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

Anonymous ACL submission

Abstract

In this paper, we hypothesize that sarcasm detection is closely associated with the emotion present in meme. Thereafter, we propose a deep multitask model to perform these two tasks in parallel, where sarcasm detection is treated as the primary task, and emotion recognition is considered as an auxiliary task.

We create a large scale dataset consisting of 7416 memes in Hindi, one of the widely spoken languages. We collect the 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. Furthermore, we propose a novel Knowledge Infusion (KI) based module which captures sentiment aware representation from a trained model using the *Memotion 2.0* dataset. Detailed empirical evaluation shows that multitasking model performs better than the single-task model. We also show that using this KI module on top of our model can boost the performance of sarcasm detection in both single- task and multi-task settings even further. We will make the resources and codes available[1]

1 Introduction

Sarcasm is an integral part in day-to-day conversations. People make use of sarcasm in conversation or writings to convey dis-likeness towards a situation or a person. Sarcasm is hard to understand because it usually uses humor in dialog (may also contain nonverbal cues) to show disapproval/dislike. Some of the negative aspects include using sarcasm as a malicious propaganda against rival parties and exaggerated achievements.

Memes are a form of multimodal media that is becoming increasingly popular on the internet. It was initially created for humor purposes. But due to the multimodality in nature, some memes help users to spread negativity in society in the form of sarcasm or dark humor (Kiela et al., 2020; Sharma et al., 2020; Suryawanshi et al., 2020). In the context of memes, detecting sarcasm is more difficult, as memes typically connect to a lot more background or contextual information. In meme, just like offensiveness detection (He et al., 2016), we cannot uncover the complex meaning of sarcasm until we know all the modalities and their contributions in sarcastic content. Also, a sarcastic sentence always has an implied negative sentiment because it intends to express contempt (Joshi et al., 2017).

For example, if there is a meme with text containing, "By showing this innocent dream of becoming the Prime Minister, those who snatched his childhood! You will feel sin..." The sentiment of this meme can, itself, be positive, negative, or neutral if we only focus on the textual part. The true sentiment can only be found by adding an image to it (c.f. Test Sample 1 in Figure 1). Once we focus on the visual part, we understand that the meme creator is considering an adult as a child to insult a person-xyz[2], which shows negative sentiment. Some memes are purely humorous, while others spread offensive content in the form of sarcasm.

In example 2 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 3 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*

[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

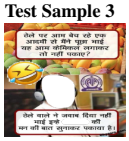
	Test Sample 1	Test Sample 2	Test Sample 3
			
Sarcasm	Highly Sarcastic	Non-sarcastic	Highly sarcastic
Emotions	insult,joy	joy	joy,insult

Figure 1: Some samples from our dataset

replied, "No, brother, it has been riped/annoyed after listening to Person-A's inner thoughts." When we look at this meme from outer perspective, it is seen that the meme was formed solely for humor purpose with no apparent twist. But, after carefully analyzing the emotion of the creator of the meme by adding the context, we 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 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 hypothesize 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.

Key contributions of our work are summarized as follows:

- *We create a high-quality and large-scale multimodal meme dataset annotated with three label to detect and also quantify the sarcasm given in a meme by utilizing* 3*-classes (non-sarcastic, mildly sarcastic, and high sarcastic) and* 13 *fine-grained emotion labels.*
- *We propose a deep neural model which simultaneously detects sarcasm and recognizes emotions in a given meme. Multitasking ensures that we exploit the emotion of the meme, which aids in detecting sarcasm more easily. We also propose a module denoted as knowledge infusion (KI) by which we leverage pre-trained sentiment-aware representation in our model.*
- *Empirical results show that the proposed KI module significantly outperforms the naive multimodal model.*

2 Related Work

According to a literature review, a multimodal approach to sarcasm detection in memes is a relatively recent trend 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. (2019a) 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. We use a list of total 126 keywords like terrorism, political memes, 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 using the collected 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 removing 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-multimodal, 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 Optical Character Recognition(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

All prior works have merely detected sarcasm (Sharma et al., 2020; Chauhan et al., 2020); however, in our work, we also attempt to quantify the sarcasm given in a meme by utilizing a 3-class classification. 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: *Non-sarcastic* meme: Textual part of the meme doesn't contain any twisted meaning. A general statement is given in the textual part of the meme, which we can quickly understand by merely reading it.(c.f. **Appendix** Table 12 Non-sarcastic meme examples.)

1: *Mildly-Sarcastic* meme: In order to understand the meme, we need to focus on both the modality, i.e. text as-well-as image part of the meme. If we can infer the twisted meaning of the meme by focusing on both text and image, it will come under a *mildly sarcastic* category. (c.f. **Appendix** Table 12 mildly sarcastic meme examples.)

2: *Highly-sarcastic* meme: The twisted meaning of a *highly sarcastic meme* is determined after adding implicit background (or, contextual) information of the meme. (c.f. **Appendix** Table 12 highly sarcastic meme examples.)

3.3.2 Emotion

Most psycho-linguistics usually claim that few primary emotions are the foundation for all other emotions. For example, 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 (i.e. *politics, religious, racist and sexist*). 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). (c.f. **Appendix** §A.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). For the annotation purpose, we employed experienced annotators with an expert-level understanding of Hindi. Additionally, we guaranteed that no annotator was biased in favor of a specific political leader, party, situation, occurrence, or caste. We ensure that our data collection is done keeping equality in mind in view of political and religious bias. We have discussed the removal of political and religious bias in detail in the **Appendix** §A.2. We recruited an experienced team of AI professionals who have delivered

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

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

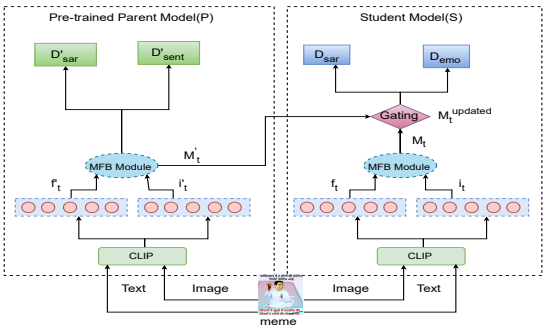


Figure 2: Schematic of our training methodology and the associated models. **Left: Parent Model (P)** Already trained and frozen model, trained on Memotion 2 dataset to detect 'Sarcasm' and 'Sentiment' using two feed forward layers D'_{sar} and D'_{sent}, respectively. **Right: Student Model (S)** It 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.

impactful AI data annotation. We only included those annotators who were familiar with the Indian scenario. At first, we have provided expert-level training based on 100 samples. For each sample, two annotators were there. In the case of a few disagreements during annotation process, we resolved it by agreeing on a common point after thorough discussions. We have mentioned a few challenges and their solution in the **Appendix** §A.3. 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 Cohen'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. Similarly, for 13 fine-grained emotion labels, the reported Krippendorff's Alpha Coefficient (krippendorff, 2011) is 0.6174 in a multilabel scenario.

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 10 and Table 11 in the **Appendix** §A.4. We have also shown the distribution of emotion tags within each sarcasm class in Figure 5 in the **Appendix** §A.4

4 Proposed Methodology

This section presents the details our proposed multitasking model 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 M_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 $L(Y_s, Y_e|M, \theta)$. The basic diagram of the proposed model is shown in Figure 2. The following section discusses our method in details:

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. *Note that CLIP is only used as a feature extractor, they are not finetuned.* We summarize the above steps by the following equation:

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

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).

We have CLIP extracted text feature (f_t) and visual features (i_t) having dimensions $\mathbb{R}^{m\times1}$ and $\mathbb{R}^{n\times1}$

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

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
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

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. Here, Knowledge Infusion (KI) is disabled.

Setup	Model	T+V				T				V			
		re	pr	f1	acc	re	pr	f1	acc	re	pr	f1	acc
STL	M^{KI}_{sar}	63.15	64.01	63.29	65.89	58.15	58.32	58.19	60.14	56.89	57.63	57.01	59.81
MTL	$M^{KI}_{sar+emo}$	63.11	65.01	63.37	66.64	58.14	60.01	57.80	62.31	57.79	60.73	57.25	62.24

Table 2: *Sarcasm head performance.* Here, Knowledge Infusion (KI) is enabled. $M^{KI}_{sar+emo}$ is statistically significant to M_{sar} ($p < 0.05$). *McNemar's test* is performed to determine statistical significance level.

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.

$$M_t = SumPool(U^T f_t \circ V^T i_t, k) \quad (2)$$

$SumPool(x, k)$ 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 training.We perform multitasking by classifying each meme instance into (i) one of three 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 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:

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

where f is a generic function used to show the 'gating' mechanism.

Given an example from our dataset, we input it to our model which 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:

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

We know that the parent model was never trained to detect emotion in meme and the obtained representation from the gating mechanism ($M_t^{updated}$) thus conflicts with the multitasking objective (simultaneously detecting sarcasm and recognizing emotion) being utilized in the student model. To compensate this issue, we tweak our training objective by replacing $M_t^{updated}$ with $M_t^{updated'}$, where it is given by:

$$M_t^{updated'} = w1 \times M_t^{updated} + w2 \times M_t \quad (5)$$

where, $w1$ and $w2$ are scalar weight parameters initialized to 0.5. This ensures that while training

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

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

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 3: Emotion head performance for multimodal (T+V) setting. *hloss* refers to Hamming Loss(Venkatesan and Er, 2014).

the student model, we take a weighted average of its hidden state representation (M_t) and the GRU gate output ($M_t^{updated}$). Initial weightage of both $M_t^{updated}$ and M_t are same and we take simple average of $M_t^{updated}$ and M_t (by setting $w1$=$w2$=0.5). The 'update' and 'reset' gates within the GRU unit captures necessary information from M_t' to enrich the 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 §A.6, we compare the performance with our proposed GRU based gating scheme with other gating approaches that also could be used.

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:

$$\begin{aligned} O_{sar} &= D_{sar}(M_t^{updated'}, activation = softmax) \\ O_{emo} &= D_{emo}(M_t^{updated'}, activation = sigmoid) \\ O_{sar} &\in \mathbb{R}^{1\times 3}; O_{emo} \in \mathbb{R}^{1\times 13} \end{aligned} \quad (6)$$

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, i.e, 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 component (c.f. §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 use

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 and Table 2, we use F1 score (F1), precision (P) and recall score (R) 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 +2.26 F1-score when $M_{sar+emo}^{KI}$ compared to $M_{sar+emo}$. This improvement could 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). Our best performing model ($M_{sar+emo}^{KI}$) obtains an F1 score of 63.37, surpassing all the baselines. This performance is also statistically significant to M_{sar} ($p<0.05$).

For emotion recognition, we demonstrate the performance for STL and MTL setups both in Table 3. 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 Comparative Analysis

We compare performance of our model to that of a set of baselines. Worse performance of the baselines compared to our proposed model can be attributed to the difference between their training processes. Though the baselines do not use the 2-step training process that our proposed model ($M_{sar+emo}^{KI}$) uses, all of the baselines developed have more parameter counts than our proposed model as the CLIP backbone of our model is not finetuned. It only acts as textual and visual feature extractor. The baselines are described below:

1. *CNN+VGG-19 ensemble*: We form an Convolutional Neural Network (CNN) (O'Shea and Nash, 2015) and VGG-19 (Simonyan and Zisserman, 2015) based ensemble model where textual part of the meme is encoded by a CNN model and VGG-19 is used to encode the visual part the meme.
2. *BiLSTM+VGG-19*: In similar fashion as the previous model, we add BiLSTM as our text encoder. This models is also trained end-to-end.
3. *mBERT+VGG-19*: In similar fashion as the previous model, we add mBERT as our text encoder. This models is also trained end-to-end.
4. *mBERT+ViT (concat)*: We build a mBERT[8] and ViT (Dosovitskiy et al., 2020) based baseline system. Textual and Visual portions from the memes are forwarded to mBERT and ViT respectively to extract textual and visual features. We concatenate those features and use an MLP at the end for classification of sarcasm. Only pre-trained weights are used in this stage without fine-tuning.
5. *mBERT+ViT (finetune)*: We further fine-tune the system elaborated in the previous point in an end-to-end setting.

All the system described above are trained for convergence with an early stopping threshold of maximum of 10 epochs on the validation set.

Baselines	Performance Metrics			
	re	pr	f1	acc
CNN+VGG-19	53.32	58.01	51.52	74.43
BiLSTM+VGG-19	51.52	51.52	51.52	51.52
mBERT+VGG-19	48.66	48.79	47.80	52.94
mBERT+ViT (concat)	54.09	56.87	53.39	58.11
mBERT+ViT (finetune)	57.31	58.01	57.18	63.59

Table 4: Performance of baseline models with respect to sarcasm detection. Our proposed models are statistically significant to all the baselines ($p<0.05$).

Chauhan et al. (2020) proposed a methodology which uses multitasking using sentiment and emotion labels to detect sarcasm in a multimodal setup, which obtained state-of-the-art in MUStARD dataset (Castro et al., 2019b). We repurpose their model for our task to compare against our proposed approach. We depict the result in Table 5

Objective	Performance Metrics			
	re	pr	f1	acc
sarcasm	17.44	33.61	23.12	52.11

Table 5: Performance for our sarcasm detection task using Chauhan et al. (2020) approach.

5.4 Detailed Analysis

To explain the feasibility of our proposed model, we perform a detailed quantitative and qualitative

[8] https://github.com/google-research/bert/blob/master/multilingual.md

		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 6: Sample test examples with predicted sarcasm label for STL and MTL models. Label definition: **2**: Highly Sarcastic, **1**: Mildly Sarcastic, **0**: Not Sarcastic.

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 7: 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 17.

T	T+V	Image	Text
×	✓		ओ 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.

analysis of some samples from the test set. In Table 6, 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 that MTL model with KI objective ($M^{KI}_{sar+emo}$) helps to capture related information from the meme to correctly predict the associated sarcasm class. For both STL and MTL setups, heatmaps of confusion matrices in shown in **Appendix** §A.8. From the confusion matrix, we identify the effectiveness of our proposed model.

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 classified by the multimodal (T+V) M_{sar} model. Similarly, 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. 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). This is discussed in detail in **Appendix** §A.12.

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 7, 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 six most common reasons why the models are failing to predict the actual class associated with the meme. (Refer **Appendix**, Table 17 for the corresponding memes.)

6 Conclusion

In this paper, we attempted to solve a challenging task of sarcasm detection in Internet memes. We have proposed a deep learning-based *multi-task 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 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 well enough in a few cases due to the lack of context knowledge. In the future, along with investigating new techniques in this direction, we will explore more about including background context to solve this problem more efficiently.

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

Petra Saskia Bayerl and Karsten Ingmar Paul. 2011. What determines inter-coder agreement in manual annotations? a meta-analytic investigation. *Computational Linguistics*, 37(4):699–725.

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

Katarina Boland, Andias Wira-Alam, and Reinhardt Messerschmidt. 2013. Creating an annotated corpus for sentiment analysis of german product reviews.

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. 2019a. Towards multimodal sarcasm detection (an _obviously_ perfect paper). *CoRR*, abs/1906.01815.

Santiago Castro, Devamanyu Hazarika, Verónica Pérez-Rosas, Roger Zimmermann, Rada Mihalcea, and Soujanya Poria. 2019b. 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.

Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. 2019. Racial bias in hate speech and abusive language detection datasets. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 25–35, Florence, Italy. Association for Computational Linguistics.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan S. Cowen, Gaurav Nemade, and Sujith Ravi. 2020. Goemotions: A dataset of fine-grained emotions. *CoRR*, abs/2005.00547.

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.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *CoRR*, abs/2010.11929.

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.

Saike He, Xiaolong Zheng, Jiaojiao Wang, Zhijun Chang, Yin Luo, and Daniel Zeng. 2016. Meme extraction and tracing in crisis events. In *2016 IEEE Conference on Intelligence and Security Informatics (ISI)*, pages 61–66.

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

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.

klaus krippendorff. 2011. Computing krippendorff's alpha-reliability.

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.

Emily Öhman. 2020. Emotion annotation: Rethinking emotion categorization. In *DHN Post-Proceedings*.

Keiron O'Shea and Ryan Nash. 2015. An introduction to convolutional neural networks. *ArXiv e-prints*.

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 visuolingual metaphor! *CoRR*, abs/2008.03781.

Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556.

Shardul Suryawanshi, Bharathi Raja Chakravarthi, Mihael Arcan, and Paul Buitelaar. 2020. Multimodal meme dataset (MultiOFF) for identifying offensive content in image and text. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 32–41, Marseille, France. European Language Resources Association (ELRA).

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.

A Appendix

A.1 Fine-grained emotion categories

In Table 9, we define all the 13 fine-grained emotion categories with the respective example which is defined in our dataset. In Table 8, We have mentioned a list of 13 emotional categories which are easily separable when we use their list of definitions.

Fear	Alarm, shock, fear, fright, horror, terror, panic, hysteria, mortification
Neglect	Alienation, isolation, neglect, loneliness, rejection, homesickness, defeat, dejection, insecurity, embarrassment, humiliation, insult, recklessness
Irritation	Aggravation, irritation, agitation, annoyance, grouchiness, grumpiness, frustration
Rage	Anger, rage, outrage, fury, wrath, hostility, ferocity, bitterness, hate, loathing, scorn, spite, vengefulness, dislike, resentment, betrayal
Disgust	Disgust, revulsion, contempt
Nervousness	Anxiety, nervousness, tenseness, uneasiness, apprehension, worry, distress, dread
Shame	Guilt, shame, regret, remorse
Disappointment	Dismay, disappointment, displeasure
Envy	Envy, jealousy
Suffering	Agony, suffering, hurt, anguish
Sadness	Depression, despair, hopelessness, gloom, glumness, sadness, unhappiness, grief, sorrow, woe, misery, melancholy
Joy	Amusement, bliss, cheerfulness, gaiety, glee, jolliness, joviality, joy, delight, enjoyment, gladness, happiness, jubilation, elation, satisfaction, ecstasy, euphoria, hope, humor
Pride	Pride, triumph

Table 8: Proposed emotions categories and list of keywords to define each emotionsin our dataset

(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.

Table 9: 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.

A.2 Removal of political and religious biases

Detecting and removing political and religious biases is an extensive research area. Any biases detected in our dataset are unintentional, and we have no intention of harming any individual or group. However, previous annotation studies show that we cannot correctly remove bias and subjectivity from the annotation process despite having some form of annotation scheme. We ensure that our data collection is generated equally and comparably in order to answer any political and religious biases queries. Furthermore, we ensure that the topic includes all potential issues in the Indian context over the previous seven years by using a keyword-based data-gathering technique. Moreover, we made sure that the terms included were inclusive of all conceivable politicians, political organizations, young politicians, extreme groups, and religions and were not prejudiced against any one group. Based on previous work done by Davidson et al. (2019) to remove biases from the dataset during annotation, in our dataset, annotators were also instructed not to make decisions based on what they believe but what the social media user wants to transmit through that meme.

A.3 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 4. 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 Figure 4. The meme says, *"Tag a friend who*

Test Sample 1

Highly sarcastic or Mildly sarcastic

Test Sample 2

joy or insult

Figure 4: Challenges during annotation

*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.

A.4 Dataset Statistics

Dataset statistics are presented in Table 10 and Table 11. We have also reported Inter-annotator agreement (IAA) of emotion class in Section 3.3.3. We noticed that IAA of emotion categories are relatively low but previous annotation tasks(Öhman, 2020; Bayerl and Paul, 2011; Boland et al., 2013) have shown that even with binary or ternary classification schemes, human annotators agree only about 70-80% of the time and the more categories there are, the harder it becomes for annotators to agree . Some emotions are also harder to detect and recognize. Demszky et al. (2020) shows how that the emotions of admiration, approval, annoyance and gratitude had the highest inter-rater correlation at around 0.6, and grief, relief, pride, nervousness, embarrassment had the lowest inter-rater correlations between 0-0.2, with a vast majority of emotions falling in the range of 0.3-0.5. Since in our work, we have used 13 fine-grained categories of emotion which is combination of explicit and contextual emotions, so we are getting relatively low IAA for emotion classification.

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

Table 10: 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 11: Emotion class distribution in our dataset

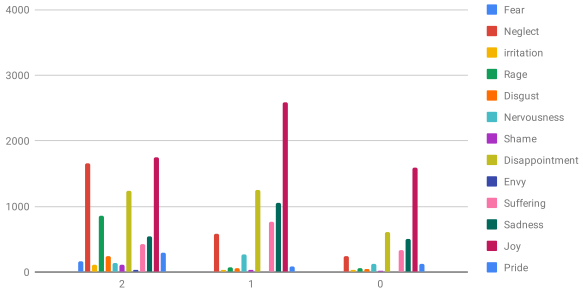


Figure 5: Distribution of fine-grained emotion categories for each sarcasm class(0,1,2). Refer Table 10 for label definition.

Non-sarcastic

Good people are not born again and again on earth..so respect me.

meme 1

The day when I stop eating and drinking after getting angry with the family members, on the same day the family members make peas and cheese.

meme 2

Turn on the fan on number 4 and take out the sheet on one of the sheets and the temperature maintenance season is going on.

meme 3

Mildly-sarcastic

online shopping,order delivery

meme 4

Look brothers!! Got solid proof today that girls always like donkeys....

meme 5

Just have to learn to be so tension free in life.

meme 6

Highly-sarcastic

Dense darkness looted convoy, a beauties, a p****!

meme 7

O Parth, shoot the arrow but on whom? Parth! You only shoot the arrow, Person XYZ himself will jump and take it in the middle.

meme 8

In 2019, the slogan of the saheb is "Leave studies. . . .Take wooden saddle", if you don't get a horse job, then sell pakodas.

meme 9

Table 12: Some data samples to understand all three categories of sarcasm

A.5 Ensemble Models

In this section, we describe different ensemble models built by weighted model averaging ensemble techniques. Instead of majority voting, we make inference by weighting the logit score of respective models (M_{sar}, $M_{sar+emo}$, M^{KI}_{sar}, $M^{KI}_{sar+emo}$). The weighting is determined on the basis of performance of the ensemble on the validation set. We analyse our ensemble 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.

To build the ensembles, we make use of all the models developed (with or without KI). To observe effects of KI technique, we form ensemble of the trained model with three setups, *viz* (i). *Ensemble with KI* (ens^{KI}) and (ii). *Ensemble without KI* (ens^{-KI}), (iii). *Ensemble with all* (ens^{all}).

In ens^{KI}, we only consider two models which were trained with knowledge infusion (KI). We consider predictions of models M^{KI}_{sar} and $M^{KI}_{sar+emo}$ 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 $+2.27\%$ in terms of F1-score (c.f Table 13). This also shows the effectiveness of our proposed KI scheme. Finally, we build an ensemble model (ens^{all}). This final model performs decently better than the 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 $+4.91\%$ in terms of F1-score.

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.

In Table 13, all these experimental results are shown. We can infer the following things from this Table: (i). *GRU* performs the best as a fusion mechanism compared to *KLD* and *cat+proj*. 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. (ii). Performance of the ensemble models generally follow the pattern: ens^{-KI} < ens^{KI} < ens^{all}. This is applicable for all of the three fusion techniques, i.e, *GRU*, *KLD* and *cat+proj*. (iii). We also empirically verify that *sar+sent* pre-training objective of the parent model could learn better representation (M'_t) than *sar* only pre-training objective, such that the performance of the student model increases.

A.6 Ablation Study

Ablation experiments mainly consist of 3 setups.
Firstly, we compare our training method to that of Sequential Finetuning.
Secondly, we show how KI-enabled models (M^{KI}_{sar} and $M^{KI}_{sar+emo}$) perform for different combinations of (i). parent pre-training objectives and (ii). Different KI fusion techniques.
Thirdly, we compare our proposed MFB module on top of CLIP with a simple concatenation followed by projection operation.

A.6.1 Setup 1

We compare whether sequential fine-tuning on our dataset after training on Memotion 2.0 could be as effective as the knowledge infusion (KI) based transfer learning setup. We perform the sequential fine-tuning using two setups:
1. KI-enabled fine-tuning: Instead of only fine-tuning sequentially, we also enable the KI mechanism using a GRU gating scheme.
2. KI-disabled fine-tuning: We only fine-tune sequentially, without enabling the KI mechanism using a GRU gating scheme. From Table 14, we see the sequential fine-tuning performance. As per the intuition, the model performance is in the following manner with respect to the F1 metric:
KI disabled+STL<KI enabled+STL<KI disabled+STL<KI enabled+MTL. Also we observe that,
(i). KI helps the sequential fine-tuning procedure.
(ii). Combining KI with sequential fine-tuning is not effective as only performing KI.

A.6.2 Setup 2

We tabulate our results for using different KI gating scheme in Table 15 under both *sar* and *sar+sent*

Ens.	KI Fusion	sent+sar				sent			
		re	pr	f1	acc	re	pr	f1	acc
ens^{KI}	*GRU*	63.73	65.56	63.97	67.11	63.10	65.00	63.33	66.50
	KL_div	62.63	64.91	62.81	66.23	62.04	64.48	62.27	65.49
	cat+proj	60.96	63.35	61.23	63.80	60.99	62.57	61.28	63.26
ens^{-KI}	-	61.61	63.69	61.70	65.29	61.61	63.69	61.70	65.29
ens^{all}	*GRU*	64.50	66.44	64.79	67.79	63.69	65.52	63.89	67.18
	KL_div	62.65	64.98	62.91	66.03	62.61	64.68	62.83	66.03
	cat+proj	62.66	64.39	62.95	65.62	62.32	63.98	62.55	65.56

Table 13: **Ensemble** result on variation of with KI and without KI ensembles and their effect on performance for the sarcasm head. *GRU obtains the best performance among all the fusion techniques and the gain is significant (p<0.01) compared to individual models (M_{sar}, $M_{sar+emo}$, M^{KI}_{sar}, $M^{KI}_{sar+emo}$).*

Obj.	Process	STL				MTL			
		re	pr	f1	acc	re	pr	f1	acc
KI enabled	*seq. finetuning*	**63.32**	**63.53**	**62.11**	**65.08**	**63.10**	63.92	**62.08**	**65.47**
KI disabled	*seq. finetuning*	60.25	61.78	60.34	63.73	61.5	**64.25**	61.74	65.02

Table 14: Performance of sequential fine-tuning process.

pretraining objective of the parent model. The following points can be drawn from the results shown in Table 15:
(i). *sar+sent* pre-training objective for the parent model is more effective for KI enabled models.
(ii). *GRU* consistently achieves better downstream performance compared to *KL_Div* and *cat+proj* techniques.

A.6.3 Setup 3

In Table 16, we test whether we could directly use the obtained textual and visual representation from the CLIP model and subsequently concatenate and 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 16. 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.

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

Table 17: Example memes shown in Table 7

A.7 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[9] framework. We use Adam(Kingma and 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

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

Obj.	KI Fusion	M^{KI}_{sar}				$M^{KI}_{sar+emo}$			
		re	pr	f1	acc	re	pr	f1	acc
sar	*GRU*	62.07	62.82	62.31	64.34	62.05	65.05	61.89	66.37
	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*	63.15	64.01	63.29	65.89	63.12	65.00	63.37	66.64
	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 15: Ablation results of two models *viz sar* only and *sar+sent* pretraining objective of parent model with different KI fusion methods. *Refer to Section A.6 for detailed description of sar+sent and sar training objective.*

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}.
4. *Random seed:* 123 for all experiments.

A.8 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^{KI}_{sar} 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}.

A.9 Training Graphs

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

A.10 Class-wise results for emotion recognition

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 18: 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 3, we show results for our secondary task of emotion recognition which is performed as a multilabel classification task.
In Table 18, we show class-wise evaluation result for each of the 13 emotion classes. All of the classes which gets poor class-wise performance has very less no. of (<50) test samples. Emotion Class *Suffering* has the highest number of test samples (1319), thus it obtains the highest performance.

A.11 Performance on Memotion 2.0 dataset

For knowledge infusion (KI), we have used Memotion 2.0 dataset to train the parent model. The parent model is trained to optimize for predicting sentiment class labels present in the dataset. In Table 19, we tabulate the results the parent model ob-

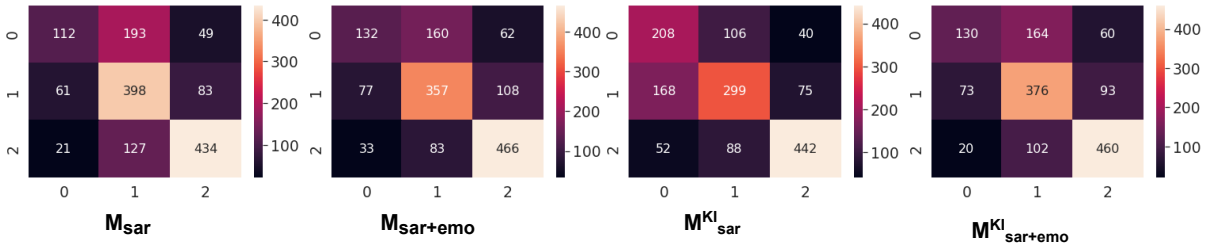


Figure 6: Heatmaps of the confusion matrix for four multimodal (T+V) models using both STL and MTL setup.

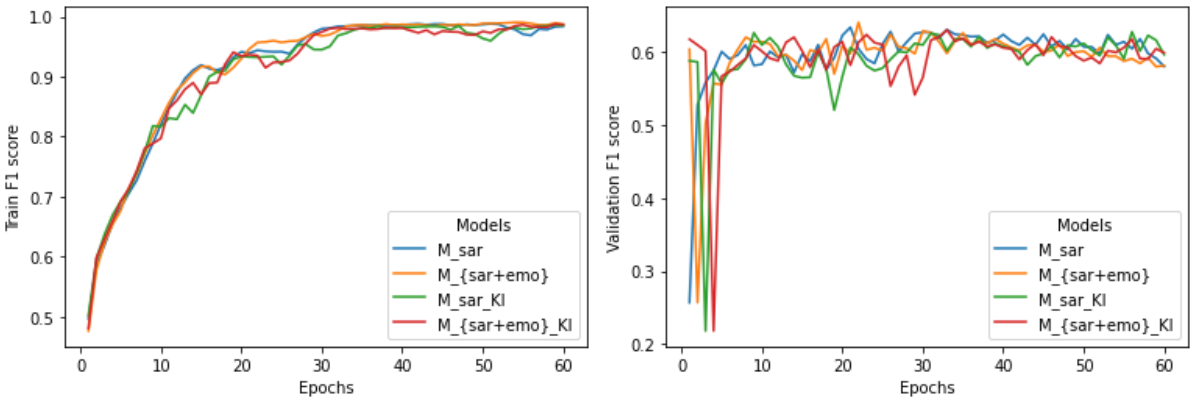


Figure 7: Training Graphs of all STL and MTL multimodal (T+V) models.

tains on Memotion 2.0 dataset for sentiment class labels.

Objective	Performance Metrics			
	re	pr	f1	acc
sentiment	32.92	32.18	32.19	51.8

Table 19: Performance of model on Memotion 2.0 dataset. No of sentiment class labels is 3.

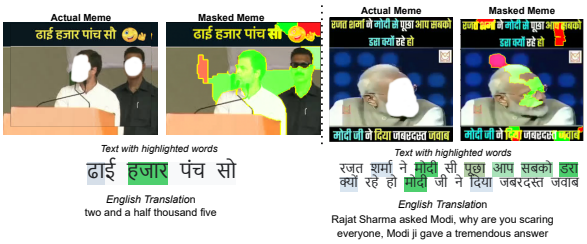


Figure 8: Examples showing visualization by LIME for multimodal (T+V) $M^{KI}_{sar+emo}$ model.

A.12 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 8, 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. Similarly 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*.

Note that faces are masked manually to remove identification of any well-known/political person in the paper. However, during training, we kept all the faces intact. Also note that upon acceptance of our paper, we aim to release our dataset in two steps.

Task	T				T+V				T+V (embedded)			
	re	pr	F1	acc	re	pr	F1	acc	re	pr	F1	acc
$M^{KI}_{sar+emo}$	58.15	58.32	58.19	60.14	63.88	63.07	62.05	66.15	63.11	65.01	63.37	66.64

Table 20: Performance of $M^{KI}_{sar+emo}$ on T(unimodal), T+V and T+V(embedded) settings.

Model	BERT+ViT				VisualBERT				M_{sar}			
	re	pr	F1	acc	re	pr	F1	acc	re	pr	F1	acc
sarcasm detection	28.19	27.21	25.68	40.70	27.50	24.90	24.27	40.97	26.91	28.32	24.43	51.03

Table 21: Comparing sarcasm detection performace on Memotion dataset by various baselines and our proposed approach.

(i). The version of the dataset that will be publicly available will not contain any mark of identification for political persons. All the faces will be masked out by automatic face recognition software. (ii). We will provide the actual dataset to the interested users by signing up a consent form. This dataset will contain actual faces without any morphing. It will be intended to be used for modeling purposes by the research community. Needless to say, we train our model on the actual dataset without the faces being masked.

A.12.1 Issues with LIME Visualization

For the examples showing LIME visualization in Figure 8, we observe that LIME sometimes focuses on textual part embedded in the meme. A meme generally contains embedded texts inside the image. We ask whether we should remove the embedded text or not for better downstream performance.To answer this question, we aim to perform an ablation study where the meme is modeled by considering two multimodal scenarios:

i) *T+V (embedded)*: This is the multimodal task where we do not remove the embedded text from the meme.

ii) *T+V*: In this multimodal scenario, the embedded text from the image part of the meme has been removed.

To guide our intuition that the texts embedded in the meme should help in classifying the meme into sarcastic/non-sarcastic class, we show performance of our best performing model ($M^{KI}_{sar+emo}$) on both *T+V (embedded) and T+V* on Table 20. In the same Table 20, we also show performance for sarcasm detection when only unimdodal text input (T) is considered. The F1 score is aligned with our intuition which follows the order:
T < T+V < T+V (embedded)

This shows that embedded text plays a key role in multimodality.

A.13 Sarcasm Performance on Memotion

In this section, we assess the capability of our proposed approach. We develop 2 state-of-the-art baselines i) BERT+ViT , ii)VisualBERT and compare their performance to that of our proposed model. This is to assess the generalization capability of our model to other datasets other than the dataset we propose in this paper. We observe that on Memotion 2.0 dataset, our model perform as par with state-of-the-art baselines. Needless to say, there are a lot less parameter in our proposed model than any of the baseline models developed.