AN ABSTRACTIVE APPROACH TO PODCAST TEXT SUMMARIZATION

FATHIMA AREEFA THASSIM

MSC

2022

Abstract

Podcasts are a medium of entertainment becoming increasingly popular. When transcribed they are a rich source of data for natural language processing tasks. Podcasts are diverse in structure, are of varying length and tend to explore topics in a conversational manner. Text summarization is a field of NLP that is complex due to the difficulty in generating summaries that are concise and grammatically correct. These factors cause generating abstractive summaries more difficult.

This research aims to generate abstractive summaries with minimal grammatical errors and minimal redundant data for podcast transcripts using deep learning models. The state-of-the-art model, GPT-Neo has been finetuned to achieve this. The research solution includes a Flask API to use the GPT-Neo finetuned model and a ReactJS based web application for efficient summarization.

Evaluation of the research model consisted of manual evaluation of generated summaries and the ROUGE score. Manual evaluation concluded that most summaries were concise and grammatically correct. However, around 20% of resulting summaries faced faults such as repetition of phrases. The ROUGE score analysis showcased that a satisfactory ROUGE-1 metric was achieved exceeding existing research.

Overall, the research provides a novel design and approach for summarization of podcast transcripts, with interesting results which could be used to further extend research in this domain.

Keywords – podcasts, text summarization, NLP, GPT-Neo