Supplementary material for "Multimodal Humor Dataset: Predicting Laughter tracks for Sitcoms"

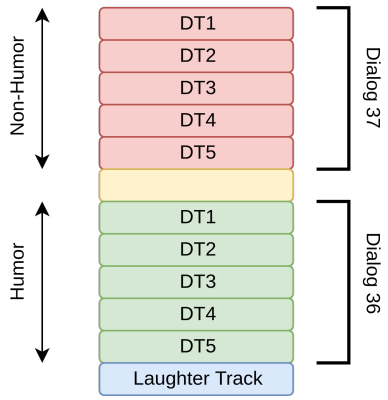


Figure 1: Figure showing the dataset creation for Dialog having 5 Dialog Turns. **DT** stands for Dialog Turn.

1. Overview

Figure 1 demonstrates the creation of the dataset. We provide relevant theoretical background about the models used in the Section 3. ROC plots of the test results obtained are shown in the Figure 4, and additional data analysis is discussed in the Section 2.

2. Data analysis

Words distribution: Figure 5 shows the words used by the Top 6 speakers in our dataset. We observe that distribution for different speakers is more or less the same.

Laughter time: Figure 3 shows the distribution of the laughter track duration across the dataset.

tSNE plot: We made a t-SNE plot by randomly selecting 1500 images from the last frames of the last dialog turns in the dataset. This is shown in Fig 6, 7. A random distribution like this hints towards absence of visual bias.

Bubble plot: A bubble plot to visualize the vocabulary of humorous and non-humorous dialogs is shown in Figure 2.

3. Model details

Below we describe the implementation of Text and Video based attention models separately:

- **Text Attention Model (TAM)** Given a sequence of dialog turns $(d_1, d_2, \ldots d_t)$, first we append special START and STOP tokens indicating start and end mark of each dialog turn. Then, we make a V dimensional one-hot vector representation for every word in the dialog turn and transform it to a real valued word representation g_{WE}, using a matrix $\theta_w \in \mathcal{R}^{l_w \times V}$ which is parameterized by a function $G(d_w, \theta_w)$. Then, the l_w dimensional word embeddings are fed to a LSTM for obtaining a l-dimensional hidden state representation h_t for each input dialog turn $d_t, t \in 1 \ldots T$. (T=5 when no. of dialog turns are 5).

  We then use a single head self-attention method to obtain attention features over each turn as follows,

  $$Attention(Q, K, V) = softmax(\frac{QK}{\sqrt{n}})V$$

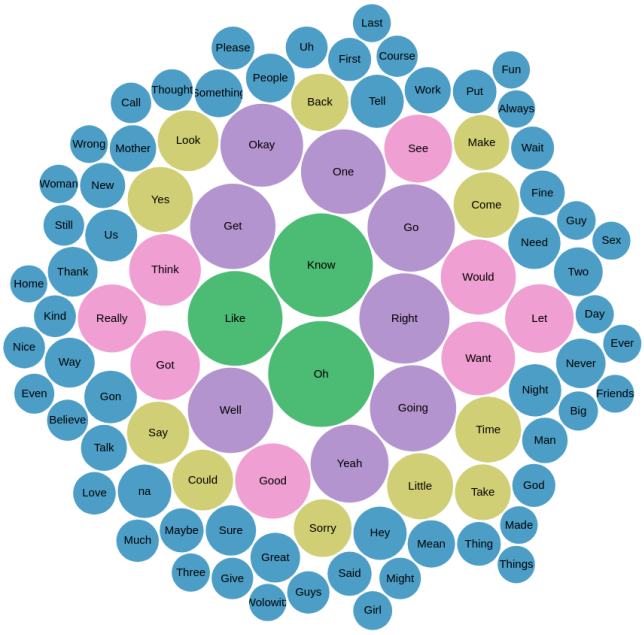
  where, the queries (Q), the keys (K) and values (V) are the encoder hidden states. We further used Transformer [3] and BERT [1] architecture to get better attention embedding for each dialog turn and refer the variants as TAM_Tran and TAM_BERT. That is,

  $$MultiHead(Q, K, V) = [h_1 \dot{h}_h]W_O$$

  where $h_i == Attention(QW_i^Q, KW_i^K, VW_i^V)$.

- **Video based Attention Model** Given an input video X_v, we obtain a vector $g_v \epsilon \mathcal{R}^{4096}$, a 4096 dimensional C3D [2] video feature embedding from the fc-7 layer of a C3D CNN network (parameterized by a function $G_i(x_v, \theta_v)$, where θ_v are the weights of the convolutional layers). We also use self attention on video turns.

(a) Humor



(b) Non-Humor

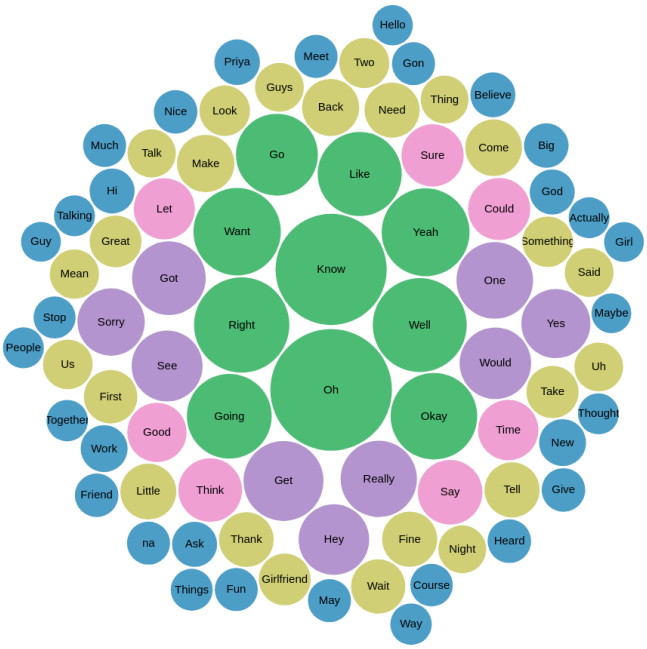


Figure 2: Bubble plots drawn to visualize the vocabulary of humorous and non-humorous Dialogs. Similarity between the two suggests that simple bag of words based humor detection methods would not give good results. (The relative size of the bubble gives measure of the frequency of the words)

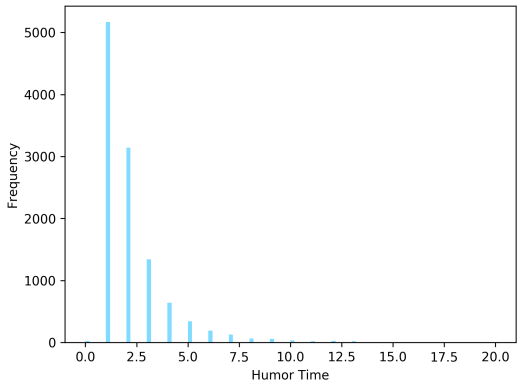


Figure 3: This figure shows the laughter track duration distribution across the dataset.

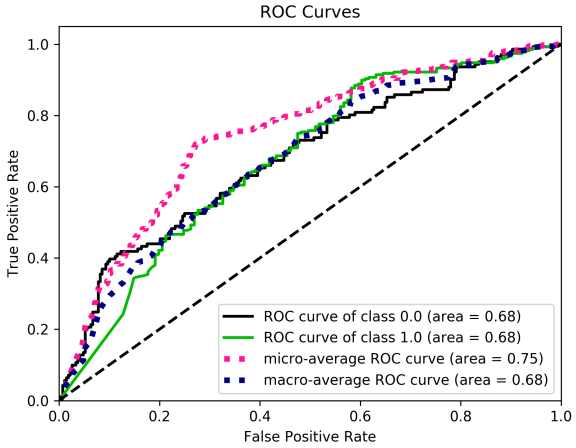


Figure 4: ROC curve for the test set. Class 1 represents the humor class, and Class 0, the non-humor class.

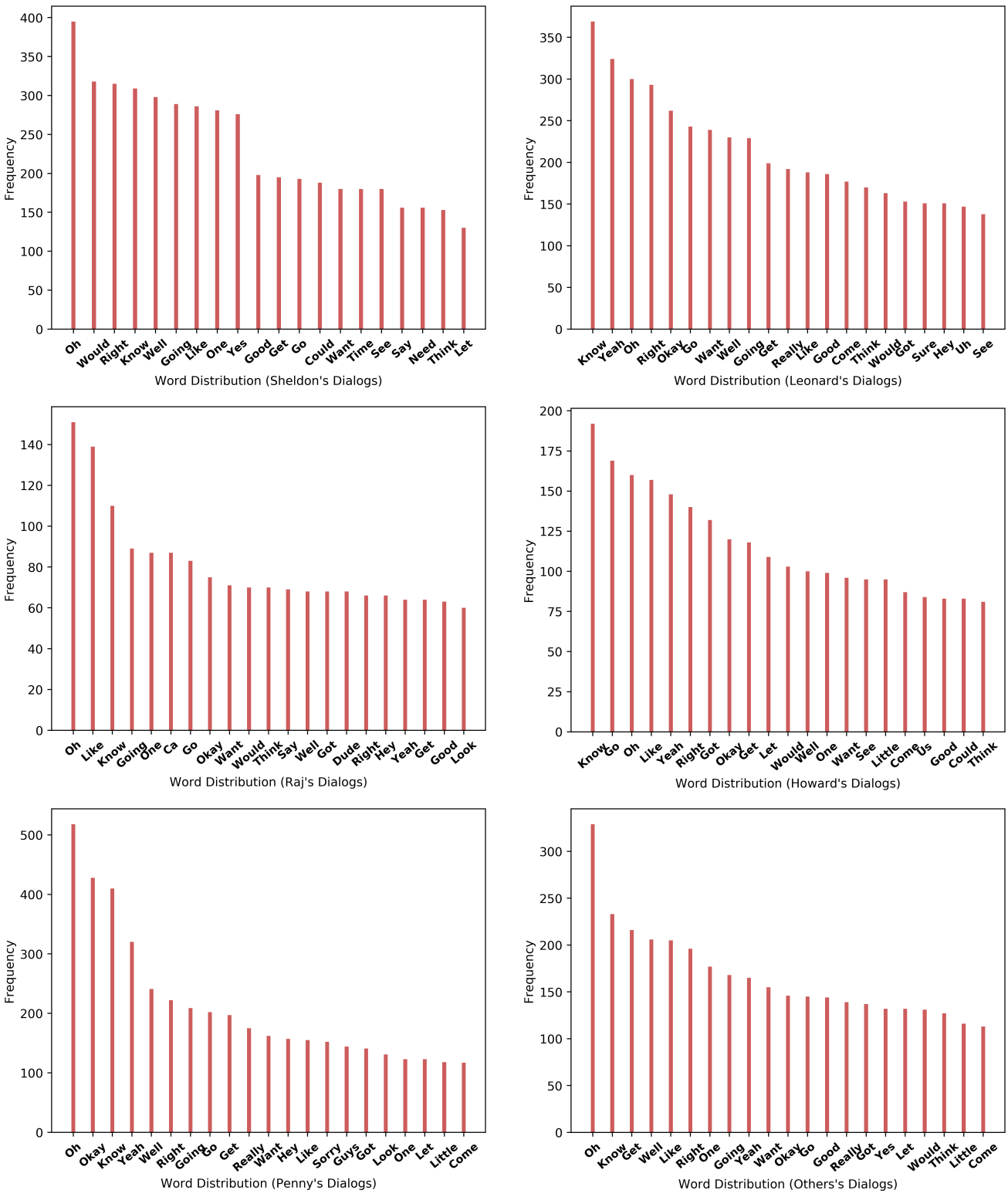


Figure 5: Bar plots drawn for the word distribution of dialogs spoken by Top 6 Speakers in our dataset.

Figure 6: A tSNE plot made by randomly selecting 1500 images (each from Humorous and Non-Humorous set) as the last frame of the visual dialogue turns. Sometimes these visual models could cheat by detecting some pattern in humorous/non-humorous visual dialogs like specific camera angle etc. The above plot hints towards its absence. To visualize the plot better, each image is represented by a dot and the corresponding plot is shown below. (Current plot is slightly scaled up to ease the visibility.)

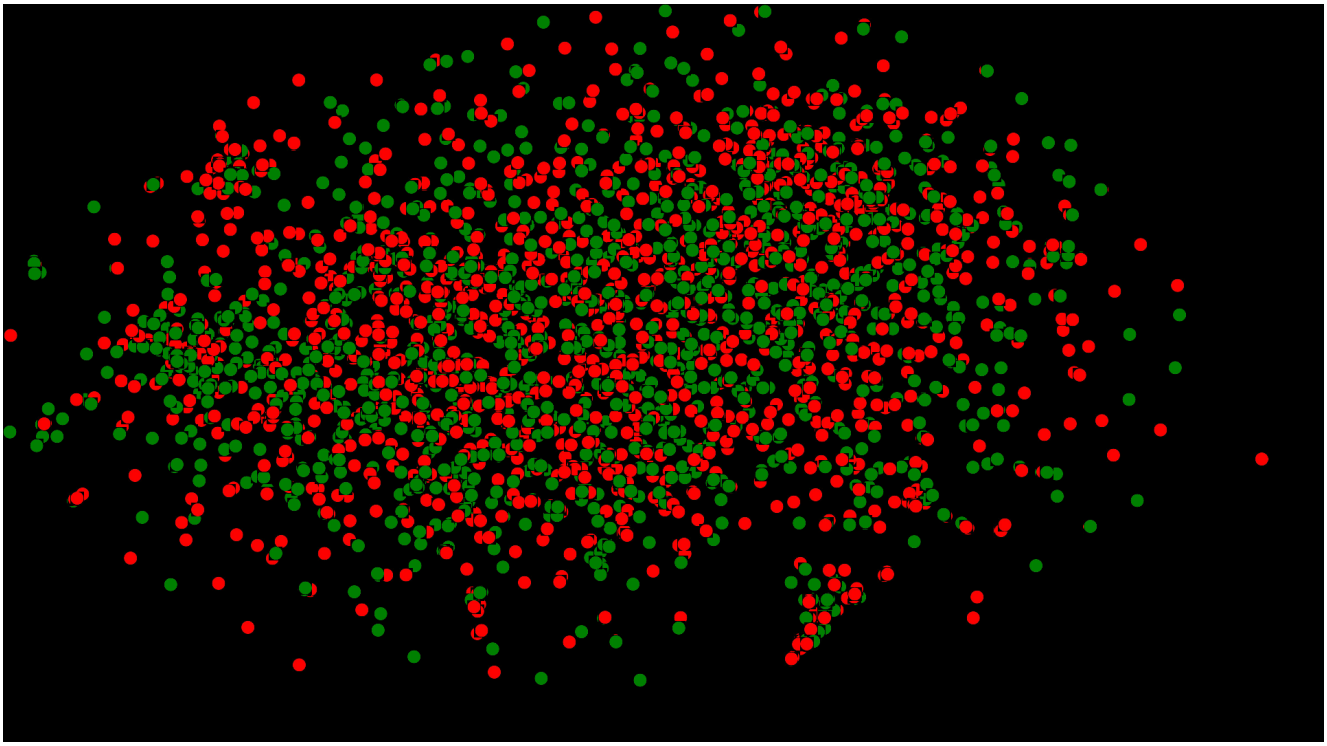
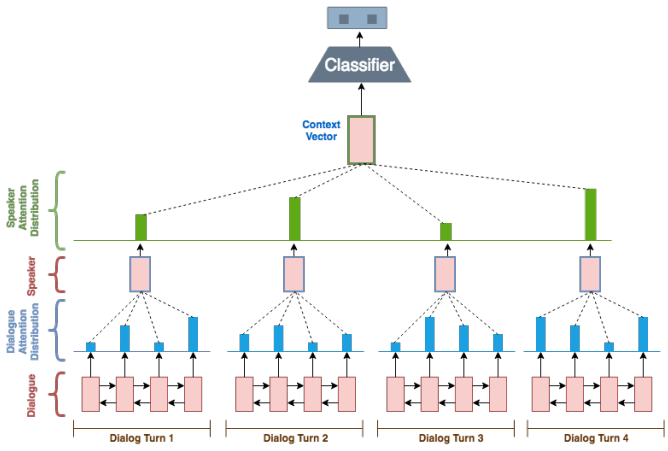


Figure 7: A green dot represents a humorous sample and red dot, a non-humorous sample. They seem to be randomly distributed, hinting towards absence of any such bias.

(a) TAM



(b) VAM

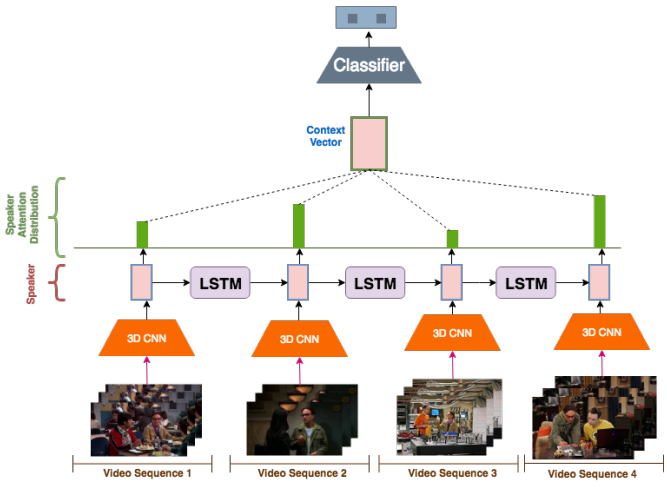


Figure 8: (a) Text based Attention Model (b) Video based Attention Model

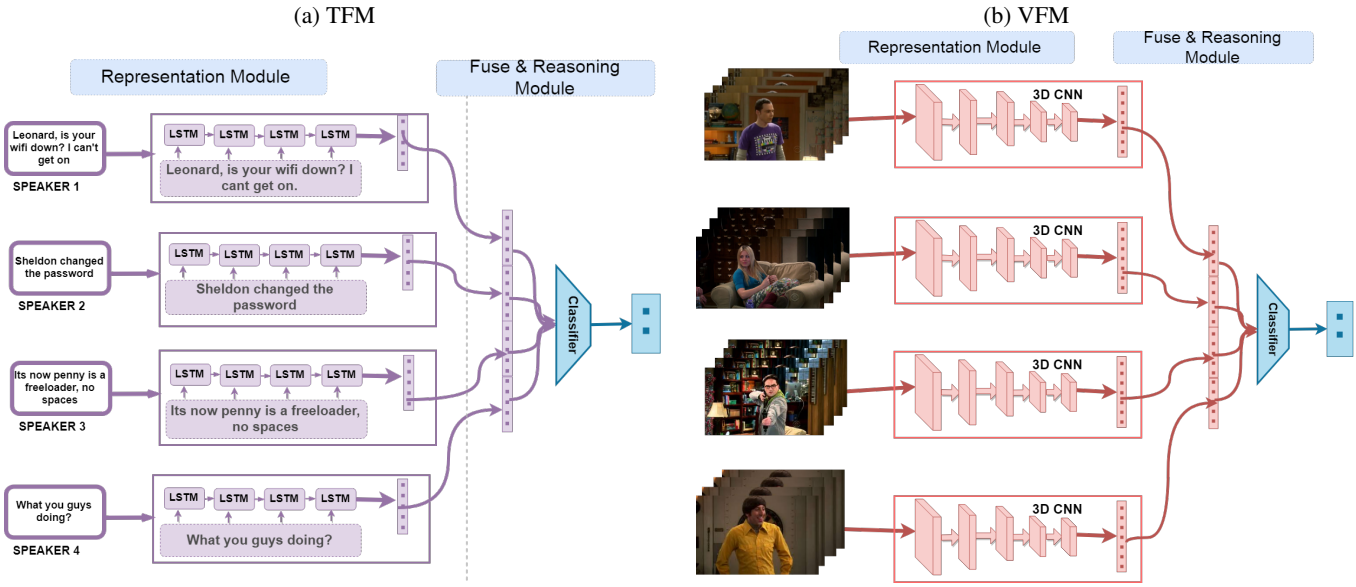


Figure 9: (a) Text based Fusion Model (b) Video based Fusion Model

References

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[2] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015.

[3] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017.