CM-Off-Meme: Code-Mixed Hindi-English Offensive Meme Detection with Multi-Task Learning by Leveraging Contextual Knowledge

Gitanjali Kumari[1*], Dibyanayan Bandyopadhyay[1*], Asif Ekbal[1], Vinutha B N[2]
[1]Department of Computer Science and Engineering, Indian Institute of Technology Patna, Bihta, India,
[2]Wipro AI Labs, Bangalore, Karnataka, India
[1]{gitanjali_2021cs03, dibyanayan_2321cs14, asif}@iitp.ac.in, [2]vinutha.narayanmurthy@wipro.com

Abstract

Detecting offensive content in internet memes is challenging as it needs additional contextual knowledge. While previous works have only focused on detecting offensive memes, classifying them further into implicit and explicit categories depending on their severity is still a challenging and underexplored area. In this work, we present an end-to-end multitask model for addressing this challenge by empirically investigating two correlated tasks simultaneously: (i) offensive meme detection and (ii) explicit-implicit offensive meme detection by leveraging the two self-supervised pre-trained models. The first pre-trained model, referred to as the "knowledge encoder," incorporates contextual knowledge of the meme. On the other hand, the second model, referred to as the "fine-grained information encoder", is trained to understand the obscure psycho-linguistic information of the meme. Our proposed model utilizes contrastive learning to integrate these two pre-trained models, resulting in a more comprehensive understanding of the meme and its potential for offensiveness. To support our approach, we create a large-scale dataset, CM-Off-Meme, as there is no publicly available such dataset for the code-mixed Hindi-English (Hinglish) domain. Empirical evaluation, including both qualitative and quantitative analysis, on the CM-Off-Meme dataset demonstrates the effectiveness of the proposed model in terms of cross-domain generalization. The sample dataset and codes are available at this link: https://www.iitp.ac.in/~ai-nlp-ml/resources.html as well as at our GitHub repository: https://github.com/Gitanjali1801/CM_MEMES.git.

Keywords: Offensive, Memes, Code-Mixed, Multitask Learning, Contrastive Learning

1. Introduction

In recent years, the proliferation of memes on social media platforms like Facebook, Twitter, and Instagram has gained significant attention due to their widespread influence and potential to shape public discourse (Hossain et al., 2022a; Rijhwani et al., 2017; Sharma et al., 2020, 2022a; Suryawanshi et al., 2020). Despite being humorous, many memes use sarcasm and dark humor to promote societal harm (Kiela et al., 2020; Kirk et al., 2021; Kumari et al., 2021). Meme analysis, is, therefore, essential for detecting offensive content (Akhtar et al., 2022), analyzing psychological responses, etc. But, detecting offensiveness in memes is particularly challenging by automated models due to the relatively weak correlation between their textual and visual modalities, exacerbated by contextual complexities, subculture, and subjectivity (Sharma et al., 2020; Bandyopadhyay et al., 2023; Kirk et al., 2021). While prior research (although not large in number) has mostly focused on finding offensive memes, classifying them further into **explicit**[1] and **implicit** offensive categories based on their severity remains a difficult and understudied problem. We hypothesize in this research that offensive memes might be both explicit and implicit. While detecting explicit offensive memes is easier due to the presence of slur words and/or visual cues that frequently indicate profanity (Refer to meme samples (a) and (b) in Figure 1), detecting an implicitly offensive meme is challenging due to the need for the presence of confounding variables such as *Background context of the meme, mental state of the meme creator* (Refer to meme samples (c) and (d) in Figure 1[2]). Figure 1 (e) shows an example of an implicit offensive meme that says, "The world thinks Person XYZ defeated Congress, they don't know me.", which is not easy to detect. The meme lacks explicit elements in its text and image that would aid our model in recognizing its offensiveness. However, the meme creator's mental state, as indicated by negative sentiment, negative emotion, and the use of sarcasm, enhance the context of the meme to "ridicule a political leader." When incorporating this additional information, our model correctly identifies this meme as implicitly offensive.

Our proposed work is motivated by the aforementioned discussion, where we adopted two-phase training of the proposed model. The first phase, known as pre-training, equips a Fine-grained Encoder (FE) to capture fine-grained details like sarcasm, emotions, and sentiment within memes while enabling a Knowledge Encoder (KE) to gain a deeper understanding of meme ground truths. Subsequently, in the second phase, we introduce a multi-task classifier that leverages the learned representations from these pre-trained encoders. We jointly incorporate supervised

[*]These authors contributed equally to this work

[1]WARNING: This paper contains meme samples that are offensive in nature.

[2]Better image visibility through zooming throughout the paper.

LREC-COLING 2024, pages 3380–3393

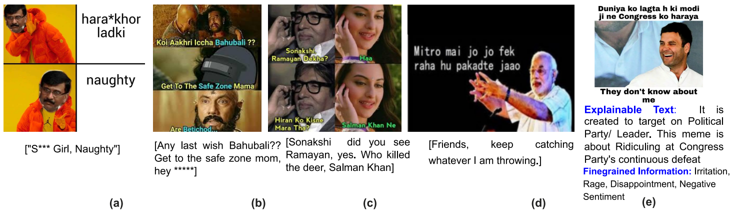

Figure 1: A few sample memes from our dataset for illustration of different types of offensive memes.

contrastive learning (SCL) and cross-entropy loss to optimize the training process further. These enhancements significantly bring instances of the same class closer in the semantic space and elevate the precision of our training methodology. In addition to this approach, we create a novel dataset of Code-Mixed Hindi-English (Hinglish) memes for four domains (i.e., political, religious, racist, and sexist). To assess the generalization capability of our model, we use leave-one-out cross-validation of domains and provide empirical evaluation results. The main **contributions** of this paper are as follows: ***(i) Dataset:*** A novel multimodal Hinglish dataset for identifying offensiveness in online memes, referred to as "CM-Off-Meme" (*Code-Mixed Hindi-English Offensive Meme*). ***(ii) Model:*** We introduce an end-to-end multitasking model, say "M_{FE}^{KE}", that effectively employs two pre-trained encoder models named (a) *knowledge encoder* and (b) *fine-grained information encoder* using supervised contrastive learning (CSL) to identify offensive memes and explicit and implicit offensive memes simultaneously. ***(iii) Analysis:*** Through an extensive empirical study conducted on the CM-Off-Meme dataset, we illustrate the effectiveness of our M_{FE}^{KE} model, with a focus on implicit offensive memes relative to baseline models.

2. Related Work

Hateful content detection. There has been quite a significant volume of prior research existing in the Natural Language Processing (NLP) community that focuses on detecting offensiveness, cyberbullying, hate speech, etc. in social media posts (Waseem and Hovy, 2016; Van Hee et al., 2018; Chatzakou et al., 2017; Chen et al., 2012; Roberts et al., 2012). There have been prior works (Wiegand, 2019; Kumar et al., 2018; Zampieri et al., 2019; Rosenthal et al., 2021) that focused on creating corpus and evaluation benchmarks for hate speech and offensiveness detection, but these are predominant in the English language. To address the challenge of predicting offensiveness in visual content only, a few attempts have also been made in the past few years (Duan et al., 2001; Fleck et al., 1996; Ganguly et al., 2017; Gandhi et al., 2019; Deselaers et al., 2008; Bosson et al., 2002; Gupta et al., 2022; Hu et al., 2007).

Multi-modality. Though most of the existing prior research on offensive content detection has primarily focused on the unimodal data (mainly on text only), incorporating multimodality (text with image), on the other hand, is still a work in progress (He et al., 2016; Hu and Flaxman, 2018; Sabat et al., 2019; Kumar et al., 2018; Tran and Cambria, 2018). The proliferation of memes and their expansion has recently attracted research on meme analysis. As a result, a few efforts have been put into meme analysis, such as focusing on hateful/offensive meme identification, etc. (Sharma et al., 2020; Kiela et al., 2020; Suryawanshi et al., 2020).

Code-mixing. Furthermore, most of the existing works in offensiveness detection in the code-mixed settings have been performed on textual data (Kamble and Joshi, 2018; Bali et al., 2014; Mathur et al., 2018; Tang et al., 2020; Bohra et al., 2018). Even though offensive meme identification for code-mixing among Dravidian languages (Tamil, Malayalam, Bengali, and Kannada) exists Hossain et al. (2022a), to the best of our knowledge, there is no publicly available dataset for English-Hindi (Hinglish) code-mixing. Following a thorough literature survey, we found no existing work uses psycho-linguistic aspects like sentiment, sarcasm, emotions, and the meme's context to determine offensiveness and identify explicit and implicit offenses in Hinglish memes. This encourages us to work in this particular domain, and the current is an initial effort to bridge this research gap.

3. Meme Corpus Creation

3.1. Data collection

we inlined our work with the existing meme analysis works (Sharma et al., 2020; Pramanick et al., 2021; Fersini et al., 2022) and used keyword-based searches to collect the publicly available memes using Google search[3]. We collected memes, which include keywords (c.f. Table 1) prominently used in India for the last 6-7 years for four domains: political, religious, racist, and sexist. It provided us with a total of 125 unique and globally popular categories. We finally retain only around 7K unique memes after removing the duplicates.

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

Political	Odd-even Rule, 2016 JNU incident, Demonetization, GST, Bihar liquor ban
Religious	Ayodhya dispute ,Fatwa,Beef ban,Hindu-muslim,,Love jihad
Racist	Darkisbeautiful, Anti Hindu, Citizenship Bill, Islamophobia, Intolerance, article370
Sexist	Dowry, LGBTQ, Aurat Azadi March, metoo, article377, No Acid, fake Feminism

Table 1: Offensive lexicons used to collect offensive memes

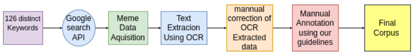

Figure 2: Data collection procedure

3.2. Annotation process

In establishing our annotation guidelines, we adopted a similar strategy to (Dimitrov et al., 2021; Bandyopadhyay et al., 2023). We divided our data annotation process into four phases: (i). Pre-processing and Text editing, (ii). Dry run, (iii). Final Annotation, and (iv). Consolidation. Our annotators, comprising AI professionals and linguists, covered a wide age range (20 to 45 years) and had a balanced gender representation. They were compensated at local rates and were explicitly instructed to remain politically and religiously neutral to ensure objectivity and avoid biases. Furthermore, to address bias, we took steps to ensure: i) The selected keywords include a broad spectrum of politicians, political organizations, young politicians, extremist groups, and religions without favoring any specific group/person, and ii) Annotators were guided to annotate memes based on the intended message by the social media user wants to deliver via that meme rather than personal beliefs.

Phase 1: Pre-processing and Text editing We manually filtered out (i) noisy memes with unclear backgrounds, (ii) non-code-mixed Hindi-English memes, and (iii) non-multi-modal memes. After that, 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 the annotators. Finally, we consider 6,967 memes for data annotation. The average meme text length for the meme samples in our dataset is 25 words (See the plot in Figure 3.)

Phase 2: Dry run This pilot stage included 200 annotated samples, which we annotated by ourselves for training annotators and quality control. We conducted a dry run on the same set to clarify label definitions and guidelines.

Phase 3: Final Annotation Following the dry run phase, we proceeded with the final annotation stage, where two annotators annotated each meme. We asked the annotators to annotate a given meme with the correct label of each layer as given in the annotation guidelines. After confirming the validity of the meme, we proceed toward the consolidation phase of the annotation.

Phase 4: Consolidation In this phase, the annotations from the final annotations are consolidated. This step was critical for maintaining quality and providing additional training for the entire team, which we found really beneficial. In the case of disagreements, we solved them by agreeing on a common point after many discussions.

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

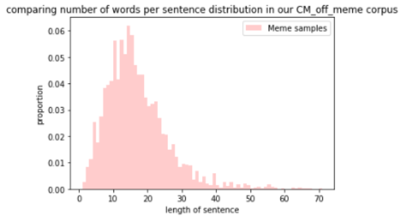

Figure 3: Distributions of the meme text length for the memes samples in our dataset

3.3. Annotation guidelines

Based on the context of memes, experts have annotated each meme with four labels: (i) Level 1: Offensive/non-offensive,(ii) Level 2: Explicit/implicit offensive, (iii) Level 3: Fine-grained information, i.e., (a) sarcasm (yes/no), (b) sentiment (positive, neutral, negative), (c) five primary emotions (each with yes/no), i.e., anger, fear, joy, sadness, and surprise. (iv) Level 4: Knowledge Text, i.e., ground truth level explanation for each meme.

(i) Level 1: Offensive/non-offensive: The offensive class has two labels: offensive and non-offensive.

Offensive meme: A meme will be categorized as offensive if it either explicitly or implicitly dehumanizes, degrades, insults, or attacks any individual or group based on attributes, such as gender, nationality, sexual orientation, ethnicity, race, skin color, health condition, otherwise non-offensive. To assess inter-rater agreement, we utilized Cohen's Kappa coefficient (Pat, 1987), a statistical metric. We attain a score of 0.7187 for this label, which shows a decent agreement between the annotators.

(ii) Level 2: Explicit/implicit offensive: Offensive class is further classified into explicit or implicit offensive.

Explicit offensive meme: A meme will be classified as explicitly offensive if it directly conveys offensiveness through text or image. For instance, it may exhibit offensiveness towards an individual or group in the image component or contain abusive language slurs or hints of offensive vocabularies such as threats, looting, killing, revenge, or imply a direct verbal assault against an individual or group (Refer to meme sample (a) and (b) in Figure 1)"

Implicit offensive meme: On the other hand, some memes may be covertly offensive. Although there are no slurs, negative sentiment-oriented words, or unpleasant visuals in the meme, if the implicit background knowledge/ underlying connotations/ implied meanings are considered, the meme becomes offensive to a person or a group. Ex: By employing metaphorical words/names like Pappu, chai wala, Shav-Sena, chowkidar, etc., or indirect references like Andhbhakt, chamcha for the blind followers of any political party, etc. (Refer to meme samples (c),(d) and (e) in Figure 1). We also acquire a Cohen's Kappa coefficient score of 0.8938 for this label.

(iii) Level 3: Fine-grained information: For the sentiment annotation, we annotate each meme based on the context. We annotate the dataset with three labels of sentiment as follows: *Positive*, *Neutral*, and *Negative*. For this level, we obtain Cohen's Kappa coefficient score of 0.6321.

Every meme sample is annotated with either one label of sarcasm: sarcastic or non-sarcastic. We attain a Cohen's Kappa coefficient score of 0.7152 for this label.

For the emotion annotation, each sample in our dataset is labeled with multiple labels of (at most three emotion classes, Emo1, Emo2, and Emo3) from the following primary emotion labels mentioned by Ekman and Cordaro (2011): anger, fear, joy, sadness, and surprise. For the emotion labels, the reported Krippendorff's Alpha Coefficient (krippendorff, 2011a) stands at 0.6174 in a multilabel context, which may appear relatively low. However, prior annotation tasks (Öhman, 2020; Bayerl and Paul, 2011; Boland et al., 2013) have demonstrated that human annotators tend to agree only around 70-80% of the time, even in scenarios with binary or ternary classification schemes. With an increasing number of categories, achieving higher agreement becomes more challenging. Given this, 0.6174 can be regarded as a strong score for inter-annotator agreement. **(iv) Level 4: Explainable Text:** All entities, including meme text, images, emojis, etc., along with the meme's context (Domain/ ground-truth reality), are to be considered for an appropriate explanation to annotate the meme for explainable/knowledge text. The minimum and maximum text length for it is set to a minimum of 5 words to a maximum of 30 words.

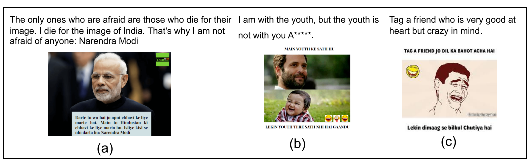

Figure 4: Sample dataset to show annotation challenges

3.3.1. Challenges During Annotations

Due to the obscure nature of memes, our annotators faced several challenges. The final class was chosen after agreeing on a common point after many discussions:

(i) Highly Opinionated Memes: Opinion-based memes from political domains are highly biased as they appear to be waging a covert campaign against the other party/leader, but they may not necessarily insult other political parties or leaders. Therefore, we annotate such memes as non-offensive (c.f. Figure 4(a)).

(ii) Funny emoticons with slur words : Memes sometimes contain offensive slur words and humorous emoticons simultaneously, making it challenging for annotation. For instance, in Figure 4(b), the presence of harsh slur words alongside humorous emoticons complicates annotation. We annotate such memes as explicitly offensive but also recognize the humorous emoticon by including "joy" in the emotion category.

(iii) Normalization of slur words: Some social media users use certain common words humorously, which has become a societal norm. For example, in Figure 4(c), a meme combines joy with slur words, making annotation challenging. It is unclear if the intention is to offend or express joy directly. We labeled memes non-offensive to align with current social media trends.

3.4. Dataset statistics and comparison with existing datasets

Our dataset, *CM-OFF-Meme*, comprises 6,967 annotated memes (c.f. Table 2) and provides a substantial resource for offensive meme research. It provides several unique advantages compared to existing datasets (c.f. Table 4) with diverse domains (political, religious, racist, sexist), Hinglish, and multimodal content (text and images) to examine offensiveness comprehensively in internet memes.

Out of domain test dataset collection We collected around 500 in Indian memes from the internet. We did not follow any particular domain to collect those memes as done for in-domain memes (c.f. Table 1). This is done to ensure that the training domains do not have anything in common with the collected memes. Further, after collection, two in-house annotators were used to filter out any memes where the training domains overlap (c.f Table 3 for the statistics of the out-of-domain dataset).

4. Methodology

We are given a set of meme samples $S \in \{T, I, E\}$, where each sample S_i includes text T_i, E_i includes meme explanation text and RGB image $I_i \in \mathbb{R}^{224\times224\times3}$. Our goal is to predict the correct label of each task, i.e., $\hat{y}_{t1} \subseteq$\{offensive, non-offensive\} and $\hat{y}_{t2} \subseteq$\{explicit, implicit\} for each S_i. The respective optimizing goal is then to learn the model weights θ and get the optimum loss $\mathcal{L}((\hat{y}_{t1}, \hat{y}_{t2}) \mid S_i, \theta)$. The overall architecture of our proposed model is shown in Figure 5. The components of our proposed architecture

Split	#Memes	Level 1		Level 2		Sentiment			Sarcasm		Emotions				
		Offensive	**Non-offensive**	**Explicit**	**Implict**	**Positive**	**Neutral**	**Negative**	**Yes**	**No**	**Fear**	**Joy**	**Surprise**	**Sadness**	**Anger**
Train	6000	4020	1980	2133	1887	961	2092	2947	3431	2569	332	1228	808	2576	410
Test	967	639	328	341	298	126	358	483	571	396	171	278	54	329	481

Table 2: Class wise data distribution of CM-OFF-meme dataset (Here test set is the in-Domain test dataset)

Level 1	#memes	Level 2	#memes
Non- offensive	352	implicit	87
Offensive	165	explicit	78

Table 3: Class-wise distribution of out-of-domain test set

Dataset	Domain	Language	Multimodal	Label	Statistics
COLD (Deng et al., 2022)	Open	Chineese	-	Offensive	37K
HASOC Fire 2020 (Mandl et al., 2021)	Open	English/German/Hindi	-	Offensive	3.7K/ 2.3K/ 2.9K
MultiOff (Suryawanshi et al., 2020)	U.S. Pre. Ele.	English	✓	Offensive	743
Hateful meme (Kiela et al., 2020)	Open	English	✓	Offensive	10K
HarMeme (Pramanick et al., 2021)	Open	English	✓	Harmful	3.5K
Memotion Analysis (Sharma et al., 2020)	Open	English	✓	Offensive	7K
MAMI (Fersini et al., 2022)	Misogynous	English	✓	Offensive	10K
MUTE (Hossain et al., 2022a)	Open	CM Eng-Ben	✓	Offensive	4K
CM-OFF-Meme (Ours)	P,R,Ra,S	Hinglish	✓	Offensive, Explicit/Implicit, Emotion, Sarcasm, Sentiment, Ground-truth Reality	6.9K

Table 4: Comparison of our dataset with some existing datasets. Here, *2016 U.S. Pre. Ele.*: U.S.Presidential Election, *CM Eng-Ben:* Code-Mixed English-Bengali, *Hinglish:* Code-Mixed Hindi-English, *P,R,Ra,S:* Political, Religious, Racist and Sexist

are discussed below.

4.1. Feature Extraction Layer

A meme sample S_i comprises of meme text $T_i = (t_{i_1}, t_{i_2}, \ldots, t_{i_k})$ and meme explanation text $E_i = (e_{i_1}, e_{i_2}, \ldots, e_{i_l})$, which are tokenized into sub-word units and projected into high-dimensional feature vectors, where k and l are the numbers of tokens in the meme text and explanation text respectively, and image I_i with regions $r_i = \{r_{i_1}, r_{i_2}, ..., r_{i_N}\}$; for $r_{i_j} \in R^N$, where N is the number of regions. These are then fed into a M3P[5] (Ni et al., 2020) pre-trained model designed to extract features by understanding text and images at a semantic level.

$$ft_i, fvt_i = M3P(t_i, r_i)\,; fe_i, fve_i = M3P(e_i, r_i)\,; \quad (1)$$

4.2. Multimodal Fusion

Our fusion module is based on Multimodal Factorized Bilinear pooling (MFB) (Yu et al., 2017).
Fusion between textual and visual features: This fusion module is comprised of two trainable weight matrices, W_1 and W_2. The following projection, followed by the sum-pooling operation, is performed in this layer.

$$M_{ti} = SumPool(W_1^T fvt_i \circ W_2^T fv_i(r)) \quad (2)$$

M_{ti} refers to the multimodal fusion between text and image.
Fusion between explanation and visual features: Another multimodal representation M_{ei} is created by passing explanation feature (fe_i) and visual features (fve_i) to another MFB module.

$$M_{ei} = SumPool(W_3^T fe_i \circ W_4^T fve_i) \quad (3)$$

4.3. Backbone Classifier

We use a fully connected layer (FCN)) with softmax activation, which takes the multimodal representation (M_{ti}) in Eq 2 as input and outputs class for Task 1 (offensiveness detection), shown in the following Equation 4:

$$\hat{y}_{t1} = P(Y_i|M_{ti}, W, b) = softmax(M_{ti}W_i + b_i) \quad (4)$$

Gating Mechanism. A non-offensive meme does not need further classification into corresponding implicit and explicit offensiveness categories. To address this, we use a gating mechanism to zero out M_{ti} when Task 1 predicts 'non-offensive.' This ensures that Task 2 gradient errors are only propagated for samples predicted as 'offensive.'

$$M_{ti}^{Masked} = Mask(M_{ti}, \hat{y}_{t1}) \quad (5)$$

Later, another FCN with softmax activation is used, which takes (M_{ti}^{Masked}) in Eq 5 as input and predicts the specific class of Task 2, i.e., $P(Y_i|M_{ti}^{Masked}, W_i, b_i)$ where W_i and b_i are the learnable weights and biases.:

$$\hat{y}_{t2} = softmax(M_{ti}^{Masked} W_i + b_i) \quad (6)$$

For both Task 1 and Task 2, we use categorical cross entropy as the loss function:

$$\mathcal{L}_{taski} = -\sum[y_{ti}\log\hat{y}_{ti} + (1 - y_{ti})\log(1 - \hat{y}_{ti})] \quad (7)$$

The final loss of our backbone multi-task model is computed by Equation 8:

$$\mathcal{L}_{classifier} = \mathcal{L}_{task1} + \mathcal{L}_{task2} \quad (8)$$

4.4. Pre-trained Encoders

4.4.1. Knowledge Enriched Encoder(KE)

This M3P-based pre-trained encoder predicts fine-grained information using explanations (E_i) and images (I_i). It classifies memes into sarcasm, sentiment, and multi-label emotion classes with task-specific layers. After training, we freeze the encoder's multimodal layers and use it to extract explanation-enriched hidden representations (h_{KEi}) for memes (S_i).

$$h_{KEi} = \{h_{k0}, h_{k1}, ..., h_{kH}\} \quad (9)$$

[5] https://github.com/microsoft/M3P

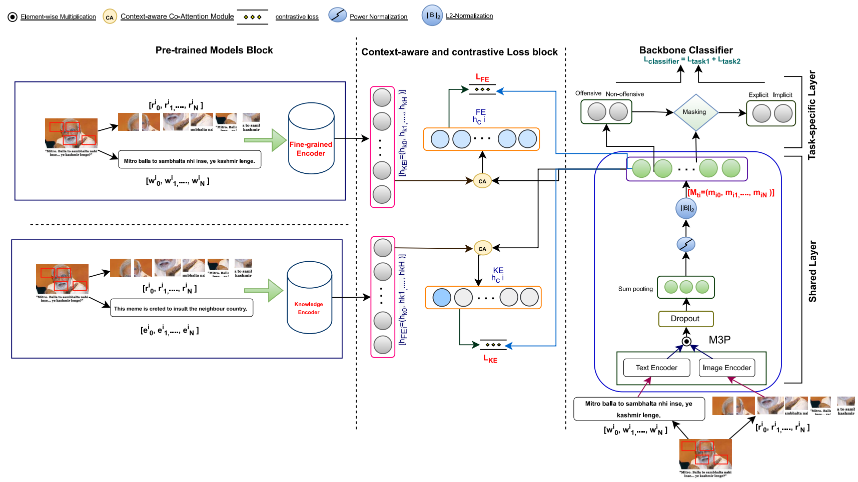

Figure 5: Our proposed multitask Model for Offensiveness identification

4.4.2. Fine-grained information Encoder (FE)

Like KE, the fine-grained information encoder model (FE) also learns to predict fine-grained hidden representation, but it takes the meme text (T_i) and an image (I_i) as input. For a given meme (S_i), we obtain another hidden representation h_{FEi} from our trained FE module.

$$h_{FEi} = \{h_{f0}, h_{f1}, ..., h_{fH}\} \tag{10}$$

4.5. Context-aware Co-Attention Module

To enhance the awareness of the offensive context in both encoder representations, we use a co-attention mechanism between hidden representations from both encoders and the extracted multimodal representation (M_{ti}). For a given hidden representation $h_{KE} \in \mathbb{R}^{(d\times H)}$ in Equation 10 and multimodal representation $M_{ti} \in \mathbb{R}^{(d\times M)}$ in Equation 2, at first we calculate an affinity matrix $A \in \mathbb{R}^{(H\times M)}$:

$$A = \tanh\left(h_{KE}^T W_b M_{ti}\right) \tag{11}$$

Afterward, we calculate the attention maps using the affinity matrix A in equation 11:

$$\begin{aligned} H_{h_{KE}} &= \tanh\left((W_t h_{KE} + (W_v M_{ti})A\right); \\ a^{h_{KE}} &= \text{softmax}(w_{KE}^T H_{h_{KE}}) \end{aligned} \tag{12}$$

Here, $W_t, W_v \in \mathbb{R}^{(k\times d)}$ and w_{KE}^T are weight matrix. $a^{h_{KE}}$ is the attention probability . After that, we calculate the attentive knowledge enriched representations h_{KE}^c, which is the weighted sum of h_{KE} feature.

$$h_{KEi}^c = \sum_{i=1}^{N} a^{h_{KE}} h_{KEi} \tag{13}$$

Similarly, for a given hidden representation h_{FEi} from the fine-grained encoder and multimodal representation M_{ti}, we calculate the context-aware fine-grained representation vector h_{FEi}^c.

4.6. Network Training

In addition to cross-entropy loss, we incorporate supervised contrastive loss (SCL) to enhance supervised learning and provide empirical evidence of its effectiveness in learning well-separated and equitable representations for each class (Shen et al., 2021; Li et al., 2023). The context-aware co-attentive representations from both the encoders (i.e., h_{KEi}^c ,h_{FEi}^c) and multimodal representations (M_{ti}) for a given meme (S_i) are assumed to describe similar contexts. These representations are aligned in the same semantic space to utilize both encoders effectively using CSL during training time.

$$\begin{aligned} \mathcal{L}_{KE} &= -\log \frac{\exp\left(\text{sim}\left(\boldsymbol{M_{ti}}, \boldsymbol{h^c_{KE}}_i\right)/\tau\right)}{\sum_{k=1,[k\neq i]}^{2N} \exp\left(\text{sim}\left(\boldsymbol{M_{ti}}, \boldsymbol{h^c_{KE}}_k\right)/\tau\right)} \\ \mathcal{L}_{FE} &= -\log \frac{\exp\left(\text{sim}\left(\boldsymbol{M_{ti}}, \boldsymbol{h^c_{FE}}_i\right)/\tau\right)}{\sum_{k=1,[k\neq i]}^{2N} \exp\left(\text{sim}\left(\boldsymbol{M_{ti}}, \boldsymbol{h^c_{FE}}_k\right)/\tau\right)} \end{aligned} \tag{14}$$

where N is the batch size, and τ is the temperature to scale the logits.

Now, to minimize the overall loss for the proposed model, $\mathcal{L}_{KE}$ and $\mathcal{L}_{FE}$ are combined along with categorical cross-entropy loss defined in Equation 7 for each task. $\mathcal{L}'_{taski} = \mathcal{L}_{taski} + \mathcal{L}_{FE} + \mathcal{L}_{KE}$. It makes the final loss of the proposed classifier as $\mathcal{L}'_{final}$ defined in the following equation 15:

$$\mathcal{L}'_{final} = \mathcal{L}'_{task1} + \mathcal{L}'_{task2} \tag{15}$$

4.7. Inference Objective

After training, our model generalizes over test data without pre-trained encoders. This design maintains performance without extra computational overhead, using the same loss as in Equation 8 during inference.

		Model	Modality		In-Domain test set				Out-of-Domain test set			
			T	I	Task 1		Task 2		Task 1		Task 2	
					Acc ↑	F1 ↑	Acc ↑	F1 ↑	Acc ↑	F1 ↑	Acc ↑	F1 ↑
Baseline	*S. Baselines*	*LSTM with Character level encoding (L_Char)*	✓		36.72	33.14	38.73	37.81	25.90	25.12	31.69	34.83
		LSTM with FastText (L_FT) (Bojanowski et al., 2016)	✓		38.91	36.21	34.81	30.83	28.09	28.19	27.77	27.85
		m-BERT (Pires et al., 2019)	✓		49.96	43.41	40.71	38.12	39.14	35.39	33.67	35.14
		VGG-19 (Simonyan and Zisserman, 2015)		✓	36.77	35.92	38.35	32.71	25.95	27.90	21.73	29.81
		ViT (Dosovitskiy et al., 2020)		✓	50.63	49.02	36.42	34.91	42.01	41.62	29.71	32.32
	M. Baselines	Char+VGG	✓	✓	49.01	47.38	33.62	32.42	38.19	39.36	26.58	29.44
		LSTM+VGG	✓	✓	39.15	41.47	40.63	35.61	28.33	33.45	33.59	32.63
		mBERT+ViT	✓	✓	55.91	50.81	38.18	35.49	39.91	39.90	31.14	33.12
		LXMERT (Tan and Bansal, 2019)	✓	✓	65.05	39.41	58.41	58.08	57.11	40.01	42.11	33.33
		VisualBERT (Li et al., 2019)	✓	✓	70.41	**69.45**	53.58	52.79	64.58	58.82	48.55	27.88
		mCLIP (Radford et al., 2021)	✓	✓	72.23	65.76	55.28	54.58	39.64	38.41	37.10	26.18
		BLIP (Li et al., 2022)	✓	✓	67.49	64.23	48.16	44.95	**65.82**	**61.24**	**63.28**	39.44
		ALBEF (Li et al., 2021)	✓	✓	67.46	62.71	49.21	47.83	49.02	47.90	49.21	36.46
Proposed & Ablation	*U. Abl*	$SU_{T}{}^{KE}_{FE}$	✓		59.21	51.84	48.72	44.62	59.03	51.81	47.93	45.63
		$SU_{V}{}^{KE}_{FE}$		✓	61.03	55.95	49.94	48.04	53.65	49.60	51.84	42.45
	S. Abl	M	✓	✓	61.35	58.6	56.01	52.03	50.53	50.58	48.97	49.05
		S^{FE}	✓	✓	**71.97**	63.11	53.43	54.41	61.15	55.09	46.39	51.43
		S_{KE}	✓	✓	69.07	66.29	57.74	55.78	58.25	58.27	50.70	52.80
		M^{KE}_{FE} (proposed)	✓	✓	70.94	67.11	**60.39**	**60.75**	60.12	59.09	53.35	**57.77**

Table 5: Results for Task 1 and Task 2 of *baselines*, *variations of the proposed model, shown as Ablation* and the proposed model. Note that each model is trained following a single-task learning setup. Here, *T*: Text, *I*: Image, *Task 1*: Offensive/Non-offensive, *Task 2*: Explicit/Implicit Offensive,*Acc*: Accuracy, *F1*: Macro F1 score.

		Model	Modality		In-Domain test set				Out-of-Domain test set			
			T	I	Task 1		Task 2		Task 1		Task 2	
					Acc ↑	F1 ↑	Acc ↑	F1 ↑	Acc ↑	F1 ↑	Acc ↑	F1 ↑
Baselines	*U. Baselines*	L_FT	✓		40.42	38.62	43.72	42.53	38.68	36.64	41.57	41.51
		L_Char	✓		50.63	49.05	47.15	44.61	47.59	47.07	43.83	43.82
		m-BERT	✓		65.92	64.56	46.83	41.74	64.43	62.58	41.03	40.78
		VGG-19		✓	58.16	49.43	49.43	47.23	56.71	47.45	46.72	46.17
		ResNet		✓	57.49	50.71	52.61	46.13	56.45	48.73	45.17	44.40
		ViT		✓	56.02	51.03	51.03	47.44	54.98	49.30	46.48	45.77
	M. Baselines	L_Char+VGG	✓	✓	53.41	49.81	49.61	43.11	52.37	47.83	42.15	42.09
		L_FT+VGG	✓	✓	42.15	48.71	46.41	45.23	41.11	47.73	44.27	44.21
		mBERT+ViT	✓	✓	70.33	68.83	45.08	45.95	**69.29**	**67.35**	43.99	43.93
		LXMERT	✓	✓	68.45	59.19	46.64	44.90	59.46	37.29	30.88	18.26
		VisualBERT	✓	✓	67.32	67.03	51.77	46.09	58.39	61.46	61.11	38.66
		mCLIP	✓	✓	72.12	66.34	54.53	53.88	41.01	40.09	38.67	27.45
		BLIP	✓	✓	70.26	64.04	48.65	46.98	63.67	60.43	**64.64**	42.00
		ALBEF	✓	✓	68.58	62.72	48.69	47.00	41.40	41.05	43.55	29.33
Ablation	*U. Abl*	$U_{T}{}^{KE}_{FE}$	✓		65.73	66.01	59.27	53.91	59.73	58.41	55.93	53.82
		$U_{V}{}^{KE}_{FE}$		✓	64.92	67.85	54.71	52.78	61.97	59.73	57.47	54.62
	M. Abl	M	✓	✓	68.14	65.25	68.14	58.88	60.42	57.17	58.72	55.35
		$M^{-gating}$	✓	✓	70.6	66.79	55.73	52.39	58.8	54.83	51.92	49.95
		M^{FE}	✓	✓	72.80	69.99	62.46	63.45	66.42	59.63	57.81	57.34
		M_{KE}	✓	✓	71.76	69.67	69.96	65.83	63.81	57.94	56.54	54.96
		M^{KE}_{FE} (proposed)	✓	✓	**73.42**	**70.33**	**67.25**	**67.37**	68.42	61.38	62.63	**60.84**

Table 6: Results for Task 1 and Task 2 of *baselines*, *variations of the proposed model, shown as Ablation* and the proposed model. Note that each model is trained following multitasking. *T*: Text, *I*: Image, *Task 1*: Offensive/Non-offensive, *Task 2*: Explicit/Implicit Offensive, *Acc*: Accuracy, *F1*: Macro-F1 score.

5. Experimental setups

5.1. Implementation Details

We evaluate our proposed architecture on our curated dataset. The optimal hyperparameters for our model are found using grid search. We chose the same set of hyperparameters to maintain consistency over all the experiments performed. We employ M3P with XLM-R (Conneau et al., 2019) tokenizer, which includes 250K BPE tokens and covers 100 languages. We use Adam optimizer (Kingma and Ba, 2015) with a learning rate of 3e-5 , $\beta1 = 0.9$, $\beta2 = 0.999$, $\epsilon = 10^{-8}$ for all the models. We train the model for 60 epochs with 64 batch sizes and early-stopping callback. A single NVIDIA Tesla GPU is used to conduct the experiments. To evaluate the model's generalization capability, we employ two types of test sets: *(i) In-domain test set*, and *(ii) Out-of-domain test set.* Details of the distribution of each test set are given in Table 2 and 3.

6. Results and Analysis

6.1. Model Result and Comparison

Main results. In Table 5 and Table 6, we show the results of our proposed model M^{KE}_{FE} and its unimodal and multimodal variations in single-task learning (STL) and multitask learning (MTL) scenarios on in-domain and out-of-domain test sets.
i) ***Performance of baseline:*** Multimodal baselines with MTL consistently perform better compared to unimodal and STL baselines for both Task-1 and Task-2 by a margin of 15-17 % F1 score. Notably, the mCLIP-based model outperformed other baselines in STL and MTL scenarios for both test sets, forming the foundation for our proposed method.
ii) ***Performance of the proposed system:*** The proposed model (M^{KE}_{FE}) performs better compared to the developed baselines consistently for both Task-1 and Task-2 for the in-domain test set. M^{KE}_{FE} performs at par VisualBERT and mBERT+ViT for the out-of-domain test set. Improvements over the best baselines for the respective tasks are statistically significant (t-test with a $p < 0.05$).
Ablations. To test the proposed architecture, we develop unimodal and multimodal variants of our proposed model. **i) *Unimodal variations:*** We develop two unimodal models, i) $U_{T}{}^{KE}_{FE}$: Here, only the textual part of the meme is considered, ii) $U_{V}{}^{KE}_{FE}$: Only visual part of the meme is considered. Note that unimodal variations ($U_{T}{}^{KE}_{FE}$, $U_{V}{}^{KE}_{FE}$) perform poorly compared to the proposed model across in-domain and out-of-domain test sets for both Task-1 and Task-2. **ii) *Multimodal variations:*** We develop four multimodal variations of our proposed model: i) M: This is our proposed model trained without pre-trained encoders (both FE and KE) and the contrastive loss. ii) $M^{-gating}$: This variation is trained without the gating mechanism and without pre-trained encoders. iii) M^{FE}: This model is only trained with the fine-grained encoder (FE). iv) M_{KE}: This model is only trained with the knowledge encoder (KE). Comparing the performance of the ablation models, M^{KE}_{FE} stands out as the most effective in detecting offensive memes in terms of all the matrices. This can be attributed to its effective utilization of both the encoders with integrated CSL loss.

6.2. Detailed Result Analysis

6.2.1. Qualitative Analysis with Case Study

Using Figure 6, we qualitatively analyze our proposed framework through the predictions obtained from different configurations of our proposed model. All the samples of Figure 6 have the gold label 'offensive.' M classifies In-domain test sample (a) as Non-offensive. This meme is implicitly offensive despite the absence of slurs and the existence of a non-offensive image because it mocks a specific political figure in context.

This context is correctly recognized by KE and FE, and our proposed model M_{FE}^{KE} classifies it as implicitly offensive. The only way to classify the in-domain test sample (b) is by applying both KE and FE. Without context ('India-Pakistan rivalry') and fine-grained information ('Negative sentiment, sarcasm'), it is impossible for M to determine whether or not this meme is offensive.

We also present two out-of-domain test instances in which our proposed model M_{FE}^{KE} accurately classifies an implicitly offensive meme. Its success can be linked to contextual relevance modeled by KE and the incorporation of fine-grained information modeled by FE. In Figure 7, we present four offensive memes from four different domains. Due to distinct training and test domains, all of these memes have been wrongly categorized as non-offensive. By including fine-grained information and context, even if the training and testing domains are distinct, all of these memes are accurately labeled as offensive. This demonstrated our proposed model's domain generalization capacity.

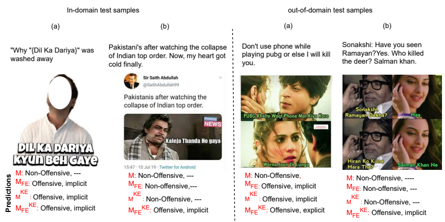

Figure 6: Case studies of the proposed model for in-domain and out-of-domain test sets. For every example meme, we show its translation at the top.

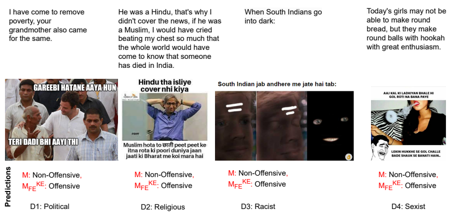

Figure 7: Case studies of the proposed model for domain generalization. For every example meme, we show its translation at the top.

6.2.2. Modality Importance

As shown in (Kiela et al., 2021), both textual and visual modalities act together to decipher the offensiveness of a meme. To analyze the multimodality's effectiveness, we also qualitatively analyzed the prediction from the unimodal and proposed model.

Failure of textual modality. Left-most meme in Figure 8 (a) is classified as non-offensive by $U_{T}{}_{FE}^{KE}$. Incorporating visual modality (which shows a man with a knife) along with text in the multimodal model helps it classify the meme as offensive correctly.

Failure of visual modality. Similarly, in the right-most meme of Figure 8 (a), $U_{V}{}_{FE}^{KE}$ fails to detect offensiveness, whereas M_{FE}^{KE} model correctly classifies it as offensive. The text modality provides information on intent and meaning through keywords, phrases, sentiments, emotions, and sarcasm.

Figure 8: **Modality Importance (a)** Test cases where unimodal systems (either text-only model $U_{T}{}_{FE}^{KE}$ or image-only model $U_{V}{}_{FE}^{KE}$) fail to correctly predict whereas proposed multimodal model M_{FE}^{KE} effectively predicted the offensive class. **Error Analysis (b)** Test cases where proposed multimodal model M_{FE}^{KE} fails

6.2.3. Domain Generalization

We tested our model's cross-domain generalizability by training it on three domains and evaluating it on others, showing results in Table 7 (Macro-F1-score). As an illustration, the first row of the table illustrates the results of training the model on domains $D1$, $D3$, and $D4$ and testing it on a domain $D2$. Model M_{FE}^{KE}, with both KE and FE, consistently outperforms other models across unseen domains. This highlights our model's cross-domain adaptability, which is essential for real-world applications.

	Task 1				**Task 2**			
	M	M_{FE}^{KE}	M_{FE}	M_{KE}	M	M_{FE}^{KE}	M_{FE}	M_{KE}
$D1 \cup D3 \cup D4$	69.99	72.8	70.18	71.34	62.46	67.96	65.73	64.25
$D2 \cup D3 \cup D4$	69.67	71.76	69.96	69.83	63.45	65.83	62.39	63.37
$D1 \cup D4 \cup D2$	68.34	71.62	69.46	72.45	62.46	67.45	62.46	63.45
$D1 \cup D2 \cup D3$	65.54	69.26	67.96	68.83	60.15	66.72	63.81	61.11

Table 7: Generalization over Domains. D1: Political, D2: Religious, D3: Racist, D4: Sexist domain data samples (In terms of F1-score)

6.3. Comparison with existing works

The results presented in Table 8 demonstrate the superiority of our proposed model, attributed to its effective utilization of contextual and fine-grained information. Notably, our proposed model M_{FE}^{KE} outperforms almost all existing models across all metrics, presenting a significant advancement for both the tasks.

6.3.1. Explainability and Diagnostics

Once our M_{KE}^{FE} model is trained, we use LIME (Locally Interpretable Model-Agnostic Explanations) to diagnose the model's prediction (Ribeiro et al., 2016). In Figure 9, we can see that for both the given test samples, either certain image regions (e.g., a person's face) or specific words in the text which contribute prominently

Models	In Domain test set				Out-of-Domain test set			
	Acc↑	F1↑	Acc↑	F1↑	Acc↑	F1↑	Acc↑	F1↑
	Task 1		Task 2		Task 1		Task 2	
Zhou and Chen (2020)	66.53	61.95	46.64	42.83	55.37	52.83	42.05	40.42
Chauhan et al. (2020)	65.38	62.41	44.81	42.91	48.53	42.57	40.61	39.72
Hossain et al. (2022b)	63.93	58.02	42.56	41.75	52.74	49.04	43.64	39.92
Sharma et al. (2022b) (i)	67.94	59.42	56.93	53.32	59.03	56.93	50.62	47.31
Sharma et al. (2022b) (ii)	66.53	57.03	58.82	52.71	62.72	55.84	54.38	50.85
M_{FE}^{KE} **(Ours)**	73.42	**70.33**	67.25	**67.37**	**68.42**	**61.38**	**62.63**	**60.84**

Table 8: Comparison of our proposed model with existing models. Performance improvement in our proposed model is statistically significant wrt all the existing models ($p<0.05$)

to the correct prediction significantly influence M_{KE}^{FE}'s accurate predictions. In contrast, the baseline *M* model struggles to utilize the contextual and fine-grained information effectively, lacking the ability to recognize offensive intent.

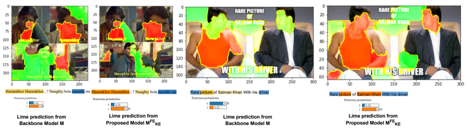

Figure 9: Visualization by LIME for baseline model M and proposed model M_{FE}^{KE}.

7. Error Analysis

Despite its high performance, our proposed model M_{KE}^{FE} still misclassifies several instances. To gain insight into these errors, we identify key reasons for misclassifications by our M_{KE}^{FE}:
(i) Overgeneralization of a few slur words: Misclassifications of non-offensive memes as offensive due to overgeneralization of certain humor-related slurs. (c.f. sample 1, Table 8(b)),
(ii) Lack of common sense knowledge: instances where the model sometimes fails to reason intuitively about everyday situations, causing misclassification (sample 2, Table 8(b)), and
(iii) Model overfitting of contextual knowledge: situations where the model overfitted and misclassifies non-offensive memes as offensive by merely the presence of a few phrases like "Mann ki Baat" (inner thoughts). (c.f. sample 3 in Table 8(b)).

8. Conclusion and Future Work

In summary, in this work, we introduce a novel end-to-end multitasking framework for offensive meme identification in Hinglish memes. Our proposed framework leverages memes' contextual knowledge and psycholinguistic aspects using two pre-trained encoders: (i) knowledge encoder and (ii) fine-grained information encoder. Subsequently, we use these encoders to create a robust classifier with state-of-the-art performance. We have performed a detailed qualitative evaluation to show the effectiveness of our approach. In future work, we plan to investigate ways to incorporate dynamic contextual knowledge in the meme classification framework in an unsupervised manner, to make the proposed model more robust and effective.

Limitations

In this paper, we discussed an effective end-to-end model for offensiveness detection in memes. While this model includes a novel knowledge encoder and a fine-grained information encoder, which subsequently obtains state-of-the-art performance for the newly created in-domain and out-of-domain Hinglish dataset, this work has some limitations. A detailed discussion of a few limitations is discussed in Section 7. In future research, we aim to address this limitation by exploring ways to improve the model's understanding of memes by incorporating more robust common sense knowledge.

Ethics and Broader Impact

Individual Privacy: Our study used publicly available memes, adhering to copyright laws and gaining Institutional Review Board (IRB) approval. We plan to make our code and data accessible for research purposes, subject to appropriate data agreement procedures, upon acceptance of our study. In this paper, we protected individuals' anonymity by replacing real names with "Person-XYZ" and anonymized faces in memes.
Biases: Detecting political and religious biases is a complex research area. Prior annotation studies have revealed challenges in completely eliminating bias and subjectivity from the annotation process, even with established annotation schemes. We want to clarify that any biases that may be identified in our dataset are unintentional, and we have no intent to harm individuals or groups. We have taken steps to ensure that our data collection is impartial and balanced, addressing potential political and religious bias concerns. To ensure relevance to the Indian context over the past seven years, we utilized a keyword-based data-collection approach. We also ensured that the keywords encompassed many political organizations, emerging leaders, extremist groups, and religions without favoring any specific group. Inlined with (Davidson et al., 2019) in bias reduction during annotation, we instructed our annotators to base their decisions not on personal beliefs but on the intended message conveyed by the social media user through each meme.
Misuse Potential: We suggest that researchers be aware that people could use the dataset we have created to filter memes unfairly based on their own prejudices or beliefs. To avoid this scenario, it is crucial to have human oversight and moderation.
Intended Use: Our dataset is designed to help researchers study offensive memes online. We hope it will be a valuable resource for researchers who use it responsibly.

Acknowledgements

The research reported in this paper is an outcome of the project "**HELIOS: Hate, Hyperpartisan, and Hyperpluralism Elicitation and Observer System,**" sponsored by Wipro AI Labs, India.

9. Bibliographical References

1987. Quantification of agreement in psychiatric diagnosis revisited. In *Archives of general psychiatry 44 2*.

Shad Akhtar, Deepanway Ghosal, Asif Ekbal, Pushpak Bhattacharyya, and Sadao Kurohashi. 2022. All-in-one: Emotion, sentiment and intensity prediction using a multi-task ensemble framework. *IEEE Transactions on Affective Computing*, 13:285–297.

Dibyanayan Bandyopadhyay, Gitanjali Kumari, Asif Ekbal, Santanu Pal, Arindam Chatterjee, and Vinutha BN. 2023. A knowledge infusion based multitasking system forăsarcasm detection inămeme. In *Advances in Information Retrieval*, pages 101–117, Cham. Springer Nature Switzerland.

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.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*.

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

Despoina Chatzakou, Nicolas Kourtellis, Jeremy Blackburn, Emiliano De Cristofaro, Gianluca Stringhini, and Athena Vakali. 2017. Mean birds: Detecting aggression and bullying on twitter. In *Proceedings of the 2017 ACM on Web Science Conference*, WebSci '17, page 1322, New York, NY, USA. Association for Computing Machinery.

Dushyant Singh Chauhan, Dhanush S R, Asif Ekbal, and Pushpak Bhattacharyya. 2020. Sentiment and emotion help sarcasm? a multi-task learning framework for multi-modal sarcasm, sentiment and emotion analysis. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4351–4360, Online. Association for Computational Linguistics.

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

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale.

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.

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.

Margaret M. Fleck, David A. Forsyth, and Christoph Bregler. 1996. Finding naked people. In *ECCV*.

Nikhil Ghanghor, Parameswari Krishnamurthy, Sajeetha Thavareesan, Ruba Priyadharshini, and Bharathi Raja Chakravarthi. 2021. IIITK@DravidianLangTech-EACL2021: Offensive language identification and meme classification in Tamil, Malayalam and Kannada. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 222–229, Kyiv. Association for Computational Linguistics.

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)*, page 6166. IEEE Press.

Eftekhar Hossain, Omar Sharif, and Mohammed Moshiul Hoque. 2022a. MUTE: A multimodal dataset for detecting hateful memes. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing: Student Research Workshop*, pages 32–39, Online. Association for Computational Linguistics.

Eftekhar Hossain, Omar Sharif, Mohammed Moshiul Hoque, M. Ali Akber Dewan, Nazmul Siddique, and

Md. Azad Hossain. 2022b. Identification of multilingual offense and troll from social media memes using weighted ensemble of multimodal features. *Journal of King Saud University - Computer and Information Sciences*, 34(9):6605–6623.

Anthony Hu and Seth Flaxman. 2018. Multimodal sentiment analysis to explore the structure of emotions. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery &; Data Mining*, KDD '18, page 350358, New York, NY, USA. Association for Computing Machinery.

Weiming Hu, Ou Wu, Zhouyao Chen, Zhouyu Fu, and Stephen J. Maybank. 2007. Recognition of pornographic web pages by classifying texts and images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29:1019–1034.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2021. The hateful memes challenge: Detecting hate speech in multimodal memes.

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

Hannah Kirk, Yennie Jun, Paulius Rauba, Gal Wachtel, Ruining Li, Xingjian Bai, Noah Broestl, Martin Doff-Sotta, Aleksandar Shtedritski, and Yuki M Asano. 2021. Memes in the wild: Assessing the generalizability of the hateful memes challenge dataset. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 26–35, Online. Association for Computational Linguistics.

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

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

Gitanjali Kumari, Amitava Das, and Asif Ekbal. 2021. Co-attention based multimodal factorized bilinear pooling for Internet memes analysis. In *Proceedings of the 18th International Conference on Natural Language Processing (ICON)*, pages 261–270, National Institute of Technology Silchar, Silchar, India. NLP Association of India (NLPAI).

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation.

Junnan Li, Ramprasaath R. Selvaraju, Akhilesh Deepak Gotmare, Shafiq Joty, Caiming Xiong, and Steven Hoi. 2021. Align before fuse: Vision and language representation learning with momentum distillation.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language. In *Arxiv*.

Yingji Li, Mengnan Du, Xin Wang, and Ying Wang. 2023. Prompt tuning pushes farther, contrastive learning pulls closer: A two-stage approach to mitigate social biases. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14254–14267, Toronto, Canada. Association for Computational Linguistics.

Minheng Ni, Haoyang Huang, Lin Su, Edward Cui, Taroon Bharti, Lijuan Wang, Jianfeng Gao, Dongdong Zhang, and Nan Duan. 2020. M3p: Learning universal representations via multitask multilingual multimodal pre-training.

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

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

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.

Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should I trust you?": Explaining the predictions of any classifier. *CoRR*, abs/1602.04938.

Shruti Rijhwani, Royal Sequiera, Monojit Choudhury, Kalika Bali, and Chandra Shekhar Maddila. 2017. Estimating code-switching on Twitter with a novel generalized word-level language detection technique. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1971–1982, Vancouver, Canada. Association for Computational Linguistics.

Kirk Roberts, Michael A. Roach, Joseph Johnson, Josh Guthrie, and Sanda M. Harabagiu. 2012. EmpaTweet: Annotating and detecting emotions on Twitter. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 3806–3813, Istanbul, Turkey. European Language Resources Association (ELRA).

Benet Oriol Sabat, Cristian Canton-Ferrer, and Xavier Giró i Nieto. 2019. Hate speech in pixels: Detection of offensive memes towards automatic moderation. *ArXiv*, abs/1910.02334.

Shivam Sharma, Md Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. 2022a. DISARM: Detecting the victims targeted by harmful memes. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1572–1588, Seattle, United States. Association for Computational Linguistics.

Shivam Sharma, Mohd Khizir Siddiqui, Md. Shad Akhtar, and Tanmoy Chakraborty. 2022b. Domain-aware self-supervised pre-training for label-efficient meme analysis. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 792–805, Online only. Association for Computational Linguistics.

Aili Shen, Xudong Han, Trevor Cohn, Timothy Baldwin, and Lea Frermann. 2021. Contrastive learning for fair representations.

Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*.

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).

Hao Tan and Mohit Bansal. 2019. Lxmert: Learning cross-modality encoder representations from transformers.

Ha Nguyen Tran and Erik Cambria. 2018. Ensemble application of ELM and GPU for real-time multimodal sentiment analysis. *Memetic Comput.*, 10(1):3–13.

Cynthia Van Hee, Gilles Jacobs, Chris Emmery, Bart Desmet, Els Lefever, Ben Verhoeven, Guy De Pauw, Walter Daelemans, and Véronique Hoste. 2018. Automatic detection of cyberbullying in social media text. *PLOS ONE*, 13(10):1–22.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on Twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93, San Diego, California. Association for Computational Linguistics.

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

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. Predicting the type and target of offensive posts in social media. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1415–1420, Minneapolis, Minnesota. Association for Computational Linguistics.

Yi Zhou and Zhenhao Chen. 2020. Multimodal learning for hateful memes detection.

10. Language Resource References

Kalika Bali, Jatin Sharma, Monojit Choudhury, and Yogarshi Vyas. 2014. "I am borrowing ya mixing ?" an analysis of English-Hindi code mixing in Facebook. In *Proceedings of the First Workshop on Computational Approaches to Code Switching*, pages 116–126, Doha, Qatar. Association for Computational Linguistics.

Aditya Bohra, Deepanshu Vijay, Vinay Singh, Syed Sarfaraz Akhtar, and Manish Shrivastava. 2018. A dataset of Hindi-English code-mixed social media text for hate speech detection. In *Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media*, pages 36–41, New Orleans, Louisiana, USA. Association for Computational Linguistics.

Alison Bosson, Gavin C. Cawley, Yi Chan, and Richard Harvey. 2002. Non-retrieval: Blocking pornographic images. In *CIVR*.

Jiawen Deng, Jingyan Zhou, Hao Sun, Chujie Zheng, Fei Mi, Helen Meng, and Minlie Huang. 2022. COLD: A benchmark for Chinese offensive language detection. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11580–11599, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Thomas Deselaers, Lexi Pimenidis, and Hermann Ney. 2008. Bag-of-visual-words models for adult image classification and filtering. In *2008 19th International Conference on Pattern Recognition*, pages 1–4.

Dimitar Dimitrov, Bishr Bin Ali, Shaden Shaar, Firoj Alam, Fabrizio Silvestri, Hamed Firooz, Preslav Nakov, and Giovanni Da San Martino. 2021. Detecting propaganda techniques in memes. In *Proceedings of the 59th Annual Meeting of the Association* *for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6603–6617, Online. Association for Computational Linguistics.

Lijuan Duan, Guoqin Cui, Wen Gao, and Hongming Zhang. 2001. Adult image detection method base-on skin color model and support vector machine.

Elisabetta Fersini, Francesca Gasparini, Giulia Rizzi, Aurora Saibene, Berta Chulvi, Paolo Rosso, Alyssa Lees, and Jeffrey Sorensen. 2022. SemEval-2022 task 5: Multimedia automatic misogyny identification. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 533–549, Seattle, United States. Association for Computational Linguistics.

Shreyansh Gandhi, Samrat Kokkula, Abon Chaudhuri, Alessandro Magnani, Theban Stanley, Behzad Ahmadi, Venkatesh Kandaswamy, Omer Ovenc, and Shie Mannor. 2019. Image matters: Detecting offensive and non-compliant content / logo in product images. *CoRR*, abs/1905.02234.

D. Ganguly, M. H. Mofrad, and A. Kovashka. 2017. Detecting sexually provocative images. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 660–668, Los Alamitos, CA, USA. IEEE Computer Society.

Kshitij Gupta, Devansh Gautam, and Radhika Mamidi. 2022. cvil: Cross-lingual training of vision-language models using knowledge distillation.

Eftekhar Hossain, Omar Sharif, and Mohammed Moshiul Hoque. 2022. MUTE: A multimodal dataset for detecting hateful memes. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing: Student Research Workshop*, pages 32–39, Online. Association for Computational Linguistics.

Satyajit Kamble and Aditya Joshi. 2018. Hate speech detection from code-mixed hindi-english tweets using deep learning models. *CoRR*, abs/1811.05145.

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. In *Advances in Neural Information Processing Systems*, volume 33, pages 2611–2624. Curran Associates, Inc.

Ritesh Kumar, Atul Kr. Ojha, Shervin Malmasi, and Marcos Zampieri. 2018. Benchmarking aggression identification in social media. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, pages 1–11, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Thomas Mandl, Sandip Modha, Anand Kumar M, and Bharathi Raja Chakravarthi. 2021. Overview of the hasoc track at fire 2020: Hate speech and offensive language identification in tamil, malayalam, hindi, english and german. In *Proceedings of the 12th Annual Meeting of the Forum for Information Retrieval Evaluation*, FIRE '20, page 2932, New York, NY, USA. Association for Computing Machinery.

Puneet Mathur, Rajiv Shah, Ramit Sawhney, and Debanjan Mahata. 2018. Detecting offensive tweets in Hindi-English code-switched language. In *Proceedings of the Sixth International Workshop on Natural Language Processing for Social Media*, pages 18–26, Melbourne, Australia. Association for Computational Linguistics.

Shraman Pramanick, Dimitar Dimitrov, Rituparna Mukherjee, Shivam Sharma, Md. Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. 2021. Detecting harmful memes and their targets. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2783–2796, Online. Association for Computational Linguistics.

Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Marcos Zampieri, and Preslav Nakov. 2021. SOLID: A large-scale semi-supervised dataset for offensive language identification. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 915–928, Online. Association for Computational Linguistics.

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! In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 759–773, Barcelona (online). International Committee for Computational Linguistics.

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).

Tiancheng Tang, Xinhuai Tang, and Tianyi Yuan. 2020. Fine-tuning bert for multi-label sentiment analysis in unbalanced code-switching text. *IEEE Access*, 8:193248–193256.

Michael Wiegand. 2019. *GermEval-2018 Corpus (DE)*. heiDATA.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. SemEval-2019 task 6: Identifying and categorizing offensive language in social media (OffensEval). In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 75–86, Minneapolis, Minnesota, USA. Association for Computational Linguistics.