MOMENTA: A Multimodal Framework for Detecting Harmful Memes and Their Targets

Shraman Pramanick[1]*, **Shivam Sharma**[2,4]*, **Dimitar Dimitrov**[3], **Md Shad Akhtar**[2], **Preslav Nakov**[5], **Tanmoy Chakraborty**[2]
[1]Johns Hopkins University [2]Indraprastha Institute of Information Technology - Delhi
[3]Sofia University [4]Wipro AI Labs, India [5]Qatar Computing Research Institute, HBKU, Doha
spraman3@jhu.edu, {shivams, shad.akhtar, tanmoy}@iiitd.ac.in
mitko.bg.ss@gmail.com, pnakov@hbku.edu.qa

Abstract

Internet memes have become powerful means to transmit political, psychological, and socio-cultural ideas. Although memes are typically humorous, recent days have witnessed an escalation of *harmful* memes used for trolling, cyberbullying, and abuse. Detecting such memes is challenging as they can be highly satirical and cryptic. Moreover, while previous work has focused on specific aspects of memes such as hate speech and propaganda, there has been little work on harm in general. Here, we aim to bridge this gap. We focus on two tasks: (*i*) *detecting harmful memes*, and (*ii*) *identifying the social entities they target*. We further extend a recently released HarMeme dataset, which covered *COVID-19*, with additional memes and a new topic: *US politics*. To solve these tasks, we propose MOMENTA (MultimOdal framework for detecting harmful MemEs aNd Their tArgets), a novel multimodal deep neural network that uses global and local perspectives to detect harmful memes. MOMENTA systematically analyzes the local and the global perspective of the input meme (in both modalities) and relates it to the background context. MOMENTA is interpretable and generalizable, and our experiments show that it outperforms several strong rivaling approaches.

1 Introduction

The growing popularity of social media platforms has given rise to a new form of multimodal entity: the *meme*, which is an image, embedded with a short piece of text. Memes are easily shared and can spread fast on the Internet, especially in social media. They are typically humorous and amusing in nature; however, by using an adroit combination of images and texts in the context of contemporary political and socio-cultural divisions, a seemingly harmless meme can easily become a multimodal source of harm.

(a) Partially harmful meme.

(b) Very harmful meme.

Figure 1: Examples of harmful memes. (a) A meme that is *partially harmful*, but is arguably not so hateful or offensive. (b) A meme, where the image and the text are not harmful when considered in isolation, but are *very harmful* when taken as a whole.

Such harmful memes can be dangerous as they can easily damage the reputation of individuals, renowned celebrities, political entities, companies, or social groups, e.g., minorities. Despite memes being so influential, their multimodal nature and camouflaged semantics makes them very challenging to analyze.

The abundant quantity, fecundity and escalating diversity of online memes has led to a growing body of research on meme analysis, which has focused on tasks such as meme emotion analysis (Sharma et al., 2020; Pramanick et al., 2021a), sarcastic meme detection (Kumar and Garg, 2019), and hateful meme detection (Kiela et al., 2020; Zhou et al., 2021b; Velioglu and Rose, 2020). Research on these problems has shown that off-the-shelf multimodal systems, which often perform well on a range of visual-linguistic tasks, struggle when applied to memes. There are a number of reasons for that. First, memes are context-dependent, and thus focusing only on the image and on the text without background knowledge about the context in which the meme was generated, as well as some background information about people, companies and events, often is not enough to understand it.

* denotes equal contribution

Second, unlike other multimodal tasks, the image and the textual content in the meme are often uncorrelated, and its overall semantics is presented holistically. Finally, real-world memes can be noisy, and the text embedded in them can be hard to extract using standard OCR tools.

The proliferation of virulent memes has stimulated research focusing on their dark sides: hate (Kiela et al., 2020) and offensiveness (Suryawanshi and Chakravarthi, 2021). Recently, Pramanick et al. (2021b) defined the notion of *harmful meme* and demonstrated its dependency on the background context. For example, the meme in Figure 1a is somewhat harmful to Joe Biden in the context of an election, but it is arguably neither hateful nor offensive. Moreover, the notion of *harm* is often apparent only when the two modalities are combined. For example, in Figure 1b, the unimodal cues are not harmful, but the meme as a whole is harmful to Donald Trump. Moreover, identifying the target of harmful memes (e.g., *Joe Biden* and *Donald Trump*) requires separate analysis, which is not prevalent for hateful or offensive memes.

With the above motivation in mind, here we aim to explore the role of background context for detecting harmful memes and for identifying the social entities they target. In particular, we make the following contributions:

- We extend our recently released HarMeme dataset (Pramanick et al., 2021b), which covered *COVID-19*, with aditional examples and a new topic (*US Politics*), thus ending up with two datasets: Harm-C and Harm-P.
- We benchmark the two datasets against ten state-of-the-art unimodal and multimodal models, and we discuss the limitations of these models.
- We propose MOMENTA, a novel multimodal framework that systematically analyzes the local and the global perspective of the input meme and relates it to the background context, with the aim of detecting subtle harmful elements.
- We perform extensive experiments on both datasets, and we show that MOMENTA outperforms the ten baselines in terms of accuracy by 1.3–2.6 points absolute for both tasks.
- Finally, we establish the generalizability and the interpretability of MOMENTA.

2 Related Work

2.1 Harm and Multimodality

Various aspects of harm, such as hate speech, misinformation, and offensiveness, have been studied in isolation. Ahn and Jang (2019) addressed harmfulness in terms of obscenity and violence using multimodal approaches involving video and images. Hirschberg et al. (2005), Kopev et al. (2019), and Dinkov et al. (2019) studied intentional deception and bias using textual and acoustic cues from the speech signal. Gogate et al. (2017) and Baly et al. (2020) designed robust systems for deception detection by combining acoustic, textual, and other information (visual, social). In recent work on detecting offensiveness in memes, Suryawanshi et al. (2020) showed improvements using an early-fusion multimodal approach that combines representations from unimodal models. Critical aspects such as prevalence of racial biases within the datasets and the modeling approaches were addressed in (Mills and Unsworth, 2018; Davidson et al., 2019; Mozafari et al., 2020; Xia et al., 2020; Zhou et al., 2021a); they characterized the biases and proposed de-biasing mechanisms for tasks such as detecting toxic/abusive language and hate speech, as well as for identifying racial prejudices.

Finally, recent research and a shared task focused on propaganda in memes (Dimitrov et al., 2021a,b), but did not target harmfulness per se.

2.2 Harm and Memes

There was a recent shared task on troll meme classification (Suryawanshi and Chakravarthi, 2021), and two tasks on hateful meme detection: (Kiela et al., 2020) and (Zhou et al., 2021b). A number of models have been developed for these tasks. Suryawanshi and Chakravarthi (2021) used a diverse set of models including logistic regression and BERT (Devlin et al., 2019). Muennighoff (2020) used a separate and a combined stream of Transformers (Vaswani et al., 2017). Velioglu and Rose (2020) used a Detectron-based representation to fine-tune Visual BERT (Li et al., 2019), along with data augmentation. Lippe et al. (2020) found UNITER (Chen et al., 2020) to be a very strong choice for multimodal content. Sandulescu (2020) used a multimodal deep ensemble, while examining both single-stream models such as ViLBERT (Lu et al., 2019), VLP (Zhou et al., 2020), and UNITER (Chen et al., 2020), and dual-stream models like LXMERT (Tan and Bansal, 2019).

Wang et al. (2021) proposed a multimodal deep neural network with semantic and task-level attention for detecting medical misinformation. Another shared task, on memotion analysis (Sharma et al., 2020), asked to recognize expressive emotions via sentiment (positive, negative, neutral), type of emotion (sarcastic, funny, offensive, motivation), and their intensity. Recently, Chandra et al. (2021) investigated antisemitism, its subtypes, and its use in memes. However, none of these studies addressed the broader concept of *harmful memes*.

In our previous work (Pramanick et al., 2021b), we defined the notion of *harmful meme*, and we differentiated it from hateful and offensive meme. We further formulated two tasks: (*i*) detecting harmful meme, and (*ii*) identifying the social entities they target. We also created HarMeme, the first large-scale dataset for harmful meme analysis. However, HarMeme contains memes related to only one topic, COVID-19. Here, we extend HarMeme with additional examples and a new topic (*US politics*). We further propose a novel multimodal framework, which systematically analyzes the local and the global perspective of the input meme (in both modalities) and relates it to the background context.

2.3 Multimodal Pretraining

Self-supervised pre-training using crossmodal and multimodal information saw an early reinstation with the work of Frome et al. (2013), where semantic information from vast unannotated textual data was leveraged to classify images in a zero-shot setup. Similarly, Natural Language Processing (NLP) recently saw the emergence of Pattern-Exploiting Learning (Schick and Schütze, 2021), which allows smaller models to outperform much larger ones such as GPT-3 (Brown et al., 2020) when fine-tuned using a very small number of examples in a few-shot learning setup.

There have been also innovations towards better multimodal systems. Ramesh et al. (2021) proposed DALL-E, a simple yet scalable Transformer that autoregressively models the text tokens with the image features as a single stream of data, towards generating images from query texts and established competitive zero-shot performance. Then, Radford et al. (2021) proposed a competitive model, CLIP, pre-trained on 400 million image–text pairs to train a joint multimodal visual-semantic embedding layer. In our experiments below, we compare our framework to CLIP and to variants thereof.

3 Defining *Harmful Meme*

Following Pramanick et al. (2021b), we abridge the definition of *harmful meme* as follows: a *multimodal unit consisting of an image and an embedded text that has the potential to cause harm to an individual, an organization, a community, or society*. Here, *harm* includes mental abuse, defamation, psycho-physiological injury, socio-economic damages, proprietary damage, emotional disturbance, compensated public image, etc.

Offensive and hateful memes are harmful, but not the other way round. Offensive memes typically aim to mock or to bully a social entity, usually by using abusive words. A hateful meme contains derogatory content, influenced by utmost bias towards an entity (e.g., an individual a community, or an organization). The harmful content in a harmful meme is often camouflaged and might require critical judgment to detect. Furthermore, the social entities attacked or targeted by harmful memes can be any individual, organization, or community, as opposed to *hateful* memes, where entities are attacked based on personal attributes.

4 Data

Here, we describe our two datasets, Harm-C and Harm-P, which consist of memes related to COVID-19 and to US politics, respectively.

4.1 Data Collection and Deduplication.

To collect potentially harmful memes, we conducted keyword-based[1] web search on different sources, mainly Google Image. To alleviate potential biases from this search, we intentionally included non-harmful examples using the same keywords. We used an extension[2] of Google Chrome to download the images. We further scraped various publicly available meme pages on Reddit, Facebook, and Instagram. Unlike the Hateful Memes Challenge (Kiela et al., 2020), which offered synthetically generated memes, our datasets contain real-world memes. To remove noise, we maintained strict filtering on the resolution of the meme images and on the readability of the meme's text as part of the collection process. The process of filtering is described in detail in Appendix B.

[1] **Example keywords for Harm-C:** *Wuhan virus memes*, *COVID vaccine memes*, *Work from home memes*; **Example keywords for Harm-P:** *Presidential debate memes*, *Election-2020 vote counting memes*, *Trump not wearing mask memes*.

[2] download-all-images.mobilefirst.me

Dataset	Split	#Memes	Harmfulness			#Memes	Target			
			Very Harmful	Partially Harmful	Harmless		Individual	Organization	Community	Society
Harm-C	Train	3,013	182	882	1,949	1,064	493	66	279	226
	Validation	177	10	51	116	61	29	3	16	13
	Test	354	21	103	230	124	59	7	32	26
	Total	3,544	213	1,036	2,295	1,249	582	75	327	265
Harm-P	Train	3,020	216	1,270	1,534	1,451	797	470	111	73
	Validation	177	17	69	91	85	70	12	2	1
	Test	355	25	148	182	170	96	54	12	8
	Total	3,552	258	1487	1,807	1,706	963	536	125	82

Table 1: Statistics about Harm-C and Harm-P. Harmful memes are also annotated with a target.

To remove duplicate memes, we used two deduplication tools[3] [4] sequentially, and we preserved the memes with the highest resolution from each group of duplicates. The final size of Harm-C (on *COVID-19*) and Harm-P(on *US politics*) is 3,544 and 3,552 memes, respectively. We used the Google Cloud Vision API to extract the textual content of the memes.

4.2 Data Annotation

We followed the annotation procedure from (Pramanick et al., 2021b). In particular, we asked the annotators to label both the presence and the intensity of harm (*harmful* vs. *partially harmful*), as well as its target:

1. **Individual:** A person, usually a celebrity (e.g., a well-known politician, an actor, etc. such as *Donald Trump*, *Greta Thunberg*).

2. **Organization:** A group of people with a particular purpose, such as a business, a government department, a company, etc. Examples include research organizations such as *WHO*, and political organizations such as the *Democratic Party*.

3. **Community:** A social unit with commonalities based on personal, professional, social, cultural, or political attributes such as religious views, country of origin, gender or gender identity, etc.

4. **Society:** When a meme promotes conspiracies or hate crimes, it is considered harmful to society.

We hired a total of 15 annotators: all of them experts in NLP or linguists, 22–40 years old, including 10 male and 5 female. We paid them fairly for their work as per the standard local pay rate.

Before the actual annotation process, we asked all annotators to go through the annotation guidelines. We further conducted several discussion sessions to evaluate whether they could understand what harmful content is and how to differentiate it from non-harmful content. The annotation process went through three stages: (*i*) a dry run, (*ii*) a final annotation, and (*iii*) a consolidation stage.

4.3 Inter-Annotator Agreement and Statistics

The inter-annotator agreement (Cohen's κ) (Bobicev and Sokolova, 2017) for harmfulness/target is 0.683/0.782 on Harm-C, and 0.675/0.790 on Harm-P, respectively. Table 1 shows statistics about the data distribution and the split into train/validation/test sets, and Figure C.3 (in the Appendix) shows statistics about the sources and the labels. Appendix C gives more detail about the target classes, the annotation guidelines, and the annotation process, and Appendix D offers some statistics about the textual content of the memes, including length distribution, and the most frequent words per dataset and per category.

5 MOMENTA: Our Proposed System

Here, we describe our system, MOMENTA for harmful meme detection and target identification. It takes a meme as input, and extracts the embedded text using Google's OCR Vision API[5]. We encode each text–image pair using CLIP (Radford et al., 2021), a pre-trained visual-linguistic model, leveraging its representations to capture the strong invariance and the overall semantics of the meme.

In addition to the CLIP features, we also identify faces and object proposals[6] (Ren et al., 2015), and we extract various attributes (see Figure 3), which define high-level topics or entities, such as *Joe Biden* and the *Republican Party*, in an image.

[3] gitlab.com/opennota/findimagedupes
[4] github.com/arsenetar/dupeguru
[5] cloud.google.com/vision/docs/ocr
[6] Rectangular bounding boxes or regions of interests (ROI) surrounding the faces and the foreground objects.

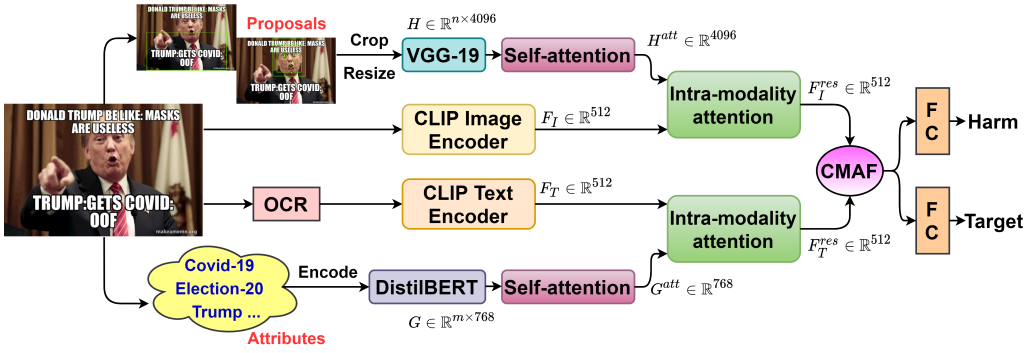


Figure 2: The architecture of our proposed model, MOMENTA.

Then, on the visual side, we encode the ROIs using the pre-trained VGG-19 model (Simonyan and Zisserman, 2015), and on the textual side, we encode the topics/entities using DistilBERT (Sanh et al., 2019). As we mentioned earlier, the analysis of harmful memes is challenging because of their abstruse nature without specific context. Thus, we hypothesize that adding object proposals and attributes would enable the model to understand the high-level concepts in the meme. Our analysis below shows that it indeed captures the appropriate background context reasonably well. Afterwards, we fuse the proposal and the attribute features together with the CLIP representations, separately for the image and for the textual representations, and we add first intramodel attentions, and then a cross-modal attention on top of them, in a hierarchical attention architecture. Finally, we use the resulting multimodal context-aware representation to predict the meme's harmfulness and its target. Figure 2 shows the overall architecture of MOMENTA; below, we explain each of its components in detail.

5.1 CLIP Representations

CLIP (Contrastive Language–Image Pre-training) addresses the generalizability issues of standard computer vision systems (Simonyan and Zisserman, 2015; He et al., 2016), which are often good for some particular tasks, but perform poorly on stress sets and other tasks (Geirhos et al., 2019; Alcorn et al., 2019; Barbu et al., 2019). CLIP is pre-trained using contrastive learning on 400M image–text pairs from the Internet. It offers excellent zero-shot capabilities due to the variety of images it has seen and the natural language supervision. In MOMENTA, given the meme's image I and its OCR-extracted text T, we extract a CLIP image embedding F_I and a CLIP text embedding F_T; both F_I and F_T are 512-dimensional vectors.

5.2 Object Proposal and Attribute Representations

Following previous studies (Wu et al., 2016; Cai et al., 2019) on image captioning and visual question answering, we introduce attributes as high-level image concepts in MOMENTA. Moreover, in addition to meme image attributes, we compute face and foreground object proposals, both of which help to capture subtle harmful contents and appropriate background context of the input meme. Figure 3 shows the detected proposals and image attributes for two example memes. For the first meme (shown on Figure 3a), attributes such as *Christopher Nolan* and *Interstellar* capture the proper context, while for the second meme (shown on Figure 3b), the detected face of Joe Biden perceives minute harmful content.

We use three separate branches of the Google Cloud Vision API to detect faces,[7] foreground objects,[8] and various image attributes.[9] Assume that given an input meme image I, the face and the object bounding boxes are $\{bb_1, bb_2, \ldots, bb_n\}$, and the attributes are $\{att_1, att_2, \ldots, att_m\}$. Each bounding box is cropped, reshaped and fed into VGG-19, which encodes it into a 4,096-dimensional representation. Next, we represent the encoded face and the detected object proposals as $H = \{h_1, h_2, \ldots, h_n\}$, where $H \in \mathbb{R}^{n \times 4096}$. Similarly, each detected attribute is encoded and fed into DistilBERT to generate a 768-dimensional representation. We represent these attributes as $G = \{g_1, g_2, \ldots, g_m\}$, where $G \in \mathbb{R}^{m \times 768}$. Note that the number of detected entities can vary.

[7] cloud.google.com/vision/docs/detecting-faces

[8] cloud.google.com/vision/docs/object-localizer

[9] cloud.google.com/vision/docs/detecting-web

Attributes: {Christopher Nolan, Interstellar, work from home, humor}

(a) Detected faces, foreground objects and image attributes for a *harmless* meme from the Harm-C dataset.

Attributes: {US election, Joe Biden, political campaigning, Macy's}

(b) Detected faces, foreground objects and image attributes for a *very harmful* meme from the Harm-P dataset.

Figure 3: Detected proposals and attributes for two different memes from Harm-C and Harm-P datasets.

5.3 Intra-Modality Attention

Next, we use self-attention over n object proposals and m image attributes to emphasize the most relevant ones for the target meme. The resulting self-attended respresentations are H^{att} and G^{att}, respectively, where $H^{att} \in \mathbb{R}^{4096}$ and $H^{att} \in \mathbb{R}^{768}$.

$$H^{att} = \mathbf{W_H} \otimes H; \quad G^{att} = \mathbf{W_G} \otimes G \qquad (1)$$

where, $W_H \in \mathbb{R}^{1\times n}$ and $W_G \in \mathbb{R}^{1\times m}$ are learnable parameters, and $\otimes$ is a matrix outer product.

Subsequently, we fuse the self-attended object proposals with the CLIP image features in an intra-modality attention module. This stage aims to combine the local image descriptions with the global semantics of the meme. Similarly, we fuse the self-attended image attributes with the CLIP text features. Overall, the local and the global features capture the semantics of the meme considering the background context. Prior to the cross-modal attention fusion (CMAF) block, the proposal and the attribute features are projected to similar dimensions using a dense layer.

$$F_I^{res} = \mathbf{W_I} \otimes [F_I, Dense(H^{att})] \qquad (2)$$

$$F_T^{att} = \mathbf{W_T} \otimes [F_T, Dense(G^{att})] \qquad (3)$$

Finally we feed the resulting image and text features, $F_I^{res}, F_T^{att} \in \mathbb{R}^{512}$ into CMAF to obtain the final multimodal meme representation.

5.4 Cross-Modality Attention Fusion

For some memes, the text modality is more relevant, while for others, the image plays a crucial role. CMAF uses an attention mechanism to fuse the representations from the textual and the visual modalities. Motivated by (Gu et al., 2018), we design our CMAF module with two major parts: modality attention generation and weighted feature concatenation. In the first part, we use a sequence of dense layers followed by a softmax layer to generate the attention scores $[a_v, a_t]$ for the two modalities.

In the second part, we weigh the original unimodal features using their respective attention scores and we concatenate them together. We also use residual connections for better gradient flow.

$$F_{Meme}^V = (1 + a_v)F_I^{res} \qquad (4)$$

$$F_{Meme}^T = (1 + a_t)F_T^{res} \qquad (5)$$

$$F_{Meme} = \mathbf{W_F} \otimes [F_{Meme}^V, F_{Meme}^T] \qquad (6)$$

where, $W_F \in \mathbb{R}^2$ is a learnable parameter, and $F_{Meme} \in \mathbb{R}^{512}$ is the final representation.

5.5 Prediction and Training Objective

We feed the final multimodal meme representation F_{Meme} into two parallel fully-connected branches for the final classification: one branch per task.

As there is class imbalance (shown in Table 1 for each dataset and task), we use focal loss (Lin et al., 2020), which down-weighs the easy examples and focuses training on the hard ones.

Finally, we train MOMENTA in a multi-task learning setup, where the loss for to target identification is considered only if the meme is *partially harmful* or *very harmful*. This is because we should not be looking for a target if the meme is not harmful.

6 Experiments

We train MOMENTA and all baselines using Pytorch on NVIDIA Tesla V100 GPU, with 32 GB dedicated memory, with CUDA-11.2 and cuDNN-8.1.1 installed. The hyper-parameter values for all models are given in Appendix A.

We experiment with Harm-C and Harm-P using a variety of state-of-the-art unimodal textual models, unimodal visual models, and multimodal models that were pre-trained on both modalities. We use three measures for evaluation: Accuracy, Macro-F1, and Macro-Averaged Mean Absolute Error (MMAE) (Baccianella et al., 2009). For the first two, higher values are better, while for MMAE, lower values are better.

Modality	Model	Harmful Meme Detection on Harm-C						Harmful Meme Detection on Harm-P					
		2-Class Classification			3-Class Classification			2-Class Classification			3-Class Classification		
		Acc ↑	F1 ↑	MMAE ↓	Acc ↑	F1 ↑	MMAE ↓	Acc ↑	F1 ↑	MMAE ↓	Acc ↑	F1 ↑	MMAE ↓
	Human†	90.68	83.55	0.1723	86.10	65.10	0.4857	94.40	88.47	0.1028	92.12	70.35	0.6274
	Majority	64.76	39.30	0.5000	64.76	26.20	1.0000	51.27	33.39	0.5000	51.27	22.59	1.0000
Text (T) Only	TextBERT	70.17	66.25	0.2911	68.93	48.72	0.5591	80.12	78.35	0.1660	74.55	54.08	0.7742
Image (I) Only	VGG19	68.12	61.86	0.3190	66.24	41.76	0.6487	70.65	70.46	0.1887	73.65	51.89	0.8466
	DenseNet-161	68.42	62.54	0.3125	65.21	42.15	0.6326	74.05	73.68	0.1845	71.80	50.98	0.8388
	ResNet-152	68.74	62.97	0.3114	65.29	43.02	0.6264	73.14	72.77	0.1800	71.02	50.64	0.8900
	ResNeXt-101	69.79	63.68	0.3029	66.55	43.68	0.6499	73.91	73.57	0.1812	71.84	51.45	0.8422
I + T (Unimodal Pre-training)	Late Fusion	73.24	70.25	0.2927	66.67	45.06	0.6077	78.26	78.50	0.1674	76.20	55.84	0.7245
	Concat BERT	71.82	71.82	0.3156	65.54	43.37	0.5976	77.25	76.38	0.1743	76.04	55.95	0.7450
	MMBT	73.48	67.12	0.3258	68.08	50.88	0.6474	82.54	80.23	0.1413	78.14	58.03	0.7008
I + T (Multimodal Pre-training)	ViLBERT CC	78.53	78.06	0.1881	**75.71**	48.82	0.5329	**87.25**	**86.03**	**0.1276**	**84.66**	**64.70**	**0.6982**
	V-BERT COCO	**81.36**	**80.13**	**0.1857**	74.01	**53.85**	**0.5303**	86.80	86.07	0.1318	84.02	63.68	0.7020
Proposed System and Variants	CLIP	74.23	73.85	0.2955	67.04	44.25	0.6228	80.55	80.25	0.1659	77.00	56.85	0.7852
	CLIP + Proposals	77.65	76.90	0.2142	70.52	45.60	0.5955	84.16	83.80	0.1556	81.06	60.65	0.7122
	CLIP + Attributes	78.10	77.64	0.2010	71.05	45.55	0.5887	84.02	83.85	0.1508	80.75	60.23	0.7058
	MOMENTA w/o CMAF	80.75	80.17	0.1896	74.85	51.25	0.5360	86.20	85.55	0.1355	83.85	63.02	0.6990
	MOMENTA	**83.82**	**82.80**	**0.1743**	**77.10**	**54.74**	**0.5132**	**89.84**	**88.26**	**0.1314**	**87.14**	**66.66**	**0.6805**
$\Delta_{\text{MOMENTA}-best\_model}$		2.46	2.67	0.0114	1.39	0.89	0.0171	2.59	2.23	0.0038	2.48	1.96	0.0177

Table 2: Performance on the two tasks. For two-class, we merge *very harmful* and *partially harmful*. †This row shows the human performance on test, and the last row shows the improvement of MOMENTA over the best baseline.

6.1 Baselines

6.1.1 Unimodal Models

▷ **Text BERT:** We use BERT (Devlin et al., 2019) as our unimodal text-only model.

▷ **VGG19, DenseNet, ResNet, ResNeXt:** For the unimodal visual-only models, we use four well-known models: VGG19 (Simonyan and Zisserman, 2015), DenseNet-161 (Huang et al., 2017), ResNet-152 (He et al., 2016), and ResNeXt-101 (Xie et al., 2017), pre-trained on ImageNet (Deng et al., 2009).

6.1.2 Multimodal Models

▷ **Late fusion:** This model uses the average prediction scores of ResNet-152 and BERT.

▷ **Concat BERT:** This model concatenates the representations from ResNet-152 and BERT, and uses a perceptron as a classifier on top of them.

▷ **MMBT:** This is a Multimodal Bitransformer (Kiela et al., 2019), capturing the intra-modal and the inter-modal dynamics of the two modalities.

▷ **ViLBERT CC:** Vision and Language BERT (Lu et al., 2019), trained on an intermediate multimodal objective (conceptual captions) (Sharma et al., 2018), is a strong model with task-agnostic joint representation of image and text.

▷ **Visual BERT COCO:** This is Visual BERT (Li et al., 2019), pre-trained on the COCO dataset (Lin et al., 2014), another strong multimodal model.

7 Experimental Results

We compare MOMENTA to unimodal textual models, unimodal visual models, and multimodal models pre-trained on both modalities.

Except for the unimodal visual models, we use the MMF framework.[10] We further explore the generalizability and the interpretability of MOMENTA.

7.1 Harmful Meme Detection

Table 2 shows the results for harmful meme detection. We start by merging the *partially harmful* and the *very harmful* classes, thus ending up with binary classification. In both datasets, *harmless* is the majority class; the majority class baseline yields accuracy of 64.76 on Harm-C and of 51.27 on Harm-P. Among the unimodal models, those using the textual modality perform better. In case of Harm-C, the accuracy of the unimodal models is 68.1–70.2, while on Harm-P, it is 70.7–80.1.

We also see that multimodal models outperform unimodal ones, and more sophisticated fusion techniques perform better. For example, late fusion, the simplest one, performs only slightly better than unimodal models, while MMBT, yields 2.5–3.3 absolute points of improvement. We also notice the effectiveness of multimodal pre-training. On Harm-C, Visual BERT COCO outperforms all other models, while on Harm-P, VilBERT CC is the best. Thus, below we will compare the performance of MOMENTA to these two models.

In the binary case, MOMENTA achieves 2.46 absolute points of improvement on Harm-C, and 2.59 points on Harm-P over the best models. The corresponding Macro-F1 scores also improve by a similar margin. We show in Section 7.4 that all modules in MOMENTA contribute to this.

[10] github.com/facebookresearch/mmf

Modality	Model	Target on Harm-C			Target on Harm-P		
		Acc ↑	F1 ↑	MMAE ↓	Acc ↑	F1 ↑	MMAE ↓
	Human†	87.55	82.01	0.3647	90.58	72.68	0.6324
	Majority	46.60	15.89	1.5000	56.47	18.05	1.5000
Text (T) only	TextBERT	69.35	55.60	0.8988	72.54	60.36	0.8895
Image (I) only	VGG19	63.48	53.60	1.0549	68.24	55.24	1.0225
	DenseNet-161	64.52	53.51	1.0065	69.40	57.95	0.9540
	ResNet-152	65.75	53.78	1.0459	68.75	57.00	0.9667
	ResNeXt-101	65.82	53.95	0.9277	70.22	59.67	0.9245
I + T (Unimodal Pretraining)	Late Fusion	72.58	58.43	0.6318	73.25	64.28	0.8541
	Concat BERT	67.74	49.77	0.8879	72.46	60.87	0.8655
	MMBT	72.58	58.35	0.6318	74.65	65.12	0.8441
I + T (Multimodal Pretraining)	ViLBERT CC	72.58	57.17	0.8035	77.25	**67.39**	**0.8410**
	V-BERT COCO	**75.81**	**65.77**	**0.5036**	**77.28**	66.90	0.8536
Proposed System and Variants	CLIP	72.47	62.14	0.6312	72.40	65.66	0.8557
	CLIP + Proposals	74.85	64.38	0.5746	75.85	66.13	0.8482
	CLIP + Attributes	74.56	61.38	0.6015	76.20	66.34	0.8491
	MOMENTA w/o CMAF	76.16	64.80	0.5422	77.54	67.25	0.8430
	MOMENTA	**77.95**	**69.65**	**0.4225**	**78.54**	**68.83**	**0.8295**
$\Delta_{\text{MOMENTA}-best\_model}$		2.14	3.88	0.0811	1.26	1.44	0.0115

Table 3: Performance for target identification. †This row shows the human performance on the test set.

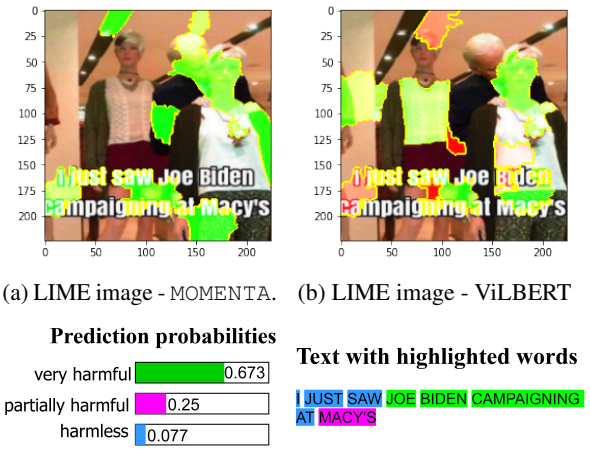
For the 3-class harmful meme detection, we see a similar trend: early-fusion models with multimodal pre-training (ViLBERT CC, V-BERT COCO) outperform unimodal and simple multimodal ones. Moreover, MOMENTA achieves an improvement of 1.39 and 2.48 points absolute over the corresponding best models on Harm-C and Harm-P.

7.2 Target Identification

Table 3 shows the performance for the 4-class target identification. Here, Harm-P is more imbalanced than Harm-C. The majority class baseline yields accuracy of 46.60 on Harm-C and of 56.47 on Harm-P. Similarly to the earlier task, unimodal models perform poorly, achieving 63.4–69.3 accuracy on Harm-C, and 68.2–72.5 on Harm-P. Adding multimodal cues with multimodal pre-training yields sizable improvements. Visual BERT COCO is the best on Harm-C, and ViLBERT CC is the best on Harm-P. However, MOMENTA outperforms the best models by 2.14 points absolute in terms of accuracy and by 3.88 points in terms of F1 score on Harm-C, and by 1.26 points of accuracy and 1.44 points of F1 on Harm-P.

7.3 Human Evaluation

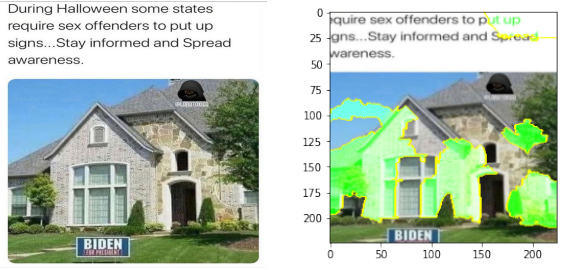
In order to understand how humans perceive these tasks compared to neural systems, we performed an extensive manual evaluation. We hired three linguists to label the test set. The results show that for all three experiments and on both datasets, MOMENTA is the best performing model; yet, it is 4.5–12 absolute points of accuracy behind human performance. The difference is more sizable for target identification, indicating the difficulty of that task and suggesting that there is a lot of room for further improvement.



(a) LIME image - MOMENTA. (b) LIME image - ViLBERT

(c) LIME text - MOMENTA.

Figure 4: Visualization of explanation as generated by LIME on both modalities for MOMENTA and ViLBERT.



(a) Misclassified meme. (b) LIME image - MOMENTA.

Figure 5: Misclassified example and explanation.

7.4 Ablation Study

Here, we present an ablation study, analyzing the contribution of each of the components of MOMENTA: proposal, attribute detection blocks, and CMAF module. The last five rows of Tables 2 and 3 show these results.

We can see in the tables that without the proposals and the attributes, CLIP performs similarly to MMBT. Then, adding the proposals improves accuracy by 2.3–4.1 points absolute for all tasks and datasets. Incorporating attributes in CLIP also improves results by a similar margin. When both are added, the accuracy improves further, outperforming CLIP by 5.5–10 points absolute in terms of accuracy and by 3.2–10.5 points absolute in terms of F1 score. The CMAF module also plays a pivotal role: we notice a drop of 1–3.8 points absolute in terms of accuracy, and 1.5–4.8 points absolute in terms of F1 score when CMAF is replaced by simple concatenation. Hence, we can conclude that each block of MOMENTA's architecture helps to boost the overall performance.

		Harm-C			Harm-P			Combined		
		H-2†	H-3‡	Tar⋆	H-2†	H-3‡	Tar⋆	H-2†	H-3‡	Tar⋆
Harm-C	ViLBERT	78.06	48.82	57.17	74.20	51.39	54.10	74.85	44.15	46.52
	V-BERT	80.13	53.85	68.77	74.56	52.87	53.46	75.04	45.20	47.66
	MOMENTA	**82.80**	**54.74**	**69.65**	80.25	61.87	58.39	81.66	49.83	50.12
Harm-P	ViLBERT	71.28	42.57	48.20	86.03	64.70	67.39	75.88	44.18	45.82
	V-BERT	72.58	45.10	54.07	86.07	63.68	66.90	76.20	45.69	47.38
	MOMENTA	76.30	50.46	58.33	**88.26**	**66.66**	**68.83**	80.75	49.70	50.28
Combined	ViLBERT	73.48	43.11	51.45	76.92	56.50	60.20	79.20	53.65	58.12
	V-BERT	74.88	46.28	60.82	76.85	56.07	58.22	80.45	53.98	58.76
	MOMENTA	79.50	51.07	62.56	81.09	62.85	61.87	**85.20**	**58.44**	**61.20**

Table 4: Transferability of the two best-performing baselines and MOMENTA on Harm-C, on Harm-P, and on the combination thereof. The models are trained on the dataset in the row and tested on the one in the column. All scores are Macro F1. H-2† is 2-class, and H-3‡ is 3-class harmful meme detection, and Tar⋆ is target identification. The blue color indicates the best transferable results for each task–dataset combination.

7.5 Transferability of MOMENTA

Table 4 shows the transferability of MOMENTA on Harm-C, on Harm-P, and on the combination thereof, compared to ViLBERT and Visual BERT. We can see that when training and testing on the same dataset, all models yield high F1 scores. However, when trained on one dataset and tested on a different one, MOMENTA yields 2.2–6.6, 1.1–9.0, and 0.9–4.2 points of absolute improvements in terms of F1 score for 2-class and 3-class harmful meme detection and for target identification, respectively. CLIP, which was pre-trained on 400M image–text pairs, contributes to the superior transferability of MOMENTA.

7.6 Side-by-Side Diagnostics

We visualize the explainability of MOMENTA and we compare it to ViLBERT using LIME (Ribeiro et al., 2016). We take the example from Figure 3b for our analysis. MOMENTA correctly classified it as *very harmful* with a dominant probability of 0.673, but ViLBERT fails. Figures 4a and 4b highlight the most important super-pixels contributing to the decision of MOMENTA and ViLBERT, respectively. We notice that the face of Joe Biden and the mannequin, which are presented in a very insulting way in this meme, contribute heavily to the prediction of MOMENTA. However, as Biden's face is partially occluded, ViLBERT cannot recognize this harmful gesture. The fine-grained face detection and the robust CLIP encoder help MOMENTA to identify this subtle harmful element. Figure 4c shows the contribution of different words in the meme's text to the prediction of MOMENTA. Overall, the word *CAMPAIGNING* and the conflicting gesture in the image make the meme *very harmful*.

7.7 Error Analysis

Figure 5a shows a *very harmful* meme on which MOMENTA fails: the image contains no harmful gestures, and the text has no harmful words. The most contributing super pixels are also spread randomly, as shown in Figure 5b. Moreover, the detected attributes, {*palace*, *summer*, *mansion*}, do not model context well. Yet, the entrenched semantics of the entire meme makes it *very harmful*.

8 Conclusion and Future Work

We introduced two large-scale datasets, Harm-C and Harm-P, for detecting harmful memes and their targets. We further proposed MOMENTA, a novel multimodal deep neural network that systematically analyzes the local and the global perspective of the input meme (in both modalities) and relates it to the background context. Extensive experiments on the two datasets showed the efficacy of MOMENTA, which outperforms ten baselines for both tasks. We further demonstrated its transferability and interpretability.

In future work, we plan to extend the datasets with more domains and languages.

Acknowledgments

The work was partially supported by a Wipro research grant, the Infosys Centre for AI, IIIT Delhi, India, 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.

Ethics and Broader Impact

Reproducibility. We present detailed hyper-parameter configurations in Table A.1 and Appendix A. The source code, and the datasets (Harm-Cand Harm-P) are available at http://github.com/LCS2-IIITD/MOMENTA

User Privacy. Our datasets only include memes, and they do not contain any user information. All the memes in our datasets were collected from publicly available web pages and there are no known copyright issues regarding them. The sources are listed in Section 4 and Figure C.2. Note that we also release links to the memes instead of the actual memes. In this way, we ensure that if a user deletes a posted meme, that meme would not be available in our datasets anymore. The same strategy was previously used by several researchers to distribute collections of tweets.

Annotation. The annotation was conducted by the experts working in NLP or linguists in India. We treated the annotators fairly and with respect. They were paid as per the standard local paying rate. Before beginning the annotation process, we requested every annotator to thoroughly go through the annotation guidelines. We further conducted several discussion sessions to make sure all annotators could understand well what harmful content is and how to differentiate it from humorous, satirical, hateful, and non-harmful content.

Biases. Any biases found in the dataset are unintentional, and we do not intend to cause harm to any group or individual. We note that determining whether a meme is harmful can be subjective, and thus it is inevitable that there would be biases in our gold-labeled data or in the label distribution. We address these concerns by collecting examples using general keywords about COVID-19, and also by following a well-defined schema, which sets explicit definitions during annotation. Our high inter-annotator agreement makes us confident that the labeling of the data is correct most of the time.

Misuse Potential. We ask researchers to be aware that our dataset can be maliciously used to unfairly moderate memes based on biases that may or may not be related to demographics and other information within the text. Intervention with human moderation would be required in order to ensure that this does not occur.

Intended Use. We release our dataset aiming to encourage research in studying harmful 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 in the appropriate manner.

Environmental Impact. Finally, we would also like to warn that the use of large-scale Transformers requires a lot of computations and the use of GPUs/TPUs for training, which contributes to global warming (Strubell et al., 2019). This is a bit less of an issue in our case, as we do not train such models from scratch; rather, we fine-tune them on relatively small datasets. Moreover, running on a CPU for inference, once the model has been fine-tuned, is perfectly feasible, and CPUs contribute much less to global warming.

References

Byeongtae Ahn and Seok-Woo Jang. 2019. Multimodal approach for multimedia injurious contents blocking. *Multimedia Tools and Applications*, 79:16459–16472.

Michael A Alcorn, Qi Li, Zhitao Gong, Chengfei Wang, Long Mai, Wei-Shinn Ku, and Anh Nguyen. 2019. Strike (with) a pose: Neural networks are easily fooled by strange poses of familiar objects. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '19, pages 4845–4854, Long Beach, California, USA.

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2009. Evaluation measures for ordinal regression. In *Proceedings of the Ninth International Conference on Intelligent Systems Design and Applications*, ISDA '09, pages 283–287, Pisa, Italy.

Ramy Baly, Georgi Karadzhov, Jisun An, Haewoon Kwak, Yoan Dinkov, Ahmed Ali, James Glass, and Preslav Nakov. 2020. What was written vs. who read it: News media profiling using text analysis and social media context. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, ACL '20, pages 3364–3374.

Andrei Barbu, David Mayo, Julian Alverio, William Luo, Christopher Wang, Dan Gutfreund, Josh Tenenbaum, and Boris Katz. 2019. ObjectNet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, volume 32 of *NeurIPS '19*, pages 9453–9463, Vancouver, Canada.

Victoria Bobicev and Marina Sokolova. 2017. Inter-annotator agreement in sentiment analysis: Machine learning perspective. In *Proceedings of the International Conference Recent Advances in Natural*

Language Processing, RANLP '17, pages 97–102, Varna, Bulgaria.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, volume 33 of *NeurIPS '20*, pages 1877–1901.

Yitao Cai, Huiyu Cai, and Xiaojun Wan. 2019. Multi-modal sarcasm detection in Twitter with hierarchical fusion model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, ACL '19, pages 2506–2515, Florence, Italy.

Mohit Chandra, Dheeraj Pailla, Himanshu Bhatia, Aadilmehdi Sanchawala, Manish Gupta, Manish Shrivastava, and Ponnurangam Kumaraguru. 2021. "Subverting the Jewtocracy": Online antisemitism detection using multimodal deep learning. In *Proceedings of the 13th ACM Web Science Conference*, WebSci '21, pages 148—157.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. UNITER: Universal image-text representation learning. In *Proceedings of the European Conference on Computer Vision*, ECCV '20, pages 104–120. Springer.

Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. 2019. Racial bias in hate speech and abusive language detection datasets. In *Proceedings of the Third Workshop on Abusive Language Online*, ALW '19, pages 25–35, Florence, Italy.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '09, pages 248–255, Miami, Florida, USA.

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. 2021a. 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.

Dimitar Dimitrov, Bishr Bin Ali, Shaden Shaar, Firoj Alam, Fabrizio Silvestri, Hamed Firooz, Preslav Nakov, and Giovanni Da San Martino. 2021b. SemEval-2021 Task 6: Detection of persuasion techniques in texts and images. In *Proceedings of the International Workshop on Semantic Evaluation*, SemEval '21, pages 70–98.

Yoan Dinkov, Ahmed Ali, Ivan Koychev, and Preslav Nakov. 2019. Predicting the leading political ideology of Youtube channels using acoustic, textual and metadata information. In *Proceedings of the 20th Annual Conference of the International Speech Communication Association*, INTERSPEECH '19, pages 501–505, Graz, Austria.

Andrea Frome, Greg S. Corrado, Jonathon Shlens, Samy Bengio, Jeffrey Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. 2013. DeViSE: A deep visual-semantic embedding model. In *Proceedings of the 26th International Conference on Neural Information Processing Systems*, NIPS '13, pages 2121–2129, Lake Tahoe, Nevada, USA.

Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. 2019. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *Proceedings of the 7th International Conference on Learning Representations*, ICLR '19, New Orleans, Louisiana, USA.

Mandar Gogate, Ahsan Adeel, and Amir Hussain. 2017. Deep learning driven multimodal fusion for automated deception detection. In *Proceedings of the 2017 IEEE Symposium Series on Computational Intelligence*, SSCI '17, pages 1–6, Honolulu, Hawaii, USA.

Yue Gu, Kangning Yang, Shiyu Fu, Shuhong Chen, Xinyu Li, and Ivan Marsic. 2018. Hybrid attention based multimodal network for spoken language classification. In *Proceedings of the 27th International Conference on Computational Linguistics*, COLING '18, pages 2379–2390, Santa Fe, New Mexico, USA.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '16, pages 770–778, Las Vegas, Nevada, USA.

Julia Hirschberg, Stefan Benus, Jason Brenier, Frank Enos, Sarah Hoffman, Sarah Gilman, Cynthia Girand, Martin Graciarena, Andreas Kathol, Laura

Michaelis, Bryan Pellom, Elizabeth Shriberg, and Andreas Stolcke. 2005. Distinguishing deceptive from non-deceptive speech. In *Proceedings of the 9th European Conference on Speech Communication and Technology*, INTERSPEECH '05, pages 1833–1836, Lisbon, Portugal.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. 2017. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '17, pages 4700–4708, Honolulu, Hawaii, USA.

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*.

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.

Daniel Kopev, Ahmed Ali, Ivan Koychev, and Preslav Nakov. 2019. Detecting deception in political debates using acoustic and textual features. In *Proceedings of the 2019 IEEE Automatic Speech Recognition and Understanding Workshop*, ASRU '19, pages 652–659, Singapore.

Akshi Kumar and Geetanjali Garg. 2019. Sarc-M: Sarcasm detection in typo-graphic memes. In *Proceedings of the International Conference on Advances in Engineering Science Management & Technology*, ICAESMT '19, Dehradun, India.

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, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2020. Focal loss for dense object detection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 42(2):318–327.

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.

Kathy A. Mills and Len Unsworth. 2018. The multimodal construction of race: a review of critical race theory research. *Language and Education*, 32(4):313–332.

Marzieh Mozafari, Reza Farahbakhsh, and Noël Crespi. 2020. Hate speech detection and racial bias mitigation in social media based on BERT model. *PLOS ONE*, 15(8):1–26.

Niklas Muennighoff. 2020. Vilio: State-of-the-art visio-linguistic models applied to hateful memes. *arXiv:2012.07788*.

Shraman Pramanick, Md Shad Akhtar, and Tanmoy Chakraborty. 2021a. Exercise? I thought you said 'extra fries': Leveraging sentence demarcations and multi-hop attention for meme affect analysis. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 15 of *ICWSM '21*, pages 513–524.

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

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.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *ICML '21*, pages 8821–8831.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Proceedings of the 28th International Conference on Neural Information Processing Systems*, NIPS '15, pages 91–99, Montreal, Canada.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 1135–1144.

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

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv:1910.01108*.

Timo Schick and Hinrich Schütze. 2021. It's not just size that matters: Small language models are also few-shot learners. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL-HLT '21, pages 2339–2352.

Chhavi Sharma, Deepesh Bhageria, William Scott, Srinivas PYKL, Amitava Das, Tanmoy Chakraborty, Viswanath Pulabaigari, and Björn Gambäck. 2020. SemEval-2020 task 8: Memotion analysis- the visuo-lingual metaphor! In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, SemEval '20, pages 759–773.

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.

Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. In *Proceedings of the 3rd 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 and Bharathi Raja Chakravarthi. 2021. Findings of the shared task on troll meme classification in Tamil. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 126–132.

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*, TRAC '20, pages 32–41.

Hao Tan and Mohit Bansal. 2019. LXMERT: Learning cross-modality encoder representations from transformers. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, EMNLP-IJCNLP '19, pages 5100–5111, Hong Kong, China.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the Annual Conference on Neural Information Processing Systems*, NeurIPS '17, pages 5998–6008, Long Beach, California, USA.

Riza Velioglu and Jewgeni Rose. 2020. Detecting hate speech in memes using multimodal deep learning approaches: Prize-winning solution to hateful memes challenge. *arXiv:2012.12975*.

Zuhui Wang, Zhaozheng Yin, and Young Anna Argyris. 2021. Detecting medical misinformation on social media using multimodal deep learning. *IEEE J. Biomed. Health Informatics*, 25(6):2193–2203.

Qi Wu, Chunhua Shen, Lingqiao Liu, Anthony Dick, and Anton Van Den Hengel. 2016. What value do explicit high level concepts have in vision to language problems? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '16, pages 203–212, Las Vegas, Nevada, USA.

Mengzhou Xia, Anjalie Field, and Yulia Tsvetkov. 2020. Demoting racial bias in hate speech detection. In *Proceedings of the Eighth International Workshop on Natural Language Processing for Social Media*, SocialNLP '20, pages 7–14.

Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. 2017. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '17, pages 1492–1500, Honolulu, Hawaii, USA.

Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason Corso, and Jianfeng Gao. 2020. Unified vision-language pre-training for image captioning and VQA. In *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence*, AAAI '20, pages 13041–13049.

Xuhui Zhou, Maarten Sap, Swabha Swayamdipta, Yejin Choi, and Noah Smith. 2021a. Challenges in automated debiasing for toxic language detection. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '21, pages 3143–3155.

Yi Zhou, Zhenhao Chen, and Huiyuan Yang. 2021b. Multimodal learning for hateful memes detection. In *Proceedings of the 2021 IEEE International Conference on Multimedia & Expo Workshops*, ICMEW '21, pages 1–6.

Appendix

A Implementation Details and Hyperparameter Values

We train all the models using Pytorch on an NVIDIA Tesla V100 GPU, with 32 GB dedicated memory, CUDA-11.2 and cuDNN-8.1.1 installed. For the unimodal models, we import all the pre-trained weights from the TORCHVISION.MODELS[11] subpackage of the PyTorch framework. We initialize the remaining weights randomly using a zero-mean Gaussian distribution with a standard deviation of 0.02. Statistics about the dataset are shown in Table 1, where we can see that there is label imbalance both for harmfulness intensity ([*Very Harmful*, *Partially Harmful*] vs. *Harmless*) and for target classification ([*Individual*, *Organization*, *Community*] vs. *Entire Society*). We address this using focal loss (FL) (Lin et al., 2020), which down-weighs the easy examples and focuses training on the hard ones. We train MOMENTA in a multi-task learning setup, where the loss due to target identification is considered only if the meme is *partially harmful* or *very harmful*.

We train all models we experiment with using the Adam optimizer (Kingma and Ba, 2015) and a negative log-likelihood loss (NLL) as the objective function. Table A.1 gives more detail about the hyper-parameters we used for training.

B Data Filtering

We apply the following fine-grained filtering criteria during data collection and annotation for the examples we include in our datasets, Harm-C and Harm-P:

1. The meme text must be in English; no other languages and no code-switching are allowed.
2. The meme text must be readable. Thus, we exclude blurry text, incomplete text, etc.
3. The meme must contain no cartoons (as they are often very hard to interpret by AI systems).
4. The meme must be multimodal, i.e., it should contain both an image and soem text.

Figure B.1 shows some example memes that we rejected during the filtering process due to them failing to satisfy some of the above criteria.

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

C Annotation Guidelines

C.1 Defining *harmful* memes

The harm can be expressed in an obvious manner such as abusing, offending, disrespecting, insulting, demeaning, or disregarding a target 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.

Harmful memes can target a social entity (e.g., an individual, an organization, a community) and can aim at calumny/vilification/defamation based on their background (bias, social background, educational background, etc.). The harm can be in the form of mental abuse, psycho-physiological injury, proprietary damage, emotional disturbance, or public image damage. A harmful meme typically attacks celebrities or well-known organizations.

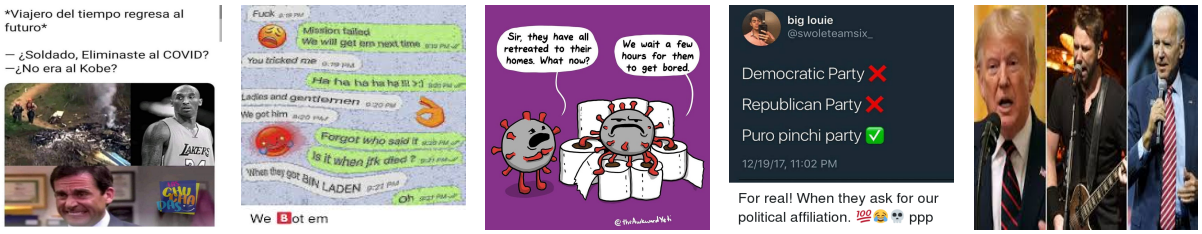
C.2 The target categories

Here are the categories for the targeted entities:

1. **Individual:** A person, usually a celebrity, e.g., a well-known politician, an actor, an artist, a scientist, an environmentalist, etc., such as *Donald Trump, Joe Biden, Vladimir Putin, Hillary Clinton, Barack Obama, Chuck Norris, Greta Thunberg*, and *Michelle Obama*.
2. **Organization:** A group of people with a particular purpose, such as a business, a government department, a company, an institution, or an association, e.g., *Facebook*, *WTO*, and the *Democratic party*.
3. **Community:** A social unit with commonalities based on personal, professional, social, cultural, or political attributes such as religious views, country of origin, gender identity, etc. Communities may share a sense of place situated in a given geographical area (e.g., a country, a village, a town, or a neighborhood) or in virtual space through communication platforms (e.g., online fora based on religion, country of origin, gender, etc.).
4. **Society:** Society as a whole. When a meme promotes conspiracies or hate crimes, it becomes harmful to the general public, i.e., to the entire society.

		Batch-size	Epochs	Learning Rate	Image Encoder	Text Encoder	#Param
Unimodal	TextBERT	16	20	0.001	-	Bert-base-uncased	110M
	VGG19	64	200	0.01	VGG19	-	138M
	DenseNet-161	32	200	0.01	DenseNet-161	-	28M
	ResNet-152	32	300	0.01	ResNet-152	-	60M
	ResNeXt-101	32	300	0.01	ResNeXt-101	-	83M
Multimodal	Late Fusion	16	20	0.0001	ResNet-152	Bert-base-uncased	170M
	Concat BERT	16	20	0.001	ResNet-152	Bert-base-uncased	170M
	MMBT	16	20	0.001	ResNet-152	Bert-base-uncased	169M
	ViLBERT CC	16	10	0.001	Faster RCNN	Bert-base-uncased	112M
	V-BERT COCO	16	10	0.001	Faster RCNN	Bert-base-uncased	247M
	MOMENTA	64	50	0.001	VGG19	DistilBERT-base-uncased	358M

Table A.1: Values of the hyperparameters for the models we experimented with.

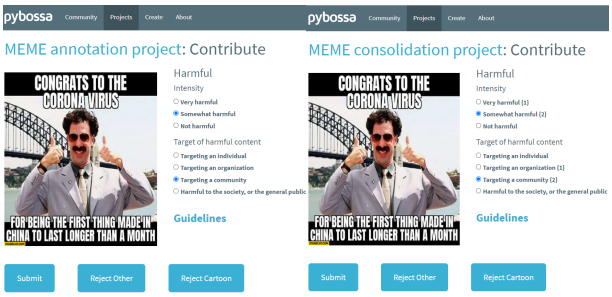


(a) (b) (c) (d) (e)

Figure B.1: Examples of memes that we filtered out as part of the annotation process and the reason for that filtering: (a) not in English, (b) blurry text, (c) cartoon, (d) text-only, and (e) image-only.

C.3 Characteristics of *harmful* memes

- Harmful memes may or may not be offensive, hateful or biased in nature.
- Harmful memes point out vices, allegations, and other negative aspects of an entity based on verified or unfounded claims or mocks.
- Harmful memes leave an open-ended connotation to the word *community*, including anti-social communities such as terrorist groups.
- The harmful content in harmful memes is often implicit and might require critical judgment to establish the potency it can cause.
- Harmful memes can be classified on multiple levels, based on the intensity of the harm caused, e.g., *very harmful*, *partially harmful*.
- A harmful meme can target multiple individuals, organizations, and/or communities at the same time. In such cases, we ask the annotators to go with their best personal choice.
- Harm can take the form of sarcasm or satire. Sarcasm is praise that is actually an insult, and involves malice, the desire to demean someone. Satire is the ironical exposure of the vices or follies of an individual, a group, an institution, an idea, or society.



(a) Annotation interface (b) Consolidation interface

Figure C.1: Snapshot of the PyBossa GUI we used.

C.4 Annotation Process

Figure C.1 shows the annotation and consolidation interface, based on the crowd-sourcing platform pybossa,[12] which we built for annotating degree of harmfulness and its target. Before starting the annotation process, we requested each annotator to thoroughly go through the annotation guidelines, and we conducted several discussion sessions to understand whether they understood what harmful content is and how to differentiate it from humorous, satirical, hateful, and non-harmful content. The average inter-annotator agreement scores in terms of Cohen's κ for the two tasks are 0.683 and 0.782 on Harm-C, 0.675 and 0.790 on Harm-P.

[12] http://pybossa.com/

Dataset	Harmfulness			Target			
	Very harmful	*Partially harmful*	*Harmless*	*Individual*	*Organization*	*Community*	*Society*
Harm-C	mask (0.0512)	trump (0.0642)	you (0.0264)	trump (0.0541)	deadline (0.0709)	china (0.0665)	mask (0.0441)
	trump (0.0404)	president (0.0273)	home (0.0263	president (0.0263)	associated (0.0709)	chinese (0.0417)	vaccine (0.0430)
	wear (0.0385)	obama (0.0262)	corona (0.0251)	donald (0.0231)	extra (0.0645)	virus (0.0361)	alcohol (0.0309)
	thinks (0.0308	donald (0.0241)	work (0.0222)	obama (0.0217)	ensure (0.0645)	wuhan (0.0359)	temperatures (0.0309)
	killed (0.0269)	virus (0.0213)	day (0.0188)	covid (0.0203)	qanon (0.0600)	cases (0.0319)	killed (0.0271)
Harm-P	photoshopped (0.0589)	democratic (0.0164)	party (0.02514)	biden (0.0331)	libertarian (0.0358)	liberals (0.0328)	crime (0.0201)
	married (0.0343)	obama (0.0158)	debate (0.0151)	joe (0.0323)	republican (0.0319)	radical (0.0325)	rights (0.0195)
	joe (0.0309)	libertarian (0.0156)	president (0.0139)	obama (0.0316)	democratic (0.0293)	islam (0.0323)	gun (0.0181)
	trump (0.0249)	republican (0.0140)	democratic (0.0111)	trump (0.0286)	green (0.0146)	black (0.0237)	taxes (0.0138)
	nazis (0.0241)	vote (0.0096)	green (0.0086)	putin (0.0080)	government (0.0097)	mexicans (0.0168)	law (0.0135)

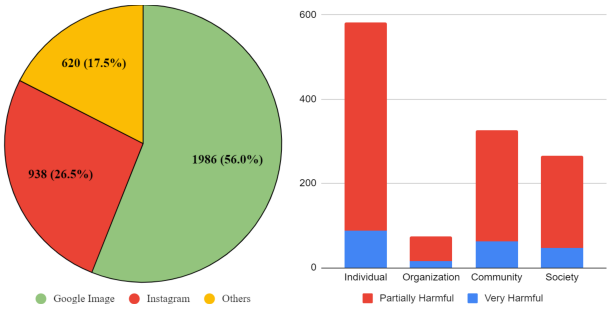
Table C.1: The top-5 most frequent words in each class for each of the two tasks and each of the two datasets. The TF.IDF score for each word is shown in parentheses.

Step 1. Training annotation. First, we took a subset of 200 memes (100 per dataset), and we asked each annotator to do annotations for degree of harmfulness and for its target. This step aimed to ensure that annotators understood the definition of harmfulness and targets. After this initial step, the average inter-annotator agreement in terms of Cohen's κ for the two tasks across all pairs of annotators was quite low: 0.241 and 0.317 on Harm-C, and 0.271 and 0.325 on Harm-P. Next, we asked the annotators to discuss their disagreements and to re-annotate the memes. This time, the average inter-annotator agreement in terms of Cohen's κ improved to 0.704 and 0.815 on Harm-C, and to 0.711 and 0.826 on Harm-P, which is quite satisfactory for both tasks. Hence, we decided we were ready to start the actual annotation process.

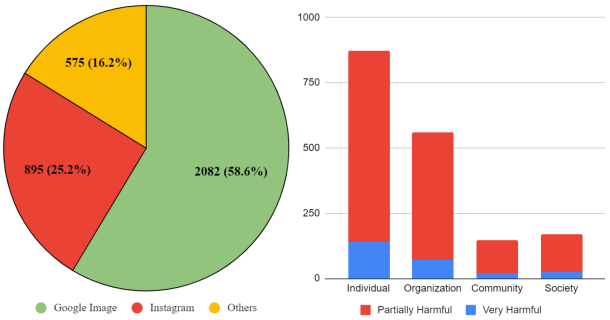
Step 2. Actual annotation. In the actual annotation stage, we divided the two datasets into five equal subsets, and we assigned three annotators per subset. This ensured that each meme was annotated three times. We further asked the annotators to reject the memes that violated any of the filtering criteria, as described in Section B above.

Step 3. Consolidation. After the above annotation, the Cohen's κ was quite good: it was 0.683 and 0.782 for Harm-C, and it was 0.675 and 0.790 for Harm-P. Yet, we observed many memes where two annotators chose the same label (e.g., *partially harmful*), but the third one had made a different choice (e.g., *very harmful*). To resolve such disagreements, in the consolidation phase, we used majority voting to decide on the final label. For cases where all three proposed labels were different, we involved an additional consolidator to help make the final decision.

Figure C.2 shows statistics about the distribution of sources and labels in the final datasets.



(a) statistics about Harm-C



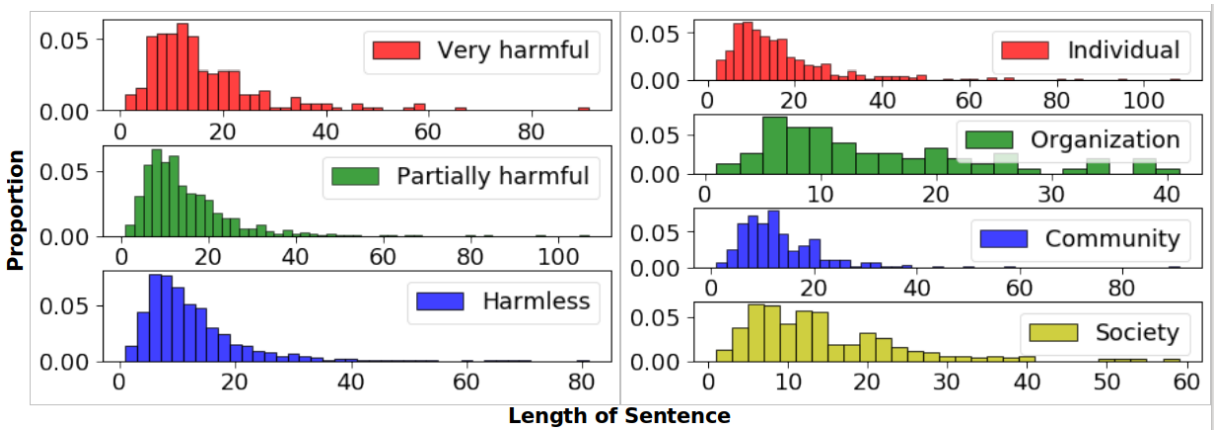
(b) statistics about Harm-P

Figure C.2: Statistics about the distribution of sources and labels in the two datasets.

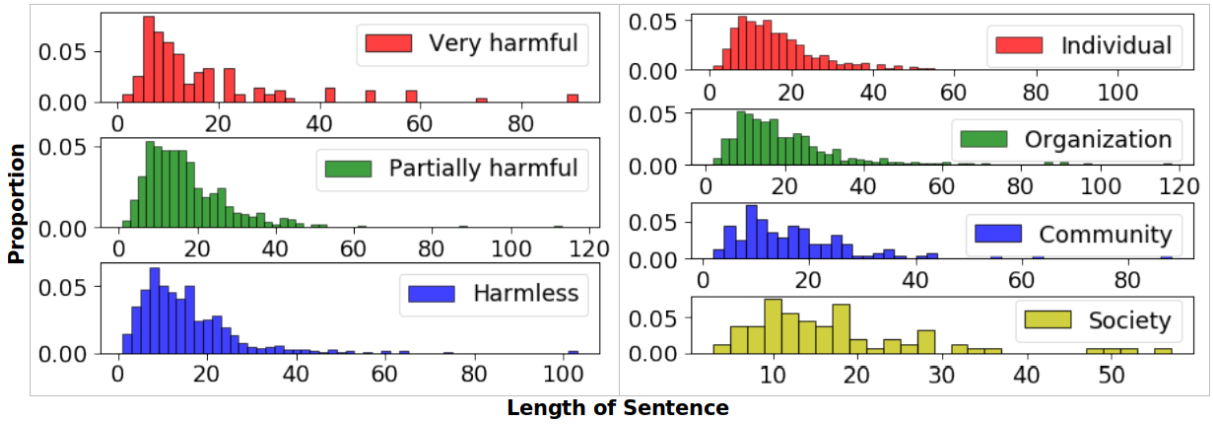
D Lexical Statistics About the Datasets

Table C.1 shows the top-5 most frequent words in the combined validation and test sets for the two datasets. We observe that for *very harmful* and *partially harmful* memes, the names of US politicians and COVID-19 oriented words are quite prominent. Moreover, we notice that the targets are dominated by words like *Trump*, *Joe*, *Obama*, *Republican*, *Wuhan*, *China*, *Islam*, etc. To alleviate the potential bias caused by the presence of such words in the text of the memes, we intentionally included harmless memes related to these individuals, groups, and entities, as we have described in Section 4 above.

Figure C.3 shows the length distribution of the meme text. We see that there are no major differences between the different classes.



(a) Meme text length distribution for Harm-C



(b) Meme text length distribution for Harm-P

Figure C.3: Normalized histograms of meme text length per class for the two datasets.