DISARM: Detecting the Victims Targeted by Harmful Memes

Shivam Sharma¹,³, Md. Shad Akhtar¹, Preslav Nakov², Tanmoy Chakraborty¹
¹Indraprastha Institute of Information Technology - Delhi, India
²Qatar Computing Research Institute, HBKU, Doha, Qatar
³Wipro 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.

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. 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., 2021a) 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.

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

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

<table>
<thead>
<tr>
<th>Split</th>
<th># Examples</th>
<th>Harmful</th>
<th>Not-harmful</th>
</tr>
</thead>
<tbody>
<tr>
<td>Train</td>
<td>3,618</td>
<td>1,206</td>
<td>2,412</td>
</tr>
<tr>
<td>Validation</td>
<td>216</td>
<td>72</td>
<td>144</td>
</tr>
<tr>
<td>Test</td>
<td>612</td>
<td>316</td>
<td>296</td>
</tr>
<tr>
<td>Total</td>
<td>4,446</td>
<td>1,594</td>
<td>2,852</td>
</tr>
</tbody>
</table>

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.

### 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.

### 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)

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):

\[^2\text{NER using SpaCy \& PID using http://github.com/ageitgey/mega_recognition.}\]
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.

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

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 $W_i$. Consequently, the number of parameters and hence the complexity, are reduced. The weight matrix $W_i$ is re-written as $W_i = U_i V_i^T$, where $U_i \in \mathbb{R}^{N \times d}$ and $V_i \in \mathbb{R}^{M \times d}$, 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:

$$ f_i = x^T W_i y = x^T U_i V_i^T y = I^T(U_i^T x \circ V_i^T y)$$  \hspace{1cm} (1) $$

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

$$ f = P^T(U^T x \circ V^T y) + b$$ \hspace{1cm} (2) $$

where $f \in \{f_i\}$, $P \in \mathbb{R}^{d \times c}$, $b \in \mathbb{R}^c$, $d$ 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:

$$ f = P^T \tanh(U^T x \circ V^T y) + b$$ \hspace{1cm} (3) $$

We slightly modify the multimodal low-rank bi-linear pooling (MMLRBP). Instead of directly projecting the input $x \in \mathbb{R}^N$ and $y \in \mathbb{R}^M$ 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 $x_{mm} \in \mathbb{R}^{N \times d}$ and $y_{mm} \in \mathbb{R}^{N \times d}$ to obtain a multimodal fused representation $f_{mm}$, as expressed below:

$$ f_{mm} = P^T \tanh(U^T x_{mm} \circ V^T y_{mm})$$ \hspace{1cm} (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 to retrieve relevant contexts, using the title and the first paragraph of the resulting top document as a context, which we call $\text{con}$.

**Contextualized-entity Representation (CE).** Towards modelling the context-enriched entity, we first obtain the embedding of the input entity $\text{ent}$. 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 $\text{ent} \in \mathbb{R}^H$, 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 $\text{ent}$ into a 512-dimensional space, which we call $e$. To augment a given entity with relevant contextual information, we fuse it with a contextual representation $c \in \mathbb{R}^{512}$ obtained by encoding the associated context ($\text{con}$) 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 $c_{\text{ent}}$:

$$ c_{\text{ent}} = P_1^T \tanh(U_1^T e \circ V_1^T c) + b$$ \hspace{1cm} (5) $$

where $c_{\text{ent}} \in \mathbb{R}^{512}$, $P_1 \in \mathbb{R}^{256 \times 512}$, $b \in \mathbb{R}^{512}$, $U_1 \in \mathbb{R}^{512 \times 256}$, and $V_1 \in \mathbb{R}^{512 \times 256}$.

**Contextualized-Text (CT) Representation.** Once we obtain the contextualized-entity embedding $c_{\text{ent}}$, we concatenate it with the BERT encoding for the combined representation of the OCR-extracted text and the entity ($o_{\text{ent}} \in \mathbb{R}^{768}$). 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 $c_{\text{ext}}$ a contextualized-text (CT) representation:

$$ c_{\text{ext}} = W_i [o_{\text{ent}}, c_{\text{ent}}] + b$$ \hspace{1cm} (6) $$

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

\footnote{https://pypi.org/project/googlesearch-python/}
Contextualized Multimodal (CMM) Representation. Once we obtain the contextualized-text representation $c_{txt} \in \mathbb{R}^{512}$, we again perform multimodal low-rank bi-linear pooling using Equation (4) to fuse it with the contextualized-image representation $c_{img} \in \mathbb{R}^{512}$, obtained using the CLIP image-encoder. The operation is expressed as follows:

$$c_{mm} = P_2^T \tanh(U_2^T c_{txt} \circ V_2^T c_{img})$$  \hspace{1cm} (7)

where $c_{mm} \in \mathbb{R}^{512}$, $P_2 \in \mathbb{R}^{256 \times 512}$, $U_2 \in \mathbb{R}^{512 \times 256}$, and $V_2 \in \mathbb{R}^{512 \times 256}$.

Notably, we learn two different projection matrices $P_1$ and $P_2$, 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 corresponding entity as either harmful or non-harmful, 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, XL-Net: 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.  
  - 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).

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.

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.

\(^4\)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
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 ViSaBERT 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.

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

<table>
<thead>
<tr>
<th>System</th>
<th>Modality</th>
<th>Approach</th>
<th>Test Set A</th>
<th>Test Set B</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Acc</td>
<td>Prec</td>
<td>Rec</td>
</tr>
<tr>
<td>DISARM</td>
<td>Unimodal</td>
<td>0.565</td>
<td>0.49</td>
<td>0.59</td>
</tr>
<tr>
<td>VGG</td>
<td>Unimodal</td>
<td>0.475</td>
<td>0.45</td>
<td>0.51</td>
</tr>
<tr>
<td>VGG</td>
<td>Multimodal</td>
<td>0.475</td>
<td>0.45</td>
<td>0.51</td>
</tr>
<tr>
<td>ViLBERT</td>
<td>Multimodal</td>
<td>0.475</td>
<td>0.45</td>
<td>0.51</td>
</tr>
<tr>
<td>VisualBERT</td>
<td>Multimodal</td>
<td>0.475</td>
<td>0.45</td>
<td>0.51</td>
</tr>
<tr>
<td>MM Transformer</td>
<td>Multimodal</td>
<td>0.475</td>
<td>0.45</td>
<td>0.51</td>
</tr>
<tr>
<td>MMBT</td>
<td>Multimodal</td>
<td>0.475</td>
<td>0.45</td>
<td>0.51</td>
</tr>
<tr>
<td>ViLBERT</td>
<td>Multimodal</td>
<td>0.475</td>
<td>0.45</td>
<td>0.51</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>System</th>
<th>Modality</th>
<th>Approach</th>
<th>Test Set B</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Acc</td>
<td>Prec</td>
</tr>
<tr>
<td>DISARM</td>
<td>Unimodal</td>
<td>0.565</td>
<td>0.49</td>
</tr>
<tr>
<td>VGG</td>
<td>Unimodal</td>
<td>0.475</td>
<td>0.45</td>
</tr>
<tr>
<td>VGG</td>
<td>Multimodal</td>
<td>0.475</td>
<td>0.45</td>
</tr>
<tr>
<td>ViLBERT</td>
<td>Multimodal</td>
<td>0.475</td>
<td>0.45</td>
</tr>
<tr>
<td>VisualBERT</td>
<td>Multimodal</td>
<td>0.475</td>
<td>0.45</td>
</tr>
<tr>
<td>MM Transformer</td>
<td>Multimodal</td>
<td>0.475</td>
<td>0.45</td>
</tr>
<tr>
<td>MMBT</td>
<td>Multimodal</td>
<td>0.475</td>
<td>0.45</td>
</tr>
<tr>
<td>ViLBERT</td>
<td>Multimodal</td>
<td>0.475</td>
<td>0.45</td>
</tr>
</tbody>
</table>

### 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 using the contextualized-multimodal modules of DISARM (see Appendix B for a detailed ablation study). 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.

---

7We 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 bi-linear pooling, respectively.
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 suggests 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 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. Finally, we 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


Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and


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, 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 hyperparameters, we experimented with using the Adam optimizer (Kingma and Ba, 2015) with a learning rate of $10^{-4}$, a weight decay of $10^{-5}$, 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.

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.

---

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.

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.
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 is being 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 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.

---

**Notes:**


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.

<table>
<thead>
<tr>
<th>Harmful meme</th>
<th>Not-harmful meme</th>
</tr>
</thead>
<tbody>
<tr>
<td>Individual</td>
<td>Organization</td>
</tr>
<tr>
<td>Individual</td>
<td>Organization</td>
</tr>
<tr>
<td>Joe Biden (333)</td>
<td>Democratic Party (184)</td>
</tr>
<tr>
<td>Donald Trump (285)</td>
<td>Republican Party (130)</td>
</tr>
<tr>
<td>Barack Obama (142)</td>
<td>Libertarian Party (44)</td>
</tr>
<tr>
<td>Hillary Clinton (35)</td>
<td>CNN (6)</td>
</tr>
<tr>
<td>Mike Pence (13)</td>
<td>Government (5)</td>
</tr>
</tbody>
</table>
| `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 include 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.

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.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 Republican 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 predominantly 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.