Zombies Eat Brains, You are Safe: A Knowledge Infusion based Multitasking System for Sarcasm Detection in Meme

Anonymous ACL submission

Abstract

Sarcasm detection is, in itself, a challenging task in the field of Natural Language Processing (NLP), and the task even becomes more complex when the target is a meme. In this paper, we first hypothesize that sarcasm detection is closely associated with emotions present in the meme. We propose a deep learning-based multitask model to perform these two tasks in parallel, where sarcasm detection is the primary, whereas emotion recognition is considered as an auxiliary task. Furthermore, we propose a novel *knowledge infusion (KI)* method to get a sentiment-aware knowledge representation on top of our multitasking model. This sentiment-aware knowledge representation is obtained from a pre-trained parent model and subsequently this representation is used via a novel *Gating Mechanism* to train our downstream multitasking model. For training and evaluation purposes, we created a large-scale dataset consisting of 7416 sample Hindi memes as there was no readily available dataset for building such multimodal systems. We collect the Hindi memes from various domains, such as *politics*, *religious*, *racist*, and *sexist*, and manually annotate each instance with three sarcasm categories, i.e., *(i) Not Sarcastic*, *ii) Mildly Sarcastic* or *iii) Highly Sarcastic* and 13 fine-grained emotion classes. We demonstrate the effectiveness of our proposed work through extensive experiments. The experimental results show that our proposed system achieves a 64.48% macro F1-score, outperforming all the baseline models. Finally, we note that our proposed system is model agnostic and can be used with any downstream model in practice. We will make the resources and codes available[1]

1 Introduction

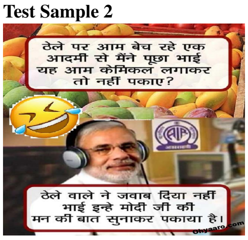
Social media platforms such as Facebook, Twitter, Instagram, etc., are interactive platforms that help in creating and sharing of information. The omnipresence of social media in the 21^{st} century established an enormous impact in different fields of society more powerfully and effectively. In day-to-day conversations, users make use of social media posts to convey dis-likeness towards a situation or a person with the help of sarcasm. Sarcasm is hard to understand because it usually uses humor in dialog (may also contain nonverbal cues) to show disapproval/dislike. Memes are the form of multimodal media that is becoming increasingly popular on the internet. It was initially created for humor purposes only. But due to the multimodality in nature, some memes help users to spread negativity in society in the form of sarcasm/dark humor. In the context of memes, detecting sarcasm is more difficult, as memes typically connect to a lot more background (or, contextual) information.

It can be easily depicted through the following examples. In example 1 of Figure 1, the meme says *"Bottles of Pepsi, Cola, Limca, Mirinda are kept in the fridge of my house, but all contain drinking water."*. In this example,the meme is serving its fundamental nature by spreading humor. The creator of this meme wants to spread joy with this meme. Therefore, we can easily infer positive sentiment associated with this meme.On the other hand, refer to example 2 of Figure 1, which is taken from the political domain. It says, *"While selling mangoes on a handcart, I asked a man, "brother, this mango is not ripe by giving chemicals." The vendor replied, "No, brother, it has been riped/annoyed after listening to Person-A's*[2] *Mann Ki Baat."* While we look at this meme from outer appearance, this can be seen that the meme was formed solely for humor purpose with no apparent twist. But, when we carefully analyze the emotion of the creator of the meme by adding the context knowledge, we

[1]Some samples of data, and the codes are available here:https://anonymous.4open.science/r/xxxxx-5222/

[2]To maintain the anonymity of any individual, we replaced actual name with Person-XYZ throughout the paper

Figure 1: Some samples from our dataset

	Test Sample 1	Test Sample 2
		
Sarcasm	Non-sarcastic	Highly sarcastic
Emotions	joy	joy,insult

observe that the meme creator is sarcastically targeting to offend Person-A. We can easily infer that the meme creator wants to insult the targeted person with the help of sarcasm. The meme creator wants to convey two obscure emotional states with the help of this meme, i.e., *insult* and *joy*. Additionally, we can infer a negative sentiment associated with the meme, amplified by the negative connotation present ('annoyed').

Given the above analysis, we observe that a trivial meme can be sarcastic too and we can be more certain of the sarcasm through the help of the associated emotions and the overall sentiment associated with the meme. Multi-modal input also helps us to understand the intent of the meme creator with more certainty. Thus with the help of multi-modal inputs and associated emotion and sentiment of the meme creator, detecting sarcasm in the meme can be an easier task. With these motivation in mind, in this paper, we propose a multitask model which can detect sarcasm in a meme with the help of emotion and sentiment. The key contributions of our work are summarized as follows:

- We create a high-quality and large-scale multimodal meme dataset annotated with sarcasm and 13 fine-grained emotion labels.
- We propose a multitasking model which simultaneously detects *sarcasm* and *emotions* in a given meme. Multitasking ensures that we exploit the emotion of the meme, which aids in detecting sarcasm more fluently. We also propose a gating mechanism denoted as knowledge infusion (KI) by which we leverage pre-trained sentiment-aware representation to our multitasking model.
- Empirical results show that the proposed KI method significantly outperforms the *naive* multimodal models.

2 Related Work

According to a literature review, a multimodal approach to sarcasm detection in memes is a relatively recent method rather than just text-based classification (Bouazizi and Tomoaki, 2016; Liu et al., 2019). (Tsur and Rappoport, 2009) proposed a semi-supervised framework for the recognition of sarcasm. They proposed a robust algorithm that utilizes features specific to (Amazon) product reviews. (Poria et al., 2016) developed pre-trained sentiment, emotion, and personality models to predict sarcasm on a text corpus through a Convolutional Neural Network, which effectively detects sarcasm. In a paper (Bouazizi and Tomoaki, 2016), researchers proposed four sets of features, i.e., sentiment-related features, punctuation-related features, syntactic and semantic features, and pattern-related features that cover the different types of sarcasm. Then, they used these features to classify tweets as sarcastic/non-sarcastic.

The use of multi-modal sources of information has recently gained significant attention to the researchers for affective computing. (Ghosal et al., 2018) proposed a recurrent neural network-based attention framework that leverages contextual information for multi-modal sentiment prediction. (Hasan et al., 2019) presented a new multi-modal dataset for humor detection called UR-FUNNY. It contains three modalities of text, vision, and acoustic. Researchers have also put their effort towards sarcasm detection in the direction of conversational AI(Joshi et al., 2016; Ghosh et al., 2017; Dong et al., 2020). For multimodal sarcasm detection in conversational AI, (Castro et al., 2019) created a new dataset, *MUStARD*, with high-quality annotations by including both multimodal and conversational context features. (Majumder et al., 2019) demonstrated that sarcasm detection could also be beneficial to sentiment analysis and designed a multitask learning framework to enhance the performance of both tasks simultaneously. Similarly, (Chauhan et al., 2020) has also shown that sarcasm can be detected with better accuracy when we know the *sarcasm* and *sentiment* of the speaker. In this paper we show that these multitasking approaches hold true in the domain of meme as well.

3 Resource Creation

3.1 Data collection

We inlined our data collection part with previous studies done on meme analysis(Sharma et al., 2020;

Kiela et al., 2020). We collect memes from various domains like politics, religion, social issues like terrorism, racism, sexism, etc. using a list of total 126 keywords like terrorism, beef ban, political memes, Ram Mandir-Babri Masjid, exams, Alok Nath memes, entertainment etc in hindi. All the memes were retrieved with the help of a browser extension called Download All Images[3] of Google's image search engine for all the collected unique keywords. We gathered memes that are freely available in the public domain to keep a strategic distance from any copyright issues. We have roughly 7k memes after deleting all the duplicates.

3.2 Data Pre-processing

The collected raw memes are (i) noisy such as background pictures are not clear, (ii) non-Hindi, i.e., meme texts are written in other languages except Hindi, and (iii) non-multi-modal, i.e., memes contain either text or visual content. Therefore, we manually discarded these memes to reduce manual data annotation effort. Next, we extracted the textual part of each meme using an open-source OCR tool: Tesseract[4] . The OCR errors are manually post-corrected by annotators. Finally, we considered 7,416 memes for data annotation.

3.3 Data Annotation

3.3.1 Sarcasm

We annotate each sample in the dataset for three labels of sarcasm *viz.* 0: Non-sarcastic meme, 1:Mildly sarcastic meme, and 2: Highly Sarcastic meme. Details of each label is as follows:

0: A very general statement is given in the textual part of the meme, which we can quickly understand by merely reading it. The meaning of the meme is not twisted at all. So, we don't need to focus either on the visual part of the meme or include implicit cultural knowledge/context of that meme.

1: At first, look at the textual part of the meme; if the meaning of the meme is twisted and we cannot get its meaning properly, then focus on the image part of the meme. If we can easily infer the twisted meaning of the meme by focusing on both text and image, it will come under a *mildly sarcastic* category.

2: A *highly sarcastic* meme is determined with the help of implicit contextual knowledge of the meme.

3.3.2 Emotion

Most psycho-linguistics usually claim that few primary emotions are the foundation for all other emotions. For example, Ekman(Ekman and Cordaro, 2011) introduced six basic emotions: anger, disgust, fear, joy, sadness, and surprise. Similarly, *The psycho-evolutionary theory of emotion*, developed by Robert Plutchik(Wilson and Lewandowska, 2012), known as the *Plutchik Wheel of Emotions*, claimed eight primary emotions: joy, sadness, acceptance, disgust, fear, anger, surprise, and anticipation. However, (Kosti et al., 2017) claimed that merely these primary emotions could not adequately represent the diverse emotional states that humans are capable of. Taking inspiration from their work, we conducted extensive psychological research on the list of 120 affective keywords collected from our pre-defined four domains. After mapping these affective keywords to their respective emotions, we came up with 13 fine-grained emotion categories for our meme dataset. We annotate every sample of the dataset for 13 fine-grained categories of emotions, *viz. Disappointment (Disap), Disgust (Disg), Envy (En), Fear (Fe), Irritation (Ir), Joy (J), Neglect (Neg), Nervousness (Ner), Pride (Pr), Rage (Ra), Sadness (Sad), Shame (Sh), and, Suffering (Su).* (Refer Appendix Section 8.1 for example of each emotion category.)

3.3.3 Annotation guidelines

We annotate all the memes of our dataset with two labels (sarcasm and emotion). We employed experienced annotators with an expert-level understanding of Hindi for this purpose. We only included those annotators who were familiar with the Indian scenario. Additionally, we guaranteed that no annotator was biased in favor of a specific political leader, party, situation, occurrence, or caste. We annotated 100 samples to serve as a quality checker while evaluating the annotators' abilities. We faced a few challenges during annotation, which we solved by agreeing on a common point after a lot of discussions. We have mentioned a few challenges and their solution in the **Appendix**. Finally, the annotation guidelines and several annotated examples were distributed to the annotators. The annotators were asked to annotate the respective sarcasm label and as many emotions as possible in their annotations for a given meme.
To assess inter-rater agreement, we utilized Co-

[3] https://download-all-images.mobilefirst.me/

[4] github.com/tesseract-ocr/tesseract

hen's Kappa coefficient (Bernadt and Emmanuel, 1993), a statistical metric. For sarcasm label, we observed Cohen's Kappa coefficient score of 0.7197, which is considered a reliable score.

3.4 Dataset Statistics

Our corpus consists of a total 7,416 memes. Its distribution across various classes and more details about the dataset are shown in Table 7 in the **Appendix**.

4 Proposed Methodology

This section presents the details our proposed multitasking architecture by which we perform two tasks in parallel, *viz.* Sarcasm detection and Emotion recognition. We also describe the knowledge infusion (KI) mechanism which is a novel addition to the multitasking model. We can formalize our current problem as: Given a sample meme M_i from our corpus which is a combination of text $T_i = (t_{i1}, t_{i2}, ..., t_{ik})$ and image V_i with the shape (224,224,3) in RGB pattern, our task is to create a multitask classifier that should simultaneously predict the correct label $Y_s \subseteq$\{Non-sarcastic,Mildly-sarcastic,highly Sarcastic\} for S_i and all possible emotion labels Y_e. The respective optimizing goal is then to learn the parameter θ and get the optimum loss function $L(Y_s, Y_e|S, \theta)$. The basic diagram of the proposed model is shown in Figure 2. Detailed discussion of our proposed method is done in the following subsections:

4.1 Feature Extraction Layer

We use memes (M) as input to our model which are comprised of an image (V) and an associated text (T). These are then input into a feature extractor module to obtain the text representation (f_t) and visual representation (i_t), respectively. For our task, we use CLIP model as the feature extractor module. Specifically, we have used Multilingual CLIP (Radford et al., 2021) [5] to obtain textual features given Hindi text. We observe the following benefits of using CLIP over other image and text based feature extractors:
CLIP is pre-trained based on contrastive learning of image and text representations which ensures those representations lie close to each other given related text and image pair. This property is exploited to obtain better text and image features from CLIP model. We summarize the above steps by the following equation:

$$\begin{aligned} T, V &\in M \\ f_t, i_t &= CLIP(T, V) \end{aligned} \tag{1}$$

4.2 Multimodal Fusion

Separate text (f_t) and visual representation (i_t) obtained from feature extraction layer are then fed into a Fusion Module to prepare a fused multimodal representation. Our fusion module is based on Multimodal Factorized Bilinear pooling (MFB) (Yu et al., 2017).
Let us assume, we have CLIP extracted text feature (f_t) and visual features (i_t) having dimensions $\mathbb{R}^{m\times 1}$ and $\mathbb{R}^{n\times 1}$ respectively. Further assume we need a multimodal representation M_t having dimension $\mathbb{R}^{o\times 1}$. MFB module is comprised of two weight matrices U and V having dimensions $\mathbb{R}^{m\times ko}$ such that the following projection followed by sum-pooling operation is performed.

$$M_t = SumPool(U^T f_t \circ V^T v_t, k) \tag{2}$$

$SumPool(x, k)$ refers to using one dimensional non-overlapped window with the size k to perform sum pooling over x.

4.3 Knowledge Infusion (KI)

We devise a simple knowledge infusion (KI) technique to enrich multimodal representation (M_t) for better performance in our downstream classification tasks. Our KI method consists of two steps: i) Obtaining a learned representation from an already trained model, ii) Utilizing the learned representation via a gating mechanism to 'enrich' M_t. The following subsections deal with the aforementioned steps in details.

4.3.1 KI Learned Representation

We fine tune a copy of our model until convergence. We use *Memotion 2.0* dataset[6] for finetuning.We perform multitasking by classifying each meme instance into (i). one of four classes for sarcasm; and (ii). one of the three classes of sentiment.[7] This is done using two task specific classification layers, D'_{sar} and D'_{sent}, respectively, on top of the shared layers.
After the model is completely trained, we freeze

[5] https://github.com/FreddeFrallan/Multilingual-CLIP

[6] https://competitions.codalab.org/competitions/35688

[7] Each meme in *Memotion 2.0* dataset is annotated with both sarcasm and sentiment classes

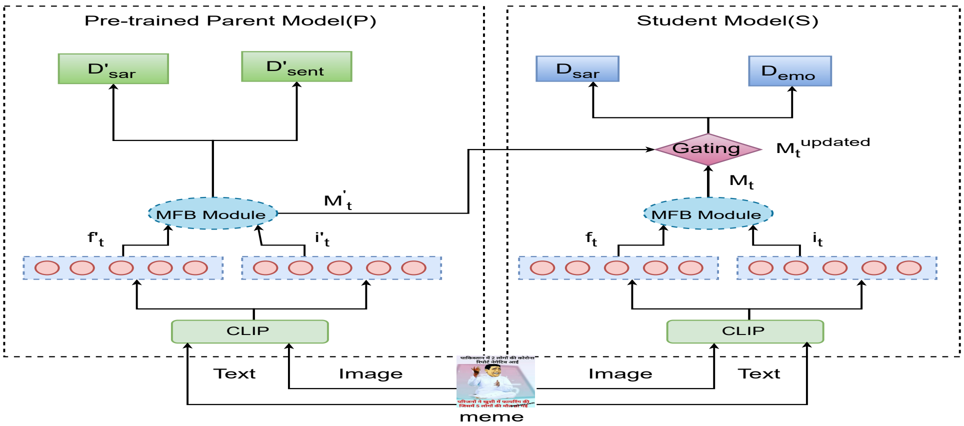


Figure 2: Schematic of our training methodology and the associated models. **Left: Parent Model (P)** Already trained and frozen model, trained on Memotion 2 dataset to detect 'Sarcasm' and 'Sentiment' using two feed forward layers D'_{sar} and D'_{sent}, respectively. **Right: Student Model (S)** It utilizes learned representation (M'_t) from the already trained model (P) shown in the left via the gating mechanism to update its hidden representation from M_t into $M_t^{updated}$. Thereafter, $M_t^{updated}$ is fed into two feed forward layers (D_{sar} and D_{emo}) associated with 'Sarcasm' and 'Emotion' respectively. Note that both of the models in *left* and *right* share the same architecture.

Setup	Model	T+V				T				V			
		re	pr	f1	acc	re	pr	f1	acc	re	pr	f1	acc
STL	M_{sar}	59.88	63.28	59.88	63.87	53.18	53.79	53.24	55.88	55.94	58.69	56.00	59.13
	M_{sar}^{KI}	63.28	62.86	62.86	64.20	54.40	54.82	54.48	56.90	55.86	57.56	56.22	59.2
MTL	$M_{sar+emo}$	61.07	62.43	61.11	64.61	53.04	54.48	53.14	55.81	56.75	62.03	56.28	60.75
	$M_{sar+emo}^{KI}$	61.71	63.96	61.86	65.35	52.95	53.36	52.94	55.75	55.84	56.39	55.90	58.72
Ensemble	ens^{-KI}	61.62	63.69	61.71	65.29	53.37	54.05	53.43	56.08	57.14	62.29	56.74	61.09
	ens^{KI}	63.60	64.23	63.79	66.17	54.83	55.12	54.87	57.44	56.56	57.66	57.64	59.74
	ens^{all}	**64.32**	**64.77**	**64.48**	**66.64**	55.38	55.94	55.46	58.05	58.06	60.60	58.04	61.63

Table 1: *Sarcasm head performance.* For both text only (T) and vision only (V) unimodal architectures, we show prformance of our proposed model for sarcasm detection. For comparison purposes, we also show multimodal (T+V) system performance.

its layers and use it to extract multimodal representation M'_t from its trained MFB module. Subsequently, M'_t is used to enrich M_t via the gating mechanism described below.

4.3.2 Gating Mechanism

Firstly, we obtain Multimodal representation (M_t) following Equation 2. Instead of feeding M_t directly into the subsequent classifier layers, we use a gating mechanism by which we pass extra information (M'_t) as needed and update M_t according to the following equation:

$$M_t^{updated} = f(M_t, M'_t) \tag{3}$$

where f is a generic function used to show the 'gating' mechanism.
Given an example from our dataset, we input it to our model and the model we have already trained on *Memotion 2.0* dataset. We extract multimodal representations M_t and M'_t from both the models. Specifically, we use a 'GRU unit' (Cho et al., 2014) to model the gating mechanism as follows:

$$M_t^{updated} = GRUCell(input = M_t, hidden = M'_t) \tag{4}$$

The 'update' and 'reset' gate within the GRU unit captures necessary information from M'_t to enrich shared multimodal representation M_t, which is then fed into task specific classification layers. Note that our gating scheme is generic and need not only be implemented using a GRU unit. In the ablation section, we compare the performance with our proposed GRU based gating scheme with other gating approaches that also could be used as well.

4.4 Classification

Our objective is divided into performing two tasks in parallel, i.e. (i). Classifying a meme into three categories, *viz.* Non-Sarcastic, Mildly-Sarcastic and Highly-Sarcastic; and (ii). Detecting the presence of thirteen fine-grained emotions. For both of these tasks, task specific classification layers are used and both of the task specific layers get

same multimodal representation from the previous 'shared' layers. Specifically, for sarcasm classification, a single feed-forward layer (D_{sar}) is used which obtains the multimodal representation (M_t) output from the previous MFB stage.
Similarly for recognizing emotion, we use another feed-forward layer (D_{emo}), which also obtains the same representation as D_{sar}.
Previous operations can be described as follows:

$$O_{sar} = D_{sar}(M_t^{updated}, activation = softmax)$$
$$O_{emo} = D_{emo}(M_t^{updated}, activation = sigmoid) \quad (5)$$
$$O_{sar} \in \mathbb{R}^{1\times3}; O_{emo} \in \mathbb{R}^{1\times13}$$

O_{sar} and O_{emo} are respectively the logit outputs associated to the D_{sar} and D_{emo} classifier heads. These output vectors are then used to calculate the respective cross entropy loss to optimize the model.

5 Results and Analysis

5.1 Models

We first evaluate our proposed architecture with unimodal inputs (Text only (T) and Vision only (V)) and compare their performance with multimodal inputs (T+V). For all of input combinations (T, V, T+V), We perform our experiments for both Single Task Learning (STL) and Multitask learning (MTL) setup. In STL setup, we only consider the model to learn to detect sarcasm in a given meme; whereas in MTL setup, the model learns from the mutual interaction of two similar tasks, viz. Sarcasm detection, and Emotion recognition. For each of STL and MTL setups, we also show the effect of knowledge infusion by training our proposed model with KI objective (c.f. Section 4.3).
STL Setup: In STL setup, we train the models to detect sarcasm in a meme by only training its D_{sar} classifier head. Furthermore, we train two separate models based on whether we use KI method or not.
1. M_{sar}: This model is trained by only optimizing its D_{sar} head for sarcasm. Also we set $M_t^{updated} = M_t$ to disable Knowledge infusion.
2. M_{sar}^{KI}: This is same as M_{sar} except KI is enabled here. We follow Equation 4 to enable KI.
MTL Setup: In MTL setup, we simultaneously train D_{sar} and D_{emo} classifier heads of the model to perform multitasking by detecting both sarcasm and emotion in a meme. Similar to the STL setup, two models are trained for STL setup too.
3. $M_{sar+emo}$: This model is an extension of M_{sar} model. It is trained by optimizing its D_{sar} head for detecting sarcasm and D_{emo} for detecting emotion. We set $M_t^{updated} = M_t$ to disable Knowledge infusion.
4. $M_{sar+emo}^{KI}$: This is same as M_{sar}^{KI} except that we train both of its classifier heads (D_{sar} and D_{emo}) to perform multitasking. We follow Equation 4 to enable KI.

5.2 Result Analysis

In this section, we show the results that outline the comparison between the single-task(STL) and multi-task (MTL) learning framework. We have used 7416 data points with a train-test split of $80 - 20$. 15% of the train set is used for validation purposes. For evaluation of sarcasm in Table 1, we use F1 score (F1), precision (pr) and recall score (re) and accuracy (acc) as the preferred metrics. In STL setup, we observe that the M_{sar}^{KI} performs better than M_{sar}. This shows enabling knowledge infusion aids the model to detect sarcasm. We observe that even the MTL setup benefits by enabling knowledge infusion (KI). This is evident from the increased performance of +0.75% in terms of F1-score when $M_{sar+emo}^{KI}$ compared to $M_{sar+emo}$. This increased performance can be attributed to the sentiment-aware hidden representation (M_t'), which helps our model perform better by transferring knowledge via the proposed gating mechanism.
We also observe that for both STL and MTL setups, the multimodal input settings(T+V) shows better performance than unimodal input settings(T or V).

To observe effects of KI technique, we form ensemble of the trained model with two setups, *viz* (i). *Ensemble with KI* (ens^{KI}) and (ii). *Ensemble without KI* (ens^{-KI}). In ens^{KI}, we only consider two models which were trained with knowledge infusion (KI). We consider predictions of models M_{sar}^{KI} and $M_{sar+emo}^{KI}$ to build the ensemble model ens^{KI}. Similarly for ens^{-KI} model, we consider M_{sar} and $M_{sar+emo}$ models to build our ensemble. We observe that ens^{KI} outperforms ens^{-KI} by $+2.1\%$ in terms of F1-score. This also shows the effectiveness of our proposed KI scheme. Finally, we build an ensemble model ens^{all} by considering predictions from all the four models in hand. This final model performs decently better than other models. It can be seen in the increased performance of the model with respect to the baseline M_{sar} model with an improvement of +4.6% in terms of F1-score.

For emotion analysis, we demonstrate the performance for STL and MTL setups both in Table 12. We observe that the model performs better in MTL setup ($M_{sar+emo}$) compared to the STL setup (M_{emo}), thus reinforcing the hypothesis of symbiosis between sarcasm and emotion.

5.3 Ablation Analysis

In this section, we analyse our models with different setups. Firstly, we observe that the generic gating mechanism shown in Equation 3 can be implemented by the following methodologies. Beside the proposed *GRU* based gating mechanism, we implement the generic gating scheme with two other methods: (i). Concatenation followed by projection (*cat+proj*) to combine M_t and M'_t and (ii). Minimize KL divergence (*KL_div*) between M_t and M'_t. We also observe that besides using different KI gating schemes, performance of the student models could also depend on the objective by which the parent model is trained. We can train the parent model with (i). *sar* objective (only detecting sarcasm) by only training its D'_{sar} classifier head; or (ii). *sar+sent* objective (detecting both sarcasm and sentiment via multitasking) by training its D'_{sar} head and D'_{sent} simultaneously.

KI Fusion	ens^{all}			
	re	pr	f1	acc
cat+proj	62.66	64.39	62.95	65.62
KL_div	62.65	64.98	62.91	66.03
GRU	**64.32**	**64.77**	**64.48**	**66.64**

Table 2: Ablation: performance of ensemble based on *sar+sent* pretraining objective of parent model. Ensemble model ens^{all} is built by weighted ensemble of M_{sar}, $M_{sar+emo}$, M^{KI}_{sar},$M^{KI}_{sar+emo}$ models. For different KI fusion, we show the effect on the ensemble above.

We also show the performance of the ensemble model (end^{all}) based on different fusion schemes in Table 3 and Table 2 for *sar* and *sar+sent* pretraining objectives of parent model, respectively.

KI Fusion	ens^{all}			
	re	pr	f1	acc
cat+proj	62.32	63.98	62.55	65.56
KL_div	62.61	64.68	62.83	66.03
GRU	**63.62**	**64.71**	**63.91**	**66.23**

Table 3: Ablation: performance of ensemble based on *sar* only pretraining objective of parent model. Ensemble model ens^{all} is built by weighted ensemble of M_{sar}, $M_{sar+emo}$, M^{KI}_{sar},$M^{KI}_{sar+emo}$ models. For different KI fusion, we show the effect on the ensemble above.

		Sample 1	Sample 2	Sample 3
				
STL	**True Label**	**2**	**1**	**0**
	M_{sar}	0	2	1
	M^{KI}_{sar}	2	0	1
MTL	$M_{sar+emo}$	1	2	2
	$M^{KI}_{sar+emo}$	2	1	0

Table 4: Sample test examples with predicted sarcasm label for STL and MTL models. Refer Table 5 for label definition.

Meme Name	sarcasm class					Possible Reason
	Act	M_{sar}	M^{KI}_{sar}	$M_{sar+emo}$	$M^{KI}_{sar+emo}$	
meme1	0	2	2	2	2	hazy picture
meme2	0	2	1	2	2	uninformative picture
meme3	0	2	2	2	2	Background Knowledge
meme4	0	1	1	1	1	Common Sense
meme5	1	2	2	2	2	Hindi words in English font
meme6	2	1	1	0	1	Code mixing

Table 5: Error Analysis: Frequent error cases and the possible reasons frequently occurring with each of them. Due to space constraint, we provide actual memes corresponding to the *Meme Name* col. in the **appendix** Table 11. Label definition: **2**: Highly Sarcastic, **1**: Mildly Sarcastic, **0**: Not Sarcastic.

We observe that when we use *GRU* as the knowledge infusion (KI) technique, ensemble performance is better compared to the *KL_div* and *cat+proj* fusion methods. This is in alignment with the intuition that the gating mechanisms inside GRU acts as a 'better' filter of which information of the parent model it should retain and discard for downstream performance of student models. We also empirically verify that *sar+sent* pretraining objective of the parent model could learn better representation (M'_t) than *sar* only pretraining objective, such that the performance of the student model increases.

5.4 Detailed Analysis

T	T+V	Image	Text
×	✓		औ भई मारो मुझे मारो । Come brother, Beat me
V	**T+V**	**Image**	**Text**
×	✓		भारत माता की जय बोलूं तो जीतने दोगे ? Will you let me win, if I say "Long Live Mother India"

Figure 3: Two examples where we show multimodal (T+V) M_{sar} model performs better than unimodal (T and V only) M_{sar} models.

To explain the feasibility of our proposed model, we performed a detailed quantitative and qualitative analysis of some samples from the test set. In Table 4, we show 3 examples with true labels of sarcasm class. We compare models for both STL and MTL setups by comparing their predicted labels with actual labels. We observe how MTL model with KI objective ($M^{KI}_{sar+emo}$) helps to capture related information from the meme to correctly predict the associated sarcasm class. We also report the confusion matrix (c.f. Fig 6) of our proposed multitasks learning model(Detailed discussion is done in **Appendix**, Section 8.6). From the confusion matrix, we identify the effectiveness of our proposed model.

Furthermore, to analyse whether the multimodality helps in the context of detecting sarcasm, we also analyse two predicted examples in Figure 3. In the first example, we see that the text only (T) model fails to detect sarcasm, whereas the multimodal (T+V) model correctly classifies it. The text *'Come brother, beat me'* alone is not sarcastic, but whenever we add Mahatma Gandhi's picture as a context, the meme becomes sarcastic. This is correctly captured by the multimodal (T+V) M_{sar} model.Similarily, in the second example, without textual context the image part is non-sarcastic and thus the vision only (V) M_{sar} model wrongly classifies this meme as non sarcastic. Adding textual context helps the multimodal model to correctly classify this meme as a sarcastic meme.

We also observe that despite the strong performance of our proposed model, it still fails to predict the sarcasm class correctly in a few cases. In Table 5, we show some of the memes with actual and predicted sarcasm labels from the multimodal (T+V) framework (M_{sar}, M^{KI}_{sar},$M_{sar+emo}$, $M^{KI}_{sar+emo}$,). We show four most common reasons why the models are failing to predict the actual class associated with the meme. (c.f **Appendix**, Table 11 for the corresponding memes.)

5.5 Explainability and Diagnostics

After the training is done, we expect the model to exploit contextual knowledge embedded in the meme to explain its prediction. To explain the prediction behavior of our model, we use a well known model-agnostic interpretability method known as LIME (Locally Interpretable Model-Agnostic Explanations) (Ribeiro et al., 2016).

In Figure 4, we show two memes and by using the LIME outputs, we explain the behavior of $M^{KI}_{sar+emo}$ model. The first meme which contains the picture of Person-A is manually labeled as *highly sarcastic* and the model correctly predicts the class. We observe that the face of Person-A is contributing mostly to the correct prediction. Simi-

Figure 4: Examples showing visualization by LIME for multimodal (T+V) $M^{KI}_{sar+emo}$ model.

larly for the second meme, the associated sarcasm label is *non sarcastic* but the model wrongly classifies it as *highly sarcastic*. We observe that the model tends to focus more on the face of Person-B to make its prediction as it did in the case of Person-A in the previous meme. By analysing examples from our dataset, we found that there is a large collection of highly sarcastic memes which contain the face of either Person-A or Person-B. Therefore, instead of leaning the underlying textual and visual semantic of a particular meme, the model gets biased by the presence of Person-B's face and the meme is incorrectly classified as *highly sarcastic*.

6 Conclusion

In this paper, we have attempted to solve a very challenging task of sarcasm detection from Internet memes. We have proposed a deep learning-based *multitask knowledge-infused(KI)* model that leverages a meme's emotions and sentiment to identify the presence of sarcasm in it. Since there was no suitable labeled dataset available for this problem, we manually created the large-scale benchmark dataset by annotating 7,416 memes for sarcasm and emotion. Quantitative and qualitative error analysis on the dataset shows the efficiency of our proposed model, which produces promising results with respect to the baseline models. Our analysis found that the model could not perform exceptionally well in a few cases due to the lack of context knowledge. In the future, along with investigating new techniques in this direction, we will also explore more fusion strategies to learn a better multimodal representation of textual and visual parts of memes jointly.

7 Ethical Section

We gathered all the memes freely available in the public domain. We followed the policies for using those data and did not violate any copyright issues. The dataset used in this paper is solely for academic research purposes.We also have got it verified from our institute review board. To maintain the anonymity of any individual, we replaced actual name with Person-XYZ throughout the paper. We employed experienced annotators with an expert-level understanding of Hindi for this purpose. The annotators are from the Indian population, and we got this data annotated from a crowdsource company following standard protocol. We only included those annotators who are familiar with the Indian scenario. Additionally, we guaranteed that no annotator was biased in favor of a specific political leader, party, situation, occurrence, or caste. Our motivation is within the scope of building a multitasking system that would restrict people who intended to spread the meme purposefully to reinforce stereotypes, wrong philosophies, personalities, and false ideologies.

References

Morris Bernadt and J Emmanuel. 1993. Diagnostic agreement in psychiatry. *The British journal of psychiatry : the journal of mental science*, 163:549–50.

Ohtsuki Bouazizi and Tomoaki. 2016. A pattern-based approach for sarcasm detection on twitter. *IEEE Access*, 4:5477–5488.

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

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.

Kyunghyun Cho, Bart van Merrienboer, Çaglar Gülçehre, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *CoRR*, abs/1406.1078.

Xiangjue Dong, Changmao Li, and Jinho D. Choi. 2020. Transformer-based context-aware sarcasm detection in conversation threads from social media. *CoRR*, abs/2005.11424.

Paul Ekman and Daniel T. Cordaro. 2011. What is meant by calling emotions basic. *Emotion Review*, 3:364 – 370.

Deepanway Ghosal, Md Shad Akhtar, Dushyant 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.

Debanjan Ghosh, Alexander Richard Fabbri, and Smaranda Muresan. 2017. The role of conversation context for sarcasm detection in online interactions. *CoRR*, abs/1707.06226.

Md. Kamrul Hasan, Wasifur Rahman, Amir Zadeh, Jianyuan Zhong, Md. Iftekhar Tanveer, Louis-Philippe Morency, and Mohammed E. Hoque. 2019. UR-FUNNY: A multimodal language dataset for understanding humor. *CoRR*, abs/1904.06618.

Aditya Joshi, Vaibhav Tripathi, Pushpak Bhattacharyya, and Mark J. Carman. 2016. Harnessing sequence labeling for sarcasm detection in dialogue from TV series 'Friends'. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 146–155, Berlin, Germany. Association for Computational Linguistics.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020. The hateful memes challenge: Detecting hate speech in multimodal memes. *CoRR*, abs/2005.04790.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Ronak Kosti, Jose M. Alvarez, Adria Recasens, and Agata Lapedriza. 2017. Emotic: Emotions in context dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2309–2317.

Liyuan Liu, Jennifer Lewis Priestley, Yiyun Zhou, Herman E. Ray, and Meng Han. 2019. A2text-net: A novel deep neural network for sarcasm detection. In *2019 IEEE First International Conference on Cognitive Machine Intelligence (CogMI)*, pages 118–126.

Navonil Majumder, Soujanya Poria, Haiyun Peng, Ni Chhaya, Erik Cambria, and Alexander Gelbukh. 2019. Sentiment and sarcasm classification with multitask learning. *IEEE Intelligent Systems*, 34:38–43.

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

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. *CoRR*, abs/2103.00020.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, pages 1135–1144.

Chhavi Sharma, Deepesh Bhageria, William Scott, Srinivas PYKL, Amitava Das, Tanmoy Chakraborty, Viswanath Pulabaigari, and Björn Gambäck. 2020. Semeval-2020 task 8: Memotion analysis - the visuo-lingual metaphor! *CoRR*, abs/2008.03781.

Oren Tsur and Ari Rappoport. 2009. Revrank: A fully unsupervised algorithm for selecting the most helpful book reviews. In *ICWSM*.

Rajasekar Venkatesan and Meng Joo Er. 2014. Multi-label classification method based on extreme learning machines. In *2014 13th International Conference on Control Automation Robotics Vision (ICARCV)*, pages 619–624.

Paul A. Wilson and Barbara Lewandowska. 2012. The nature of emotions.

Zhou Yu, Jun Yu, Jianping Fan, and Dacheng Tao. 2017. Multi-modal factorized bilinear pooling with co-attention learning for visual question answering. *CoRR*, abs/1708.01471.

8 Appendix

8.1 Fine-grained emotion categories

In the Table 6, we have defined all 13 fine-grained emotion categories with the respective example which is defined in our dataset.

Table 6: Examples of all 13 fine-grained emotion categories defined in section 3.3.2. For each category, we provide a sample in which that emotion outweighs other emotions. Additionally, we mentioned which modality (textual, visual, or a combination of the two) is more involved in unveiling the underlying emotion.

(1)Pride
Due to Text

Fear is the one who dies for his image. And I die for the image of India. That's why I am not afraid of anyone.

(2) Rage
Due to both

**** you only said, take the *prodical* science. There is a lot of scope ahead.

(3)Envy
Due to Text

O Partha, let's go arrows. But on whom? You just shoot. Person-C himself will settle and take it in the middle.

(4) Disgust
Due to Text

We have a simple funda, whenever we talk about ourselves, entangle the public by raising religious issues like love-jihad,Triple Talaq, Mandir Masjid, Loudspeaker, Hindu-Muslim, Temple Mosque, Loudspeaker

(5) Suffering
Due to Text

I am not afraid of slaps, sir, I am afraid of love. You let it be sister, I have got a slap, I know.

(6)Joy
Due to both

If you go to see someone's newly built house, you should praise him a lot so that you can also get an invitation to its dinner party.

(7) Fear
Due to Image

Now you will be trimmed.

(8) Neglect
Due to Text

Person-A is because of ancestors, and Person-C because of fools.

(9) Irritation
Due to Text

"Theft will increase due to the construction of 4-lane highway, 1000 trees will be cut, pollution will increase":Person-Y. This is a stigma in the name of the journalist. No work is done in the country, they have to be criticized.

(10) Nervousness
Due to Image

Logic in Hindi serials, given the death of extinguished husband.

(11)Shame
Due to both

Saheb's slogan in 2019. "Leave studies, take embroidery" Wooden saddle, Horse on the saddle. If you do not get a job, then sell pakora.

(12)Disappointment
Due to Text

We have NASA. We have a destroyer.

(13)Sadness
Due to both

By 2024, no one will remain poor, some will die of corona, some will die of hunger.Some will die of hatred, those who survive will die of debt. Then our sahib will have this fun together with his friends.

Figure 5: Challenges during annotation

Test Sample 1

Highly sarcastic or Mildly sarcastic

Test Sample 2

joy or insult

8.2 Challenges

The presence of incongruity that gives rise to sarcasm also raises many challenges during data annotations. Additionally, emotion detection in a meme is challenging due to the obscure nature of memes. During annotation, we faced a few challenges, which we resolved after many discussions. We have listed here a few challenges we faced during data annotation.

- Certain issues have grown so ubiquitous that they are no longer twisted for humans in today's world. For example, consider 1st meme in Table 5. It says,*"Go to hell, but not in the crowd."* The term *crowd* has been used in relation to covid-19. As a result, these memes should be classified as *mildly sarcastic* or *highly sarcastic*. We decided to annotate these memes as *highly sarcastic* without being biased towards any issues. Even though these words are general for humans, the model will not know its contextual knowledge.
- The annotation difficulty is exacerbated by the fact that social media users frequently use few words. For example, consider 1st meme in the Table 5. The meme says, *"Tag a friend who is good at heart but a bada** in mind."* The existence of joy alongside slur words makes annotation difficult since it can't articulate if the meme maker is attempting to offend the target directly with slur words or is just conveying joy.

8.3 Dataset Statistics

Dataset statistics are presented in Table 7 and Table 8.

8.4 Extended Ablation Study

In Table 10, we test whether we could directly use the obtained textual and visual representation from the CLIP model and subsequently concatenate and

classes	instance	% distribution
Non-Sarcastic(0)	1798	24.25
Mildly Sarcastic(1)	2770	37.35
Highly Sarcastic(2)	2848	38.40

Table 7: Data statistics of our annotated corpus for Sarcasm

Emotions	Disa	Disg	En	Fe	Ir	J	Neg	Ner	Pr	Ra	Sad	Sh	Su
Instances	3099	350	51	186	169	5940	2488	526	508	992	2095	151	1531

Table 8: Emotion class distribution in our dataset

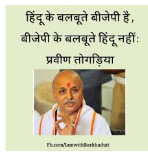
project them to obtain the multimodal representation. We further ask whether this approach could perform better than our proposed MFB as the fusion module. These results are tabulated in Table 10. We infer from the results that, simple methods such as concatenation followed by projection performs worse than using sophisticated method like MFB as multimodal fusion module. We tabulate our results for using different KI gating scheme in Table 9 under both *sar* and *sar+sent* pretraining objective of the parent model.

Fusion	M_{sar}				$M_{sar+emo}$			
	re	pr	f1	acc	re	pr	f1	acc
Concat	58.89	62.83	58.59	62.99	58.98	62.54	58.58	63.12
MFB	59.88	63.28	59.88	63.87	61.07	62.43	61.11	64.21

Table 10: Ablation: effect of concatenation (**Concat**) vs MFB module (**MFB**) for STL (M_{sar}) and MTL ($M_{sar+emo}$) schemes.

meme1

meme2



meme3

meme4

meme5

meme6

Table 11: Example memes shown in Table 5

8.5 Experimental setup

We evaluate our proposed architecture on our curated dataset. The optimal hyperparameters for our model are found using grid search and to maintain consistency over all the experiments performed, we choose same set of hyperparameters.
Our proposed model is implemented using Pytorch Lightning[8] framework. We use Adam(Kingma and

[8] https://www.pytorchlightning.ai/

Obj.	KI Fusion	M_{sar}^{KI}				$M_{sar+emo}^{KI}$			
		re	pr	f1	acc	re	pr	f1	acc
sar	*GRU*	62.68	63.75	62.91	64.74	62.41	64.40	62.61	65.42
	KL_div	61.85	64.11	62.06	65.29	61.14	64.25	61.00	65.30
	cat+proj	60.70	61.87	60.89	62.31	59.63	64.08	59.24	64.07
sar+sent	*GRU*	63.28	62.86	62.86	64.20	61.71	63.96	61.86	65.35
	KL_div	61.75	64.33	62.00	65.15	62.34	64.67	62.49	66.00
	cat+proj	61.12	62.28	61.31	64.20	60.86	63.58	61.20	63.59

Table 9: Ablation results of two models *viz sar* only and *sar+sent* pretraining objective of parent model with different KI fusion methods. *Refer Section 5.3 for detailed description of sar+sent and sar training objective.*

Ba, 2015) as the optimizer for the model. Softmax and Sigmoid activations are used for the sarcasm classifier head (D_{sar}) and emotion classifier head (D_{emo}), respectively.

We have used 7416 data points to split those into *train set*, validation set and test set. Original data point is first split into $80 - 20$ parts to create train-test split. We have used 15% of the train set as the validation set while training the model.

All of the models are trained until convergence. We have used early stopping based on validation set performance. The training stops if the validation set performance does not increase after consecutive 10 epochs. A single NVIDIA Tesla GPU is used to conduct the experiments.

To compare the models in equal footing a same set of hyper-parameters are used across each experiment.

1. *Optimizer:* Adam (lr=5e-3)
2. *Batch Size:* 128
3. *Loss function:* Cat. cross-entropy for training D_{sar} and binary cross-entropy for training D_{emo}.

8.6 Visualization of Confusion Matrix

In figure 6, we visualize the heatmaps of the confusion matrix for all the multimodal models to compare their classwise prediction. From the visualization, we observe that for *Non-Sarcastic* class, M_{sar}^{KI} correctly classifies 208 examples and thus it gets the highest class wise accuracy for the class *Non-Sarcastic*. Similarly for classes *Mildly Sarcastic* and *Highly Sarcastic*, models M_{sar} and $M_{sar+emo}$ perform the best respectively. This entails that for each classes, each of this model possess a substantial contribution resulting in performance gain of the weighted ensemble model ens^{all}.

8.7 Training Graphs

We plot F1 score of all our models (M_{sar}, $M_{sar+emo}$, M_{sar}^{KI} and $M_{sar+emo}^{KI}$) with respect to no. of epochs. In figure 7, these results are shown.

8.8 Results for Emotion

Task	M_{emo}				$M_{sar+emo}$			
	re	pr	F1	hloss	re	pr	F1	hloss
Emo. Recognition	46.93	75.36	57.84	12.88	51.07	71.11	59.46	13.11

Table 12: Emotion head performance for multimodal (T+V) setting.

Categories	$M_{sar+emo}$			M_{emo}		
	re	pr	F1	re	pr	F1
Disappointment	0	0	0	0.0	0.0	0.0
Disgust	78	38	52	65	56	61
Envy	100	2	0.4	100	2	0.5
Fear	69	12	20	46	17	25
Irritation	100	2	0.1	100	3	0.1
Joy	0	0	0	0	0	0
Neglect	0	0	0	0	0	0
Nervousness	57	38	55	53	44	48
Pride	44	19	27	55	35	43
Rage	46	75	53	44	72	51
Sadness	54	27	36	49	17	25
Shame	46	75	57	55	35	43
Suffering	89	91	90	89	89	89

Table 13: Class-wise emotion head performance for multimodal (T+V) setting.

Besides precision score (pr), recall score (re) and F1 score (F1), for emotion recognition, we additionally use hamming loss (Venkatesan and Er, 2014) to report performance score.

In Table 12, we show results for our secondary task of Emotion recognition which is performed as a multilabel classification task.

In Table 13, we show class-wise result for each of the 13 emotion classes. All of the classes which gets poor class-wise performance has very less no.

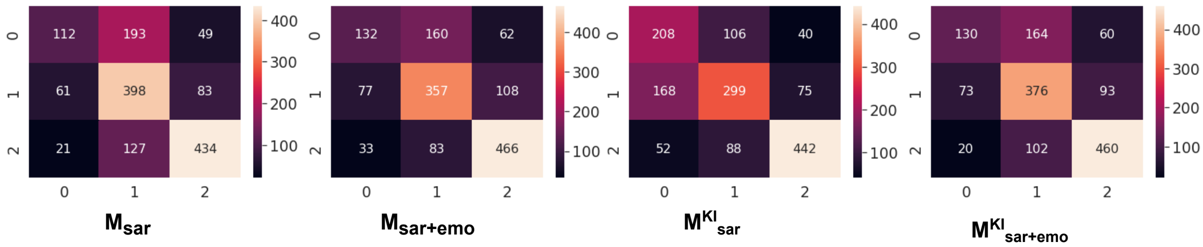


Figure 6: Heatmaps of the confusion matrix for four multimodal (T+V) models using both STL and MTL setup.

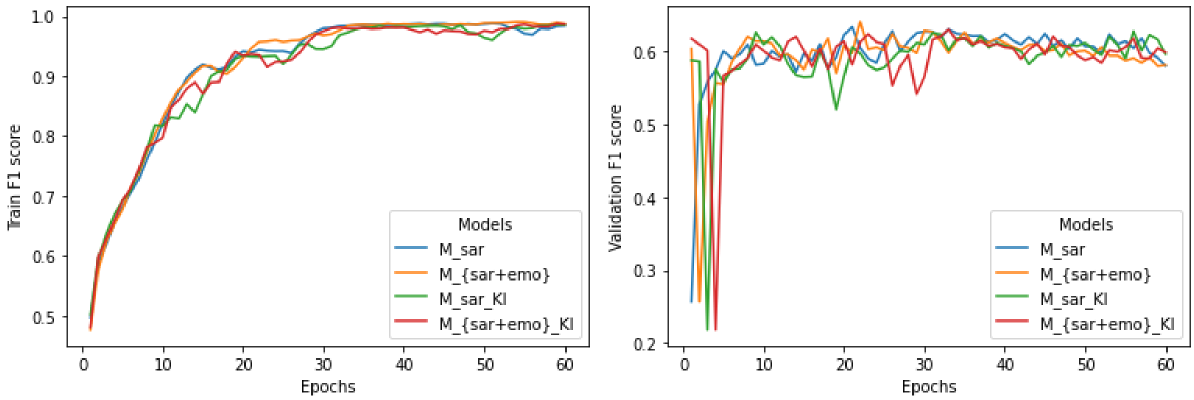


Figure 7: Training Graphs of all STL and MTL multimodal (T+V) models.

of (<50) test samples. Emotion Class *Suffering* has the highest number of test samples (1319), thus it obtains the highest performance.