DISARM: Detecting the Victims Targeted by Harmful Memes

Anonymous ACL submission

Abstract

Internet memes have emerged as an increasingly popular means of communication on the web. Although memes are typically intended to elicit humour, they have been increasingly used to spread hatred, trolling, and cyberbullying, as well as to target specific individuals, communities, or society on political, socio-cultural, and psychological grounds. While previous work has focused on detecting harmful, hateful, and offensive memes in general, identifying whom these memes attack (i.e., the 'victims') remains a challenging and underexplored area. We attempt to address this problem in this paper. To this end, we create a dataset in which we annotate each meme with its victim(s) such as the name of the targeted person(s), organization(s), and community(ies). We then propose DISARM (Detecting vIctimS targeted by hARmful Memes), a framework that uses named-entity recognition and person identification to detect all entities a meme is referring to, and then, incorporates a novel contextualized multimodal deep neural network to classify whether the meme intends to harm these entities. We perform several systematic experiments on three different test sets, corresponding to entities that are (i) all seen while training, (ii) not seen as a harmful target while training, and (iii) not seen at all while training. The evaluation shows that DISARM significantly outperforms 10 unimodal and multimodal systems. Finally, we demonstrate that DISARM is interpretable and comparatively more generalizable and that it can reduce the relative error rate of harmful target identification by up to 9% absolute over multimodal baseline systems.

1 Introduction

Social media platforms offer the freedom and the means to express deeply ingrained sentiments, which can be done using diverse and multimodal content such as memes. Besides being popularly used to express benign humour, Internet memes are also misused to incite extreme reactions, hatred, and to spread disinformation on a massive scale. Numerous recent efforts have attempted to characterize harmfulness (Pramanick et al., 2021b), hate speech (Kiela et al., 2020), offensiveness (Suryawanshi et al., 2020), etc. within memes. Most of these efforts have been directed towards detecting such malicious influence within memes, but there has been little work on identifying *whom the memes target*. Besides detecting whether a meme is harmful, it is often important to know whether the meme contains an entity that is particularly targeted in a harmful way. This motivates us to address the problem of detecting the entities that meme targets in a harmful way.

(a) Harmful reference **(b)** Harmless reference

Figure 1: (a) A meme that targets Justin Trudeau in a *harmful* way, with a communal angle. (b) A *non-harmful* mention of Justin Trudeau, as a benign humor.

The harmful targeting in memes is often done using satirical, sarcastic, or humorous elements. This involves either explicit or implicit ways to imply harm. Such stealth techniques are often used to implicate an individual, an organization, a community, or society, in general. For example, Fig. 1a depicts Justin Trudeau as *communally biased* – *against* Canadians – while favoring alleged *killings by* Muslims, whereas Fig. 1b shows a benign meme expressing subtle humour. Essentially, the meme in Fig. 1a *harmfully* targets *Justin Trudeau* directly, while causing indirect harm to *Canadians* and to *Muslims* as well. Also, a large number of memes require some additional background context for holistic comprehension. Hence, some challenges that indicate how intricate it is for an automated system to accurately detect harmful targeting in memes are the following: (i) insufficient *background context*, (ii) complexity posed by the *implicit* harm, and (iii) keyword *bias* in a supervised setting.

We aim to address the task of harmful target detection from memes by posing it as an open-ended task. The end-to-end solution primarily requires (i) identification of the entities mentioned within a meme, and (ii) a multimodal framework that helps in detecting whether the referenced entity is being harmfully targeted in a given meme. Essentially, we perform systematic contextualization of the multimodal information presented within memes, by first performing intra-modal fusion between external knowledge-based *contextualized-entity* and *embedded-harmfulness* in memes. This is followed by cross-modal fusion of contextualized textual and visual modalities using low-rank bi-linear pooling, as a contextualized-multimodal feature. We evaluate using three-level stress testing towards assessing their generalizability.

We aim to address the aforementioned requirements, and we make the following contributions[1]:

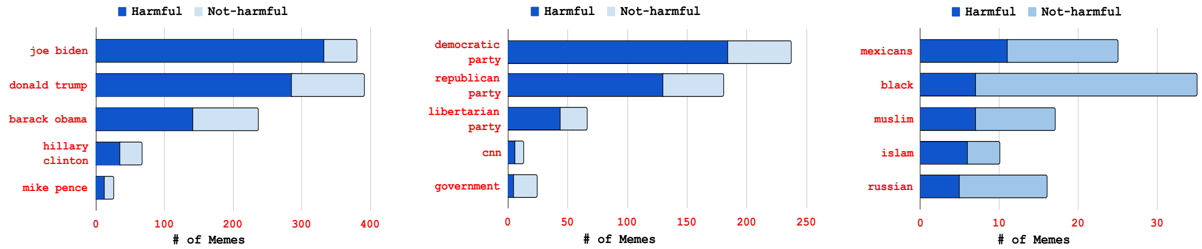
1. We introduce a novel task of detecting harmful targets within a meme.
2. We create a new dataset, by extending Harm-P (Pramanick et al., 2021b) via re-annotating the memes for the fine-grained entities they target.
3. We propose `DISARM`, a novel multimodal neural architecture that models contextualized multimodal features, towards detecting the harmful targeting in memes.
4. We empirically showcase that `DISARM` outperforms 10 unimodal and multimodal baselines by 4%, 7%, and 13% increment in the macro-F1 scores in three different evaluation setups.
5. We finally discuss `DISARM`'s generalizability and interpretability.

2 Related Work

Misconduct on Social Media. The rise in misconduct on social media has brought a range of related studies under active investigation. Some forms of online misconduct include rumors (Zhou et al., 2019), fake news (Aldwairi and Alwahedi, 2018; Shu et al., 2017), misinformation (Ribeiro et al., 2021), disinformation (Alam et al., 2021), hate speech (MacAvaney et al., 2019a; Zhang and Luo, 2018), trolling (Cook et al., 2018), and cyber-bullying (Kowalski et al., 2014; Kim et al., 2021). Some notable work in this direction includes stance (Graells-Garrido et al., 2020) and rumour veracity prediction, explored in a multi-task learning framework (Kumar and Carley, 2019), wherein the authors proposed a Tree LSTM for characterizing online conversations. Wu and Liu (2018) explored user and social network feature embeddings towards classifying a message trajectory as genuine vs. fake. User's mood along with the online contextual discourse was studied by Cheng et al. (2017) to demonstrate better modelling for trolling behaviour prediction in contrast with using just the user's behavioural history. Relia et al. (2019) studied the synergy between discrimination based on race, ethnicity and national origin in the physical and in the virtual space.

Studies Focusing on Memes. Recent efforts have shown interest in incorporating extra contextual information for meme analysis. Shang et al. (2021a) proposed knowledge-enriched graph neural networks that use common-sense knowledge for offensive memes detection. Pramanick et al. (2021a) focused on detecting COVID-19 related harmful memes and highlighted the challenge of inherent biases within existing multimodal systems. Pramanick et al. (2021b) further released another dataset for US Politics and proposed a multimodal framework for harmful meme detection. The Hateful Memes detection challenge by Facebook (Kiela et al., 2020) introduced the task of classifying a meme as either hateful or non-hateful. Different approaches such as feature augmentation, attention mechanism, and multimodal loss re-weighting were attempted (Das et al., 2020; Sandulescu, 2020; Zhou and Chen, 2020; Lippe et al., 2020). Sabat et al. (2019) studied hateful memes by highlighting the importance of visual cues such as structural template, graphic modality, causal depiction, etc. Interesting approaches such as web-entity detection along with fair face classification (Karkkainen and Joo, 2021) and semi-supervised learning-based classification (Zhong, 2020) were also used for the hateful meme classification task. Other noteworthy work includes implicit models and topic modelling of multimodal

[1]The source codes and dataset are uploaded in the supplementary.



(a) Individual **(b)** Organization **(c)** Community

Figure 2: Comparison plots of top-5 *harmfully referenced* entities, for their harmful/not-harmful referencing in our dataset.

Split	# Samples	Category	
		Harmful	Not-harmful
Train	3618	1206	2412
Validation	216	72	144
Test	612	316	296

Table 1: Summary of Ext-Harm-P

cues for detecting offence analogy (Shang et al., 2021b) and hatefully discriminatory (Mittos et al., 2020) memes. Wang et al. (2020) argued that online attention can be garnered immensely via fauxtography content, which could eventually evolve towards becoming memes that go viral. Several datasets including the ones about offence, hate speech, harmfulness, etc. have been proposed (Suryawanshi et al., 2020; Kiela et al., 2020; Pramanick et al., 2021a,b; Gomez et al., 2019).

Most of these studies attempt to address classification tasks in a constrained setting. However, to the best of our knowledge, none of them addressed the task of detecting the specific targets of hate, offence, harm, etc. We intend to explore precisely this task in this work for harmful memes.

3 Dataset

The Harm-P dataset (Pramanick et al., 2021b) consists of $3,552$ US politics memes. Each meme is annotated with its harmful label and the social entity that it targets. The target entities are coarsely classified into four social groups – individual, organization, community, and the general public. While these coarse classes provide an overall nature of targets, we feel the need to identify the targeted person, organization, or community in a fine-grained fashion. All the memes in this dataset are on the same topic, and they target well-known personalities or organizations. To this end, we manually re-annotated this dataset with the name of the persons, the organizations, and the communities that the harmful memes target.

Extending Harm-P (Ext-Harm-P). Towards generalizability, we extend Harm-P by reformulating existing train/test splits, as shown in Table 1. We call the resulting dataset Ext-Harm-P. For training, we use the *harmful* memes provided as part of the original annotations in the dataset (Pramanick et al., 2021b) and re-annotate them for the fine-grained entities being targeted harmfully as positive examples (*harmful* targets). This is matched with twice as many negative examples (*not-harmful* targets). For negative targets, top-2 entities that have the highest lexical similarity with the meme text are selected (Ferreira et al., 2016). This ensures very similarly, if not the same (due to OCR-induced noise) entities referenced within a meme, thereby facilitating a confounding effect (Kiela et al., 2020) as well. The *overall* test set is created by considering *all* entities referenced within memes. Entities are first extracted automatically using names entity recognition (NER) and person identification (PID)[2]. This is followed by manual annotation of the test set to address noisy assignments.

Data Annotation After extracting the entities automatically, we manually annotate the test set memes by refining the noisy entities with the help of detailed annotation guidelines. Additional details about the annotation process are included in Appendix D

Analyzing Harmful Targeting in Memes. Since all memes in Ext-Harm-P are about *US Politics*, a large number of them refer popular entities like *Joe Biden* and *Donald Trump*, both harmfully and harmlessly. For such harmful references, the trade-off with their harmless counterparts is observed to vary across *individuals*, *organizations*, and *communities* categories, as shown in Fig. 2. The top-5 harmfully referenced

[2]NER using SpaCy & PID using http://github.com/ageitgey/face_recognition.

individuals and *organizations* are observed to be subjected to a higher amount of harm, as against the support they garner. This could be due to infrequent reaction from such high profile entities, to online targeting. In contrast, the stacked plots for the top-5 harmfully targeted communities (Fig. 2c) either depict relatively higher support or harmless referencing/discussion on social media for *communities* like *Mexicans, Black, Muslim, Islam*, and *Russian*.

4 Proposed Approach

DISARM, as depicted in Fig. 3, models the fusion of textual and visual modalities, explicitly enriched via contextualised representations by leveraging CLIP Radford et al. (2021). At first, valid entities are extracted automatically, are part of the train/val set creation. Then for each meme, we first obtain the *contextualized-entity* (CE) representation by fusing the CLIP (Radford et al., 2021) encoded context and the entity representation. CE is then fused with BERT-based (Devlin et al., 2019) *embedded-harmfulness* (EH) encoding fine-tuned over OCR-extracted text and entities as inputs. We call the fusion output *contextualized-text* (CT) representation. CT is then fused with the *contextualized-image* (CI) representation, obtained using the CLIP encoder for image. We, henceforth, refer to the resulting representation as the *contextualized multimodal* (CMM) representation. We slightly modify multimodal low-rank bi-linear pooling (Kim et al., 2017), to fuse joint embedding space representations of input features. This approach not only captures complex cross-modal features, but also provides an efficient fusion mechanism towards obtaining context-enriched features. Finally, CMM is used to train a classification head for our task. We describe each module in more detail below.

Low-rank Bi-linear Pooling (LRBP). We begin by revisiting *low-rank bi-linear pooling* to set the necessary background. Due to many parameters in bi-linear models, Pirsiavash et al. (2009) suggested a low-rank bi-linear (LRB) approach to reduce the rank of the weight matrix $\mathbf{W}_i$. Consequently parameters, and hence the complexity is reduced. The weight matrix $\mathbf{W}_i$ is re-written as $\mathbf{W}_i = \mathbf{U}_i\mathbf{V}_i^T$, where $\mathbf{U}_i \in \mathbb{R}^{N\times d}$ and $\mathbf{V}_i \in \mathbb{R}^{M\times d}$, effectively putting an upper bound of $\min(N, M)$ on the value of d. Therefore, the low-rank bi-linear models can be expressed as follows:

$$f_i = \mathbf{x}^T\mathbf{W}_i\mathbf{y} = \mathbf{x}^T\mathbf{U}_i\mathbf{V}_i^T\mathbf{y} = \mathbb{1}^T(\mathbf{U}_i^T\mathbf{x} \circ \mathbf{V}_i^T\mathbf{y}) \quad (1)$$

where $\mathbb{1} \in \mathbb{R}^d$: column vector of ones, and $\circ$: Hadamard product. f_i in Equation 1 can be further re-written to obtain $\mathbf{f}$ as follows:

$$\mathbf{f} = \mathbf{P}^T(\mathbf{U}^T\mathbf{x} \circ \mathbf{V}^T\mathbf{y}) + \mathbf{b} \quad (2)$$

where $\mathbf{f} \in \{f_i\}$, $\mathbf{P} \in \mathbb{R}^{d\times c}$, $\mathbf{b} \in \mathbb{R}^c$. d and c: output and LRB hyper-parameters.

Following (Kim et al., 2017), we introduce a non-linear activation based formulation for the LRBP. Kim et al. (2017) argued that non-linearity both before and after the Hadamard product complicates the gradient computation. This, addition to Equation 2, can be represented as follows:

$$\mathbf{f} = \mathbf{P}^T\tanh(\mathbf{U}^T\mathbf{x} \circ \mathbf{V}^T\mathbf{y}) + \mathbf{b} \quad (3)$$

We slightly modify multimodal low-rank bi-linear pooling (MMLRBP). Instead of directly projecting the input $\mathbf{x} \in \mathbb{R}^N$ and $\mathbf{y} \in \mathbb{R}^M$ to a lower dimension d, we first project the input modalities in a joint space (N). We then perform LRBP as expressed in Equation 3, by using jointly embedded representations $\mathbf{x}_{mm} \in \mathbb{R}^{N\times d}$ and $\mathbf{y}_{mm} \in \mathbb{R}^{N\times d}$ to obtain a multimodal fused feature $\mathbf{f}_{mm}$, as expressed below:

$$\mathbf{f}_{mm} = \mathbf{P}^T\tanh(\mathbf{U}^T\mathbf{x}_{mm} \circ \mathbf{V}^T\mathbf{y}_{mm}) \quad (4)$$

Structured Context. Towards modelling auxiliary knowledge, we curate *contexts* for the memes in Ext-Harm-P. First, we use meme text as the search query[3] to retrieve relevant contexts. We treat the title and the first paragraph from the top resulting document, towards modelling the required *context* and represent it as *con*.

Contextualized-entity Representation (CE). Towards modelling the context enriched entity, we first obtain the embedding of a given entity *ent*. Since we have a finite set of entities referenced in the memes in our dataset, we perform a lookup in the embedding matrix ($\in \mathbb{R}^{V\times H}$) to obtain the corresponding entity embedding $\mathbf{ent} \in \mathbb{R}^H$, with $H = 300$ being the embedding dimension and V is the vocabulary size. The embedding matrix is jointly trained from *scratch*, during training. We project the obtained entity

[3] https://pypi.org/project/googlesearch-python/

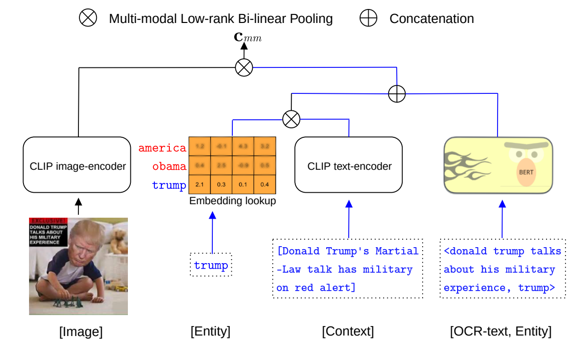


Figure 3: Architecture of DISARM (our proposed approach). $\mathbf{c}_{mm}$ is the multimodal feature used for classification.

representation **ent** into 512 dimensional space, and we call it **e**. To augment a given entity with relevant contextual information, we fuse it with contextual representation $\mathbf{c} \in \mathbb{R}^{512}$, obtained by encoding the associated context (*con*) using CLIP text-encoder (Radford et al., 2021). We perform this fusion using our adaptation of multimodal low-rank bi-linear pooling (Equation 4). This gives contextualized-entity (CE) representation $\mathbf{c}_{ent}$ as shown below:

$$\mathbf{c}_{ent} = \mathbf{P}_1^T \tanh(\mathbf{U}_1^T \mathbf{e} \circ \mathbf{V}_1^T \mathbf{c}) + \mathbf{b} \quad (5)$$

where $\mathbf{c}_{ent} \in \mathbb{R}^{512}$, $\mathbf{P}_1 \in \mathbb{R}^{256\times512}$, $\mathbf{b} \in \mathbb{R}^{512}$, $\mathbf{U}_1 \in \mathbb{R}^{512\times256}$ and $\mathbf{V}_1 \in \mathbb{R}^{512\times256}$.

Contextualized-Text (CT) Representation. Once we obtain the contextualized-entity embedding $\mathbf{c}_{ent}$, we concatenate it with the BERT encoding for the combined representation of the OCR-extracted text and the entity ($\mathbf{o}_{ent} \in \mathbb{R}^{768}$). We call this encoding *embedded-harmfulness* (EH) representation. The concatenated feature $\in \mathbb{R}^{1280}$ is then projected non-linearly into a lower dimension using a dense layer of size 512. We term the resultant vector $\mathbf{c}_{txt}$ as *contextualized-text* (CT) representation.

$$\mathbf{c}_{txt} = \mathbf{W}_i[\mathbf{o}_{ent}, \mathbf{c}_{ent}] + b_i \quad (6)$$

where $\mathbf{W} \in \mathbb{R}^{1280\times512}$.

Contextualized Multimodal (CMM) Representation. Once we obtain the contextualized-text representation $\mathbf{c}_{txt} \in \mathbb{R}^{512}$, we again perform multimodal low-rank bi-linear pooling using Equation 4 to fuse it with the contextualized-image representation $\mathbf{c}_{img} \in \mathbb{R}^{512}$, obtained using CLIP image-encoder (Radford et al., 2021). The operation is expressed as

$$\mathbf{c}_{mm} = \mathbf{P}_2^T \tanh(\mathbf{U}_2^T \mathbf{c}_{txt} \circ \mathbf{V}_2^T \mathbf{c}_{img}) \quad (7)$$

where $\mathbf{c}_{mm} \in \mathbb{R}^{512}$, $\mathbf{P}_2 \in \mathbb{R}^{256\times512}$, $\mathbf{U}_2 \in \mathbb{R}^{512\times256}$ and $\mathbf{V}_2 \in \mathbb{R}^{512\times256}$. Notably, we learn two different projection matrices $\mathbf{P}_1$ and $\mathbf{P}_2$, for the two fusion operations performed as part of Equations 5 and 7, respectively since the fused representations at the respective steps are obtained using different modality-specific interactions.

Classification Head. Towards modelling the binary classification for a given meme and a corresponding entity as either harmful or non-harmful, we use a shallow multi-layer perceptron with a single dense layer of size 256, which represents a condensed representation for classification. We finally map this layer to a single dimension output via a sigmoid activation. We use binary cross-entropy for the back-propagated loss.

5 Experiments

We train DISARM and all unimodal baselines using PyTorch and multimodal baselines using the MMF framework[4] [5]. We experiment with various state-of-the-art unimodal (image/text-only) and multimodal baseline systems, including the ones that are pre-trained using multimodal datasets such as MS COCO (Lin et al., 2014) and CC (Sharma et al., 2018). For evaluation, we use commonly used metrics such as accuracy, precision, recall (including their class-wise scores) along with F1 score, and we macro-average them. The harmful class recall is relevant for our study as it characterizes the model performance, towards detecting *harmfully* targeting memes correctly. The results reported are averaged across five independent runs.

Evaluation Strategy. Towards examining a realistic setting, we pose our evaluation strategy as an open-class one. We train all the systems with the set having positive (harmful) samples and twice as many negative (not-harmful) samples. We then evaluate using open-class testing, for all referenced entities (some possibly unseen during training) per meme, effectively making the evaluation more realistic. To this end, we formulate three testing scenarios as follows, with their Harmful (H) and Not-harmful(N) sample counts:

(a) **Test set A (316H, 296NH)** – Includes examples with entities *seen* during training.

[4] github.com/facebookresearch/mmf
[5] Additional details along with the hyper-parameters are reported in Appendix A.

(b) **Test set B (27H, 94NH)** – The examples in this set correspond to the entities that are *unseen* as *harmful*, during training.

(c) **Test set C (16H, 76NH)** – Only entities that are *unseen* as either harmful or not-harmful during the training are considered.

Baseline Models. Our baselines include both unimodal and multimodal models as follows:

– *Unimodal Systems*: ▶ **VGG16, VIT:** For the unimodal (image-only) systems, we use two well-known models: VGG16 (Simonyan and Zisserman, 2015) and VIT (Vision Transformers) that emulate a Transformer based application jointly over textual tokens and image patches (Dosovitskiy et al., 2021). ▶ **GRU, XLNet:** For the unimodal (text-only) systems, we use GRU (Cho et al., 2014), which adaptively captures temporal dependencies, and XLNet (Yang et al., 2020), which implements a generalized auto-regressive pre-training strategy.

– *Multimodal Systems*: ▶ **MMF Transformer:** This is a multimodal Transformer model that utilizes visual and language tokens with self-attention[6]. ▶ **MMBT:** Multimodal Bitransformer (Kiela et al., 2019) captures the intra-modal and the inter-modal dynamics of the two modalities. ▶ **ViLBERT CC:** Vision and Language BERT (Lu et al., 2019), pre-trained for conceptual captions (Sharma et al., 2018) based pretext task, is a strong model with task-agnostic joint representation of images and text. ▶ **Visual BERT COCO:** Visual BERT (Li et al., 2019), pre-trained on the MS COCO dataset (Lin et al., 2014).

Experimental Results. We compare the performance of several unimodal and multimodal systems (pre-trained and otherwise) and DISARM along-with its variants. All systems are evaluated using the 3-way testing strategy described above. We then perform ablation studies over *contextualized-entity*, its fusion with *embedded-harmfulness* resulting into *contextualized-text* and the final fusion with *contextualized-image*, yielding the *contextualized-multimodal* modules of DISARM[7,8]. This is followed by interpretability analysis. Finally, we discuss the limitations of DISARM by performing error analysis[9].

All Entities Seen During Training: Towards unimodal text-only baseline evaluation, the GRU-based system yields a relatively lower *harmful* recall 0.74 along-with an overall better F1 0.75, in comparison to XLNet's 0.82 and a lower F1 of 0.67, as shown in Table 2. The lower *harmful* precision 0.65 and *not-harmful* recall of 0.52 contribute to the lower F1 score for XLNet. Amongst image-only unimodal systems, VGG-based (image-only) system performs better with *not-harmful* recall 0.81, but is poor for detecting the harmful memes correctly with a lower *harmful* recall value of 0.68. On the other hand, VIT has a relatively better *harmful* class recall 0.74. Overall, the unimodal results (Table 2) indicate the efficacy of self-attention processing of the input modality as compared to that for convolution-based operation for images and RNN (GRU) sequence modeling for text.

Multimodally pre-trained models such as VisualBERT (MS COCO (Lin et al., 2014)) and ViLBERT (Conceptual Captions (Sharma et al., 2018)), yield moderate F1 scores of 0.70 and 0.68, and *harmful* recall values of 0.78 and 0.77, respectively (Table 2). Fresh training facilitates more meaningful results in favour of *not-harmful* precision (0.78 and 0.78 respectively) and *harmful* recall (0.84 and 0.82 respectively). Overall, ViLBERT yields the most balanced performance with 0.75 F1 score. It can be inferred from these results (Table 2) that multimodal pre-training could leverage domain relevance.

Multimodal low-rank bi-linear pooling is observed to distinctly enhance the performance by 4% and 6% F1 scores. The improvements can be attributed to the fusion of the CE and EH representations, respectively with CI, instead of a simple concatenation (Table 2). This is more prominent for CE with 0.78 F1, effectively implying the importance of the background context. Finally, DISARM is observed to yield a balanced performance with 0.78 F1 score, having a reasonable precision of 0.74 for non-harmful and the best

[6] http://mmf.sh/docs/notes/model_zoo

[7] We use abbreviations CE, CT, CI, CMM, EH, and MMLRBP for *contextualized* representations of entity, text, image, multimodal feature, embedded-harmfulness and multimodal low-rank bi-linear pooling, respectively.

[8] Ablation study is reported in Appendix C

[9] Error analysis is discussed in Appendix B

System	Modality	Approach	Test Set A Acc	Test Set A Prec	Test Set A Rec	Test Set A F1	Not-harmful P	Not-harmful R	Harmful P	Harmful R
Baselines	Unimodal	XLNet Text-only	0.6765	0.69	0.67	0.6663	0.73	0.52	0.65	0.82
		VGG Image-only	0.7451	0.75	0.745	0.7438	0.71	0.81	0.79	0.68
		GRU Text-only	0.7484	0.745	0.75	0.7473	0.73	0.76	0.76	0.74
		VIT Image only	0.7647	0.765	0.765	0.7642	0.74	0.79	0.79	0.74
	Multimodal	ViLBERT CC	0.6895	0.69	0.685	0.6835	0.71	0.6	0.67	0.77
		MM Transformer	0.6993	0.71	0.695	0.6926	0.75	0.57	0.67	0.82
		VisualBERT	0.7026	0.725	0.69	0.6918	**0.78**	0.54	0.67	0.84
		VisualBERT – COCO	0.7059	0.71	0.7	0.7014	0.73	0.62	0.69	0.78
		MMBT	0.7157	0.72	0.71	0.7121	0.74	0.64	0.7	0.78
		ViLBERT	0.7516	0.755	0.75	0.7495	**0.78**	0.68	0.73	0.82
Prop. system & variants		CE + CI (concat)	0.7353	0.74	0.735	0.7361	0.71	0.77	0.77	0.7
		CE + CI (MMLRBP)	**0.781**	**0.785**	0.78	0.7790	0.74	**0.84**	**0.83**	0.72
		EH + CI (concat)	0.6634	0.665	0.66	0.6609	0.67	0.6	0.66	0.72
		EH + CI (MMLRBP)	0.7255	0.73	0.725	0.7260	0.74	0.67	0.72	0.78
		DISARM	**0.781**	0.74	**0.835**	**0.7845**	0.74	0.81	0.74	**0.86**
$\Delta_{(\texttt{DISARM}--ViLBERT)\times 100}(\%)$			↑ 2.94%	↓ 1.5%	↑ 8%	↑ 3.5%	↓ 4%	↑ 13%	↑ 1%	↑ 4%

System	Modality	Approach	Test Set B Acc	Test Set B Prec	Test Set B Rec	Test Set B F1	Not-harmful P	Not-harmful R	Harmful P	Harmful R
Baselines	Unimodal	XLNet Text-only	0.5041	0.425	0.405	0.4060	0.72	0.59	0.13	0.22
		VGG Image-only	0.5455	0.42	0.405	0.4101	0.73	0.66	0.11	0.15
		GRU Text-only	0.5455	0.43	0.42	0.4210	0.73	0.65	0.13	0.19
		VIT Image only	0.5207	0.525	0.535	0.4843	0.8	0.51	0.25	0.56
	Multimodal	ViLBERT CC	0.438	0.535	0.53	0.4302	0.82	0.35	0.25	**0.71**
		MM Transformer	**0.7769**	0.53	0.575	0.5032	0.78	0.51	0.28	0.64
		VisualBERT	0.5537	0.545	0.565	0.5108	0.82	0.54	0.27	0.59
		VisualBERT – COCO	0.5785	0.53	0.545	0.5147	0.8	0.61	0.26	0.48
		MMBT	0.6116	0.54	0.55	0.5310	0.81	0.66	0.27	0.44
		ViLBERT	0.6612	0.58	0.595	0.5782	**0.83**	0.71	0.33	0.48
Prop. system & variants		CE + CI (concat)	0.4793	0.46	0.44	0.4230	0.74	0.51	0.18	0.37
		CE + CI (MMLRBP)	0.562	0.535	0.545	0.5079	0.81	0.57	0.26	0.52
		EH + CI (concat)	0.5868	0.505	0.51	0.4964	0.78	0.65	0.23	0.37
		EH + CI (MMLRBP)	0.6612	0.545	0.555	0.5470	0.8	0.74	0.29	0.37
		DISARM	0.74	**0.605**	**0.74**	**0.6498**	**0.83**	**0.79**	**0.38**	0.69
$\Delta_{(\texttt{DISARM}--ViLBERT)\times 100}(\%)$			↑ 7.88%	↑ 2.5%	↑ 14.5%	↑ 7.16%	–	↑ 8%	↑ 5%	↑ 21%

Table 2: Performance comparison of unimodal and multimodal baselines vs DISARM (and its variants) on Test Set A and B.

Sys		Approach	Acc	Prec	Rec	F1	Not-harmful P	Not-harmful R	Harmful P	Harmful R
Baselines	Unimodal	GRU Text-only	0.478	0.45	0.41	0.394	0.78	0.51	0.12	0.31
		VIT Image only	0.532	0.435	0.4	0.403	0.78	0.61	0.09	0.19
		XLNet Text-only	0.445	0.51	0.515	0.415	0.84	0.41	0.18	0.62
		VGG Image-only	0.532	0.45	0.42	0.414	0.79	0.59	0.11	0.25
	Multimodal	ViLBERT CC	0.358	0.53	0.49	0.350	0.87	0.26	0.19	**0.72**
		VisualBERT	0.478	0.535	0.56	0.442	0.87	0.43	0.2	0.69
		MM Transformer	0.510	0.505	0.505	0.448	0.83	0.51	0.18	0.5
		ViLBERT	0.608	0.525	0.54	0.505	0.84	0.64	0.21	0.44
		VisualBERT – COCO	**0.771**	0.525	0.515	0.511	0.83	**0.91**	0.22	0.12
		MMBT	0.587	0.55	0.575	0.514	0.87	0.59	0.23	0.56
Prop. system & variants		CE + CI (concat)	0.456	0.495	0.495	0.412	0.82	0.43	0.17	0.56
		CE + CI (MMLRBP)	0.532	0.55	0.595	0.485	**0.88**	0.5	0.22	0.69
		EH + CI (concat)	0.532	0.48	0.475	0.442	0.81	0.57	0.15	0.38
		EH + CI (MMLRBP)	0.619	0.5	0.495	0.483	0.83	0.68	0.17	0.31
		DISARM	0.739	**0.61**	**0.73**	**0.641**	0.86	0.76	**0.36**	0.7
$\Delta_{(\texttt{DISARM}--MMBT)\times 100}(\%)$			↑ 15.21%	↑ 6%	↑ 15.5%	↑ 12.66%	↓ 1%	↑ 17%	↑ 13%	14%

Table 3: Performance comparison of unimodal and multimodal baselines vs DISARM (and its variants) on Test Set C.

recall of 0.86 for the harmful categories, respectively.

All Entities Unseen as Harmful Targets During Training: With Test Set B, the evaluation is made slightly more challenging (Table 2) in terms of the entities to be assessed, as these were never seen as part of the training process as *harmful*. Unimodal systems mostly perform poorly in terms of both precision and recall for *harmful* class, with the exception of XLNet (Table 2) with *harmful* class recall as 0.56. For the multimodal baselines, the performance of the systems that are pre-trained using COCO (VisualBERT) and CC (ViLBERT) yields moderate recall of 0.64 and 0.71 for the *harmful* class in contrast to what we saw for Test Set A in Table 2. This could be due to additional commonsense reasoning facilitated by such systems, on a test set that is more open-ended compared to Test Set A. Their non-pre-trained versions along with MM Transformer and MMBT achieve better F1 scores, but with low *harmful* class recall.

Multimodal fusion using MMLRBP is observed (Table 2) to obtain an improved *harmful* class recall for CE (0.52) and lower values for EH (0.37) based fusion with CI, respectively. This reconfirms the utility of context. In comparison, DISARM yields a balanced F1 score of 0.6498 with the best precision values 0.83 and 0.38, along with decent recall values of 0.79 and 0.69 for *not-harmful* and *harmful* memes, respectively.

All Entities Unseen During Training: The results decline in this scenario (similarly to Test Set B), except for the *harmful* class recall score for XLNet (0.62), as shown in Table 3. In the current scenario (Test Set C), none of the entities being assessed during testing is seen during the training phase. For multimodal baselines, we see a similar trend for VisualBERT (COCO) and ViLBERT (CC), with the *harmful* class recall of 0.72 for ViLBERT (CC) being significantly better than 0.12 for VisualBERT (COCO). This again emphasizes the need for the affinity between the pre-training dataset and the downstream task at hand. In general, the precision for the *harmful* class is very low.

We observe (Table 3) significant increase in the *harmful* class recall for MMLRBP-based multimodal fusion of CI with CE (0.69%), as against a decrease in the same with EH (0.31%). In comparison to all other systems, DISARM yields a low, yet the best *harmful* precision value of 0.36 and a moderate recall value of 0.70, as can be observed in Table 3. Also, besides yielding reasonable precision and recall values of 0.86 and 0.76, respectively, for the *not-harmful* class, DISARM exhibits better average precision, recall, and F1 scores of 0.61, 0.73 and 0.64, respectively.

Generalizability of DISARM. The generalizability of DISARM follows from the characteristic modelling and context-based fusion. Although there is still scope for improvement in terms of the performance and generalizability beyond US Politics, DISARM demonstrates the potential for detecting harmful targeting for a diverse set of entities. Specifically, the three-way testing setup inherently captures the efficacy with which DISARM

(a) L-AT **(b)** MM-AT-CLIP **(c)** V-AT-DISARM **(d)** V-AT-ViLBERT

Target Candidate→`democratic party`

Context→Politics tears families apart during bruising political season, when many Americans drop friends and family members who have different political views.

Figure 4: Comparison of the attention-maps for DISARM [(a), (b) & (c)] and ViLBERT [(d)] using BertViz and Grad-CAM.

can detect *unseen* harmful targets. Prediction for entities *completely* unseen during training is observed to be better as compared to when they are not seen as just *harmful* targets (Table 2 and 3), which could be due to the induced bias and limited training data. This could be addressed by training with a balanced dataset at scale. Overall, we argue that DISARM reveals encouraging generalizability with its performance on unseen entities by performing best with 0.6498 and 0.6412 macro-F1 scores, as compared to ViLBERT's 0.5782 and MMBT's 0.5146, for Test Sets B and C, respectively.

Comparative Diagnosis. Despite marginally better *harmful* recall of ViLBERT (CC) for Test Sets B (Table 2) and C (Table 3), the overall balanced performance of DISARM appears to be reasonably justified based on the comparative interpretability analysis between the attention maps for both the systems. Fig. 4 shows attention maps for an example meme. It depicts a meme that is *correctly* predicted for *harmfully* targeting *democratic party* by DISARM and incorrectly by ViLBERT. As visualised in Fig. 4a, harmfully-inclined word *killing* effectively attends not only to *baby*, but also to *democrats* and *racist*. The relevance is depicted via different color schemes and intensities, respectively. Interestingly, *killing* also attends to *democratic party*, both as part of OCR-extracted text and the target-candidate, jointly encoded by BERT. Multimodal attention being leveraged by DISARM depicted (via CLIP encoder) in Fig. 4b, demonstrates the utility of contextualised attention over *male* figure depicted, who represents insinuation of *democratic party*. Also, DISARM has a relatively focused field of vision, shown in Fig. 4c as compared to a relatively scattered one for ViLBERT (Fig. 4d). This demonstrates a better multimodal modelling capacity of DISARM as compared to that of ViLBERT.

6 Conclusion and Future Work

In this work, we introduced a novel task of detecting victimized entities within harmful memes and highlighted the inherent challenges involved. Towards addressing this open-ended task, we extended Harm-P with target entities for each harmful meme. We then proposed a novel multimodal deep neural framework, called DISARM that employs an adaptation of multimodal low-rank bilinear pooling-based fusion strategy at different levels of feature abstraction. We showed that DISARM outperforms various uni/multi-modal baselines in three different scenarios by 4%, 7%, and 13% increments in the macro-F1 score, respectively. Also, DISARM achieved a relative error rate reduction of 9% over the best baseline. We further emphasized the utility of different components of DISARM through ablations studies. We also elaborated on the generalizability of DISARM and established its modelling efficacy over that of ViLBERT via. interpretability analysis. We finally analysed the shortcomings in DISARM that lead to incorrect harmful target predictions. Through this work, we made an attempt towards eliciting a few inherent challenges pertaining to the task at hand – augmenting relevant context, effectively fusing multiple modalities, and pre-training. This reinstates the required motivation and leaves scope for future investigations in this direction.

Ethics and Broader Impact

Reproducibility. We present detailed hyper-parameter configurations in Appendix A and Table 4. We commit to releasing the dataset and the source code upon the acceptance of this paper.

User Privacy. The meme content and the associated information doesn't include any personal information. Issues related to copyright are addressed as part of the dataset source.

Annotation. The annotation was conducted by experts working in NLP or linguists in India. We treated the annotators fairly and with respect. They were paid as per the standard local paying rate. Before beginning the annotation process, we requested every annotator to thoroughly go through the annotation guidelines. We further conducted several discussion sessions to make sure all annotators could understand well what harmful targeting is and how to differentiate it from not-harmful or benign references.

Biases. Any biases found in the dataset are unintentional, and we do not intend to cause harm to any group or individual. We acknowledge that detecting harmfulness can be subjective, and thus it is inevitable that there would be biases in our gold-labelled data or the label distribution. This is addressed by working on a dataset that is created using general keywords about US Politics, and also by following a well-defined schema, which sets explicit definitions during annotation.

Misuse Potential. We state that this dataset can be potentially used for ill-intended purposes, like biased targeting of individuals/communities/organizations, etc. that may or may not be related to demographics and other information within the text. Intervention with human moderation would be required to ensure that this does not occur.

Intended Use. We make use of the existing dataset in our work in line with the intended usage prescribed by its creators and solely for research purposes. This applies in its entirety to its further usage as well. We commit to releasing our dataset aiming to encourage research in studying harmful targeting in memes on the web. We distribute the dataset for research purposes only, without a license for commercial use. We believe that it represents a useful resource when used appropriately.

Environmental Impact. Finally, due to the requirement of GPUs/TPUs large-scale Transformers require a lot of computations, contributing to global warming (Strubell et al., 2019). However, in our case, we do not train such models from scratch; rather, we fine-tune them on relatively small datasets. Moreover, running on a CPU for inference, once the model has been fine-tuned, is perfectly feasible, and CPUs contribute much less to global warming.

References

Firoj Alam, Stefano Cresci, Tanmoy Chakraborty, Fabrizio Silvestri, Dimiter Dimitrov, Giovanni Da San Martino, Shaden Shaar, Hamed Firooz, and Preslav Nakov. 2021. A survey on multimodal disinformation detection. *CoRR*, abs/2103.12541.

Monther Aldwairi and Ali Alwahedi. 2018. Detecting fake news in social media networks. *Procedia Computer Science*, 141:215–222. The 9th International Conference on Emerging Ubiquitous Systems and Pervasive Networks (EUSPN-2018) / The 8th International Conference on Current and Future Trends of Information and Communication Technologies in Healthcare (ICTH-2018) / Affiliated Workshops.

Justin Cheng, Michael Bernstein, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. 2017. Anyone can become a troll: Causes of trolling behavior in online discussions. In *Proc. of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, CSCW '17, page 1217–1230, New York, NY, USA. Assoc. for Computing Machinery.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proc. of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar. Assoc. for Computational Linguistics.

Christine Cook, Juliette Schaafsma, and Marjolijn Antheunis. 2018. Under the bridge: An in-depth examination of online trolling in the gaming context. *New Media & Society*, 20(9):3323–3340. PMID: 30581367.

Abhishek Das, Japsimar Singh Wahi, and Siyao Li. 2020. Detecting hate speech in multi-modal memes.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proc. of the 2019 Conference of the North American Chapter of the Assoc. for Computational Linguistics: Human Language Technologies*, NAACL-HLT '19, pages 4171–4186, Minneapolis, Minnesota, USA.

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. 2021. An image is worth 16x16 words: Transformers for image recognition at scale.

Rafael Ferreira, Rafael Dueire Lins, Steven J. Simske, Fred Freitas, and Marcelo Riss. 2016. Assessing sentence similarity through lexical, syntactic and semantic analysis. *Computer Speech & Language*, 39:1–28.

Raul Gomez, Jaume Gibert, Lluis Gomez, and Dimosthenis Karatzas. 2019. Exploring hate speech detection in multimodal publications.

Eduardo Graells-Garrido, Ricardo Baeza-Yates, and Mounia Lalmas. 2020. Every colour you are: Stance prediction and turnaround in controversial issues. In *WebSci '20: 12th ACM Conference on Web Science, Southampton, UK, July 6-10, 2020*, pages 174–183. ACM.

Kimmo Karkkainen and Jungseock Joo. 2021. Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *Proc. of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, WACV '21, pages 1548–1558.

Douwe Kiela, Suvrat Bhooshan, Hamed Firooz, Ethan Perez, and Davide Testuggine. 2019. Supervised multimodal bitransformers for classifying images and text. In *Proc. of the NeurIPS Workshop on Visually Grounded Interaction and Language*, ViGIL '19, Vancouver, Canada.

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 *Proc. of the 34th International Conference on Neural Information Processing Systems*, volume 33 of *NeurIPS '20*.

Jin-Hwa Kim, Kyoung-Woon On, Woosang Lim, Jeonghee Kim, Jung-Woo Ha, and Byoung-Tak Zhang. 2017. Hadamard product for low-rank bilinear pooling.

Seunghyun Kim, Afsaneh Razi, Gianluca Stringhini, Pamela J. Wisniewski, and Munmun De Choudhury. 2021. A human-centered systematic literature review of cyberbullying detection algorithms. *Proc. ACM Hum. Comput. Interact.*, 5(CSCW2):1–34.

Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proc. of the 3rd International Conference on Learning Representations*, ICLR '15, San Diego, California, USA.

Robin Kowalski, Gary Giumetti, Amber Schroeder, and Micah Lattanner. 2014. Bullying in the digital age: A critical review and meta-analysis of cyberbullying research among youth. *Psychological bulletin*, 140.

Sumeet Kumar and Kathleen Carley. 2019. Tree LSTMs with convolution units to predict stance and rumor veracity in social media conversations. In *Proc. of the 57th Annual Meeting of the Assoc. for Computational Linguistics*, pages 5047–5058, Florence, Italy. Assoc. for Computational Linguistics.

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. *arXiv:1908.03557*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. 2015. Microsoft coco: Common objects in context.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft COCO: Common objects in context. In *Proc. of the European Conference on Computer Vision*, ECCV '14, pages 740–755, Zurich, Switzerland.

Phillip Lippe, Nithin Holla, Shantanu Chandra, Santhosh Rajamanickam, Georgios Antoniou, Ekaterina Shutova, and Helen Yannakoudakis. 2020. A multimodal framework for the detection of hateful memes. *arXiv:2012.12871*.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Proc. of the Conference on Neural Information Processing Systems*, NeurIPS '19, pages 13–23, Vancouver, Canada.

Sean MacAvaney, Hao-Ren Yao, Eugene Yang, Katina Russell, Nazli Goharian, and Ophir Frieder. 2019a. Hate speech detection: Challenges and solutions. *PLOS ONE*, 14(8):1–16.

Sean MacAvaney, Hao-Ren Yao, Eugene Yang, Katina Russell, Nazli Goharian, and Ophir Frieder. 2019b. Hate speech detection: Challenges and solutions. *PLOS ONE*, 14(8):1–16.

Alexandros Mittos, Savvas Zannettou, Jeremy Blackburn, and Emiliano De Cristofaro. 2020. "and We Will Fight for Our Race!" A Measurement Study of Genetic Testing Conversations on Reddit and 4chan. In *Proc. of the Fourteenth International AAAI Conference on Web and Social Media*, ICWSM '20, pages 452–463, Atlanta, Georgia, USA.

Hamed Pirsiavash, Deva Ramanan, and Charless Fowlkes. 2009. Bilinear classifiers for visual recognition. In *Advances in Neural Information Processing Systems*, volume 22. Curran Associates, Inc.

Shraman Pramanick, Dimitar Dimitrov, Rituparna Mukherjee, Shivam Sharma, Md. Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. 2021a. Detecting harmful memes and their targets. In *Findings of the Assoc. for Computational Linguistics*, ACL-IJCNLP '21, pages 2783–2796.

Shraman Pramanick, Shivam Sharma, Dimitar Dimitrov, Md. Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. 2021b. MOMENTA: A multimodal framework for detecting harmful memes and their targets. In *Findings of the Assoc. for Computational Linguistics: EMNLP 2021*, pages 4439–4455, Punta Cana, Dominican Republic. Assoc. for Computational Linguistics.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *Proc. of the 38th International Conference on Machine Learning*, ICML '21, pages 8748–8763.

Kunal Relia, Zhengyi Li, Stephanie H. Cook, and Rumi Chunara. 2019. Race, ethnicity and national origin-based discrimination in social media and hate crimes across 100 u.s. cities. *Proc. of the International AAAI Conference on Web and Social Media*, 13(01):417–427.

Bárbara Gomes Ribeiro, Manoel Horta Ribeiro, Virgílio A. F. Almeida, and Wagner Meira Jr. 2021. Follow the money: Analyzing @slpng_giants_pt's strategy to combat misinformation. *CoRR*, abs/2105.07523.

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

Vlad Sandulescu. 2020. Detecting hateful memes using a multimodal deep ensemble. *arXiv:2012.13235*.

Morgan Klaus Scheuerman, Jialun Aaron Jiang, Casey Fiesler, and Jed R. Brubaker. 2021. A framework of severity for harmful content online. *Proc. ACM Hum.-Comput. Interact.*, 5(CSCW2).

Lanyu Shang, Christina Youn, Yuheng Zha, Yang Zhang, and Dong Wang. 2021a. Knowmeme: A knowledge-enriched graph neural network solution to offensive meme detection. In *2021 IEEE 17th International Conference on eScience (eScience)*, pages 186–195.

Lanyu Shang, Yang Zhang, Yuheng Zha, Yingxi Chen, Christina Youn, and Dong Wang. 2021b. Aomd: An analogy-aware approach to offensive meme detection on social media.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proc. of the 56th Annual Meeting of the Assoc. for Computational Linguistics*, ACL '18, pages 2556–2565, Melbourne, Australia.

Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake news detection on social media: A data mining perspective. *SIGKDD Explor. Newsl.*, 19(1):22–36.

Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in NLP. In *Proc. of the 57th Annual Meeting of the Assoc. for Computational Linguistics*, ACL '19, pages 3645–3650, Florence, Italy.

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

Yuping Wang, Fatemeh Tahmasbi, Jeremy Blackburn, Barry Bradlyn, Emiliano De Cristofaro, David Magerman, Savvas Zannettou, and Gianluca Stringhini. 2020. Understanding the Use of Fauxtography on Social Media. *arxiv:2009.11792*.

Liang Wu and Huan Liu. 2018. Tracing fake-news footprints: Characterizing social media messages by how they propagate. In *Proc. of the Eleventh ACM International Conference on Web Search and Data Mining*, WSDM '18, page 637–645, New York, NY, USA. Assoc. for Computing Machinery.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2020. Xlnet: Generalized autoregressive pretraining for language understanding.

Ziqi Zhang and Lei Luo. 2018. Hate speech detection: A solved problem? the challenging case of long tail on twitter.

Xiayu Zhong. 2020. Classification of Multimodal Hate Speech – The Winning Solution of Hateful Memes Challenge. *arXiv e-prints*, page arXiv:2012.01002.

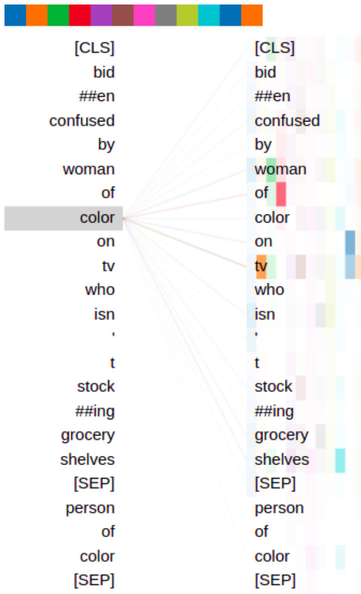
Kaimin Zhou, Chang Shu, Binyang Li, and Jey Han Lau. 2019. Early rumour detection. In *Proc. of the 2019 Conf. of the North American Chapter of the Assoc. for Comp. Ling.: Human Lang. Tech., Volume 1 (Long and Short Papers)*, pages 1614–1623, Minneapolis, Minnesota. Assoc. for Comp. Ling.

Yi Zhou and Zhenhao Chen. 2020. Multimodal Learning for Hateful Memes Detection. *arxiv:2011.12870*.

Appendix

		BS	#Epochs	LR	V-Enc	T-Enc	#Param
UM	GRU	32	25	0.0001	-	`bert`	2M
	XLNet	16	20	0.0001	-	`xlnet`	116M
	VGG16	32	25	0.0001	VGG16	-	117M
	ViT	16	20	0.0001	`vit`	-	86M
MM	MMFT	16	20	0.001	ResNet-152	`bert`	170M
	MMBT	16	20	0.001	ResNet-152	`bert`	169M
	ViLBERT*	16	10	0.001	Faster RCNN	`bert`	112M
	V-BERT*	16	10	0.001	Faster RCNN	`bert`	247M
	DISARM	16	30	0.0001	`vit`	`bert`	111M

Table 4: Hyperparameters summary. [BS→Batch Size; LR→Learning Rate; V/T-Enc→Vision/Text-Encoder; `vit`→`vit-base-patch16-224-in21k`; `bert:`→`bert-base-uncased`; `xlnet`→`xlnet-base-uncased`].



(a) L-AT

(b) MM-AT-CLIP

(c) V-AT-DISARM

(d) V-AT-ViLBERT

Target Candidate→`person of color`

Context→`During the evening of the VP debates, Joe Biden settled down on his soft couch with a glass of warm milk to watch this.`

Figure 5: Comparison of attention-maps for a miclassification, between DISARM [(a), (b) & (c)] and ViLBERT [(d)] using BertViz and Grad-CAM.

A Implementation Details and Hyperparameter Values

We train all the models using PyTorch on an actively dedicated NVIDIA Tesla V100 GPU, with 32 GB dedicated memory, CUDA-11.2 and cuDNN-8.1.1 installed. For the unimodal models, we import all the pre-trained weights from the TORCHVISION.MODELS[10], a sub-package of the PyTorch framework. We initialize the remaining weights randomly using a zero-mean Gaussian distribution with a standard deviation of 0.02. We train DISARM in a setup considering only *harmful* class data from Harm-P (Pramanick et al., 2021b). We extend it by manually annotating for *harmful* targets, followed by including *not-harmful* samples using automated entity extraction (textual and visual) strategies for train/val splits and manual annotation (for both harmful and not-harmful) for the test split.

We train all models we experiment with, using the Adam optimizer (Kingma and Ba, 2015) with a learning rate of $1e^{-4}$, with a weight decay of $1e^{-5}$ and a Binary Cross-Entropy (BCE) loss as the objective function. We extensively fine-tuned our experimental setups, based upon different architectural requirements to finalise on afore-mentioned hyper-parameters. We also use early-stopping for saving the best intermediate checkpoints as well. Table 4 gives more detail about the hyper-parameters we used for training. On average, it took approx. 2:30 hours to train a multi-modal neural model.

B Error Analysis

It is evident from the results shown in Table 2 and 3, that DISARM still has short-comings. Examples like the one shown in Fig. 5 are seemingly *harmless*, both textually and visually, but imply serious *harm* to a *person of color* in an implicit way. Such complexity can be challenging to model, without providing additional context like *people of colour face racial discrimination all over the world*. This is also analogous to a fundamental challenge associated with detecting implicit hate (MacAvaney et al., 2019b). Despite modelling contextual information explicitly in DISARM, it misclassifies this meme. Although the context obtained for this meme pertains to its content (Fig. 5), it does not relate to *global racial prejudice*, which is key to ascertaining it as a harmfully targeting meme. Moreover, besides context, visuals and the mes-

[10]http://pytorch.org/docs/stable/torchvision/models.html

Approach	Test Set A					Test Set B					Test Set C				
	F1	Not-harmful		Harmful		F1	Not-harmful		Harmful		F1	Not-harmful		Harmful	
		P	R	P	R		P	R	P	R		P	R	P	R
CE	0.7411	0.71	0.78	0.77	0.71	0.4847	0.78	**0.95**	0.29	0.07	0.4829	0.83	**0.93**	0.17	0.06
EH	0.7250	0.75	0.66	0.71	0.79	0.5544	0.81	0.72	0.3	0.41	0.5658	0.88	0.68	0.27	0.56
CI	0.7729	0.74	0.82	0.81	0.73	0.5174	0.79	0.89	0.29	0.15	0.5314	0.84	0.87	0.23	0.19
CE + EH	0.7406	0.71	0.78	0.78	0.7	0.5775	0.82	0.74	0.33	0.44	0.5840	**0.89**	0.7	0.29	0.57
CE + CI (concat)	0.7361	0.71	0.77	0.77	0.7	0.4230	0.74	0.51	0.18	0.37	0.4125	0.82	0.43	0.17	0.56
CE + CI (MMLRBP)	0.7790	0.74	**0.84**	**0.83**	0.72	0.5079	0.81	0.57	0.26	0.52	0.4857	0.88	0.5	0.22	0.69
EH + CI (concat)	0.6609	0.67	0.6	0.66	0.72	0.4964	0.78	0.65	0.23	0.37	0.4421	0.81	0.57	0.15	0.38
EH + CI (MMLRBP)	0.7260	0.74	0.67	0.72	0.78	0.5470	0.8	0.74	0.29	0.37	0.4836	0.83	0.68	0.17	0.31
DISARM	**0.7845**	0.74	0.81	0.74	**0.86**	**0.6498**	**0.83**	0.79	**0.38**	**0.69**	**0.6412**	0.86	0.76	**0.36**	**0.7**

Table 5: Ablation results for DISARM and its variants for Test Sets A, B and C.

sage embedded within the meme do not convey definite harm when considered in isolation. This error can be inferred clearly from the embedded-harmfulness, contextualised-visuals, and the visuals being attended by DISARM as depicted in Fig. 5a, 5b and 5c respectively. On the other hand, as shown in the visual attention plot for ViLBERT in Fig. 5d, the field of view being attended to encompasses the visuals of *Kamala Harris*, who is the *person of colour* being primarily targeted through the meme. Besides the distinct attention on the primary target-candidate within the meme, ViLBERT could have leveraged the pre-training it received from Conceptual Captions (CC) (Sharma et al., 2018), a dataset known for its diverse coverage of complex textual descriptions. This essentially highlights the importance of multimodal pre-training using the dataset that is not as generic as MS COCO (Lin et al., 2015), but facilitate modelling of the complex real-world multimodal information, especially for tasks related to memes.

C Ablation Study

In this section, we present some ablation studies for CE, EH, CT and CI based sub-modules of DISARM, examined in isolation and combinations, and finally for DISARM using CMM representation.

Test Set A: As observed in the comparisons made with the other baseline systems for the Test Set A in Table 2, the overall range of the F1 scores is relatively higher with the least value observed to be 0.66 for XLNet (text-only) model, the results for unimodal systems is satisfactory with values of 0.74, 0.73, and 0.77 for CE EH, and CI based unimodal systems, respectively. For multimodal systems, we can observe distinct lead for the MMLRBP-based fusion strategy, for both CE and EH based systems over the concatenation-based approach, except for EH's recall drop by 7%. Finally DISARM yields the best overall F1 score of 0.7845.

Test Set B: With *context* not having any harmfulness cues for a given meme, the unimodal CE based module performs the he worst with 0.48 F1 and 0.07 *harmful* recall, in the open-ended setting of Test Set B. In contrast, EH yields an impressive F1 score of 0.55 and a *harmful* recall of 0.41. This relative gain of 7% in the F1 score could be due to the presence of explicit harmfulness cues. The complementary effect of considering contextual information can be inferred from the joint modeling of CE and EH, to obtained CT, that enhances the F1 and *harmful* recall by 2% and 3%, respectively (Table 5). Unimodal assessment of CI performs moderately with 0.51 F1 score, but with a poor *harmful* recall of 0.15. MMLRBP, towards joint-modeling of CE and CI yields a significant boost of the *harmful* recall value 0.52 (Table 5). On the other hand, MMLRBPbased fusion of EH and CI yields 0.54 F1 score, which is 1% below that for the unimodal EH system. This emphasizes the importance of accurately modeling the embedded harmfulness, besides *augmenting* with additional context. The complementary effects of CE, EH, and CI are observed for DISARM with a balanced F1 score of 0.65 and a competitive *harmful* recall value of 0.69.

Test Set C: As observed in the previous scenario (Test Set B), the unimodal models for CE yield a low F1 score of 0.48 and the worst *harmful* recall value of 0.06. Much better performance is observed for unimodal setups involving the EH and its joint modelling with CE with an improved F1 score of 0.56 and 0.58, along with the *harmful* recall score of 0.56 and 0.57, respectively. CI based unimodal evaluation again yields a moderate F1 score of 0.53 (Table 5), along with a poor *harmful* recall of 0.19, which shows its insufficiency for modelling harmful targeting on its own. For

(a) Harmful analogy **(b)** Sensitive visuals **(c)** Political grounds **(d)** Religious grounds **(e)** National threat

Figure 6: Examples of memes depicting different types ((a)–(e)) of *harmful* targeting.

multimodal setups, the joint modelling of CE and CI benefits from MMLRBP based fusion, yielding a gain of 7% and 13% in F1 and *harmful* recall, respectively. This confirms the importance of contextual multimodal semantic alignment. Correspondingly, joint multimodal modelling of EH and CI regresses the unimodal affinity within the EH. Finally, DISARM outperforms all other systems in this category with the best F1 score of 0.64, with a decent *harmful* recall score of 0.7.

The results reported in this work are for the comparison and analysis of the most optimal set of design and baseline choices. We have performed extensive experiments as part of preliminary investigations, with different contextual modelling strategies, attention mechanisms, modelling choices, etc., to reach a conclusive architectural configuration, that indicates promise towards addressing the task of target detection from harmful memes to a certain extent.

D Annotation Guidelines

Before discussing details about the annotation process, revisiting the definition of *harmful* memes would set the pretext towards consideration of *harmful* targeting and *not-harmful* referencing. According to Pramanick et al. (2021b), abuse, offence, disrespect, insult or insinuation of a targeted entity or any socio-cultural or political ideology, belief, principle, or doctrine associated with that entity amounts to the expression of harm.

Another common understanding[11,12,13] about the harmful content is that it could be anything online which causes a person distress. It is an extremely subjective phenomenon, wherein what maybe be harmful to some, might not be considered an issue by someone else. This makes it significantly challenging to characterize and hence study it via the computational lens.

Based on a survey of 52 participants, Scheuerman et al. (2021) defines online harm to be any violating content that results in any (or a combination) of four categories: (i) physical harm, (ii) emotional harm, (iii) relational harm and (iv) financial harm.

With this pretext, we define below 2 types of referencing that we have investigated in our work, within the context of internet memes: (i) *harmful* (ii) *not-harmful*

D.1 Reference types

Harmful. The understanding about harmful referencing (*targeting*) in memes, can be sourced back to the definition of harmful memes by Pramanick et al. (2021b), wherein a social entity is subjected to some form of ill-treatment like mental abuse, psycho-physiological injury, proprietary damage, emotional disturbance, or public image damage, based on their background (bias, social background, educational background, etc.) by a meme author.

Not-harmful. Not-harmful referencing in memes is any benign mention (or depiction) of a social entity via humour, limerick, harmless pun or any content that does not cause distress. Any reference that is *not* harmful, comes under this category.

D.2 Characteristics of harmful targeting

There are several factors that collectively facilitate characterisation of *harmful* targeting in memes. Few are enlisted below:

1. A prominent way of harmfully targeting an entity in a meme is by leveraging sarcastically

[11] https://reportharmfulcontent.com/advice/other/further-advice/harmful-content-online-an-explainer

[12] https://swgfl.org.uk/services/report-harmful-content

[13] https://saferinternet.org.uk/report-harmful-content

Harmful meme			Not-harmful meme		
Individual	**Organization**	**Community**	**Individual**	**Organization**	**Community**
joe biden (333)	democratic party (184)	mexicans (11)	donald trump (106)	green party (189)	trump supporters (86)
donald trump (285)	republican party (130)	black (7)	republican voter (102)	biden camp (162)	white (50)
barack obama (142)	libertarian party (44)	muslim (7)	barack obama (94)	communist party (114)	african american (47)
hillary clinton (35)	cnn (6)	islam (6)	joe biden (47)	america (64)	democrat officials (45)
mike pence (13)	government (5)	russian (5)	alexandria ocasio cortez (44)	trump administration (52)	republican (44)

Table 6: The top-5 most frequently referenced entities in each harmfulness class and target categories. The total count for each word is shown in parentheses.

harmful analogies, framed via either textual or visual instruments (Fig. 6a).

2. There could be multiple entities being harmfully targeted within a meme as depicted in Fig. 7. Hence, annotators were asked to provide all targets as harmful, without exception.
3. Harmful targeting within a meme could have visual depictions, that are either gory, violent, graphically sensitive or pornographic (Fig. 6b).
4. Any meme that insinuates an entity on either social, political, professional, religious grounds, can cause harm (Fig. 6c and 6d).
5. Any meme that implies an explicit/implicit threat to an individual, community, national or international entity is harmful (Fig. 6d and 6e).
6. Whenever there is any ambiguity regarding the harmfulness of any reference being made, we request authors to proceed with the best of their understanding.

D.3 Annotation process

Annotators were requested to follow 4 standard steps towards annotating each meme as enlisted below, to ensure consistency in the approach adopted. We consider an example, depicted in Fig. 7 to demonstrate the steps taken while annotating. Annotators were requested to:

1. Understand a meme and its background context clearly. The argument being made in the example meme, depicted in Fig. 7a is reasonably self-explanatory, due to its descriptiveness.
2. Enlist all the valid entities that are referenced within a given meme. For the sample meme (Fig. 7), valid entities are bill clinton, hillary clinton, white house, donald trump and democrat.
3. Assign the suitable entities from the list, the label harmful, annotating a positive case for *harmful* targeting. *bill clinton, hillary clinton and democrat* are being framed in the meme argument, for exhibiting hypocrisy over the appointment of close relatives for a high profile role.
4. Finally, assign *harmless* references under not-harmful category. *donald trump and white house* would be annotated as a harmless reference, as they aren't the subject of implied insinuation.

(a) A meme referencing harmful & not-harmful entities.

Candidates→`bill clinton, hillary clinton, white house, donald trump, democrat`

Harmful→`bill clinton, hillary clinton, democrat`

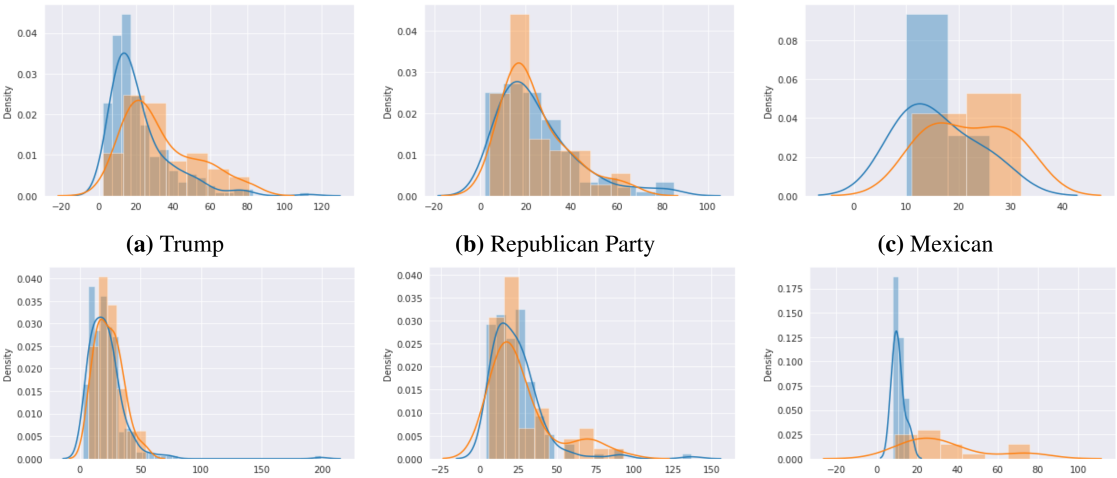
Not-harmful→`white house, donald trump`

Figure 7: A sample meme, along with the *candidate* entities, *harmful* targets and *not-harmful* references.

E Ext-Harm-P Characteristics

E.1 Lexical Analysis

Interestingly, a significant number of memes are disseminated making references to popular *individuals* like *Joe Biden, Donald Trump, etc.*, as can



(a) Trump **(b)** Republican Party **(c)** Mexican

(d) Biden **(e)** Democratic Party **(f)** Black

Figure 8: Distributions of the OCR's length for the memes of top-5 harmful references. Harmful (Blue) and Not-harmful (Orange). The depiction is for Individual: (a) and (d), Organization: (b) and (e) and Community: (c) and (f).

be observed for individual sub-category (for both harmful and not-harmful categories), in Table 6. It can be noticed for *harmful–organization* in Table 6, top-5 harmfully targeted organizations include top-2 leading political organizations (*democratic and republican party*), which are of significant political relevance, followed by *libertarian party*, a media house (*CNN*) and finally *government*. Whereas, non-harmfully referenced organizations includes *biden camp* and *trump administration*, that are mostly leveraged for harmfully targeting (or otherwise) the associated public figure. Finally, communities like *mexicans, black, muslim, islam and russian* are often immensely prejudiced online. Whereas, non-harmfully targeted communities like *trump supporters and african american* are not targeted as often as the aforementioned ones Table 6.

This largely emphasizes the inherent bias that multimodal content like memes implies, which has a direct influence on the efficacy of machine/deep learning-based systems. The reasons for this bias are mostly linked to societal behaviour at the organic level, and the limitations posed by current techniques to process such data. Distinct mutual exclusion for harmful vs. not-harmful categories for community shows the inherent bias that could pose a challenge, even for the best multimodal deep neural systems. The high pervasiveness of a few prominent keywords could effectively lead to increasing bias towards them for specific eventualities. Whereas, the significant overlap observed in Table 6 for the enlisted entities, between harmful and not-harmful individuals, highlight the need for sophisticated multimodal systems that can effectively reason towards making a complex decision like detecting harmful targeting within memes.

E.2 Meme-message Length Analysis

Most of the *harmful* memes are observed to be created using texts of length $16 - 18$ (Fig. 8). Whereas, *not-harmful* meme-text lengths are have a relatively higher std.-dev., possibly due to diversity of *not-harmful* messages. Trump and Republic party have meme-text length distributions similar for *not-harmful* category; skewing left, but gradually decreasing towards the right. This suggests varying content generation pattern amongst meme creators (Fig. 8). Meme-text length distribution for Biden closely approximates a normal distribution with the low std.-dev. Both the categories would pre-dominantly entail creating memes with shorter text lengths, due to the popularity of *Biden* amongst humorous content creators. A similar trend could be seen for the democratic party as well, where most of the samples are observed to be falling within $50 - 75$ meme-text length range. The overall harmful and not-harmful meme-text length distribution is observed to be fairly distributed across different meme-text lengths for *mexican*. Whereas, the amount of harm intended towards *black* community is observed to be significantly more, as compared to moderately distributed *not-harmful* memes depicted by the corresponding meme-text length distribution in Fig. 8.