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

Shraman Pramanick\textsuperscript{1}; Shivam Sharma\textsuperscript{2,\textdagger}; Dimitar Dimitrov\textsuperscript{3}; Md Shad Akhtar\textsuperscript{2}, Preslav Nakov\textsuperscript{5}, Tanmoy Chakraborty\textsuperscript{2}

\textsuperscript{1}Johns Hopkins University \textsuperscript{2}Indraprastha Institute of Information Technology - Delhi \textsuperscript{3}Sofia University \textsuperscript{4}Wipro AI Labs, India \textsuperscript{5}Qatar Computing Research Institute, HBKU, Doha

spraman3@jhu.edu,\{shivams, shad.akhtar, tanmoy\}@iitd.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.

\textsuperscript{\textdagger}denotes equal contribution

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.

(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.
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 additional 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 cross-modal 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-based1 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 extension2 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
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) (Bobic 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.
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 intramodal 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, foreground objects, and various image attributes. 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.
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 representations are $H^{att}$ and $G^{att}$, respectively, where $H^{att} \in \mathbb{R}^{1024}$ and $H^{att} \in \mathbb{R}^{768}$.

\[
H^{att} = W_H \otimes H; \quad G^{att} = W_G \otimes G \quad (1)
\]

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.

\[
F^{res}_{t} = W_t \otimes [F_t, Dense(H^{att})] \quad (2)
\]

\[
F^{att}_{t} = W_T \otimes [F_T, Dense(G^{att})] \quad (3)
\]

Finally, we feed the resulting image and text features, $F^{res}_{t}, F^{att}_{t} \in \mathbb{R}^{512}$ 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 weight the original unimodal features using their respective attention scores and we concatenate them together. We also use residual connections for better gradient flow.

\[
F^{V}_{Meme} = (1 + a_v) F^{res}_{V} \quad (4)
\]

\[
F^{T}_{Meme} = (1 + a_t) F^{res}_{T} \quad (5)
\]

\[
F_{Meme} = W_F \otimes [F^{V}_{Meme}, F^{T}_{Meme}] \quad (6)
\]

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 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.
Table 2: Performance on the two tasks. For two-class, we merge very harmful and partially harmful. \textsuperscript{1}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 ResNetXt-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 (Kietla 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.\textsuperscript{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.
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.

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.
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 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. 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.

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 hyperparameter 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


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\footnote{http://pytorch.org/docs/stable/torchvision/models.html} 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 some 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.

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

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

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 antisocial 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.

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,\(^\text{12}\) 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 $\kappa$ for the two tasks are 0.683 and 0.782 on Harm-C, 0.675 and 0.790 on Harm-P.

\(^{12}\text{http://pybossa.com/}\)
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 $\kappa$ 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 $\kappa$ 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 $\kappa$ 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 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.
Figure C.3: Normalized histograms of meme text length per class for the two datasets.