Analysis of Stand-Up Comedians

The Science behind the Art

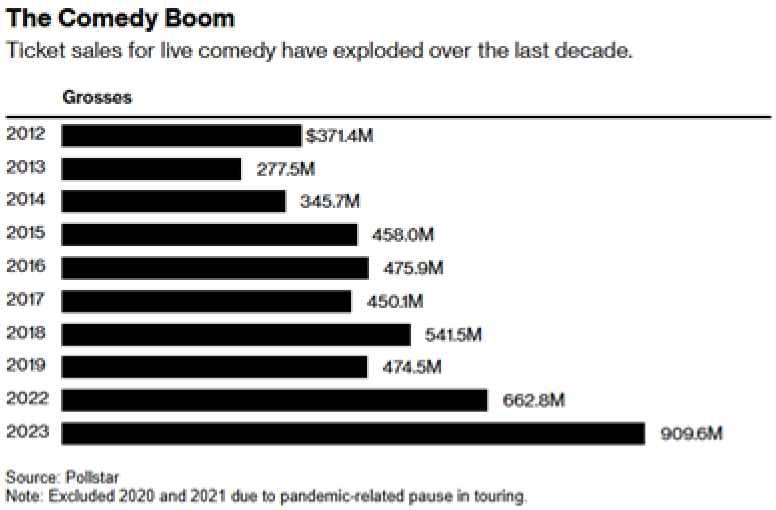
Kedar Godbole
Alex Imhoff
Tara Mary Joseph
Jyoti Kumari
Aman Sharma
MSBA Fall 2024

Agenda

1. Overview and Purpose
2. Data and the YouTube API
3. EDA - Text and Audio
4. Modeling - Text and Audio
5. Combining Text and Audio Models
6. Conclusion, Insights, and Future Scope
7. Q&A

Overview and Purpose

- Stand-Up Comedy industry tripled over the past 10 years
- More opportunities for comedians to get noticed and earn revenue through YouTube, Netflix specials, podcasts, etc.
- However, talented comics with good content continue to struggle with delivery



Using the YouTube-DL API to Gather Data

- YouTube-DL API allows for caption and audio downloading
- Clean-up “extra” in the captions
- Identify instances of laughter/cheering
- APIs used:
  - Youtube-transcript-api → Captions & Audio
  - Pydub → convert raw data to mp3 audio

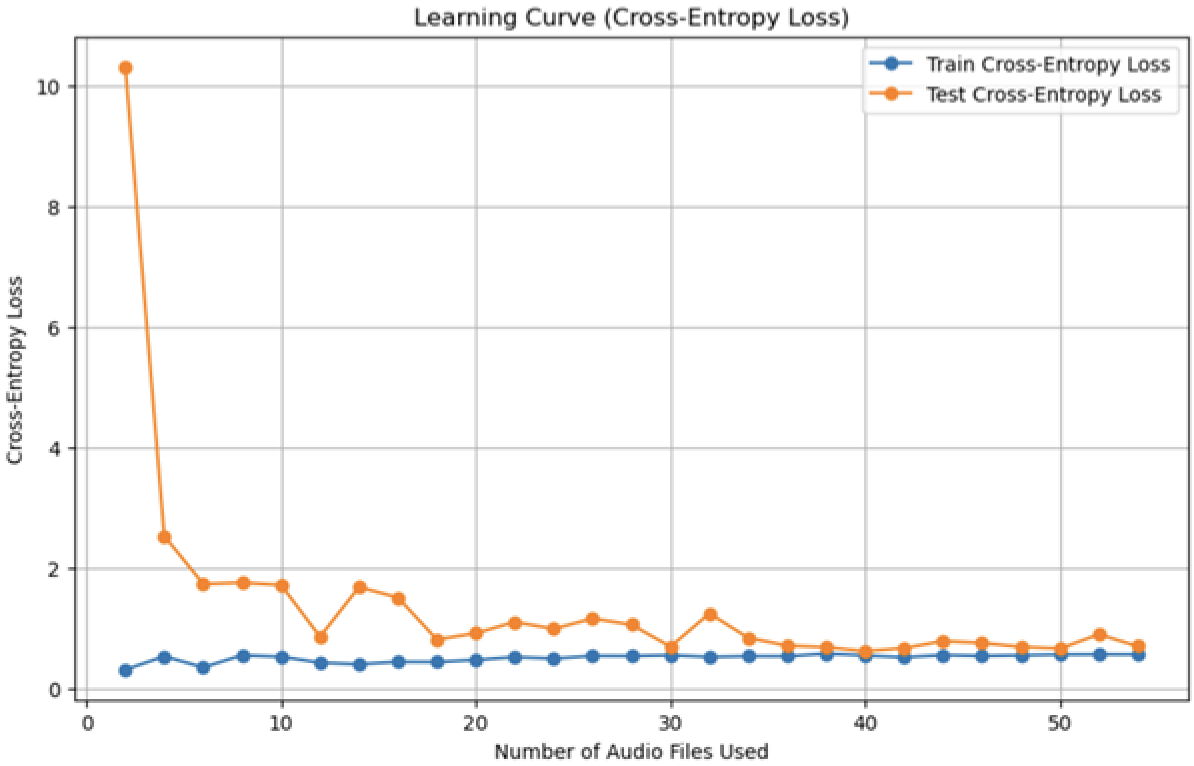
```
00:03:27.774 --> 00:03:30.710 align:start size:91% position:9%
"HEY, WHY DID WE HAVE
MY SECOND BIRTHDAY PARTY
AT THE PYRAMIDS OF GEZA?"

00:03:30.710 --> 00:03:33.546
[LAUGHTER]
```

YouTube Data - Challenges

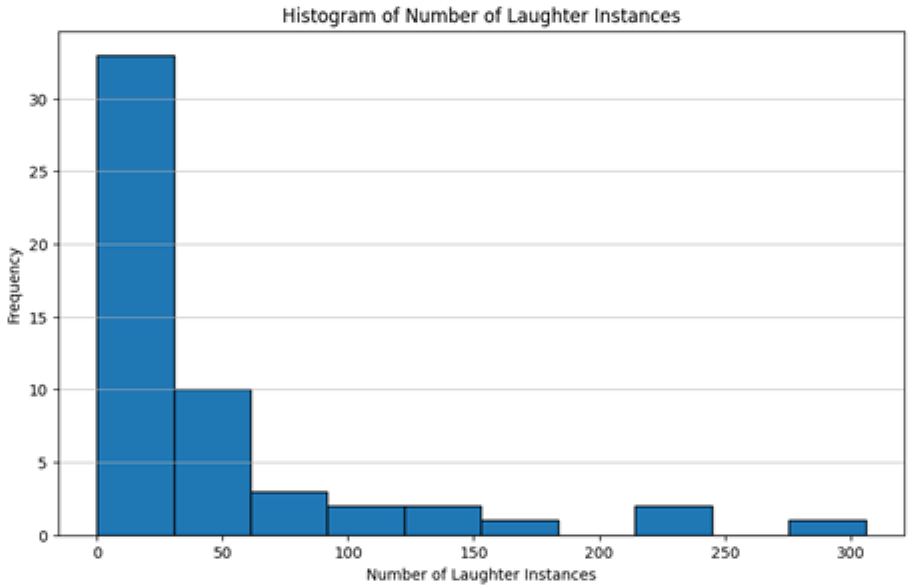
- Manual vs Automatic Captions
- Swear Words
- Converting to Usable Format (Audio)
- Timestamps
- Language (American vs British vs Australian)
- 3rd Party API - No Association with YouTube

```
No subtitles found for https://www.youtube.com/watch?v=oLhZVRphhew . Skipping.
No subtitles found for https://www.youtube.com/watch?v=8qfndbEYroE . Skipping.
No subtitles found for https://www.youtube.com/watch?v=kZZez42HWvU . Skipping.
Download and cleaning completed!
Skipped videos:
https://www.youtube.com/watch?v=oLhZVRphhew
https://www.youtube.com/watch?v=8qfndbEYroE
https://www.youtube.com/watch?v=kZZez42HWvU
```

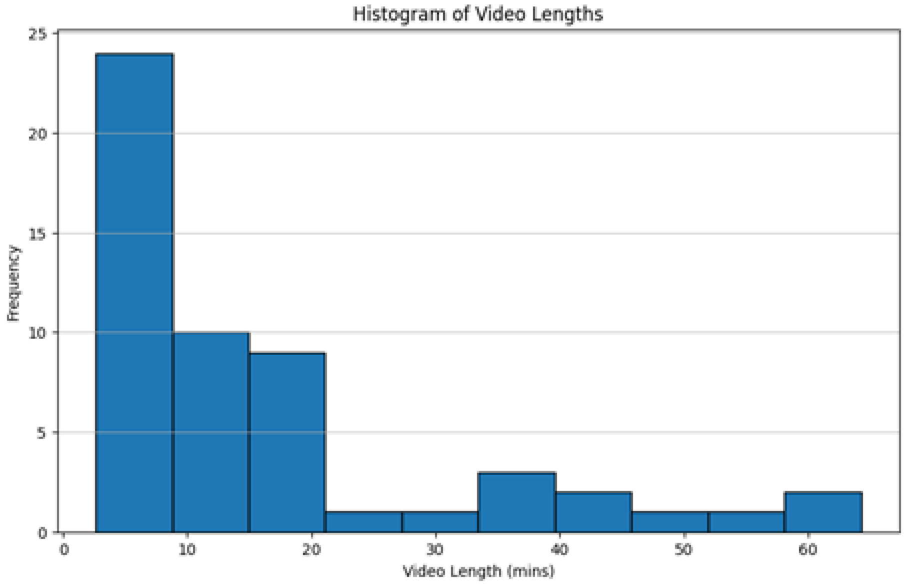
Learning Curves



EDA - Laughs and Video Lengths



~46-47 laughs per Video



Avg Vid Length is ~16-17 mins

Text Analytics

Our target variable - ‘laughter’ is the row that contains the laughter annotation and the previous three lines which act as context

subtitle	laughter
i've just been standing up straight.	1
and you know what i learned?	1
standing up straight hurts my back.	1
(audience laughing)	1
this is very uncomfortable.	0
where all the bones meet, pain.	0
you know who i think had it right?	1
that dude in the middle	1
of the evolutionary chart.	1
(audience laughing)	1

Context

Context

An example from video ‘A Man Of Average Intelligence.’ - Zoltan Kaszas

“the sentence immediately preceding the marker was considered the “punchline”, while the three sentences preceding the punchline were considered the “setup”. ”

Methodology inspired from *: Analyzing Humor by Turano et. al. (2022)*

Text Analytics

A simple approach with feature engineering based on the text of the subtitles

Phonetic - *alliteration , assonance , consonance*

Ambiguity - *sense combination , sense farthest , sense closest, average senses*

Humor - *entity types, antonyms, entity count , profanity*

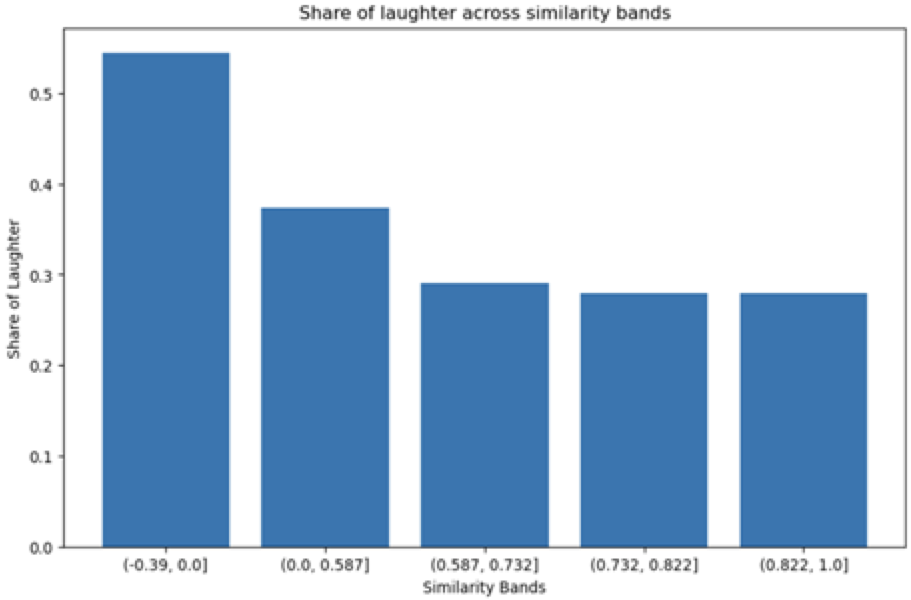
Similarity - *cosine similarity between pairwise vectors (based on lookback)*

Sentiment - *sentiment compound score*

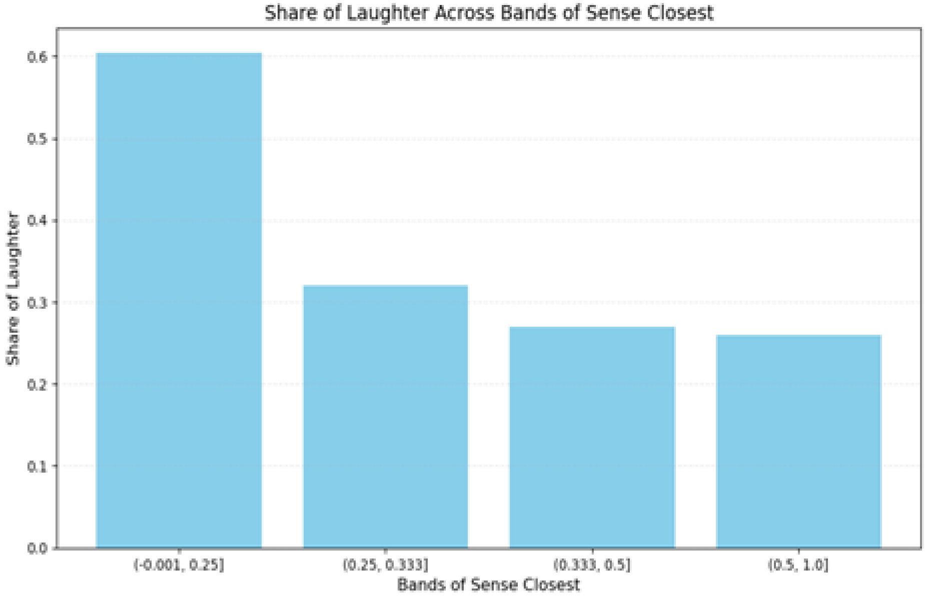
Punctuations - *exclamations , question marks , pauses*

Text Analytics - EDA

Low Similarity, Big Laughs

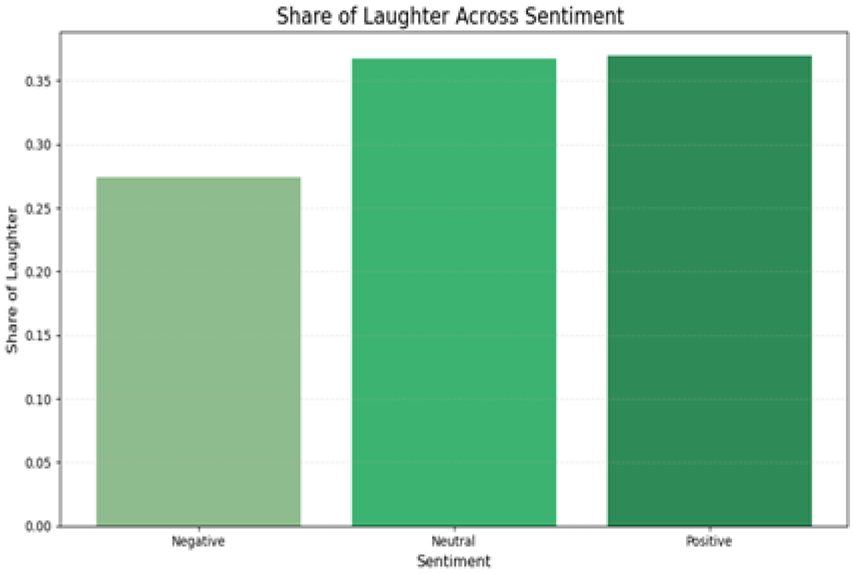


The role of "unexpectedness" or "contrast" in driving laughter

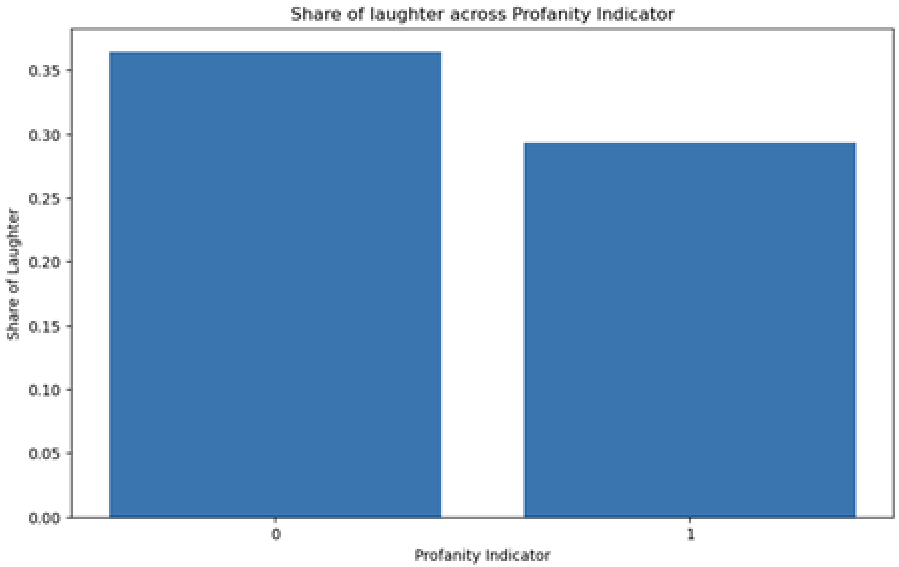


Text Analytics - EDA

Laughter Across Sentiment Types

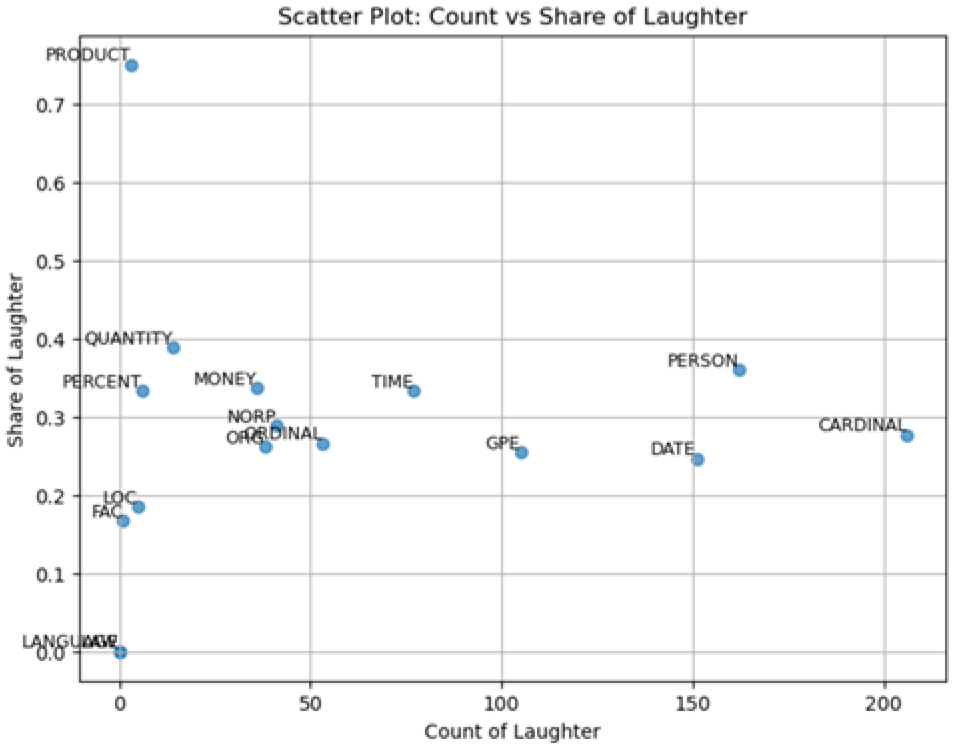


Profanity and Punchlines: Does it enhance humor?



Text Analytics - EDA

- Does the presence of certain entity types correlate with laughter?"

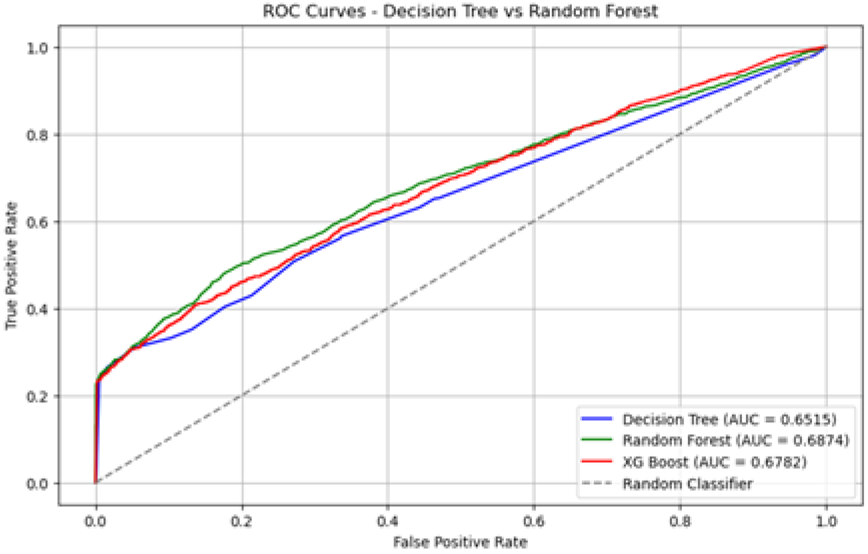


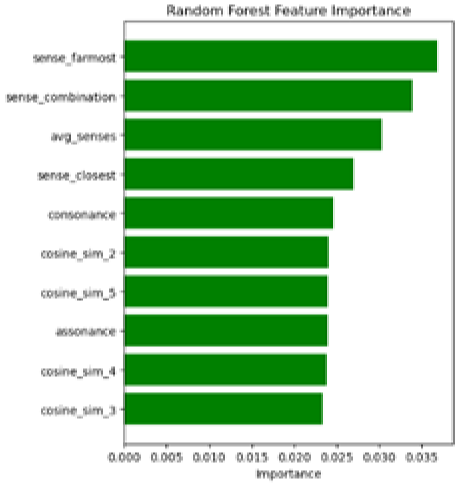
An example of label = PERSON

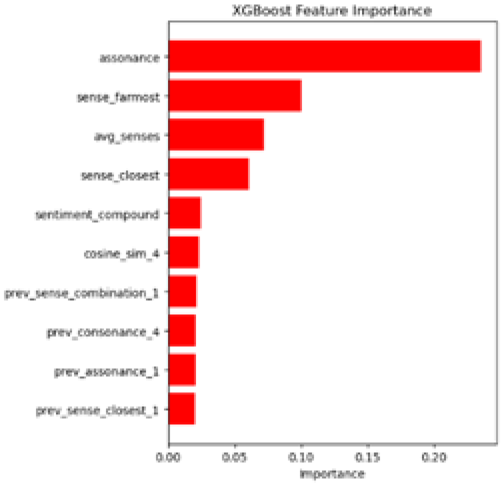
"I think Lil Wayne will die before me
Because he drinks a lot of cough
syrup with no symptoms
[laughter] "

From video 'Live from Chicago' - Hannibal Buressï

Text Analytics : Feature based tree models







Next Step: Leverage transformers to better capture contextual and nuanced complexities in humor.

Text Analytics : Deep Networks

Target variable definition

Old Approach:

- Mark 3 lines preceding the marker as context [1]

subtitle	laughter
i've just been standing up straight.	1
and you know what i learned?	1
standing up straight hurts my back.	1
(audience laughing)	1
this is very uncomfortable.	0
where all the bones meet, pain.	0
you know who i think had it right?	1
that dude in the middle	1
of the evolutionary chart.	1
(audience laughing)	1

Context

Context

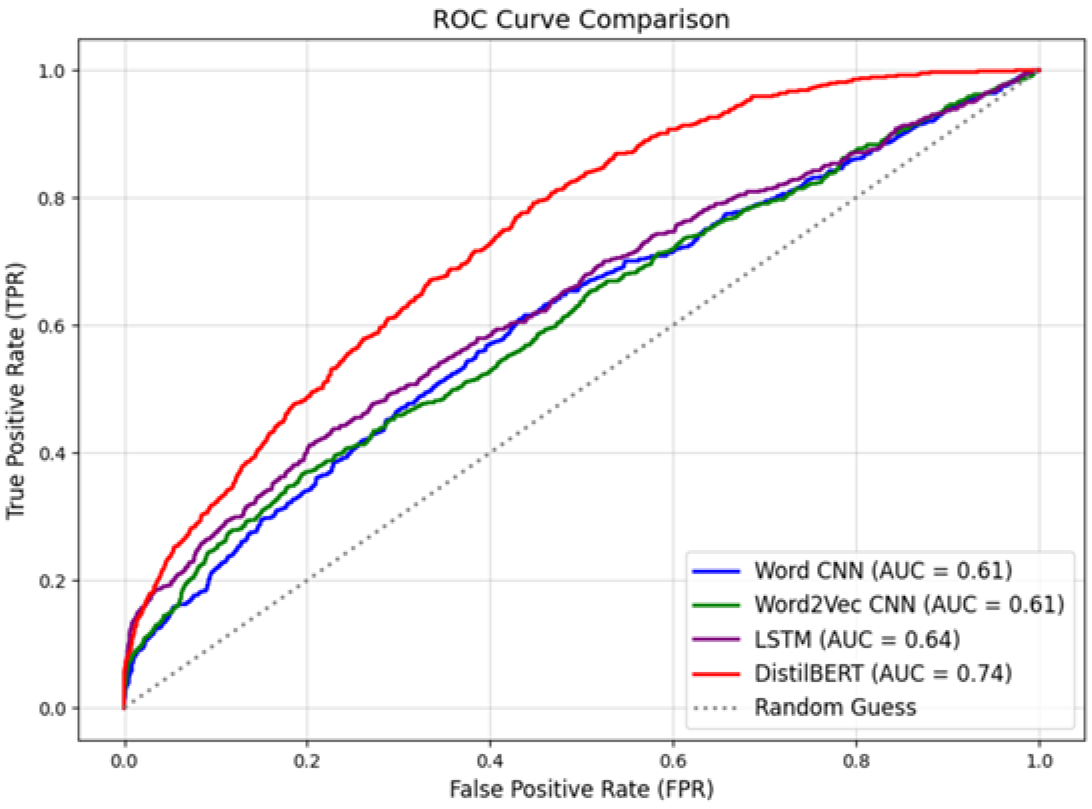
New Approach:

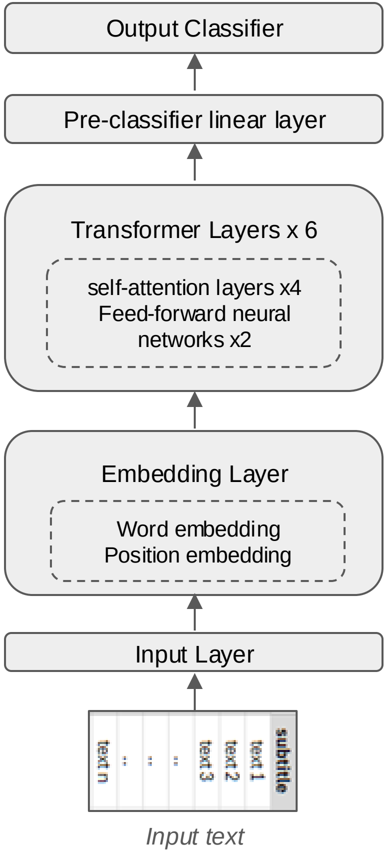
- Mark line only with the marker as context [1]
- If word count in a sentence < 5, then combine it with previous chunk.

subtitle	laughter
Do you remember the first time	0
you put on a mask?	0
I went, is this what my breath smells like? (audience laughing)	1
I owe a lot of people an apology. (audience laughing)	1
You ever burp wearing a mask? (audience laughing)	1

An example from video 'I'm Nervous, Insecure and Squishy' - Mark Normand

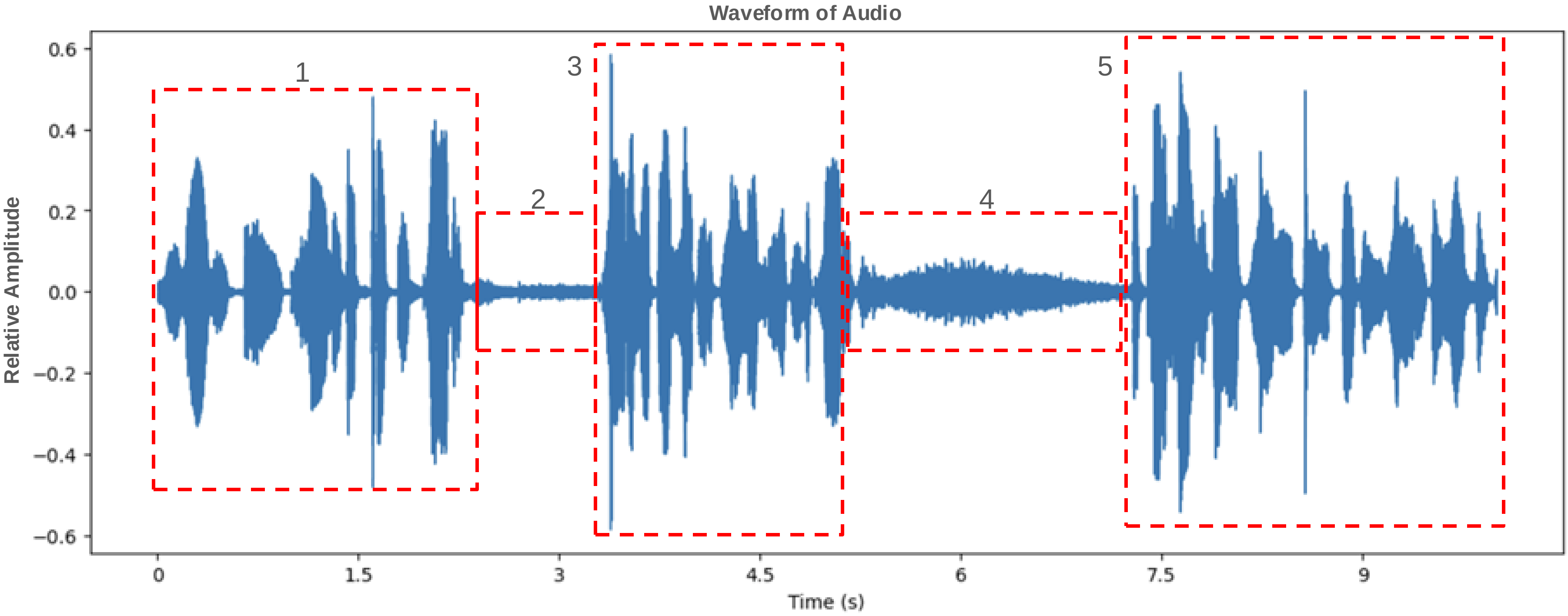
Text Analytics : Performance



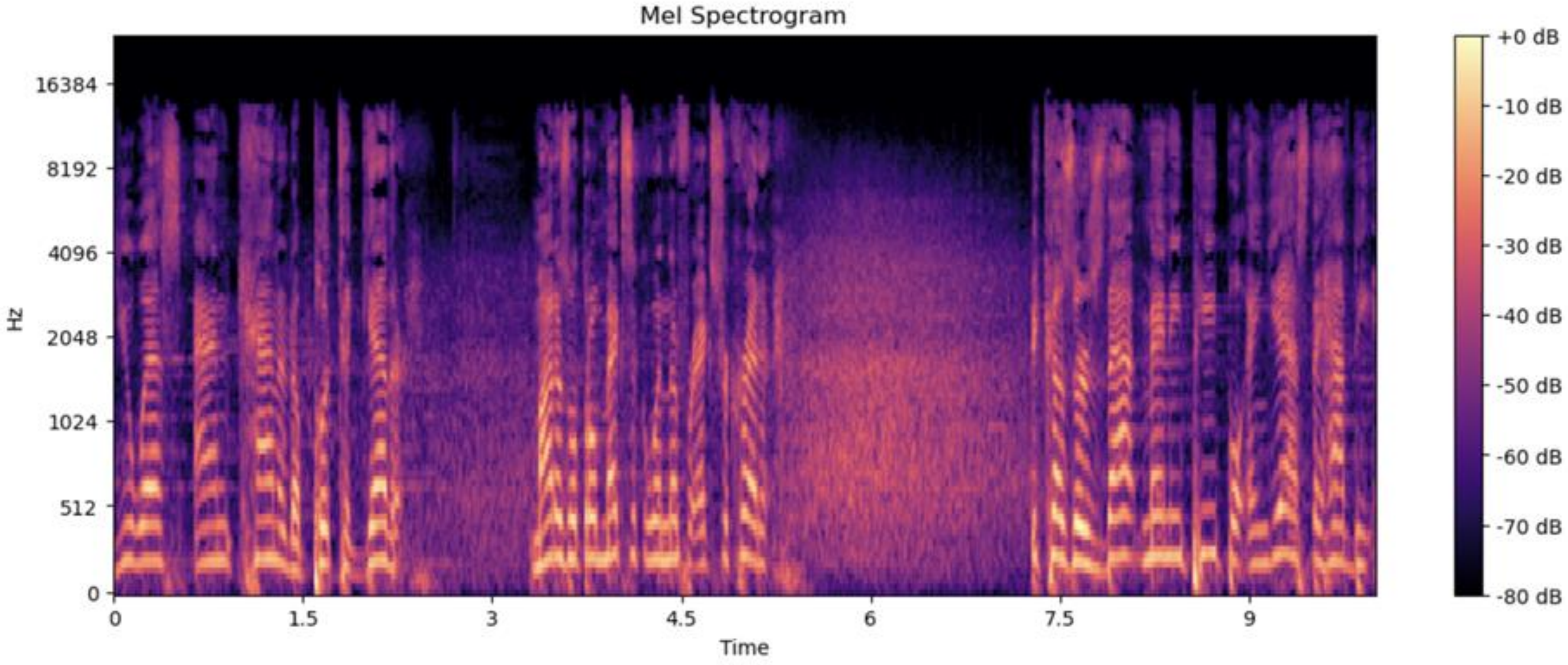


Audio Analytics

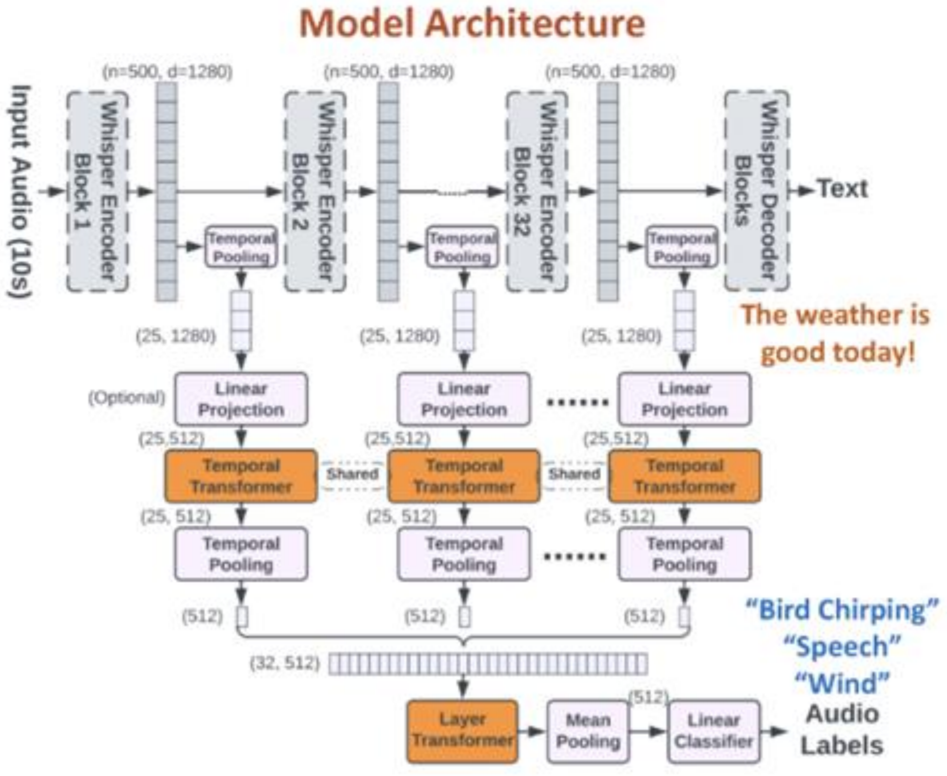
Which of the five sections below represent audio laughter?



Using higher dimensional representations doesn't make the task easier.
Turns out, it is challenging to come up with a deterministic rule to annotate laughter in audio



Annotating audio with Whisper AT *



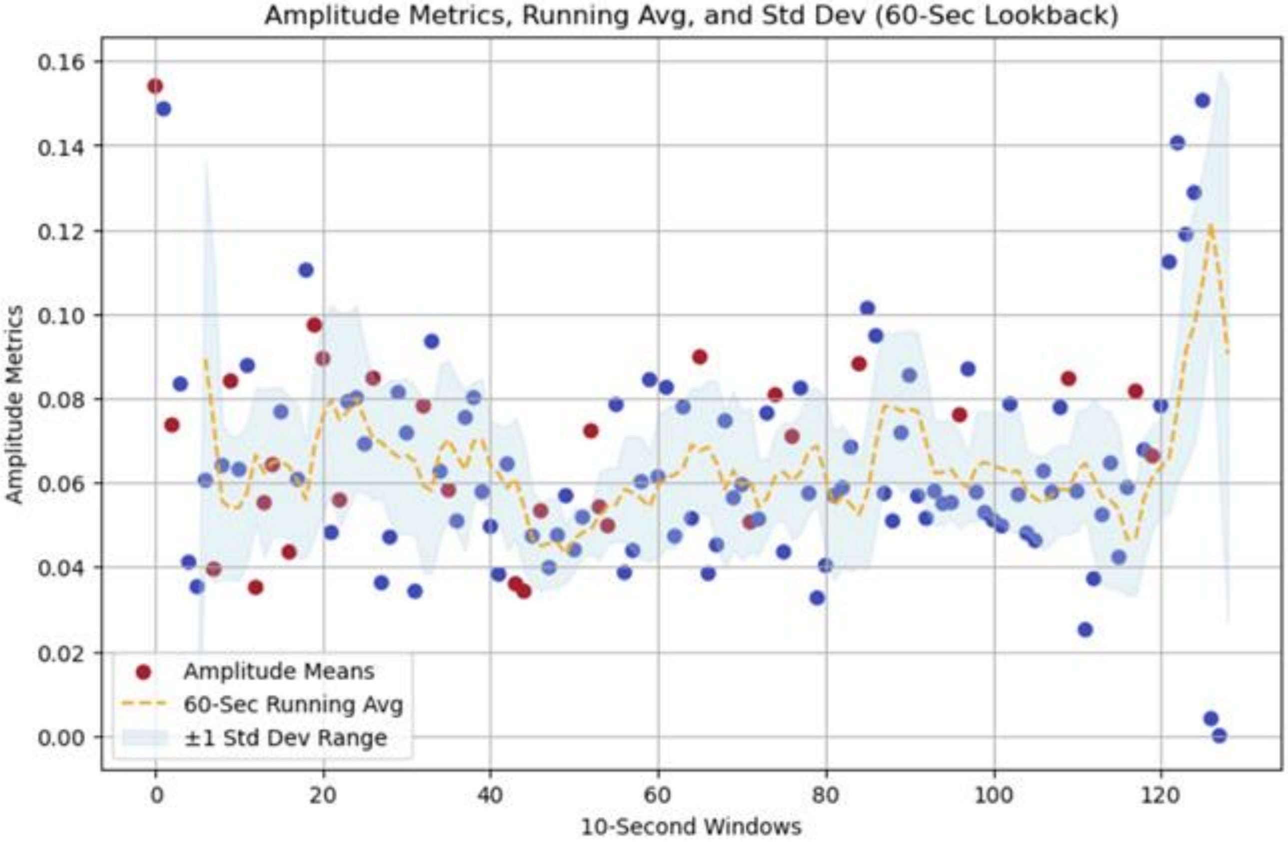
- Whisper - AT is built by fine tuning Open AI's ASR model Whisper

- The model, reliably, assigns audio tags {clapping, chirping etc} from a 527 classes for every disjoint **10 second interval**

- We used this, state of the art, model for annotating Audio with **laughter tags**

** Reference : Whisper - AT: A Unified Audio Tagger and ASR Model*
Y.Gong et. al. MIT- IBM Watson Labs

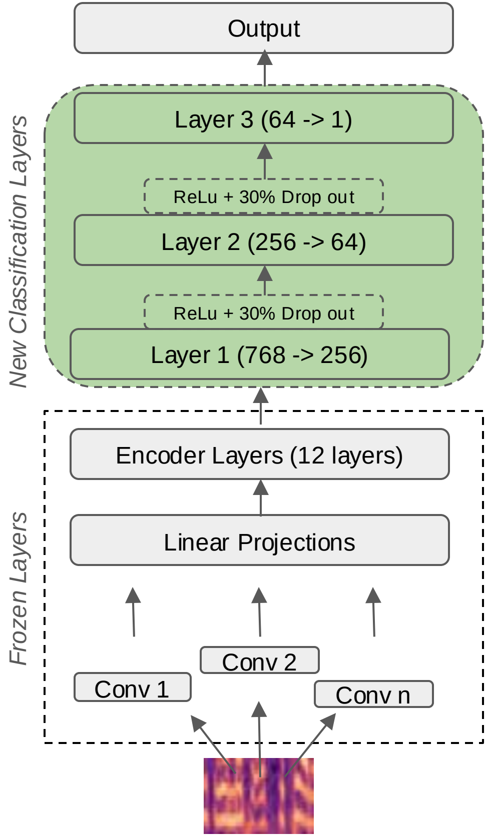
Audio - EDA



Laughter in next 10 seconds interval

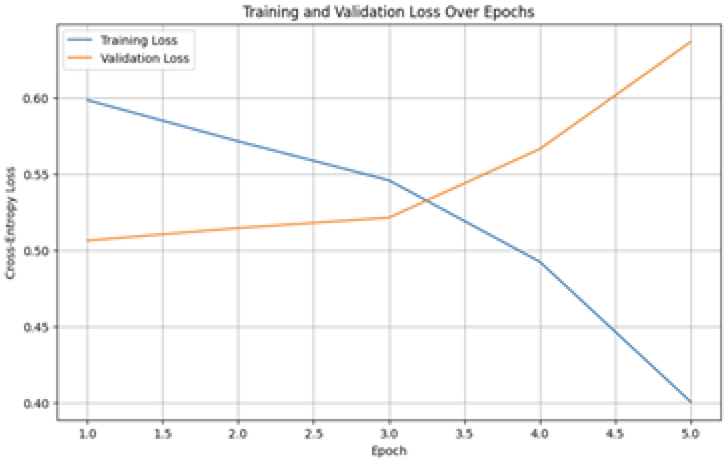
No Laughter in next 10 seconds interval

Audio - AST : Audio Spectrogram Transformer



Input Spectrogram

Model	Performance
CNN 3 convolution layers	AUC 50% *Pred = Prior of majority class*
AST Only Classifier Layer Trained	AUC 57%



AST overfits quickly

*AST - Audio Spectrogram Transformer , Yuan Gong et. al.

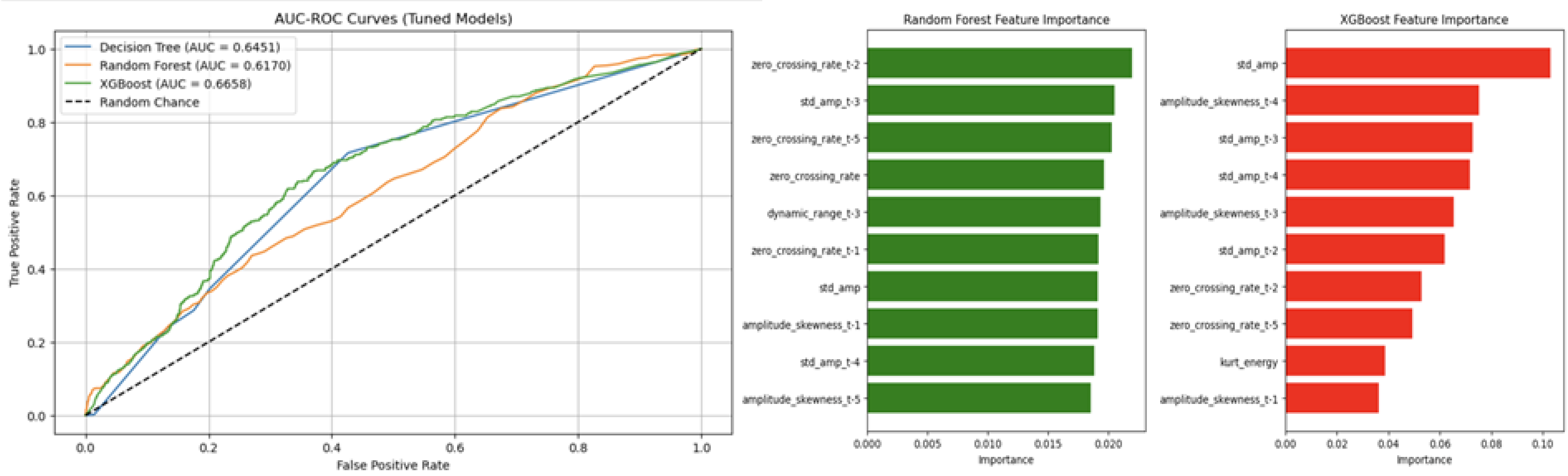
Audio - Feature Engineering

A simpler modeling approach with engineered features, designed to capture speech modulations:

1. Dynamic amplitude range (measure of loudness)
2. Zero crossing rate (key feature in percussion sounds)
3. Skewness & Kurtosis of Amplitude
4. Energy Decay
5. Total pause duration etc, AND

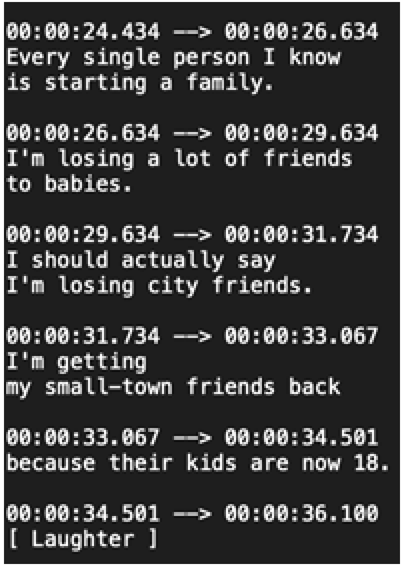
   Lookback features (last 50 seconds / 5 windows)

Audio - Performance Comparison



For audio, simpler models are doing better than complex transformer architectures.

Combining Predictions (Methodology)


00:00:24.434 --> 00:00:26.634
Every single person I know
is starting a family.

00:00:26.634 --> 00:00:29.634
I'm losing a lot of friends
to babies.

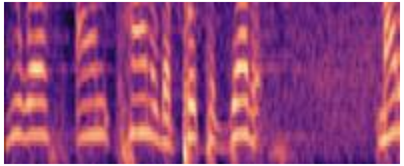
00:00:29.634 --> 00:00:31.734
I should actually say
I'm losing city friends.

00:00:31.734 --> 00:00:33.067
I'm getting
my small-town friends back

00:00:33.067 --> 00:00:34.501
because their kids are now 18.

00:00:34.501 --> 00:00:36.100
[Laughter]

Subtitles



Spectrogram

Start	End	Pred_1	target
0:00:24.434	00:00:26.634	0.20	0
...	...	...	...
00:00:33.067	00:00:34.501	0.80	1

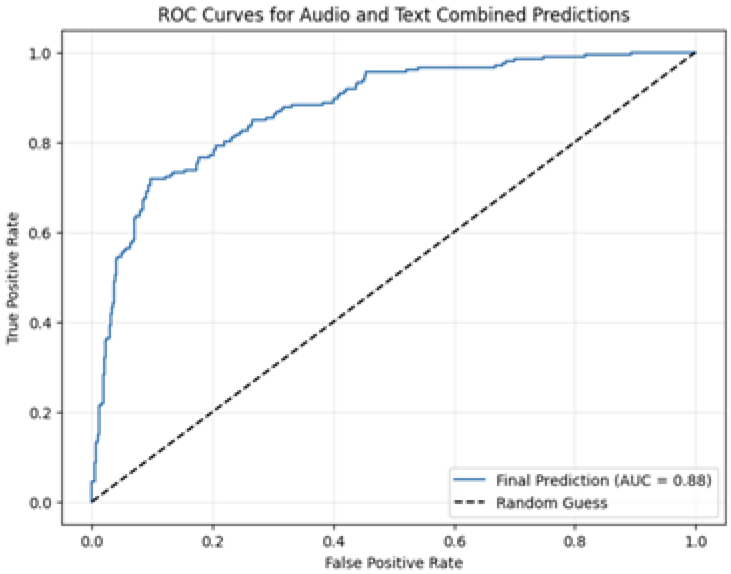
Start (s)	End (s)	Pred_2*
20	30	0.40
30	40	0.60

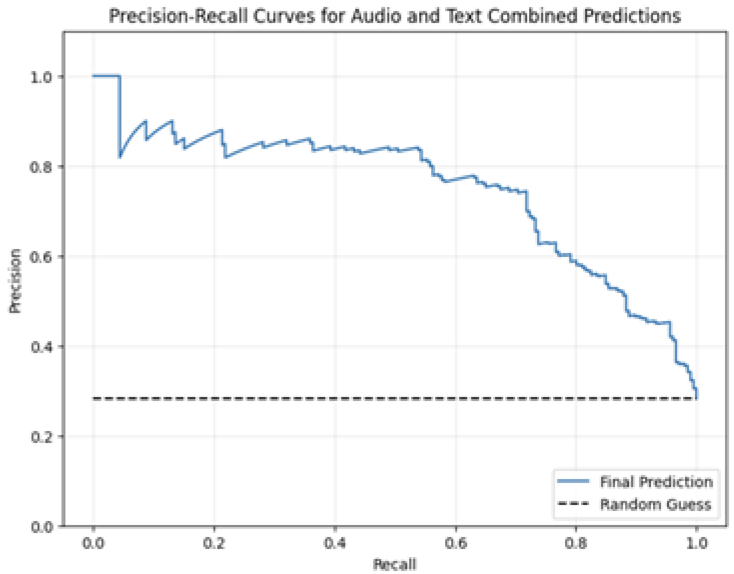
Combined Predictions

Start	End	Pred1	Pred2	Pred_f	target
0:00:24.434	00:00:26.634	0.20	0.40	0.30	0
...	...	...	...	...	...
00:00:33.067	00:00:34.501	0.80	0.60	0.70	1

**Pred_2 in audio data capture whether people laughed in the next 10 seconds*

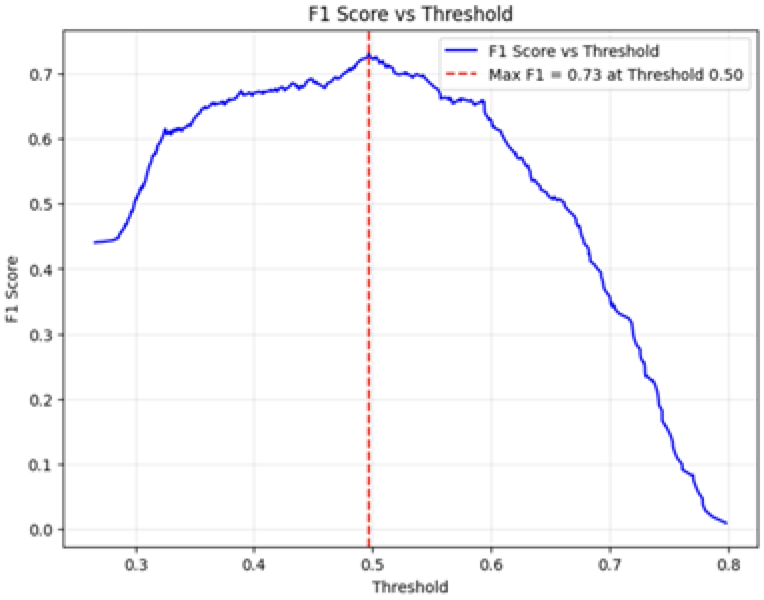
Combining Predictions (Performance)





Data & Model	AUC
Audio (RF)	66%
Text (Transformer - Distillbert)	74%
Combined (Separate reserved test data of 13 videos ; Simple averaging of models)	88%

Benchmarking against research



F1-score of our combined model at optimal threshold is comparable to leading research

	Logistic Regression				Naive Bayes				Random Forest			
	acc.	prec.	recall	F1	acc.	prec.	recall	F1	acc.	prec.	recall	F1
BoW	0.58	0.58	0.58	0.58	0.59	0.60	0.59	0.58	0.71	0.72	0.71	0.71
BoW+SocialFeats	0.61	0.61	0.61	0.61	0.57	0.57	0.57	0.56	0.72	0.73	0.72	0.71
BoW+LingFeats	0.69	0.69	0.69	0.69	0.61	0.61	0.61	0.61	0.71	0.73	0.71	0.70
BoW+Social+LingFeats	0.70	0.70	0.70	0.70	0.58	0.59	0.58	0.57	0.72	0.73	0.72	0.71
TFIDF	0.63	0.63	0.63	0.63	0.61	0.61	0.61	0.61	0.71	0.73	0.71	0.71
TFIDF+SocialFeats	0.65	0.65	0.65	0.65	0.64	0.64	0.64	0.64	0.71	0.73	0.71	0.71
TDIDF+LingFeats	0.70	0.71	0.70	0.70	0.66	0.66	0.66	0.66	0.71	0.73	0.71	0.71
TFIDF+Social+LingFeats	0.71	0.72	0.71	0.71	0.67	0.67	0.67	0.67	0.72	0.73	0.72	0.71

Table 1: Performance of the classifiers on the different models.

Reference: Analyzing Humor by Turano et. al. (2022)

Classifier and features	Accuracy	Precision	Recall	F-score
CRF n-grams	61.8	56.8	45.1	50.2
CRF language features	67.8	**67.5**	47.8	56.0
CRF n-grams + language features	65.9	61.2	55.3	58.1
LSTM	63.1	56.7	58.7	57.6
LSTM + high level features	**70.0**	66.7	**59.4**	**62.9**

Table 1: Results, percentage.

Reference: LSTM Framework for Predicting Humor by Bertero et. al. (2016)

Future Scope

- Modeling Scope
  - Fine tune frozen transformer on audio data, extract audio vectors, use for classification
  - Whisper-Small with attention layer to combine spectrograms

- Production Scope
  - **End Goal** - Come up with a tool for struggling comedians to upload their mock scripts or practice audio set, and the tool is able to detect which of the jokes would land (make audience laugh)

Takeaways

- Take your chances with all models!
  - Audio → Simpler model was better
  - Text → Complicated model was better
- Mind Your Language!
  - Profanity does NOT necessarily lead to more laughs
  - Profanity can also limit your audience
- Computation is Expensive and Time-Consuming!
  - Transformers
  - Converting audio data to mp3 format
  - Neural network

Q&A

Thank-you for listening!

Appendix