Multilingual Podcast Summarization using Longformers

Edgar Tanaka
Spotify
edgart@spotify.com

Ann Clifton
Spotify
aclifton@spotify.com

Md. Iftekhar Tanveer
Spotify
iftekhart@spotify.com

***Abstract*—Most literature on automated summarization, including podcast summarization, has been restricted to the English language. At the same time, podcasts are now an important form of media in many countries and in many languages and therefore, it is crucial that we expand the problem of podcast summarization to a wider range of languages. In this work, we explore the application of multilingual models to the task of summarizing podcasts in English and Portuguese. We compare various training scenarios including adapting a Longformer encoder, cross-lingual and cross-task transfer learning and we demonstrate that a unified model fine-tuned to multilingual data can perform on par with dedicated models that are fine-tuned monolingually. As a result, our models significantly outperform the TREC baseline based on the first minute of each episode.**

***Index Terms*—podcasts, summarization, multilingual, longformer**

I. Introduction

Text summarization has long been researched within the Natural Language Processing (NLP) community. However, most works have focused on summarizing news articles while the summarization of podcasts still remains fairly unexplored. Summarizing podcasts is a challenging task due to a number of reasons. Firstly, there are many podcast formats such as interviews, debates, monologues. Secondly, podcast transcripts are noisy as the audio often contains fillers and overlapping speakers, and the resulting transcripts contain errors with ASR and inferred punctuation. Lastly, these transcripts are often very long, whereas state-of-the-art models are generally trained on short texts and can ingest only a limited number of tokens.

The Podcast Summarization Track in TREC 2020 has encouraged the research in this area but it has been restricted to the English language. At the same time, previous works have demonstrated that single models trained on multiple languages provide competitive performance when compared against monolingual models [16] [15] [6]. In this work, we aim to explore the problem of multilingual podcast summarization by training a model to summarize podcasts in Portuguese and English; we have selected these two languages for proof of concept, but we note that nothing about our approach restricts it to just two languages. We have trained two models: (1) one finetuned on podcasts data (English and Portuguese) and (2) another finetuned first on news articles and then on podcasts data (English and Portuguese).

Previous works have used BART [10] to summarize podcasts in English: as a baseline [3] and in a summarization model using Longformer attention [8]. We have decided to use mBART-50[1] which is a multilingual version of BART pre-trained in 50 languages including English and Portuguese. Starting with a model pre-trained in these two languages was a requirement for our work as training a new language from scratch requires massive computing power and data. mBART-50 was also a good fit because, even though it was evaluated as a machine translator, it is still a sequence-to-sequence model trained in many languages.

Another model we would like to use as a starting point is XL-SUM [6]. It is a massive multilingual summarization model trained in 44 languages, including English and Portuguese. Due to time restrictions, we have only used XL-SUM's dataset with BBC news articles and summaries in this work. In the future, we would like to finetune the XL-SUM model with podcasts data.

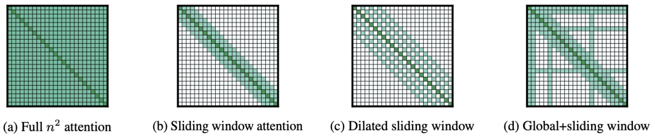
In addition to exploring the multilingual perspective on summarization, we also wanted to investigate the benefits of using the Longformer attention mechanism. Transformer-based models are unable to process long documents due to their self-attention mechanism, which scales quadratically with the sequence length [2], and thus most large-scale pretrained models can only accept inputs much shorter than the average transcript length. To address this problem, Iz Beltagy et al proposed Longformer [2], a transformer-based model with an attention mechanism that scales linearly with sequence length, making it possible to process documents with thousands of tokens or more. While BERT-based pretrained models typically have a 512 or 1024 token limit, Longformer can process up to 16K tokens.

Figure 1 illustrates the different attention mechanisms proposed by Iz Beltagy et al [2].

This paper's major contribution is in providing a multilingual alternative to the podcast summarization problem which can be compared to other monolingual models submitted to TREC 2021. In addition, we also investigate how this stacks with increasing the text input size with Longformer [2] attention mechanism.

[1] https://huggingface.co/facebook/mbart-large-50

Fig. 1. Comparing (a) the full self-attention pattern and (b)(c)(d) the configuration of attention patterns in Longformer.



(a) Full n^2 attention (b) Sliding window attention (c) Dilated sliding window (d) Global+sliding window

Source: Longformer: The Long-Document Transformer [2]

II. Datasets

Since we are experimenting in a multilingual space, we have decided to work with two languages: Portuguese and English. We have also decided to work with two distinct types of data: news articles provided by the XL-SUM dataset and podcast transcripts.

A. English Podcasts dataset

This dataset consists of over 100,000 podcast episodes in English. This includes nearly 60,000 hours of audio and accompanying transcripts. It also includes metadata such as creator-provided descriptions [3] and was published last year to the research community[2] in the 2020 Text Retrieval Conference (TREC) Podcasts Track.

To train the summarization model, we have reused the same dataset from the TREC 2020 submission described in [8] which is a subset of the original podcasts dataset [3]. We have also reused the same splits (train, dev, test). To build this dataset, the following filters were used:

- Removed episodes where the creator descriptions are either too long or too short with the boundary conditions set to between 10 and 1300 characters.
- Applied a TF-IDF vectorization of the descriptions which were compared to each other using the cosine distance. Any data points with too similar descriptions (threshold 0.95) were filtered out.
- Removed boilerplate sentences from the creator descriptions using a sentence classifier based on BERT [4]. This classifier was trained using a small set of 1000 manually labeled episodes [14].

B. Portuguese Podcasts dataset

This dataset composed of more than 100,000 podcast episodes in Portuguese (both PT-PT and PT-BR) was built internally at Spotify.

We have followed the same process used by the English podcasts dataset [3] but filtering for Portuguese content instead of English content. The following filters were used:

- The language of the show specified in the metadata must be Portuguese (PT-BR or PT-PT)
- The episodes are all Spotify owned-and-operated, for copyright reasons.
- Since the metadata language tags are noisy, the episode descriptions must also be identified as Portuguese by the langid[3] python package.
- The episode must have more than 50% of speech over its duration; a proprietary speech detection algorithm was used to determine this. This filters out podcasts that are more music, white noise, or meditation than speech.

Once the episodes were selected, they were sent to transcription using the Speech-to-Text service provided by the Azure platform.

The Portuguese Podcasts dataset is not yet available to the general public but we do plan to release it in 2022.

C. XL-SUM dataset

We have used the BBC news articles and their summaries provided by the XL-SUM dataset[4]. This dataset contains article-summary pairs in 44 languages but we only used the articles in Portuguese and English.

In order to keep data in both languages balanced, we have downsampled the English articles so that they matched the number of articles in Portuguese. No other filters were applied.

D. Removing extraneous content

To remove extraneous content such as ads or boilerplate from episode descriptions, we manually annotated sentences from 1000 episode descriptions as either "extraneous" or "not extraneous". We then used these labeled data to train a binary classifier to detect extraneous content in the manner described in [14] and used it to clean the creator-provided descriptions both in Portuguese and English podcasts.

Here are some examples of extraneous content found in episode descriptions:

- "Send in voice message http://anchor.com/foobar"
- Requests for followers on any social media
- Advertisements
- List of technical staff (producer, editor, sound technician…)
- Hashtags to characterize or promote the content
- Credits to the soundtrack used during the episode
- Time marks such as "0:30 <topic 1> 1:25 <topic 2> 5:40 <topic 3>"
- License information such as Creative Commons license
- List of participants described by their Twitter or Instagram usernames

To evaluate the accuracy of this binary classifier, we have analyzed 100 episode descriptions that went through the cleaning process. Out of 100 examples, 2 still had extraneous content, 3 had non-extraneous content removed and 95 were correctly cleaned.

III. Methodology

Our methodology was comprised of the following steps:

1) Prepare the data as described in section "Datasets".

[2] https://podcastsdataset.byspotify.com/

[3] https://pypi.org/project/langid/

[4] https://github.com/csebuetnlp/xl-sum

2) Convert the mBART-50 model into a Longformer [2] version.
3) Finetune model to the summarization task using news article-summary pairs. (for the double finetuned variant only)
4) Finetune model to the podcast summarization using podcast transcriptions and descriptions.
5) Evaluate model.

We have chosen the mBART-50 [12] model because it has been already pre-trained in 50 languages including Portuguese and English. Then, we changed the source code to adapt it to a Longformer version which allowed the expansion of the input size from 1024 to 4096 tokens. The notebook available in [1] was used as a reference for this code change. Our hypothesis here is that the model will generate more complete summaries by being exposed to more of the full input text. This extension seems particularly crucial for podcasts given that transcripts are much longer than a news article.

The next step was to finetune the model to the task of podcast summarization. We developed two distinct models with different finetuning strategies:

- The first one called "Unicamp1" was finetuned only using podcasts data. Although mBART is a sequence-to-sequence model, it is initially trained to the machine translation task so we wanted to verify if finetuning it directly into the podcast summarization task would be successful. The training was configured for early stopping once the loss function had not improved after 3 validation checkpoints.
- The second one called "Unicamp2" was finetuned initially on news articles from the XLSUM dataset [6] and then subsequently finetuned using podcasts data. The intuition here is that in the first round of finetuning, the model should learn how to summarize using high-quality news article-summary pairs. In other words, we would expect mBART to transition from a neural machine translation model to a summarization model. With the second round of finetuning, the model would then learn how to summarize podcast transcripts specifically. The training was set to early stop once the ROUGE score didn't improve after 3 validation checks.

In both cases, Portuguese and English data was intermingled in order to avoid catastrophic forgetting [5].

IV. Results and Discussion

A. Human evaluation

From the 1000 podcast episodes in the test set, 193 were selected at random to be evaluated by NIST assessors. There were two types of evaluation: (1) summary quality based on a 4-point scale: bad, fair, good and excellent, and (2) eight yes/no questions were answered evaluating the summaries submitted. These results were all compared against a 1st-minute baseline where the summary is simply the 1st minute of the episode transcript.

The summary quality was defined in these terms:

- Excellent: the summary accurately conveys all the most important attributes of the episode, which could include topical content, genre, and participants. In addition to giving an accurate representation of the content, it contains almost no redundant material which is not needed when deciding whether to listen. It is also coherent, comprehensible, and has no grammatical errors.
- Good: the summary conveys most of the most important attributes and gives the reader a reasonable sense of what the episode contains with little redundant material which is not needed when deciding whether to listen. Occasional grammatical or coherence errors are acceptable.
- Fair: the summary conveys some attributes of the content but gives the reader an imperfect or incomplete sense of what the episode contains. It may contain redundant material which is not needed when deciding whether to listen and may contain repetitions or broken sentences.
- Bad: the summary does not convey any of the most important content items of the episode or gives the reader an incorrect or incomprehensible sense of what the episode contains. It may contain a large amount of redundant information that is not needed when deciding whether to listen to the episode.

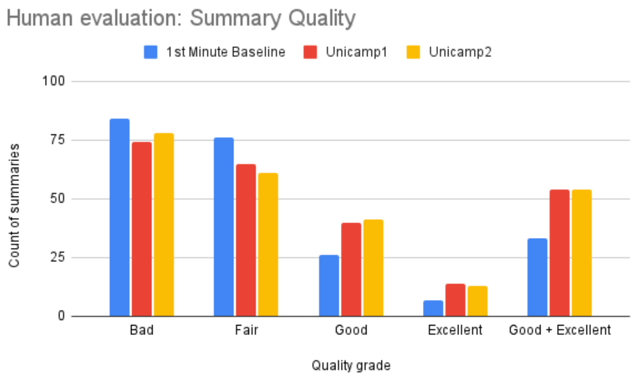


Fig. 2. Overall quality scores. '1st Minute Baseline' refers to the TREC-provided baseline of the first minute of speech.

In figure IV-A, we have the absolute count of summaries for each grade. Our models (both Unicamp1 and Unicamp 2) produced 63% more Good/Excellent summaries than the 1st Minute Baseline while producing 13% less Bad/Fair summaries. At the same time, we did not notice any significant differences between runs Unicamp1 and Unicamp2. In fact, when we bucket together Bad and Fair summaries, both models produce the exact same count of summaries. The same happens when we bucket together Good and Excellent summaries.

In the second part of the evaluation, NIST assessors answered the following yes/no questions to evaluate the content in the summary:

- **Q1**: Does the summary include names of the main people (hosts, guests, characters) involved or mentioned in the podcast?

- **Q2**: Does the summary give any additional information about the people mentioned (such as their job titles, biographies, personal background, etc)?
- **Q3**: Does the summary include the main topic(s) of the podcast?
- **Q4**: Does the summary tell you anything about the format of the podcast; e.g. whether it's an interview, whether it's a chat between friends, a monologue, etc
- **Q5**: Does the summary give you more context on the title of the podcast?
- **Q6**: Does the summary contain redundant information?
- **Q7**: Is the summary written in good English?
- **Q8**: Are the start and end of the summary good sentence and paragraph start and end points?

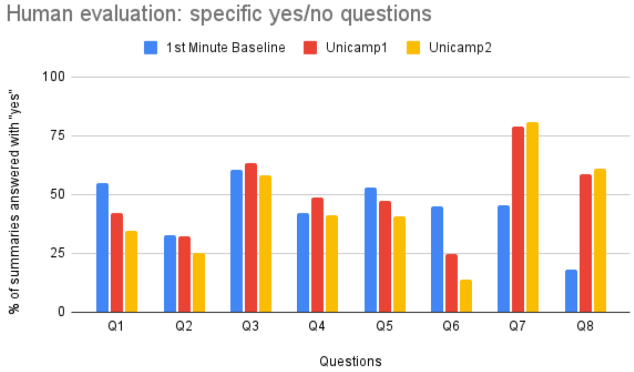


Fig. 3. Averages per question. '1st Minute Baseline' refers to the TREC-provided baseline of the first minute of speech.

In figure IV-A, we see the percentage of summaries where the answer was "yes".

For questions 1 and 2, our runs generated summaries that were equal to or worse than the 1st-minute baseline. This is an indication that our models have not learned to include the names of the main people in the episode or provide additional information about them. In this context, we consider that any results matching the 1st-minute baseline indicate poor performance given its simplicity. This requires further investigation. However, we find this result surprising given the fact that the annotators judged our systems' summaries to be overall Good or Excellent more often than the 1st-minute baseline, indicating that these qualities in isolation may not predicate the reader's experience of quality.

Questions 3 and 5 are very important as they evaluate if the model was capable of surfacing the main topic and context around it. For (3) questions, our models provided summaries that were roughly on par with the 1st-minute baseline; for (5) they were slightly worse.

In question 4, the results show that our model performs similarly to the 1st-minute baseline. A hypothesis is that a lot of the episode descriptions do not provide their format in the description.

Our models outperform the 1st-minute baseline in questions 6, 7 and 8. SOTA (state-of-the-art) transformer-based models have been successful in generating fluent text. Questions 7 and 8 are basically assessing fluency so we can understand how a summary generated by a neural model is better than simply cutting a piece of a possibly noisy podcast transcript.

TABLE I
ROUGE SCORES FOR INTERNAL TEST SET OF 4511 ENGLISH PODCAST EPISODES.

	R1-F	**R2-F**	**RL-F**
First Minute baseline	0.1723	0.0303	0.1545
TextRank Top 5 sentences	0.1401	0.0161	0.1183
TextRank Top 2 sentences	0.1407	0.0145	0.1144
XLSUM vanilla	0.1174	0.0156	0.1036
MBART vanilla	0.1579	0.0272	0.1400
MBART + finetuned PT podcasts	0.0407	0.0046	0.0385
MBART + finetuned EN podcasts	**0.1862**	**0.0563**	**0.1663**
MBART + finetuned PT/EN podcasts	**0.1859**	0.0499	**0.1657**
LongMBART vanilla	0.1620	0.0280	0.1440
LongMBART + finetuned PT podcasts	0.0341	0.0043	0.0327
LongMBART + finetuned EN podcasts	0.1845	0.0521	0.1633
LongMBART + finetuned PT/EN podcasts	0.1812	0.0482	0.1620
LongMBART + finetuned XL-SUM + finetuned PT/EN podcasts	0.1844	**0.0553**	0.1650

Overall, we conclude that the 1st-minute baseline was strong for informational adequacy and that our models' advantage was in coherence and fluency.

B. Automated evaluation: ROUGE scores

In this section, we evaluate our models using the ROUGE metric [11]. This method does not require any human intervention and relies on a gold reference summary to measure the performance of each model. Since gold reference summaries do not exist for this dataset, we use the filtered creator descriptions (without extraneous content as defined in subsection II-D) as proxies for the reference summaries. We note that the creator descriptions are noisier than gold reference summaries, which must be taken into account when interpreting the ROUGE scores. However, in [7], the authors note that the ranking over systems induced by ROUGE scores correlated with human judgments of quality. Table I presents the ROUGE scores when evaluating only on a test set of 4511 English podcasts. Table II presents the ROUGE scores when evaluating only on a test set of 5073 Portuguese podcasts. Lastly, Table III presents the ROUGE scores for the combination of both test sets in English and Portuguese.

The **First Minute** baseline extracts the initial segment of the transcript up until the first minute of the episode and considers that as the summary. This is the same baseline used by the human evaluation.

TextRank Top 2 sentences and **TextRank Top 5 sentences** use TextRank [13] which is an unsupervised extractive summarization model. TextRank Top 2 sentences uses the top 2 sentences to compose the summary. TextRank Top 5 sentences uses the top 5 sentences.

XLSUM vanilla[5] is the mT5 model finetuned to the summarization task with the XLSUM dataset. This is a multilingual

[5] https://huggingface.co/csebuetnlp/mT5_multilingual_XLSum

TABLE II
ROUGE SCORES FOR INTERNAL TEST SET OF 5073 PORTUGUESE PODCAST EPISODES.

	R1-F	R2-F	RL-F
First Minute baseline	0.1674	0.0327	0.1397
TextRank Top 5 sentences	0.1169	0.0143	0.0959
TextRank Top 2 sentences	0.1335	0.0139	0.1058
XLSUM vanilla	0.1120	0.0159	0.0951
MBART vanilla	0.1586	0.0277	0.1342
MBART + finetuned PT podcasts	**0.1886**	**0.0516**	**0.1634**
MBART + finetuned EN podcasts	0.0393	0.0067	0.0369
MBART + finetuned PT/EN podcasts	**0.1835**	0.0501	**0.1598**
LongMBART vanilla	0.1136	0.0119	0.1021
LongMBART + finetuned PT podcasts	0.1826	**0.0501**	0.1598
LongMBART + finetuned EN podcasts	0.0280	0.0046	0.0266
LongMBART + finetuned PT/EN podcasts	0.1761	0.0491	0.1536
LongMBART + finetuned XL-SUM + finetuned PT/EN podcasts	0.1764	0.0481	0.1535

TABLE III
ROUGE SCORES FOR INTERNAL TEST SET OF 5073 PORTUGUESE PODCAST EPISODES COMBINED WITH INTERNAL SET OF 4511 ENGLISH PODCAST EPISODES.

	R1-F	R2-F	RL-F
First Minute baseline	0.1697	0.0316	0.1466
TextRank Top 5 sentences	0.1278	0.0152	0.1064
TextRank Top 2 sentences	0.1369	0.0142	0.1099
XLSUM vanilla	0.1146	0.0157	0.0991
MBART vanilla	0.1583	0.0275	0.1369
MBART + finetuned PT podcasts	0.1191	0.0295	0.1047
MBART + finetuned EN podcasts	0.1084	0.0300	0.0978
MBART + finetuned PT/EN podcasts	**0.1846**	**0.0500**	**0.1625**
LongMBART vanilla	0.1364	0.0195	0.1218
LongMBART + finetuned PT podcasts	0.1128	0.0286	0.1001
LongMBART + finetuned EN podcasts	0.1018	0.0270	0.0911
LongMBART + finetuned PT/EN podcasts	0.1785	0.0487	0.1576
LongMBART + finetuned XL-SUM + finetuned PT/EN podcasts	**0.1802**	**0.0515**	**0.1589**

summarization trained in 45 languages including Portuguese and English. There was no finetuning with podcasts data.

MBART vanilla is the MBART-50[6] model without any finetuning. Note that this is machine translation model and not a summarization model. **MBART + finetuned PT podcasts** is MBART-50 finetuned only to Portuguese podcasts. **MBART + finetuned EN podcasts** is MBART-50 finetuned only to English podcasts. **MBART + finetuned PT/EN podcasts** is MBART-50 finetuned to both English and Portuguese podcasts intermingled.

LongMBART is the same as **Unicamp1** as explained in section III. However, we have experimented with the same variants as done with the MBART-50: finetuned only to Portuguese podcasts, only to English podcasts and finally to both intermingled.

LongMBART + finetuned XL-SUM + finetuned PT/EN podcasts follows the training protocol of **Unicamp2** as explained in section III. We used English and Portuguese data in an intermingled fashion for both finetuning rounds.

When interpreting the results from Table I and II, we notice that the MBART models finetuned monolingually have produced the highest ROUGE scores and outperformed the First Minute Baseline. Generally speaking, the best results came from the MBART and LongMBART models finetuned to the same language (either monolingually or bilingually) as the test set's language. It is also worth noting that the difference between finetuning monolingually and bilingually is marginal which leads us to conclude that learning to summarize in an additional language does not come at the cost of worse performance in a first language.

When we interpret the results from Table III, we can clearly see the importance of finetuning on both languages. Any of the monolingually-finetuned models have performed poorly here.

It came as a surprise that LongMBART model did not lead to higher ROUGE scores when compared to the MBART model. By converting the MBART model into a Longformer version, we increased the input text size limit from 512 tokens to 4096 tokens. Our initial hypothesis was that passing more information (i.e. more text from the transcript) to the model would lead to a better summary. However, the LongMBART model ended up producing summaries with ROUGE scores on par with the MBART model finetuned. As future work, we may want to evaluate if the text after the initial 512 tokens is irrelevant to composing a better summary, i.e. a summary that resembles the gold reference.

Relatedly, as already noted during this year's TREC Overview presentation, the First Minute baseline is a competitive baseline. This finding can be explained by the layout bias [9] present in podcasts. Similar to news articles, episodes tend to start with a brief summary of the overall content of each episode. The hosts usually present the guests (if any), they mention the topics to be covered and provide some background context for the listeners. The very same information is usually provided in the episode description which is used as our gold reference. This layout bias could also account for the fact that the Longformer-based model did not outperform the MBART model.

Another unexpected outcome was the fact that the XL-SUM vanilla model performed worse than MBART vanilla. Given that XL-SUM is a model finetuned on the summarization task, our expectation was that it would lead to higher ROUGE scores when compared to MBART which is a neural machine translation model. While observing the examples in Tables VI and IV, we noticed that MBART is mostly copying the beginning of a transcript and therefore producing a summary similar to the First Minute baseline. MBART in this case is simply acting as a translation model where the source and target language are the same. Thus, we surmise that the competitive performance of the vanilla MBART model is again an artifact of the layout bias previously discussed, and would not necessarily generalize to other datasets.

When we analyze the models finetuned monolingually, unsurprisingly, ROUGE scores were the highest when the language of the test set and the training set were the same. On the other hand, when they were different, ROUGE scores were the lowest. These low scores are due to the fact these models tended to produce text only in the language seen during its finetuning process. For example, a model finetuned only

[6] https://huggingface.co/facebook/mbart-large-50

on Portuguese podcasts, tended to produce text in Portuguese even when the input was in English. And vice-versa. We can observe this phenomenon for models "LongMBART + finetuned PT podcasts" and "MBART + finetuned PT podcasts" in Table VII and model "LongMBART + finetuned EN podcasts" in Table V.

V. Conclusion

We have presented the results of summarizing podcasts in Portuguese and English with multilingual summarization models. We have experimented with a number of summarization techniques, models and finetuning strategies. We conclude that there are no significant ROUGE score improvements when extending the text input size of MBART or when adding an extra finetuning round on XL-SUM data. Nevertheless, our models outperformed the first-minute baseline both according to human evaluation and ROUGE scores. Our best model produced 63% more summaries with Good/Excellent grades according to NIST assessors when compared against the first-minute baseline. Using ROUGE-L F-score as our metric, MBART finetuned on PT/EN podcasts led to a gain of 1.5 points over the first-minute baseline. This paper's major contribution is to serve as a first step in the study of multilingual podcast summarization and share important findings based on the various experiments conducted.

References

[1] Iz Beltagy. Roberta to longformer: build a "long" version of pretrained models. https://github.com/allenai/longformer/blob/master/scripts/convert_model_to_long.ipynb. Accessed: 2021-07-18.

[2] Iz Beltagy, Matthew E. Peters, and Arman Cohan. Longformer: The long-document transformer, 2020.

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

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[5] Robert M. French. Catastrophic forgetting in connectionist networks. *Trends in Cognitive Sciences*, 3(4):128–135, 1999.

[6] Tahmid Hasan, Abhik Bhattacharjee, Md Saiful Islam, Kazi Samin, Yuan-Fang Li, Yong-Bin Kang, M. Sohel Rahman, and Rifat Shahriyar. Xl-sum: Large-scale multilingual abstractive summarization for 44 languages, 2021.

[7] Rosie Jones, Ben Carterette, Ann Clifton, Maria Eskevich, Gareth J. F. Jones, Jussi Karlgren, Aasish Pappu, Sravana Reddy, and Yongze Yu. Trec 2020 podcasts track overview, 2021.

[8] Hannes Karlbom and A. Clifton. Abstract podcast summarization using bart with longformer attention. In *TREC*, 2020.

[9] Wojciech Kryściński, Nitish Shirish Keskar, Bryan McCann, Caiming Xiong, and Richard Socher. Neural text summarization: A critical evaluation, 2019.

[10] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension, 2019.

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

[12] Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. Multilingual denoising pre-training for neural machine translation, 2020.

[13] Rada Mihalcea and Paul Tarau. TextRank: Bringing order into text. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 404–411, Barcelona, Spain, July 2004. Association for Computational Linguistics.

[14] Sravana Reddy, Yongze Yu, Aasish Pappu, Aswin Sivaraman, Rezvaneh Rezapour, and Rosie Jones. Detecting extraneous content in podcasts, 2021.

[15] Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. mt5: A massively multilingual pre-trained text-to-text transformer, 2021.

[16] Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich. Improving massively multilingual neural machine translation and zero-shot translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1628–1639, Online, July 2020. Association for Computational Linguistics.

TABLE IV

Summaries generated by each non-finetuned model for a given episode in Portuguese.

Model	Predicted summary
Episode description by creator	Andrei Spacov, sócio e economista chefe da Exploritas, gestora de recursos independentes com foco em renda variável e fixa no Brasil, falou sobre as expectativas do mercado doméstico e internacional para 2020, os desafios fiscais no Brasil e na Argentina, o cenário para o crescimento global, eleições americanas e também os principais riscos para o investidor se atentar neste ano.
First Minute baseline	hoje eu converso com andreas pacote economista chefe da explore tas uma gestora de recursos independentes com foco em renda fixa e variável no brasil e na américa latina a gente vai falar sobre as expectativas do mercado doméstico e internacional para dois mil e vinte os desafios fiscais no brasil e na argentina o cenário para o crescimento global e as eleições americanas e por aí vai se inscreva no nosso canal e fique conosco andrei recentemente você se tornou sócio da explorer tas com você também entrou o edson sarti que é um gestor e um trader bastante experiente né é foi um ano bastante turbulento para as portas no ano passado principalmente por conta dos choques com argentina conta para mim o que que muda com a entrada de vocês bom obrigado pela oportunidade de estar aqui é compartilhando com vocês a nossa visão é o explore todas desde o início ele ele é um fundo que tem
TextRank Top 2 sentences	O consenso né entre os economistas é que é esse tipo de imposto ele é regressivo é no sentido de que é por exemplo pobre faz muita ele consome bastante para parte da sua renda ele faz muito a transação é financeira EE ele acabaria pagando é mais do que o rico não é é por conta disso apesar de ser em valores bem pequenininhos não é é tem essa questão da agressividade do lado bom da do imposto como de transações financeiras é. É essa tendência mundial já vem de muito tempo né ela é muito estudada tem a ver com é questões é democrata demográficas não é desculpa é as pessoas é tão vivendo mais aumenta expectativa de vida então as pessoas tem que aumentar a poupança ao longo da vida é então é é um fenômeno longo e está muito difícil ver isso se revertendo rapidamente né se a gente olhar para para as inflações nos países desenvolvidos.
TextRank Top 5 sentences	Ou uma eleição com mais sal não é com um cara mais esquerda como Sanders ou é Elizabeth Warren isso isso é coisa rápida é entre fevereiro e meados de março que a gente vai saber é isso vai determinar como é que se comporta o resto do ano né é mas certamente vai ser um tema que vai vai vai acompanhar os mercados aí esse ano não é difícil não ser diferente agora falando em economia doméstica é o crescimento do PIB brasileiro tu achas que ele vai conseguir se descolar? É sumido não é isso é o Chile especificamente é em relação a todos né por exemplo na Colômbia agora dia 21 a gente tem de 21 de janeiro é tem marcada e uma já 11 manifestação grande tudo mais que a gente tem que acompanhar de perto como é que vai ser é mas o fato é que acalmaram bem né e principalmente no Chile em que que isso ganhou proporções muito grandes né é para para para as nossas teses de investimento né o histórico do nosso fun? É está é atacando uma parte desses gastos obrigatórios a outra parte é que é muito importante atacar é os gastos com pessoal é funcionalismo público que é exatamente o que essas essa reforma administrativa essa PEC emergencial fazem né então é é eu acho que assim para efeito de mercado essas reformas podem ter um impacto parecido com o que teve o teto de gastos lá atrás né que foi bem bastante positivo? O consenso né entre os economistas é que é esse tipo de imposto ele é regressivo é no sentido de que é por exemplo pobre faz muita ele consome bastante para parte da sua renda ele faz muito a transação é financeira EE ele acabaria pagando é mais do que o rico não é é por conta disso apesar de ser em valores bem pequenininhos não é é tem essa questão da agressividade do lado bom da do imposto como de transações financeiras é. É essa tendência mundial já vem de muito tempo né ela é muito estudada tem a ver com é questões é democrata demográficas não é desculpa é as pessoas é tão vivendo mais aumenta expectativa de vida então as pessoas tem que aumentar a poupança ao longo da vida é então é é um fenômeno longo e está muito difícil ver isso se revertendo rapidamente né se a gente olhar para para as inflações nos países desenvolvidos.
MBART vanilla	Hoje Hoje eu converso com Andreas pacote economista chefe da explore TAS uma gestora de recursos independentes com foco em renda fixa e variável no Brasil e na América Latina a gente vai falar sobre as expectativas do mercado doméstico e Internacional para 2020 os desafios fiscais no Brasil e na Argentina o cenário para o crescimento global e as eleições americanas e por aí vai se inscreva no nosso canal e fique conosco. Andrei recentemente você se tornou sócio da Explorer TAS com você também entrou o Edson sarti que é um gestor e um trader bastante experiente né é foi um ano bastante turbulento para as portas no ano passado principalmente por conta dos choques com Argentina conta conta para mim o que que que muda com a entrada de vocês. Bom obrigado pela oportunidade de estar aqui é. Compartilhando com vocês a nossa visão.
XLSUM vanilla	O ano passado foi um ano bastante turbulento para as portas.

TABLE V
SUMMARIES GENERATED BY EACH FINETUNED MODEL FOR A GIVEN EPISODE IN PORTUGUESE.

Model	**Predicted summary**
Episode description by creator	Andrei Spacov, sócio e economista chefe da Exploritas, gestora de recursos independentes com foco em renda variável e fixa no Brasil, falou sobre as expectativas do mercado doméstico e internacional para 2020, os desafios fiscais no Brasil e na Argentina, o cenário para o crescimento global, eleições americanas e também os principais riscos para o investidor se atentar neste ano.
LongMBART + finetuned PT podcasts	No episódio de hoje, eu converso com o Andreas Paquito, Economista-chefe da Exploritas, uma gestora de recursos independentes com foco em renda fixa e variável no Brasil e na América Latina. Falamos sobre as expectativas do mercado doméstico e internacional para 2020, os desafios fiscais no Brasil e na Argentina, o cenário para o crescimento global e as eleições americanas.
LongMBART + finetuned EN podcasts	In this episode, I spoke with Andreas Pak economist, chefe da Explore TAS, about the expectations of mercado doméstico e Internacional para 2020, os desafios fiscais no Brasil e na Argentina, o cenário para o crecimiento global e as eleições americanas.
LongMBART + finetuned PT/EN podcasts	Andreas Paco, Economista-Chefe da Explorer TAS, fala sobre as expectativas do mercado doméstico e internacional para 2020, os desafios fiscais no Brasil e na Argentina, cenário para o crescimento global e eleições americanas.
MBART + finetuned PT podcasts	O economista-chefe da Exploritas, Andrea Pacote, conversou com o economista-chefe da Exploritas, Edson Sarti, sobre as expectativas do mercado doméstico e internacional para 2020, os desafios fiscais no Brasil e na Argentina, o cenário para o crescimento global e as eleições americanas.
MBART + finetuned EN podcasts	Andreas pacote economista and Andreas pacote economista chefe Andreas pacote economista Chefe da Explo Explo Explo Explo Explo Explo Exploration TAS, uma gestora de recursos independentes with a Andreas pacote economista Chefe da Explo Explo Explo Explo Explo Explo Explo Exploration TAS, is a co co co co co co co co Gabriel Andreas pacote economista and Andreas pacote economista Chefe da Explo Explo Explo Explo Explo Explo Explo Explo Exploration TAS, a gestora of ER Andreas pacote economista and Andreas pacote economista Chefe da Explo Explo Explo Explo Exploration TAS, a gestora of explore explore explore TAS, a gestora de recursos independents, uma gestora de recursos independientes, with a geo
MBART + finetuned PT/EN podcasts	Neste episódio, o economista-chefe da Explorer TAS, Andreas Pacote, fala sobre as expectativas do mercado doméstico e internacional para 2020, os desafios fiscais no Brasil e na Argentina, o cenário para o crescimento global e as eleições americanas.

TABLE VI
SUMMARIES GENERATED BY NON-FINETUNED MODELS FOR A GIVEN EPISODE IN ENGLISH.

Model	**Predicted summary**
Episode description by creator	In this episode, Dr. Lisa and Tom Gleason introduce listeners to the fundamentals of sound healing, including the origins and benefits of this therapeutic practice. — This episode is sponsored by · Anchor: The easiest way to make a podcast. https://anchor.fm/app
First Minute baseline	We've been having so much fun making these podcasts. If you're thinking about making a podcast you should think about anchor anchor is the easiest way to make a podcast. Let me explain a little bit about this creation tool. It's free these tools allow you to record and edit your podcast right from your phone or your computer and then anchor distributes your podcast for you, so it can be heard on Spotify Apple podcast and all other major podcasting platforms. And here's the best part you can make money from your Podcast with no minimum listenership. It's everything you need to podcast in one place. Just go ahead and download the free anchor app or go to Anchor dot f m– to get started. Hello everyone, and thank you for tuning in to Good Vibration sound healing the Art and Science of vibro acoustic sound therapy. I really appreciate
TextRank Top 2 sentences	So a sound healing was just kind of a natural progression from my music and I started to experiment a little with sound frequencies in songs and things of that nature and I just got really excited about the power of sound and we all know that a song can certainly touch Us in such a deep way and it's the same thing for for for sound healing as well. In fact, we were trying not to he ate music and not trying to organize sound but to distill sound down to notes and use those very intentionally and specifically and what I thought was really interesting was the fact that sometimes what didn't sound necessarily musical had some of the highest healing properties right such as gongs and things like that that just reverberates so deeply, but it was very difficult for me because I immediately when I hear Sam Sound of any kind, I immediately tried to give it Melody and Harmony and I create around that in a musical way.
TextRank Top 5 sentences	We'll talk a little bit more about myths the tools the benefits and of course the science behind this amazing therapy before we get into talking more about sound healing or sound therapy. So a sound healing was just kind of a natural progression from my music and I started to experiment a little with sound frequencies in songs and things of that nature and I just got really excited about the power of sound and we all know that a song can certainly touch Us in such a deep way and it's the same thing for for for sound healing as well. In fact, we were trying not to he ate music and not trying to organize sound but to distill sound down to notes and use those very intentionally and specifically and what I thought was really interesting was the fact that sometimes what didn't sound necessarily musical had some of the highest healing properties right such as gongs and things like that that just reverberates so deeply, but it was very difficult for me because I immediately when I hear Sam Sound of any kind, I immediately tried to give it Melody and Harmony and I create around that in a musical way. We looked at a variety of training organizations from organizations in California to New York City and we kept coming back to a smaller training organization in New York state and in particular in the area that we spend a lot of time up in the Catskills and we Sound Stage Academy of sound healing and Tom and I that's kind of our happy place going up to the Catskills. I think it's the only practice that I can think of that is fully reciprocal right like while we are practicing and Tom and I do sessions one-on-one with clients or we'll do a session where Tom and I worked together on a client which is, you know, really a like a full bonus kind of session and then we also work in group sound therapy or what we call sound baths where a large group of clients will come together.
MBART vanilla	We've been having so much fun making these podcasts. If you're thinking about making a podcast you should think about anchor anchor is the easiest way to make a podcast. Let me explain a little bit about this creation tool.
XLSUM vanilla	This is a full transcript of BBC Radio 4's Welcome to the Sound of Music series.

TABLE VII
SUMMARIES GENERATED BY FINETUNED MODELS FOR A GIVEN EPISODE IN ENGLISH.

Model	**Predicted summary**
Episode description by creator	In this episode, Dr. Lisa and Tom Gleason introduce listeners to the fundamentals of sound healing, including the origins and benefits of this therapeutic practice. — This episode is sponsored by · Anchor: The easiest way to make a podcast. https://anchor.fm/app
LongMBART + finetuned PT podcasts	O que é Sound Healing? O
LongMBART + finetuned EN podcasts	Welcome to Good Vibration, Sound Healing, the Art and Science of Vibro Acoustic Sound Therapy. In this episode, Dr. Lisa Gleason and Dr. Tom Gleason discuss all aspects of vibrational sound healing and try to help you gain an understanding of how to achieve personally a body in balance and an overall sense of well-being.
LongMBART + finetuned PT/EN podcasts	In our first episode, Tom and Dr. Lisa discuss the science behind Sound Healing and the use of vibrational sound healing.
MBART + finetuned PT podcasts	Neste episódio, Tom Gleason e Lisa Gleason falam sobre a importância do vibracional sound para a saúde e bem-being.
MBART + finetuned EN podcasts	Welcome to Good Vibration Sound Healing, the Art and Science of Vibrational Sound Therapy. I really appreciate you spending some time with us today. My name is Tom Gleason, and I'm here with my wife Dr. Lisa Gleason, and we are both certified sound practitioners and extremely passionate about this subject. Check in these shows we're going to be discussing in depth all aspects of vibrational sound in healing and try to help you gain an understanding of how to achieve personally a body in balance and overall sense of wellbeing.
MBART + finetuned PT/EN podcasts	Welcome to the first episode of Good Vibration Sound Healing with Dr. Tom Gleason and Dr. Lisa Gleason. In this episode, we introduce the concept of sound and vibration ultherapy, and talk about the various applications over time. We talk a little bit more about myths, tools, benefits, and of course the science behind this amazing therapy. Before we get into talking more about sound healing or sound therapy, we talk a little bit about how we came into this practice.