# "So You Think You're Funny?": Rating the Humour Quotient in Standup Comedy

**Anirudh Mittal, Pranav Jeevan P, Prerak Gandhi, Diptesh Kanojia and Pushpak Bhattacharyya**

**The 2021 Conference on Empirical Methods in Natural Language Processing**




Presentation for EMNLP 2021 | 7th-11th Nov Punta Cana, Dominican Republic

# Motivation

## Can a computer understand humor?

**Standup comedy is a permutation and combination of words that elicit humor**
**NLP is the science that can measure that**

**TARS in Interstellar- manually change degree of humour in robot**

# What are we doing differently?

Automatic humour rating **(non-binary)**

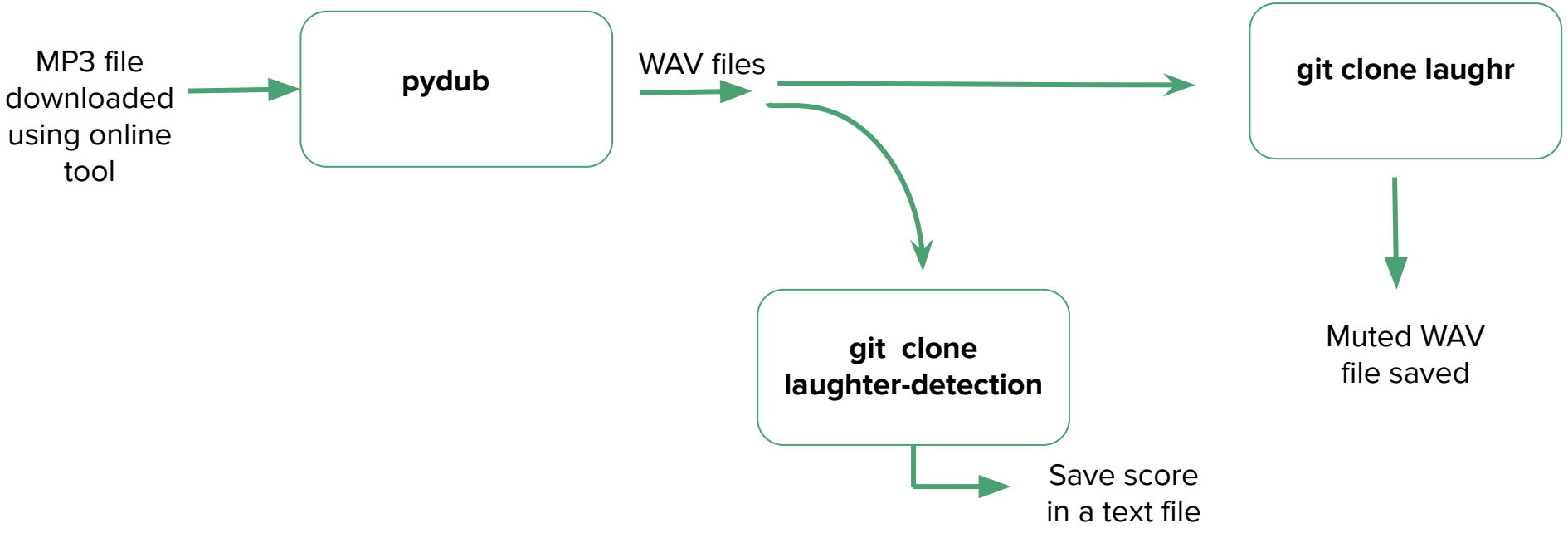
First **multi-modal** dataset for **standup comedy (using audio and text)**

**Scoring mechanism** for funniness from audience feedback

# Dataset - Open Mic

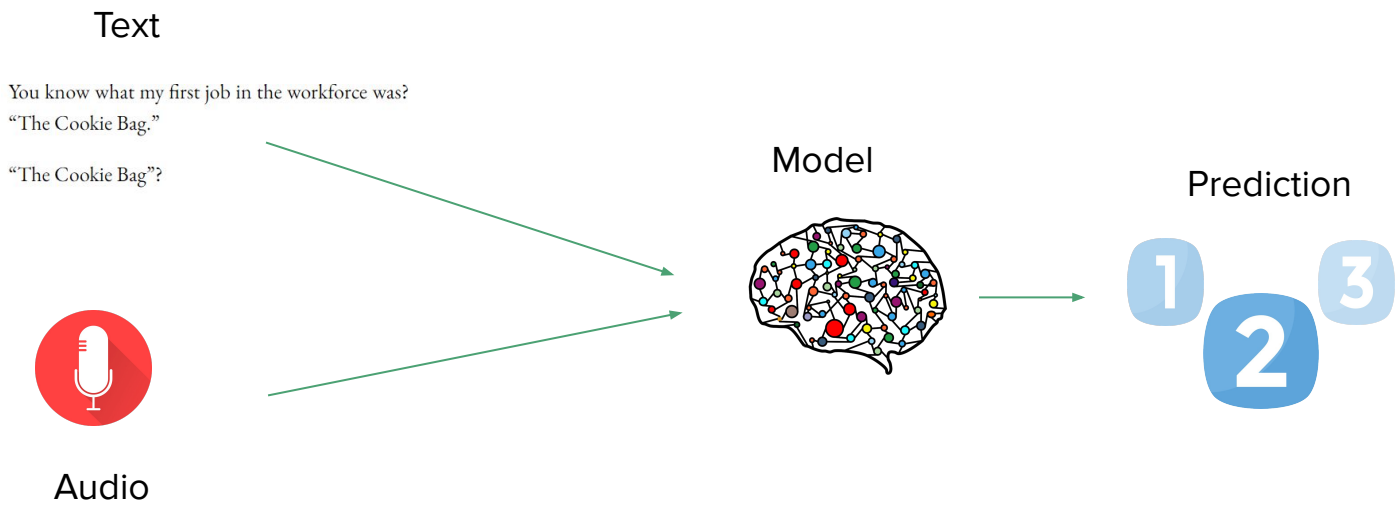
- Over **40 hours** of English standup comedy audio from 36 shows of 32 comedians
  - **927** two minute stand up audio clips
- For unfunny data, we used TED talks
  - **128** two minute TED talk audio clips
- Diverse genders, nationalities, cultures
- Audio divided into smaller **independent clips** of about 2 mins.
- Corresponding transcript stored
- **1055** data points in Dataset

# Data preparation



# Overview

Input and Output



# Funniness score

- Extract the audience laugh and find its length
- For any clip,

$$\text{Normalised score} = \frac{\sum \text{length of all audience laughter in the clip}}{\text{Length of the clip}}$$

- Norm scores are then compiled and the classes are assigned as shown in table

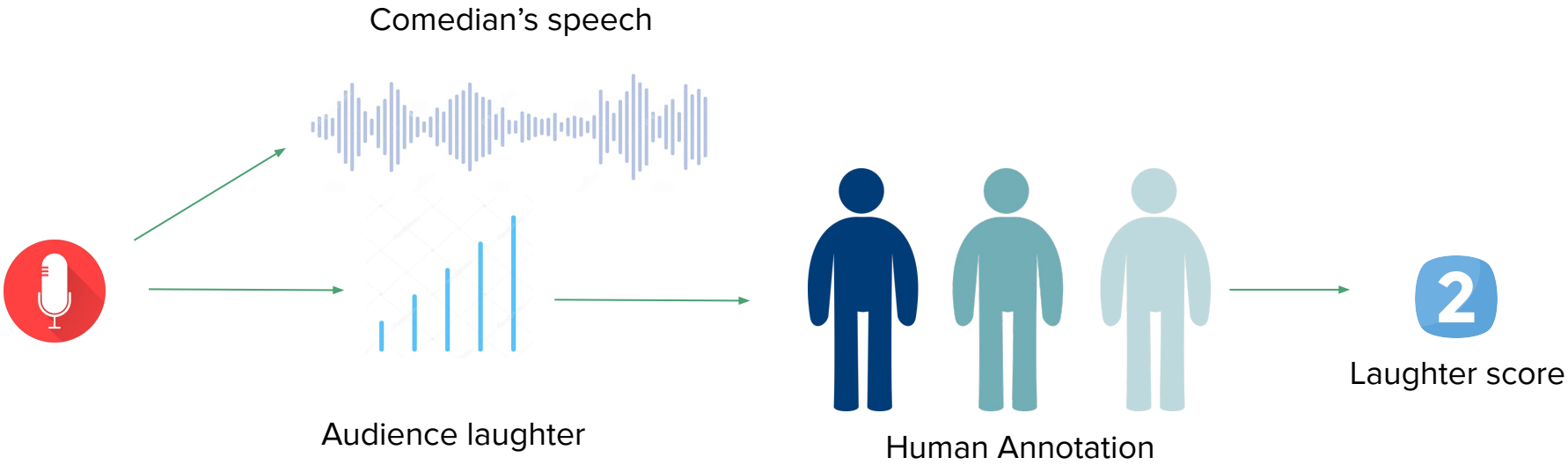
| Rating | # Clips | Scoring Criteria |
|---|---|---|
| 4 | 233 | $\text{score} > \mu + 0.75\sigma$ |
| 3 | 185 | $\mu + 0.75\sigma \geq \text{score} > \mu$ |
| 2 | 256 | $\mu \geq \text{score} > \mu - 0.75\sigma$ |
| 1 | 253 | $\mu - 0.75\sigma \geq \text{score} > 0$ |
| 0 | 128 | $\text{score} = 0$ |

Table 1: Number of clips and the scoring criteria for assigning humour rating to each clip based on the mean ($\mu$) and standard deviation ($\sigma$) of the scores

Note: The TED data was directly assigned to class 0 as it doesn't contain any laughs.

Source: Jon Gillick and Marcin Wlodarczak. 2019. laughter- detection.

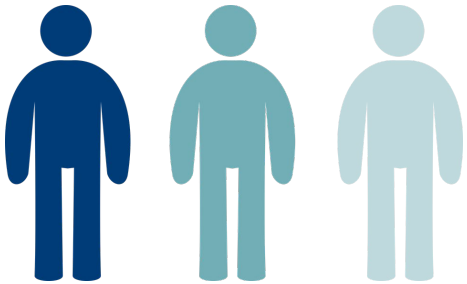
# Humor annotation



# Validation of scoring mechanism

| Pairwise Agreement | |
|---|---|
| Annotators A and B | 0.643 |
| Annotators B and C | 0.926 |
| Annotators C and A | 0.611 |
| Average pairwise Cohen's Kappa | 0.634 |
| Fleiss' Kappa | 0.632 |
| **Krippendorff's alpha** | 0.632 |

Table 2: Inter-Annotator Agreement (Fleiss' Kappa and Krippendorff's alpha) values along with pairwise agreement among the annotators



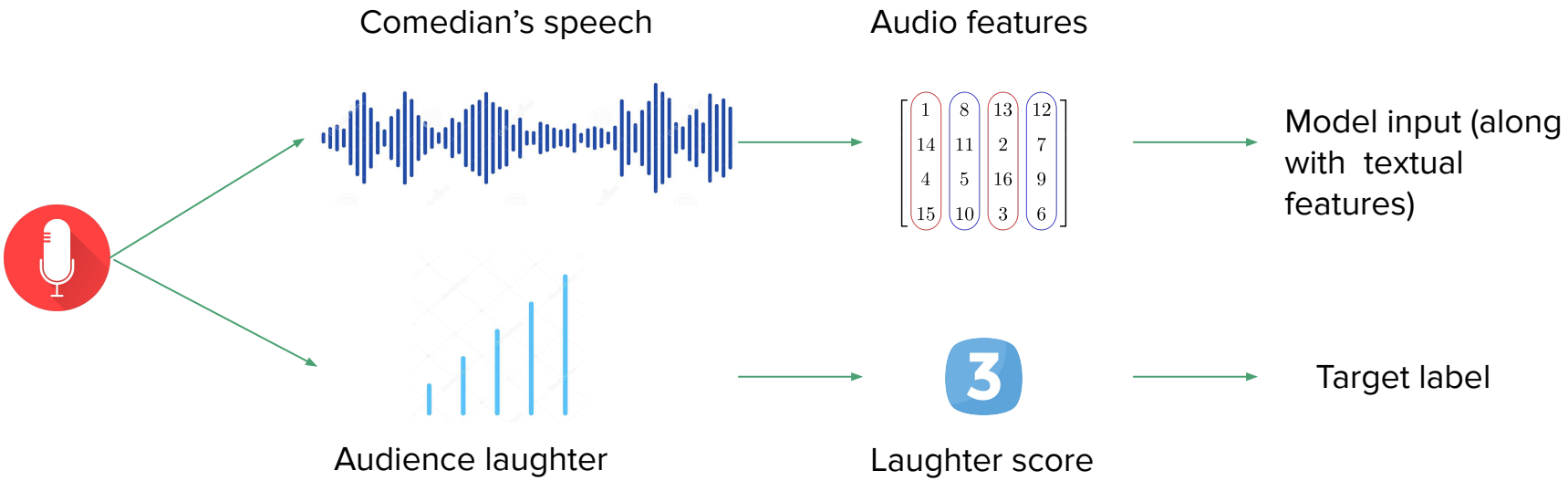
| Annotaters | QWK |
|---|---|
| Human A | 0.659 |
| Human B | 0.562 |
| Human C | 0.563 |
| **Average** | 0.595 |

Agreement of human annotation scores with the automatic scoring mechanism

Source: Jacob Cohen. 1968. Weighted kappa: Psychological bulletin, 70(4):213

# Experiment setup and methodology

Muting laughter




Source: Jeff Green. 2018. Sitcom laughtrack mute tool

# Experiment setup and methodology

Network Architecture

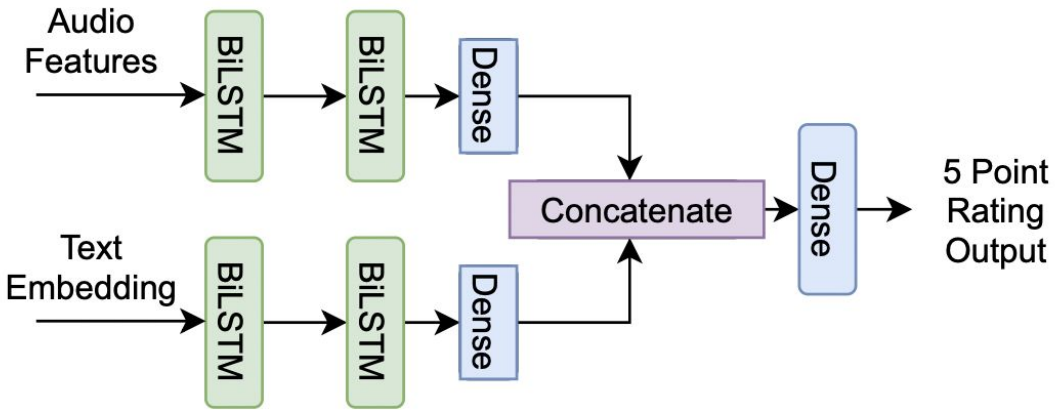



Figure 1: Neural Network Architecture

# Experiment setup and methodology

Audio features

- Features:
  1. MFCCs
  2. RMS energy
  3. Spectogram
- Information about the speech:
  1. Volume
  2. Intonation
  3. Emotion of the speaker
- 33 dimension vector for each time sample
- Maximum sequence length for audio embeddings = 8000

Source: Brian McFee et al. 2020. librosa/librosa: 0.8.0.

# Experiment setup and methodology

Textual features

Convert text i.e. Content to vectors for the model

$BERT_{base}$

DistilBERT

$RoBERT_{large}$

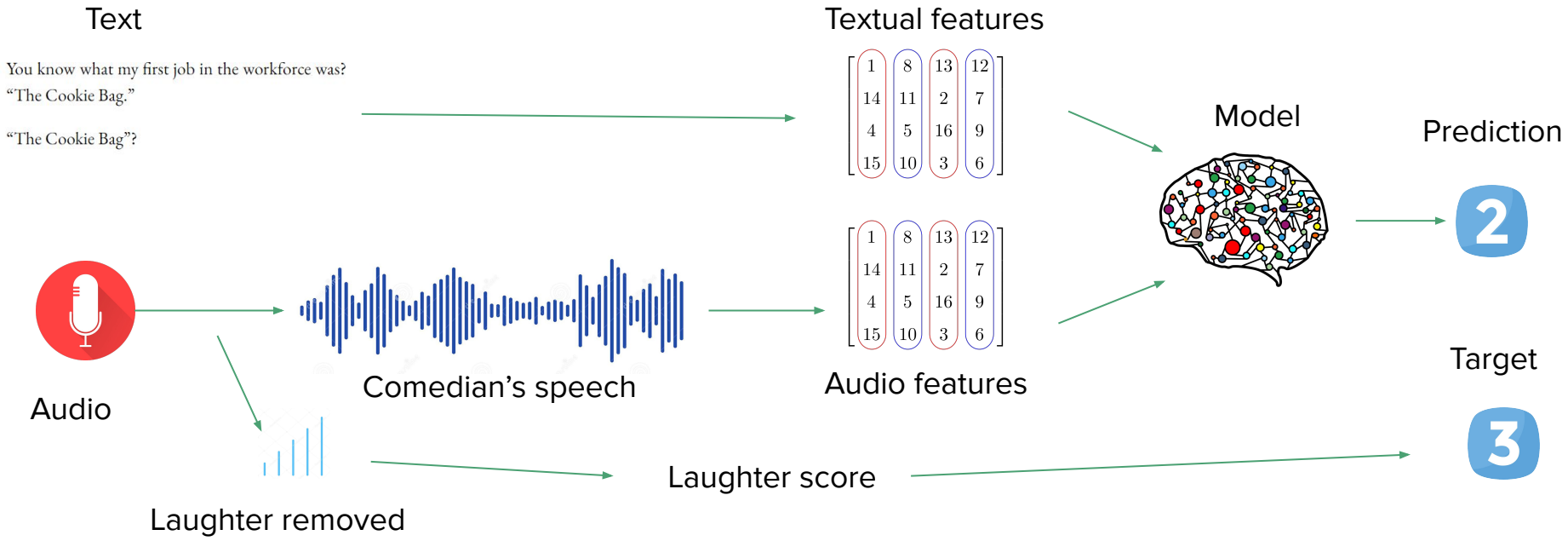
$BERT_{large}$

$RoBERT_{base}$

XLM

# Experiment setup and methodology

Methodology



# Results

| Textual Features | QWK |
|---|---|
| GloVe | 0.691 |
| $\text{BERT}_{base}$ | 0.722 |
| $\text{BERT}_{large}$ | 0.796 |
| DistilBERT | 0.721 |
| $\text{RoBERTa}_{base}$ | 0.775 |
| **$\text{RoBERTa}_{large}$** | 0.813 |
| XLM | 0.714 |

**$\text{RoBERTa}_{\text{large}}$** outperforms all the other language models. This is because it is **pre-trained** on datasets that contain text in a **story-like** format similar to stand-up comedy text.

**$\text{RoBERTa}_{\text{large}}$** and **$\text{BERT}_{\text{large}}$** can distinguish different **levels of humourousness** quite well and they show the highest accuracy in identifying the **non-funny clips**. A larger neural network would need a dataset of significant size to train which shows that our dataset is reasonably sized.

# Error analysis

- In cases of error in assigning ratings to the intermediate funny clips (2-3), the assigned ratings are not off by more than one rating point which correlates with human annotators' incorrect predictions among classes 2-3 and 3-4.
- Most misclassified clips by the model belong to the following categories:
  - Sarcasm and irony
  - Morbid jokes/Dark humour
  - Metaphoric comparisons
- Misclassification among annotators arises due to subjectivity towards topics such as:
  - Country-specificity
  - Bias against country/race
  - Insensitivity towards females

*The last point is validated by observing that such clips are consistently less scored by Annotator A (female) than Annotator B and C (males)

# Conclusion

We train a machine learning model to predict funniness of a standup comedy clip given its audio and text.

We train several textual embeddings to compare accuracy.

We also release our dataset for further analysis called Open Mic.

# THANK YOU