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

Anirudh Mittal†, Pranav Jeevan◇, Prerak Gandhi♣, Diptesh Kanojia‡, Pushpak Bhattacharyya⋆

†,◇,♣,⋆Indian Institute of Technology Bombay, India; ‡Centre for Translation Studies, University of Surrey, United Kingdom

Key Questions

- "How can you automatically rate humour?"
- "Can a machine measure the funniness of a comedy clip?"

Introduction

- Creating datasets for automatic measurement of humour quotient is difficult due to multiple possible interpretations of the content.
- We create a multi-modal humour-annotated dataset (∼40 hours) using stand-up comedy clips.
- We devise a novel scoring mechanism to annotate the training data with a humour quotient score using the audience's laughter.
- The normalized duration (laughter duration divided by the clip duration) of laughter in each clip is used to compute this humour coefficient score on a five-point scale (0-4).
- This method of scoring is validated by comparing with manually annotated scores, wherein a quadratic weighted kappa of 0.6 is obtained.
- We use this dataset to train a model that provides a "funniness" score, on a five-point scale, given the audio and its corresponding text.
- We compare various neural language models for the task of humour-rating and achieve an accuracy of 0.813 in terms of Quadratic Weighted Kappa (QWK).

Dataset - Open Mic

Total Datapoints: 1055 Total hours: 45

We release our dataset 'Open Mic'. 36 English language standup comedy shows from 32 comedians from diverse categories of gender, nationality, and culture, are segmented manually into 927 $\sim$ 2 minute long clips. We also create text files with the transcript for each audio clip. We collect data for "unfunny" samples from TED talk audio clips and segment them into 128 $\sim$ 2 minute audio clips and create text files of their transcript.

Scoring Humour Quotient

The sum of the duration of all the laugh intervals is detected from each clip. Then we divide the sum with the duration of the clip. We use a Likert-scale to regard for the subjectivity in human opinion on each clip. The mean μ and standard deviation σ of all the scores are calculated.

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

Three human annotators (2 males, 1 female) between the ages of 21-33 are assigned to rate the humour quotient in our dataset.

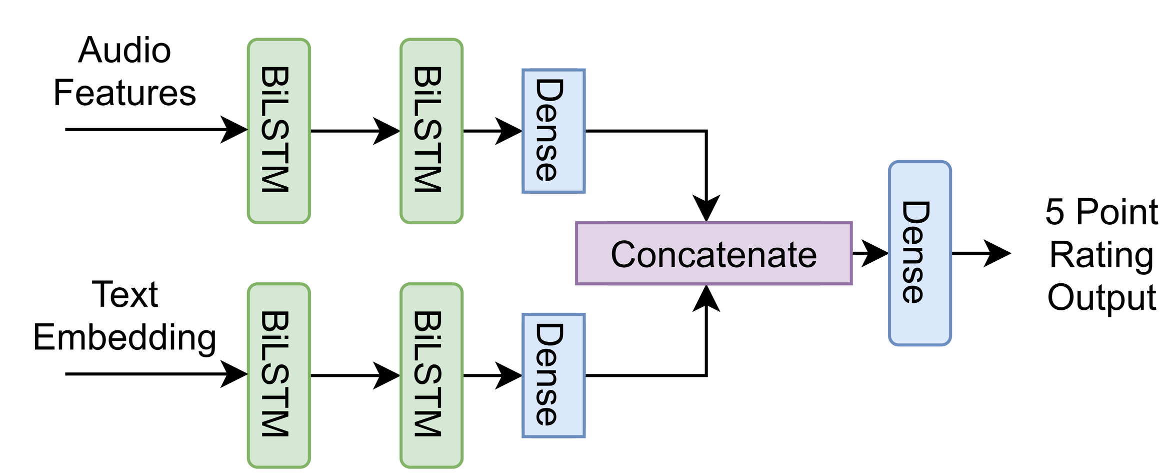
Extracting Audio Features

We remove the audience laughter and isolate the speaker's voice from each clip. Audio features such as MFCCs, RMS energy, and Spectrogram are extracted from the laughter-muted clips. These 3 feature tensors are concatenated to create a single feature of dimension 33 for each time sample. These features convey information about the volume, intonation, and emotion of the speaker, which are important for humour.

Extracting Text Features

We use the textual features extracted from various language models such as BERT$_{base}$, BERT$_{large}$, XLM, DistilBERT, RoBERTa$_{base}$ and RoBERTa$_{large}$ to ensure that the context of each joke is retained. As baseline textual features, we use GloVe embeddings.

Network Architecture



Annotator Agreement

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

Results

Annotaters	QWK
Human A	0.659
Human B	0.562
Human C	0.563
Average	0.595
Textual Features	**QWK**
GloVe	0.691
BERT$_{base}$	0.722
BERT$_{large}$	0.796
DistilBERT	0.721
RoBERTa$_{base}$	0.775
RoBERTa$_{large}$	0.813
XLM	0.714

Observations

- Since RoBERTa is pre-trained on datasets that contain text in a story-like format similar to standup comedy text, RoBERTa$_{large}$ can be seen performing better than all the other textual features.
- Upon futher probing our best-performing model with an ablation test, we observe that audio-based features (0.66 QWK) outperform text-based features (0.48 QWK).
- Our model can identify non-funny clips and most funny clips with very high accuracy. The the assigned ratings are not off by more than one rating point in cases of error.
- Sarcastic and ironic statements, "dark humour", and subtle comparisons that generate human laughter are given low scores by our model

Conclusion

- We propose a novel scoring mechanism to show that humour rating can be automated using audience laughter, which concurs well with the humour perception of humans.
- We create a multi-modal (audio & text) dataset for the task of humour rating
- Our evaluation shows that our scoring mechanism can be emulated with the help of pre-existing language models and traditional audio features.

Dataset & Code Repository

https://github.com/TheExtraSemiColon/AI-OpenMic