A Sentiment and Emotion aware Multimodal Multiparty Humor Recognition in Multilingual Conversational Setting

Dushyant Singh Chauhan[1][*]; Gopendra Vikram Singh[1], Aseem Arora[1], Asif Ekbal[1], and Pushpak Bhattacharyya[2]

Department of Computer Science and Engineering
[1] Indian Institute of Technology Patna, India
{1821cs17,gopendra_1921cs15,aseem_1911mc02,asif}@iitp.ac.in
[2] Indian Institute of Technology Bombay, India
pb@cse.iitb.ac.in

Abstract

In this paper, we hypothesize that humor is closely related to sentiment and emotions. Also, due to the tremendous growth in multilingual content, there is a great demand for building models and systems that support multilingual information access. To this end, we first extend the recently released *Multimodal Multiparty Hindi Humor* (M2H2) dataset by adding parallel English utterances corresponding to Hindi utterances and then annotating each utterance with sentiment and emotion classes. We name it *Sentiment, Humor, and Emotion aware Multilingual Multimodal Multiparty Dataset* (SHEMuD). Therefore, we propose a multitask framework wherein the primary task is humor detection, and the auxiliary tasks are sentiment and emotion identification. We design a multitasking framework wherein we first propose a *Context Transformer* to capture the deep contextual relationships with the input utterances. We then propose a *Sentiment and Emotion aware Embedding* (SE-Embedding) to get the overall representation of a particular emotion and sentiment *w.r.t.* the specific humor situation. Experimental results on the *SHEMuD* show the efficacy of our approach and shows that multitask learning offers an improvement over the single-task framework for both monolingual (4.86 points ↑ in Hindi and 5.9 points ↑ in English in F1-score) and multilingual (5.17 points ↑ in F1-score) setting.

1 Introduction

Humor (Ritschel et al., 2019, 2020; Song et al., 2021; Chauhan et al., 2021) is an essential aspect of daily conversation, and people try to provoke humor in their talks. Warren et al. (2018) defined humor as *"the nature of experiences to induce laughter and provide amusement"*. Humor is a tool by which anyone can draw the audience's attention. Even in a formal conversation, humor may make a person look more attractive and thus may lead to a better conclusion.

Irrespective of its relation to intelligence, humor is inherently a challenging problem to understand. To understand humor, we also take some additional information into consideration i.e., sentiment (Ghosal et al., 2018; Chauhan et al., 2019, 2020a, 2022) and emotion (Russell and Barrett, 1999; Pagé Fortin and Chaib-draa, 2019; Chauhan et al., 2020b) to help humor detection in conversations. Sentiment and emotion are two aspects that help each other, which has already been shown in (Akhtar et al., 2019). We hypothesize that sentiment and emotion do not only help each other (Akhtar et al., 2019) but also help humor. For example, it is difficult to decide whether the following utterance "अरे बुढापे में मैं तुम्हें खुद ले के जाता। (Oh in old age I would have taken you myself)" is humorous or not. But, careful observation of the sentiment (positive) and emotion (happy) of the speaker helps us understand that the speaker is in a funny mood and trying to mock his wife.

Sometimes it is difficult for the global audience to understand any local language like Hindi, so this is where multilingual comes into the picture. Also, multilinguality provides freedom for the model to understand humor; e.g., if the model is unable to understand the one language's word (say Hindi) for a particular utterance, then other languages (say English) can give a better clue to understand humor. Thus, we propose a multitask framework in a multilingual setting wherein the primary task is humor detection, and the auxiliary tasks are sentiment and emotion identification.

The main contributions of our proposed research are as follows;

- We first extend recently released M2H2 dataset by adding parallel English utterances corresponding to Hindi utterances and then annotate each utterance with sentiment and emotion classes. We name it *Sentiment, Humor,*

[*] The first two authors contributed equally to this work and are jointly the first authors.

and Emotion aware Multilingual Multimodal Multiparty Dataset (SHEMuD).

- We propose a *Context Transformer* to capture the deep contextual relationship with input utterance.
- We also propose a *Sentiment and Emotion Embedding* (SE-Embedding) to obtain the overall representation of a particular emotion and sentiment *w.r.t.* the specific humor situation.
- We present new state-of-the-art systems for Humor detection on *SHEMuD*.

2 Related Work

We have already discussed that sometimes just going through the utterance text is not enough to understand humor. The other modalities, such as visual and acoustic, can provide additional cues. In natural language processing, understanding humor from these modalities comes under the boundaries of multimodal language. Humor may be found in almost every human-to-human encounter. Some works (Hasan et al., 2019; Fallianda et al., 2018; Ritschel et al., 2019, 2020; Mirnig et al., 2017; Piata, 2020; Song et al., 2021; Vasquez and Aslan, 2021; Sabur et al., 2020; Veronika, 2020; Yang et al., 2019) on multimodal humor have already been done.

Hasan et al. (2019) proposed the UR-FUNNY dataset to aid in the comprehension of multimodal language utilized in the expression of humor. The author has also shown the importance of multimodality over unimodality. Humor can be created with any modality. Fallianda et al. (2018) looked at how humor was created using only verbal media, both verbal and visual media, and exclusively visual media. The data was gathered from 74 political comic strips published in Kompas newspaper. The author used the General Theory of Verbal Humor (GTVH) for humor analysis.

Higher pitch or loudness, a faster speaking cadence, or considerable pauses are not characteristics of conversational humor. In the paper, Attardo et al. (2011) discovered that when people are laughing or smiling, they are more likely to be humorous. El Khatib (2020) did two experiments: the first experiment looks at whether untrained people can recognize the structural aspects of humor in a multimodal text. The second research investigates the sequence in which textual and visual inputs are processed (tweets). In another work, Ritschel et al. (2019) discussed the quality of a conversation and how conversation partners see one other and are influenced by irony and irony-related humor.

The Humor Styles Questionnaire (HSQ) distinguishes four types of humor that can be good or damaging to one's mental health. Schneider et al. (2018) studied to compile research that used the HSQ to analyze the relationships between distinct humor types and four distinct aspects of mental health (self-esteem, life satisfaction, optimism, depression). Recently, Chauhan et al. (2021) proposed Multimodal Multiparty Hindi Humor (M2H2) dataset for conversations which was the very first of its kind. Then, the authors employed two strong baseline setups, viz. MISA[1] w/ DialogueRNN and MISA w/ bcLSTM. They used MISA for fusion and DialogueRNN & bcLSTM for understanding the contextual relationship among words.

In comparison to the existing systems, we first created the *SHEMuD* by manually annotating each utterance in the English language and then labeling each utterance with sentiment and emotion classes. After that, we design a multitasking framework wherein we first propose a *Context Transformer* to capture the deep contextual relationships with the input utterances. We then propose a *Sentiment and Emotion aware Embedding* (SE-Embedding) to get the overall representation of a particular emotion and sentiment *w.r.t.* the specific humor situation. Further, to the best of our knowledge, this is the first attempt at solving the sentiment and emotion aware humor detection in a multilingual conversational setting. Empirical results on the *SHEMuD* dataset demonstrate that the baselines yield better performance over the state-of-the-art systems.

3 Datasets

Chauhan et al. (2021) proposed the *Multimodal Multiparty Hindi Humor* (M2H2) dataset which contains 6,191 utterances spoken by 43 speakers from 13 episodes of a very popular TV series *"Shrimaan Shrimati Phir Se"* where each episode is divided into multiple scenes, and each scene is divided into multiple utterances. Each utterance is annotated with humor/non-humor labels and encompasses acoustic, visual, and textual modalities.

We extend the M2H2 dataset (Chauhan et al.,

[1] MISA:Modality-Invariant and-Specific Representations for Multimodal Sentiment Analysis

	Hindi Utterances	English Utterances	Humor	Sentiment	Emotion
1	अरे बुढापे में मैं तुम्हें खुद ले के जाता।	Oh in old age I would have taken you myself.	humor	positive	happy
2	ये लोग मुझे बहुत मार रहे हैं।	These guys are beating me hard	humor	positive	sad
3	ये दिलरुबा कहता है ना कि पिछले जन्म में वो छोटूमल मोटूमल करोडपति का एकलौता बेटा था, राइट?	This Dilruba says that he was the only son of Chhotumal Motumal Crorepati in his previous life, right?	non-humor	neutral	neutral
4	अरे भाई, मेरी असली वाली गन यहीं छूट गैइ।	Hey brother, my original gun is left here	non-humor	neutral	neutral
5	स्टाफ वालों! मेरा मतलब	Staff guys! I mean	humor	negative	anger

Table 1: Some samples from annotated dataset

2021) by manually annotating each Hindi utterance with the English language, making it a multilingual dataset (*SHEMuD*), and then annotating each utterance with sentiment and emotion classes. For sentiment annotation, we consider three sentiment classes, namely *positive, negative* and *neutral*. While for emotion annotation, we annotate the dataset with seven emotion values[2], *viz.* anger, disgust, fear, happy, sad, surprised, and neutral (c.f. Table 1). We show statistics of *SHEMuD* in Table 2.

Statistics	Train	Dev	Test
#Utterances	5000	500	691
#Humor (H)	1830	155	104
#Non-humor (NH)	3170	345	587
#Positive (Ps)	1048	102	197
#Neutral (Nu)	2488	245	307
#Negative (Ng)	1464	153	187
#Anger (Ag)	811	105	130
#Disgust (Dg)	65	19	22
#Fear (Fr)	93	30	40
#Happy (Ha)	648	65	102
#Sad (Sd)	300	55	60
#Surprise (Sp)	306	81	119
#Neutral (Nu)	2777	145	218

Table 2: Dataset statistics for *SHEMuD*

Please note that humor, many times, is related to the language as well as the culture. So, we cannot guarantee that if a Hindi utterance is humorous, then English will be humorous, but sharing information across the languages helps each other in humor prediction.

3.1 Annotation Guidelines

We employ four Ph.D. students highly proficient in the Hindi and English languages with prior experience in labeling *sentiment and emotion*. The guidelines for annotation were explained to the annotators before starting the annotation process (c.f. Table 1). The annotators were given data without humor labels and asked first to add English utterances corresponding to Hindi utterances and then annotate every utterance with one emotion and the sentiment by seeing the context utterances before annotation. The majority voting scheme was used for selecting the final emotion and sentiment. We achieve an overall Fleiss' (Fleiss, 1971) kappa score of 0.83, which is considered to be reliable.

3.2 Feature Extraction

For monolingual Hindi and English textual features, we take the pre-trained Hindi and English embedding using `XLM-Roberta`[3] (Conneau et al., 2019) (XLM-R). While, for multilinguality, we first train XLM-R on both English and Hindi (shared embedding) and then we test the result on both Hindi and English. Thus, we extract multilingual embedding for Hindi and English.

Similarly, for visual feature extraction, we use `3D-ResNeXt-101`[4] (Xie et al., 2017) which is pre-trained on Kinetics at a rate of 1.5 features per second and a resolution of 112. We use `openSMILE`[5] (Eyben et al., 2010) for acoustic feature extraction. It can extract Low-Level Descriptors (LLD) and change them using different filters, functions, and transformations. We use a tonal low-level features group of openSMILE to extract the features.

4 Proposed Methodology

This section describes our proposed model, where we aim to leverage multimodal sentiment and emotion information for solving multimodal humor detection in a multitask framework. We depict the overall architecture in Figure 1. The extended dataset and source code are available at `https://www.iitp.ac.in/~ai-nlp-ml/resources.html#SHEMuD`

[2] One emotion per utterance

[3] `https://huggingface.co/xlm-roberta-base`

[4] `https://github.com/kaiqiangh/extracting-video-features-ResNeXt`

[5] `https://github.com/audeering/opensmile`

4.1 Context Transformer

We propose a Context Transformer (Con_{trans}) to capture the deep contextual relationship with input utterance. As we know, contextual utterances might give essential information when identifying an input utterance. This necessitates a model that accounts for such interdependencies and the impact they may have on the target utterance. We use a Transformer based approach to capture the flow of informative triggers across the utterances.

Let us assume, the unimodal features have dimension p and each utterance is thus represented by a feature vector $d_{l,s} \in \mathcal{R}^p$, where s represents the s^{th} utterance in a conversation l. For a conversation, we collect the feature vectors of the utterances in D_L, which is as follows;

$$D_L = [d_{l,1}, d_{l,2}, \ldots, d_{l,J_l}] \in \mathcal{R}^{J_l \times p} \quad (1)$$

where J_l represents the number of utterances in a conversation. The matrix D_L serves as input to the Transformer. To learn the representation of D_L, we first map it into the continuous space U_c;

$$\begin{aligned} U_c &= u^i_1, u^i_2, \ldots, u^i_{|D_L|} \\ u^i_j &= u(d_{l,1}) + p_j \end{aligned} \quad (2)$$

where $u(d_{l,1})$ & p_j are the utterance vectors and positional embedding of every utterance, respectively. We adopt sine-cosine positional embedding (Vaswani et al., 2017) as it performs better and does not introduce additional trainable parameters.

The contextual encoder converts U_c into a list of hidden representations $h^i_1, h^i_2, \ldots, h^i_{|U_i|}$. We use the last hidden representation $h^i_{|U_i|}$ (i.e., the representation at the EOS token) as the contextual representation of the utterance. Therefore, the final contextual representation of the D_L;

$$D_L = h_i = h^i_{|D_L|} + p_i \quad (3)$$

Please note that words and sentences share the same positional embedding matrix.

For multimodal inputs, we simply concatenate the input embeddings and pass them to the Con_{trans} to capture the deep contextual relationship with the input utterance. We pass the output of Con_{trans} to *Linear Layer* then output of this is fed to two *Softmax* layers for sentiment and emotion prediction, respectively.

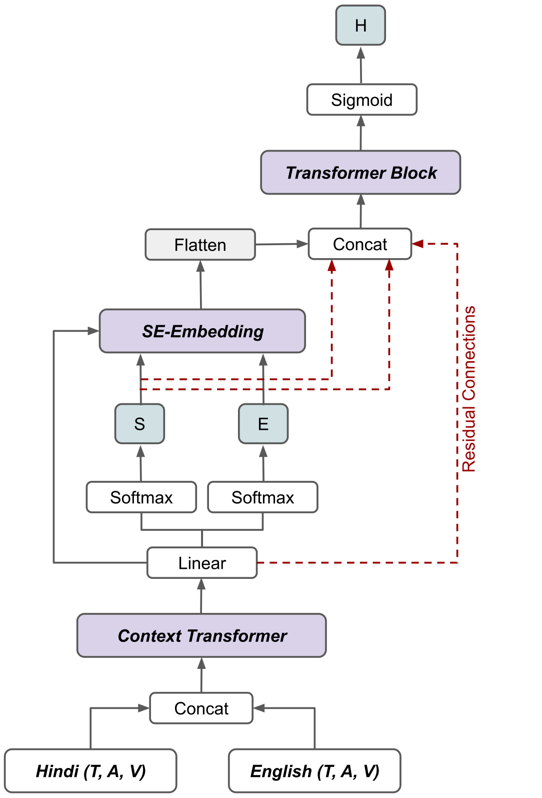


Figure 1: A contextual transformer based proposed framework for humor detection where *SE-Embedding* stands for Sentiment and Emotion Embedding.

$$\begin{aligned} cat_{TAV} &= Concat(T, A, V) \\ Con_{rel} &= Con_{trans}(cat_{TAV}) \\ L_{vec} &= Linear(Con_{rel}) \\ S_p &= Softmax(L_{vec}) \\ E_p &= Softmax(L_{vec}) \end{aligned} \quad (4)$$

We use a weighted loss function (Kendall et al., 2018) to teach the model how to weigh the task-specific losses. We define the loss as L_{total};

$$L_{total} = \sum_i w_i L_i \quad (5)$$

where *i* defines the different tasks i.e., Sentiment and Emotion. Initially we define w with **xavier-norm** for each task.

4.2 Sentiment and Emotion Embedding

We then propose SE-Embedding (SE_{emb}) to obtain the overall representation of a particular emotion and sentiment *w.r.t.* the specific humor situation. We take output of Linear Layer ($vec \in \mathcal{R}^k$), sentiment prediction ($S_p \in \mathcal{R}^3$), and emotion prediction ($E_p \in \mathcal{R}^7$) as input.

$$O_{rep} = SE_{emb}(L_{vec}, S_p, E_p) \quad (6)$$

We then concatenate S_p & E_p and obtain $SE_p \in \mathcal{R}^{10}$. After that we multiply SE_p & L_{vec} and get sentiment-emotion aware matrix ($SE_{mat} \in \mathcal{R}^{10\times k}$).

$$SE_{mat} = Concat(S_p, E_p) \times L_{vec} \quad (7)$$

We initially define SE_{emb} with random weights and later we apply convolve function (*) between SE_{emb} & SE_{mat} and update the weights of SE_{emb}.

$$SE_{emb} = SE_{emb} * SE_{mat} \quad (8)$$

We use the convolve function to obtain the discrete, linear convolution of SE_{emb} & SE_{mat}. Let us say there are two one-dimensional input arrays f & g. So, we compute convolve function between f & g;

$$(f * g)[n] = \sum_{n=-\infty}^{+\infty} f[m]g[n-m] \quad (9)$$

Let us suppose the model predicts the positive sentiment and happy emotion, then only corresponding vectors from SE_{emb} will go forward because non-predicted (n_p) vectors from SE_{emb} may confuse the model. To achieve this, we convert all the non-predicted vectors to -1 by first multiplying with 0 and then adding -1.

$$SE_{emb} = \forall n_p \ (SE_{emb} \times 0 - 1) \quad (10)$$

Then, we flatten the SE_{emb}

$$F_{out} = Flatten(SE_{emb}) \quad (11)$$

Motivated by the residual skip connection network (He et al., 2016), we concatenate the output of *Linear Layer* (L_{vec}), S_p, and E_p with the flatten output and pass to the Transformer Block ($Trans_b$) to learn the relationship between humor and sentiment & emotion.

$$\begin{aligned} cat_{out} &= Concat(F_{out}, L_{vec}, S_p, E_p) \\ trans_{out} &= Trans_b(cat_{out}) \end{aligned} \quad (12)$$

At last, we pass the output of *Transformer Block* to the *Sigmoid* layer for humor detection.

$$H_p = Sigmoid(trans_{out}) \quad (13)$$

We use *binary cross-entropy* loss for humor detection. Please note that two losses are backpropagated, one from humor and another from sentiment & emotion.

5 Experimental Setup

We evaluate our proposed model on the *SHEMuD*. We perform a grid search to find the optimal hyperparameters, which are shown in Table 3. We take five[6] utterances as context for a particular input utterance. We implement our proposed model on the Python-based PyTorch deep learning library. As the evaluation metric, we employ precision (P), recall (R), and F1-score (F1) for sentiment, emotion, and humor recognition.

Parameters	**Proposed**
Transformer Encoder	2 Layers (Context Transformer, Transformer Block)
Dense layer	1024N, D=0.3
Activations	ReLu
Optimizer	Adam (*lr=1e-3*)
Outputs	*Softmax* (Sent and Emo) *Sigmoid* (Humor)
Batch	16
Epochs	60

Table 3: Model configurations

6 Results and Analysis

We divide our experiments into two parts, i.e., monolingual and multilingual;

6.1 Monolingual (Multitask vs Single-task)

In monolingual experiments, we evaluate our proposed model for Hindi and English languages separately. We perform experiments for unitask (Humor), bitask (Humor+Sentiment and Humor+Emotion), and tritask (Humor+Sentiment+Emotion). We show experimental results in Table 4. For Hindi tritask (H+S+E), our proposed model shows an improvement of 4.53 points in precision, 4.1 points in recall, and 3.98 points in F1-score over bitask[7] while 4.55 points in precision, 4.41 points in recall, and 4.86 points in F1-score over unitask.

We get a similar improvement for English tritask (H+S+E). For English tritask (H+S+E), our proposed model shows an improvement of 3.63 points in precision, 4.46 points in recall, and 4.2 points in F1-score over bitask[8] while 5.63 points in preci-

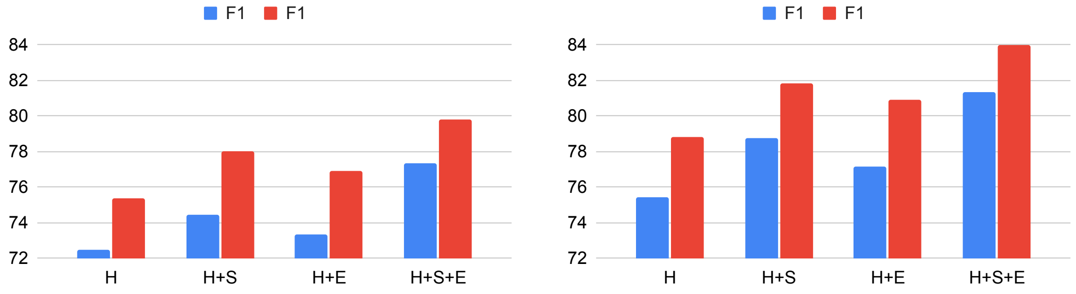
[6] For fair comparison with M2H2 (2021) as they took five utterances as context.

[7] We get a maximum improvement over H+E

[8] We get a maximum improvement for H+E

	Monolingual						*Multilingual*					
	Hindi			*English*			*Hindi*			*English*		
Labels	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
H	71.82	73.91	72.47	74.89	76.64	75.42	74.84	75.89	75.36	78.32	79.25	78.78
H+S	73.42	74.91	74.46	77.63	79.32	78.74	77.11	78.94	78.01	81.22	82.43	81.79
H+E	72.14	74.22	73.35	76.89	77.68	77.12	76.34	77.49	76.91	80.49	81.37	80.92
H+S+E	**76.37**	**78.32**	**77.33**	**80.52**	**82.14**	**81.32**	**78.69**	**80.94**	**79.79**	**83.15**	**84.78**	**83.95**

(a) Experiment results of our proposed model for monolingual and multilingual setting



(b) Monolingual (Hindi) vs Multilingual (Hindi)

(c) Monolingual (English) vs Multilingual (English)

Table 4: Experiment results and bar-chart representations of our proposed model for monolingual and multilingual setting, where H, S, and E represent the humor, sentiment, and emotion, respectively.

	Monolingual						*Multilingual*					
	Hindi			*English*			*Hindi*			*English*		
Labels	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
T	69.61	72.89	71.21	74.64	76.32	75.47	72.54	73.26	72.89	76.87	78.89	77.86
A	59.71	64.89	62.19	59.71	64.89	62.19	59.71	64.89	62.19	59.71	64.89	62.19
V	58.84	60.57	59.61	58.84	60.57	59.61	58.84	60.57	59.61	58.84	60.57	59.61
T+V	73.77	75.13	74.41	77.71	78.94	78.94	75.97	76.21	75.58	78.67	80.11	79.38
T+A	73.61	76.89	75.21	75.72	77.63	76.65	75.54	76.89	76.20	81.11	82.31	81.70
A+V	70.14	72.36	70.24	70.14	72.36	70.24	70.14	72.36	70.24	70.14	72.36	70.24
T+V+A	**76.37**	**78.32**	**77.33**	**80.52**	**82.14**	**81.32**	**78.69**	**80.94**	**79.79**	**83.15**	**84.78**	**83.95**

Table 5: Modality-wise results of our proposed tritask model (H+S+E) for monolingual and multilingual setting, where T, A, and V represent the text, acoustic, and visual, respectively.

sion, 5.5 points in recall, and 5.9 points in F1-score over unitask.

6.2 Multilingual (Multitask vs. Single-task)

In multilingual experiments, we evaluate our proposed model in both languages (i.e., Hindi and English). We perform similar experiments as monolingual setup, i.e., unitask (Humor), bitask (Humor+Sentiment and Humor+Emotion), and tritask (Humor+Sentiment+Emotion). We show experimental results in Table 4. For Hindi tritask (H+S+E), our proposed model shows an improvement of 2.35 points in precision, 3.45 points in recall, and 2.88 points in F1-score over bitask[9] while 3.85 points in precision, 5.05 points in recall, and 4.43 points in F1-score over unitask.

We get a similar improvement for English tritask (H+S+E). For English tritask (H+S+E), our proposed model shows an improvement of 2.66 points in precision, 3.41 points in recall, and 3.03 points in F1-score over bitask[10] while 4.83 points in precision, 5.53 points in recall, and 5.17 points in F1-score over unitask.

For both monolingual and multilingual setups, we observe that tritask (H+S+E) performance is better than unitask and bitask for Hindi and English. Thus, we can say that sentiment and emotion help to understand humor. We also show the modality-wise results of our proposed tritask model (H+S+E) for monolingual and multilingual setting in Table 5.

7 Comparative Analysis

We compare the results obtained from our proposed model against the existing model M2H2 (Chauhan et al., 2021). We evaluate our proposed framework with all the possible combinations i.e., unitask *(H)*,

[9] We get a maximum improvement over H+E

[10] We get a maximum improvement over H+E

	Monolingual											
	Hindi						*English*					
	M2H2(2021)			*Proposed*			*M2H2(2021)*			*Proposed*		
Labels	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
H	71.21	72.11	71.67	**71.82**	**73.91**	**72.47**	72.33	74.42	73.35	**74.89**	**76.64**	**75.42**
H+S	73.31	74.21	73.92	**73.42**	**74.91**	**74.46**	77.51	77.79	77.64	**77.63**	**79.32**	**78.74**
H+E	71.11	72.91	72.41	**72.14**	**74.22**	**73.35**	75.13	75.94	75.73	**76.89**	**77.68**	**77.12**
H+S+E	74.71	75.91	75.31	**76.37**	**78.32**	**77.33**	78.71	79.81	79.25	**80.52**	**82.14**	**81.32**

Table 6: Comparative analysis: comparison between M2H2 (2021) and our proposed model in monolingual setting

	Multilingual											
	Hindi						*English*					
	M2H2(2021)			*Proposed*			*M2H2(2021)*			*Proposed*		
Labels	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
H	73.47	74.39	73.92	**74.84**	**75.89**	**75.36**	75.21	76.43	75.81	**78.32**	**79.25**	**78.78**
H+S	75.37	76.74	76.04	**77.11**	**78.94**	**78.01**	79.76	80.21	79.98	**81.22**	**82.43**	**81.79**
H+E	74.55	76.21	75.37	**76.34**	**77.49**	**76.91**	77.86	79.48	78.66	**80.49**	**81.37**	**80.92**
H+S+E	76.49	78.61	77.67	**78.69**	**80.94**	**79.79**	81.28	82.49	81.88	**83.15**	**84.78**	**83.95**

Table 7: Comparative analysis: comparison between M2H2 (2021) and our proposed model in multilingual setting

bitask (*H+S and H+E*) and tritask (*H+S+E*).

Please note that as the presented model in M2H2 (2021) was the only unitask (H) model without multilingual, so we made some changes in the model to make it a multitask model and multilingual. Thereafter we obtain the results for unitask (H), bitask (H+S and H+E) and tritask (H+S+E) for both monolingual and multilingual which are also shown in Table 6 and Table 7. We divide this into two parts; i) monolingual and ii) multilingual;

7.1 Monolingual SOTA vs. Proposed

For Hindi, our the proposed tritask model achieves the best precision of 76.37% (1.66 points ⇡), recall of 78.32% (2.41 points ⇡) and F1-score of 77.33% (2.02 points ⇡) compared to precision of 74.71%, recall of 75.91%, and F1-score of 75.31% of the M2H2 (2021). We observe that our proposed model performs better than the state-of-the-art system, M2H2 (2021). We show the results in Table 6.

Similarly, for English, our the proposed tritask model achieves the best precision of 80.52% (1.81 points ⇡), recall of 82.14% (2.33 points ⇡) and F1-score of 81.32% (2.07 points ⇡) compared to precision of 78.71%, recall of 79.81%, and F1-score of 79.25% of the M2H2 (2021). We observe that our proposed model performs better than the state-of-the-art system, M2H2 (2021).

7.2 Multilingual SOTA vs. Proposed

Similar to monolingual, we observe a similar trend of improvement for multilingual experiments. We show the multilingual experiment results in Table 7. For Hindi, our the proposed tritask model achieves the best precision of 78.69% (2.2 points ⇡), recall of 80.94% (2.33 points ⇡) and F1-score of 79.79% (2.12 points ⇡) compared to precision of 76.49%, recall of 78.61%, and F1-score of 77.67% of the M2H2 (2021). We observe that our proposed model performs better than the state-of-the-art system, M2H2 (2021).

Similarly, for English, our the proposed tritask model achieves the best precision of 83.15% (1.87 points ⇡), recall of 84.78% (2.29 points ⇡) and F1-score of 83.95% (2.07 points ⇡) compared to precision of 81.28%, recall of 82.49%, and F1-score of 81.88% of the M2H2 (2021). We observe that our proposed model performs better than the state-of-the-art system, M2H2 (2021).

8 Ablation Study

We perform an ablation study to show the efficacy of *SE-Embedding* for both monolingual and multilingual. For monolingual, we perform experiments with and without *SE-Embedding*. We show the ablation experimental results in Table 8. As per the result, we can see when *SE-Embedding* is used with models then it shows significant improvement rather than without *SE-Embedding*.

Similarly, for multilingual, we perform experiments with and without *SE-Embedding*. As per the result (c.f. Table 8), we can see when *SE-Embedding* is used with models then it shows significant improvement rather than without *SE-Embedding*.

	Monolingual						*Multilingual*					
	Hindi			*English*			*Hindi*			*English*		
Setup	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
w/o SE-Embedding	74.23	77.89	76.01	78.85	80.31	79.42	76.90	78.74	77.80	81.31	82.44	81.87
w/ SE-Embedding	**76.37**	**78.32**	**77.33**	**80.52**	**82.14**	**81.32**	**78.69**	**80.94**	**79.79**	**83.15**	**84.78**	**83.95**

Table 8: Ablation study: comparison between proposed tritask model (H+S+E) w/ *SE-Embedding* and w/o *SE-Embedding*

	Hindi Utterances	**English Utterances**	**True Label**	**M2H2 (2021) (H+S+E) (Multilingual)**	**Proposed (H+S+E) (Multilingual)**
1	ये लोग मुझे बहुत मार रहे हैं।	These guys are beating me hard	H	NH	NH,Ng,Sd
2	हाय, टोटो!	hey, Toto!	NH	H	NH,Nu,Nu
3	ये पिंकी लाई।	Pinky brought this	H	NH	H,Nu,Nu
4	अरे भाई, मेरी असली वाली गन यहीं छूट गैइ।	Hey brother, my original gun is left here	NH	H	NH,Nu,Nu
5	कौन सी असली है कौन सी नकली है।	which is real which is fake	NH	H	NH,Nu,Ha

Table 9: Error analysis: some correct and incorrect predicted samples by our proposed tritask model (H+S+E) in multilingual setting.

9 Error Analysis

We perform a detailed analysis to realize the model's strengths and weaknesses. To better analyze and justify our proposed model in terms of quality, we use the same samples (for multilingual setup), which were wrongly predicted and depicted in the paper M2H2 (2021). We show these samples in Table 9. The main motivation for taking the same samples for error analysis is to show the performance of our proposed model over state-of-the-art system. We see that our proposed model predicts all the utterances correctly except the first utterance.

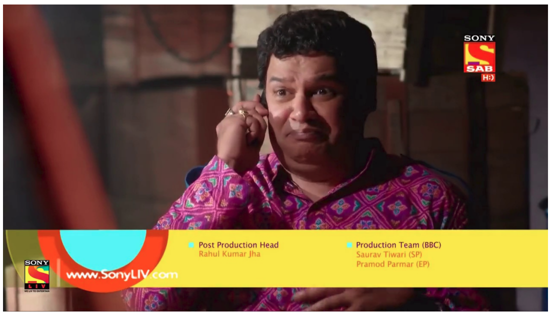


Figure 2: The visual frame of Dilruba for saying "ये लोग मुझे बहुत मार रहे हैं। (These guys are beating me hard)", which shows the sad visual expression

We observe that for the first utterance "ये लोग मुझे बहुत मार रहे हैं। (These guys are beating me hard)", our proposed model predict negative sentiment and sad emotion because of the word "beating" and Dilruba's facial expression (c.f. Figure 2) and sad tone. Also, the context utterances of that utterance are negative. So, this is the reason that our proposed model wrongly predict this utterance.

10 Conclusion

In this paper, we have proposed an effective deep learning-based multitask model for humor detection (primary task) with auxiliary tasks, i.e., sentiment and emotion, in a multilingual conversational setting. As there was no suitable data available for this problem, we first extend recently released M2H2 dataset by adding parallel English utterances corresponding to Hindi utterances and made it a multilingual dataset. Then, we annotate each utterance with sentiment and emotion classes. We have proposed a multitasking framework wherein we propose a *Context Transformer* to capture the deep contextual relationships with the input utterances. We have also successfully established a *Sentiment and Emotion Embedding* (SE-Embedding) which gets the overall representation of a particular emotion and sentiment *w.r.t.* the specific humor situation. Experimental results on the *SHEMuD* have shown that the multitask system yields better performance over the single-task framework.

In the future, we would like to extend our work towards the multiparty dialogue generation in Hindi with the help of humor, sentiment, emotion and speaker information.

11 Ethical Declaration

The *SHEMuD* dataset is freely available and will be used only for the purpose of academic research. We create *SHEMuD* by extending the M2H2 dataset with English utterances and labeling each utterance with sentiment and emotion classes. The annotation for extending the dataset was done by human experts, who are the regular employee of our research group. There are no other issues to declare.

Acknowledgement

Dushyant Singh Chauhan acknowledges the support of the Prime Minister Research Fellowship (PMRF), Govt. of India. Asif Ekbal acknowledges the Young Faculty Research Fellowship (YFRF), supported by Visvesvaraya Ph.D. Scheme for Electronics and IT, Ministry of Electronics and Information Technology (Meit/8Y), Government of India, being implemented by Digital India Corporation (formerly Media Lab Asia).

References

Md Shad Akhtar, Dushyant Singh Chauhan, Deepanway Ghosal, Soujanya Poria, Asif Ekbal, and Pushpak Bhattacharyya. 2019. Multi-task learning for multi-modal emotion recognition and sentiment analysis. *arXiv preprint arXiv:1905.05812*.

Salvatore Attardo, Lucy Pickering, and Amanda Baker. 2011. Prosodic and multimodal markers of humor in conversation. *Pragmatics & Cognition*, 19(2):224–247.

Dushyant Singh Chauhan, Md Shad Akhtar, Asif Ekbal, and Pushpak Bhattacharyya. 2019. Context-aware interactive attention for multi-modal sentiment and emotion analysis. 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 5651–5661, Hong Kong, China. Association for Computational Linguistics.

Dushyant Singh Chauhan, Asif Ekbal, and Pushpak Bhattacharyya. 2022. An efficient fusion mechanism for multimodal low-resource setting. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, page 2583–2588, New York, NY, USA. Association for Computing Machinery.

Dushyant Singh Chauhan, Dhanush S R, Asif Ekbal, and Pushpak Bhattacharyya. 2020a. All-in-one: A deep attentive multi-task learning framework for humour, sarcasm, offensive, motivation, and sentiment on memes. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 281–290, Suzhou, China. Association for Computational Linguistics.

Dushyant Singh Chauhan, Dhanush S R, Asif Ekbal, and Pushpak Bhattacharyya. 2020b. Sentiment and emotion help sarcasm? a multi-task learning framework for multi-modal sarcasm, sentiment and emotion analysis. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4351–4360, Online. Association for Computational Linguistics.

Dushyant Singh Chauhan, Gopendra Vikram Singh, Navonil Majumder, Amir Zadeh, Asif Ekbal, Pushpak Bhattacharyya, Louis-philippe Morency, and Soujanya Poria. 2021. M2h2: A multimodal multiparty hindi dataset for humor recognition in conversations. In *Proceedings of the 2021 International Conference on Multimodal Interaction*, pages 773–777.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Nabiha El Khatib. 2020. *Multimodal Processing of Humorous Tweets*. Ph.D. thesis, Texas A&M University-Commerce.

Florian Eyben, Martin Wöllmer, and Björn Schuller. 2010. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 1459–1462. ACM.

Fallianda Fallianda, Rani Yuni Astiti, and Zulvy Alivia Hanim. 2018. Analyzing humor in newspaper comic strips using verbal-visual analysis. *Lingua Cultura*, 12(4):383–388.

Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many rater. *Psychological Bulletin*, 76:378–382.

Deepanway Ghosal, Md Shad Akhtar, Dushyant Singh Chauhan, Soujanya Poria, Asif Ekbal, and Pushpak Bhattacharyya. 2018. Contextual inter-modal attention for multi-modal sentiment analysis. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3454–3466, Brussels, Belgium. Association for Computational Linguistics.

Md Kamrul Hasan, Wasifur Rahman, Amir Zadeh, Jianyuan Zhong, Md Iftekhar Tanveer, Louis-Philippe Morency, et al. 2019. Ur-funny: A multimodal language dataset for understanding humor. *arXiv preprint arXiv:1904.06618*.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Alex Kendall, Yarin Gal, and Roberto Cipolla. 2018. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7482–7491.

Nicole Mirnig, Gerald Stollnberger, Manuel Giuliani, and Manfred Tscheligi. 2017. Elements of humor: How humans perceive verbal and non-verbal aspects of humorous robot behavior. In *Proceedings of the Companion of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, pages 211–212.

Mathieu Pagé Fortin and Brahim Chaib-draa. 2019. Multimodal multitask emotion recognition using images, texts and tags. In *Proceedings of the ACM Workshop on Crossmodal Learning and Application*, pages 3–10. ACM.

Anna Piata. 2020. Stylistic humor across modalities: The case of classical art memes. *Internet Pragmatics*, 3(2):174–201.

Hannes Ritschel, Ilhan Aslan, David Sedlbauer, and Elisabeth André. 2019. Irony man: augmenting a social robot with the ability to use irony in multimodal communication with humans.

Hannes Ritschel, Thomas Kiderle, Klaus Weber, Florian Lingenfelser, Tobias Baur, and Elisabeth André. 2020. Multimodal joke generation and paralinguistic personalization for a socially-aware robot. In *International Conference on Practical Applications of Agents and Multi-Agent Systems*, pages 278–290. Springer.

James A Russell and Lisa Feldman Barrett. 1999. Core affect, prototypical emotional episodes, and other things called emotion: dissecting the elephant. *Journal of personality and social psychology*, 76(5):805.

Andy Jefferson Sabur, Retno Purwani Sari, and Tatan Tawami. 2020. Identification of the multimodal structure of humor in an animated superhero film. In *International Conference on Language, Linguistics, and Literature (COLALITE) 2020*.

Martha Schneider, Martin Voracek, and Ulrich S Tran. 2018. “a joke a day keeps the doctor away?” meta-analytical evidence of differential associations of habitual humor styles with mental health. *Scandinavian journal of psychology*, 59(3):289–300.

Kwangok Song, Kyle M Williams, Diane L Schallert, and Alina Adonyi Pruitt. 2021. Humor in multimodal language use: Students’ response to a dialogic, social-networking online assignment. *Linguistics and Education*, 63:100903.

Camilla Vasquez and Erhan Aslan. 2021. “cats be outside, how about meow”: multimodal humor and creativity in an internet meme. *Journal of Pragmatics*, 171:101–117.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Zhigailova Veronika. 2020. Multimodal discourse analysis of humor in picture books for children.

Caleb Warren, Adam Barsky, and A Peter McGraw. 2018. Humor, comedy, and consumer behavior. *Journal of Consumer Research*, 45(3):529–552.

Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. 2017. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500.

Zixiaofan Yang, Lin Ai, and Julia Hirschberg. 2019. Multimodal indicators of humor in videos. In *2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, pages 538–543. IEEE.