Exploring Abstractive Text Summarisation for Podcasts: A Comparative Study of BART and T5 Models

Parth Saxena **Mahmoud El-Haj**
School of Computing and Communications
Lancaster University
Lancaster, United Kingdom
parth.s.1909@gmail.com m.el-haj@lancaster.ac.uk

Abstract

Podcasts have become increasingly popular in recent years, resulting in a massive amount of audio content being produced every day. Efficient summarisation of podcast episodes can enable better content management and discovery for users. In this paper, we explore the use of abstractive text summarisation methods to generate high-quality summaries of podcast episodes. We use pre-trained models, BART and T5, to fine-tune on a dataset of Spotify's 100K podcast. We evaluate our models using automated metrics and human evaluation, and find that the BART model fine-tuned on the podcast dataset achieved a higher ROUGE-1 and ROUGE-L score compared to other models, while the T5 model performed better in terms of semantic meaning. The human evaluation indicates that both models produced high-quality summaries that were well received by participants. Our study demonstrates the effectiveness of abstractive summarisation methods for podcast episodes and offers insights for improving the summarisation of audio content.

1 Introduction

Podcasts are a rapidly growing and popular medium for consuming knowledge and entertainment through spoken audio files that can be streamed or downloaded. With the podcast industry reaching new heights, there is a need for innovative and computationally effective methods for processing, analysing, and summarising podcast content. This need is further highlighted by the fact that Spotify[1] acquired Anchor and Gimlet Media, two leading podcast companies, for $340 million in 2018 and has since invested approximately $500 million in this industry, demonstrating the growth and importance of podcasting (Sullivan, 2019).

The abundance of data generated in various forms worldwide necessitates the need for methods to compress and understand this data, enabling the discovery of content available globally. One of these methods is automatic text summarisation, a technique used to shorten lengthy input texts to small coherent texts that convey the original meaning, reducing reading time and facilitating faster processing of research information. The field of Natural Language Processing (NLP) has witnessed significant progress in the development and research of text summarisation, resulting in an active research area in both Information Retrieval (IR) and NLP. The successful integration of transformer architecture and attention mechanisms has also contributed significantly to the advancement of text summarisation techniques.

The podcast domain presents unique challenges in text summarisation due to its conversational nature and informal language use, which could cause difficulties in extracting salient information. Furthermore, podcast episodes can be lengthy, making it a daunting task for listeners to find and identify those of interest. Summarising podcast transcripts into short, readable text can alleviate these issues, allowing podcast listeners to make more informed decisions on which episodes to listen to and enabling easier retrieval of information (Tulley, 2011).

The two fundamental approaches in text summarisation are extractive and abstractive summarisation. Extractive summarisation involves selecting the most relevant sentences from the input text to form a summary, while abstractive summarisation involves generating new sentences to capture the content of the input. Although both approaches can be used to summarise podcast transcripts, abstractive text summarisation is more challenging and requires the use of sophisticated techniques (El-Haj, 2012).

This paper focuses on generating automated summaries of podcast episodes using abstractive text summarisation to provide users with a concise summary of the podcast's content. The paper aims to

[1] www.spotify.com

Proceedings of Recent Advances in Natural Language Processing, pages 1023–1033
Varna, Sep 4–6, 2023
https://doi.org/10.26615/978-954-452-092-2_110

explore how state-of-the-art NLP (SOTA) models can generate summaries that convey the essence of the podcast and investigate the effectiveness of these summaries for podcast listeners. To achieve these objectives, the paper presents a systematic review of the background research on summarisation, the podcast domain, and related work in this field. It also analyses a dataset of podcast transcripts to develop a conceptual framework for the methods and techniques applied in this study. Finally, the paper examines the results and findings of the study, including the state-of-the-art evaluation metrics adopted for this study, and presents a conclusion that discusses the study's limitations, potential future work, and reflection on the study.

2 Background

Recent years have seen significant advancements in the field of automatic text summarisation, with the development of new techniques and models for generating summaries. Two main approaches are commonly used in text summarisation: extractive and abstractive summarisation. Extractive summarisation involves selecting the most relevant sentences from the source text, while abstractive summarisation generates a summary by rephrasing the text into new sentences. While extractive summarisation can be simpler and more efficient, abstractive summarisation has the potential to produce more informative and coherent summaries (Zmandar et al., 2021; El-Haj et al., 2010).

In recent years, the use of deep learning models, such as transformers, has led to significant improvements in the quality of abstractive text summarisation. Models such as BART, T5, and GPT-3 have achieved state-of-the-art performance on summarisation tasks, demonstrating the potential of these models for generating high-quality summaries. These models are pre-trained on large amounts of text data and can be fine-tuned on task-specific datasets to generate summaries that capture the essential content of the source text (Lewis et al., 2019; Xue et al., 2021; Brown et al., 2020)

The task of summarising podcasts has gained attention in recent years, with several studies proposing approaches for automated podcast summarisation. Laban et al. (2022) proposed an interactive summarisation approach for news podcasts, which allows users to engage with the summarisation process by providing feedback. This approach uses extractive summarisation to select important sentences from the podcast transcript and then generates a summary from these sentences. The system then allows users to rate the summary and provide feedback, which is used to refine the summary for future users.

Vartakavi et al. (2021) proposed an extractive summarisation approach for podcast episodes, where the summary is generated by selecting the most important sentences from the podcast transcript. They used a graph-based ranking algorithm to score each sentence based on its importance, and then selected the top sentences to form the summary. The system was evaluated on a dataset of 100 podcast episodes and achieved a ROUGE-2 score of 0.26.

Risne and Siitova (2019) introduced both extractive and abstractive summarisation approaches for news and podcast data using transfer learning. They fine-tuned BERT and GPT-2 models on a dataset of news articles and podcast transcripts to perform extractive and abstractive summarisation, respectively. The authors reported that the abstractive model outperformed the extractive model in terms of ROUGE scores on both news and podcast data.

Vartakavi and Garg (2020) presented an extractive summarisation approach for podcast episodes that uses sentence embeddings to score the importance of each sentence. The system selects the top sentences based on their scores to generate the summary. The authors evaluated their system on a dataset of 400 podcast episodes and achieved a ROUGE-2 score of 0.30.

Karlbom (2021) proposed an abstractive summarisation approach for podcast episodes, which uses a transformer-based model to generate summaries. The system was trained on a dataset of podcast transcripts and evaluated on a separate test set. The author reported that the system was able to generate coherent and informative summaries of the podcast episodes.

In summary, podcast summarisation has gained attention in recent years, with both extractive and abstractive approaches proposed for the task. Extractive approaches are simpler to implement but often produce less informative summaries, while abstractive approaches require more complex models but can produce more informative summaries.

3 Dataset

The podcast domain is notably distinct from other domains such as news in terms of its style and struc-

ture. To conduct this study, a dataset was obtained from Spotify, which has created the most extensive corpus of transcribed data and audio files in this emerging domain. The corpus is comprised of over 100,000 podcast episodes, amounting to almost 60,000 hours of speech. The transcriptions were generated using Google Cloud Platform's Speech-to-text API (GCP-API), revealing a unique and noisy set of data that has yet to be fully explored in the field of NLP. Additional information about the dataset can be found in (Clifton et al., 2020a). Podcasts are structured in different ways such as scripted and unscripted monologues, interviews, conversations, debate, and includes clips of non-speech audio material. This dataset includes a diverse range of topics, subject matter, speaking styles, and formats, comprising both audio files and transcripts of podcast episodes in Portuguese and English. However, the study solely focuses on summarizing English-language podcasts. Future work could incorporate audio files and transcripts in Portuguese. Additionally, the dataset features metadata, such as descriptions provided by the creators, which can serve as labeled data or reference summaries for summarization.taset covers a wide range of topics, speaking styles, and formats, and includes both audio files and transcripts of podcast episodes in Portuguese, although this study focuses solely on summarising English-language podcast episodes. Furthermore, the dataset provides metadata, including descriptions provided by creators, which are used as labelled data for the summarisation task.

3.1 Analysis of Dataset

The podcast dataset used in this study was obtained from Spotify, which holds one of the largest and most extensive collections of podcasts, including more than 5 million podcast titles[2]. The podcast dataset used in this paper consists of 105,360 podcast episodes that have been transcribed using the Google Cloud Platform's Speech-to-Text API[3]. The resulting dataset comprises a big corpus of approximately 60,000 hours of spoken audio and over 600 million tokens (Clifton et al., 2020b). The transcripts in the dataset have an average length of just under 6000 words, varying from a small number of extremely short episodes to as long as 45,000 words. The majority of the transcripts, approximately two-thirds of them, fall within the range of 1,000 to around 10,000 words. There is also a small percentage, about 1% or 1000 episodes, consisting of very short trailers used to promote the creator's content.

The average episode duration in the dataset is around 33.8 minutes (Figure 1), and the average number of transcribed words in an episode is 5,700. This shows that the documents in the dataset are considerably longer than typical summarisation data. Each show, on average, contains five episodes (Figure 2), with a median of two episodes per show. The dataset covers a wide range of topics and subject matter, including Comedy, Sports, Health & Fitness, Society & Culture, Business, and Education, among others. The dataset is significantly large and varied, making it an ideal resource for the development and evaluation of summarisation techniques in the podcast domain.

This dataset is publicly available and can be accessed through the Spotify API or by contacting Spotify's data research team.

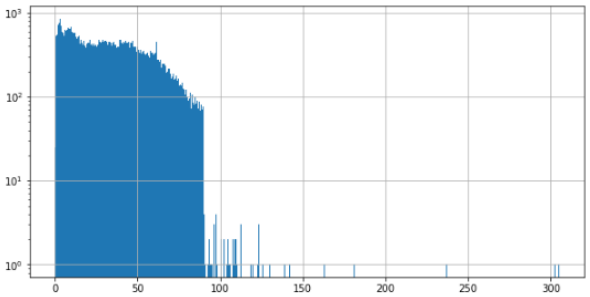

Figure 1: Average duration of episodes.

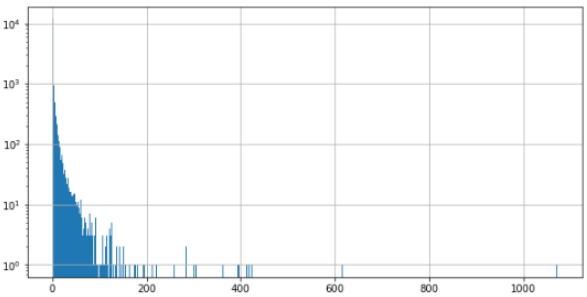

Figure 2: Number of episodes per show.

3.2 Challenges with the Dataset

The podcast dataset used in this study presents several challenges due to its nature and characteristics. Firstly, the transcripts of the audio files are automatically transcribed from the GCP-API, which

[2]Spotify Newsroom February 2023 https://newsroom.spotify.com

[3]https://cloud.google.com/text-to-speech

makes them prone to speech recognition errors.As a result, the dataset is inherently noisy, which can make it more challenging to extract meaningful information despite the post-editing process. In the context of summarisation, having multiple reference summaries for a single source document is beneficial for the models. However, in this particular dataset, we encounter a limitation where only a single summary is available for each episode, provided by the creator. This restriction places a heavy reliance on the creator's provided summary as the sole reference for generating abstractive summaries. Consequently, this limitation may result in a lack of diversity and alternative perspectives in the generated summaries.

Secondly, as podcasts are conversational in nature, they have disfluencies and redundancies in the spoken text. These conversational elements can make it more difficult to accurately interpret the data, especially when compared to data from other domains.

Thirdly, the podcast documents are significantly longer than typical summarisation data, which presents a challenge for SOTA models due to the limitation on the number of tokens. This can make it more difficult to generate high-quality summaries that capture the essence of the episode while remaining concise.

Finally, the descriptions provided by the creators vary widely in quality and often contain sponsorship details that are not intended to act as summaries of the episode. These descriptions were used as labelled data for the summarisation task, highlighting the need for users to be able to read summaries that give an overview of the episode.

These challenges underline the need for sophisticated approaches and techniques to accurately summarise podcast episodes, and this study aims to address these challenges by exploring the use of abstractive text summarisation techniques on this unique dataset.

4 Design and Methodology

The primary objective of this study is to develop an abstractive summarisation system that generates a concise and informative summary of podcast episodes, enabling users to make an informed decision about which podcast to listen to. The ideal summary should accurately convey the essence and most important attributes of the episode, including topical content, participants, and genre, and it should be easily readable on a smartphone with less than 200 words (Liu and Wang, 2022). To achieve this, we aim to fine-tune state-of-the-art transformer-based models (e.g. T5 and Bart) on podcast data. To achieve this we use podcast transcripts rather than audio files to fine-tune the models. Audio data has not been selected for this project due to several reasons. One of the main considerations is the significant variability in audio quality across different podcast episodes. The audio content ranges from professionally produced podcasts with high-quality audio to amateur podcasts that exhibit a wide variety of audio quality. Additionally, the dataset includes episodes self-published through a phone application, further introducing variations in the quality and equipment used by the creators. Given these factors, opting for transcript data ensures a more consistent and standardised input for the summarisation task. Our research pipeline follows a sequential approach that involves several processes to ensure effective summarisation. The process includes text pre-processing, model fine-tuning, and summary generation. This study evaluates the quality of the generated summaries using state-of-the-art evaluation metrics and investigates user attitudes towards the produced summaries. The ultimate goal is to improve the accessibility of podcast content by providing a concise summary that saves users' time and effort in selecting podcasts to listen to.

4.1 Data Pre-processing

The quality of reference summaries is vital for training accurate and reliable models for summarisation. However, episode descriptions provided by podcast creators varied in quality and often contained noisy information, such as sponsorships and promotional content. To improve the accuracy of training data, we filtered out low-quality descriptions dominated by emojis, URLs, advertisements, and promotions.

In addition to filtering, we employed the TextRank algorithm, to identify the most relevant sentences in a text (Mihalcea and Tarau, 2004). Our method was employed on both the descriptions and transcripts of podcast episodes. We aimed to pinpoint crucial keywords and to assess the quality of each episode's description by calculating precision, recall, and F1 scores. To differentiate between low and high-quality descriptions, we set a filter based on precision scores. In particular, episodes that achieved precision scores greater than 0.88 were

labeled as high-quality, whereas those with lower scores were deemed low-quality. Although this cut-off point could be adjusted with further testing, we selected a higher value due to computational resource constraints.

Utilizing the TextRank algorithm, we enhanced the accuracy and relevance of our reference summaries. This resulted in more reliable and applicable summaries for model training. Although this processing step of filtering and applying TextRank reduced the number of dependable episodes for training, it guaranteed the preservation of accuracy and reliability in our summarizer. Moreover, it enabled the effective use of transfer learning.

Implementing these measures allowed us to generate a top-quality dataset of reference summaries to train our summarization model. This empowered us to create precise and succinct summaries for podcast episodes.

4.2 Summarisation Approach

The abstractive approach to text summarisation involves the use of neural methods to generate a condensed representation of documents. A number of approaches have been developed in recent years, which are surveyed in (Lin and Ng, 2019).

For this study, we utilised two state-of-the-art transformer-based models: BART and T5. The BART model (Lewis et al., 2019) is a pre-trained sequence-to-sequence (seq2seq) model that uses a denoising autoencoder to generate summaries. The architecture is based on the transformer model, which has proven highly effective for machine translation tasks (Vaswani et al., 2017). In particular, the BART model is fine-tuned on news summarisation data such as CNN/DailyMail or XSum (Lewis et al., 2019) before being fine-tuned on our podcast dataset. We used the BART-LARGE variant, which contains 12 layers of transformer blocks in both the encoder and decoder. To learn more about the BART model and how to access it using HuggingFace[4]

The T5 model (Xue et al., 2021) is also a transformer-based encoder-decoder model that has been pre-trained on a variety of unsupervised and supervised tasks. One of its key features is the ability to convert NLP problems into a text-to-text format, which makes it highly versatile. For our study, we fine-tuned the T5-BASE model, which has a total of 220 million parameters. To learn more about the T5 model and how to access it using HuggingFace[5]

Both models were fine-tuned on our cleaned and filtered dataset using the episode descriptions provided by podcast creators as training summaries and ground truth summaries. Hyperparameters for the models were chosen based on their effectiveness in prior research, as well as on our specific goals for this project.

The key hyperparameters for both models were as follows: It's worth noting that some of the hyperparameter values used for BART and T5 are default values that are known to work well for their respective architectures.

1. $Maximum\ length$: As the aim of the project was to generate summaries that could be easily read on a smartphone screen, we set a maximum length of 150 characters (Liu and Wang, 2022).

2. $Early\ stopping$: Enabling early stopping helped to prevent overfitting during training.

3. $Length\ penalty$: We used a length penalty of 2 for the BART model and 1 for the T5 model to discourage the models from generating excessively long summaries.

4. $No\ repeat\ n-gram\ size$: To avoid generating repetitive content in the summaries, we set the n-gram size to 3, which ensures that a trigram cannot be generated more than once in the summary.

5. $Num\ beams$: We used a value of 2 for the T5 model and 4 for the BART model. This hyperparameter keeps track of the number of steps taken while the model generates a sequence. Larger values typically generate better summaries, but at the cost of slower processing speeds.

6. $Learning\ rate$: We set the learning rate to 1e-4 for the T5 model and 3e-5 for the BART model to allow the models to converge without overfitting.

7. $Optimiser$: We used the Adam optimiser for the T5 model and the Ranger optimiser for the BART model to compare the performance of different optimisers with different learning rates.

[4]For BART: https://rb.gy/ehvcj

[5]For T5: https://rb.gy/xa3d5

8. *Epochs*: 2 (T5) and 3 (BART): Our models were trained for a total of 3 epochs, but because they were pre-trained, we didn't need to fine-tune them for an extended period of time. This is because we were able to take advantage of transfer learning. The number of epochs was chosen based on early stopping, which was enabled to prevent overfitting. Early stopping was set to 5 epochs, but the models converged in fewer epochs. Therefore, we chose to stop at 2 epochs for T5 and 3 epochs for BART, which provided good results without overfitting the models.

Other hyperparameter values were set to default because the creator descriptions varied greatly in quality so optimising hyperparameters was not worthwhile. Apart from validation, the fine-tuned models were tested with the official TREC 2020 (Clifton et al., 2020a) test set which consists of 1,027 podcast episodes.

We selected hyperparameters based on best practices and previous research. For instance, we set a maximum length for the summaries to ensure they would be easy to read on a smartphone screen. We also used a length penalty to discourage excessively long summaries, and a limit on the n-gram size to prevent repetitive content. We selected the number of beams based on the trade-off between summary quality and processing speed. Additionally, we used standard optimizers and learning rates to help the models converge without overfitting. Finally, we chose the number of epochs based on early stopping as explained by Bai et al. (2021), which we enabled to prevent overfitting. These hyperparameters were tested on a dataset of podcast episodes and validated using standard evaluation metrics.

5 Evaluation

In this section, we present the results of our evaluation, which was conducted on the test set consisting of 1,027 podcast episodes. In addition to the automated metrics, we conducted a human evaluation to gain a better understanding of how people interpret the generated summaries, especially since the creator-provided descriptions were of poor quality.

For automated metrics, we used ROUGE (Lin, 2004) and BERTScore (Zhang et al., 2020). ROUGE score automatically determines the quality of a summary by comparing it to reference summaries. It does this by counting the number of overlapping units, such as word sequences, n-grams, and word pairs between the sets of summaries (Lin, 2004). However, since the aim of abstractive text summarisation is to generate new sentences in the final summary, this metric may not be appropriate. Therefore, we also used BERTScore to better understand the semantic meaning of the summaries. BERTScore computes a similarity score for each token in the candidate sentence with each token in the reference sentence using contextual word embedding. This metric has been shown to correlate well with human judgement (Zhang et al., 2020). Two variants of BERTScore were used: one that utilises IDF weightings and another that does not.

BERTScore		
Model	**F1 Score(%)**	
IDF Weighting	**Yes**	**No**
BART Fine-tuned	80.25	**82.21**
T5 Fine-tuned	**80.43**	82.17
T5 Base	76.74	79.06

Table 1: F1 Measure of BERT Score.

The F1 measure of BERTScore and the F1 measure of ROUGE-1, ROUGE-2, and ROUGE-L are summarised in Table 1 and Table 2, respectively. These results were used to compare the models as well as the pre-trained model.

Model	**R1-F**	**R2-F**	**RL-F**
BART FT	**19.16**	**4.43**	**17.06**
T5 FT	16.88	3.27	14.76
T5 Base	16.55	1.60	14.45

Table 2: F1 Measure of ROUGE Scores.
FT: Fine-Tuned.

5.1 Human Evaluation

In order to evaluate how humans judge summaries, we conducted a qualitative evaluation. A total of 50 participants were recruited as volunteers for the study, with ages ranging from 18 to 50 years old. A questionnaire was distributed to the participants as part of the study. To gather more information about the participants and their views on the importance of podcast episode summaries, there were some questions regarding their demographics at the beginning of the study such as occupation and asked participants about their podcast listening habits. Participants were also asked whether they believe a summary of a podcast episode is important and how an accurate summary would be beneficial to

them. For this study, it took participants on average 15-20 minutes to complete the questionnaire. Participants were not compensated for their time as their participation was completely voluntary.

A questionnaire was distributed to compare the summaries generated by models and determine their quality. For comparison, participants were required to rank summaries generated by the fine-tuned models and the pre-trained models. They were provided with some information about the episode, such as a link to the episode and necessary metadata. This information was sufficient to determine the episode's context and comprehend the summary. The next set of questions were aimed at determining the quality of the generated summaries based on the Excellent, Good, Fair and Bad scale as shown in Table 5 in the Appendix. The details of each scale were given to the participants, and moreover, participants were asked to describe the reason for their choice of selection. This provided more details into the human evaluation of the project. Figure 3 shows an excerpt from the questionnaire that illustrates the type of questions posed to participants.

5.2 Analysis of Results

The results in Table 2 indicate that the BART model fine-tuned on the podcast dataset achieved a higher ROUGE-1 and ROUGE-L score compared to other models. Similarly, the T5 fine-tuned model outperformed its baseline, as evidenced by its higher ROUGE-1 and ROUGE-L scores. While the BART model fine-tuned on the podcast dataset showed higher ROUGE scores, the difference between the two fine-tuned models was minimal when analysing BERTScores in Table 1. When calculating the semantic meaning of generated summaries with IDF weighting set to true, the T5 fine-tuned model performed better than both the BART model by 0.18 percentage points and the T5 Base by 3.69 percentage points, indicating a strong correlation between the meaning of the generated summaries and the descriptions provided by the creators.

The results of the human evaluation revealed that both the BART and T5 fine-tuned models produced high-quality summaries that were well received by participants. The majority of participants rated the summaries generated by both models as Good or Excellent. This suggests that the summaries were coherent, accurate and provided a meaningful overview of the content of the podcast episode.

The results of the human evaluation (Table 3) indicate that there was minimal difference between the performance of the BART and T5 models, both fine-tuned on the podcast dataset, and that their results were highly comparable. The BART model produced summaries that achieved an Excellent rating of 39.29% and a Good rating of 35.7%, with only 3.57% rated as Bad. These results indicate that the generated summaries were of high quality and were well received by the participants. The T5 fine-tuned model can be similarly described as it obtained a majority of Good ratings, with 44.6% of participants rating it as such.

In contrast, the baseline T5 model (T5 Base) had a high percentage of Bad ratings at 64.3%, indicating that it struggled to capture the meaning and context of the podcast episodes. This highlights the importance of fine-tuning on domain-specific data for generating high-quality summaries.

Participants were also asked to rank the summaries generated by the models for an episode. The evaluation results (Table 4) show that 60.22% of participants ranked the summaries generated by the fine-tuned BART model as first, while 50.55% ranked the fine-tuned T5 as second, and the T5 baseline model was ranked as third by 73.45% of the participants. These findings suggest that both fine-tuned models, particularly BART, generated summaries that were of high quality compared to the baseline. Participants praised the fine-tuned models' summaries for being "very concise and accurate", for "grabbing the reader's attention" and "containing accurate descriptions of the content that were easy to read." On the other hand, participants described the baseline model's summary as too long and poorly formatted. Table 6 provides example of summaries generated by the models and the metadata of an episode.

Overall, the evaluation results demonstrate the effectiveness of both the BART and T5 models for summarising podcast episodes. The BART model performed well in terms of ROUGE scores, while the T5 model excelled in capturing the semantic meaning of the summaries. The human evaluation confirmed that the generated summaries were of high quality and provided a meaningful overview of the podcast episode. The success of these models could have significant practical applications, such as assisting listeners in choosing which episodes to listen to or summarising podcasts for users with

limited time.

Model	E	G	F	B
BART FT	**39.3%**	35.7%	21.4%	3.57%
T5 FT	25.0%	**44.6%**	**25.0%**	5.4%
T5 Base	8.9%	7.1%	19.7%	**64.3%**

Table 3: The mean percentage of the quality of the generated summaries.

Model	1	2	3
BART FT	**60.22%**	29.64%	10.15%
T5 FT	35.07%	**50.55%**	14.38%
T5 Base	13.60%	12.96%	**73.45%**

Table 4: The table displays the mean percentage of the ranking, which compares three models based on a scale of 1 (best) to 3 (worst).

6 Conclusion

In conclusion, this paper presents a study on the summarisation of podcast episodes using abstractive methods. We explored the use of the BART and T5 models, fine-tuned on a dataset of podcast episode descriptions, and evaluated their performance using automated metrics and human judgement. Our results showed that the fine-tuned models outperformed their pre-trained counterparts and achieved high scores in both ROUGE and BERTScore metrics. Moreover, the human evaluation indicated that the generated summaries were of high quality and well-received by participants. Overall, our findings demonstrate the potential of using abstractive summarisation for podcasts, providing listeners with a quick and accurate summary of episodes.With the help of abstractive text summarisation, podcast creators can implement this technology to automatically generate descriptions for their episodes, which was a manual process in the podcasting industry, helping them save time and allowing the users to read high-quality descriptions for their favourite podcasts. In light of these findings, it is clear that the podcast domain could greatly benefit from the use of NLP technology in generating accurate and concise summaries of audio content. This could help users better manage and discover relevant content, while also making podcast episodes more accessible to individuals with hearing impairments or language barriers in the future. Future work could explore the use of additional features or fine-tuning methods to further improve the performance of the summarisation models on podcast data. Another aspect that can be explored is finding methods to tackle disfluencies in spoken text.Overall, this project has provided valuable insights into the application of NLP in the podcast domain and the potential for improving the accessibility and usability of podcast content.

References

Yingbin Bai, Erkun Yang, Bo Han, Yanhua Yang, Jiatong Li, Yinian Mao, Gang Niu, and Tongliang Liu. 2021. Understanding and improving early stopping for learning with noisy labels. *Advances in Neural Information Processing Systems*, 34:24392–24403.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Ann Clifton, Sravana Reddy, Yongze Yu, Aasish Pappu, Rezvaneh Rezapour, Hamed Bonab, Maria Eskevich, Gareth Jones, Jussi Karlgren, Ben Carterette, and Rosie Jones. 2020a. 100,000 podcasts: A spoken English document corpus. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5903–5917, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Ann Clifton, Sravana Reddy, Yongze Yu, Aasish Pappu, Rezvaneh Rezapour, Hamed Bonab, Maria Eskevich, Gareth Jones, Jussi Karlgren, Ben Carterette, et al. 2020b. 100,000 podcasts: A spoken english document corpus. pages 5903–5917.

Mahmoud El-Haj. 2012. *Arabic multi-document text summarisation*. Ph.D. thesis, University of Essex.

Mahmoud El-Haj, Udo Kruschwitz, and Chris Fox. 2010. Using mechanical turk to create a corpus of arabic summaries.

Hannes Karlbom. 2021. Abstractive summarization of podcast transcriptions.

Philippe Laban, Elicia Ye, Srujay Korlakunta, John Canny, and Marti Hearst. 2022. Newspod: Automatic and interactive news podcasts. In *27th International Conference on Intelligent User Interfaces*, pages 691–706.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. *arXiv:1910.13461 [cs, stat]*. ArXiv: 1910.13461.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Hui-Ching Lin and Vincent Ng. 2019. Abstractive summarization: A survey of the state of the art. In *AAAI Conference on Artificial Intelligence*.

Zhendong Liu and Beihai Wang. 2022. Research on text visual effect of multimedia courseware for mobile online learning. In *Man-Machine-Environment System Engineering: Proceedings of the 21st International Conference on MMESE: Commemorative Conference for the 110th Anniversary of Xuesen Qian's Birth and the 40th Anniversary of Founding of Man-Machine-Environment System Engineering 21*, pages 841–847. Springer.

Rada Mihalcea and Paul Tarau. 2004. Textrank: Bringing order into text. In *Proceedings of the 2004 conference on empirical methods in natural language processing*, pages 404–411.

Victor Risne and ADÉLE Siitova. 2019. Text summarization using transfer learnin: Extractive and abstractive summarization using bert and gpt-2 on news and podcast data.

John L Sullivan. 2019. The platforms of podcasting: Past and present. *Social media+ society*, 5(4):2056305119880002.

Christine Tulley. 2011. Itext reconfigured: The rise of the podcast. *Journal of Business and Technical Communication*, 25(3):256–275.

Aneesh Vartakavi and Amanmeet Garg. 2020. Podsumm–podcast audio summarization. *arXiv preprint arXiv:2009.10315*.

Aneesh Vartakavi, Amanmeet Garg, and Zafar Rafii. 2021. Audio summarization for podcasts. In *2021 29th European Signal Processing Conference (EUSIPCO)*, pages 431–435. IEEE.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. *arXiv:2010.11934 [cs]*. ArXiv: 2010.11934.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. *arXiv:1904.09675 [cs]*. ArXiv: 1904.09675.

Nadhem Zmandar, Abhishek Singh, Mahmoud El-Haj, and Paul Rayson. 2021. Joint abstractive and extractive method for long financial document summarization. In *Proceedings of the 3rd Financial Narrative Processing Workshop*, pages 99–105.

Appendix: Guidelines for Human Evaluation

Excellent	Accurately conveys all the most important attributes of the episode, which could include topical content, genre, and participants. It contains almost no redundant material which isn't needed when deciding whether to listen.
Good	Conveys most of the most important attributes and gives the reader a reasonable sense of what the episode contains. Does not need to be fully coherent or well edited. It contains little redundant material which isn't needed when deciding whether to listen.
Fair	Conveys some attributes of the content but gives the reader an imperfect or incomplete sense of what the episode contains. It may contain some redundant material which isn't needed when deciding whether to listen.
Bad	Does not convey any of the most important content items of the episode or gives the reader an incorrect sense of what the episode contains. It may contain a lot of redundant information that isn't needed when

Table 5: Guidelines for evaluation of summaries (Clifton et al., 2020b).

Show Name	Skincare Sommeliers
Episode Name	The Great Exfoliation Debate
Creator Description	Tune in as Natalie and Jessica debate physical vs. chemical exfoliation options, and see what our ultimate verdict is on the best type and specific products we love!
BART FT	In today's episode, Natalie and Jessica debate the benefits of different types of exfoliation. We also do our wine pairing of Santa Julia, a Malbec and cabernet sauvignon blend from 2016. Find out which one is best for you!
T5 FT	We talk about how to get into exfoliation and why we love it. We also talk about our favorite products from the past week, like a charcoal and salt scrub.
T5 Base	I love the way they work. a little bit of it's not cute. you can get an extra strength version for like $88. Yeah, we're going to talk about and on our podcast that your Natalie this is really kind of my yeah. right? Okay. So first off so let's start with Malbec. also just these are pretty good. but. And...

Table 6: Sample Details of podcast episode and generated summaries.

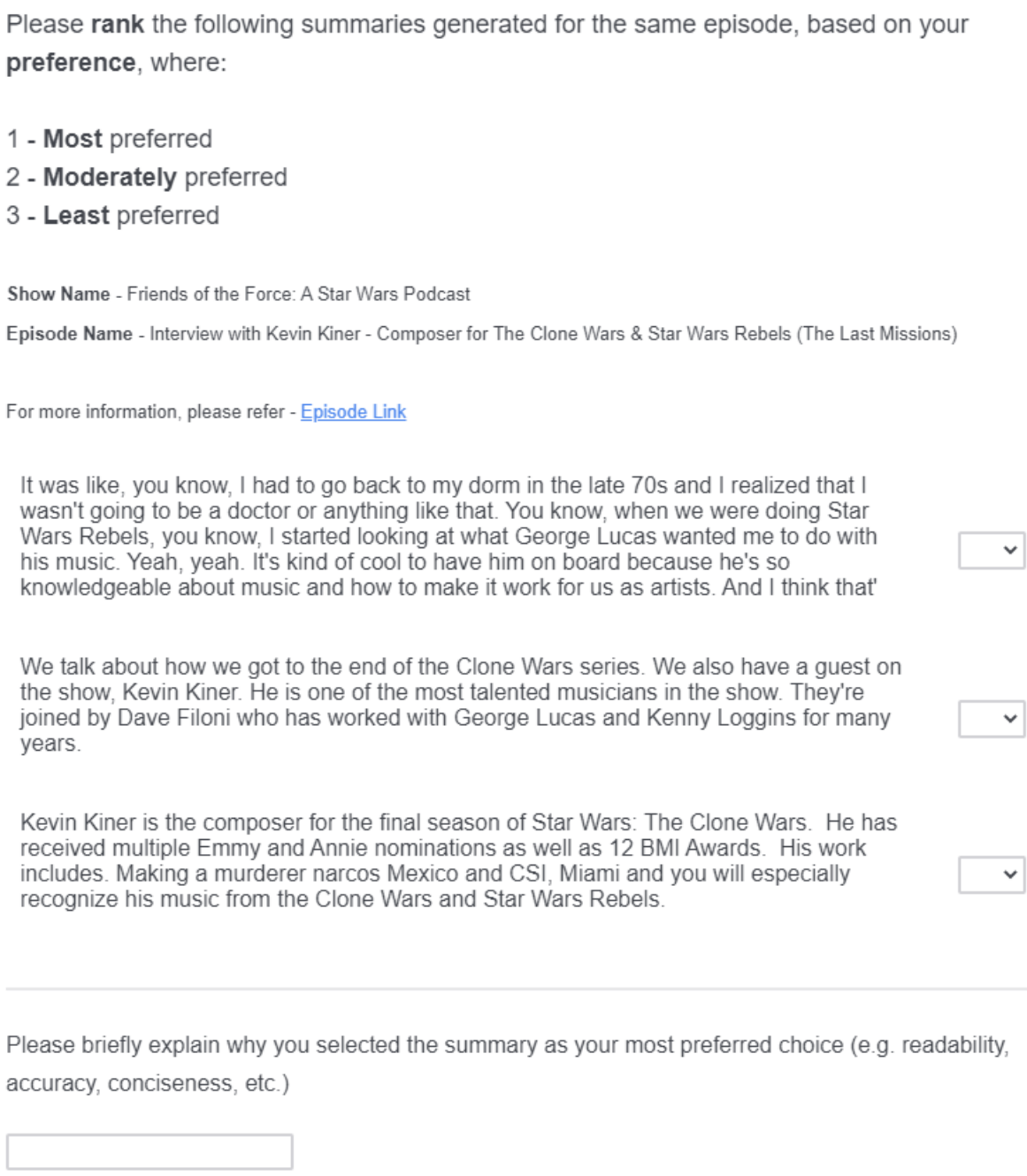

Figure 3: One of the questions from qualitative evaluation. This questions asks participants to rank the summaries generated by the models, from 1 (best) to 3 (worst).