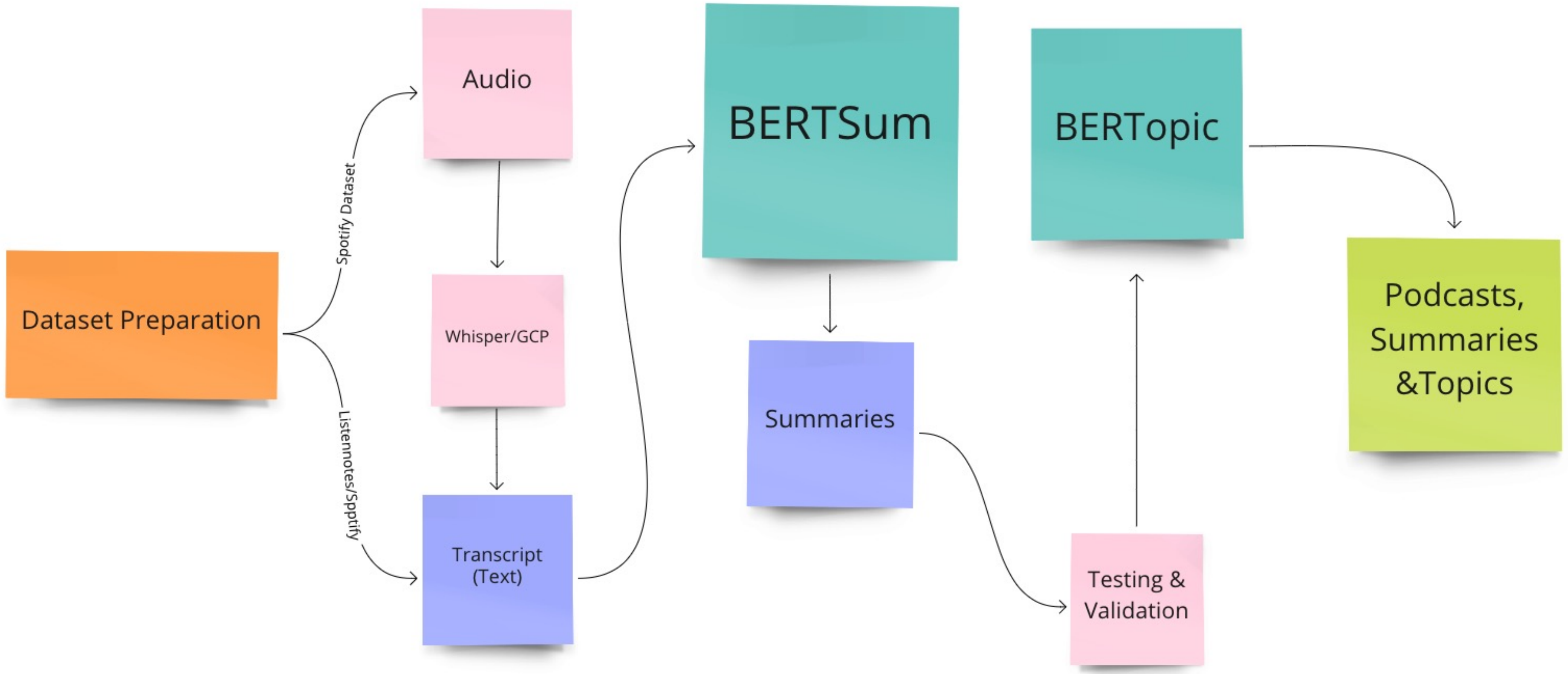
- **Saves Time:** This is particularly useful for busy individuals who may not have the time to listen to a full podcast episode.
- **Helps in decision making:** By reading a summary, listeners can quickly determine if the content of an episode is relevant to their interests and needs.
- **Better retention of information:** This is particularly helpful when the episode covers complex or technical topics, and a summary can break down the information into more easily digestible parts.

Topic analysis of Podcast episodes

Topic modeling is typically done using statistical algorithms that analyze the frequency and distribution of words and phrases within a dataset. It allows us to discover the underlying themes or topics that are present in a collection of documents or texts.

- **Identifying relevant topics for listeners:** By analyzing the language used in podcast episode titles and descriptions, topic modeling algorithms can identify the main topics and themes covered in each episode. This can help listeners quickly find podcasts that cover topics of interest to them.

Workflow


Dataset Preparation
Spotify Dataset
Listennotes/Spptify
Audio
Whisper/GCP
Transcript (Text)
BERTSum
Summaries
Testing & Validation
BERTopic
Podcasts, Summaries &Topics

References

- https://arxiv.org/pdf/2004.04270.pdf - Spotfiy Podcast Dataset
- https://arxiv.org/pdf/2203.05794.pdf - BERTopic: Neural topic modeling with a class-based TF-IDF procedure
- https://arxiv.org/pdf/1903.10318.pdf - Fine-tune BERT for Extractive Summarization