Spotify Podcast Disagreement Detection

Statement of Work

Prepared by	Connor Capitolo, *connorcapitolo@g.harvard.edu* Max Li, *manli@fas.harvard.edu* Morris Reeves, *morrisreeves@g.harvard.edu* Jiahui Tang, *jiahuitang@g.harvard.edu*
Prepared for	Rosie Jones, *rjones@spotify.com* Jussi Karlgren, *jkarlgren@spotify.com*

Background

Our capstone project partner is Spotify, a Swedish music, podcast, audio streaming and media service provider. It is the largest audio streaming subscription service in the world, with a mission "to unlock the potential of human creativity by giving millions of creative artists the opportunity to live off their art and empower billions of fans with the opportunity to enjoy and get inspired by the music."[1] In other words, it serves as a matchmaker between content producers and consumers.

Although audio data has been frequently used in the areas of voice commands and transcription, there has been limited exploration into its relation to speakers' stances. For example, by identifying disagreements in podcasts, we would be able to better understand narrative arcs (e.g. exposition, climax, resolution) in successful and popular podcasts while empowering companies to build recommendation systems that better match users with their listening preferences. In one way, we could cluster and group pools of podcasts with similar topics and genres (for example, debate-heavy or controversial episodes with disagreements and disputes). In another way, we could give hosts insightful suggestions to switch their narrative style or forms of moderation to conduct better audio search engine optimization, so that they could match target listeners' preferences. This would result in a smoother user experience, enhance user utility and retention, and also reciprocally drive larger traffic and audiences for podcast creators, creating a dynamic ecosystem among platform creators and users.

Problem Statement

[1] https://newsroom.spotify.com/company-info/

Our goal is to create a classification model that identifies periods of disagreement in a podcast episode using the audio file and automatic transcription of the episode in order to unlock a new metric for comparisons across podcasts.

To achieve this goal, we will enrich a subset of the existing Spotify Podcasts dataset (described in the next section) with time-stamped labels indicating the presence or absence of disagreement. A disagreement detection model would allow Spotify to better identify similarity between podcast episodes and improve podcast recommendations (e.g. based on user preferences for 'contentious' podcasts or based on implied sentiment for/against a topic of interest).

Approaches and considerations

As baseline models, we plan to use logistic regression on transformed features (e.g. mel spectrogram, sentiment, dialogue act tags) for binary classification separately on the text and audio data, followed by more sophisticated sequence models including RNN and LSTM models. We plan to use F1 score, precision, and recall as evaluation metrics; these metrics are consistent with related literature on disagreement, such as Allen, Carenini, and Ng (2014) and Wang, Yaman, Precoda, Richey, and Raymond (2011).[2] As a stretch goal and time permitting, we may explore multimodal versions of these models which predict on the combined text and audio data. Foreseeable challenges include subjectivity of disagreement annotations, annotation alignment between the audio and text data, missing values in the text data in the case of overlapping speakers, speaker diarization quality, and sparsity of positive identifications of disagreement.

A successful project would include these outcomes: an enriched podcasts dataset with disagreement labels for a subset of podcast episodes, scripts which output disagreement predictions, and a report summarizing the strengths, pitfalls, and potential associated with audio-based disagreement classification. Items beyond the scope of this project include a comprehensive manual annotation of the podcast dataset, predictions of the 'magnitude' of disagreement, and improvement of transcription or diarization quality.

Resources

We will be using the Spotify Podcasts Dataset as our main source of data. This dataset comprises approximately 100,000 podcast episodes in the form of raw audio files, accompanied by 50,000 hours of automatically transcribed audio. Due to the large size of the audio (~2TB) and transcript (~13GB) data, it is imperative that we use cloud storage to have one central location where all our data can be accessed by team members, as well as cloud computing to process and model in a timely and scalable manner. One potential avenue is using AWS: we would create an S3 bucket for the data storage with the necessary permission levels to meet Spotify's data usage agreement, and use either EC2 or Sagemaker to perform experimentation and modeling. A more cost effective option for the team would be using the Harvard Faculty of

[2] https://emnlp2014.org/papers/pdf/EMNLP2014124.pdf, https://aclanthology.org/P11-2065.pdf

Arts & Sciences Research Computing (FASRC) compute cluster, known as Cannon.[3] As Harvard students conducting a class research project, we should have access to the 60PB of storage and the CPU/GPU nodes available on the cluster; however, this approach may add unforeseen administrative overhead that can take too much time given our project timeline.

We are also considering the use of University College London's speech datasets that are freely available for researchers and students (https://www.phon.ucl.ac.uk/resource/data.html). This may provide nice supplemental data to test our models on a different corpus of audio data.

A final resource we have identified is Prodigy[4], an annotation tool that will allow us to more easily annotate Spotify's audio files and identify specific timestamps of disagreement. We are also looking into using it for active learning as more audio files are labeled. Fortunately, Prodigy has already provided us with a 3-month research license for use.

High-Level Project Stages

We plan to divide our project into four major stages of exploration:

1. **Audio annotation**: We will use Prodigy to manually annotate a few podcasts by disagreement types in this stage to explore the data.
2. **Exploratory data analysis and visualization**: Since our dataset contains very large volumes of text and audio, we would need to pre-process and clean the data before performing further analysis. In addition, we will visualize the data to grasp a rough understanding of how the podcasts are distributed in length and genre, which could be helpful to the sampling process of disagreement detection.
3. **Text sentiment exploration**: This aims to help us find a subset of podcasts with high negative sentiment scores to explore if this may be a useful feature for prediction of disagreement, to see if there are any noticeable genre clusters in negative sentiment, and to guide genre areas it may be worth focusing on for manual annotation. We also hope to assess the quality of the transcript data and provide some intuition into the occurrence of disagreement in the audio data. If for some reason stage 4 seems unrealistic given our limited time frame, we will perform a more traditional sentiment analysis on the text data.
4. **Constructing disagreement detection model for audio data:** at this stage, we would hope to use machine learning methods to automatically detect or annotate disagreements in podcasts. This could become an ambitious goal as we further explore the data and study more relevant research, so we are aware that this may be beyond the scope of our investigation.

[3] https://www.rc.fas.harvard.edu/about/cluster-architecture/
[4] https://prodi.gy/docs/audio-video

Project Timeline

Sprint ending	Tentative milestone or goal

9/23	• **Logistics**:
	  ○ Slack channel setup (with partner) + weekly calendar invite
	  ○ Request genre data, dialogue acts model name
	• **Annotation sprint:**
	  ○ Goal: Agree upon definition of disagreement
	  ○ Download subset of audio files through Box
	  ■ Test different packages for audio annotation
	• **Annotation plan:**
	  ○ Goal:
	  ■ Agree upon packages/software for audio annotation
	  ■ Decide granularity of annotation
	  ■ Decide sampling method for podcasts
	• **Ignite Talk slides**

9/30	• **Database storage** (S3 bucket setup):
	  ○ Download on S3
	  ○ Group AWS account w/ IAM users
	  ○ EC2 instance or SageMaker
	• **Text sentiment** exploration:
	  ○ Goal:
	  ■ Understand alignment between genres and subsets/clusters of podcasts with high negative sentiment scores
	  ■ EDA / gauging quality of transcript data
	• **Enrichment sprint:**
	  ○ Goal: annotate subset of podcasts w/ disagreement labels
	• **Literature review**

10/7	• **Milestone 1**:
	  ○ **Summary Report** and **Technical Report** with:
	  ■ Description of work to date
	  ■ Summary metrics on annotation progress (counts, examples, and challenges)
	  ■ Results of EDA (text sentiment clusters by category)

10/28	• **Milestone 2:**
	  ○ **Updated Summary Report and Technical Report** with**:**
	  ■ **Preliminary audio classification model** results (trained on annotated data)
	  ■ (Tentative) **preliminary text model** results

11/18	● **Milestone 3:** ○ **Updated Summary Report and Technical Report** with**:** ■ **Final audio model** results ■ (Tentative) updated text model results
12/16	● **Final Deliverables:** ○ Research paper (DEC 15) ○ Impact statement (DEC 15)