Detecting Harmful Memes and Their Targets

Shraman Pramanik[1], **Dimiter Dimitrov**[2], **Rituparna Mukherjee**[1], **Shivam Sharma**[1,3], **Md. Shad Akhtar**[1], **Preslav Nakov**[4], **Tanmoy Chakraborty**[1]
[1]Indraprastha Institute of Information Technology - Delhi, India
[2]Sofia University, Bulgaria
[3]Wipro AI Labs, India
[4]Qatar Computing Research Institute, HBKU, Doha, Qatar
{shramanp, shivams, shad.akhtar, tanmoy}@iiitd.ac.in
mitko.bg.ss@gmail.com, ritumukherjee23@gmail.com, pnakov@hbku.edu.qa

Abstract

Among the various modes of communication in social media, the use of Internet memes has emerged as a powerful means to convey political, psychological, and socio-cultural opinions. Although memes are typically humorous in nature, recent days have witnessed a proliferation of *harmful memes* targeted to abuse various social entities. As most harmful memes are highly satirical and abstruse without appropriate contexts, off-the-shelf multimodal models may not be adequate to understand their underlying semantics. In this work, we propose two novel problem formulations: *detecting harmful memes* and *the social entities that these harmful memes target*. To this end, we present HarMeme, the first benchmark dataset, containing $3,544$ memes related to COVID-19. Each meme went through a rigorous two-stage annotation process. In the first stage, we labeled a meme as *very harmful*, *partially harmful*, or *harmless*; in the second stage, we further annotated the type of target(s) that each harmful meme points to: *individual*, *organization*, *community*, or *society/general public/other*. The evaluation results using ten unimodal and multimodal models highlight the importance of using multimodal signals for both tasks. We further discuss the limitations of these models and we argue that more research is needed to address these problems.

1 Introduction

The growing popularity of social media has led to the rise of multimodal content as a way to express ideas and emotions. As a result, a brand new type of message was born: *meme*. A meme is typically formed by an image and a short piece of text on top of it, embedded as part of the image. Memes are typically innocent and designed to look funny.

Over time, memes started being used for harmful purposes in the context of contemporary political and socio-cultural events, targeting individuals, groups, businesses, and society as a whole. At the same time, their multimodal nature and often camouflaged semantics make their analysis highly challenging (Sabat et al., 2019).

Meme analysis. The proliferation of memes online and their increasing importance have led to a growing body of research on meme analysis (Sharma et al., 2020a; Reis et al., 2020; Pramanick et al., 2021). It has also been shown that off-the-shelf multimodal tools may be inadequate to unfold the underlying semantics of a meme as (*i*) memes are often context-dependent, (*ii*) the visual and the textual content are often uncorrelated, and (*iii*) meme images are mostly morphed, and the embedded text is sometimes hard to extract using standard OCR tools (Bonheme and Grzes, 2020).

The dark side of memes. Recently, there has been a lot of effort to explore the dark side of memes, e.g., focusing on hate (Kiela et al., 2020) and offensive (Suryawanshi et al., 2020) memes. However, the harm a meme can cause can be much broader. For instance, the meme[1] in Figure 1c is neither hateful nor offensive, but it is harmful to the media shown on the top left (ABC, CNN, etc.), as it compares them to China, suggesting that they adopt strong censorship policies. In short, the scope of harmful meme detection is *much broader*, and it may encompass other aspects such as cyberbullying, fake news, etc. Moreover, harmful memes have a target (e.g., news organization such as ABC and CNN in our previous example), which requires separate analysis not only to decipher their underlying semantics, but also to help with the explainability of the detection models.

WARNING: This paper contains meme examples and words that are offensive in nature.

[1]In order to avoid potential copyright issues, all memes we show in this paper are our own recreation of existing memes, using images with clear licenses.

Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 2783–2796
August 1–6, 2021.

(a) [0] (b) [2, 0] (c) [1, 1] (d) [2, 2] (e) [2, 3]

Figure 1: Examples from our HarMeme dataset. The labels are in the format [Intensity, Target]. For Intensity, {0, 1, 2} correspond to *harmless*, *partially harmful*, and *very harmful*, respectively. For Target, {0, 1, 2, 3} correspond to *individual*, *organization*, *community*, and *society*, respectively. Examples 1b and 1c are harmful, but neither hateful, nor offensive. Example 1d is both harmful and offensive. Source (a); Source (b); Source (c) 1, Source (c) 2, Source (c) 3; Source (d); Source (e) 1, Source (e) 2, Source (e) 3; License 1 License 2.

Our contributions. In this paper, we study harmful memes, and we formulate two problems. **Problem 1 (Harmful meme detection)**: Given a meme, detect whether it is *very harmful*, *partially harmful*, or *harmless*. **Problem 2 (Target identification of harmful memes)**: Given a harmful meme, identify whether it targets an *individual*, an *organization*, a *community/country*, or the *society/general public/others*. To this end, we develop a novel dataset, HarMeme, containing 3,544 real memes related to COVID-19, which we collected from the web and carefully annotated. Figure 1 shows several examples of memes from our collection, whether they are harmful, as well as the types of their targets. We prepare detailed annotation guidelines for both tasks. We further experiment with ten state-of-the-art unimodal and multimodal models for benchmarking the two problems. Our experiments demonstrate that a systematic combination of multimodal signals is needed to tackle these problems. Interpreting the models further reveals some of the biases that the best multimodal model exhibits, leading to the drop in performance. Finally, we argue that off-the-shelf models are inadequate in this context and that there is a need for specialized models

Our contributions can be summarized as follows:

- We study two new problems: (*i*) detecting harmful memes and (*ii*) detecting their targets.
- We release a new benchmark dataset, HarMeme, developed based on comprehensive annotation guidelines.
- We perform initial experiments with state-of-the-art textual, visual, and multimodal models to establish the baselines. We further discuss the limitations of these models.

Reproducibility. The full dataset and the source code of the baseline models are available at

http://github.com/di-dimitrov/harmeme

The appendix contains the values of the hyperparameters and the detailed annotation guidelines.

2 Related Work

Below, we present an overview of the datasets and the methods used for multimodal meme analysis.

Hate speech detection in memes. Sabat et al. (2019) developed a collection of 5,020 memes for hate speech detection. Similarly, the Hateful Memes Challenge by Facebook introduced a dataset consisting of $10k+$ memes, annotated as hateful or non-hateful (Kiela et al., 2020). The memes were generated *artificially*, so that they resemble real ones shared on social media, along with "benign confounders." As part of this challenge, an array of approaches with different architectures and features have been tried, including Visual BERT, ViLBERT, VLP, UNITER, LXMERT, VILLA, ERNIE-Vil, Oscar and other Transformers (Li et al., 2019; Su et al., 2020; Zhou et al., 2020; Tan and Bansal, 2019; Gan et al., 2020; Yu et al., 2021; Li et al., 2020; Vaswani et al., 2017; Lippe et al., 2020; Zhu, 2020; Muennighoff, 2020). Other approaches include multimodal feature augmentation and cross-modal attention mechanism using inferred image descriptions (Das et al., 2020; Sandulescu, 2020; Zhou and Chen, 2020), as well as up-sampling confounders and loss re-weighting to complement multimodality (Lippe et al., 2020), web entity detection along with fair face classification (Karkkainen and Joo, 2021) from memes (Zhu, 2020), cross-validation ensemble learning and semi-supervised learning (Zhong, 2020) to improve robustness.

Meme sentiment/emotion analysis. Hu and Flaxman (2018) developed the TUMBLR dataset for emotion analysis, consisting of image–text pairs along with associated tags, by collecting posts from the TUMBLR platform. Thang Duong et al. (2017) prepared a multimodal dataset containing images, titles, upvotes, downvotes, #comments, etc., all collected from Reddit. Recently, SemEval-2020 Task 9 on Memotion Analysis (Sharma et al., 2020a) introduced a dataset of $10k$ memes, annotated with sentiment, emotions, and emotion intensity. Most participating systems in this challenge used fusion of visual and textual features computed using models such as Inception, ResNet, CNN, VGG-16 and DenseNet for image representation (Morishita et al., 2020; Sharma et al., 2020b; Yuan et al., 2020), and BERT, XLNet, LSTM, GRU and DistilBERT for text representation (Liu et al., 2020; Gundapu and Mamidi, 2020). Due to class imbalance in the dataset, approaches such as GMM and Training Signal Annealing (TSA) were also found useful. Morishita et al. (2020); Bonheme and Grzes (2020); Guo et al. (2020); Sharma et al. (2020b) proposed ensemble learning, whereas Gundapu and Mamidi (2020); De la Peña Sarracén et al. (2020) and several others used multimodal approaches. A few others leveraged transfer-learning using pre-trained models such as BERT (Devlin et al., 2019), VGG-16 (Simonyan and Zisserman, 2015), and ResNet (He et al., 2016). Finally, state-of-the-art results for all three tasks —sentiment classification, emotion classification and emotion quantification on this dataset,— were reported by Pramanick et al. (2021), who proposed a deep neural model that combines sentence demarcation and multi-hop attention. They also studied the interpretability of the model using the LIME framework (Ribeiro et al., 2016).

Meme propagation. Dupuis and Williams (2019) surveyed personality traits of social media users who are more active in spreading misinformation in the form of memes. Crovitz and Moran (2020) studied the characteristics of memes as a vehicle for spreading potential misinformation and disinformation. Zannettou et al. (2020a) discussed the quantitative aspects of large-scale dissemination of racist and hateful memes among polarized communities on platforms such as 4chan's /pol/. Ling et al. (2021) examined the artistic composition and the aesthetics of memes, the subjects they communicate, and the potential for virality. Based on this analysis, they manually annotated 50 memes as viral vs. non-viral. Zannettou et al. (2020b) analyzed the "Happy merchant" memes and showed how online fringe communities influence their spread to mainstream social networking platforms. They reported reasonable agreement for most manually annotated labels, and established a characterization for meme virality.

Other studies on memes. Reis et al. (2020) built a dataset of memes related to the 2018 and the 2019 election in Brazil ($34k$ images, $17k$ users) and India ($810k$ images, $63k$ users) with focus on misinformation. Another dataset of 950 memes targeted the propaganda techniques used in memes (Dimitrov et al., 2021a), which was also featured as a shared that at SemEval-2021 (Dimitrov et al., 2021b). Leskovec et al. (2009) introduced a dataset of 96 million memes collected from various links and blog posts between August 2008 and April 2009 for tracking the most frequently appearing stories, phrases, and information. Topic modeling of textual and visual cues of hate and racially abusive multi-modal content over sites such as 4chan was studied for scenarios that leverage genetic testing to claim superiority over minorities (Mittos et al., 2020). Zannettou et al. (2020a) examined the content of meme images and online posting activities to identify the probability of occurrence of one event in a specific background process, affecting the occurrence of other events in the rest of the processes, also known as Hawkes process (Hawkes, 1971), within the context of online posting of trolls. Wang et al. (2020) observed that fauxtographic content tends to attract more attention, and established how such content becomes a meme in social media. Finally, there is a recent survey on multi-modal disinformation detection (Alam et al., 2021).

Differences with existing studies. Hate speech detection in multimodal memes (Kiela et al., 2020) is the closest work to ours. However, we are substantially different from it and from other related studies as (*i*) we deal with *harmful* meme detection, which is a more *general* problem than *hateful* meme detection; (*ii*) along with harmful meme detection, we also identify the *entities that the harmful meme targets*; (*iii*) our HarMeme comprises *real-world memes* posted on the web as opposed to using synthetic memes as in (Kiela et al., 2020); and (*iv*) we present a unique dataset and benchmark results for harmful meme detection and for identifying the target of harmful memes.

3 *Harmful* Meme: Definition

Here, we define *harmful memes* as follows: *multimodal units consisting of an image and a piece of text embedded that has the potential to cause harm to an individual, an organization, a community, or the society more generally*. Here, *harm* includes mental abuse, defamation, psycho-physiological injury, proprietary damage, emotional disturbance, and compensated public image.

Harmful vs. hateful/offensive. *Harmful* is a more general term than *offensive* and *hateful*: *offensive* and *hateful* memes are *harmful*, but not all *harmful* memes are *offensive* or *hateful*. For instance, the memes in Figures 1b and 1c are neither offensive nor hateful, but harmful to *Donald Trump* and to *news media* such as CNN, respectively. Offensive memes typically aim to mock or to bully a social entity. A hateful meme contains offensive content that targets an entity (e.g., an individual, a community, or an organization) based on its personal/sensitive attributes such as gender, ethnicity, religion, nationality, sexual orientation, color, race, country of origin, and/or immigration status. The *harmful* content in a harmful meme is often camouflaged and might require critical judgment to establish its potencial to do hard. Moreover, 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 Dataset

Below, we describe the data collection, the annotation process and the guidelines, and we give detailed statistics about the HarMeme dataset.

4.1 Data Collection and Deduplication

To collect potentially harmful memes in the context of COVID-19, we searched using different services, mainly *Google Image Search*. We used keywords such as *Wuhan Virus Memes*, *US Election and COVID Memes*, *COVID Vaccine Memes*, *Work From Home Memes*, *Trump Not Wearing Mask Memes*. We then used an extension[2] of Google Chrome to download the memes. We further scraped various publicly available groups on *Instagram* for meme collection. Note that, adhering to the terms of social media, we did not use content from any private/restricted pages.

[2] http://download-all-images.mobilefirst.me/

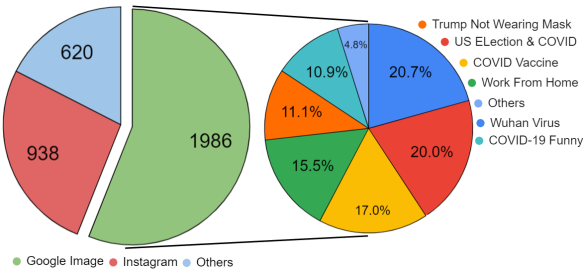


Figure 2: Statistics about the HarMeme dataset. On the left, we show the distribution by source, while on the right, we show the percentage of memes collected by corresponding keywords in Google Image Search.

Unlike the Hateful Memes Challenge (Kiela et al., 2020), which used synthetically generated memes, our HarMeme dataset contains *original memes* that were actually shared in social media. As all memes were gathered from real sources, we maintained strict filtering criteria[3] on the resolution of meme images and on the readability of the meme text during the collection process. We ended up collecting $5,027$ memes. However, as we collected memes from independent sources, we had some duplicates. We thus used two efficient de-duplication repositories[4] [5] sequentially, and we preserved the memes with the highest resolution from each group of duplicates. We removed $1,483$ duplicate memes, thus ending up with a dataset of $3,544$. Although we tried to collect only harmful memes, the dataset contained memes with various levels of harmfulness, which we manually labeled during the annotation process, as discussed in Section 4.3. We further used Google's OCR Vision API[6] to extract the textual content of each meme.

4.2 Annotation Guidelines

As discussed in Section 3, we consider a meme as harmful only if it is implicitly or explicitly intended to cause *harm* to an entity, depending on the personal, political, social, educational or industrial background of that entity. The intended *harm* can be expressed in an obvious manner such as by abusing, offending, disrespecting, insulting, demeaning, or disregarding the entity or any sociocultural or political ideology, belief, principle, or doctrine associated with that entity. Likewise, the *harm* can also be in the form of a more subtle attack such as mocking or ridiculing a person or an idea.

[3] Details are given in Appendix B.3.
[4] gitlab.com/opennota/findimagedupes
[5] http://github.com/arsenetar/dupeguru
[6] http://cloud.google.com/vision

We asked the annotators to label the intensity of the harm as *harmful* or *partially harmful*, depending upon the context and the ingrained explication of the meme. Moreover, we formally defined four different classes of targets and compiled well-defined guidelines[7] that the annotators adhered to while manually annotating the memes. The four target entities are as follows (c.f. Figure 1):

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, Michelle Obama*).

2. **Organization:** An organization is a group of people with a particular purpose, such as a business, a governmental department, a company, an institution or an association, comprising more than one person, and having a particular purpose, such as research organizations (e.g., *WTO, Google*) and political organizations (e.g., *the Democratic Party*).

3. **Community:** A community is 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 forums based on religion, country of origin, gender).

4. **Society:** When a meme promotes conspiracies or hate crimes, it becomes harmful to the general public, i.e., to the entire society.

During the process of collection and annotation, we rejected memes based on the following four criteria: (*i*) the meme text is in code-mixed or non-English language; (*ii*) the meme text is not readable (e.g., blurry text, incomplete text, etc.); (*iii*) the meme is unimodal, containing only textual or visual content; (*iv*) the meme contains cartoons (we added this last criterion as cartoons can be hard to analyze by AI systems).

[7]More details of the annotation guidelines are presented in Appendix B.

(a) Annotation interface

(b) Consolidation interface

Figure 3: Snapshot of the PyBossa GUI used for annotation and consolidation.

4.3 Annotation Process

For the annotation process, we had 15 annotators, including professional linguists and researchers in Natural Language Processing (NLP): 10 of them were male and the other 5 were female, and their age ranged between 24–45 years. We used the PyBossa[8] crowdsourcing framework for our annotations (c.f. Figure 3). We split the annotators into five groups of three people, and each group annotated a different subset of the data. Each annotator spent about 8.5 minutes on average to annotate one meme. At first, we trained our annotators with the definition of harmful memes and their targets, along with the annotation guidelines. To achieve quality annotation, our main focus was to make sure that the annotators were able to understand well what harmful content is and how to differentiate it from humorous, satirical, hateful, and non-harmful content.

[8]http://pybossa.com/

	Phase	Annotators		κ
Harmful meme detection	Trial Annotation	α_1	α_2	0.29
		α_1	α_3	0.34
		α_2	α_3	0.26
	Final Annotation	α_1	α_2	0.67
		α_1	α_3	0.75
		α_2	α_3	0.72
Target identification	Trial Annotation	α_1	α_2	0.35
		α_1	α_3	0.38
		α_2	α_3	0.39
	Final Annotation	α_1	α_2	0.77
		α_1	α_3	0.83
		α_2	α_3	0.79

Table 1: Cohen's κ agreement during different phases of annotation for each task: harmful meme detection (3-class classification) and target identification (4-class classification) of harmful memes.

Dry run. We conducted a dry run on a subset of 200 memes, which helped the annotators understand well the definitions of harmful memes and targets, as well as to eliminate the uncertainties about the annotation guidelines. Let α_i be a single annotator. For the preliminary data, we computed the inter-annotator agreement in terms of Cohen's κ (Bobicev and Sokolova, 2017) for three randomly chosen annotators $\alpha_{[1,2,3]}$ for each meme for both tasks. The results are shown in Table 1. We can see that the score is low for both tasks (0.295 and 0.373), which is expected for the initial dry run. With the progression of the annotation phases, we observed much higher agreement, thus confirming that the dry run helped to train the annotators.

Final annotation. After the dry run, we started the final annotation process. Figure 3a shows an example annotation of the PyBossa annotation platform. We asked the annotators to check whether a given meme falls under the four rejection criteria as given in the annotation guidelines. After confirming the validity of the meme, it was rated by three annotators for both tasks.

Consolidation. In the consolidation phase, for high agreements, we used majority voting to decide the final label, and we added a fourth annotator otherwise. Table 2 shows statistics about the labels and the data splits. After the final annotation, Cohen's κ increased to 0.695 and 0.797 for the two tasks, which is moderate and high agreement, respectively. These scores show the difficulty and the variability in gauging the *harmfulness* by human experts. For example, we found memes where two annotators independently chose *partially harmful*, but the third annotator annotated it as *very harmful*.

4.4 Lexical Analysis of HarMeme

Figure 4 shows the length distribution of the meme text for both tasks, and Table 3 shows the top-5 most frequent words in the union of the validation and the test sets. We can see that names of politicians and words related to COVID-19 are frequent in *very harmful* and *partially harmful* memes. For the target of the harmful memes, we notice the presence of various class-specific words such as *president, trump, obama, china*. These words often incorporate bias in the machine learning models, which makes the dataset more challenging and difficult to learn from (see Section 6.4 for more detail).

5 Benchmarking HarMeme dataset

We provide benchmark evaluations on HarMeme with a variety of state-of-the-art unimodal textual models, unimodal visual models, and models using both modalities. Except for unimodal visual models, we use MMF (Multimodal Framework)[9] to conduct the necessary experiments.

5.1 Unimodal Models

▷ **Text BERT:** We use textual BERT (Devlin et al., 2019) as the unimodal text-only model.
▷ **VGG19, DenseNet, ResNet, ResNeXt:** For the unimodal visual-only models, we used four different 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 the ImageNet (Deng et al., 2009) dataset. We extracted the feature maps from the last pooling layer of each architecture and fed them to a fully connected layer.

5.2 Multimodal Models

▷ **Late Fusion:** This model uses the mean score of pre-trained unimodal ResNet-152 and BERT.
▷ **Concat BERT:** It concatenates the features extracted by pre-trained unimodal ResNet-152 and text BERT, and uses a simple MLP as the classifier.
▷ **MMBT:** Supervised Multimodal Bitransformers (Kiela et al., 2019) is a multimodal architecture that inherently captures the intra-modal and the inter-modal dynamics within various input modalities.
▷ **ViLBERT CC:** Vision and Language BERT (ViLBERT) (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 + text.

[9] github.com/facebookresearch/mmf

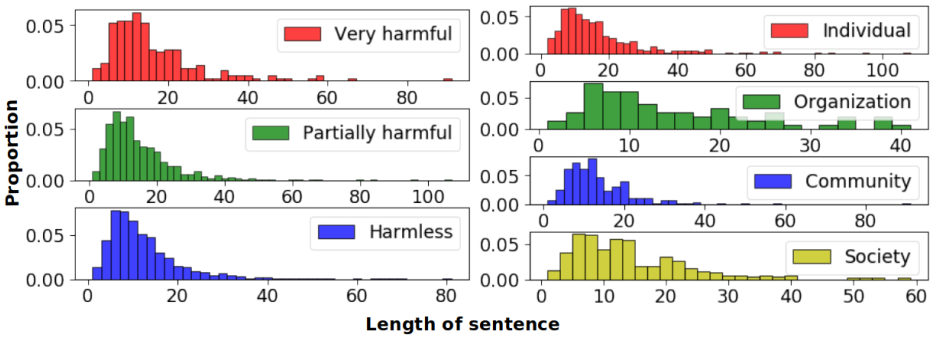


Figure 4: Histogram of the length of the meme' text for each class: for harmfulness on the left, and for the target of harmful memes on the right.

	#Memes	Harmfulness			#Memes	Target			
		Very Harmful	Partially Harmful	Harmless		Individual	Organization	Community	Society
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

Table 2: Statistics about the HarMeme dataset. The memes belonging to the *very harmful* and the *partially harmful* categories are annotated with one of the following four targets: *individual*, *organization*, *community*, or *society*.

Harmfulness			Target			
Very Harmful	**Partially Harmful**	**Harmless**	**Individual**	**Organization**	**Community**	**Society**
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)

Table 3: Top-5 most frequent words per class. The tf-idf score per word is given within parenthesis.

▷ **Visual BERT COCO:** Visual BERT (V-BERT) (Li et al., 2019) pre-trained on the multimodal COCO dataset (Lin et al., 2014) is another strong multimodal model used for a broad range of vision and language tasks.

6 Experimental Results

Below, we report the performance of the models described in the previous section for each of the two tasks. We further discuss some biases that negatively impact performance. Appendix A gives additional details about training and the values of the hyper-parameters we used in our experiments.

Evaluation measures We used six evaluation measures: Accuracy, Precision, Recall, Macro-averaged F1, Mean Absolute Error (MAE), and Macro-Averaged Mean Absolute Error (MMAE) (Baccianella et al., 2009). For the first four measures, higher values are better, while for the last two, lower values are better. Since the test set is imbalanced, measures like macro F1 and MMAE are more relevant.

6.1 Harmful Meme Detection

Table 4 shows the results for the harmful meme detection task. We start our experiments by merging the *very hateful* and the *partially hateful* classes, thus turning the problem into an easier *binary classification*. Afterwards, we perform the 3-class classification task. Since the test set is imbalanced, the majority class baseline achieves 64.76% accuracy. We observe that the unimodal visual models perform only marginally better than the majority class baseline, which indicates that they are insufficient to learn the underlying semantics of the memes.

Moving down the table, we see that the unimodal text model is marginally better than the visual models. Then, for multimodal models, the performance improves noticeably, and more sophisticated fusion techniques yield better results. We also notice the effectiveness of multimodal pre-training over unimodal pre-training, which supports the recent findings by Singh et al. (2020). While both ViLBERT CC and V-BERT COCO perform similarly, the latter achieves better Macro F1 and MMAE, which are the most relevant measures.

Modality	Model	Harmful Meme Detection											
		2-Class Classification						3-Class Classification					
		Acc ↑	P ↑	R ↑	F1 ↑	MAE ↓	MMAE ↓	Acc ↑	P ↑	R ↑	F1 ↑	MAE ↓	MMAE ↓
	Human†	90.68	84.35	84.19	83.55	0.1760	0.1723	86.10	67.35	65.84	65.10	0.2484	0.4857
	Majority	64.76	32.38	50.00	39.30	0.3524	0.5000	64.76	21.58	33.33	26.20	0.4125	1.0
Text Only	TextBERT	70.17	65.96	66.38	66.25	0.3173	0.2911	68.93	48.49	49.15	48.72	0.3250	0.5591
Image Only	VGG19	68.12	60.25	61.23	61.86	0.3204	0.3190	66.24	40.95	44.02	41.76	0.3198	0.6487
	DenseNet-161	68.42	61.08	62.10	62.54	0.3202	0.3125	65.21	41.88	44.25	42.15	0.3102	0.6326
	ResNet-152	68.74	61.86	62.89	62.97	0.3188	0.3114	65.29	41.95	44.32	43.02	0.3047	0.6264
	ResNeXt-101	69.79	62.32	63.26	63.68	0.3175	0.3029	66.55	42.62	44.87	43.68	0.3036	0.6499
Image + Text (Unimodal Pre-training)	Late Fusion	73.24	70.28	70.36	70.25	0.3167	0.2927	66.67	44.96	50.02	45.06	0.3850	0.6077
	Concat BERT	71.82	71.58	72.23	71.82	0.3033	0.3156	65.54	42.29	45.42	43.37	0.3881	0.5976
	MMBT	73.48	68.89	68.95	67.12	0.3101	0.3258	68.08	51.72	51.94	50.88	0.3403	0.6474
Image + Text (Multimodal Pre-training)	ViLBERT CC	78.53	78.62	**81.41**	78.06	0.2279	0.1881	**75.71**	48.89	49.21	48.82	**0.2763**	0.5329
	V-BERT COCO	**81.36**	**79.55**	81.19	**80.13**	**0.1972**	**0.1857**	74.01	**56.35**	**54.79**	**53.85**	0.3063	**0.5303**

Table 4: Performance for harmful meme detection. For two-class classification, we merge *very harmful* and *partially harmful* into a single class. † This row reports the human accuracy on the test set.

Modality	Model	Target Identification of Harmful Memes					
		Acc ↑	P ↑	R ↑	F1 ↑	MAE ↓	MMAE ↓
	Human†	87.55	82.28	84.15	82.01	0.7866	0.3647
	Majority	46.60	11.65	25.00	15.89	1.2201	1.5000
Text (T) only	TextBERT	69.35	55.60	54.37	55.60	1.1612	0.8988
Image (I) only	VGG19	63.48	53.85	54.02	53.60	1.1687	1.0549
	DenseNet-161	64.52	53.96	53.95	53.51	1.1655	1.0065
	ResNet-152	65.75	54.25	54.13	53.78	1.1628	1.0459
	ResNeXt-101	65.82	54.47	54.20	53.95	1.1616	0.9277
I + T (Unimmodal Pre-training)	Late Fusion	72.58	58.43	58.83	58.43	1.1476	0.6318
	Concat BERT	67.74	54.79	49.65	49.77	1.1377	0.8879
	MMBT	72.58	58.43	58.83	58.35	1.1476	0.6318
I + T (Multimodal Pre-training)	ViLBERT CC	72.58	59.92	55.78	57.17	1.1671	0.8035
	V-BERT COCO	**75.81**	**66.29**	**69.09**	**65.77**	**1.1078**	**0.5036**

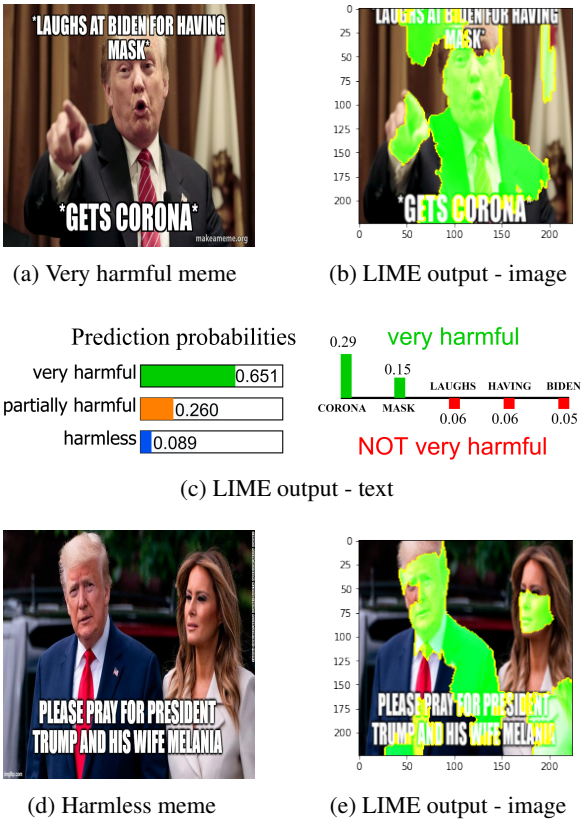
Table 5: Performance for target identification of harmful memes (†human accuracy on the test set).

6.2 Target Identification for Harmful Memes

Table 5 shows the results for the target identification task. This is an imbalanced 4-class classification problem, and the majority class baseline yields 46.60% accuracy. The unimodal models perform relatively better here, achieving $63\% - 70\%$ accuracy; their F1 Macro and MMAE scores are also above the majority class. However, the overall performance of the unimodal models is poor. Incorporating multimodal signals with fine-grained fusion improves the results substantially, and advanced multimodal fusion techniques with multimodal pre-training perform much better than simple late fusion with unimodal pre-training. Moreover, V-BERT COCO outperforms ViLBERT CC by 8% of F1 score and by nearly 0.3 of MMAE.

6.3 Human Evaluation

To understand how human subjects perceive these tasks, we further hired a different set of experts (not the annotators) to label the test set. We observed $86\% - 91\%$ accuracy on average for both tasks, which is much higher than V-BERT, the best-performing model. This shows that their is a potential for enriched multimodal models that better understand the ingrained semantics of the memes.



(a) Very harmful meme (b) LIME output - image

(c) LIME output - text

(d) Harmless meme (e) LIME output - image

Figure 5: Example of explanation by LIME on both visual and textual modalities and visualization of bias in V-BERT for both tasks.

6.4 Side-by-side Diagnostics and Anecdotes

Since the HarMeme dataset was compiled of memes related to COVID-19, we expected that models with enriched contextual knowledge and sophisticated technique would have superior performance. Thus, to comprehend the interpretability of V-BERT (the best model), we used LIME (Locally Interpretable Model-Agnostic Explanations) (Ribeiro et al., 2016), a consistent model-agnostic explainer to interpret the predictions.

We chose two memes from the test set to analyze the potential explanability of V-BERT. The first meme, which is shown in Figure 5a, was manually labeled as *very harmful*, and V-BERT successfully classified it, with prediction probabilities of 0.651, 0.260, and 0.089 corresponding to the *very harmful*, the *partially harmful*, and the *harmless* classes respectively. Figure 5b highlights the most contributing super-pixels to the *very harmful* (green) class. As expected, the face of Donald Trump, as highlighted by the green pixels, prominently contributed to the prediction. Figure 5c demonstrates the contribution of different meme words to the model prediction. We can see that words like *CORONA* and *MASK* have significant contributions to the *very harmful* class, thus supporting the lexical analysis of HarMeme as shown in Table 3.

The second meme, which is shown in Figure 5d, was manually labeled as *harmless*, but V-BERT incorrectly predicted it to be *very harmful*. Figure 5e shows that, similarly to the previous example, the face of Donald Trump contributed to the prediction of the model. We looked closer into our dataset, and we found that it contained many memes with the image of Donald Trump, and that the majority of these memes fall under the *very harmful* category and targeted and individual. Therefore, instead of leaning the underlying semantics of one particular meme, the model easily got biased by the presence of Donald Trump's image and blindly classified the meme as *very harmful*.

7 Conclusion and Future Work

We presented HarMeme, the first large-scale benchmark dataset, containing 3,544 memes, related to COVID-19, with annotations for degree of harmfulness (*very harmful*, *partially harmful*, or *harmless*), as well as for the target of the harm (an *individual*, an *organization*, a *community*, or *society*). The evaluation results using several unimodal and multimodal models highlighted the importance of modeling the multimodal signal (for both tasks) —(*i*) detecting harmful memes and (*ii*) detecting their targets—, and indicated the need for more sophisticated methods. We also analyzed the best model and identified its limitations.

In future work, we plan to design new multimodal models and to extend HarMeme with examples from other topics, as well as to other languages. Alleviating the biases in the dataset and in the models are other important research directions.

Ethics and Broader Impact

User Privacy. Our dataset only includes memes and it does not contain any user information.

Biases. Any biases found in the dataset are unintentional, and we do not intend to do 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 assignment of the schema to 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 present our dataset 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.

Acknowledgments

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

References

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

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.

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.

Lisa Bonheme and Marek Grzes. 2020. SESAM at SemEval-2020 task 8: Investigating the relationship between image and text in sentiment analysis of memes. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, SemEval '20, pages 804–816, Barcelona, Spain.

Darren Crovitz and Clarice Moran. 2020. Analyzing Disruptive Memes in an Age of International Interference. *The English Journal*, 109(4):62–69.

Abhishek Das, Japsimar Singh Wahi, and Siyao Li. 2020. Detecting hate speech in multi-modal memes. *arXiv 2012.14891*.

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, FL, 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, MN, 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.

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.

Marc J. Dupuis and Andrew Williams. 2019. The spread of disinformation on the Web: An examination of memes on social networking. In *Proceedings of the 2019 IEEE SmartWorld, Ubiquitous Intelligence Computing, Advanced Trusted Computing, Scalable Computing Communications, Cloud Big Data Computing, Internet of People and Smart City Innovation*, SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI '19, pages 1412–1418.

Zhe Gan, Yen-Chun Chen, Linjie Li, Chen Zhu, Yu Cheng, and Jingjing Liu. 2020. Large-scale adversarial training for vision-and-language representation learning. In *Proceedings of the 34th Conference on Neural Information Processing Systems*, NeurIPS '20, pages 1–15, Vancouver, Canada.

Sunil Gundapu and Radhika Mamidi. 2020. Gundapusunil at SemEval-2020 task 8: Multimodal Memotion Analysis. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, SemEval '20, pages 1112–1119, Barcelona, Spain.

Yingmei Guo, Jinfa Huang, Yanlong Dong, and Mingxing Xu. 2020. Guoym at SemEval-2020 task 8: Ensemble-based classification of visuolingual metaphor in memes. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, SemEval '20, pages 1120–1125, Barcelona, Spain.

Alan Hawkes. 1971. Spectra of some self-exciting and mutually exciting point processes. *Biometrika*, 58:83–90.

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, NV, USA.

Anthony Hu and Seth Flaxman. 2018. Multimodal sentiment analysis to explore the structure of emotions. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, ACMKDD '18, pages 350–358, London, UK.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q. Weinberger. 2017. Densely connected convolutional networks. In *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '17, pages 2261–2269, Honolulu, HI, USA.

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

Douwe Kiela, Suvrat Bhooshan, Hamed Firooz, and Davide Testuggine. 2019. Supervised multimodal bitransformers for classifying images and text. *arXiv 1909.02950*.

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 *Advances in Neural Information Processing Systems*, volume 33 of *NeurIPS '20*, pages 2611–2624, Vancouver, Canada.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations*, ICLR '14, San Diego, CA, USA.

Jure Leskovec, Lars Backstrom, and Jon Kleinberg. 2009. Meme-tracking and the dynamics of the news cycle. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '09, page 497–506, New York, USA.

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*.

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *Proceedings of the 16th European Computer Vision Conference*, ECCV '20, pages 121–137, Glasgow, UK.

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.

Chen Ling, Ihab AbuHilal, Jeremy Blackburn, Emiliano De Cristofaro, Savvas Zannettou, and Gianluca Stringhini. 2021. Dissecting the meme magic: Understanding indicators of virality in image memes. In *Proceedings of the 24th ACM Conference on Computer-Supported Cooperative Work and Social Computing*, CSCW '21.

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*.

Zehao Liu, Emmanuel Osei-Brefo, Siyuan Chen, and Huizhi Liang. 2020. UoR at SemEval-2020 task 8: Gaussian mixture modelling (GMM) based sampling approach for multi-modal memotion analysis. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, SemEval '20, pages 1201–1207, Barcelona, Spain.

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.

Alexandros Mittos, Savvas Zannettou, Jeremy Blackburn, and Emiliano De Cristofaro. 2020. "And we will fight for our race!" A measurement study of genetic testing conversations on Reddit and 4chan. In *Proceedings of the Fourteenth International AAAI Conference on Web and Social Media*, ICWSM '20, pages 452–463, Atlanta, GA, USA.

Terufumi Morishita, Gaku Morio, Shota Horiguchi, Hiroaki Ozaki, and Toshinori Miyoshi. 2020. Hitachi at SemEval-2020 task 8: Simple but effective modality ensemble for meme emotion recognition. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, SemEval '20, pages 1126–1134, Barcelona, Spain.

Niklas Muennighoff. 2020. Vilio: State-of-the-art Visio-Linguistic Models applied to Hateful Memes. *arxiv 2012.07788*.

Gretel Liz De la Peña Sarracén, Paolo Rosso, and Anastasia Giachanou. 2020. PRHLT-UPV at SemEval-2020 task 8: Study of multimodal techniques for memes analysis. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, SemEval '20, pages 908–915, Barcelona, Spain.

Shraman Pramanick, Md Shad Akhtar, and Tanmoy Chakraborty. 2021. Exercise? I thought you said 'extra fries': Leveraging sentence demarcations and multi-hop attention for meme affect analysis. *proceedings of the Fifteenth International AAAI Conference on Web and Social Media*.

Julio C. S. Reis, Philipe Melo, Kiran Garimella, Jussara M. Almeida, Dean Eckles, and Fabrício Benevenuto. 2020. A dataset of fact-checked images shared on WhatsApp during the Brazilian and Indian elections. In *Proceedings of the International AAAI Conference on Web and Social Media*, ICWSM '20, pages 903–908.

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, San Francisco, CA, USA.

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

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

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

Mayukh Sharma, Ilanthenral Kandasamy, and W.b. Vasantha. 2020b. Memebusters at SemEval-2020 task 8: Feature fusion model for sentiment analysis on memes using transfer learning. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, SemEval '20, pages 1163–1171, Barcelona, Spain.

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, CA, USA.

Amanpreet Singh, Vedanuj Goswami, and Devi Parikh. 2020. Are we pretraining it right? Digging deeper into visio-linguistic pretraining. *arXiv 2004.08744*.

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.

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2020. VL-BERT: pre-training of generic visual-linguistic representations. In *Proceedings of the 8th International Conference on Learning Representations*, ICLR '20, Addis Ababa, Ethiopia.

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*, LREC-TRAC '20, pages 32–41, Marseille, France.

Hao Tan and Mohit Bansal. 2019. LXMERT: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, EMNLP-IJCNLP '17, pages 5100–5111, Hong Kong, China.

Chi Thang Duong, Remi Lebret, and Karl Aberer. 2017. Multimodal classification for analysing social media. *arXiv 1708.02099*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz 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, CA, USA.

Yuping Wang, Fatemeh Tahmasbi, Jeremy Blackburn, Barry Bradlyn, Emiliano De Cristofaro, David Magerman, Savvas Zannettou, and Gianluca Stringhini. 2020. Understanding the use of fauxtography on social media. *arXiv 2009.11792*.

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, HI, USA.

Fei Yu, Jiji Tang, Weichong Yin, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. 2021. ERNIE-ViL: Knowledge enhanced vision-language representations through scene graph. In *Proceedings of the Thirty-Fifth AAAI Conference on Artificial Intelligence*, AAAI '21, pages 3208–3216.

Li Yuan, Jin Wang, and Xuejie Zhang. 2020. YNU-HPCC at SemEval-2020 task 8: Using a parallel-channel model for memotion analysis. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, SemEval '20, pages 916–921, Barcelona, Spain.

Savvas Zannettou, Tristan Caulfield, Barry Bradlyn, Emiliano De Cristofaro, Gianluca Stringhini, and Jeremy Blackburn. 2020a. Characterizing the use of images in state-sponsored information warfare operations by Russian trolls on Twitter. In *Proceedings of the Fourteenth International AAAI Conference on Web and Social Media*, ICWSM '20, pages 774–785, Atlanta, GA, USA.

Savvas Zannettou, Joel Finkelstein, Barry Bradlyn, and Jeremy Blackburn. 2020b. A quantitative approach to understanding online antisemitism. In *Proceedings of the Fourteenth International AAAI Conference on Web and Social Media*, ICWSM '20, pages 786–797, Atlanta, GA, USA.

Xiayu Zhong. 2020. Classification of multimodal hate speech – the winning solution of hateful memes challenge. *arXiv 2012.01002*.

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

Yi Zhou and Zhenhao Chen. 2020. Multimodal learning for hateful memes detection. *arXiv 2011.12870*.

Ron Zhu. 2020. Enhance multimodal transformer with external label and in-domain pretrain: Hateful meme challenge winning solution. *arXiv 2012.08290*.

A Implementation Details and Hyper-Parameter Values

We trained all the models using the Pytorch framework on an NVIDIA Tesla T4 GPU with 16 GB of dedicated memory and with CUDA-10 and cuDNN-11 installed. For the unimodal models, we imported all the pre-trained weights from the TORCHVISION.MODELS[10] subpackage of PyTorch. We initialized the non pre-trained weights randomly with a zero-mean Gaussian distribution with a standard deviation of 0.02. To minimize the impact of the label imbalance in the loss calculation, we assigned larger weights to the minority class. We trained our models using the Adam optimizer (Kingma and Ba, 2014) and the negative log-likelihood loss as the objective function. Table A.1 gives the values of all hyper-parameters we used for training.

We trained the models end-to-end for the two classification tasks, i.e., the memes that were classified as *Very Harmful* or *Partially Harmful* in the first classification stage were sent to the second stage for target identification.

B Annotation Guidelines

B.1 What do we mean by *harmful* memes?

The entrenched meaning of harmful memes is targeted towards a social entity (e.g., an individual, an organization, a community, etc.), likely to cause calumny/vilification/defamation depending on their background (bias, social background, educational background, etc.). The *harm* caused by a meme can be in the form of mental abuse, psycho-physiological injury, proprietary damage, emotional disturbance, compensated public image. A harmful meme typically attacks celebrities or well-known organizations, with the intent to expose their professional demeanor.

Characteristics of *harmful* memes:

- Harmful memes may or may not be offensive, hateful, or biased in nature.
- Harmful memes expose 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 its potential to do harm.
- Harmful memes can be classified at multiple levels, based on the intensity of the harm they could cause, e.g., *very harmful* or *partially harmful*.
- One harmful meme can target multiple individuals, organizations, and/or communities at the same time. In that case, we asked the annotators to go with the best personal judgment.
- Harm can be expressed in the form of sarcasm and/or political satire. Sarcasm is praise that is actually an insult; sarcasm generally involves malice, the desire to put someone down. On the other hand, satire is the ironical exposure of the vices or the follies of an individual, a group, an institution, an idea, the society, etc., usually with the aim to correcting it.

B.2 What is the difference between *organization* and *community*?

An organization is a group of people with a particular purpose, such as a business or a government department. Examples include a company, an institution, or an association comprising one or more people with a particular purpose, e.g., a research organization, a political organization, etc.

On the other hand, a community is a social unit (a group of living things) with a commonality such as norms, religion, values, ideology customs, or identity. 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 the virtual space through communication platforms.

B.3 When do we *reject* a meme?

We apply the following rejection criteria during the process of data collection and annotation:

1. The meme's text is code-mixed or not in English.
2. The meme's text is not readable. (e.g., blurry text, incomplete text, etc.)
3. The meme is unimodal in nature, containing only textual or only visual content.
4. The meme contains a cartoon.

Figure B.1 shows some rejected memes.

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

	Models	Hyper-parameters					
		Batch Size	Epochs	Learning Rate	Image Encoder	Text Encoder	#Parameters
Unimodal	TextBERT	16	100	0.001	-	Bert-base-uncased	110,683,414
	VGG19	64	200	0.01	VGG19	-	138,357,544
	DenseNet-161	32	200	0.01	DenseNet-161	-	28,681,538
	ResNet-152	32	300	0.01	ResNet-152	-	60,192,808
	ResNeXt-101	32	300	0.01	ResNeXt-101	-	83,455,272
Multimodal	Late Fusion	16	200	0.0001	ResNet-152	Bert-base-uncased	170,983,752
	Concat BERT	16	200	0.001	ResNet-152	Bert-base-uncased	170,982,214
	MMBT	16	200	0.001	ResNet-152	Bert-base-uncased	169,808,726
	ViLBERT CC	16	100	0.001	Faster RCNN	Bert-base-uncased	112,044,290
	V-BERT COCO	16	100	0.001	Faster RCNN	Bert-base-uncased	247,782,404

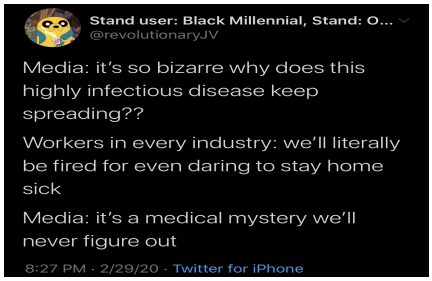
Table A.1: The values of the hyper-parameters of all our models.

(a) Non-English (Hindi) meme. Source License

(b) Unreadable meme. Source License

(c) Meme with a cartoon. Source License

(d) Meme without textual modality. Source License



(e) Meme without visual modality. Source License

Figure B.1: Examples of memes that we rejected during the process of data collection and annotation.