# Multimodal Multiparty Humour Recognition

**Shivangana Rawat**
Indian Institute of Technology
Hyderabad, Telangana, India
cs20mtech12001@iith.ac.in

**Pranoy Panda**
Indian Institute of Technology
Hyderabad, Telangana, India
cs20mtech12002@iith.ac.in

## Abstract

Recognising humor is a challenging sentiment analysis task as it requires a great of understanding of the context, and the information present in different modalities like video, audio and text. In this work we tackle the problem of multimodal humor recognition in a low resource language i.e. Hindi. We propose to extract representation from different modalities using transformer models(for text) and 3D CNN(for video), and strategically fuse them before classification. We also use past utterances information to understand the context of the current utterance which ultimately leads to better humor recognition. Through ablation studies of different components of our method, we justify the requirement of each of them. Our final model achieves an 8.6% improvement on F1 score and 1.6% improvement on precision w.r.t. our baseline method. We also open source our code in *this github repository*.

## 1 Introduction

Multimodal sentiment or emotion analysis is an important area of applied machine learning research which brings together Vision, NLP and Speech communities together. Understanding emotion is innate to humans but is an extremely challenging task for machines as it requires information extraction from multiple modalities such as vision, audio and language, and accumulating it over time. Cues for understanding emotion are distributed across modalities and therefore using all of the modalities is pertinent for designing a good sentiment analysis system.

Humor as a sentiment is one of the most difficult emotions to detect. This is because of its subject nature and strong dependence on the context. Humor like other emotions can be induced via different modalities, such as visually(for eg. Charlie Chaplin movies), vocally(for eg. Stand up comedy or radio shows) and a mixture of both(for eg. SitComs). Jokes delivered vocally are explicit, meaning they are self contained and could be understood by themselves, but comedy sitcoms often require understanding of the nature of the speaker and the context of the situation. Due to its wide range of variability across modalities and its strong dependence on context, humor recognition is a very challenging and interesting problem to tackle in the space of sentiment analysis and it is also the problem of interest for our work.

Humor recognition is particularly challenging in languages such as Hindi. Hindi a low-resource language with morphological richness which makes understanding Hindi text and audio challenging. In this work, we are operating in this challenging setting and focus on humor recognition in a dataset(Chauhan et al., 2021) build from the Hindi Sitcom "Shrimaan Shrimati Phir Se"(available on YouTube). The dataset is structured w.r.t to utterances. Utterance is a part of the dialogue where one party speaks without taking any breathes/pauses. As our dataset is multimodal therefore each utterance is represented by three components audio, video and text. Further details of the dataset are provided in 3.2.

In this work, we utilized a three stage pipeline as shown in Figure 1 for humor recognition. The first stage deals with extracting representations from the different modalities. The second stage deals with modality fusion to exploit relationship between the representation of different modalities. The third and final stage deals with context modelling and classification which ensures we use the past utterance information for recognising humor.

## 2 Related Work

There have been a few datasets (Hasan et al., 2019), (Patro et al., 2021), (Wu et al., 2021) on multimodal humor recognition for English language in the recent past. All of these datasets have video, audio and text for each utterance.

Recently, M2H2 dataset(Chauhan et al., 2021)

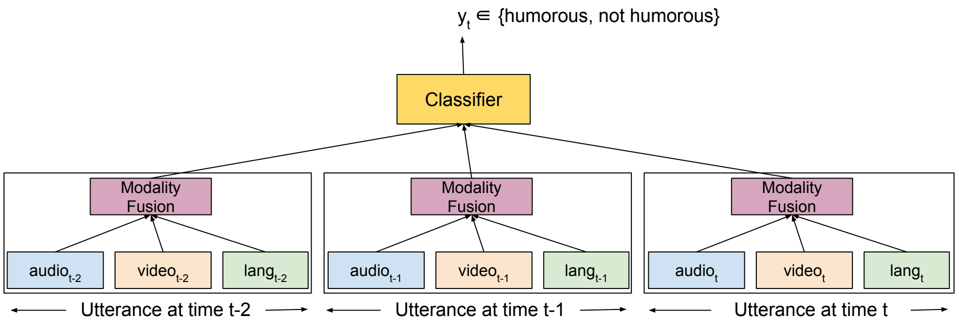

Figure 1: This is the pipeline of our work. Utterances from the past 2 time steps are used to predict humor at current time step. Past 2 time steps are shown just for illustration purposes, we use 5 time steps information in our experiments.

was released which is the first humor recognition dataset in Hindi which is a low resource language. The dataset comprises of 13 episodes from a comedy SitCom "Shrimaan Shrimati Phir Se". The dataset has 136 scenes(based on context) and 6191 utterances in total. There are in total 41 speakers and each utterance is provided with the speaker information. The authors of the paper also implement strong baselines but we do not compare our results with them as no train test split is provided for the experiments.

The aim of this work is to study the problem of multimodal humor recognition on a low-resource language(Hindi) based M2H2 dataset, and implement algorithms to address different parts of this problem. The different parts that we explore in this work are shown in Figure 2. Our final model combines solutions to all these parts(Figure 3).

## 3 Multimodal Multiparty Humour Recognition

In this section we describe the problem statement for our multimodal multiparty humour recognition problem and the dataset which we use for all our experiments.

### 3.1 Problem description

In multiparty humour recognition we assume that we have $m$ parties or speakers, $\{1, ..., m\}$ and at a particular time instance $t$, we have an utterance $u_t$. An utterance is a part of a dialogue where one party speaks without any breaths or pauses. In a multimodal sceneario, we can have more than one signals to describe an utterance. In our case, we consider 3 signals, text, video and audio so the utterance $u_t$ can be written as a combination of the 3 signals such that $u_t = [u_t^{text}, u_t^{video}, u_t^{audio}]$. Here $u_t^{text}$ is the text content, $u_t^{video}$ is the video content, $u_t^{audio}$ is the audio content all corresponding to the utterance $u_t$. There also exists a speaker corresponding to each of the utterances $u_t$ so there exists a mapping function $f$ which maps the utterance $u_t$ to one of the m speakers such that $u_t^{speaker} = f(u_t)$ where $u_t^{speaker} \in \{1, ..., m\}$.

### 3.2 Dataset

The M2H2 dataset by Chauhan et al. (2021) is a multimodal multiparty dataset. The dataset contains 6191 utterances in Hindi fron the TV show *"Shrimaan Shrimati Phir Se"*. The utterances are grouped into scenes based on context. The dataset overall contains 41 speakers and 136 scenes. The dataset does not provide a test set so we use our own train-test split. The train set contains 104 scenes whereas the test set contains 32 scenes. The dataset contains only 2 classes humour and non-humour. There are 2089 utterances in the humorous class and 4102 utterances in the non-humorous class. The ratio of the number of utterances in the humorous class to the number of utterances in the nun-humorous class is approximately 1:2. The dataset suffers from the class imbalance problem with respect to the number of utterances in each class.

## 4 Method

This section describes the approach which we explored to solve the problem at hand. Figure 2 outlines the different components of our approach. We further go on the describe each of the different components in detail.

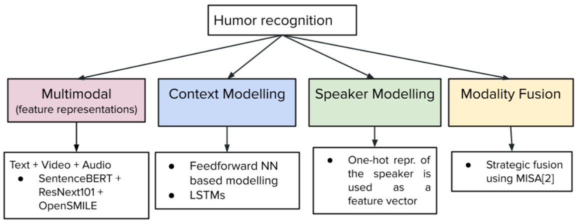
Humor recognition
Multimodal
(feature representations)
Context Modelling
Speaker Modelling
Modality Fusion
Text + Video + Audio
SentenceBERT + ResNext101 + OpenSMILE
Feedforward NN based modelling
LSTMs
One-hot repr. of the speaker is used as a feature vector
Strategic fusion using MISA[2]

Figure 2: Flowchart

## 4.1 Data pre-processing and Feature extraction

Here we describe the techniques we use to pre-process the data and the different representation techniques we use for obtaining the feature representations of the different modalities.

### 4.1.1 Text Modality

For the text modality we try 3 different feature representation. We first make use of FastText by Bojanowski et al. (2017) which is an improved version of Word2Vec by Mikolov et al. (2013). We pre-process the text by removing stop words and specific symbols used in Hindi. The representations obtained for each utterance by using FastText are 300 in dimension.

We then go on to use contextual models such as BERT by Devlin et al. (2018) for obtaining the text representations. BERT models are bidirectional transformer models which provide unsupervised language representations and are pretrained on a large corpus of data. We use IndicBERT by Kakwani et al. (2020) which is a multilingual ALBERT model by Lan et al. (2019). It exclusively covers 12 major Indian languages. Kakwani et al. (2020) create a custom dataset for diverse tasks such as Article Genre Classification and Headline Prediction and IndicBERT was trained for these tasks over datasets which had approximately 9 billion tokens.

In order to obtain embedding vectors for an utterance or a sentence, vanilla BERT style models require us to have a representation of length ($seqlen$ x 768) which might turn out to be very large when the $seqlen$ is large. In order to avoid using such large feature representations for each utterance, we make use of the SentenceBERT model. SentenceBERT makes use of siamese and triplet network architecture whose embeddings can be used to compute semantic similarity information between 2 sentences by using distance metrics like cosine similarity which could be useful for our task of classification. The representatons obtained for each utterance using SentenceBERT are 512 in dimension.

### 4.1.2 Video Modality

Each video contains different number of frames so we sample 16 frames from each of the videos to create corresponding videos which are all equal in length. We then pass all the videos through a ResNext(Hara et al., 2017) 3D CNN [1] in order to extract features. The extracted features have a dimension of 2048.

### 4.1.3 Audio Modality

For extracting the audio features, we make use of opensmile [2] library. For each of the audio files opensmile gives us feature representations of 65 dimension for the different subsections of the audio file. The number of subsections is different for all the audio files. We average the features of the different subsections to get an overall feature representation for the audio file and obtain the final 65 dimensional representation.

## 4.2 Context Modelling

Humor has a strong dependence on the context of the scene. A humorous scene seen out of context can seem to be ordinary. Therefore, we use different techniques to use context information for predicting humor. We explore LSTM[ref to paper], and GRU for context modelling. The input to these context modelling units was a sequence of fused representation(of different modalities) that were extracted from a sequence of utterances in a scene. The final output was the probability of the last utterance in the sequence being humorous.

## 4.3 Speaker Modelling

Speaker information i.e. the speaker ID is a powerful signal for detecting emotion(here humor). It is because if a prior is learned over the speakers w.r.t. the "funnyness" of every person in the dataset, then the posterior of detecting humor given we know the speaker would have a high predictive performance w.r.t. modelling without the speaker information. For example, if a comedian is involved in a conversation on a TV show then it is very likely that he/she

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

[2] https://github.com/audeering/opensmile-python

is going to say something funny, therefore knowledge of the identity of the speaker of an utterance could help predict humor better. We use a one-hot encoded vector for representing each speaker.

### 4.4 Modality fusion

Different modalities capture different aspects of humor, for example funny facial expressions are captured by video, jokes are captured by text and subtle tone and voice changes are captured by audio. Also, across modalities there are certain commonalities which need to be exploited as they also capture important signals, for example when a person is laughing his/her audio and video both reflect the same thing. One can also argue that these common signals are less noisy as they are present in multiple modalities. The above two aspects, that is, modality specific information and modality invariant information are effectively exploited by MISA(Hazarika et al., 2020) for effective multimodal fusion. MISA maps each modality representation to two different latent spaces:(i) Modality-invariant space and (ii) Modality-specific space. The latent representation of each modality in the modality-invariant space should be similar as they have to capture commonalities, therefore MISA minimizes central moment discrepancy between them. Now, to ensure that the latent space representation of each modality in the modality-specific space captures different things(about the input) than the modality-invariant space, MISA enforces a soft orthogonality constraint between the two representations. Finally, MISA also has a reconstruction loss in order to ensure no relevant information is lost in the modality-invariant and modality-specific space.

The final architecture of our method which combines modality fusion via MISA, context modelling using LSTM cell and speaker state modelling using a one-hot encoded representation of the speaker is given in Figure 3.

## 5 Results and Discussion

### 5.1 Metrics

The standard metric for classification, classification accuracy used in  gives us the percentage of the correctly classified examples but does not give a holistic picture of the model performance in cases when the dataset is imbalanced. When the dataset is imbalanced, the classification accuracy can be high even though the actual model performance is low. Thus in our cases we use the F1-score as our metric.

F1-score is the harmonic mean of the precision and recall. Precision gives the fraction of relevant instances among the retrieved instances whereas recall gives the fraction of relevant instances which were retrieved and can be viewed as the sensitivity in binary classification.

### 5.2 Experimental Setting

For all our experiments we use pre-trained models for feature extraction. We do not fine tune the large models such as SentenceBERT(SBERT) and 3D CNN on our dataset since it is computationally infeasible for us to train such models. We extract features from these pre-trained models and pass them through our classification model. The classification model is trained from scratch. We use the Adam optimizer and a batch size of 4 for our experiments.

### 5.3 Component wise Analysis

In this sub section we focus on different components that we explored and what performance gains we achieved by adding each of them in our pipeline.

#### 5.3.1 Multi-Modal Representations

| Method | F1 | Precision | Recall |
|---|---|---|---|
| IndicBERT (text) | 0.59 | 0.60 | 0.65 |
| SBERT (text) | **0.59** | **0.62** | **0.66** |
| FastText (text) | 0.53 | 0.55 | 0.66 |
| ResNext101 3D CNN (video) | 0.55 | 0.57 | 0.66 |
| openSmile (audio) | 0.53 | 0.44 | 0.66 |

Table 1: Performance with uni-modal data

We used IndicBERT, FastText and SBERT for extracting representations from text. IndicBERT and FastText give word embeddings and in order to generate embedding for utterances which contained multiple words, we took the average of the word embeddings. For video and text we used ResNext 3D and OpenSmile respectively. It can be observed from Table 1 that SBERT performs the best among the text modality representations, therefore we use it for all our further experiments.

Table 2 shows an ablation study where we can see that even with naive fusion of modalities i.e. simple concatenation of features(followed by a classifier), tri-modal gives better results than bi-modal and uni-modal representations.

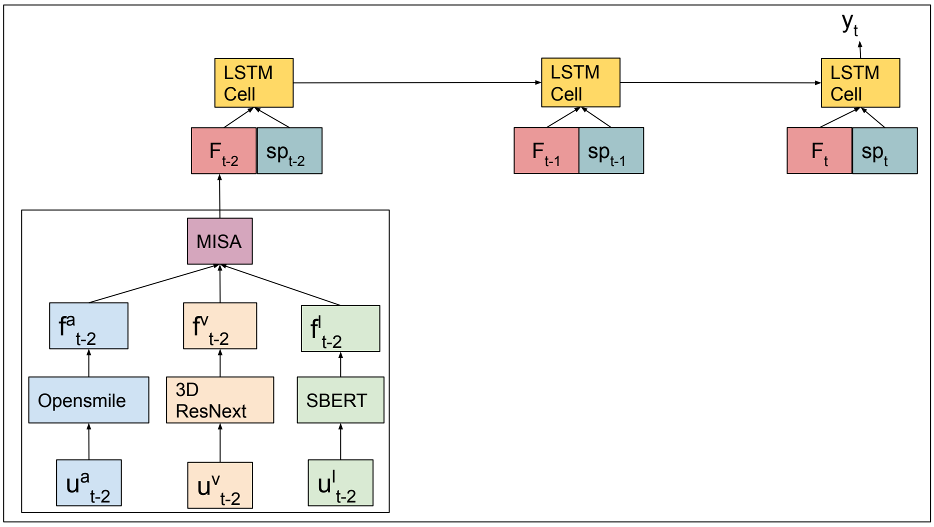

Figure 3: Final architecture of our humor recognition system. $F_t$ is the fused multimodal representation, $sp_t$ is the speaker ID representation and $y_t$ is the output node which estimates the probability of the utterance at time $t$ being humorous.

| Method | F1 | Precision | Recall |
|---|---|---|---|
| SBERT + ResNext101 3D CNN | **0.60** | 0.60 | 0.64 |
| SBERT + openSmile | 0.55 | 0.61 | 0.66 |
| ResNext101 3D CNN + openSmile | 0.53 | 0.55 | 0.66 |
| SBERT + 3D CNN + openSmile | 0.58 | **0.62** | **0.67** |

Table 2: Performance with multiple modalities. First three rows are for bi-modal setting and the fourth/last row is for the tri-modal setting.

### 5.3.2 Context modelling

In order to choose between different recurrent neural network architectures, we perform experiments on the text representations. We can observe from Table 3 that LSTM architecture gives the best performance overall, therefore we use it for all our future experiments for context modelling. Although the improvement is marginal, but as LSTM cell provides one point improvement on F1 score over not using context modelling, we use LSTM.

### 5.3.3 Modality fusion

MISA operates on each individual utterances and provides a 4 percent increase in precision on the F1 score w.r.t. simple concatenation of feature representation of different modalities, as can be observed in the top part of Table 4. We also use two kinds of context modelling methods i.e. using NN and LSTM along with the fused representation of MISA. The LSTM based architecture gives the best performance overall as can be observed in the bottom part of Table 4.

| Method | F1 | Precision | Recall |
|---|---|---|---|
| SBERT (No context modelling) | 0.59 | **0.62** | **0.65** |
| SBERT + RNN | 0.59 | 0.60 | 0.59 |
| SBERT + LSTM | **0.60** | **0.62** | **0.65** |
| SBERT + GRU | 0.59 | 0.59 | 0.60 |
| SBERT + BiLSTM | 0.60 | 0.60 | 0.63 |

Table 3: Performance with context modelling

| Method | F1 | Precision | Recall |
|---|---|---|---|
| SBERT + 3D CNN + openSmile (simple concat) | 0.58 | 0.62 | 0.67 |
| MISA | 0.62 | 0.62 | 0.65 |
| MISA with NN based context modelling | 0.62 | **0.65** | 0.62 |
| MISA with LSTM based context modelling | **0.63** | 0.62 | **0.66** |

Table 4: Performance with Modality Fusion

### 5.3.4 Speaker modelling

The speaker information might help us to get relevant information about humour. We have 41 speakers in our dataset. We model the speaker information for as a one hot vectors. Table 5 shows the results by including the speaker information.

| Method | F1 | Precision | Recall |
|---|---|---|---|
| SBERT + 3D CNN + openSmile (simple concat) | 0.58 | 0.62 | 0.67 |
| SBERT + 3D CNN + openSmile + Speaker (simple concat) | **0.60** | 0.60 | 0.65 |

Table 5: Performance with speaker information

## 5.4 Final Results

Table 6 gives the final results for our work. Here we can observe that modality fusion along with sequence modelling gives us the best results.

| Method | F1 | Precision | Recall |
|---|---|---|---|
| SBERT + 3D CNN + openSmile | 0.58 | 0.62 | **0.67** |
| SBERT + 3D CNN + openSmile + Speaker | 0.60 | 0.60 | 0.65 |
| MISA + LSTM based context modelling + Speaker | **0.63** | **0.63** | 0.66 |

Table 6: Summary of performance of different methods

# 6 Conclusion and Future work

In this paper, we study the effect of different components of the multimodal multiparty humour recognition pipeline. We show that text is an important modality as compared to video and audio for our dataset and leveraging signals from different modalities performs better than using only a single modality. We also show that modality fusion techniques gives a performance boost as compared to concatenation of individual modalities. Our best performing model uses context modelling with modality fusion using MISA.

For our future work, we plan on using sophisticated context modelling techniques rather than using simple LSTM networks. We feel that improving the video representations by using facial emotion recognition models and audio representations by using models such as Wav2Vec can help us to improve the performance of our model. We would also like to use models pre-trained on other datasets and leverage cross-lingual humour datasets to improve the performance of our dataset owing to the small size of our dataset for training deep learning models.

# 7 Acknowledgements

We would like to thank Suvodip Dey and Dr. Maunendra Sankar Desarkar for their guidance and valuable feedback.

# References

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

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.

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

Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. 2017. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? *arXiv preprint*, arXiv:1711.09577.

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*.

Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. 2020. Misa: Modality-invariant and-specific representations for multimodal sentiment analysis. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1122–1131.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, NC Gokul, Avik Bhattacharyya, Mitesh M Khapra, and Pratyush Kumar. 2020. inlpsuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for indian languages. In *Proceedings of the 2020 Conference on*

*Empirical Methods in Natural Language Processing: Findings*, pages 4948–4961.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Badri N Patro, Mayank Lunayach, Deepankar Srivastava, Hunar Singh, Vinay P Namboodiri, et al. 2021. Multimodal humor dataset: Predicting laughter tracks for sitcoms. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 576–585.

Jiaming Wu, Hongfei Lin, Liang Yang, and Bo Xu. 2021. Mumor: A multimodal dataset for humor detection in conversations. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 619–627. Springer.