“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 (0-1).
• 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 (0-4).

Dataset - Open Mic

Total Datapoints: 1055 Total hours: 45
 We release our dataset ‘Open Mic’. 36 English language stand-up comedy shows from 32 comedians from diverse categories of gender, nationality, and culture, are segmented manually into 927 ~ 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 ~ 2 minute audio clips and create text files of their transcript.

Extraction of Text Features

We use the textual features extracted from various language models such as BERTbase, BERTlarge, XLM, DistilBERT, RoBERTa-base and RoBERTalarge to ensure that the context of each joke is retained. As baseline textual features, we use GloVe embeddings.

Extraction of 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 vector 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.

Dataset & Code Repository

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

Observations

• Since RoBERTa is pre-trained on datasets that contain text in a story-like format similar to standup comedy text, RoBERTalarge can be seen performing better than all the other textual features.
• Upon further 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 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.