Sentiment and Emotion help Sarcasm? A Multi-task Learning Framework for Multi-Modal Sarcasm, Sentiment and Emotion Analysis

Dushyant Singh Chauhan, Dhanush S R, Asif Ekbal and Pushpak Bhattacharyya
Department of Computer Science & Engineering
Indian Institute of Technology Patna
Patna, Bihar, India-801106
{1821CS17, dhanush.cs16, asif, pb}@iitp.ac.in

Abstract

In this paper, we hypothesize that sarcasm is closely related to sentiment and emotion, and thereby propose a multi-task deep learning framework to solve all these three problems simultaneously in a multi-modal conversational scenario. We, at first, manually annotate the recently released multi-modal MUStARD sarcasm dataset with sentiment and emotion classes, both implicit and explicit. For multi-tasking, we propose two attention mechanisms, *viz.* Inter-segment Inter-modal Attention (I_e-Attention) and Intra-segment Inter-modal Attention (I_a-Attention). The main motivation of I_e-Attention is to learn the relationship between the different segments of the sentence across the modalities. In contrast, I_a-Attention focuses within the same segment of the sentence across the modalities. Finally, representations from both the attentions are concatenated and shared across the five classes (i.e., sarcasm, implicit sentiment, explicit sentiment, implicit emotion, explicit emotion) for multi-tasking. Experimental results on the extended version of the MUStARD dataset show the efficacy of our proposed approach for sarcasm detection over the existing state-of-the-art systems. The evaluation also shows that the proposed multi-task framework yields better performance for the primary task, i.e., sarcasm detection, with the help of two secondary tasks, emotion and sentiment analysis.

1 Introduction

Sarcasm is an essential aspect of daily conversation, and it adds more fun to the language. Oscar Wilde, an Irish poet-playwright, quotes, *"Sarcasm is the lowest form of wit, but the highest form of intelligence"*. Irrespective of its relation with intelligence, sarcasm is often challenging to understand.

Sarcasm is often used to convey thinly veiled disapproval humorously. This can be easily depicted through the following example, *"This is so good, that I am gonna enjoy it in the balcony. I can enjoy my view, whilst I enjoy my dessert."* This utterance, at an outer glance, conveys that the speaker is extremely pleased with his dessert and wants to elevate the experience by enjoying it in the balcony. But, careful observation of the sentiment and emotion of the speaker helps us understand that the speaker is disgusted with the dessert and has a negative sentiment during the utterance (c.f. Figure 1). This is where sentiment and emotion come into the picture. Sentiment, emotion and sarcasm are highly intertwined, and one helps in the understanding of the others better.

Figure 1: Example to show that sentiment and emotion of the speaker can influence sarcasm detection

Even though sentiment, emotion, and sarcasm are related, sarcasm was treated separately from its other counterparts in the past due to its complexity and its high dependency on the context. Moreover, multi-modal input helps the model to understand the intent and the sentiment of the speaker with more certainty. Thus in the context of a dialogue, multi-modal data such as video (acoustic + visual) along with text helps to understand the sentiment and emotion of the speaker, and in turn, helps to detect sarcasm in the conversation.

In this paper, we exploit these relationships, and make use of sentiment and emotion of the speaker for predicting sarcasm, specifically for the task, in a multi-modal conversational context. The main contributions and/or attributes of our proposed research are as follows: **(a).** *we propose a multi-task*

learning framework for multi-modal sarcasm, sentiment, and emotion analysis. We leverage the utility of sentiment and emotion of the speaker to predict sarcasm. In our multi-task framework, sarcasm is treated as the primary task, whereas emotion analysis and sentiment analysis are considered as the secondary tasks. ***(b).*** *We also propose two attention mechanisms viz. I_e-Attention and I_a-Attention to better combine the information across the modalities to effectively classify sarcasm, sentiment, and emotion.* ***(c).*** *We annotate the recently released Sarcasm dataset, MUStARD with sentiment and emotion classes (both implicit and explicit), and* ***(d).*** *We present the state-of-the-art for sarcasm prediction in multi-modal scenario.*

2 Related Work

A survey of the literature suggests that a multi-modal approach towards sarcasm detection is a fairly new approach rather than a text-based classification. Traditionally, rule-based classification (Joshi et al., 2017; Veale and Hao, 2010) approaches were used for sarcasm detection. Poria et al. (2016) have exploited sentiment and emotion features extracted from the pre-trained models for sentiment, emotion, and personality on a text corpus, and use them to predict sarcasm through a Convolutional Neural Network.

In recent times, the use of multi-modal sources of information has gained significant attention to the researchers for affective computing. Mai et al. (2019) proposed a new two-level strategy (*Divide, Conquer, and Combine*) for feature fusion through a Hierarchical Feature Fusion Network for multi-modal affective computing. Chauhan et al. (2019) exploits the interaction between a pair of modalities through an application of Inter-modal Interaction Module (IIM) that closely follows the concepts of an auto-encoder for the multi-modal sentiment and emotion analysis. Ghosal et al. (2018) proposed a contextual inter-modal attention based framework for multi-modal sentiment classification. In other work (Akhtar et al., 2019), an attention-based multi-task learning framework has been introduced for sentiment and emotion recognition.

Although multi-modal sources of information (e.g., audio, visual, along with text) offers more evidence in detecting sarcasm, this has not been attempted much, one of the main reasons being the non-availability of multi-modal datasets. Recently, researchers (Castro et al., 2019) have started exploiting multi-modal sources of information for sarcasm detection. It is true that the modalities like acoustic and visual often provide more evidences about the context of the utterance in comparison to text. For sarcasm detection, the very first multi-modal dataset named as MUStARD has been very recently released by Castro et al. (2019), where the authors used a Support Vector Machine (SVM) classifier for sarcasm detection.

In our current work, we at first extend the MUStARD dataset (Castro et al., 2019) by manually labeling each utterance with sentiment and emotion labels. Thereafter, we propose a deep learning based approach along with two attention mechanisms (I_e-Attention and I_a-Attention) to leverage the sentiment and emotion for predicting sarcasm in a multi-modal multi-task framework. Further, to the best of our knowledge, this is the very first attempt at solving the multi-modal sarcasm detection problem in a deep multi-task framework. We demonstrate through a detailed empirical evaluation that sarcasm detection can be improved significantly if we are successful in leveraging the knowledge of emotion and sentiment using an effective multi-task framework.

3 Dataset

The MUStARD (Castro et al., 2019) dataset consists of conversational audio-visual utterances (total of 3.68 hours in length). This dataset consists of 690 samples, and each sample consists of utterance accompanied by its context and sarcasm label. The samples were collected from 4 popular TV Series *viz.*, Friends, The Big Bang Theory, The Golden Girls, and Sarcasmaholics Anonymous and manually annotated for the sarcasm label. The dataset is balanced with an equal number of samples for both sarcastic and non-sarcastic labels. The utterance in each sample consists of a single sentence, while the context associated with it consists of multiple sentences that precede the corresponding utterance in the dialogue. We manually re-annotated this dataset to introduce sentiment and emotion labels in addition to sarcasm. We define two kinds of emotion and sentiment values *viz.*, implicit and explicit, which are discussed in the following subsections.

3.1 Sentiment

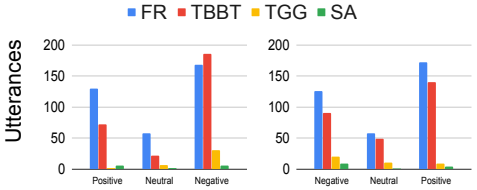
For sentiment annotation of an utterance, we consider both implicit and explicit affect information. The *implicit sentiment* of an utterance is determined

with the help of context. Whereas, *explicit sentiment* of an utterance is determined directly from itself, and no external knowledge from the context is required to infer it. We consider three sentiment classes, namely *positive, negative* and *neutral*. For the example in Figure 1, the implicit sentiment would be *Negative*, whereas explicit sentiment is *Positive*.

Table 1 shows the overall ratio of *implicit* and *explicit sentiment* labels, respectively. Whereas, Figure 2a and Figure 2b depict the show-wise ratio and distribution of each label.

Implicit Sentiment			**Explicit Sentiment**		
Neg	***Neu***	***Pos***	***Neg***	***Neu***	***Pos***
391	89	210	246	119	325

Table 1: Sentiment distribution.



(a) *Show-wise-IS.* (b) *Show-wise-ES.*

Figure 2: Distribution of implicit sentiment (IS) and explicit sentiment (ES).

3.2 Emotion

Like sentiment, we annotate each sentence on the context and utterance for the *implicit* and *explicit emotion*. We annotate the dataset for 9 emotion values, *viz.* anger (An), excited (Ex), fear (Fr), sad (Sd), surprised (Sp), frustrated (Fs), happy (Hp), neutral (Neu) and disgust (Dg). Each utterance and context sentence are annotated, and these can have multiple labels per sentence for both *implicit* and *explicit emotion*. In the example of Figure 1, the implicit emotion of the speaker would be *disgust* while the explicit emotion is *happy*.

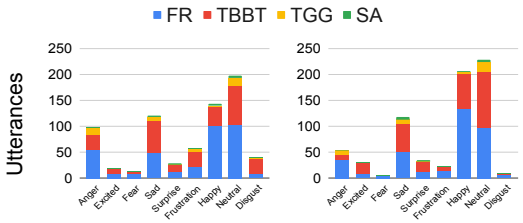
Table 2 shows the overall ratio of *implicit* and *explicit emotion* labels, respectively. Whereas Figure 3a and Figure 3b depict the show-wise ratio and distribution of each label.

3.3 Annotation Guidelines

We annotate all the samples with four labels (implicit sentiment/emotion and explicit sentiment/emotion). We employ three graduate students highly proficient in the English language with prior

Explicit Emotion								
An	***Ex***	***Fr***	***Sd***	***Sp***	***Fs***	***Hp***	***Neu***	***Dg***
54	30	6	118	35	23	206	228	10
Implicit Emotion								
An	***Ex***	***Fr***	***Sd***	***Sp***	***Fs***	***Hp***	***Neu***	***Dg***
97	18	14	121	29	57	143	198	39

Table 2: Emotion distribution.



(a) *Show-wise-IE.* (b) *Show-wise-EE.*

Figure 3: Distribution of implicit emotion (IE) and explicit emotion (EE).

experience in labeling *sentiment, emotion*, and *sarcasm*. The guidelines for annotation, along with some examples, were explained to the annotators before starting the annotation process.

The annotators were asked to annotate every utterance with as many emotions present in the utterance as possible, along with the sentiment. Initially, the dataset was annotated for explicit labels, with only the utterances provided to the annotators. Later, for the implicit labels, we also made the corresponding context video available to provide the relevant information for each sample. This method helps the annotators to resolve the ambiguity between the implicit and explicit labels. A majority voting scheme was used for selecting the final emotion and sentiment. We achieve an overall Fleiss' (Fleiss, 1971) kappa score of 0.81, which is considered to be reliable.

4 Proposed Methodology

In this section, we describe our proposed methodology, where we aim to leverage the multi-modal sentiment and emotion information for solving the problem of multi-modal sarcasm detection in a multi-task framework. We propose a segment-wise inter-modal attention based framework for our task. We depict the overall architecture in Figure 4. The extended dataset with annotation guidelines and source code are available at `http://www.iitp.ac.in/~ai-nlp-ml/resources.html`.

Each sample in the dataset consists of an utterance (u) accompanied by its context (c) and labels

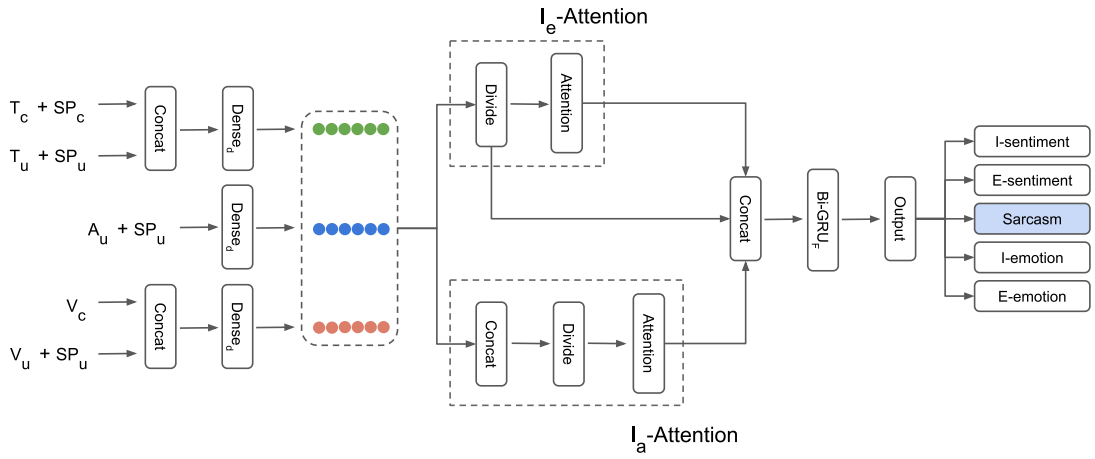


Figure 4: Overall architecture of the proposed multi-modal sarcasm detection framework.

(*sarcasm, implicit sentiment, explicit sentiment, implicit emotion*, and *explicit emotion*). The context associated with the utterance consists of multiple sentences (say, *N*) that precede the corresponding utterance in the dialogue. Each utterance and its' context is associated with its' speaker *i.e., speaker of utterance (SP_u)* and *speaker of context (SP_c)*, respectively. We represent SP_u and SP_c by using a one-hot vector embedding.

We divide our proposed methodology into three subsections *i.e., Input Layer, Attention Mechanism* and *Output Layer*, which are described below:

4.1 Input Layer

The proposed model takes multi-modal inputs *i.e., text (T), acoustic (A)*, and *visual (V)*. We describe the utterance and its' context for all the modalities below:

4.1.1 Text

Utterance: Let us assume, in an utterance, there n_t number of words $w_{1:n_t} = w_1, ..., w_{n_t}$, where $w_j \in \mathbb{R}^{d_t}$, d_t = 300, and w_js are obtained using *fastText* word embeddings (Joulin et al., 2016). The utterance is then passed through a bi-directional Gated Recurrent Unit (Cho et al., 2014) ($BiGRU_T$[1]) to learn the contextual relationship between the words. We apply the attention over the output of $BiGRU_T$ to extract the important contributing words *w.r.t.* sarcasm. Finally, we apply $BiGRU_F$[2] to extract the sentence level features. We then concatenate the speaker information of the utterance with the output of $BiGRU_F$. This is denoted by $T_u + SP_u$, where T_u denotes the utterance for the text modality and SP_u denotes the speaker for that particular utterance.

Context: There are N_c number of sentences in the context where each sentence has n_{tc} words. For each sentence, words are passed through $BiGRU_F$ to learn the contextual relationship between the words, and to obtain the sentence-wise representation. Then, we apply self-attention over the output of $BiGRU_F$ to extract the important contributing sentences for the utterance. Finally, we concatenate the speaker information with each sentence and pass through the $BiGRU_F$ to obtain the $T_c + SP_c$, where T_c denotes the context of the text modality, and SP_c denotes the speaker of that context.

4.1.2 Visual

Utterance: Let us assume there are n_v number of visual frames *w.r.t.* an utterance. We take the average of all frames to extract the sentence level information for the visual modality (Castro et al., 2019), and concatenate this with the speaker information. This is denoted as $V_u + SP_u$, where $V_u \in \mathbb{R}^{d_v}$ and d_v = 2048.

Context: Given n_{vc} number of visual frames *w.r.t.* all the sentences, we take the average of all the visual frames (Castro et al., 2019) to extract the context level information, and denote this as V_c. As sentence-wise visual frames are not provided in the dataset, speaker information is not considered.

4.1.3 Acoustic

Utterance: Given n_a number of frames for the acoustic *w.r.t.* an utterance, we take the average of all the frames to extract the sentence level in-

[1] $BiGRU_T$ refers to the Bi-directional GRU units where output from all the time steps are forwarded in the model.

[2] $BiGRU_F$ refers to the Bi-directional GRU units where output from the last time step is forwarded in the model.

formation (Castro et al., 2019), and concatenate with the speaker of the utterance. We denote this as $A_u + SP_u$, where $A_u \in \mathbb{R}^{d_a}$ and d_a = 283 corresponds to the utterance of the acoustic modality.

Context: For text, we concatenate the utterance ($T_u + SP_u$) with its context ($T_c + SP_c$). For visual, we concatenate the utterance ($V_u + SP_u$) with its context (V_c) while for acoustic, we consider only the utterance $A_u + SP_u$ (c.f. Figure 4). We do not consider any context information of the acoustics as it often contains information of many speakers, background noise, and noise due to laughter cues (which is not a part of the conversation). Hence, it might be difficult to disambiguate this with the laughter part of the conversation. Whereas, in the case of visual modality, it majorly contains the image of the speaker along with sentiment and emotion information. Thus, visual will not have a similar kind of problem as acoustic.

It is also to be noted that for a fair comparison with the state-of-the-art system (Castro et al., 2019), we take the average of the acoustic and visual features across the sentences.

4.2 Attention Mechanism

In any multi-modal information analysis, it is crucial to identify the important feature segments from each modality, so that when these are combined together can improve the overall performance. Here, we propose two attention mechanisms: (i). Inter-segment Inter-modal Attention (I_e-Attention), and (ii). Intra-segment Inter-modal Attention (I_a-Attention).

First, we pass the input representation from all the three modalities through a fully-connected layer ($Dense_d$) to obtain the feature vector of length (d). These feature vectors are then forwarded to the aforementioned attention mechanisms.

4.2.1 Inter-segment Inter-modal Attention

For each modality, we first split the feature vector into k-segments to extract the fine level information. We aim to learn the relationship between the feature vector of a segment of an utterance in one modality and feature vector of the another segment of the same utterance in another modality through this mechanism (c.f. Figure 5). Then, an I_e-Attention is applied among the segments for every possible pair of modalities *viz.,* TV, VT, TA, AT, AV, and VA. The overall procedure of I_e-Attention is depicted in Algorithm 1.

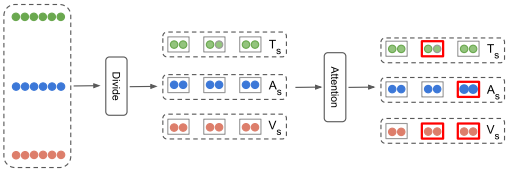


Figure 5: Procedure of the proposed I_e-Attention Mechanism.

Algorithm 1

procedure I_e-ATTENTION(X, Y)
  for $s \in 1, ..., d/k$ **do** ▷ $s = segment$
    $S_x[g] = X[k * s, k * s + k]$ ▷ $X \in \mathbb{R}^{sk}$
    $S_y[g] = Y[k * s, k * s + k]$ ▷ $Y \in \mathbb{R}^{sk}$
  return ATTENTION(S_x, S_y)

procedure I_a-ATTENTION(X, Y, Z)
  $R = concatenate(X, Y, Z)$
  for $s \in 1, ..., d/k$ **do** ▷ $s = segment$
    $S_r[s] = R[k * s, k * s + k]$ ▷ $X \in \mathbb{R}^{sk}$
  return ATTENTION(S_r, S_r)

procedure ATTENTION(B, C)
  /*Cross-Segment Correlation*/
  $M \leftarrow B.B^T$
  /*Cross-Segment Inter-modal Attention*/
  for $i, j \in 1, ..., L$ **do** ▷ $L = length(M)$
    $P(i,j) \leftarrow \frac{e^{M(i,j)}}{\sum_{l=1}^{g} e^{M(i,l)}}$
  $O \leftarrow P.C$
  /*Multiplicative gating*/
  return $O \odot B$ ▷ Element-wise mult.

4.2.2 Intra-segment Inter-modal Attention

For each utterance, we first concatenate the feature vectors (*i.e.,* $\in \mathbb{R}^d$) obtained from the three modalities *i.e.,* $\in \mathbb{R}^{3 \times d}$ (c.f. Figure 6) and then split the feature vector into k-segments *(i.e.,* $\in \mathbb{R}^{3 \times \frac{d}{k}}$*)*. Now, we have a mixed representation of all the modalities, i.e. visual, audio and text. The aim is, for a specific segment of any particular utterance, to establish the relationship between the feature vectors obtained from the different modalities.

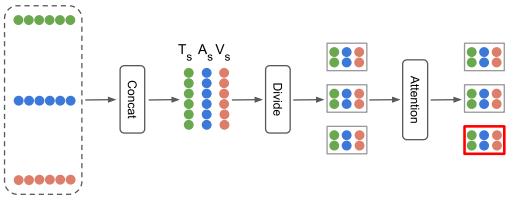


Figure 6: Procedure of the proposed I_a-Attention mechanism

4.3 Output Layer

Motivated by the residual skip connection (He et al., 2016), the outputs of I_e-Attention and I_a-Attention along with the representations of individual modalities are concatenated (c.f Figure 4). Finally, the concatenated representation is shared across the five branches of our proposed network *(i.e., sarcasm, I-sentiment, E-sentiment, I-emotion, & E-emotion)* corresponding to three tasks, classification for the prediction (one for each task in the multi-task framework). Sarcasm and sentiment branches contain a *Softmax* layer for the final classification, while the emotion branch contains a *Sigmoid* layer for the classification. The shared representation will receive gradients of error from the five branches (sarcasm, I-sentiment, E-sentiment, I-emotion, & E-emotion), and accordingly adjusts the weights of the models. Thus, the shared representations will not be biased to any particular task, and it will assist the model in achieving better generalization for the multiple tasks.

5 Experiments and Analysis

We divide the whole process into four categories: **i).** utterance without context without speaker (i.e., we do not use the information of context and its' speaker with utterance); **ii).** utterance with context without speaker (i.e., we use the context information with utterance but not speaker information); **iii).** utterance without context with speaker (i.e., we use the speaker information with utterance but not context information); and **iv).** utterance with context with speaker (i.e., we use the context and its' speaker information with utterance).

5.1 Experimental Setup

We perform all the experiments for the setup *utterances without context and speaker information* (case i). Hence, even though the sentiment and emotion labels were annotated for both the context and utterance, we use the labels associated with utterances only for our experiments.

Our experimental setup is mainly divided into two main parts (Castro et al., 2019):

- **Speaker Independent Setup:** In this experiment, samples from The Big Bang Theory, The Golden Girls, and Sarcasmaholics Anonymous were considered for the training, and samples from the Friends Series were considered as the test set. Following this step, we were able to reduce the effect of the speaker in the model.
- **Speaker Dependent Setup:** This setup corresponds to the five-fold cross-validation experiments, where each fold contains samples taken randomly in a stratified manner from all the series.

We evaluate our proposed model on the multimodal sarcasm dataset[3], which we extended by incorporating both emotion and sentiment values. We perform *grid search* to find the optimal hyper-parameters (c.f. Table 3). Though we aim for a generic hyper-parameter configuration for all the experiments, in some cases, a different choice of the parameter has a significant effect. Therefore, we choose different parameters for a different set of experiments.

Parameters	***Speaker Dependent***	***Speaker Independent***
Bi-GRU	2×200 *neurons*, *dropout=0.3*	
Dense layer	200 *neurons*, *dropout=0.3*	
Activations	*ReLu*	
Optimizer	*Adam* (*lr=0.001*)	
Output	*Softmax* (Sent) & *Sigmoid* (Emo)	
Loss	*Categorical cross-entropy* (Sent) *Binary cross-entropy* (Emo)	
Batch	*32*	
Epochs	*200*	
#Segments (k)	50	25

Table 3: Model configurations

We implement our proposed model on the Python-based Keras deep learning library. As the evaluation metric, we employ precision (P), recall (R), and F1-score (F1) for sarcasm detection. We use *Adam* as an optimizer, *Softmax* as a classifier for sarcasm and sentiment classification, and the *categorical cross-entropy* as a loss function. For emotion recognition, we use *Sigmoid* as an activation function and optimize the *binary cross-entropy* as the loss.

5.2 Results and Analysis

We evaluate our proposed architecture with all the possible input combinations i.e. bi-modal (*T+V, T+A, A+V*) and tri-modal (*T+V+A*). We do not consider uni-modal inputs (*T, A, V*) because our proposed attention mechanism requires at least two modalities. We show the obtained results in Table 4, that outlines the comparison between the multi-task (MTL) and single-task (STL) learning frameworks

[3]https://github.com/soujanyaporia/MUStARD

Labels			T + V			T + A			A + V			T + A + V		
			P	R	F1	P	R	F1	P	R	F1	P	R	F1
Speaker Dependent	*STL*	***Sar***	71.52	70.61	69.32	64.20	64.20	63.88	71.90	71.01	70.64	72.08	71.62	72.01
	MTL	***Sar + Sent***	69.65	69.42	69.33	64.09	60.72	58.21	72.20	71.45	71.18	72.52	71.73	72.07
		Sar + Emo	71.76	70.86	70.54	**65.76**	**65.65**	**65.60**	72.60	71.59	71.25	72.76	71.88	72.11
		Sar + Sent + Emo	**72.76**	**71.88**	**71.61**	62.23	61.15	59.61	**72.73**	**71.88**	**71.81**	**73.40**	**72.75**	**72.57**
Speaker Independent	*STL*	***Sar***	60.11	60.18	60.16	58.23	57.69	57.91	60.44	60.96	60.52	65.98	65.45	65.60
	MTL	***Sar + Sent***	62.74	62.92	62.81	59.25	59.55	52.89	61.60	60.95	61.14	66.97	63.76	63.68
		Sar + Emo	65.11	65.16	65.13	**59.59**	59.55	**59.58**	63.19	63.76	62.91	66.35	65.44	65.63
		Sar + Sent + Emo	**65.48**	**65.48**	**65.67**	59.13	**59.98**	50.27	**65.59**	**63.76**	**63.90**	**69.53**	**66.01**	**65.90**

Table 4: Single Task vs Multi Task: *Without Context and Without Speaker* information.

without taking context and speaker information into consideration. We observe that Tri-modal *(T+A+V)* shows better performance over the *bi-modal* setups.

For *STL*, experiments with only sarcasm class are used, whereas for *MTL*, we use three sets of experiments, i.e. sarcasm with sentiment (*Sar + Sent*), sarcasm with emotion (*Sar + Emo*), and sarcasm with sentiment and emotion (*Sar + Sent + Emo*). For sarcasm classification, we observe that multi-task learning with sentiment and emotion together shows better performance for both the setups (*i.e., speaker dependent* and *speaker independent*) over the Single-task learning framework. It is evident from the empirical evaluation, that both sentiment and emotion assist sarcasm through the sharing of knowledge, and hence *MTL* framework yields better prediction compared to the *STL* framework (c.f. Table 4).

We also show the results for the single-task *(T+A+V)* experiments under speaker-dependent and speaker-independent setups for sentiment and emotion. These results can be considered as *baseline* for the same. The detailed description of sentiment and emotion are described in Section 3.1 and Section 3.2, respectively.

For Sentiment Analysis, the results are shown in Table 5.

Speaker Dependent					
Implicit Sentiment			***Explicit Sentiment***		
P	R	F1	P	R	F1
49.27	57.39	49.12	48.32	52.46	48.11
Speaker Independent					
P	R	F1	P	R	F1
47.05	49.15	40.99	47.73	50.0	45.24

Table 5: Results for Single-task experiments for Sentiment analysis (*T+A+V*).

Similarly, for emotion analysis, the results are shown in Table 6. Along with it, results from the single-Task experiments for each emotion under implicit emotion and explicit emotion for Speaker Dependent and Speaker Independent setups are shown in Table 7 and Table 8, respectively. As each utterance can have multiple emotion labels, we take all the emotions whose respective values are above a threshold. We optimize and cross-validate the evaluation metrics and set the threshold as 0.5 0.45 for speaker-dependent and speaker-independent setups, respectively.

Speaker Dependent					
Implicit Sentiment			***Explicit Sentiment***		
P	R	F1	P	R	F1
80.66	88.51	83.57	85.01	88.90	85.12
Speaker Independent					
P	R	F1	P	R	F1
81.77	88.29	83.88	83.64	88.35	84.37

Table 6: Results for Single-task experiments for Emotion analysis (*T+A+V*).

Speaker Dependent						
Setup	***Implicit Emotion***			***Explicit Emotion***		
	P	R	F1	P	R	F1
An	74.0	85.9	79.5	85.0	92.2	88.4
Ex	94.9	97.3	96.1	91.5	95.6	93.5
Fr	95.9	97.8	96.9	98.3	99.1	98.7
Sd	68.0	82.3	74.5	72.1	83.0	75.5
Sp	91.8	95.8	93.7	90.1	94.9	92.5
Fs	84.2	91.7	87.8	93.4	96.7	95.0
Hp	67.1	79.5	71.4	66.6	71.7	66.5
Neu	60.9	71.6	60.5	70.9	68.3	58.1
Dg	89.0	94.3	91.6	97.1	98.5	97.8

Table 7: Emotion-wise results for Single-Task experiments - Speaker Dependent setup.

Speaker Independent						
Setup	***Implicit Emotion***			***Explicit Emotion***		
	P	R	F1	P	R	F1
An	72.0	84.8	77.9	81.3	90.1	85.5
Ex	95.6	97.7	96.6	94.5	97.2	95.8
Fr	95.0	97.5	96.2	97.8	98.9	98.3
Sd	74.8	86.5	80.3	72.9	85.4	78.7
Sp	92.8	96.3	94.5	93.4	96.6	94.9
Fs	88.5	94.1	91.2	91.7	95.8	93.7
Hp	65.6	71.6	60.8	67.4	62.9	49.6
Neu	50.9	71.3	59.4	52.9	72.7	61.3
Dg	94.5	97.2	95.8	96.1	98.0	97.0

Table 8: Emotion-wise results for Single-Task experiments - Speaker Independent setup.

We further evaluate our proposed model by incorporating context and speaker information to form the three combinations of experiments *viz.*, ***With Context Without Speaker***, ***Without Context With Speaker***, ***With Context and Speaker*** (c.f. Table 9). The experiment without context and without speaker information is same as the tri-modal setup in Table 4. The maximum improvement (1-5% ↑) in performance is observed when the speaker information alone is incorporated in the tri-modal setup. Whereas in Speaker Independent Setup, incorporating both context and speaker information significantly improves the performance (1-5% ↑).

Setups		*Speaker Dependent*			*Speaker Independent*		
Context	Speaker	P	R	F1	P	R	F1
✗	✗	73.40	72.75	72.57	69.53	66.01	65.90
✗	✓	**77.09**	**76.67**	**76.57**	74.69	74.43	74.51
✓	✗	72.34	71.88	71.74	71.51	71.35	70.46
✓	✓	76.07	75.79	75.72	**74.88**	**75.01**	**74.72**

Table 9: Results for different combination of Context-Speaker Experiments

To understand the contribution of I_e-Attention and I_a-Attention towards the performance of the model, an ablation study was performed without the attention-mechanisms (c.f. Table 10).

Setup	*Speaker Dependent*			*Speaker Independent*		
	P	R	F1	P	R	F1
W/o Attention	71.53	69.71	69.02	60.53	61.23	60.44
Proposed	**73.40**	**72.75**	**72.57**	**69.53**	**66.01**	**65.90**

Table 10: Ablation study: Proposed Attention v/s Without Attention.

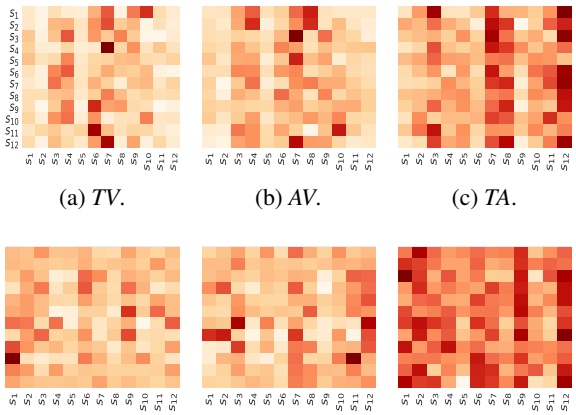
5.3 Comparative Analysis

We compare, under the similar experimental setups, the results obtained in our proposed model (without context and speaker) against the existing models called as baseline (Castro et al., 2019), which also made use of the same dataset. The comparative analysis is shown in Table 11. For tri-modal experiments, our proposed multi-modal multi-task framework achieves the best precision of 73.40% (1.5% ↑), recall of 72.75% (1.4% ↑) and F1-score of 72.57% (1.1% ↑) for proposed multi-task model (Sar + Sent + Emo) as compared to precision of 71.9%, recall of 71.4%, F1-score of 71.5% of the state-of-the-art system. We observe that both sentiment and emotion help to improve the efficiency of sarcasm detection. Similarly, for the Speaker Independent setup, we obtain an improvement of 5.2% in precision, 3.4 % in recall, and 3.1% in F1-score.

We perform statistical significance test (*paired T-test*) on the obtained results and observe that performance improvement in the proposed model over the state-of-the-art is significant with 95% confidence (i.e. p-value< 0.05).

5.4 Error Analysis

We analyze the attention weights to understand the learning behavior of the proposed framework. We take an utterance *i.e., "I love that you take pride in your looks, even when I have to pee in the morning, and you're in there spending an hour on your hair."* (c.f Table 12) from the dataset which is a sarcastic utterance. The MTL (Sar + Sent + Emo) correctly classifies this utterance as sarcastic, while the STL (Sar) predicts it as non-sarcastic. In this utterance, we feel that the speaker is pleased and happy (explicit emotion) where he is angry (implicit emotion) on the other person and is expressing that anger sarcastically.



(a) *TV*. (b) *AV*. (c) *TA*.

(d) *VT*. (e) *VA*. (f) *AT*.

Figure 7: Heatmaps for the all combinations of modalities for I_e-Attention.

We analyze the heatmaps of the attention weights (I_e*-Attention* and I_a*-Attention*) for the above utterance. Each cell of heatmaps for I_e-Attention (c.f. Figure 7) represents the different segments of the sentence across the modalities. Cell (*i,j*) of the heatmap for the modalities (say, *TV*) represents the influence of s_j of visual on s_i of textual modality, in predicting the output (where s_i represents i^{th} segment of the feature vector from the respective modality). In Figure 7a, for the first segment of the utterance (*i.e.,* s_1) of text modality, the model puts more attention weights to different segments of the utterance (*i.e.,* s_6, s_7, s_9, and s_{10}) of visual modality to classify the

Setup	Model	T + V			T + A			A + V			T + A + V		
		P	R	F1	P	R	F1	P	R	F1	P	R	F1
Speaker Dependent	***Baseline***	72.0	71.6	71.6	**66.6**	**66.2**	**66.2**	66.2	65.7	65.7	71.9	71.4	71.5
	Proposed Model	**72.8**	**71.9**	**71.6**	62.2	61.2	59.6	**72.7**	**71.9**	**71.8**	**73.4**	**72.8**	**72.6**
	T-test	-	-	-	-	-	-	-	-	-	0.0023	0.0098	0.0056
Speaker Independent	***Baseline***	62.2	61.5	61.7	**64.7**	**62.9**	**63.1**	64.1	61.8	61.9	64.3	62.6	62.8
	Proposed Model	**65.5**	**65.5**	**65.7**	59.1	60.0	50.3	**65.6**	**63.8**	**63.9**	**69.5**	**66.0**	**65.9**
	T-test	-	-	-	-	-	-	-	-	-	0.0002	0.0006	0.0012

Table 11: Comparative Analysis of the proposed approach with recent state-of-the-art systems. We evaluated on extended, publicly available MUSTARD dataset (Castro et al., 2019). Significance test p-values< 0.05

	Utterances	Sarcasm (T+A+V)		
		Actual	STL	MTL
1	*Oh yeah ok, including the waffles last week, you now owe me, seventeen zillion dollars.*	S	NS	S
2	*I love that you take pride in your looks, even when I have to pee in the morning, and you're in there spending an hour on your hair.*	S	NS	S
3	*Now?! - No, after my tongue has swollen to the size of a brisket!*	S	S	S
4	*There's no hurry. Tell them more about their secret love for each other.*	NS	S	NS
5	*I'm not saying that you're not fun. You're the most fun person I know.*	NS	S	NS
6	*Well, I'm sorry, too, but there's just no room for you in my wallet.*	S	S	S

Table 12: Comparison between multi-task learning (Sar + Sent + Emo) and single-task learning (Sar) frameworks for tri-modal (T+A+V) inputs. Few error cases where MTL framework performs better than the STL framework.

utterance correctly. Similarly, for I_a-Attention, each cell(*i,j*) of the heatmap (c.f. Figure 8) signifies the influence of s_j on s_i in predicting the output (where s_i represents i^{th} segment of the concatenated feature vector from all modalities). We observe that for a particular segment of the utterance (*say* s_6), the model puts more weights to itself rather than the others.

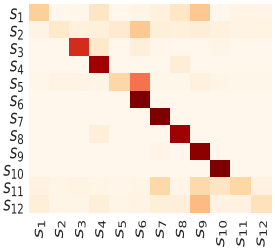


Figure 8: I_a-*Attention*.

We also observe that in the bi-modal (T+A) experiment (c.f. Table 4) our model does not perform at par for the three tasks, (i.e. sarcasm, sentiment, and emotion) together. This may be attributed to the reason of not incorporating the visual information that contains rich affect cues in the forms of sentiment and emotion. Hence, the introduction of sentiment in the T+A setting might be confusing the model.

6 Conclusion

In this paper, we have proposed an effective deep learning-based multi-task model to simultaneously solve all the three problems, *viz.* sentiment analysis, emotion analysis and sarcasm detection. As there was no suitable labeled data available for this problem, we have created the dataset by manually annotating an existing dataset of sarcasm with sentiment and emotion labels. we have introduced two attention mechanisms *(i.e., I_e-Attention and I_a-Attention)*, and incorporated the significance of context and speaker information *w.r.t.* sarcasm. Empirical evaluation results on the extended version of the MUStARD dataset suggests the efficacy of the proposed model for sarcasm analysis over the existing state-of-the-art systems. The evaluation also showed that the proposed multi-tasking framework achieves better performance for the primary task, i.e. sarcasm detection, with the help of emotion analysis and sentiment analysis, the two secondary tasks in our setting.

During our analysis, we found that the dataset is not big enough for a complex framework to learn from. Along with investigating new techniques, we hope that assembling a bigger curated dataset with quality annotations will help in better performance.

7 Acknowledgement

The research reported here is partially supported by SkyMap Global India Private Limited. Asif Ekbal acknowledges the Young Faculty Research Fellowship (YFRF), supported by Visvesvaraya PhD 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. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 370–379, Minneapolis, Minnesota. Association for Computational Linguistics.

Santiago Castro, Devamanyu Hazarika, Verónica Pérez-Rosas, Roger Zimmermann, Rada Mihalcea, and Soujanya Poria. 2019. Towards multimodal sarcasm detection (an _obviously_ perfect paper). *arXiv preprint arXiv:1906.01815*.

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.

KyungHyun Cho, Bart van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder-decoder approaches. *CoRR*, abs/1409.1259.

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.

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.

Aditya Joshi, Pushpak Bhattacharyya, and Mark J. Carman. 2017. Automatic sarcasm detection: A survey. *ACM Comput. Surv.*, 50(5):73:1–73:22.

Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. 2016. Fasttext.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*.

Sijie Mai, Haifeng Hu, and Songlong Xing. 2019. Divide, conquer and combine: Hierarchical feature fusion network with local and global perspectives for multimodal affective computing. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 481–492.

Soujanya Poria, Erik Cambria, Devamanyu Hazarika, and Prateek Vij. 2016. A deeper look into sarcastic tweets using deep convolutional neural networks. *arXiv preprint arXiv:1610.08815*.

Tony Veale and Yanfen Hao. 2010. Detecting ironic intent in creative comparisons. In *ECAI*, volume 215, pages 765–770.