Disagreement Detection

Milestone 2

Connor Capitolo, Max Li, Morris Reeves, Jiahui Tang

Harvard John A. Paulson
School of Engineering
and Applied Sciences

Overview

- Problem statement and scope of work
- Data Exploration
- Audio Analysis
- Text Analysis
- Next Steps
- Q&A

What are we trying to achieve?

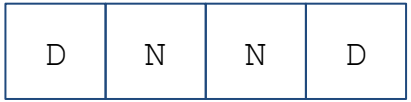
Generate insights about disagreement based on the Spotify English-Language 100k Podcast Dataset

Inputs

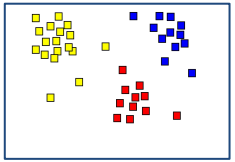
Audio, text, and metadata

Outputs

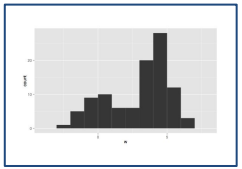
Disagreement/Non-disagreement



Insights



Can we find disagreement clusters based on the transcripts?



Does a certain pitch mean a high likelihood of disagreement?

Why is disagreement detection in podcasts important & interesting?

- **Better User Experience**
  - Tailored recommendations

podcast

- **Drive audience engagement**
- Improve style or moderation

Spotify®

- **User happier + Creators Happier = Spotify Happier**
- Expand platform ecosystem

How much has disagreement been explored in text and audio data?

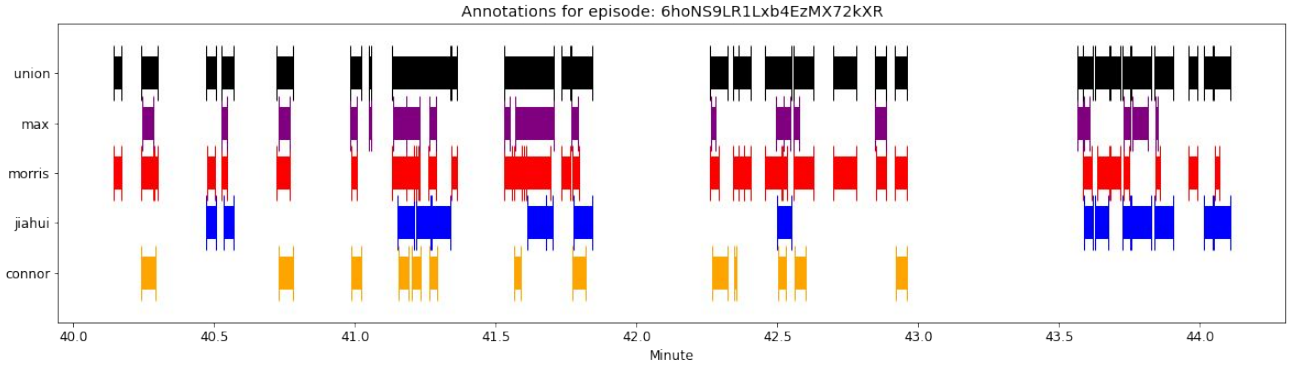
- Literature Review: Some disagreement/stance detection on text data, but not much on audio
- *Disagreement Detection on podcast audio data is super new*
- F1 metric: Balance weight on precision and recall

$$\text{F1 Score} = \frac{2 \times (\text{Precision} \times \text{Recall})}{\text{Precision} + \text{Recall}}$$

Definition of Disagreement

- Generate disagreement labels that can be used for modeling
- Definition: a speaker is **directly applying a contradiction or rejection** of another person's idea where it is **immediately perceptible** by the listener
    - Rule of thumb: If you handed this podcast to your mother, she would know it's disagreement

Dog clip: a couple discussing what it means to have a dog when single



Data

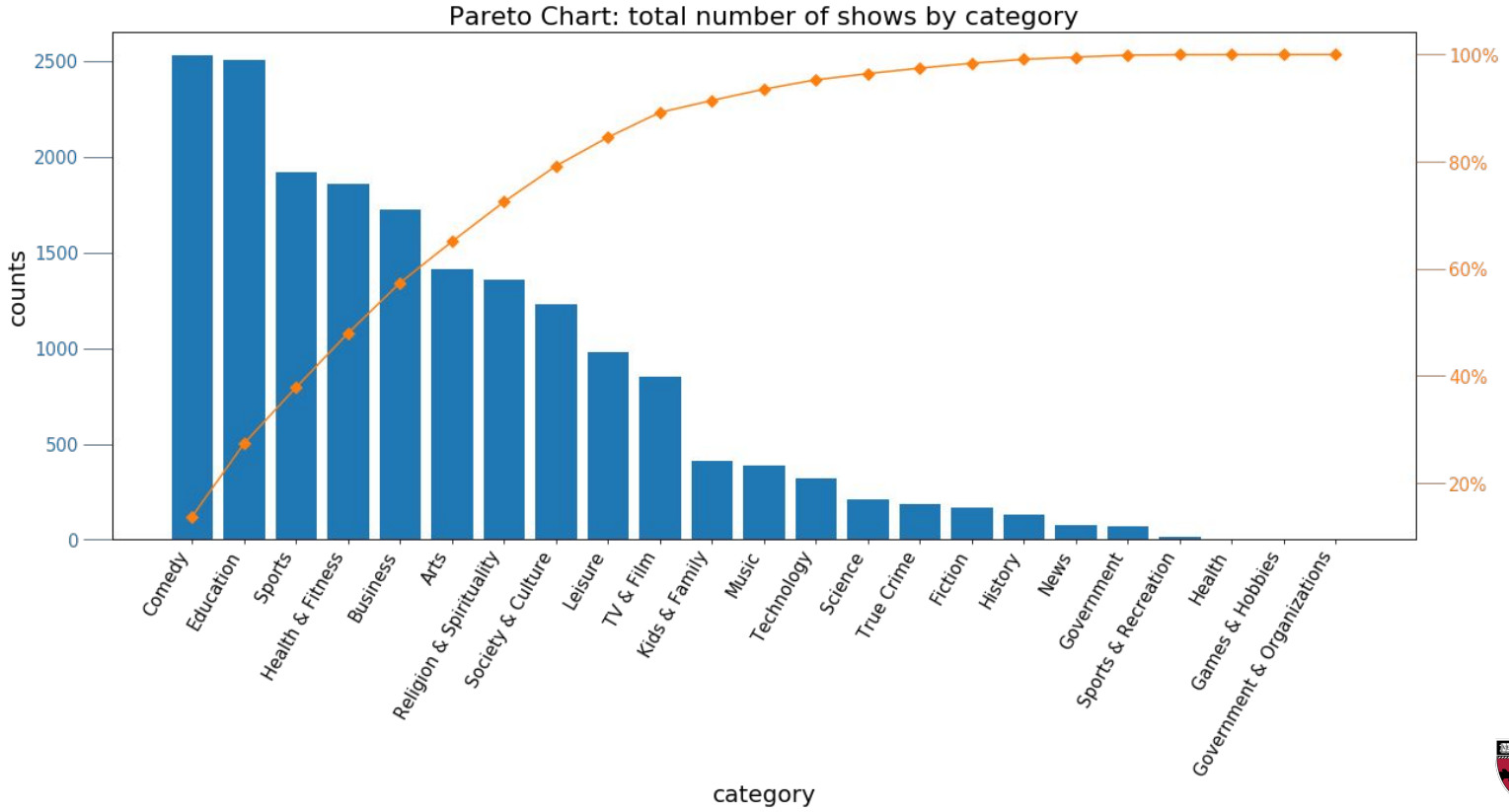
105,360 episodes, 18,360 shows

- ~ 50k hours of audio
- > 600M words
- 23 genres
- Jan 1, 2019 to March 1, 2020
- English language specific
- Professional + amateur creators

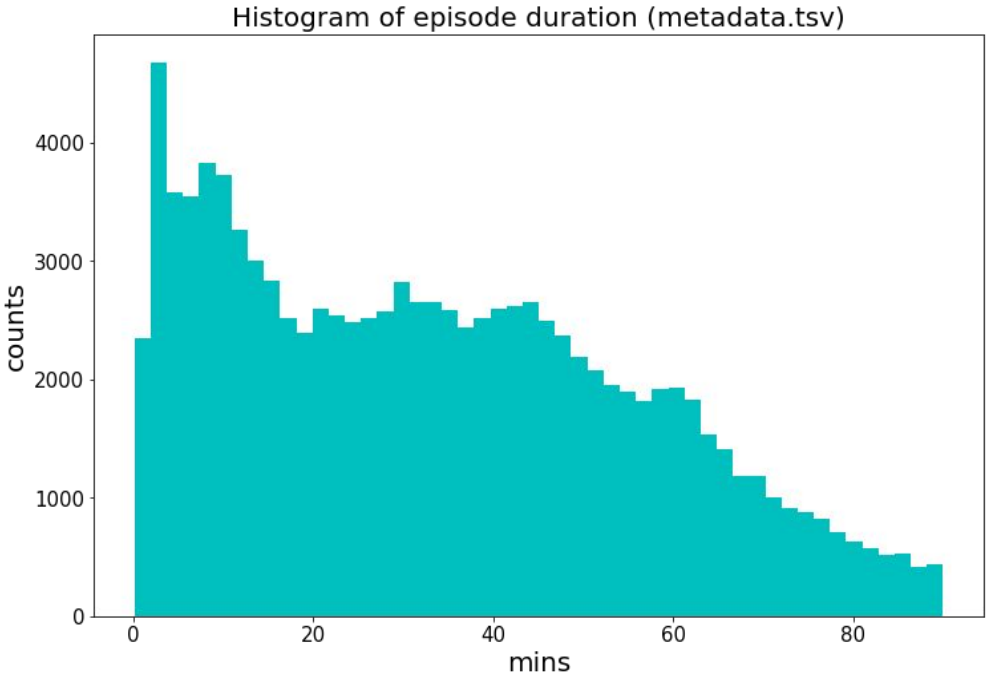
Dataset statistics

	min	average	max
minutes	<1	31.6	305.0
words	11	5,728	43,504

Data - podcast genre

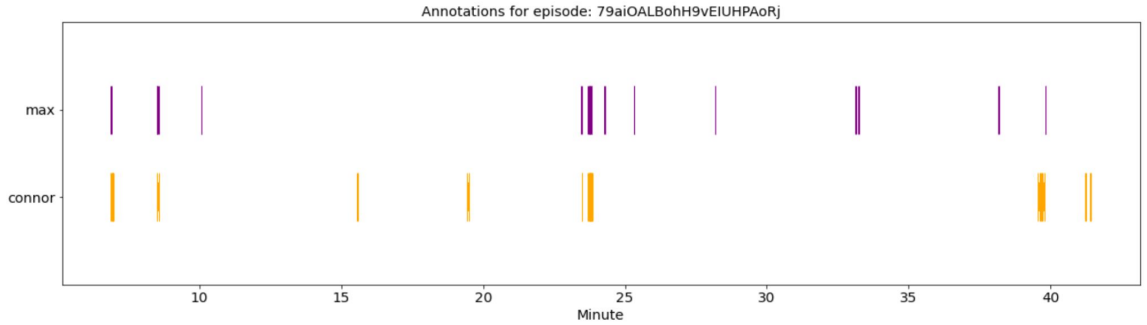


Data - episode duration

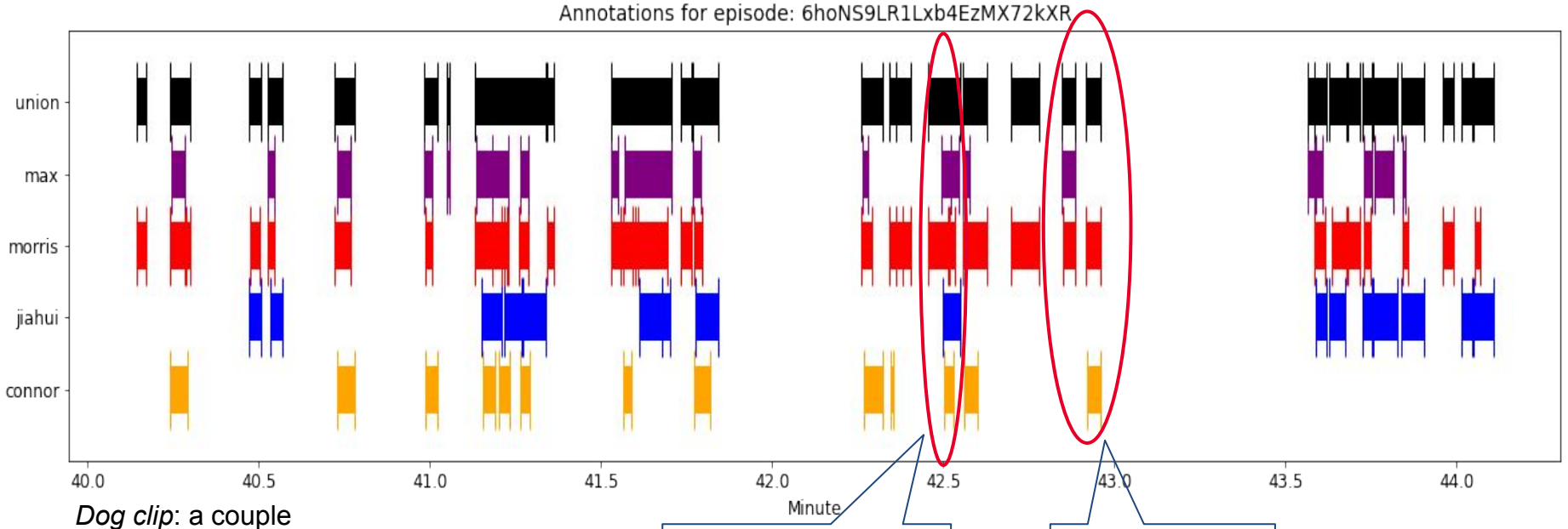


Sparsity of Disagreement

Dating + Food episode: 4 people discussing dating and restaurants in NYC



Scoping down to a 5-minute example...

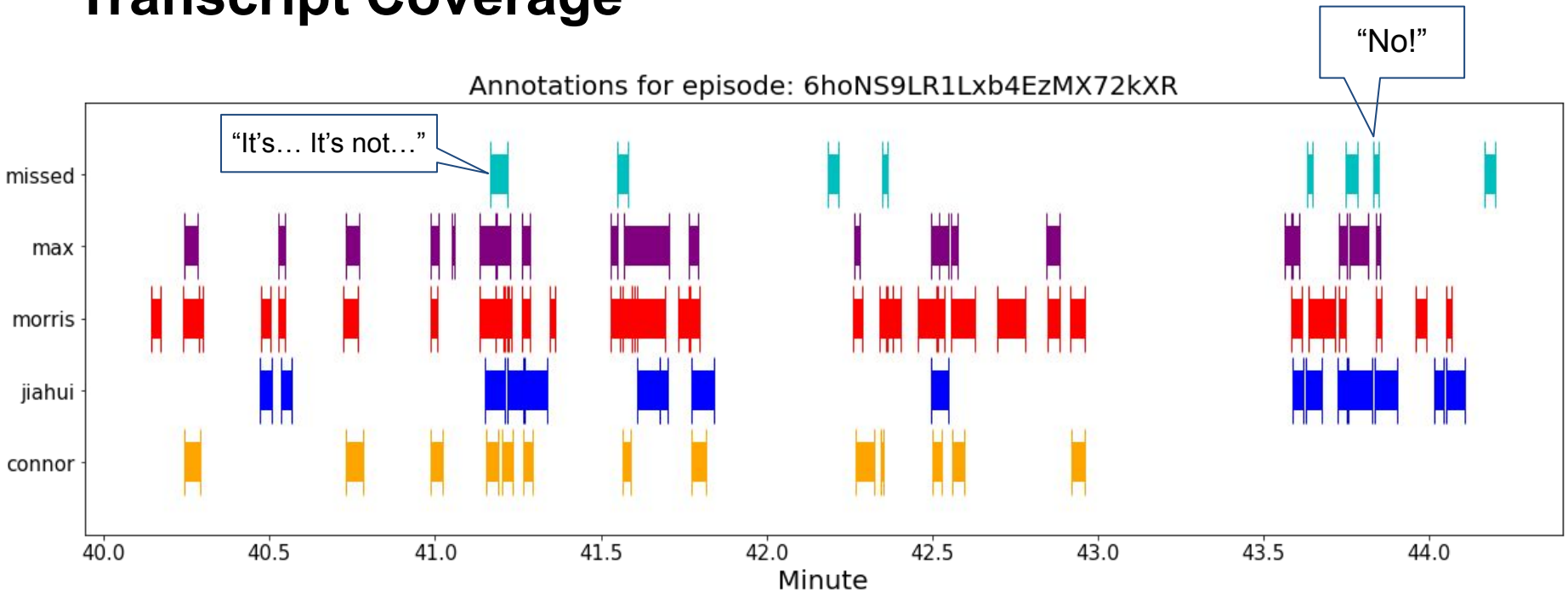


Dog clip: a couple discussing what it means to have a dog when single

"...I totally disagree with that"

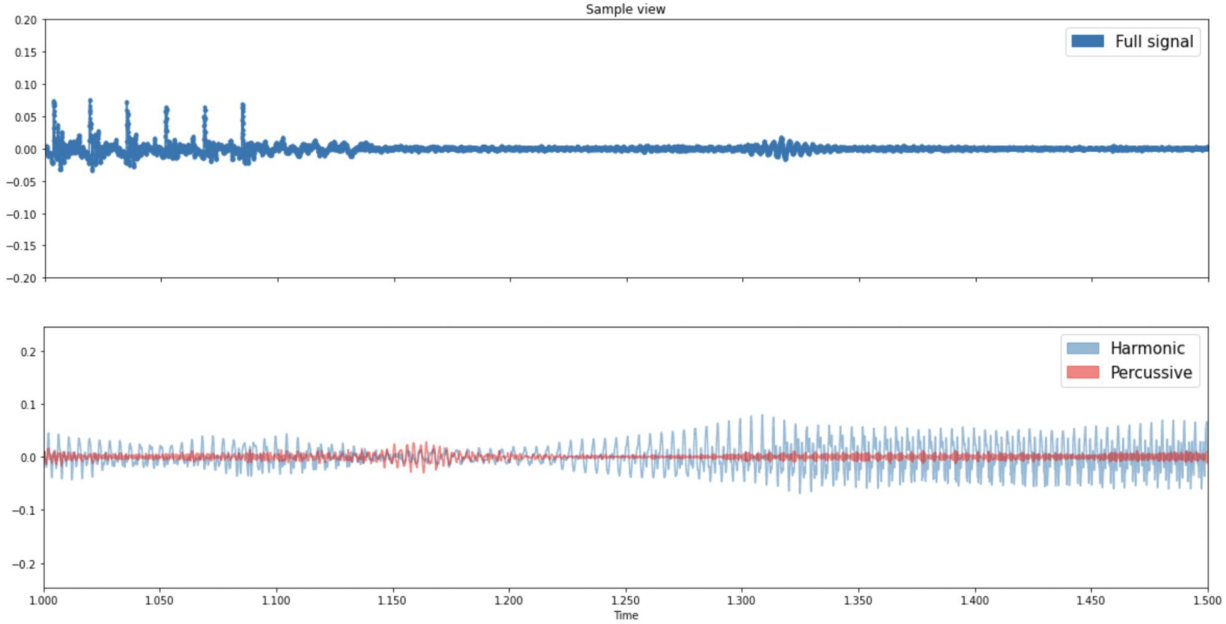
"First of all…"
"Hmmm…"

Transcript Coverage



Audio - Librosa

- Librosa is a Python package for **music and audio** analysis
- WaveForm (envelope form)
  - Monophonic/Stereo
  - Harmonic + Percussive Components



Dog clip: a couple discussing what it means to have a dog when single

Audio Analysis - Data Augmentation

- **Audio augmentation:** higher speed, slower, and different tones, or add noises etc
- ***Features***: add noise, change_tone, slow_down, speed; AddGaussianNoise, TimeStretch, PitchShift, Shift
- **Python Libraries**: Audiomentations and Pydiogment
- A simple demonstration of augmentation: https://share.streamlit.io/phrasenmaeher/audio-transformation-visualization/main/visualize_transformation.py

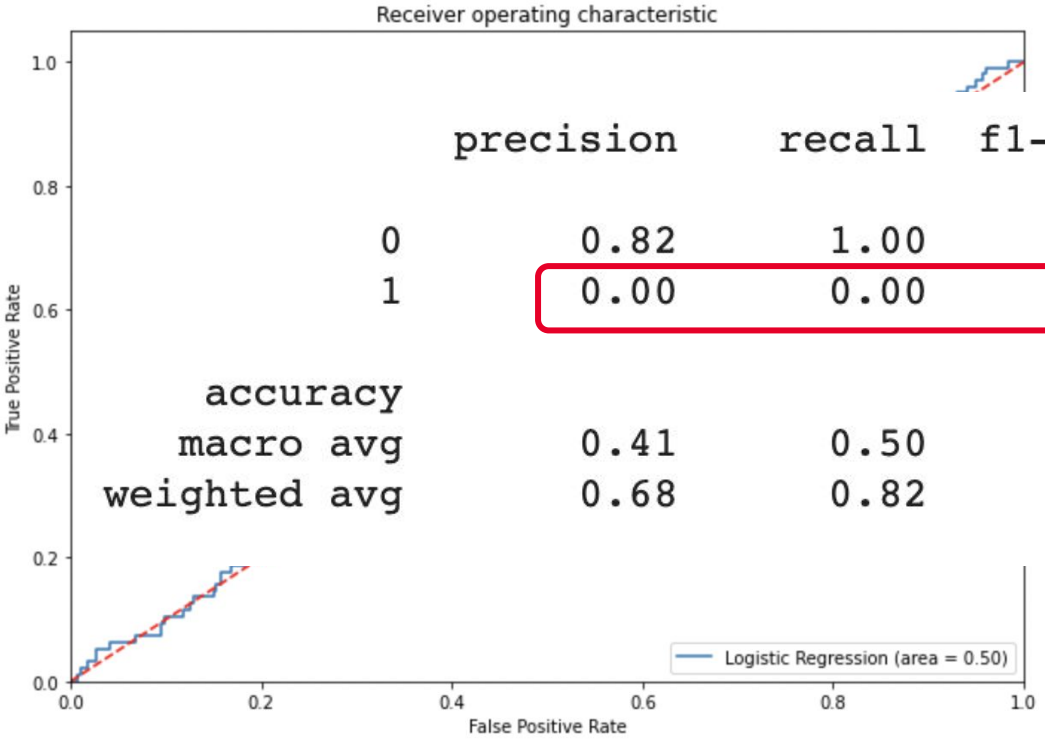
Audio Features - Wavelet Transformation

- The Wavelet Transform (WT) is a technique for analyzing signals.
- The advantage that wavelet transform has over Fast Fourier transform (FFT) is that it can capture **spectral and temporal** information simultaneously. A signal is convolved with a set of wavelets at different scales and positions.
- **Continuous Wavelet Transform (CWT)Discrete Wavelet Transform (DWT)**
  - The DWT is only discrete in the scale domain, not in the time-domain.

Baseline Model

- Logistic Regression Model
- Input X: *Original Waveform, Augmented Waveform, Variance of Aggregated Original Waveform, Approximate/Detailed Coefficient of DWT,* each recorded at different timestamp with interval of ***1 sec***
- Label *y*: Disagreement/Non-disagreement

Baseline Model

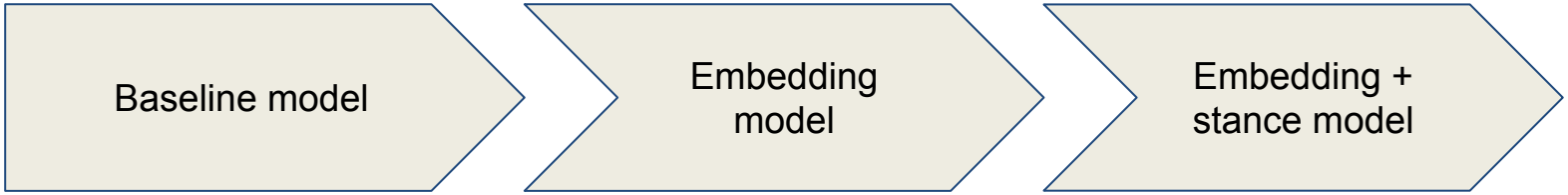


	precision	recall	f1-score	support
0	0.82	1.00	0.90	24679
1	0.00	0.00	0.00	5321
accuracy			0.82	30000
macro avg	0.41	0.50	0.45	30000
weighted avg	0.68	0.82	0.74	30000

Thoughts

- *Inverse DWT* will give back the original information, we should avoid using highly correlated features
- More data points and annotations are needed!
- Advanced models (i.e. LSTM, RNN)
- Consider using *time lag* in time series data as features

Text Transcript Modeling

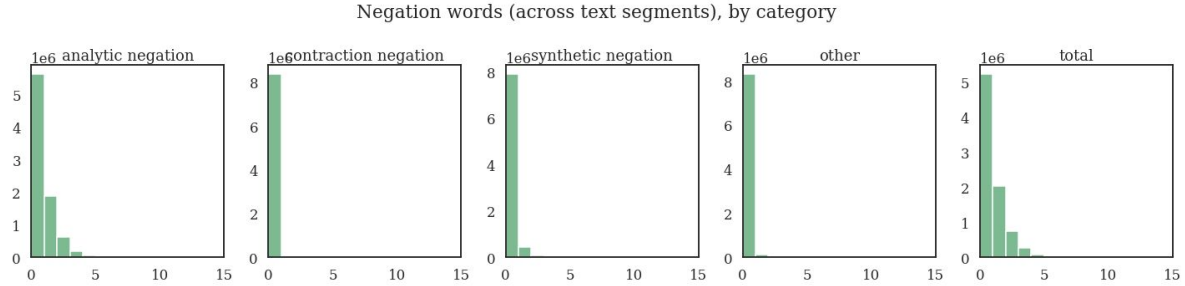


- 1-dimensional features: count of negative lexicon words
- n-dimensional features: map text segments to (SBERT) embeddings capturing semantics
- map text segments to embeddings capturing semantics: project to “negative” subspace

Text Transcript: Baseline Model

Hypothesis: more “negation words” in transcript segment → higher probability of disagreement

Category	Words
Analytic negation	no, not
Contraction negation	ain’t, aren’t, didn’t, shouldn’t, etc.
Synthetic negation	neither, never, nor, none, nobody, noone, no-one
Other	disagree, incorrect, wrong, ridiculous, absurd



Text Transcript: Baseline Model

Model: prediction based on count of “negation words” > threshold

Lexicon counts

transcript segments

analytic negation	contraction negation	synthetic negation	other
2	0	0	0
0	0	0	0
0	0	0	1
0	0	0	0
5	0	0	0
0	0	0	0
2	0	0	0
0	0	0	0
3	0	1	1
1	0	0	0

Prediction: *count > threshold*

Results

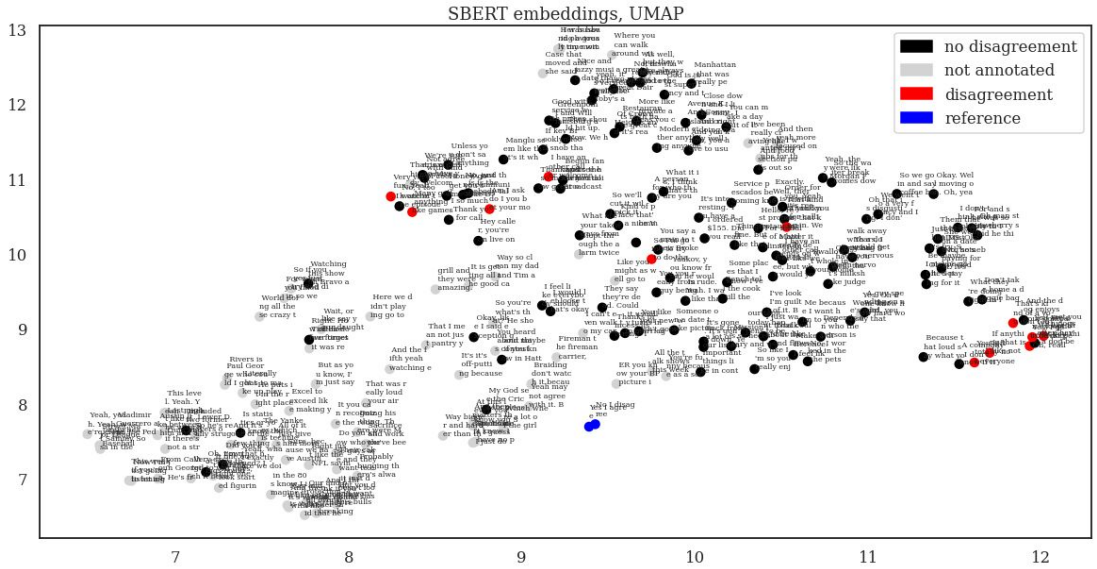
(Dog disagreement episode vs. test episode)

	Precision	Recall	F1	Predicted
“No/not”	**1.00**	0.50	0.67	3
Lexicon	**1.00**	0.33	0.50	2
“No/not”	**0.27**	0.29	0.28	15
Lexicon	**0.33**	0.14	0.20	6

Text Transcript: Embedding Model

Hypothesis: segments w/ embeddings near “No I disagree” → higher prob. of disagreement

Embeddings

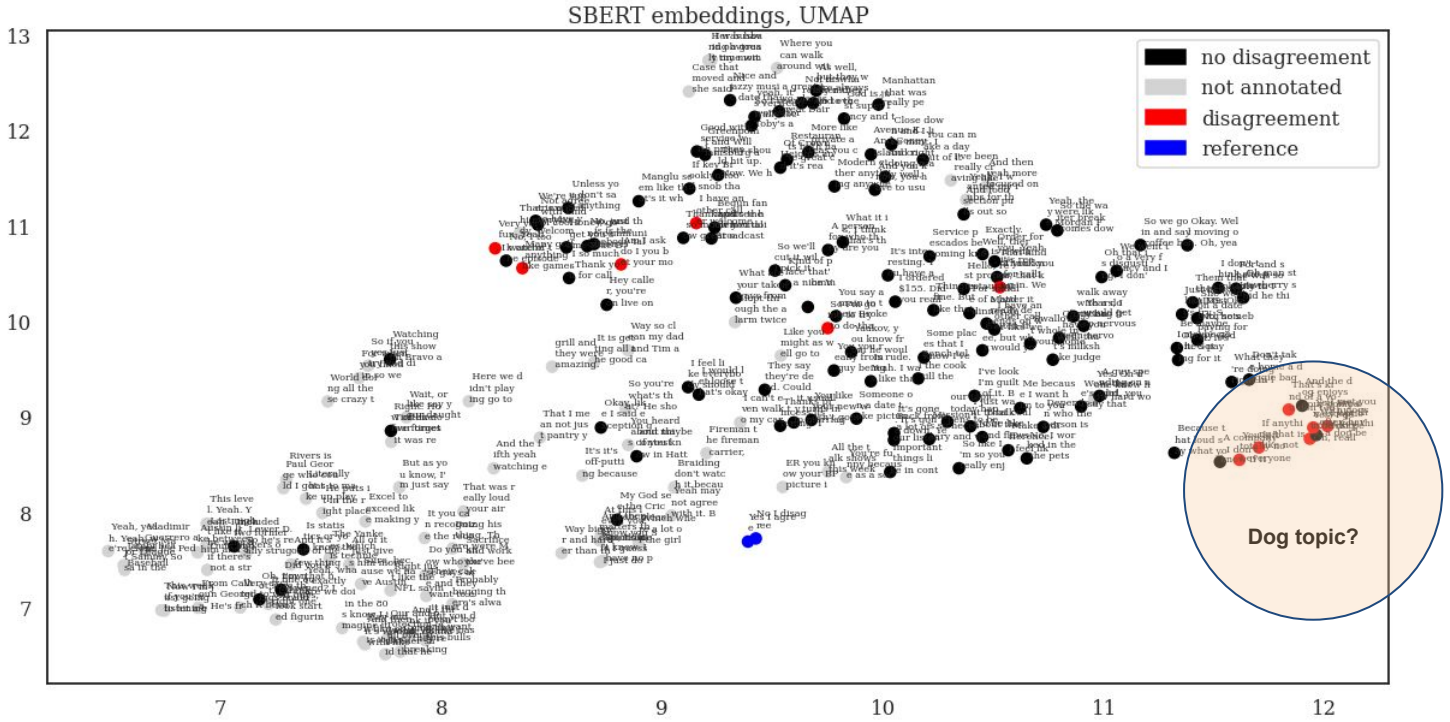


Results

(Dog disagreement episode vs. test episode)

Precision	Recall	F1	Predicted
0.67	0.67	0.67	6
0.24	0.36	0.29	21

Text Transcript: Embedding Model



Next Steps

- Audio + text specific modeling
- Scaling with AWS (set up permissions/users, S3, Sagemaker)
- Hypothesis: best results = audio + text

Thank You