All-in-One: A Deep Attentive Multi-task Learning Framework for Humour, Sarcasm, Offensive, Motivation, and Sentiment on Memes

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 aim at learning the relationships and similarities of a variety of tasks, such as humour detection, sarcasm detection, offensive content detection, motivational content detection and sentiment analysis on a somewhat complicated form of information, *i.e.,* memes. We propose a multi-task, multi-modal deep learning framework to solve multiple tasks simultaneously. For multi-tasking, we propose two attention-like mechanisms *viz.,* Inter-task Relationship Module (*iTRM*) and Inter-class Relationship Module (*iCRM*). The main motivation of *iTRM* is to learn the relationship between the tasks to realize how they help each other. In contrast, *iCRM* develops relations between the different classes of tasks. Finally, representations from both the attentions are concatenated and shared across the five tasks (i.e., humour, sarcasm, offensive, motivational, and sentiment) for multi-tasking. We use the recently released dataset in the Memotion Analysis task @ SemEval 2020, which consists of memes annotated for the classes as mentioned above. Empirical results on Memotion dataset show the efficacy of our proposed approach over the existing state-of-the-art systems (Baseline and SemEval 2020 winner). The evaluation also indicates that the proposed multi-task framework yields better performance over the single-task learning.

1 Introduction

The content and form of content shared on online social media platforms have changed rapidly over time. Currently, one of the most popular forms of media shared on such platforms is *'Memes'*. According to its definition from Oxford Dictionary, a meme is a piece of data, often in the form of images, text or videos that carry cultural information through an imitable phenomenon with a mimicked theme, that is shared (sometimes with slight modification) rapidly by internet users.

Every meme can be associated with five affect values, namely *humour* (Hu), *sarcastic* (Sar), *offensive* (Off), *motivational* (Mo), and *sentiment* (Sent). Hence, in a broad sense, memes can be categorized into four *intersecting* sets *viz.* humorous memes, sarcastic memes, offensive memes, and motivational memes.

Humour refers to the quality of being amusing or comic. Formally, humour is defined as the nature of experiences to induce laughter and provide amusement. Humourous memes are the most popular and widely used on social media platforms. An example for humourous memes is shown in Figure 1a.

Sarcasm is often used to convey thinly veiled disapproval humorously. A sarcastic meme is a meme where an incongruity exists between the intended meaning and the way it is expressed. These are generally used to express dissatisfaction or to veil insult through humour. As we can see in Figure 1a, the person on the right is made fun of, without explicitly expressing it, which is a typical example of a sarcastic meme.

Offensive content include a lot of insulting, derogatory terms. It is contrary to the moral sense or good. As social media expands, offensive language has become a huge headache to maintain sanity on social media. As memes are growing to become more and more popular, detecting offensive memes on such platforms is becoming an important and challenging task. Figure 1a, Figure 1c and Figure 1d are the instances of Offensive memes.

Motivation is derived from the word 'motive' which means needs or desires within the individuals. It is the process of stimulating people to actions to achieve their goals. By its definition, motivational memes are those that benefit a certain group of people to achieve their plans or goals. Motivation can be both either positive or negative.

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
December 4 - 7, 2020.

(a) *Humour, sarcasm, offensive*. (b) *Motivational, positive*. (c) *Sarcasm, offensive, Negative*. (d) *Sarcasm, offensive, Funny*.

Figure 1: Few examples from the *Memotion* dataset to show the inter-dependency between different tasks.

However, we usually consider motivation in a positive sense. Figure 1b is an excellent example for the positive motivation.

Sentiment analysis refers to the process of computationally identifying and categorizing opinions expressed in a piece of communication, especially to determine whether the writer's attitude towards a particular topic, product, etc. is *positive, negative*, or *neutral*. This has been a very prominent and important task in Natural Language Processing. Sentiment analysis on memes refers to the task of systematically extracting its emotional tone in understanding the opinion expressed by the meme. Figure 1b is an example for positive sentiment towards the government and Figure 1c for negative sentiment towards *Ph.D.* in Electrical Engineering.

Generally, specific labels of one task have a strong relation to the other labels of sarcasm, offensive, humour or motivational tasks. Through proper representation, training, and evaluation, these relations can be modelled to help each other for better classification. For example, in Figure 1b, just by seeing text, the meme can be either sarcastic or motivational, but the image in the meme confirms that this has an overall positive sentiment and hence motivational. Similarly, in Figure 1c, knowing that the meme is sarcastic and has a negative sentiment makes it highly probable to being offensive.

As seen above, humorous, motivational, offensive, and sarcastic nature of the memes are closely related. Thus, a multi-task learning framework would be extremely beneficial in such scenarios. In this paper, we exploit these relationships and similarities in the tasks of humour detection, sarcasm detection, offensive content detection, motivational content detection, and sentiment in a multi-task manner. The main contributions and/or attributes are as follows: **(a).** *We propose a multi-task multi-modal deep learning framework to leverage the utility of each task to help each other in a multi-task framework;* ***(b).*** *We propose two attention mechanisms viz. iTRM and iCRM to better understand the relationship between the tasks and between the classes of tasks, respectively; and* ***(c).*** *We present the state-of-the-art results for meme prediction in the multi-modal scenario.*

2 Related Work

Sentiment analysis and its related tasks, such as humour detection, sarcasm detection, and offensive content detection, are the topics of interest due to their needs in recent times. There has been a phenomenal growth in multi-modal information sources in social media, such as audio, video, and text. Multi-modal information analysis has attracted the attention of researchers and developers due to their complexity, and multi-tasking has been of keen interest in the field of affect analysis.

Humour: Early feature-based models attempt to solve humour include the models based on word overlap with jokes, presence of ambiguity, and word overlap with common idioms (Sjöbergh and Araki, 2007), human-centeredness, and negative polarity (Mihalcea and Pulman, 2007). Some of the recent multi-modal approaches include utilizing information from the various modalities, such as acoustic, visual, and text, using deep learning models (Bertero and Fung, 2016; Yang et al., 2019; Swamy et al., 2020). Yang et al. (2020) employs a paragraph decomposition technique coupled with fine-tuning BERT (Devlin et al., 2018) model for humour detection on three languages (Chinese, Spanish and Russian).

Sarcasm: Starting from the traditional approaches, such as rule-based methods (Veale and Hao, 2010), lexical features (Carvalho et al., 2009), and incongruity (Joshi et al., 2015) to all the way up to multi-modal deep learning techniques (Schifanella et al., 2016), sarcasm detection has been showing its presence. Castro et al. (2019) created a multi-modal conversational dataset, MUStARD from the famous TV shows, and provided baseline SVM approaches for sarcasm detection. Recently, Chauhan et al. (2020) proposed a multi-task learning framework for multi-modal sarcasm, sentiment and emotion analysis to explore how sentiment and emotion helps sarcasm. The author used the MUStARD dataset and extended the MUStARD dataset with *sentiment* (implicit and explicit) and *emotion* (implicit and explicit) labels.

Offensive: Razavi et al. (2010) used a three-level classification model taking advantage of various features from statistical models and rule-based patterns and various dictionary-based features. Chen et al. (2012) proposed a feature-based Lexical Syntactic Feature (LSF) architecture to detect the offensive contents. Gomez et al. (2020) created a multi-modal hate-speech dataset from Twitter *(MMHS150K)* to introduce a deep-learning-based multi-modal Textual Kernels Model (TKM) and compare it with various existing deep learning architectures on the proposed MMHS150K dataset.

Motivation: Swieczkowska et al. (2020) proposes a novel chaining method of neural networks for identifying motivational texts where the output from one model is passed on to the second model.

Sentiment: An important task to leverage multi-modality information effectively is to combine them using various strategies. Mai et al. (2019) employs a hierarchical feature fusion strategy, *Divide, Conquer*, and *Combine* for affective computing. Chauhan et al. (2019) uses the Inter-modal Interaction Module (IIM) to combine information from a pair of modalities for multi-modal sentiment and emotion analysis. Some of the other techniques include a contextual inter-modal attention based framework for multi-modal sentiment classification (Ghosal et al., 2018; Akhtar et al., 2019).

Multi-task: Some of the early attempts to correlate the tasks like sarcasm, humour, and offensive statements include a features based classification using various syntactic and semantic features, such as frequency of words, the intensity of adverbs and adjectives, the gap between positive and negative terms, the structure of the sentence, synonyms and others (Barbieri and Saggion, 2014). More recently, Badlani et al. (2019) proposed a convolution-based model to extract the embedding by fine-tuning the same for the tasks of sentiment, sarcasm, humour, and hate-speech and then concatenating these representations to be used in a sentiment classifier.

In our current work, we propose a multi-task multi-modal deep learning framework to simultaneously solve the tasks of sarcasm, humour, offensive, and motivational on memes. Further, to the best of our knowledge, this is the very first attempt at solving the multi-modal affect analysis on memes in a multi-task deep learning framework. We demonstrate through a detailed empirical evaluation that a multi-task learning framework can improve the performance of individual tasks over a single task learning framework.

3 Proposed Methodology

We propose an attention-based deep learning model to solve the problem of multi-task affect analysis of memes. The inputs to the model are the meme itself and the manually corrected text extracted through OCR. The overall architecture is depicted in Figure 2. The source code is available at `http://www.iitp.ac.in/~ai-nlp-ml/resources.html`.

3.1 Input Layer:

We now describe the input features for our proposed model.

3.1.1 Text Input

Given N number of samples, where each sample is associated with meme image and the corresponding text. Let us assume, in each sample, there are n_T number of words $w_{1:n_T} = w_1, ..., w_{n_T}$, where $w_j \in \mathbb{R}^{d_T}$, d_T = 768, and w_j is obtained using *BERT* (Devlin et al., 2018). The maximum number of words for i^{th} sample across the dataset is 189.

3.1.2 Image Input

Image is the prime component of any meme and contains the majority of the information. To leverage this information effectively, feature vectors from average pooling layer (avgpool) of the ImageNet pre-trained *ResNet-152* (He et al., 2016) image classification model are extracted. Each image is first pre-processed by resizing to 224×224 and then normalized. The extracted feature vector for image of i^{th} sample is represented by $V_i \in \mathbb{R}^{d_v}$ and d_v = 2048.

3.2 Attention Modules

These vectors are concatenated and then passed through a set of four dense layers to obtain the vectors of equal length d represented by $TV_t \in \mathbb{R}^d$,

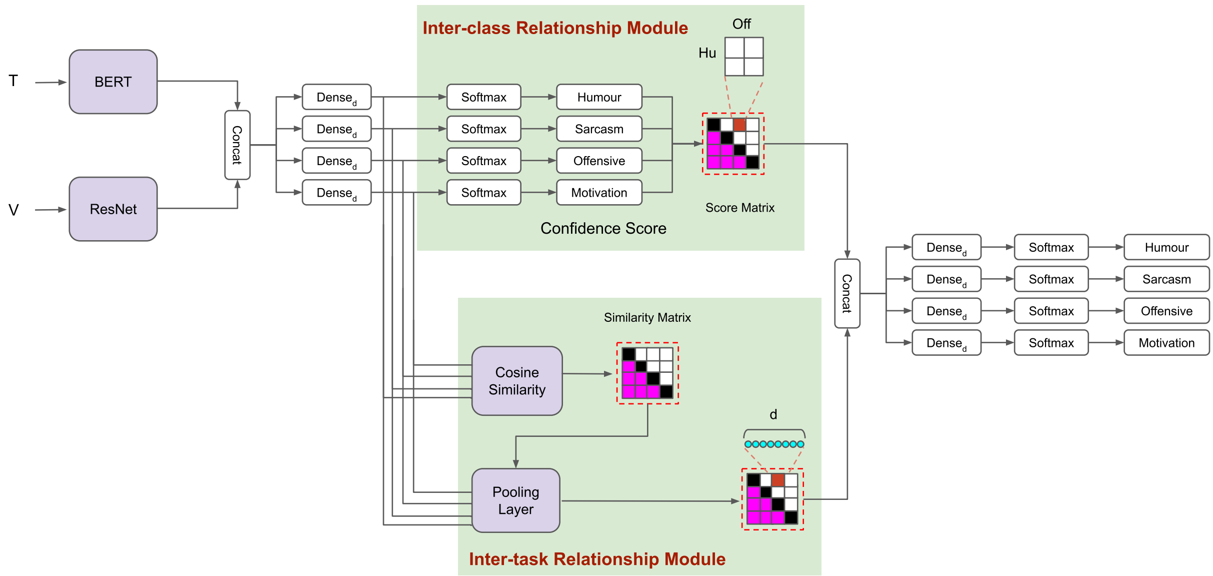


Figure 2: Overall architecture of the proposed multi-modal multi-task affect analysis framework for Memes. Here **V** refers to the *Meme Image* and **T** refers to the *text extracted from the Meme*.

where t is a task $\in$ {humour, sarcasm, offensive, motivational}. These vectors are then passed through the Inter-class Relationship Module and Inter-task Relationship module. The output is then concatenated and passed through another set of four dense layers, and a layer of softmax is applied to obtain the final output.

3.2.1 Inter-class Relationship Module

This module is used to learn the relationship between the classes of all the tasks. This is done by passing TV_t through another dense layer and softmax (*confidence score*). For each task, we first group all the classes into two classes for the hierarchical classification of the sample. At this level, the sample is labelled with either positive or negative for all the tasks. For instance, a sample will be labelled as either sarcastic or not_sarcastic for sarcasm tasks. A loss is back-propagated using these confidence scores for the corresponding tasks. This is done in order to control each dense layer so that it aligns with the respective tasks. Meanwhile, a dot-product of the softmax scores of each task is obtained and used to form the *Score Matrix*. This is then flattened and passed forward.

3.2.2 Inter-task Relationship Module

While the above module is used to find the correlation between the individual classes, this module is used to find the relationship between the different tasks in the model. This is done by initially finding the cosine-similarity between TV_t vectors. And a pooling layer is used to collect information between the tasks and then normalized by the corresponding cosine-similarity score. The output from the pooling layer is then flattened and passed forward.

3.3 Output Unit

The flattened vectors from $iTRM$ and $iCRM$ are concatenated and then branched into four dense layers for each task. This is then forwarded through a softmax layer to obtain the final output for each task, and the loss is back-propagated to learn the parameters. In this layer, the information from both *iCRM* and *iTRM* modules will be leveraged and used to predict the final outcome.*Please note that, there are two sets of loss used in the model, one in the iCRM module and second at the end the of Output Unit.*

4 Dataset

We perform experiments using the dataset released in the Memotion Analysis 1.0 @SemEval 2020 Task (Sharma et al., 2020)[1]. This dataset consists of 6992 samples. Each sample consists of an image, corrected text extracted from the meme, and the five labels associated with the five tasks, *viz.*, *Humour, Sarcasm, Offensive, Motivational, and Overall Sentiment*. The distribution of the classes associated with each of the five tasks with label is shown in Table 1 and Table 2.

[1] https://competitions.codalab.org/competitions/20629

Task	*Classes*	*Count*	*RC (%)*	*T-A*
Sent	very_negative	1033	17.34	N_g
	negative	3127	52.48	
	neutral	2201	36.94	N_u
	positive	480	8.06	P_s
	very_positive	151	2.53	

Table 1: Dataset Distribution of Task-A, where *RC* and *T-A* denotes the relative count and abbreviation for labels of Task-A, respectively.

Task	*Classes*	*Count*	*RC (%)*	*T-C*	*T-B*
Hu	not_funny	1651	30.91	N_f	N_h
	funny	2452	45.91	F_n	H_m
	very_funny	2238	41.90	V_f	
	hilarious	651	12.19	H_r	
Sar	not_sarcastic	1544	22.08	N_s	N_s
	general	3507	50.16	G_r	S_r
	twisted_meaning	1547	22.13	T_m	
	very_twisted	394	5.64	V_t	
Off	not_offensive	2713	38.80	N_o	N_o
	slight	2592	37.07	S_g	O_f
	very_offensive	1466	20.97	V_o	
	hateful_offensive	221	3.16	H_o	
Mo	not_motivational	4525	64.72	N_m	N_m
	motivational	2467	35.28	M_o	M_o

Table 2: Dataset Distribution of Task-B and Task-C, where *RC*, *T-B* and *T-C* denotes the relative count, abbreviation for labels of Task-B, and abbreviation for labels of Task-C respectively.

We address 5 multi-modal affective analysis problems, namely *humour classification, sarcasm classification, offensive classification, motivational classification*, and *sentiment classification*.

A. Humour classification: There are four classes associated with the humour task, namely not_funny, funny, very_funny, and hilarious, which are labelled as 0, 1, 2, and 3, respectively.

B. Sarcasm classification: There are four classes associated with the sarcasm task, namely not_sarcastic, general, twisted_meaning, and very_twisted which are labelled as 0, 1, 2, and 3 respectively.

C. Offensive classification: There are four classes associated with the offensive task, namely not_offensive, slight, very_offensive, and hateful_offensive which are labelled as 0, 1, 2, and 3, respectively.

D. Motivational classification: There are two classes associated with the motivational task, namely not_motivational and motivational, which are labelled as 0 and 1, respectively.

E. Sentiment classification: There are five classes associated with the sentiment task, namely very_negative, negative, neutral, positive, and very_positive, which are labelled as 0, 1, 2, 3, and 4, respectively.

5 Experimental setup

In accordance with the SemEval 2020 (Sharma et al., 2020), the project is organized into three sets of tasks[2].

- **Task A: Sentiment Classification:** In this task, memes are classified into 3 classes *viz.*, -1 (negative, very_negative), 0 (neutral) and +1 (positive, very_positive).

- **Task B: Binary-class Classification:** In this set of tasks, the memes are classified as follows (c.f. T-B in Table 2);

  1. **Humour** (funny, very_funny, hilarious) and Non-humour (not_funny).
  2. **Sarcasm** (general, twisted_meaning, very_twisted) and Non-sarcasm (non_sarcastic)
  3. **Offensive** (slight, very_offensive, hateful_offensive) and Non-Offensive (not_offensive), and
  4. **Motivational** (motivational) and Non-motivational (not_motivational).

- **Task C: Multi-class Classification:** In this set of task, the original labels are used as described in the dataset (c.f. T-C in Table 2) for the tasks of Humour, Sarcasm, Offensive and Motivational.

Please note that, in Task A, as it is not a multi-task scenario, *iCRM* and *iTRM* are not applicable. For all the other sets of tasks, the entire network is shown in Figure 2.

We evaluate our proposed model on the multi-modal Memotion dataset. 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.

[2] `https://competitions.codalab.org/competitions/20629#learn_the_details-task-labels-format`

Parameters	*Task-A*	*Task-B*	*Task-C*
Activations	*ReLu*		
Optimizer	*Adam (lr=0.001)*		
Output	*Softmax*		
Loss	*Categorical cross-entropy*		
Batch	*16*		
Epochs	*30*		
Dropout-p	0.3	0.5	0.7
#*neurons*(Dense)	50	200	200

Table 3: Model configurations

We implement our proposed model on the open source machine learning library PyTorch[3]. Hugging Face[4] library is used for BERT implementation. As the evaluation metric, we employ precision (P), recall (R), macro-F1 ($\mathrm{M_a}$-F1), and micro-F1 ($\mathrm{M_i}$- F1) for all the tasks *i.e., humour, sarcasm, offensive, motivational, and sentiment*. We use *Adam* as an optimizer, *Softmax* as a classifier, and the *categorical cross-entropy* as a loss function for all the tasks.

6 Results and Analysis

We evaluate our proposed architecture with bimodal inputs *(i.e., text and visual)*. We show the obtained results for Task-A *(i.e., sentiment analysis)* in Table 4.

Labels	*Task-A*			
	P	**R**	**$\mathrm{M_a}$-F1**	**$\mathrm{M_i}$-F1**
Sentiment	36.99	35.70	35.81	50.58

Table 4: Memes: Sentiment Classification (Task A)

Task-B has four different tasks, *i.e., humour, sarcasm, offensive, and sentiment* with binary-class labels (c.f. binary-class classification in Section 5). The results are shown in Table 5.

Labels	*Task-B (Binary Classification)*							
	STL				**MTL**			
	P	**R**	**$\mathrm{M_a}$-F1**	**$\mathrm{M_i}$-F1**	**P**	**R**	**$\mathrm{M_a}$-F1**	**$\mathrm{M_i}$-F1**
Hu	55.44	53.77	53.74	71.29	**55.52**	**53.84**	**53.84**	**71.29**
Sa	51.94	51.34	50.98	70.76	**52.99**	**52.48**	**52.52**	**70.94**
Of	**52.33**	**52.19**	**52.13**	**56.28**	51.35	51.37	51.36	54.10
Mo	53.56	53.49	53.51	57.18	**55.86**	**56.44**	**56.12**	**57.44**

Table 5: Memes: Single-task vs Multi-task (Task B)

Task-C has also four different tasks, *i.e., humour, sarcasm, offensive, and sentiment* with multi-class labels (c.f. multi-class classification in Section 5). The results are shown in Table 6.

[3]https://pytorch.org/

[4]https://github.com/huggingface/transformers

Labels	*Task-C (Multi-class Classification)*							
	STL				**MTL**			
	P	**R**	**$\mathrm{M_a}$-F1**	**$\mathrm{M_i}$-F1**	**P**	**R**	**$\mathrm{M_a}$-F1**	**$\mathrm{M_i}$-F1**
Hu	26.83	26.89	26.75	29.76	**27.23**	**27.29**	**27.03**	**32.00**
Sa	25.16	26.71	25.74	36.52	**26.30**	**27.33**	**26.80**	**39.94**
Of	**27.21**	**27.30**	**26.93**	35.30	25.05	26.04	25.53	**35.94**
Mo	53.32	52.89	52.65	58.46	**54.14**	**53.31**	**53.72**	**59.79**

Table 6: Memes: Single-task vs Multi-task (Task C)

In both the tasks B and C, we outline the comparison between the multi-task (MTL) and single-task (STL) learning frameworks in Table 5 and Table 6. We observe that MTL shows better performance over the STL setups.

For the offensive task, we find that STL performs better than MTL. We hypothesize that this is due to the model getting confused between the offensive and sarcastic (or humorous) memes. From Table 9, under Sarcasm, we can see that for the class V_t, MTL predicts a few samples as sarcastic, whereas in actuality it belongs to the other classes. However, we can see a decrease in performance for class H_o under Offensive. This is due to the lack of a larger dataset for the complex model to disambiguate the same. In the example, *BRB...GOT TO TAKE CARE OF SOME SH*T IN UKRAIN* (c.f. Figure 1d), the actual set of labels are F_n, G_n, S_g, N_m. The predicted labels in STL are V_f, G_n, S_g, M_o and in MTL are V_f, T_m, V_o, M_o. This is supposed to be slightly offensive but got it confused with the sarcastic.

7 Comparative Analysis

We compare the results obtained in our proposed model against the baseline model and SemEval 2020 winner, which also made use of the same dataset. The comparative analysis is shown in Table 7. Our proposed multi-modal framework achieves the best macro-F1 of 35.8% (0.4% $\uparrow$) and micro-F1 of 50.6% (1.9% $\uparrow$) as compared to macro-F1 of 35.4% and micro-F1 of 48.7% of the state-of-the-art system (i.e., SemEval 2020 Winner) for Task-A. Similarly, for Task-B, we obtain the macro-F1 of 53.5% (1.7% $\uparrow$) and micro-F1 of 63.4% (2.0% $\uparrow$) as compared to the macro-F1 of 51.8% and micro-F1 of 61.4% of the state-of-the-art system, whereas for Task-C, we obtain the macro-F1 of 33.3% (1.1% $\uparrow$) and micro-F1 of 41.9% (4.1% $\uparrow$) as compared to the macro-F1 of 32.2% and micro-F1 of 37.8% of the state-of-the-art system.

It is evident from Table 5 and Table 6 that multi-task learning framework successfully leverages the

Systems	Task A		Task B		Task C	
	M_a-F1	M_i-F1	M_a-F1	M_i-F1	M_a-F1	M_i-F1
Baseline	21.76	30.77	50.02	56.86	30.08	33.28
SE'20 Winner	35.46	48.72	51.83	61.44	32.24	37.79
Proposed	**35.81**	**50.58**	**53.46**	**63.44**	**33.27**	**41.92**

Table 7: Comparative Analysis of the proposed approach with recent state-of-the-art systems. Here, SE'20 denotes the SemEval 2020 winner, and 'Proposed' refers to the models described in the paper for the respective tasks.

Sentiment			
	N_g	N_u	P_s
N_g	17	19	127
N_u	25	170	399
P_s	58	290	763

(a) Task-A

Setups	*Humour*			*Sarcasm*			*Offensive*			*Motivational*		
		N_h	H_m		N_s	S_r		N_o	O_f		N_m	M_o
STL	N_h	91	354	N_s	68	353	N_o	252	455	N_m	801	387
	H_m	185	1248	S_a	196	1261	O_f	366	805	M_o	417	273
MTL	N_h	92	353	N_s	90	331	N_o	285	422	N_m	801	387
	H_m	186	1247	S_a	239	1218	O_f	440	731	M_o	431	259

(b) Task-B

Table 8: Confusion Matrix for Task-A and Task-B (Refer Table 1 and Table 2 for Label definitions).

Setups	*Humour*					*Sarcasm*					*Offensive*					*Motivational*		
		N_f	F_n	V_f	H_r		N_s	G_r	T_m	V_t		N_o	S_g	V_o	H_o		N_m	M_o
STL	N_f	122	143	130	50	N_s	117	182	122	0	N_o	254	307	111	35	N_m	878	310
	F_n	140	218	205	91	G_r	234	427	276	0	S_g	224	340	105	40			
	V_f	129	201	193	82	T_m	94	188	142	0	V_o	109	198	62	18	M_o	470	220
	H_r	36	65	47	26	V_t	19	52	25	0	H_o	20	37	11	7			
MTL	N_f	147	147	136	21	N_s	125	206	87	3	N_o	350	219	138	0	N_m	924	264
	F_n	173	240	208	33	G_r	222	525	172	18	S_g	330	250	129	0			
	V_f	172	195	204	34	T_m	112	210	100	2	V_o	181	131	75	0	M_o	491	199
	H_r	51	70	43	10	V_t	23	57	16	0	H_o	43	22	10	0			

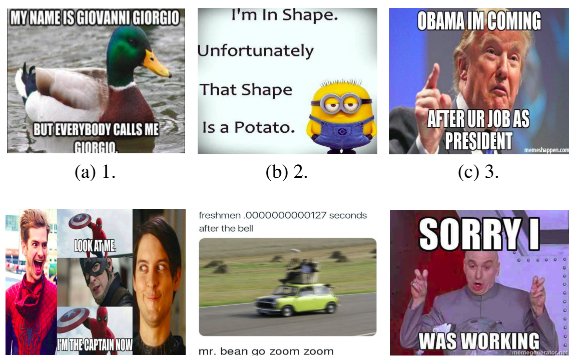
Table 9: Confusion Matrix for Task C (Refer Table 2 for Label definitions).

inter-dependence between all the tasks in improving the overall performance in comparison to the single-task learning. We also show the confusion matrices corresponding to each set of tasks in Table 8a, Table 8b, and Table 9, respectively.

8 Error Analysis

We perform error analysis (i.e. for Task-C) on the predictions of our proposed model. We take some utterances (c.f. Table 10) with corresponding image (c.f. Figure 3), where we show that *MTL* is predicting correct while *STL* is not able to predict the right labels.

We also present the attention heatmaps for *iCRM* and *iTRM* of the multi-task learning framework in Figure 4 and Figure 5, respectively. We take the fifth utterance from Table 10 (c.f. Figure 3e) to illustrate the heatmap. For *iCRM* (c.f. Figure 4), there are six matrices which show the inter-dependency between humour and sarcasm *(Hu-Sar)*, humour and offensive *(Hu-Off)*, humour and motivational *(Hu-Mo)*, sarcasm and offensive *(sar-off)*, sarcasm and motivational *(Sar-Mo)*, and offensive and motivational *(Off-Mo)*, respectively, where the light shade to dark shade shows the amount of contributions in ascending sequence.



(a) 1. (b) 2. (c) 3. (d) 4. (e) 5. (f) 6.

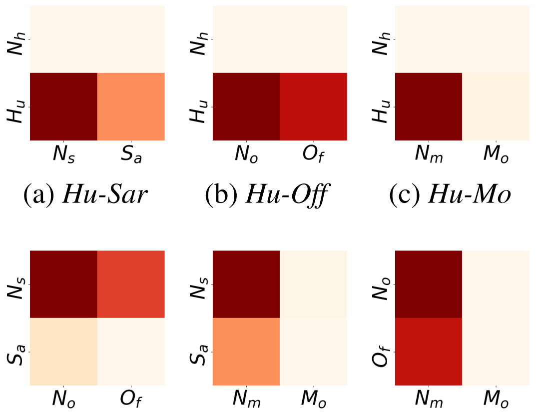
Figure 3: Few examples for Human Error Analysis corresponding to Table 10.

The main objective of *iCRM* is to develop the relationship between the classes of tasks. Figure 4 shows the established relationship between the tasks. We see the established relationship between the classes of tasks in Figure 4. For predicting the fifth utterance correctly in Table 10, humour and

	Utterances	*STL* Hu	Sar	Off	Mo	*MTL* Hu	Sar	Off	Mo
1	*my name is giovanni giorgio but everybody calls me giorgio.*	N_f	G_r	N_o	N_m	V_f	T_m	V_o	M_o
2	*i'm in shape. unfortunately that shape is a potato*	V_f	N_s	N_o	M_o	F_n	G_r	S_g	N_m
3	*obama i'm coming after ur job as president memeshappen.Com*	F_n	G_r	N_o	N_m	V_f	T_m	V_o	M_o
4	*look at me I'm the captain now.*	V_f	T_m	V_o	M_o	F_n	G_r	S_g	N_m
5	*freshmen .0000000000127 seconds after the bell mr. bean go zoom zoom.*	H_r	N_s	S_g	M_o	F_n	G_r	M_o	N_m
6	*sorry i was working.*	F_n	T_m	V_o	M_o	V_f	G_r	S_g	N_m

Table 10: Comparison between multi-task learning and single-task learning frameworks .Few error cases where MTL framework performs better than the STL framework.

not sarcasm (Figure 4a), humour and not offensive (Figure 4b) etc. are helping each other.



(a) *Hu-Sar* (b) *Hu-Off* (c) *Hu-Mo*

(d) *Sar-Off* (e) *Sar-Mo* (f) *Off-Mo*

Figure 4: *iCRM* attention for Figure 3e under Task C

Similarly, the main objective of *iTRM* is to develop the relationship between the tasks. Figure 5 shows the established relationship between the tasks, and we see that attention put more weight on sarcasm and offensive pair while less weight on humour and sarcasm. It is clear from the definition of sarcasm and humour (c.f. Section 1) that both of them have a very different meaning when used in a sentence while the actual sentence looks similar. Hence sarcasm and humour is found not be helping each other.

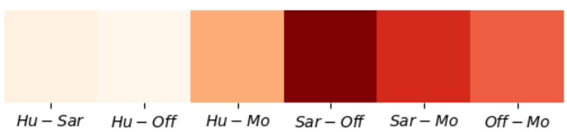


Figure 5: *iTRM* attention for Figure 3e under Task C

9 Conclusion and Future Work

In this paper, we have successfully established the concept of obtaining effective relationships between inter-tasks and between inter-classes for multi-modal affect analysis. We have proposed a deep attentive multi-task learning framework which helps to obtain very effective inter-tasks and inter-classes relationship. To capture the interdependence, we have proposed two attention-like mechanisms *viz.,* Inter-task Relationship Module (*iTRM*) and Inter-class Relationship Module (*iCRM*). The main motivation of *iTRM* is to learn the relationship between the tasks, i.e. which task helps the other tasks. In contrast, *iCRM* develops the relations between the classes of tasks. We have evaluated our proposed approach on a recently published Memotion dataset. Experimental results suggest the efficacy of the proposed model over the existing state-of-the-art systems (Baseline and SemEval 2020 winner). The evaluation shows that the proposed multi-task framework yields better performance over single-task learning.

The dataset used for the experiments is relatively small for training an effective deep learning model and is heavily biased. Therefore, assembling a large, and more balance dataset with quality annotations is an important job. Moreover, the memes are a complicated form of data which includes both text and image that repeat over numerous memes (meme templates). Hence quality representation of memes for affect analysis is challenging future work.

Acknowledgement

The research reported here is partially supported by SkyMap Global India Private Limited. Dushyant Singh Chauhan acknowledges the support of Prime Minister Research Fellowship (PMRF), Govt. of India. 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 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.

Rohan Badlani, Nishit Asnani, and Manan Rai. 2019. Disambiguating sentiment: An ensemble of humour, sarcasm, and hate speech features for sentiment classification. *W-NUT 2019*, page 337.

Francesco Barbieri and Horacio Saggion. 2014. Automatic detection of irony and humour in twitter. In *ICCC*, pages 155–162.

Dario Bertero and Pascale Fung. 2016. Multimodal deep neural nets for detecting humor in tv sitcoms. In *2016 IEEE Spoken Language Technology Workshop (SLT)*, pages 383–390. IEEE.

Paula Carvalho, Luís Sarmento, Mário J Silva, and Eugénio De Oliveira. 2009. Clues for detecting irony in user-generated contents: oh...!! it's" so easy";-. In *Proceedings of the 1st international CIKM workshop on Topic-sentiment analysis for mass opinion*, pages 53–56.

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.

Dushyant Singh Chauhan, Dhanush S R, Asif Ekbal, and Pushpak Bhattacharyya. 2020. 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.

Ying Chen, Yilu Zhou, Sencun Zhu, and Heng Xu. 2012. Detecting offensive language in social media to protect adolescent online safety. In *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing*, pages 71–80. IEEE.

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*.

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.

Raul Gomez, Jaume Gibert, Lluis Gomez, and Dimosthenis Karatzas. 2020. Exploring hate speech detection in multimodal publications. In *The IEEE Winter Conference on Applications of Computer Vision*, pages 1470–1478.

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, Vinita Sharma, and Pushpak Bhattacharyya. 2015. Harnessing context incongruity for sarcasm detection. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 757–762.

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.

Rada Mihalcea and Stephen Pulman. 2007. Characterizing humour: An exploration of features in humorous texts. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 337–347. Springer.

Amir H Razavi, Diana Inkpen, Sasha Uritsky, and Stan Matwin. 2010. Offensive language detection using multi-level classification. In *Canadian Conference on Artificial Intelligence*, pages 16–27. Springer.

Rossano Schifanella, Paloma de Juan, Joel Tetreault, and Liangliang Cao. 2016. Detecting sarcasm in multimodal social platforms. In *Proceedings of the 24th ACM international conference on Multimedia*, pages 1136–1145.

Chhavi Sharma, Deepesh Bhageria, William Paka, Scott, Srinivas P Y K L, Amitava Das, Tanmoy Chakraborty, and Björn Gambäck. 2020. SemEval-2020 Task 8: Memotion Analysis-The Visuo-Lingual Metaphor! In *Proceedings of the 14th International Workshop on Semantic Evaluation (SemEval-2020)*, Barcelona, Spain. Association for Computational Linguistics.

Jonas Sjöbergh and Kenji Araki. 2007. Recognizing humor without recognizing meaning. In *International Workshop on Fuzzy Logic and Applications*, pages 469–476. Springer.

Steve Durairaj Swamy, Shubham Laddha, Basil Abdussalam, Debayan Datta, and Anupam Jamatia. 2020. Nit-agartala-nlp-team at semeval-2020 task 8: Building multimodal classifiers to tackle internet humor. *arXiv preprint arXiv:2005.06943*.

Patrycja Swieczkowska, Rafal Rzepka, and Kenji Araki. 2020. Stepwise noise elimination for better motivational and advisory texts classification. *Journal of Advanced Computational Intelligence and Intelligent Informatics*, 24(1):156–168.

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

Hao Yang, Yao Deng, Minghan Wang, Ying Qin, and Shiliang Sun. 2020. Humor detection based on paragraph decomposition and bert fine-tuning.

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.