DISARM: Detecting the Victims Targeted by Harmful Memes

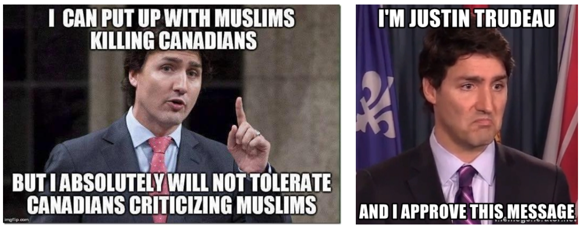
Shivam Sharma[1,3], **Md. Shad Akhtar**[1], **Preslav Nakov**[2], **Tanmoy Chakraborty**[1]
[1]Indraprastha Institute of Information Technology - Delhi, India
[2]Qatar Computing Research Institute, HBKU, Doha, Qatar
[3]Wipro AI Labs, India
{shivams, shad.akhtar, tanmoy}@iiitd.ac.in
pnakov@hbku.edu.qa

Abstract

Internet memes have emerged as an increasingly popular means of communication on the Web. Although 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, identifying whom they attack remains a challenging and under-explored area. Here we aim to bridge this gap. In particular, we create a dataset where 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 test setups, corresponding to entities that are (a) all seen while training, (b) not seen as a harmful target on training, and (c) not seen at all on training. The evaluation results show that DISARM significantly outperforms ten unimodal and multimodal systems. Finally, we show that DISARM is interpretable and comparatively more generalizable and that it can reduce the relative error rate for harmful target identification by up to 9 points absolute over several strong multimodal rivals.

1 Introduction

Social media 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 have also been misused to incite extreme reactions, hatred, and to spread disinformation on a massive scale.


I CAN PUT UP WITH MUSLIMS KILLING CANADIANS
BUT I ABSOLUTELY WILL NOT TOLERATE CANADIANS CRITICIZING MUSLIMS
I'M JUSTIN TRUDEAU
AND I APPROVE THIS MESSAGE

(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.

Numerous recent efforts have attempted to characterize harmfulness (Pramanick et al., 2021b), hate speech (Kiela et al., 2020), and offensiveness (Suryawanshi et al., 2020) within memes. Most of these efforts have been directed towards detecting 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 is the task we are addressing here: detecting the entities that a meme targets in a harmful way.

Harmful targeting in memes is often done using satire, sarcasm, or humour in an explicit or an implicit way, aiming at attacking an individual, an organization, a community, or society in general. For example, Fig. 1a depicts Justin Trudeau, the Prime Minister of Canada, as *communally biased against* Canadians, while favoring alleged *killings by* Muslims, whereas Fig. 1b shows an arguably benign meme of the same person 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. Note that in many cases interpreting memes and their harmful intent requires some additional background knowledge for the meme to be understood properly.

Hence, an automated system for detecting the entities targeted by harmful memes faces two major challenges: (*i*) insufficient *background context*, (*ii*) complexity posed by the *implicit* harm, and (*iii*) keyword *bias* in a supervised setting.

To address these challenges, here we aim to address the task of harmful target detection in memes by formulating it as an open-ended task, where a meme can target an entity not seen on training. An end-to-end solution requires (*i*) identifying the entities referred to in the meme, and (*ii*) deciding whether each of these entities is being targeted in a harmful way. To address these two tasks, we perform systematic contextualization of the multimodal information presented within the meme by first performing intra-modal fusion between an external knowledge-based *contextualized-entity* and the *textually-embedded harmfulness* in the meme, which is followed by cross-modal fusion of the contextualized textual and visual modalities using low-rank bi-linear pooling, resulting in an enriched multimodal representation. We evaluate our model using three-level stress-testing to better assess its generalizability to unseen targets.

We create a dataset, and we propose an experimental setup and a model to address the aforementioned requirements, making the following contributions:[1]:

1. We introduce the novel task of detecting the entities targeted by harmful memes.
2. We create a new dataset for this new task, Ext-Harm-P, by extending Harm-P (Pramanick et al., 2021b) via re-annotating each harmful meme with the entity it targets.
3. We propose DISARM, a novel multimodal neural architecture that uses an expressive contextualized representation for detecting harmful targeting in memes.
4. We empirically showcase that DISARM outperforms ten unimodal and multimodal models by several points absolute in terms of macro-F1 scores in three different evaluation setups.
5. Finally, we discuss DISARM's generalizability and interpretability.

2 Related Work

Misconduct on Social Media. The rise in misconduct on social media is a prominent research topic. Some forms of online misconduct include rumours (Zhou et al., 2019), fake news (Aldwairi and Alwahedi, 2018; Shu et al., 2017; Nguyen et al., 2020), misinformation (Ribeiro et al., 2021; Shaar et al., 2022), disinformation (Alam et al., 2021; Hardalov et al., 2022), hate speech (MacAvaney et al., 2019; Zhang and Luo, 2019; Zampieri et al., 2020), 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, 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 representations for classifying a message as genuine vs. fake. Cheng et al. (2017) studied user's mood along with the online contextual discourse and demonstrated that it helps for trolling behaviour prediction on top of 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 additional 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 posed by the inherent biases within the existing multimodal systems. Pramanick et al. (2021b) released another dataset focusing on 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 hateful vs. non-hateful. Different approaches such as feature augmentation, attention mechanism, and multimodal loss re-weighting were attempted (Das et al., 2020; Sandulescu, 2020; Zhou et al., 2021; Lippe et al., 2020) as part of this task. Oriol et al. (2019) studied hateful memes by highlighting the importance of visual cues such as structural template, graphic modality, causal depiction, etc.

[1] The source code and the dataset can be found here https://github.com/LCS2-IIITD/DISARM.

Split	# Examples	Category-wise # Samples.	
		Harmful	Not-harmful
Train	3,618	1,206	2,412
Validation	216	72	144
Test	612	316	296
Total	4,446	1,594	2,852

Table 1: Summary of Ext-Harm-P, with overall and category-wise # of samples.

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 research includes using implicit models, e.g., topic modelling and multimodal cues, for detecting offensive analogy (Shang et al., 2021b) and hateful discrimination (Mittos et al., 2020) in memes. Wang et al. (2021) argued that online attention can be garnered immensely via fauxtography, which could eventually evolve towards turning into memes that potentially go viral. To support research on these topics, several datasets for offensiveness, hate speech, and harmfulness detection have been created (Suryawanshi et al., 2020; Kiela et al., 2020; Pramanick et al., 2021a,b; Gomez et al., 2020; Dimitrov et al., 2021; Sharma et al., 2022).

Most of the above studies attempted to address classification tasks in a constrained setting. However, to the best of our knowledge, none of them targeted the task of detecting the specific entities that are being targeted. Here, we aim to bridge this gap with focus on detecting the specific entities targeted by a given harmful meme.

3 Dataset

The Harm-P dataset (Pramanick et al., 2021b) consists of 3,552 memes about US politics. Each meme is annotated with its harmful label and the social entity that it targets. The targeted entities are coarsely classified into four social groups: individual, organization, community, and the general public. While these coarse classes provide an abstract view of the targets, identifying the *specific* targeted person, organization, or community in a fine-grained fashion is also crucial, and this is our focus here. All the memes in this dataset broadly pertain to *US Politics domain*, and they target well-known personalities or organizations. To this end, we manually re-annotated the memes in this dataset with the specific people, organizations, and communities that they target.

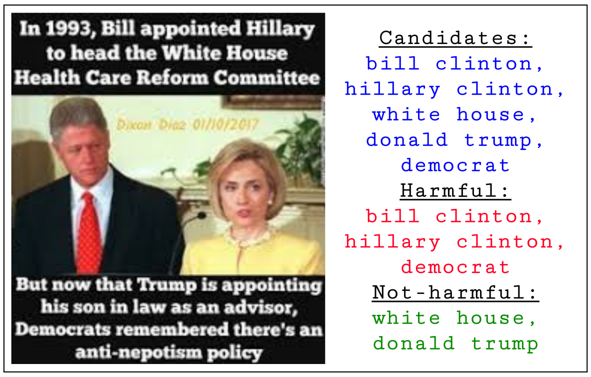
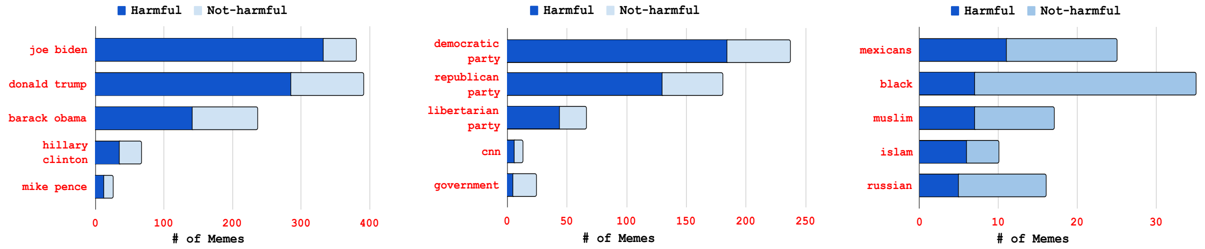

In 1993, Bill appointed Hillary to head the White House Health Care Reform Committee
Dixon Diaz 01/10/2017
But now that Trump is appointing his son in law as an advisor, Democrats remembered there's an anti-nepotism policy
Candidates:
bill clinton, hillary clinton, white house, donald trump, democrat
Harmful:
bill clinton, hillary clinton, democrat
Not-harmful:
white house, donald trump

Figure 2: Example meme, along with the candidate entities, harmful targets, and non-harmful references.

Extending Harm-P (Ext-Harm-P). Towards generalizability, we extend Harm-P by redesigning the existing data splits as shown in Table 1. We call the resulting dataset Ext-Harm-P. It contains a total of 4,446 examples including 1,594 harmful and 2,852 non-harmful; both categories have references to a number of entities. For training, we use the *harmful* memes provided as part of the original dataset (Pramanick et al., 2021b), which we re-annotate for the fine-grained entities that are being targeted harmfully as positive samples (*harmful* targets). This is matched with twice as many negative samples (*not-harmful* targets). For negative targets, we use the top-2 entities from the original entity lexicon, which are not labeled for harmfulness and have the highest lexical similarity with the meme text (Ferreira et al., 2016). This at least ensures lexical similarity with the entities referenced within a meme, thereby facilitating a confounding effect (Kiela et al., 2020) as well. For the test set, *all* the entities are first extracted automatically using named entity recognition (NER) and person identification (PID)[2]. This is followed by manual annotation of the test set.

Dataset Annotation Process Since assessing the harmfulness of memes is a highly subjective task, our annotators were requested to follow four key steps when annotating each meme, aiming to ensure label consistency. The example in Fig. 2 demonstrates the steps taken while annotating: we first identify the candidate entities, and then we decide whether a given entity is targeted in a harmful way. We asked our annotators to do the following (additional details about the annotation process are given in Appendix D):

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



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

Figure 3: Comparison plots for the top-5 *harmfully referenced* entities, for their harmful/non-harmful referencing in our dataset.

1. Understand the meme and its background context.

2. List all the valid *candidate* entities that are referenced in the meme. For the example on Fig. 2, the valid entities are *Bill Clinton, Hillary Clinton, White House, Donald Trump*, and *Democrat*.

3. Assign the relevant entities as *harmful*. For the example on Fig. 2, *Bill Clinton, Hillary Clinton*, and *Democrat* are targeted in the meme for influencing the appointment of their kin on government positions.

4. Finally, assign *harmless* references to entities under the non-harmful category. In the example on Fig. 2, *Donald Trump* and *White House* would be annotated as *non-harmful*.

We had three annotators and a consolidator. The inter-annotator agreement before consolidation had a Fleiss Kappa of 0.48 (moderate agreement), and after consolidation it increased to 0.64 (substantial agreement).

Analyzing Harmful Targeting in Memes. The memes in Ext-Harm-P are about *US Politics*, and thus they prominently feature entities such as *Joe Biden* and *Donald Trump*, both harmfully and harmlessly. The ratio between these types of referencing varies across *individuals*, *organizations*, and *communities*. We can see in Fig. 3 that the top-5 harmfully referenced *individuals* and *organizations* are observed to be subjected to a more relative harm (normalized by the number of occurrences of these entities in memes). However, the stacked plots for the top-5 harmfully targeted communities *Mexicans, Black, Muslim, Islam*, and *Russian* in Fig. 3c show relatively less harm targeting these communities.

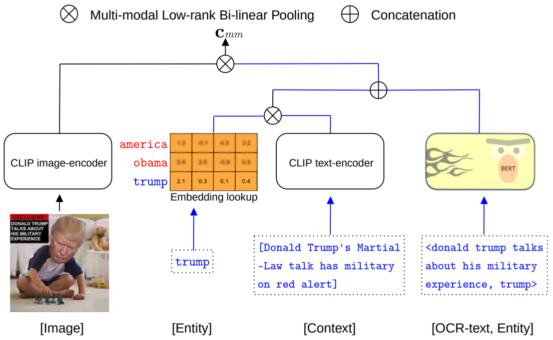

Multi-modal Low-rank Bi-linear Pooling
Concatenation
$\mathbf{c}_{mm}$
CLIP image-encoder
america
obama
trump
Embedding lookup
CLIP text-encoder
BERT
DONALD TRUMP TALKS ABOUT HIS MILITARY EXPERIENCE
trump
[Donald Trump's Martial -Law talk has military on red alert]
<donald trump talks about his military experience, trump>
[Image]
[Entity]
[Context]
[OCR-text, Entity]

Figure 4: The architecture of our proposed approach DISARM. Here, $\mathbf{c}_{mm}$ is the multimodal representation used for the final classification.

4 Proposed Approach

Our proposed model DISARM, as depicted in Fig. 4, is based on a fusion of the textual and the visual modalities, explicitly enriched via contextualised representations by leveraging CLIP (Radford et al., 2021). We chose CLIP as a preferred encoder module for contextualization, due to its impressive zero-shot multimodal embedding capabilities. At first, valid entities are extracted automatically, as part of the process of creating training/validation sets. Then, for each meme, we first obtain the *contextualized-entity* (CE) representation by fusing the CLIP-encoded context and the entity representation. CE is then fused with BERT-based (Devlin et al., 2019) *embedded-harmfulness* (EH) encoding fine-tuned on the OCR-extracted text and entities as inputs. We call the resulting fusion output a *contextualized-text* (CT) representation. CT is then fused with the *contextualized-image* (CI) representation, obtained using the CLIP encoder for the image. We, henceforth, refer to the resulting enriched representation as the *contextualized multimodal* (CMM) representation. We modify the multimodal low-rank bi-linear pooling (Kim et al., 2017) to fuse the input representation into a joint space.

This approach, as can be seen in the subsequent sections below, not only can capture complex cross-modal interactions, but it also provides an efficient fusion mechanism towards obtaining a context-enriched representation. Finally, we use this representation to train a classifier for our task. We describe each module in detail below.

Low-rank Bi-linear Pooling (LRBP). We begin by revisiting *low-rank bi-linear pooling* to set the necessary background. Due to the 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, the number of parameters and hence the complexity, are 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 $\mathbb{1} \in \mathbb{R}^d$ is a column vector of ones, and $\circ$ is Hadamard product. f_i in Equation (1) can be further re-written to obtain $\mathbf{f}$ as follows:

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

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

We further introduce a non-linear activation formulation for LRBP, following Kim et al. (2017), who 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:

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

We slightly modify the multimodal low-rank bi-linear pooling (MMLRBP). Instead of directly projecting the input $\mathbf{x} \in \mathbb{R}^N$ and $\mathbf{y} \in \mathbb{R}^M$ into 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 representation $\mathbf{f}_{mm}$, as expressed below:

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

Structured Context. Towards modelling auxiliary knowledge, we curate *contexts* for the memes in Ext-Harm-P. First, we use the meme text as a search query[3] to retrieve relevant contexts, using the title and the first paragraph of the resulting top document as a *context*, which we call con.

Contextualized-entity Representation (CE). Towards modelling the context-enriched entity, we first obtain the embedding of the input entity ent. Since we have a finite set of entities referenced in the memes in our training dataset, we perform a lookup in the embedding matrix from $\mathbb{R}^{V\times H}$ to obtain the corresponding entity embedding $\mathbf{ent} \in \mathbb{R}^H$, with $H = 300$ being the embedding dimension and V the vocabulary size. We train the embedding matrix from *scratch* as part of the overall training of our model. We project the obtained entity representation $\mathbf{ent}$ into a 512-dimensional space, which we call $\mathbf{e}$. To augment a given entity with relevant contextual information, we fuse it with a contextual representation $\mathbf{c} \in \mathbb{R}^{512}$ obtained by encoding the associated context (con) using CLIP. We perform this fusion using our adaptation of the multimodal low-rank bi-linear pooling as defined by Equation (4). This yields the following contextualized-entity (CE) representation $\mathbf{c}_{ent}$:

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

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

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 an *embedded-harmfulness* (EH) representation. The concatenated representation from $\mathbb{R}^{1280}$ is then projected non-linearly into a lower dimension using a dense layer of size 512. We call the resulting vector $\mathbf{c}_{txt}$ a *contextualized-text* (CT) representation:

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

where $\mathbf{W} \in \mathbb{R}^{1280\times 512}$.

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

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 the CLIP image-encoder. The operation is expressed as follows:

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

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 experiment with various unimodal (image/text-only) and multimodal models, including such pre-trained on multimodal datasets such as MS COCO (Lin et al., 2014) and CC (Sharma et al., 2018). We train `DISARM` and all unimodal baselines using PyTorch, while for the multimodal baselines, we use the MMF framework.[4] [5]

5.1 Evaluation Measures

For evaluation, we use commonly used macro-average versions of accuracy, precision, recall, and F1 score. For example, we discuss the harmful class recall, which is relevant for our study as it characterizes the model's performance at detecting *harmfully* targeting memes. All results we report are averaged over five independent runs.

Evaluation Strategy. With the aim of having a realistic setting, we pose our evaluation strategy as an open-class one. We train all systems using under-sampling of the entities that were not targeted in a harmful way: using all positive (harmful) examples and twice as many negative (non-harmful) ones. We then perform an open-class testing, for all referenced entities (some possibly unseen on training) per meme, effectively making the evaluation more realistic. To this end, we formulate three testing scenarios as follows, with their Harmful (H) and Non-harmful (N) counts:

1. **Test set A (316H, 296N)**: All examples in this dataset are about entities that were *seen* during training.
2. **Test set B (27H, 94N)**: The examples in this set are about entities that were *not seen* as *harmful* during training.
3. **Test set C (16H, 76N)**: All examples are about entities that were completely *unseen* during training.

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., 2019), which implements a generalized auto-regressive pre-training strategy.

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

[4] `github.com/facebookresearch/mmf`

[5] Additional details along with the values of the hyper-parameters are given in Appendix A.

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

System	Modality	Approach	Test Set A								Test Set B							
			Acc	**Prec**	**Rec**	**F1**	Not-harmful		Harmful		**Acc**	**Prec**	**Rec**	**F1**	Not-harmful		Harmful	
							P	**R**	**P**	**R**					**P**	**R**	**P**	**R**
Baselines	Unimodal	XLNet Text-only	0.6765	0.69	0.67	0.6663	0.73	0.52	0.65	0.82	0.5041	0.425	0.405	0.4060	0.72	0.59	0.13	0.22
		VGG Image-only	0.7451	0.75	0.745	0.7438	0.71	0.81	0.79	0.68	0.5455	0.42	0.405	0.4101	0.73	0.66	0.11	0.15
		GRU Text-only	0.7484	0.745	0.75	0.7473	0.73	0.76	0.76	0.74	0.5455	0.43	0.42	0.4210	0.73	0.65	0.13	0.19
		VIT Image only	0.7647	0.765	0.765	0.7642	0.74	0.79	0.79	0.74	0.5207	0.525	0.535	0.4843	0.8	0.51	0.25	0.56
	Multimodal	ViLBERT CC	0.6895	0.69	0.685	0.6835	0.71	0.6	0.67	0.77	0.438	0.535	0.53	0.4302	0.82	0.35	0.25	**0.71**
		MM Transformer	0.6993	0.71	0.695	0.6926	0.75	0.57	0.67	0.82	**0.7769**	0.53	0.575	0.5032	0.78	0.51	0.28	0.64
		VisualBERT	0.7026	0.725	0.69	0.6918	**0.78**	0.54	0.67	0.84	0.5537	0.545	0.565	0.5108	0.82	0.54	0.27	0.59
		VisualBERT – COCO	0.7059	0.71	0.7	0.7014	0.73	0.62	0.69	0.78	0.5785	0.53	0.545	0.5147	0.8	0.61	0.26	0.48
		MMBT	0.7157	0.72	0.71	0.7121	0.74	0.64	0.7	0.78	0.6116	0.54	0.55	0.5310	0.81	0.66	0.27	0.44
		ViLBERT	0.7516	0.755	0.75	0.7495	**0.78**	0.68	0.73	0.82	0.6612	0.58	0.595	0.5782	**0.83**	0.71	0.33	0.48
Prop. system & variants		CE + CI (concat)	0.7353	0.74	0.735	0.7361	0.71	0.77	0.77	0.7	0.4793	0.46	0.44	0.4230	0.74	0.51	0.18	0.37
		CE + CI (MMLRBP)	**0.781**	**0.785**	0.78	0.7790	0.74	**0.84**	**0.83**	0.72	0.562	0.535	0.545	0.5079	0.81	0.57	0.26	0.52
		EH + CI (concat)	0.6634	0.665	0.66	0.6609	0.67	0.6	0.66	0.72	0.5868	0.505	0.51	0.4964	0.78	0.65	0.23	0.37
		EH + CI (MMLRBP)	0.7255	0.73	0.725	0.7260	0.74	0.67	0.72	0.78	0.6612	0.545	0.555	0.5470	0.8	0.74	0.29	0.37
		DISARM	**0.781**	0.74	**0.835**	**0.7845**	0.74	0.81	0.74	**0.86**	0.74	**0.605**	**0.74**	**0.6498**	**0.83**	**0.79**	**0.38**	0.69
$\Delta_{(\texttt{DISARM}-ViLBERT)\times 100}(\%)$			↑ 2.94%	↓ 1.5%	↑ 8%	↑ 3.5%	↓ 4%	↑ 13%	↑ 1%	↑ 4%	↑ 7.88%	↑ 2.5%	↑ 14.5%	↑ 7.16%	–	↑ 8%	↑ 5%	↑ 21%

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

Sys		Approach	Acc	Prec	Rec	F1	Not-harmful		Harmful	
							P	**R**	**P**	**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 models vs. DISARM (and its variants) on Test Set C.

Experimental Results. We compare the performance of several unimodal and multimodal systems (pre-trained or trained from scratch) vs. DISARM and its variants. All systems are evaluated using the 3-way testing strategy described above. We then perform ablation studies on representations that use the *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 (see Appendix B for a detailed ablation study).[7] This is followed by interpretability analysis. Finally, we discuss the limitations of DISARM by performing error analysis (details in Appendix C).

All Entities Seen During Training: In our unimodal text-only baseline experiments, the GRU-based system yields a relatively lower *harmful* recall of 0.74 compared to XLNet's 0.82, but a better overall F1 score of 0.75 vs. 0.67 for XLNet, as shown in Table 2. The lower *harmful* precision of 0.65 and the *not-harmful* recall of 0.52 contribute to the lower F1 score for XLNet.

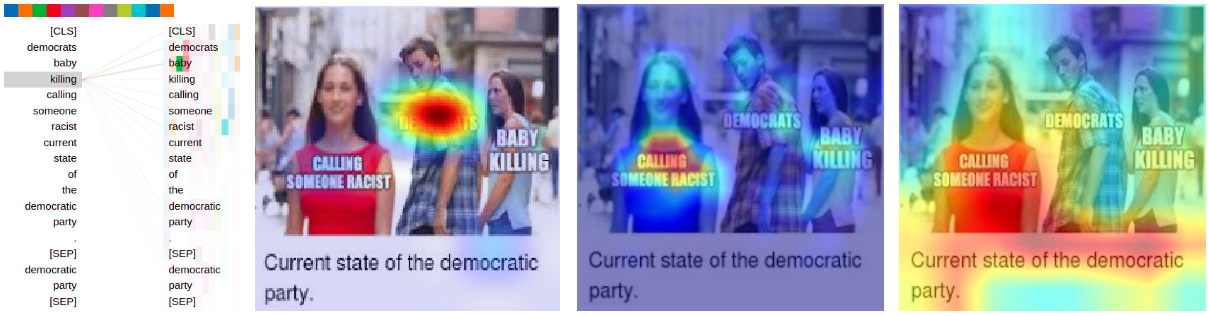
Among the image-only unimodal systems, VGG performs better with a *non-harmful* recall of 0.81, but its poor performance for detecting harmful memes yields a lower *harmful* recall of 0.68. At the same time, VIT has a relatively better *harmful* recall of 0.74. Overall, the unimodal results (see Table 2) indicate the efficacy of self-attention over convolution for images and RNN (GRU) sequence modeling for text.

Multimodally pre-trained models such as VisualBERT and ViLBERT yield moderate F1 scores of 0.70 and 0.68, and *harmful* recall of 0.78 and 0.77, respectively (see Table 2). Fresh training facilitates more meaningful results in favour of *non-harmful* precision of 0.78 for both models, and *harmful* recall of 0.84 and 0.82 for VisualBERT and ViLBERT, respectively. Overall, ViLBERT yields the most balanced performance of 0.75 in terms of F1 score. It can be inferred from these results (see Table 2) that multimodal pre-training leverages domain relevance.

We can see in Table 2 that multimodal low-rank bi-linear pooling distinctly enhances the performance in terms of F1 score. The improvements can be attributed to the fusion of the CE and EH representations, respectively, with CI, instead of a simple concatenation. This is more prominent for CE with an F1 score of 0.78, which shows the importance of modeling the background context. Finally, DISARM yields a balanced F1 score of 0.78, with a reasonable precision of 0.74 for *non-harmful* category, and the best recall of 0.86 for the *harmful* category.

All Entities Unseen as Harmful Targets During Training: With Test Set B, the evaluation is slightly more challenging in terms of the entities to be assessed, as these were never seen at training time as *harmful*.

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



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

Unimodal systems perform poorly on the *harmful* class, with the exception of XLNet (see Table 2), where the *harmful* class recall as 0.56. For the multimodal baselines, systems pre-trained using COCO (VisualBERT) and CC (ViLBERT) yield a 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 common-sense reasoning helping such systems, on a test set that is more open-ended compared to Test Set A. Their non-pre-trained versions along with the MM Transformer and MMBT achieve better F1 scores, but with low *harmful* recall.

Multimodal fusion using MMLRBP improves the *harmful* class recall for CE to 0.52, but yields lower values of 0.37 for EH fusion with CI (see Table 2). This reconfirms the utility of the context. In comparison, DISARM yields a balanced F1 score of 0.65 with the best precision of 0.83 and 0.38, along with decent recall of 0.79 and 0.69 for *non-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 of 0.62 for XLNet, as shown in Table 3. In the current scenario (Test Set C), none of the entities being assessed at 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 the 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 (see Table 3) sizable boost for the *harmful* class recall for MMLRBP-based multimodal fusion of CI with CE (0.69%), against a decrease with EH (0.31%). In comparison, DISARM yields a low, yet the best *harmful* precision of 0.36, and a moderate recall of 0.70 (see Table 3). Moreover, besides yielding reasonable precision and recall of 0.86 and 0.76 for the *non-harmful* class, DISARM achieves 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 its characteristic modelling and context-based fusion. DISARM demonstrates an ability to detect harmful targeting for a diverse set of entities. Specifically, the three-way testing setup inherently captures the efficacy with which DISARM can detect *unseen* harmful targets. The prediction for entities *completely* unseen on training yields better results (see Tables 2 and 3), and suggests possibly induced bias in the former scenario. Moreover, it is a direct consequence of the fact that we were able to incorporate only a limited set of the 246 potential targets. Overall, we argue that DISARM generalizes well for unseen entities with 0.65 and 0.64 macro-F1 scores, as compared to ViLBERT's 0.58 and MMBT's 0.51, for Test Sets B and C, respectively.

Comparative Diagnosis. Despite the marginally better *harmful* recall for ViLBERT (CC) on Test Set B (see Table 2) and Test Set C (see Table 3), the overall balanced performance of DISARM appears to be reasonably justified based on the comparative interpretability analysis between the attention maps for the two systems.

Fig. 5 shows the attention maps for an example meme. It depicts a meme that is *correctly* predicted to *harmfully* target the *Democratic Party* by DISARM and incorrectly by ViLBERT. As visualised in Fig. 5a, the 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 the *Democratic Party*, both as part of the OCR-extracted text and the target-candidate, jointly encoded by BERT. The multimodal attention leveraged by DISARM is depicted (via the CLIP encoder) in Fig. 5b, demonstrating the utility of contextualised attention over the *male* figure that represents an attack on the *Democratic Party*. Also, DISARM has a relatively focused field of vision, as shown in Fig. 5c, as compared to a relatively scattered one for ViLBERT (see Fig. 5d). This suggest a better multimodal modelling capacity for DISARM as compared to ViLBERT.

6 Conclusion and Future Work

We introduced the novel task of detecting the targeted entities within harmful memes and we 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, which uses an adaptation of multimodal low-rank bi-linear pooling-based fusion strategy at different levels of representation abstraction. We showed that DISARM outperforms various uni/multi-modal baselines in three different scenarios by 4%, 7%, and 13% increments in terms of macro-F1 score, respectively. Moreover, DISARM achieved a relative error rate reduction of 9% over the best baseline. We further emphasized the utility of different components of DISARM through ablation studies. We also elaborated on the generalizability of DISARM, thus confirming its modelling superiority over ViLBERT via interpretability analysis. We finally analysed the shortcomings in DISARM that lead to incorrect harmful target predictions.

In the present work, we made an attempt to elicit some inherent challenges pertaining to the task at hand: augmenting the relevant context, effectively fusing multiple modalities, and pre-training. Yet, we also leave a lot of space for future research for this novel task formulation.

Ethics and Broader Impact

Reproducibility. We present detailed hyper-parameter configurations in Appendix A and Table 4. The source code, and the dataset Ext-Harm-P are available at https://github.com/LCS2-IIITD/DISARM

User Privacy. The information depicted/used does not include any personal information. Copyright aspects are attributed to the dataset source.

Annotation. The annotation was conducted by NLP experts or linguists in India, who were fairly treated and were duly compensated. We conducted several discussion sessions to make sure all annotators could understand the distinction between harmful vs. non-harmful referencing.

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 in 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 for annotation.

Misuse Potential. Our dataset can be potentially used for ill-intended purposes, such as 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, large-scale models require a lot of computations, which contribute 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 a relatively small dataset.

Acknowledgments

The work was partially supported by a Wipro research grant, Ramanujan Fellowship, the Infosys Centre for AI, IIIT Delhi, and ihub-Anubhuti-iiitd Foundation, set up under the NM-ICPS scheme of the Department of Science and Technology, India. It is also part of the Tanbih mega-project, developed at the Qatar Computing Research Institute, HBKU, which aims to limit the impact of "fake news," propaganda, and media bias by making users aware of what they are reading, thus promoting media literacy and critical thinking.

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. *arXiv 2103.12541*.

Monther Aldwairi and Ali Alwahedi. 2018. Detecting Fake News in Social Media Networks. *Procedia Computer Science*, 141:215–222.

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 *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, CSCW '17, pages 1217–1230, Portland, Oregon, USA. Association 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 *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, EMNLP '14, pages 1724–1734, Doha, Qatar. Association 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. *arXiv/2012.14891*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL-HLT '19, pages 4171–4186, Minneapolis, Minnesota, USA.

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 Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, ACL-IJCNLP '21, pages 6603–6617.

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. In *Proceedings of the 9th International Conference on Learning Representations*, ICLR '21, Vienna, Austria.

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. 2020. Exploring Hate Speech Detection in Multimodal Publications. In *Proceedings of the 2020 IEEE Winter Conference on Applications of Computer Vision*, WACV '20, pages 99–1467, Snowmass Village, CO, USA.

Eduardo Graells-Garrido, Ricardo Baeza-Yates, and Mounia Lalmas. 2020. Every Colour You Are: Stance Prediction and Turnaround in Controversial Issues. In *Proceedings of the 12th ACM Conference on Web Science*, WebSci '20, pages 174–183, Southampton, UK. ACM.

Momchil Hardalov, Arnav Arora, Preslav Nakov, and Isabelle Augenstein. 2022. A survey on stance detection for mis- and disinformation identification. In *Findings of NAACL 2022*, Seattle, Washington, USA.

Kimmo Karkkainen and Jungseock Joo. 2021. Fairface: Face Attribute Dataset for Balanced Race, Gender, and Age for Bias Measurement and Mitigation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 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 *Proceedings 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 *Proceedings 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. In *Proceedings of the 5th International Conference on Learning Representations*, ICLR '17, Toulon, France.

Seunghyun Kim, Afsaneh Razi, Gianluca Stringhini, Pamela J. Wisniewski, and Munmun De Choudhury. 2021. A Human-Centered Systematic Literature Review of Cyberbullying Detection Algorithms. *Proceedings ACM Hum. Comput. Interact.*, 5(CSCW2):1–34.

Diederik P Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *Proceedings 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 *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, ACL '19, pages 5047–5058, Florence, Italy. Association 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, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft COCO: Common Objects in Context. In *Proceedings 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 *Proceedings 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. 2019. 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 *Proceedings of the Fourteenth International AAAI Conference on Web and Social Media*, ICWSM '20, pages 452–463, Atlanta, Georgia, USA.

Van-Hoang Nguyen, Kazunari Sugiyama, Preslav Nakov, and Min-Yen Kan. 2020. FANG: Leveraging social context for fake news detection using graph representation. In *Proceedings of the 29th ACM International Conference on Information and Knowledge Management*, CIKM '20, pages 1165–1174.

Benet Oriol, Cristian Canton-Ferrer, and Xavier Giró i Nieto. 2019. Hate Speech in Pixels: Detection of Offensive Memes towards Automatic Moderation. In *Proceedings of the NeurIPS 2019 Workshop on AI for Social Good*, Vancouver, Canada.

Hamed Pirsiavash, Deva Ramanan, and Charless Fowlkes. 2009. Bilinear classifiers for visual recognition. In *Advances in Neural Information Processing Systems: Proceedings of the International Conference on Neural Information Processing Systems*, volume 22, pages 1482–1490, Vancouver British Columbia Canada. 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 Association for Computational Linguistics*, ACL-IJCNLP '21, pages 2783–2796, Bangkok, Thailand.

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 Association for Computational Linguistics: EMNLP 2021*, EMNLP 21, pages 4439–4455, Punta Cana, Dominican Republic. 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, et al. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings 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. *Proceedings 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.

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. *Proceedings ACM Hum.-Comput. Interact.*, 5(CSCW2).

Shaden Shaar, Firoj Alam, Giovanni Da San Martino, and Preslav Nakov. 2022. The role of context in detecting previously fact-checked claims. In *Findings of the Association for Computational Linguistics: NAACL-HLT 2022*, NAACL-HLT '22, Seattle, Washington, USA.

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 *Proceedings of the 2021 IEEE 17th International Conference on eScience*, eScience '21, 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. *Information Processing & Management*, 58(5):102664.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual Captions: A Cleaned, Hypernymed, Image Alt-text Dataset For Automatic Image Captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, ACL '18, pages 2556–2565, Melbourne, Australia.

Shivam Sharma, Firoj Alam, Md. Shad Akhtar, Dimitar Dimitrov, Giovanni Da San Martino, Hamed Firooz, Alon Halevy, Fabrizio Silvestri, Preslav Nakov, and Tanmoy Chakraborty. 2022. Detecting and understanding harmful memes: A survey. In *Proceedings of the 31st International Joint Conference on Artificial Intelligence*, IJCAI-ECAI '22, Vienna, Austria.

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. In *Proceedings of the International Conference on Learning Representations*, ICLR '15, San Diego, California, USA.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and Policy Considerations for Deep Learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association 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 *Proceedings 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. 2021. Understanding the Use of Fauxtography on Social Media. *Proceedings of the International AAAI Conference on Web and Social Media*, 15(1):776–786.

Liang Wu and Huan Liu. 2018. Tracing Fake-News Footprints: Characterizing Social Media Messages by How They Propagate. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, WSDM '18, pages 637–645, Marina Del Rey, CA, USA. Association for Computing Machinery.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. XLNet: Generalized Autoregressive Pretraining for Language Understanding. In *Advances in Neural Information Processing Systems*, volume 32 of *NIPS '19*, Vancouver, BC, Canada.

Marcos Zampieri, Preslav Nakov, Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Hamdy Mubarak, Leon Derczynski, Zeses Pitenis, and Çağrı Çöltekin. 2020. SemEval-2020 task 12: Multilingual offensive language identification in social media (OffensEval 2020). In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, SemEval '20, pages 1425–1447, Barcelona (online). International Committee for Computational Linguistics.

Ziqi Zhang and Lei Luo. 2019. Hate Speech Detection: A Solved Problem? The Challenging Case of Long Tail on Twitter. *Semantic Web*, 10(5):925–945.

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

Kaimin Zhou, Chang Shu, Binyang Li, and Jey Han Lau. 2019. Early Rumour Detection. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL-HLT '19, pages 1614–1623, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Yi Zhou, Zhenhao Chen, and Huiyuan Yang. 2021. Multimodal Learning for Hateful Memes Detection. In *Proceedings of the 2021 IEEE International Conference on Multimedia & Expo Workshops*, ICMEW '21, pages 1–6, Shenzhen, China.

Appendix

A Implementation Details and Hyper-parameter Values

We trained all our models using PyTorch on NVIDIA Tesla V100 GPU, with 32 GB dedicated memory, CUDA-11.2 and cuDNN-8.1.1 installed. For the unimodal models, we imported all the pre-trained weights from the TORCHVISION.MODELS[8], a sub-package of the PyTorch framework. We initialized 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 extended it by manually annotating for *harmful* targets, followed by including *non-harmful* examples using automated entity extraction (textual and visual) strategies for training/validation splits and manual annotation (for both harmful and non-harmful) for the test split.

When training our models and exploring various values for the different model hyper-parameters, we experimented with using the Adam optimizer (Kingma and Ba, 2015) with a learning rate of $1e^{-4}$, 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 select the best hyper-parameter values. We also used early stopping for saving the best intermediate checkpoints. Table 4 gives more detail about the hyper-parameters we used for training. On average, it took approximately 2.5 hours to train a multi-modal neural model.

		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].

B Ablation Study

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

B.1 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 lowest value being 0.66 for XLNet (text-only) model. The results for unimodal systems, as can be observed in Table 5, is satisfactory with values of 0.74, 0.73, and 0.77 for CE EH, and CI unimodal systems, respectively. For multimodal systems, we can observe distinct lead for the MMLRBP-based fusion strategy, for both CE and EH systems over the concatenation-based approach, except for EH's recall drop by 7%. Finally DISARM yields the best overall F1 score of 0.78.

B.2 Test Set B

With *context* not having any harmfulness cues for a given meme when considered in isolation, the unimodal CE module performs the worst with 0.48 F1 score, and 0.07 recall for the *harmful* class, 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 terms of 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 score and the *harmful* recall by 2% and 3%, respectively (see Table 5). Unimodal assessment of CI performs moderately with an F1 score of 0.51, but with a poor *harmful* recall of 0.15. MMLRBP, towards joint-modeling of CE and CI yields a significant boost in the *harmful* recall to 0.52 (see Table 5). On the other hand, MMLRBP-based 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. A complementary impact of CE, EH, and CI is observed for DISARM with a balanced F1 score of 0.6 and a competitive *harmful* recall value of 0.69.

[8] 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.

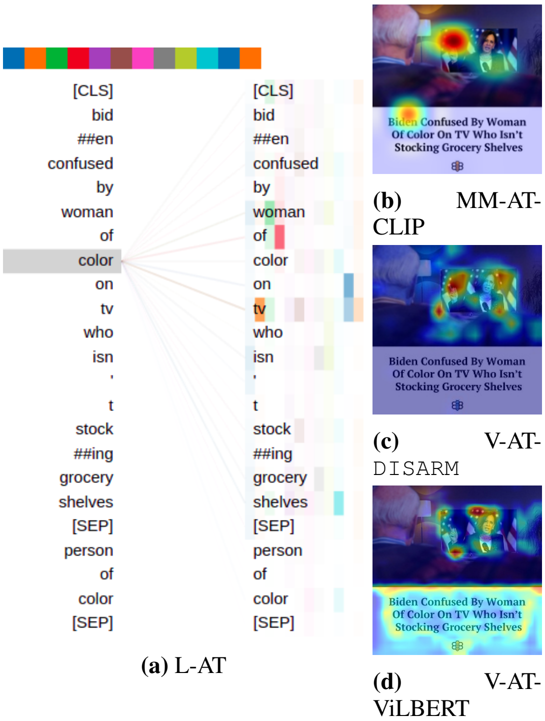
B.3 Test Set C

As observed in the previous scenario, 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 including EH, and its joint modelling with CE with improved F1 scores of 0.56 and 0.58, respectively, 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 (see Table 5), along with a poor *harmful* recall of 0.19, which shows its inadequacy to model harmful targeting on its own. For multimodal setups, the joint modelling of CE and CI benefits from MMLRBP based fusion, yielding a gain of 7% and 13% in terms of F1 score 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 experimental results here are for comparison and analysis of the optimal set of design and baseline choices. We should note that we performed extensive experiments as part of our preliminary investigation, with different contextual modelling strategies, attention mechanisms, modelling choices, etc., to reach a conclusive architectural configuration that show promise for addressing the task of target detection in harmful memes.

C Error Analysis

It is evident from the results shown in Tables 2 and 3 that DISARM still has shortcomings. Examples like the one shown in Fig. 6 are seemingly *harmless*, both textually and visually, but imply serious *harm* to a *person of color* in an implicit way.



(a) L-AT

(b) MM-AT-CLIP

(c) V-AT-DISARM

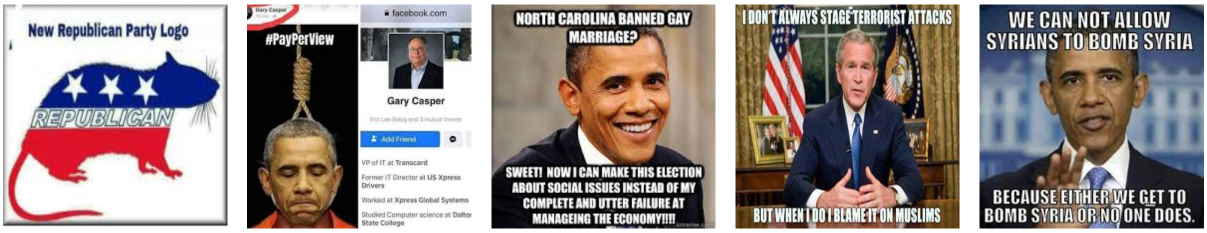
(d) V-AT-ViLBERT

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

This kind of complexity can be challenging to model without providing additional context about the meme 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., 2019). In this particular example, despite modelling contextual information explicitly in DISARM, it misclassifies this meme anyway.


New Republican Party Logo
REPUBLICAN
#PayPerView
facebook.com
Gary Casper
Add Friend
VP of IT at Transcard
Former IT Director at US Xpress Drivers
Worked at Xpress Global Systems
Studied Computer science at Dalton State College
NORTH CAROLINA BANNED GAY MARRIAGE?
SWEET! NOW I CAN MAKE THIS ELECTION ABOUT SOCIAL ISSUES INSTEAD OF MY COMPLETE AND UTTER FAILURE AT MANAGEING THE ECONOMY!!!!
I DON'T ALWAYS STAGE TERRORIST ATTACKS
BUT WHEN I DO I BLAME IT ON MUSLIMS
WE CAN NOT ALLOW SYRIANS TO BOMB SYRIA
BECAUSE EITHER WE GET TO BOMB SYRIA OR NO ONE DOES.

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

Figure 7: Examples of memes depicting different types (a)–(e) of *harmful* targeting.

Even though the context obtained for this meme pertains to its content (see Fig. 6), 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 message 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. 6a, Fig. 6b, and Fig. 6c, respectively. On the other hand, as shown in the visual attention plot for ViLBERT in Fig. 6d, the field of view that is being attended encompasses the visuals of *Kamala Harris*, who is the *person of colour* that i sbeing primarily targeted by 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 making use of multimodal pre-training using the dataset that is not as generic as MS COCO (Lin et al., 2014), but facilitates modelling of the complex real-world multimodal information, especially for tasks related to memes.

D Annotation Guidelines

Before discussing some details about the annotation process, revisiting the definition of *harmful* memes would set the pretext towards consideration of *harmful* targeting and *non-harmful* referencing. According to Pramanick et al. (2021b), a harm can be expressed as an abuse, an offence, a disrespect, an insult, or an insinuation of a targeted entity or any socio-cultural or political ideology, belief, principle, or doctrine associated with that entity. The harm can also be in the form of a more subtle attack such as mocking or ridiculing a person or an idea.

Another common understanding[9,10,11] about the harmful content is that it could be anything online that causes distress. It is an extremely subjective phenomenon, wherein what maybe be harmful to some might not be considered an issue by others. This makes it significantly challenging to characterize and hence to 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 the following four categories: (*i*) physical harm, (*ii*) emotional harm, (*iii*) relational harm, and (*iv*) financial harm. With this in mind, we define two types of referencing that we have investigated in our work within the context of internet memes: (*i*) *harmful* and (*ii*) *non-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 such as 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. Non-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 falls under this category.

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

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

[11] 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 their target categories. The total frequency for each word is shown in parentheses.

D.2 Characteristics of Harmful Targeting

There are several factors that collectively facilitate the characterisation of *harmful* targeting in memes. Here are some:

1. A prominent way of harmfully targeting an entity in a meme is by leveraging sarcastically harmful analogies, framed via either textual or visual instruments (see Fig. 7a).

2. There could be multiple entities being harmfully targeted within a meme as depicted in Fig. 2. Hence, annotators were asked to provide all such targets as harmful, with no exceptions.

3. A harmful targeting within a meme could have visual depictions that are either gory, violent, graphically sensitive, or pornographic (see Fig. 7b).

4. Any meme that insinuates an entity on either social, political, professional, religious grounds, can cause harm (see Fig. 7c and 7d).

5. Any meme that implies an explicit/implicit threat to an individual, a community, a national or an international entity is harmful (see Fig. 7d and 7e).

6. Whenever there is any ambiguity regarding the harmfulness of any reference being made, we requested the annotators to proceed following the best of their understanding.

E Ext-Harm-P Characteristics

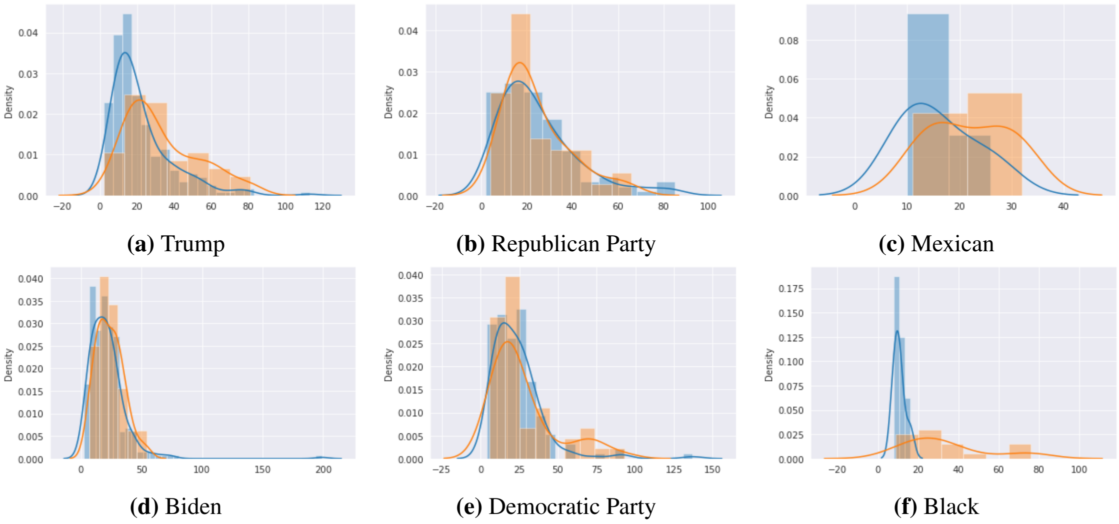
Below, we perform some analysis of the lexical content of the length of the meme text.

E.1 Lexical Analysis

Interestingly, a significant number of memes are disseminated making references to popular *individuals* such as *Joe Biden, Donald Trump, etc.*, as can be observed for individual sub-categories (for both harmful and non-harmful memes) in Table 6.

We can see in Table 6 that for *harmful–organization*, the top-5 harmfully targeted organizations include the top-2 leading political organizations in the USA (the *Democratic Party* and the *Republican Party*), which are of significant political relevance, followed by the *Libertarian Party*, a media outlet (*CNN*), and finally the generic *government*. At the same time, non-harmfully referenced organizations includes the *Biden camp* and the *Trump administration*, which are mostly leveraged for harmfully targeting (or otherwise) the associated public figure. Finally, communities such as *Mexicans, Black, Muslim, Islam*, and *Russian* are often immensely prejudiced against online, and thus also in our meme dataset. At the same time, non-harmfully targeted communities such as the *Trump supporters* and the *African Americans* are not targeted as often as the aforementioned ones, as we can see in Table 6.

The above analysis of the lexical content of the memes in our datasets largely emphasizes the inherent bias that multimodal content such as memes can exhibit, which in turn can have direct influence on the efficacy of machine/deep learning-based systems for detecting the entities targeted by harmful memes. 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. The mutual exclusion for harmful vs. non-harmful categories for community shows the inherent bias that could pose a challenge, even for the best multi-modal deep neural systems. The high pervasiveness of a few prominent keywords could effectively lead to increasing bias towards them for specific cases. At the same time, the significant overlap observed in Table 6 for the enlisted entities, between harmful and not-harmful individuals, highlights the need for sophisticated multi-modal systems that can effectively reason towards making a complex decision like detecting harmful targeting within memes, rather than exploit the biases towards certain entities in the training data.



(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 non-harmful (Orange). The depiction is for Individual: (a) and (d); Organization: (b) and (e); and Community: (c) and (f).

E.2 Meme-Message Length Analysis

Most of the *harmful* memes are observed to be created using texts of length 16–18 (see Fig. 8). At the same time, *not-harmful* meme-text lengths have a relatively higher standard deviation, possibly due to the diversity of *non-harmful* messages. *Trump* and the *Republic Party* have meme-text length distributions similar to the *non-harmful* category: skewing left, but gradually decreasing towards the right. This suggests a varying content generation pattern amongst meme creators (see Fig. 8). The meme-text length distribution for *Biden* closely approximates a normal distribution with a low standard deviation. Both categories would pre-dominantly entail creating memes with shorter text lengths, possibly 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 instances fall within the 50–75 meme-text length range. The overall harmful and non-harmful meme-text length distribution is observed to be fairly distributed across different meme-text lengths for *Mexican*. At the same time, the amount of harm intended towards the *Black* community is observed to be significantly higher, as compared to moderately distributed *non-harmful* memes depicted by the corresponding meme-text length distribution in Fig. 8.