Disagreement Detection in Spotify Podcast Episodes

Connor Capitolo
Harvard University
connorcapitolo@g.harvard.edu

Max Li
Harvard University
manli@g.harvard.edu

Morris Reeves
Harvard University
morrisreeves@g.harvard.edu

Jiahui Tang
Harvard University
jiahuitang@g.harvard.edu

Abstract

We study the detection of disagreement in Spotify podcast episodes based on audio data and automatic transcriptions. We apply unsupervised and supervised approaches for disagreement detection to evaluate whether combining prosodic and lexical features can improve detection relative to utilizing each dataset individually. We also explore which avenues of feature engineering may be promising for disagreement detection and assess the utility of automatic transcripts.

1 Introduction

With the rise of social media platforms and online forums, the public sharing of opinions to wide audiences is easier than ever; disagreement is both ubiquitous yet also nebulous to strictly define. Disagreement can be instructive both at a dialogue and societal level: topics which attract disagreement can reflect pressing societal issues or engaging topic areas, whereas within a dialogue, disagreement can reflect the stances of individual speakers. Our goal is to generate insights about disagreement based on the Spotify English-Language Podcast Dataset using its audio files, corresponding automatic transcriptions, and metadata. Disagreement detection is highly beneficial to streaming services as it can lead to higher retention rates and active engagement by its listeners through more tailored recommendations, while also increasing audience traffic and creating a more loyal fan base for podcast creators; in short, disagreement detection can lead to a better podcast user and creator experience.

2 Related Works

2.1 Domains and media

Previous work on disagreement detection has focused primarily on modeling text data using dependency relations and lexical features such as n-grams. (Xu et al., 2019) utilize pairs of tweets with stance labels (Favour/Against/None) grouped by topic from the SemEval-2016 dataset, defining disagreement as pairs whose stance labels differ. In the realm of online forums, (Gokcen and de Marneffe, 2015) utilize quote-response pairs with agreement scores between -5 and 5 from the Internet Argument Corpus (IAC). Similarly, in addition to the IAC dataset, (Wang and Cardie, 2014) use the AAWD (Authority and Alignment in Wikipedia Discussions) dataset consisting of annotations of disagreement/agreement/neutral on English Wikipedia Talk pages.

Nonetheless, using audio and text features for disagreement detection is not without precedent. Existing works which employ audio data for disagreement detection utilize low-dimensional prosodic features such as speech rate and fundamental frequency. For example, (Hillard et al., 2003) use transcribed ICSI meeting recordings, manually annotated with positive, backchannel, negative, and other labels; statistics on pause duration, speaker-normalized vowel duration, and fundamental frequency are employed as features alongside word-based features including negative and positive keyword counts. Similarly, (Wang et al., 2011) analyze English broadcast conversation data from the DARPA GALE program annotated for 11 (dis)agreement-related labels, utilizing prosodic features including pause, duration, speech rate, and pitch alongside lexical features.

In this context, the Spotify Podcasts dataset stands out in both the volume and diversity of its audio content (across various genres and podcast creators), and presents an opportunity to utilize both prosodic and lexical features to explore characteristics of human speech associated with disagreement. In particular, (Wang et al., 2011) find sarcasm to be a common source of error in disagreement classification; the large volume of combined

text and audio data in the Spotify Podcasts dataset may enable progress in this domain. We aim to contribute to the existing literature by utilizing higher dimensional audio features such as the spectrogram representation, as well as combining features from both the audio and transcripts for disagreement detection.

2.2 Definitions of disagreement

A central challenge to disagreement annotation is the inherent subjectivity in annotations: while some works in the literature provide explicit definitions of disagreement, most implicitly define disagreement through data processing decisions.

The most similar definition of disagreement in the literature, which we adopt for our annotations, is that of (Wang et al., 2011), who define disagreement as follows:

> "(dis)agreement occurs when a responding speaker agrees with, accepts, or disagrees with or rejects, a statement or proposition by a first speaker"

To this definition, we add that the disagreement must be immediately perceptible by a listener; for example, we exclude sighs or laughter which might otherwise be construed as the onset of disagreement upon rewinding. This definition notably (and deliberately) excludes single-speaker podcast episodes, as well as audio segments in which speakers explain the reasons for their disagreement, and audio segments in which speakers respond to soundbites.

Regarding the granularity of disagreement windows, (Wang et al., 2011) and (Wang and Cardie, 2014) define disagreement at the utterance (sub-turn) level, whereas (Hillard et al., 2003) annotate disagreement at the "spurt" level (period of speech by a single speaker, with no pauses longer than 0.5 seconds). We define the granularity at the utterance level in order to allow for detection of sub-windows of disagreement within long uninterrupted speech segments, which is not uncommon in podcasts. Given the dialogic nature of most podcast episodes, this definition naturally diverges from that of (Xu et al., 2019) in which disagreement is non-dialogic (independent, differing, stance-bearing tweets on a topic).

3 Podcast Metadata

The Spotify Podcast Dataset consists of $105,360$ episodes across $18,360$ shows, amounting to around $50,000$ hours of audio and over 600 million words. The podcasts are sampled from Jan 1, 2019 to March 1 2020, filtered to the English language. Since there is a mix of both professional and amateur creators, the podcasts are diverse in terms of audio quality, topics, and structural formats.

We looked into three different types of data in the dataset to explore their features and distributions. The first is the automatic transcriptions data, which contain speaker tags (based on automatic diarization) and timestamps for each word in the transcripts of the podcasts. It is worth noting that the speaker tags are automatically generated by Google's Speech To Text API, which can be inaccurate especially when multiple speakers are talking simultaneously. The second type of data is the audio from the podcasts, which will be discussed in future sections.

The third type of data is the metadata of the podcasts. Some of the useful features from the metadata file summarized by Spotify include: *episode_title, author, category, subcategory, show_name, show_description, publisher, language, episode_name, episode_description, duration*. Note that the metadata is provided by the creators of the shows, so some features might not be very reliable. For instance, creators may incorrectly label their podcasts as a more popular genre to gain a larger audience.

We performed some exploratory data analysis on the metadata of the podcasts, focusing on some of the features mentioned above. In Figure 1, we have a Pareto chart of the total number of shows for each category.

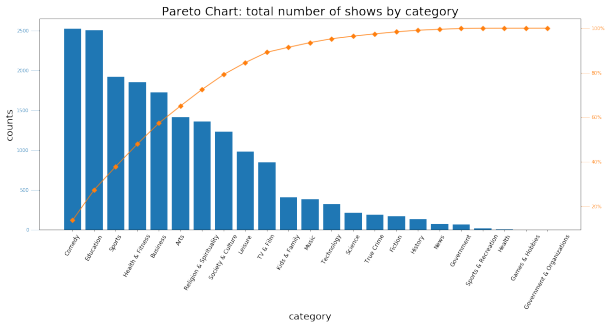

Figure 1: Pareto Chart of Total Number of Shows by Category

There are 23 major categories for the shows in the metadata. We plotted the total number of shows for each category. We can see that among the $18,360$ shows, Comedy and Education have the most number of shows, while Sports and Health &

Fitness ranked the third and the fourth, respectively. In terms of detecting disagreement, we expect Comedy and Sports podcasts with multiple speakers to contain many disagreement segments. We need to verify if the abundance of shows in these categories would yield an abundance of disagreements in the podcasts.

Also, we noticed that some categories with the least amount of shows have similar names as the top ones, such as Sports & recreation and Health. As mentioned before, creators may label their podcasts as a more popular genre to get a larger audience. It is possible that the first group of creators preferred one tag over a similar one (for instance, choosing Sports over Sports & Recreation), and then subsequent creators followed these increasingly popular ones.

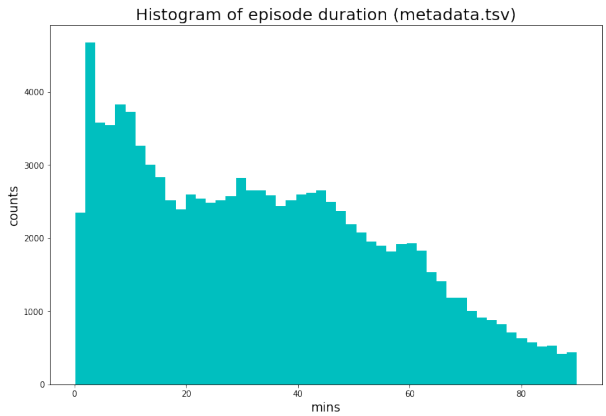

Figure 2: Histogram of Episode Duration. Only episodes shorter than 90 minutes are included for better visualization

Looking at Figure 2, we take the same approach as the data description paper (Clifton et al., 2020) to only visualize the episodes that are less than 90 minutes long (there are 125 episodes that exceeded 90 minutes in length that are not shown in graph). We see that most episodes are less than 20 minutes long; in general, the number of episodes decreases as the length increases. Also, a large number of podcast episodes last less than 5 minutes. Many of these are trailers or music remixes, in which we would not expect to find a high concentration of disagreements. Overall, this plot mainly serves to see if our data aligns with the description.

Looking at Figure 3, we see that a vast majority of the shows have less than 5 episodes with a few shows that have over 100 episodes. Relating to the distribution of episode duration, we found out that the shows with large number of episodes are

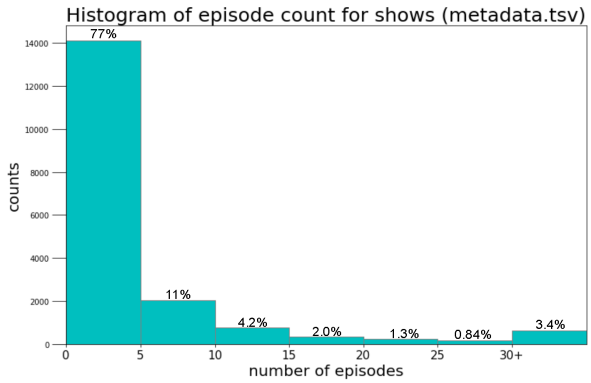

Figure 3: Histogram of Number of Episodes per Show

mainly shows for music remixes.

As discussed at the start of this section, we would like to use metadata as a guide towards finding podcasts with disagreements so that we can focus specific ones to annotate. Our first attempt was to search for the word "disagreement" in the podcast descriptions and listen to those. This is not the most efficient approach, and we have discussed potential unsupervised learning approaches that might be better able to find podcasts with disagreements; however, just our manual approach was able to find podcasts that contain disagreements.

4 Audio Analysis

The following audio and text analysis sections will be revised as analysis progresses; below is purely for status reporting purposes but will likely not be part of the final research paper.

We used the Python package Librosa to explore what kinds of information we could obtain from audio data. Unlike other data sets that naturally come with features, we needed to explore what level and granularity of information we could construct from each audio data file and how to pre-process raw data for model building. Librosa allows for transforming .ogg audio files into waveform metrics and sample rates.

The following analyses are based on a single selected annotated episode which contained many segments of disagreement between a man and woman on the topic of dog ownership, which was a particular 10 minute segment of the entire podcast episode since it contained high levels of disagreement that met our definition. Three different types of graph (waveform, spectrogram, chromaplot) were visualized to compare whether there is any significant difference between regions of disagree-

ment and non-disagreement.

In the waveform plot, we have time series amplitude metrics with an envelope view in monophonic, stereo, harmonic and percussive decomposition views. There is no obvious visual difference when we compare disagreement with non-disagreement sub-segments of the episode.

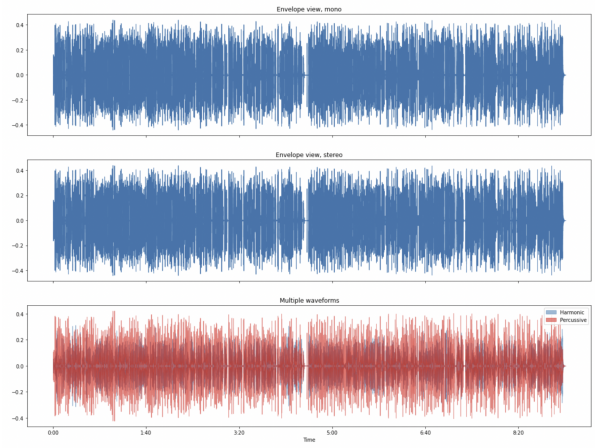

Figure 4: Envelop View of Waveform

We also encoded audio information from signals into time-frequency (Hertz) by plotting a spectrogram, with the x axis being the time series, and y being the frequency measurement of the audio.

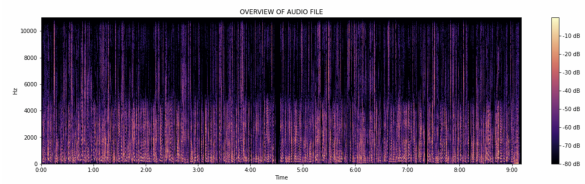

Figure 5: Time-Frequency Spectrogram

For pitch class information, we plotted a Chromagram for the episode, which gives us information about pitch class profile and pitch intensities. It could be particularly useful for analyzing audio whose pitch could be meaningfully categorized.

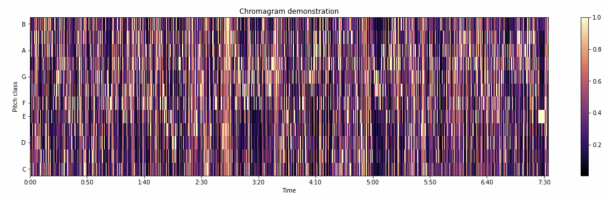

Figure 6: Chroma Plot with Pitch Class Intensity

We also explored the YAMNet embeddings for the same podcast episode, to see if there is any association between our manual disagreement annotations and these vector representations of audio segments. YAMNet is a deep net that employs the MobileNetV1 and predicts 521 audio event classes (trained on AudioSet-YouTube corpus). We utilized pretrained embeddings available through Spotify; there is an embedding of dimension 1024 for each 0.48 seconds chunk. These chunks are visualized via t-SNE in Figure 7 together with our annotations. There is no obvious separability based on the scatterplot when visualized in two dimensions, which we hypothesize is due to: (1) 0.48 seconds being too short a window to meaningfully expect differences in audio characteristics, (2) only a subset of YAMNet event classes may be relevant for disagreement, and even then only in rare cases (e.g. shouting would be expected to be limited to only the most extreme of disagreements), and (3) there may be higher dimensional structure to the data that is not discernible from the figure.

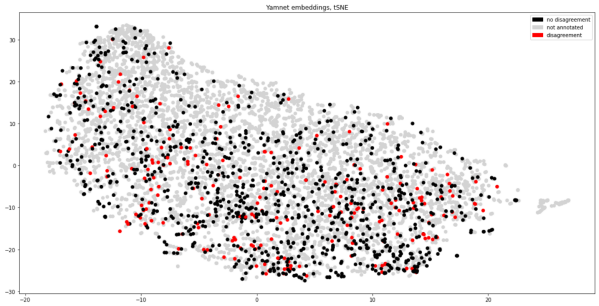

Figure 7: YAMNet Embeddings (tSNE Projection) for single annotated episode

5 Text Sentiment Analysis

Text transcripts can provide additional features beyond those directly available in the audio transcripts since they encode semantics and lexical features such as phrases associated with disagreement (e.g. “First of all..”).

The text data are automatically segmented based on speaker pauses in Google’s Speech to Text API. As preliminary text analysis, we first explored whether there is any visually apparent relationship between binary sentiment scores (which we hypothesize might be associated with disagreement) and the averaged word embeddings of each text segment. We utilize a pretrained sentiment model (DistilBERT base uncased finetuned SST-2) provided in the huggingface library, and utilize Google’s pretrained 300-dimensional word2vec vectors (trained on the Google News dataset of approx. 100 bil-

lion words). This is visualized using PCA for a single podcast episode (the dog-related disagreement clip referenced in the audio analysis section) in Figure 8 below; each point corresponds to a text segment (as segmented by Google's speech to text API), colored by our manual annotation (red for disagreement).

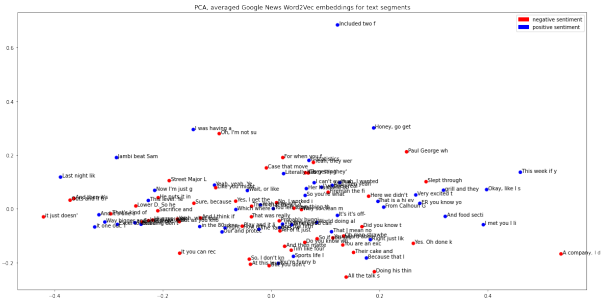

Figure 8: PCA, Averaged Word2Vec Vectors for Single Podcast Episode

This preliminary view exhibits no obvious association between the word vectors and sentiment scores, which we hypothesize is due to several factors. Firstly, the automatic segmentation of text may not faithfully capture turns between speakers; secondly, disagreement is relative to a previous statement by a speaker, so 'negative' statements may actually comprise agreement if, for example, the previous speaker shares the same negative view on a topic. More fundamentally, the word vectors utilized are not contextualized embeddings, and moreover averaging word vectors may not provide a meaningful summarization of text segments, particularly for longer text segments. Finally, the word vectors are themselves relatively dated in the literature (from 2013) and therefore the OOV issue (words for which there are no embeddings) may be prominent, particularly in unstructured conversations on topical issues as in podcast episodes.

As an initial baseline model, we also utilized cosine similarity between the average word embedding of a user-provided query ("No, I completely disagree with you") and the averaged word embeddings for each segment (in the same dog-related disagreement episode), filtering to text segments above a certain threshold. Filtering to cosine similarity above 0.7, three of the five results returned indeed corresponded to manually annotated disagreement windows, suggesting that unsupervised approaches utilizing similarity metrics to annotated segments are a promising potential avenue for disagreement classification.

Going forward, we hope to improve this analysis by utilizing richer representations of the text segments, smarter discretization of podcast episodes into utterances, and combining the text with the audio embeddings.

References

Ann Clifton, Aasish Pappu, Sravana Reddy, Yongze Yu, Jussi Karlgren, Ben Carterette, and Rosie Jones. 2020. The spotify podcast dataset.

Ajda Gokcen and Marie-Catherine de Marneffe. 2015. I do not disagree: leveraging monolingual alignment to detect disagreement in dialogue. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 94–99, Beijing, China. Association for Computational Linguistics.

Dustin Hillard, Mari Ostendorf, and Elizabeth Shriberg. 2003. Detection of agreement vs. disagreement in meetings: Training with unlabeled data. In *Companion Volume of the Proceedings of HLT-NAACL 2003 - Short Papers*, pages 34–36.

Lu Wang and Claire Cardie. 2014. Improving agreement and disagreement identification in online discussions with a socially-tuned sentiment lexicon. In *Proceedings of the 5th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 97–106, Baltimore, Maryland. Association for Computational Linguistics.

Wen Wang, Sibel Yaman, Kristin Precoda, Colleen Richey, and Geoffrey Raymond. 2011. Detection of agreement and disagreement in broadcast conversations. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 374–378, Portland, Oregon, USA. Association for Computational Linguistics.

Chang Xu, Cecile Paris, Surya Nepal, and Ross Sparks. 2019. Recognising agreement and disagreement between stances with reason comparing networks. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4665–4671, Florence, Italy. Association for Computational Linguistics.