Neural LOLgorithm:
Deep Learning for Multi-Modal Stand-Up Comedy Data

Gautam Gupta
Dept. of DSAI
IIIT Naya Raipur
Chhattisgarh, India
gautam21102@iiitnr.edu.in

Mrinal Bhan
Dept. of DSAI
IIIT Naya Raipur
Chhattisgarh, India
mrinal21102@iiitnr.edu.in

Ajay Kumar
Dept. of CSE
IIIT Naya Raipur
Chhattisgarh, India
ajay21100@iiitnr.edu.in

Kavita Jaiswal
Dept. of CSE
IIIT Naya Raipur
Chhattisgarh, India
kavita@iiitnr.edu.in

***Abstract*— Humor plays a crucial role in facilitating communication among people. However, it can be highly subjective and may depend on cultural context and stereotypes. Additionally, some forms of humor may involve cultural appropriation, which can be offensive to certain groups.These factors, along with the inherently subjective nature of humor, make it challenging for machines to accurately classify and rate the funniness of content. Furthermore, there is a lack of sufficient data and resources for training algorithms in this task.We propose rating humor on a ten-point ratio scale. We create the first multi modal and dynamic dataset using standup comedy clips from YouTube and other media platforms and compute the humor quotient of each clip using both audio and textual features. Our goal is todetect humor by preserving all three textual, audio and video features to have better prediction scores, while current methodologies are limited to using only one of the three.**

Since humor 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 largeraudience.

I. INTRODUCTION

Understanding humor is a quintessential human task that is not so well understood using currently prevalent AI (Artificial intelligence) systems. At times, even cultural appropriation is used to convey humour, which can be offensive to minority cultures. The factors listed above, along with the underlying subjectivity in humour render the task of rating humour, difficult for machines. In this paper we focus on stand-up comedy to make the dataset to understand and improve upon the already done research in field of computational humour.

Stand-up comedy is a show or performance in which a comedian performs original jokes on stage in front of a live audience to make them laugh. The jokes are scripted and have setups and punchlines. The average stand-up comedy show gets four to six laughs a minute from theaudience. We aim to build a model that rates the funniness of these clips based on the audio, video and textual features of these stand-up clips available on multiple OTT platforms and YouTube. We use different open source toolkits to analyze the laughter characteristics of clips extracted from the video as well as the facial expressions and body movements because humour is no only limited only to audio and text. We train our neural networks-based model based on this extracted data and provide a score which ranges from 1.00 to 10.00.

.

II. LITERATURE REVIEW

Most of the previous work on computational humour has been towards the detectionof humour. Smaller joke formats like one-liners which have just a single line of context, have been used. [1] A Sentiment and Emotion aware Multimodal Multiparty Humor Recognition in Multilingual Conversational Setting - Dushyant Singh Chauhan, etc makes use of the Multimodal Multiparty Hindi Humor (M2H2) dataset comprising audio features. [2] Making People Laugh like a Pro": Analyzing Humor Through Stand-Up Comedy - Beatrice Turano, Carlo Strapparava used (SCRIPTS) consists of only text features. [3] Multimodal Humor Dataset: Predicting Laughter tracks for Sitcoms - Badri N. Patro,etc. use audio and language features todetect humour in The Big Bang Theory sitcomdialogues. 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 [4] "Survey on Computational Humour" - Diptesh Kanojia, Pushpak Bhattacharyya wasbased on humour only from short jokes,tweets,etc. [5] Hasan et al. (2019) built a multi-modal dataset that uses text, audio, and video inputs for humour detection, but applied a binary classification model. There are existing datasets that rate the humour in tweets and Reddit

posts [6] “So You Think You’re Funny?”: Rating the Humour Quotient in Standup Comedy Anirudh Mittal Pranav Jeevan Prerak Gandhi Diptesh Kanojia Pushpak Bhattacharyya manually annotated scores, wherein a quadratic weighted kappa of 0.6is obtained. Using that dataset to train a model that provides a “funniness” score, on afive-point scale, given the audio and its corresponding text [7] “Multimodal Humor Dataset” : Predicting Laughter tracks for Sitcoms - Mayank Lunayuch, predicts timestamps for laughter based on conversations of fictional characters only. [8] Zixiaofan Yang, Bingyan Hu, and Julia Hirschberg 2019b. Predicting Humor by Learning from TimeAligned Comments used time-aligned user comments for generating automated humour labels for multimodal humour identification tasks.

III. METHODOLOGY

We use various media platforms like Netflix, Amazon, YouTube for the collection of multilingual videos. Laughter is detected and the time stamps are labelled manually segregating them into short clips containing pure laughter audio. The extracted laughter clips are then analysed and the mean and standard deviation are calculated for various characteristics of the audio like pitch, intensity and amplitude.

.

A. Dataset Descriptions

The dataset contains the attributes of the various length clips made from a particular video like the time stamps for when the laughter starts and ends in the clip, laughter score values, frequency of a segment, the time stamps for start and end of the clip .

video	laughter_start	laughter_end	laughter_value	start_segment_s	end_segment_s	start_segment_frame	end_segment_frame
VirDas	5	8	7	0	5	1	150
VirDas	10	12	7	8	10	241	300
VirDas	20	21	7	12	20	361	600
VirDas	47	48	7	21	47	631	1410
VirDas	57	62	7	48	57	1441	1710
VirDas	65	66	8	62	65	1861	1950
VirDas	68	72	8	66	68	1981	2040
VirDas	82	83	8	72	82	2161	2460
VirDas	95	96	8	83	95	2491	2850
VirDas	103	104	8	96	103	2881	3090
VirDas	110	111	7	104	110	3121	3300
VirDas	115	116	7	111	115	3331	3450
VirDas	127	131	7	116	127	3481	3810
VirDas	144	145	7	131	144	3931	4320
VirDas	154	155	7	145	154	4351	4620
VirDas	157	158	5	155	157	4651	4710
VirDas	166	168	5	158	166	4741	4980
VirDas	173	174	5	168	173	5041	5190
VirDas	177	181	5	174	177	5221	5310

Fig : Dataset

B. Audio Feature extraction

The next step is to gather the materials required to createour model. To do this, the classification characteristics ofeach audio sample were visually represented using the same techniques as the high-precision classification of photographs.

Mel frequency Cepstral coefficients (MFCCs): Using the MFCC, we were able to retrieve resources for each audio file in the collection, resultingin a picture display for each audio sample. We might also train the classifier using these images.

$$C_n = \sum_{k-1}^{k} (\log D_k) \cos\left[m\left(k - \frac{1}{2}\right)\frac{\pi}{k}\right]$$

where m = 0, 1… k- 1

YAMNet: It is a pre-trained neural network that employs the MobileNetV1 depth-wise and separable convolution architecture. It can use an audio waveform as input and make independent predictions for each of the 521 audio events from the Audio Set corpus. Internally, the model extracts "frames" from the audio signal and processes batches of these frames. This version of the model uses frames that are 0.96 second long and extracts one frame every 0.48 seconds .The model accepts a 1-D float32 Tensor or NumPy array containing a waveform of arbitrary length, represented as single-channel (mono) 16 kHz samples in the range [-1.0, +1.0].The model returns 3 outputs, including the class scores, embeddings (which you will use for transfer learning), and the log mel spectrogram.

Librosa: is valuable Python music and sound investigation library that helps programming designers to fabricate applications for working with sound and music document designs utilizing Python. This Python bundle for music and sound examination is essentially utilized when we work with sound information, like in the music age (utilizing Lstm's), Automatic Speech Recognition. The library upholds a few elements connected with sound records handling and extraction like burden sound from a circle, register of different spectrogram portrayals, symphonious percussive source detachment, conventional spectrogram decay, stacks and translates the sound, Time-space sound handling, successive demonstrating, coordinating consonant percussive partition, beat-simultaneous and some more.

GloVe,: coined from Global Vectors, is a model for distributed word representation. The model is an unsupervised learning algorithm for obtaining vector representations for words. This is achieved by mapping words into a meaningful space where the distance between words is related to semantic similarity.[1] Training is performed on aggregated global word-word co-occurrence statistics from a corpus, and the resulting representations showcase interesting linear substructures of the word vector space. As log-bilinear regression model for unsupervised learning of word representations, it combines the features of two model families, namely the global matrix factorization and local context window methods

C. Video Features Extraction

We have used **OpenPose** which is a real-time multi-person system to jointly detect human body, hand, facial, and foot key-points (in total 135 key-points) on single image.

OpenFace: OF is an open source tool intended for computer vision and machine learning researchers, the affective computing community and people interested in building

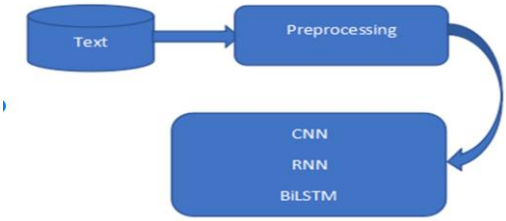
interactive applications based on facial behavior analysis. It is the first open source tool capable of facial landmark detection, head pose estimation, facial action unit recognition, and eye-gaze estimation.

D. Text Feature Extraction

BERT: BERT (Bi-directional Encoder Representations for Transformer) makes use of a Transformer that learns contextual relations between words in a sentence/text. The transformer includes 2 separate mechanisms: an encoder that reads the text input and a decoder that generates a prediction for any given task.

RoBERTa: Robustly Optimized BERT Pretraining Approach has almost similar architecture as compare to BERT, but in order to improve the results on BERT architecture, the authors made some simple design changes in its architecture and training procedure. Training with bigger batch sizes & longer sequences: Originally BERT is trained for 1M steps with a batch size of 256 sequences. In this paper, the authors trained the model with 125 steps of 2K sequences and 31K steps with 8k sequences of batch size improving perplexity on masked language modelling objective and as well as end-task accuracy. Large batches are also easier to parallelize via distributed parallel training. In BERT architecture, the masking is performed once during data pre-processing, resulting in a single static mask. To avoid using the single static mask, training data is duplicated and masked 10 times, each time with a different mask strategy over 40 epochs thus having 4 epochs with the same mask. This strategy is compared with dynamic masking in which different masking is generated every time we pass data into the model.

Word2Vec: W2V creates vectors of the words that are distributed numerical representations of word features – these word features could comprise of words that represent the context of the individual words present in our vocabulary. Word embeddings eventually help in establishing the association of a word with another similar meaning word through the created vectors.



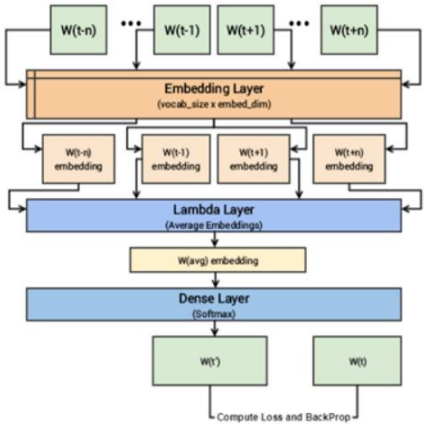


Fig III: Word2Vec Architecture

Generative Pre-trained Transformer 3 (GPT-3): GPT-3 is an autoregressive language model released by OpenAI in 2020 that uses deep learning to produce human-like text. When given a prompt, it will generate text that continues the prompt. The architecture is a decoder-only transformer network with a 2048-token-long context and then-unprecedented size of 175 billion parameters, requiring 800GB to store. The model was trained using generative pre-training; it is trained to predict what the next token is based on previous tokens. The model demonstrated strong zero-shot and few-shot learning on many tasks.

E. Calculation Of Target Variable

After all of the different features about the laughter clips have been extracted like start time, end time, mean pitch ,mean intensity, standard pitch, standard intensity we use the following formula to generate a "laughter value"

$$laughter\_value = \frac{\left(pitch - \sum_{i=1}^{n} \frac{pitch}{n}\right)}{Sd\_pitch} + \frac{\left(intensity - \sum_{i=1}^{n} \frac{intensity}{n}\right)}{Sd\_intensity}$$

We take the sum of all the laughter values generated and this gives us the "laughter score". Since the length of each video is different the laughter scores need to be normalized by diving them by the length of the clip and a "Target Variable" is acquired.

F. Model Training

LSTM: LSTM(Long Short Term Memory) extension of RNN can attain the best speech recognition accuracy to date. As a result, even though feature extraction is not carried out similarly to CARM, we recommend adopting LSTM for activity recognition rather than other traditional machine learning techniques, such as HMM. There are two benefits to LSTM usage. – First, the LSTM can extract the features automatically; in other words, there is no necessity to preprocess the data. Second, LSTM can hold temporal state

information of the activity, i.e., LSTM has the potential to distinguish similar activities like “Lie down” and “Fall.” Since “Lie down” consists of “Sit down” and “Fall,” the memory of LSTM can help in recognition of these activities.

A general LSTM unit is composed of a cell, an input gate, an output gate, and a forget gate. The cell remembers values over arbitrary time intervals, and three gates regulate the flow of information into and out of the cell. LSTM is well-suited to classify, process, and predict the time series given of unknown duration.

G. *Model Architecture*

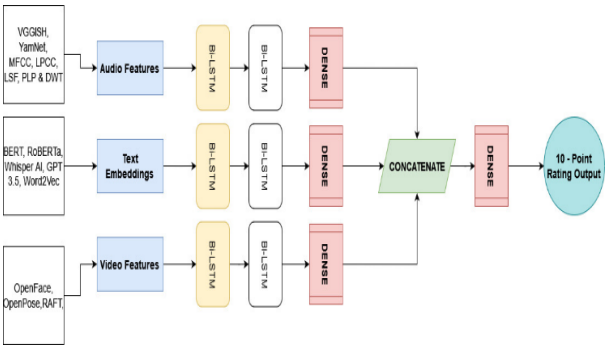


Fig III: Flow Diagram of Model

The text embeddings, audio and video 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 . The output from these two pathways is then concatenated and fed to a classifier that outputs a 10-point rating.

IV. RESULTS

Fig IV: Rating example

The model was trained on multimodal dataset and a humour rating was given on a 10-point scale i.e. (0-4): Neutral, (5-6): Funny, (7-8) Humorous and (9-10): Hilarious. In Fig IV, an example is showed to depict the humour rating scale in a comic scene ranging from 4 (Neutral) to 9 (Hilarious). For Textual embeddings, different models such as GloVE, BERT and RoBERTa were used showing a QWK score between 0.691 and 0.775.

Textual Features	Quadratic Weighted Kappa (QWK)
GloVe	0.691
BERT *base*	0.712
BERT *large*	0.796
DistilBERT	0.721
RoBERTa*base*	0.775
RoBERTa*large*	0.813

V. CONCLUSION

The project proposes a novel scoring mechanism for automated humour rating using audience laughter. A multi-model dataset which includes audio, video and text features is created for the task of humour rating, and the features are passed through the LSTM model to give the score . The scoring mechanism can be emulated using pre-existing language models and traditional audio features, with neural network-based experiments. The dataset will be released for further research and future work includes expanding the dataset and conducting experiments to compare the contribution of different features.

VI. FUTURE SCOPE

The dataset includes various features such as language, material, humor, and other characteristics of comedians that can be used for analysing and comparing the contributions of different features towards creating humor. Furthermore, we proposes expanding the dataset by adding more diverse comedians in terms of language, material, humor, etc. This would increase the range of the dataset and provide a wider range of data to train and test various humor generation algorithms. In inclusion of old styles of mime and slapstick comedy, which may not necessarily have textual features but can be analyzed based on video features. This can be achieved by including acts from comedians like Rowan Atkinson and Charlie Chaplin. Finally, the dataset can also be used to learn about humor marketing. This can be achieved by analyzing the preferences and patterns of the audience towards humor in marketing and advertising. By understanding the audience's preferences and attitudes towards humor, businesses can design more effective marketing campaigns that resonate with their target audience.

VII. REFFERRENCES

[1]. Dushyant Singh Chauhan, Gopendra Vikram Singh, Aseem Arora, Asif Ekbal, and Pushpak Bhattacharyya. 2022: “A Sentiment and Emotion Aware Multimodal Multiparty Humor Recognition in Multilingual Conversational Setting.” In Proceedings of the 29th International Conference on

Computational Linguistics, pages 6752–6761, Gyeongju, Republic of Korea. International Committee on Computational Linguistics

[2]. Beatrice Turano and Carlo Strapparava. 2022. “Making People Laugh like a Pro: Analysing Humor Through Stand-Up Comedy.” In Proceedings of the Thirteenth Language Resources and Evaluation Conference, pages 5206–5211, Marseille, France. European Language Resources Association.

[3]. Patro, Badri & Lunayach, Mayank & Srivastava, Deepankar & Sarvesh, Sarvesh & Singh, Hunar & Namboodiri, Vinay. (2021). “Multimodal Humor Dataset: Predicting Laughter tracks for Sitcoms.” 576-585. 10.1109/WACV48630.2021.00062.

[4]. “Survey on Computational Humour” - Anirudh Mittal, Diptesh Kanojia, Pushpak Bhattacharyya (Department of Computer Science and Engineering, IIT Bombay, University of Surrey)

[5]. Mittal, Anirudh & P., Pranav Jeevan & Gandhi, Prerak & Kanojia, Diptesh & Bhattacharyya, Pushpak. (2021). "So You Think You're Funny?": Rating the Humour Quotient in Standup Comedy

[6]. Patro, Badri N., Mayank Lunayach, Iit Kanpur, Deepankar Srivastava and Vinay P. Namboodiri. “Multimodal Humor Dataset: Predicting Laughter tracks for Sitcoms.” 2021 IEEE Winter Conference on Applications of Computer Vision (WACV) (2021): 576-585.

[7]. Annamoradnejad, Issa & Zoghi, Gohar. (2020). ColBERT: Using BERT Sentence Embedding for Humor Detection.

[8]. Aggarwal, Akshita & Wadhawan, Anshul & Chaudhary, Anshima & Maurya, Kavita. (2020). "Did you really mean what you said?" : Sarcasm Detection in Hindi-English Code-Mixed Data using Bilingual Word Embeddings.

[9]. Winters, T. (2021). Computers Learning Humor Is No Joke. Harvard Data Science Review, 3(2). https://doi.org/10.1162/99608f92.f13a2337

[10].https://stephenjkaplan.github.io/2020/09/18/standup-comedy-recommender/

[11]. https://github.com/adich23/Deep-Humor

[12]. https://github.com/nwams/nlp-stand-up-comedy

[13].https://github.com/singhya/MultimodalStandUpComedy Analysis