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

Anirudh Mittal[†], **Pranav Jeevan**[◇], **Prerak Gandhi**[♣], **Diptesh Kanojia**[‡], **Pushpak Bhattacharyya**[⋆]
[†,◇,♣,⋆]Indian Institute of Technology Bombay, Mumbai
[‡]Centre for Translation Studies, University of Surrey, United Kingdom
[†,♣,⋆]{anirudhmittal,prerakgandhi,pb}@cse.iitb.ac.in
[◇]pranavjp@ee.iitb.ac.in
[‡]d.kanojia@surrey.ac.uk

Abstract

Computational Humour (CH) has attracted the interest of Natural Language Processing and Computational Linguistics communities. Creating datasets for automatic measurement of humour quotient is difficult due to multiple possible interpretations of the content. In this work, 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). Our "Open Mic" dataset is released for further research along with the code.

1 Introduction

Humour is one of the most important lubricants of communication between people. Humour is subjective and, at times, also requires cultural knowledge as humour is often dependent on stereotypes in a culture or a country. At times, even cultural appropriation is used to convey humour, which can be offensive to minority cultures[1] (Rosenthal et al., 2015; Kuipers, 2017). The factors listed above, along with the underlying subjectivity in humour render the task of rating humour, difficult for machines (Meaney, 2020). The task of humour classification suffers due to this subjectivity and the lack of datasets that rate the "funniness" of content.

In this paper, we propose rating humour on a scale of zero to four. We create the first multi-modal dataset[2] using standup comedy clips and compute the humour quotient of each clip using the audience laughter. The validity of our scoring criteria is verified by finding the overall agreement between human annotation and automated scores. We use the audio and text-based signals to process this multi-modal data to generate 'humour ratings'. Since humour annotation is subjective, even the data annotated by humans might not provide an objective measure. We reduce this subjectivity by taking laughter feedback from a larger audience. To the best of our knowledge, no previous literature has proposed an automatically humour-rated multi-modal dataset and used it in ML model-building to automatically obtain the humor score.

Standup comedy is an art form where the delivery of humour has a much larger context, and there are multiple jokes and multiple related punchlines in the same story. The resulting laughter from the audience depends on various factors, including the understanding of the context, delivery, and tonality of the comic. Standup comedy seems to be an ideal choice for a humour rating dataset as it inherently contains some feedback in terms of the audience laughter. We believe a smaller context window restricts computational models, but we know this is not the case for the human audience. Hence, our approach *utilises live audience laughter as a measure to rate the humour quotient* in the data created. We also believe that such an approach can generate insights into what aspects of stories and their delivery make them funny.

Our humour rating model is partly inspired by the character "TARS" from the movie "Interstellar", which generates funny responses based on adjustable humour setting (Nolan, 2014). An essential step in developing such a machine that can adjust its "funniness" is to create a model that can

[†]Corresponding Author
[1]Racism in Comedy: An opinion piece.
[2]Dataset and Code

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 10073–10079
November 7–11, 2021.

recognize and rate the "funniness" of a joke. With this work, we aim to release a dataset that can help researchers shed light on the humour quotient of a particular text. **The key contributions of this paper are:** (a) Creation and public release of an automatically rated multi-modal dataset based on English standup comedy clips and (b) Manual evaluation of this dataset along with humour-rating quotient defined on a Likert-scale (Likert, 1932).

2 Related Work

Most of the previous work on computational humour has been towards the detection of humour. Smaller joke formats like one-liners which have just a single line of context, have been used (Hetzron, 1991). Language models like BERT are used for generating sentence embeddings, which have been shown to outperform other architectures in humour detection on short texts (Annamoradnejad, 2020). Since humour depends on how the speaker's voice changes, the audio features, and language features have been used as inputs for machine learning models for humour detection. Bertero and Fung (2016) use audio and language features to detect humour in The Big Bang Theory sitcom dialogues. Park et al. (2018) passed audio and language features from a conversation dataset into an RNN to create a chatbot that can detect and respond to humour. Hasan et al. (2019) built a multi-modal dataset that uses text, audio, and video inputs for humour detection. There are existing datasets that rate the humour in tweets and Reddit posts, with the help of human annotators (Miller et al., 2020; Castro et al., 2018; Weller and Seppi, 2019). Creating human-annotated datasets is costly in terms of both time and money and has been one of the noted issues for creating humour datasets. Yang et al. (2019a,b) used time-aligned user comments for generating automated humour labels for multi-modal humour identification tasks and found good agreement with manually annotated data. However, none of the previously existing datasets are created with standup comedy clips.

We present the first multi-modal dataset that uses a non-binary rating system. We use standup comedy clips which makes our dataset scalable and diverse. The dataset is novel in terms of the use of long contextual jokes ($\sim$ 2 mins) and audience laughter which helps annotate the funniness in each clip in an automated manner.

3 Dataset Acquisition and Pre-processing

In this section, we describe the creation of our multi-modal dataset and the manual evaluation performed with the help of human annotators.

We gather 36 English language standup comedy shows from 32 comedians available on the web, where the length of each original clip is $\sim$ 1 hour. We further segment them manually into 927 $\sim$ 2 minute long clips. The standups are chosen based on the clarity of the audience feedback laughter. We choose comics from diverse categories of gender, nationality, and culture to ensure representation and reduce bias. While segmenting them, we ignore the clips, which results in laughter on interaction with the audience/personal jokes. We also create text files with the transcript for each audio clip from multiple online sources (Tra, 2020). We collect data for "unfunny" samples by gathering TED talk audio clips with similar speech delivery modes like standup comedy. We also segment them into 128 $\sim$ 2 minute audio clips and create text files of their transcript[3].

Clips were manually trimmed from the complete audio such that the entire context for the joke is available within the clip. This results in the overall set of $\sim$ 2 minute clips described above. Finally, we acquire 519 $\sim$ 2 minute audio clips and corresponding transcript text files in our dataset. The train-test split is chosen to be 70-30.

3.1 Laughter Detection

To find the humour quotient rating of each clip, we use the feedback of the audience laughter as discussed above. We measure the intensity and recorded time intervals of audience laughter in the clip (Gillick and Wlodarczak, 2019). We modify this library to output *the sum of the duration, of all laughs in the clip*. Based on hyperparameter tuning, we set the threshold parameters, adjusting the minimum probability for laughter detection to 0.7. Further, the minimum laughter duration parameter is set to 0.1. This allows us to get the humour quotient from the total duration of the audience laughter in the clip.

3.2 Scoring Humour Quotient

The sum of the duration of all the laugh intervals is detected from each clip. Since longer clips tend to have more jokes and hence a higher score, we eliminate this bias by dividing the sum with the

[3] TED Talks

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 $= 0$

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

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. A rule for assigning a 5 point rating (0-4) for each clip is devised as shown in Table 1 (Column 3). The number of samples for each class in our rating system is shown in Table 1.

3.3 Human Annotation

Three human annotators (2 males, 1 female) between the ages of 21-33 are assigned to rate the humour quotient in our dataset. The annotators are instructed to rate each clip based *solely on the audience laughter feedback* rather than their perception of the humour quotient of the clip. This allows the annotators to be *unbiased towards a particular comedian or humour genre*. The annotations were performed in a *closed-room environment, without any external noise*.

4 Experiment Setup and Methodology

In this section, we describe the features used for the humour rating prediction task along with the additional pre-processing in detail.

4.1 Network Architecture

The text embeddings and audio features are given as input to separate Bi-LSTM layers followed by separate, Dense layers (Graves, Alex and Fernández, Santiago and Schmidhuber, Jürgen, 2005) as shown in Figure 1. The output from these two pathways is then concatenated and fed to a classifier that outputs one-hot encoding of the 5-point rating.

4.2 Muting Laughter

Before extracting audio features, we remove the audience laughter and isolate the speaker's voice from each clip. Retaining the audience laughter may enable a neural network to utilize it and predict a score without using information from the text

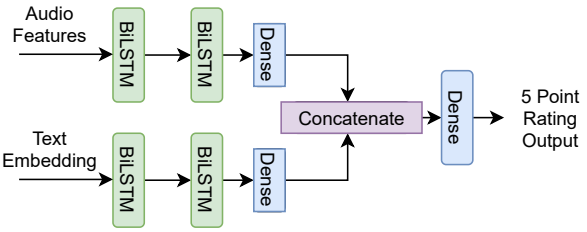

Figure 1: Neural Network Architecture

and other audio features. We envision creating a system that can predict the funniness of any clip. Such clips will not have laughter as an indicator, so we train and test our model on the muted clips. Please note that laughter is extracted separately to generate the funniness score (Section 3.1). We use Green (2018) to mute audience laughter from audio segments, thus, resulting in clips that are then used for extracting audio features.

4.3 Audio Features

Audio features such as MFCCs, RMS energy, and Spectrogram are extracted from the laughter-muted clips (McFee et al., 2020). These 3 feature tensors are concatenated to create a single feature of dimension 33 for each time sample. The maximum sequence length for audio embeddings was set as 8000. The clips with a lesser duration were padded with zeroes for uniformity. These features convey information about the volume, intonation, and emotion of the speaker, which are important for humour.

4.4 Textual Features

Additionally, we use the textual features extracted from various language models to ensure that the context of each joke is retained. We use BERT-derived models to generate contextual embeddings for each clip, which ensure attention over the entire text of the clip (Wolf et al., 2020). BERT-derived models can process sequences of token length 512; thus, we employ them for the entire transcript of each $\sim$ 2 minute clip. We sum the output of the final 4 layers from these models to obtain a clip embedding (Alammar, 2018).

As baseline textual features, we use GloVe embeddings (Pennington et al., 2014). For obtaining textual features, we experiment with BERT$_{base}$, BERT$_{large}$, XLM, DistilBERT, RoBERTa$_{base}$ and RoBERTa$_{large}$ to generate text embeddings (Devlin et al., 2018; Lample and Conneau, 2019; Sanh et al., 2019; Liu et al., 2019).

4.5 Methodology

The audio features and textual features are fed as input to the network for obtaining an output rating on the scale of 0-4. To evaluate our approach for scoring each clip, we obtain Cohen's weighted Kappa with quadratic weights, *i.e.,* Quadratic Weighted Kappa score (QWK) (Cohen, 1968) between our scoring mechanism (Table 1) and the model output. We use QWK as a scoring mechanism because, unlike accuracy and F-Score, it considers that the system may randomly assign a particular label to a clip. The QWK score also penalizes mismatches more than linear or unweighted Kappa by taking the quadratic weights into account. Additionally, we validate the scores provided by our scoring mechanism by obtaining QWK with the human annotation performed.

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

5 Results

In Table 2, we show the Krippendorff alpha, Fleiss' Kappa, and pairwise agreement between human annotators (Krippendorff, 2004; Cohen, 1960; Fleiss et al., 1971). The inter-annotator agreement between any two annotators is above 0.60, which signifies "substantial" agreement between them (Fleiss et al., 2003). We evaluate our scoring mechanism by comparing it with the manually annotated data by human annotators, as shown in Table 3. An average QWK of 0.595 was observed, indicating significant agreement with them (Vanbelle, 2014).

Table 3 also shows the QWK among the neural network outputs with our scoring mechanism. With the neural network output, we see a significant agreement when compared with our scoring mechanism. Even the GloVe-based model performs reasonably well when matched with our scoring mechanism. Embeddings created from

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

Table 3: (a) Quadratic Weighted Kappa (QWK) scores between the scores provided by human annotators, and our scoring mechanism (b) QWK scores between the various language models combined with neural networks, and our scoring mechanism.

BERT-derived language models showed considerable improvement from baseline performance. RoBERTa$_{large}$ outperforms all the other language models and shows an improvement of 12% points over the baseline GloVe score. Since RoBERTa is pre-trained on datasets that contain text in a story-like format similar to standup comedy text (Liu et al., 2019), RoBERTa$_{large}$ can be seen performing better than all the other textual features. Analysing the confusion matrix of these models shows that RoBERTa$_{large}$ and BERT$_{large}$ can distinguish different levels of humourousness quite well. They show the highest accuracy in identifying the non-funny clips. DistilBERT could not perform as well as BERT$_{large}$ because humour needs better quality text embeddings to understand the full context, which DistilBERT cannot provide due to the lower number of parameters in the model.

Larger models with embedding dimensions of 1024 (BERT$_{large}$, RoBERTa$_{large}$) and 2048 (XLM) performed better than smaller models. A larger neural network would need a dataset of significant size to train, which also shows that our dataset is reasonably sized. For BERT$_{base}$, when we increased the Bi-LSTMs in the initial layer from 256 to 512, we see a slight improvement in the Quadratic Weighted Kappa value which shows that larger embeddings need a bigger neural network to classify accurately.

We further probe our best-performing model with an ablation test and observe that audio-based

features (0.66 QWK) outperform text-based features (0.48 QWK). This contradicts what was observed by Hasan et al. (2019) as humour in standup often depends on the tonality and the well-enunciated punchlines.

6 Discussion

Analysis of the predicted ratings show that our model can identify non-funny clips and most funny clips with very high accuracy. In cases of error in assigning ratings to the intermediate funny clips, the assigned ratings are not off by more than one rating point, for *e.g.,* a clip rated 3 is assigned a rating of 2 or 4. This error should not be considered as a failure of our model since assigning a precise funniness rating in a definite way to intermediate funny clips is hard even for humans. So it is reasonable to expect our model to commit similar errors in assigning the ratings as a human would. In the individual confusion matrices obtained for both feature sets, we observe the maximum incorrect predictions among the classes 2/3 and 3/4. We correlate these results with the human annotation and observe that even human annotators differ mostly in these two classes. All our annotators observed that clips with a moderate amount of laughter could be rated either as 2 or as 3, since such annotations are difficult to be discerned to a particular class.

Additionally, we observe that there are only $16/1055$ cases where none of the ratings of the three human annotators match with each other. Out of which, only one rating differs (4, 1, 2) with a difference of $>= 2$. In the other 15 ratings, the difference between the highest and lowest human ratings is $<= 2$ (*e.g.,* 4-2-3). There are around $\sim$400 cases where 2 annotators fully agree. The rest of the $\sim$600 ratings are where all three annotators fully agree in their ratings for each clip shown to them.

As we evaluate the clips misclassified by our model, we observe that 1) sarcastic and ironic statements generate human laughter, but our model does not detect it, 2) a certain kind of jokes which are morbid also categorized as "dark humour" is consistently classified with lower scores, whereas there is a lot of human laughter generated during such jokes, 3) subtle comparisons, for example, the usage of internet to smoking where the comedian tries to imply that both are harmful to health; are classified as "mildly funny (1 or 2)" by our model.

We further evaluate clips with human annotation score difference > 2. Despite providing detailed guidelines which required our annotators to focus only on audience laughter, they could have possibly focused on the content. Due to this subjectivity, we believe that our annotators may have misclassified a few clips. We trace the reasons to 1) country-specificity, thereby leading to less comprehension by the annotator, or 2) insensitivity towards the feelings of females, or 3) bias against a country/race which again leads to the diminished absorption of the joke. This observation validates our initial discussion on the subjectivity of humorous content, along with the observation that Annotator A (female) has consistently scored such clips lower than Annotator B and C (males). However, we also observed that the audience laughter in such clips is more consistent with the scores provided by Annotators B and C.

7 Conclusion and Future Work

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-model (audio & text) dataset for the task of humour rating. With the help of three human annotators, we manually evaluate our scoring mechanism and show a substantial agreement in terms of QWK. Our evaluation shows that our scoring mechanism can be emulated with the help of pre-existing language models and traditional audio features. Our neural network-based experiments show that the output obtained using various language models like RoBERTa show an agreement with our scoring mechanism. Despite the inherent subjectivity in humour and its different perceptions among humans, we propose a method to rate humour and release this dataset under the CC-BY-SA-NC 4.0 license for further research.

In the future, we would like to evaluate this scoring mechanism with the help of more human annotators. We aim to extend the dataset with the help of more standup comedy clips. Further experiments can be conducted to compare the contribution of audio, video and text features with a more detailed analysis. We would also like to perform experiments by including more audio features like Line Spectral Frequencies, Zero-Crossing rate, and Delta Coefficients. With the release of this dataset, we hope that research in computational humour can be taken further.

References

2020. Stand-Up Comedy Transcripts.

Jay Alammar. 2018. The Illustrated BERT, ELMo, and co. (How NLP Cracked Transfer Learning).

Issa Annamoradnejad. 2020. ColBERT: Using BERT Sentence Embedding for Humor Detection.

Dario Bertero and Pascale Fung. 2016. Deep Learning of Audio and Language Features for Humor Prediction. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 496–501, Portorož, Slovenia. European Language Resources Association (ELRA).

Santiago Castro, Luis Chiruzzo, Aiala Rosá, Diego Garat, and Guillermo Moncecchi. 2018. A Crowd-Annotated Spanish Corpus for Humor Analysis. In *Proceedings of the Sixth International Workshop on Natural Language Processing for Social Media*, pages 7–11, Melbourne, Australia. Association for Computational Linguistics.

J. Cohen. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37.

Jacob Cohen. 1968. Weighted kappa: nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin*, 70(4):213.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

J.L. Fleiss et al. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

Joseph Fleiss, Bruce Levin, and Myunghee Paik. 2003. In Statistical Methods for Rates and Proportions. *Statistical Methods for Rates and Proportions*, 203.

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

Graves, Alex and Fernández, Santiago and Schmidhuber, Jürgen. 2005. Bidirectional lstm networks for improved phoneme classification and recognition. pages 799–804.

Jeff Green. 2018. Sitcom laughtrack mute tool.

Md Kamrul Hasan, Wasifur Rahman, AmirAli Bagher Zadeh, Jianyuan Zhong, Md Iftekhar Tanveer, Louis-Philippe Morency, and Mohammed (Ehsan) Hoque. 2019. UR-FUNNY: A multimodal language dataset for understanding humor. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2046–2056, Hong Kong, China. Association for Computational Linguistics.

R. Hetzron. 1991. On the structure of punchlines. *Humor: International Journal of Humor Research*, 4:61–108.

Klaus Krippendorff. 2004. *Content Analysis: An Introduction to Its Methodology (second edition)*. Sage Publications.

Giselinde Kuipers. 2017. *In The Anatomy of Laughter*, chapter Humour styles and class cultures: Highbrow humour and lowbrow humour in the netherlands. Routledge.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *Advances in Neural Information Processing Systems (NeurIPS)*.

Rensis Likert. 1932. A technique for the measurement of attitudes. *Archives of psychology*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *CoRR*, abs/1907.11692.

Brian McFee, Vincent Lostanlen, Alexandros Metsai, Matt McVicar, Stefan Balke, Carl Thomé, Colin Raffel, Frank Zalkow, Ayoub Malek, Dana, Kyungyun Lee, Oriol Nieto, Jack Mason, Dan Ellis, Eric Battenberg, Scott Seyfarth, Ryuichi Yamamoto, Keunwoo Choi, viktorandreevichmorozov, Josh Moore, Rachel Bittner, Shunsuke Hidaka, Ziyao Wei, nullmightybofo, Darío Hereñú, Fabian-Robert Stöter, Pius Friesch, Adam Weiss, Matt Vollrath, and Taewoon Kim. 2020. librosa/librosa: 0.8.0.

J. A. Meaney. 2020. Crossing the Line: Where do Demographic Variables Fit into Humor Detection? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 176–181, Online. Association for Computational Linguistics.

Tristan Miller, Erik-Lân Do Dinh, Edwin Simpson, and Iryna Gurevych. 2020. OFAI-UKP at HAHA@IberLEF2019: Predicting the Humorousness of Tweets Using Gaussian Process Preference Learning.

Christopher Nolan. 2014. Interstellar.

Kate M. Park, Annie Hu, and Natalie Muenster. 2018. Laughbot: Detecting Humor in Spoken Language with Language and Audio Cues.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global Vectors for Word Representation. volume 14, pages 1532–1543.

A. Rosenthal, David Bindman, and A.W.B. Randolph. 2015. *No laughing matter: Visual humor in ideas of race, nationality, and ethnicity*.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108.

Sophie Vanbelle. 2014. A New Interpretation of the Weighted Kappa Coefficients. *Psychometrika*.

Orion Weller and Kevin Seppi. 2019. Humor Detection: A Transformer Gets the Last Laugh.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Zixiaofan Yang, , L. Ai, and Julia Hirschberg. 2019a. Multimodal Indicators of Humor in Videos. In *2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, pages 538–543.

Zixiaofan Yang, Bingyan Hu, and Julia Hirschberg. 2019b. Predicting Humor by Learning from Time-Aligned Comments. In *Proc. Interspeech 2019*, pages 496–500.