Decoding the underlying meaning of multimodal hateful memes

Ming Shan HEE

Wen Haw CHONG
Singapore Management University, whchong.2013@phdis.smu.edu.sg

Roy Ka-Wei LEE

Citation

Decoding the Underlying Meaning of Multimodal Hateful Memes

Ming Shan Hee[1] , **Wen-Haw Chong**[2] and **Roy Ka-Wei Lee**[1]

[1]Singapore University of Technology and Design
[2]Singapore Management University
mingshan_hee@mymail.sutd.edu.sg, whchong.2013@phdis.smu.edu.sg, roy_lee@sutd.edu.sg

Abstract

Recent studies have proposed models that yielded promising performance for the hateful meme classification task. Nevertheless, these proposed models do not generate interpretable explanations that uncover the underlying meaning and support the classification output. A major reason for the lack of explainable hateful meme methods is the absence of a hateful meme dataset that contains ground truth explanations for benchmarking or training. Intuitively, having such explanations can educate and assist content moderators in interpreting and removing flagged hateful memes. This paper address this research gap by introducing **Hat**eful meme with **Re**asons **D**ataset (HatReD), which is a new multimodal hateful meme dataset annotated with the underlying hateful contextual reasons. We also define a new conditional generation task that aims to automatically generate underlying reasons to explain hateful memes and establish the baseline performance of state-of-the-art pre-trained language models on this task. We further demonstrate the usefulness of HatReD by analyzing the challenges of the new conditional generation task in explaining memes in seen and unseen domains. The dataset and benchmark models are made available here: https://github.com/Social-AI-Studio/HatRed

Disclaimer: This paper contains discriminatory content that may be disturbing to some readers.

1 Introduction

Internet memes are viral content spread among online communities. While most memes are often humorous and benign, hateful memes, which attack a target group or individual based on characteristics such as race, gender, and religion, have become a growing concern. As part of its effort to moderate the spread of hateful memes, Facebook recently launched the "*hateful meme challenge*" [Kiela *et al.*, 2020]. The challenge released a dataset with 10K+ hateful memes to encourage submissions of automated solutions to detect hateful memes. This led to the development of various multimodal deep learning approaches for hateful meme classifications [Yang *et al.*, 2022; Lee *et al.*, 2021]. Other studies have also contributed to the growing effort to curb hateful memes by collecting and releasing large hateful meme datasets to support the training and evaluation of hateful meme classification models [Suryawanshi *et al.*, 2020; Gasparini *et al.*, 2021; Pramanick *et al.*, 2021a; Sharma *et al.*, 2023; Sharma *et al.*, 2022b].

However, existing studies have primarily focused on performing hateful meme classification (i.e., predicting if a given meme is hateful) with limited explanations for its prediction. Providing explanations for the detected hate meme is integral to the content moderation process. The content moderators and users may want to understand why a particular meme is flagged as hateful. Nevertheless, explaining hateful memes is challenging as it requires a combination of information from various modalities and specific socio-cultural knowledge [Kiela *et al.*, 2021]. Consider the hateful meme in Figure 1. To explain the hateful meme, one will need the socio-cultural knowledge that the girl in the image is Anne Frank, and realize the textual reference refers to the gas poisoning of Jews during the Holocaust.

Recognizing the importance of providing contextual reasons for the predicted hateful memes, recent studies have performed fine-grained analysis to classify the type of attacks [Mathias *et al.*, 2021] and infer the targets being attacked [Pramanick *et al.*, 2021a; Sharma *et al.*, 2022a]. However, such fine-grained analysis may still be inadequate for content moderators to understand hateful memes. These analyses often only predict general protected characteristics (e.g., race) but not the specific social target attacked (e.g., Jews). Furthermore, having informative reasons in natural sentences, such as the example provided in Figure 1, would make it easier for content moderators to comprehend the hateful memes.

In a recent study, Elsherief [2021] collected a large textual dataset to support implicit hate speech classification. The dataset contains hateful textual posts annotated with their corresponding *implied statements*, which could be seen as a form of explanation to aid content moderators in understanding the implicit hate speech. The availability of ground truth explanations also allows researchers to apply and explore training Pre-trained Language Models (PLMs) such as GPT-2 for explanation generation. Ideally, we would also like to generate the underlying reasons for why a flagged hateful meme is considered hateful. To the best of our knowledge, there is no current dataset to facilitate this exploration.

Detected Web Entities
- Madame Tussauds - Berlin - London - **Anne Frank** - Zentrum Wax sculpture - Museum

Reason
Disrespects the jews by downplaying the burning of jews during the holocaust

Figure 1: Example of a hateful meme in HatReD.

Research Objectives. To address the research gaps, we propose a new conditional generation task that aims to generate the underlying reasons to explain hateful memes automatically. We constructed the **Hat**eful meme **Re**asoning **D**ataset (HatReD), which is a new multimodal hateful memes dataset annotated with the underlying hateful contextual reasons, to support the proposed task. Specifically, we carefully design a framework to annotate Facebook's *Fine-Grained Hateful Memes* dataset [Mathias *et al.*, 2021] with the underlying hateful reasons. We fine-tune PLMs on HatReD and conduct extensive experiments to evaluate the PLM's performance and limitations on the new generation task. Finally, we also demonstrate the usefulness of HatReD by evaluating the fine-tuned PLMs' ability to generate the explanations for hateful memes in an unseen misogynous meme dataset.

Contributions. We summarize our contributions as follows: (1) We construct HatReD, which is a multimodal hateful meme dataset annotated with underlying hateful contextual reasons. To the best of our knowledge, this is the first hateful meme dataset with written explanations. (2) We introduce a new conditional generation task that aims to generate underlying reasons to explain hateful memes automatically. We conduct extensive experiments to establish the task baselines using state-of-the-art PLMs. (3) We analyze the challenges of generating the generation new task and demonstrate the usefulness of HatReD in explaining memes in seen and unseen domains.

2 Related Works

2.1 Hateful Meme Datasets

Hateful meme classification is an emerging research topic made popular by the availability of several hateful meme datasets. Table 1 summarizes the hateful meme datasets released over the last few years. All the datasets contain class labels that support the hateful meme classification task. For instance, the memes in *Facebook Hateful Meme Challenge* dataset are labeled "*hateful*" or "*non-hateful*" [Kiela *et al.*, 2020]. Similarly, Suryawanshi [2020] collected a small dataset of politics-related memes from Tumblr and annotated the memes as "*offensive*" or "*non-offensive*". Besides the class labels that facilitate hateful meme classification, some datasets have also provided supplementary infor-

mation on hateful memes. For example, Pramanick [2021a] collected a dataset containing COVID-19 related memes. The researchers annotated the harmfulness of the memes and the types of target (e.g., *individual*, *organization*, and *community*) attacked in the harmful memes. Mathias [2021] extended the Facebook Hateful Meme Challenge dataset by annotating the types of attack (e.g., *Dehumanizing*) and target type (e.g., *race*) attacked in the hateful memes. While the supplementary information could provide additional contexts to the hateful memes, it is still inadequate in informatively explaining the hateful memes.

Recent studies have attempted to identify and explain the subtle hateful connotations of hate speech. Sap [2020] developed the *Social Bias Frame*, a pragmatic framework that can capture knowledge regarding the biased implications of hate speech, such as its group reference and implied statement. Elsherief [2021] subsequently extended the *Social Bias Frame* to include implicit hate speech, which has a broader scope than social bias and stereotypes. Nevertheless, these existing studies have mainly focused on explaining text-based hate speeches. In this study, we aim to fill the research gap by proposing a new hateful meme dataset that includes informative reasons to explain the background contexts in hateful memes.

2.2 Hateful Meme Classification

Hateful meme classification is an emerging multimodal task that has gain popularity in recent years. Existing studies have explored *classic two-stream models* that combine the text and visual features to classify the hateful memes [Kiela *et al.*, 2020; Suryawanshi *et al.*, 2020], and fine-tuning large scale pre-trained multimodal models for the multimodal classification task [Lippe *et al.*, 2020; Muennighoff, 2020; Pramanick *et al.*, 2021b; Yang *et al.*, 2022; Lee *et al.*, 2021; Cao *et al.*, 2022; Sharma *et al.*, 2022b]. Nevertheless, most of the existing studies have focused on the hateful meme classification task without providing any explanation for the hateful memes. A recent study proposed a post-hoc explanation framework to examine the visual-text slur grounding learned by pre-trained multimodal models trained to perform the hateful meme classification task [Hee *et al.*, 2022]. However, the framework still falls short in providing informative reasons to explain hateful memes. We postulate that the primary reason for underwhelming research studies explaining hateful memes is the lack of a dataset. Therefore, we propose a multimodal hateful meme dataset with informative reasons to encourage researchers to contribute solutions in this space. Specifically, this study will provide the benchmark dataset for the hateful meme explanation task and comprehensively evaluate state-of-the-art PLMs' capabilities to generate natural language reasons for hateful memes.

3 HatReD Dataset

In this study, we propose HatReD[1], a new multimodal hateful meme explanation dataset. Specifically, we recruited four native English speakers to annotate the underlying reasons for

---

[1]Note that researchers will have to agree with Facebook's data access agreement to download the memes

Work	Domain	Size	Num. Hateful/Off.	Target/Group	Attack Type	Explanations
Kiela [2020]	Multiple Groups	10000	$3,266^*$			
Suryawanshi [2020]	Politics	743	305			
Mathias [2021]	Multiple Groups	10,000	$3,253^*$	✓	✓	
Pramanick [2021a]	COVID-19	3,544	1,249	✓		
Gasparini [2021]	Misogyny	800	400			
Fersini [2022]	Misogyny	11,000	$5,504^*$			
HatReD (Ours)	Multiple Groups	3,228	3,228	✓	✓	✓

Table 1: Summary of hateful meme datasets. All the datasets contain class labels that support the hateful meme classification task. However, none of the existing datasets provide explanation for the hateful context. HatReD is the first dataset that include free-text explanations for multi-modal memes in Hearst-style templates. * indicates that the hateful and/or offensives memes in the test set are excluded, as the test set are not made publicly available.

the hate speeches found in Facebook's *Fine-Grained Hateful Memes* dataset [Mathias *et al.*, 2021]. To the best of our knowledge, this is the first multimodal hateful meme dataset annotated with hateful contextual reasons. In the subsequent sections, we will discuss the dataset construction process and provide a preliminary analysis of HatReD.

3.1 Dataset Construction

Fine-Grained Hateful Memes is a large-scale multimodal memes benchmark dataset that contains five standard kinds of incitement to hatred, including sexual, racial, religious, nationality and disability hatred. The selection of these curated memes conforms with the community standards on hate speech employed by Facebook [2], which presents a pragmatic view of hate speech in memes.

Dataset and Annotation Preparation
The main challenge of explaining hateful memes is that the explanation often requires knowledge of relevant socio-cultural backgrounds and societal prejudices. For example, recognizing that the presence of rainbow-striped flags may indicate a connection to LGBTQ movement, such as Pride Day. Therefore, to assist annotators in explaining hateful memes, we used the Google Web Detect API to extract web entities from the meme. The extracted web entities could provide the additional socio-cultural context for the images used in the memes. For instance, the API will return *"Anne Frank"* for the image used in Figure 1. The annotators are also encouraged to search about the unfamiliar extracted web entities and slurs in meme text on external knowledge bases such as Wikipedia and Hatebase[3]. For instance, annotators without prior knowledge of *Anne Frank* can search Wikipedia for information about her and events related to her, such as the Holocaust. Through this process, the annotators can deepen their knowledge of relevant cultural backgrounds and societal prejudices over multiple iterations of annotation and improve their annotation.

Reason Annotation
We trained the four annotators to produce high-quality reasons for each hateful meme. We present annotators with the meme, the social characteristics of the attacked target (e.g.,

---

[2]https://www.facebook.com/communitystandards/hate_speech
[3]https://hatebase.org/

	Fluency	**Relevance**
Average Score	*4.97*	*4.81*

Table 2: Human evaluation results on annotated reasons

Social Category	**# Social Targets**	**# Reasons**
Sex	3	673
Race	7	884
Religion	8	1,188
Nationality	34	328
Disability	2	231
total	*54*	*3,304*

Table 3: The distribution of social targets and annotated reasons within each social category in HatReD.

nationality), the type of attack (e.g., contempt), and the extracted web entities. Annotators have to identify and explain the hate speech with three primary goals: (i) the annotated reasons should specify and cover all implied hate speeches in the meme, (ii) the annotated reasons should accurately express and reflect the underlying hate implication, and (iii) the annotated reasons should be fluent and grammatically correct. We also ensure the annotated reasons are consistent across the annotators by requesting the reasons to be written in one of the two following Hearst-like patterns: (i) *<verb> <target> <predicate>* or (ii) *use of derogatory terms against <target> <predicate>*, where *<target>* represents the attacked social target and *<predicate>* highlights the hateful implication.

Annotation Quality Control
We conducted four trial annotations to ensure that the annotators were competent and proficient for the task. In each trial, 20 unique hateful memes are sampled for each annotator, and the annotators are tasked to craft the hateful reasons. At the end of each trial, the annotators will assess the quality of hateful reasons written by other annotators. The annotated reasons are evaluated based on the following criteria:

- *Fluency*: Rate the structural and grammatical correctness of the reasons using a 5-point Likert scale. 1: unreadable reasons with too many grammatical errors, 5: well-written reasons with no grammatical errors.

- *Relevance*: Rate the relevance of the reasons using a 5-point Likert scale. 1: reasons misrepresent the implied hate speech, 5: reasons accurately reflect the implied hate speech

At the end of each iteration, we present the evaluation ratings of their hateful reasons to the annotators and discuss how to improve the poorly rated hateful reasons. These discussions helped our annotators to improve the quality of their annotation.

3.2 Corpus Analysis

In total, HatReD dataset contains 3,304 annotated reasons for 3,228 hateful memes. Some memes may have multiple annotated reasons because they attack multiple social groups. The minimum explanation length is 5, the mean explanation length is 13.62, and the maximum is 31.

To examine the quality of the annotated reasons, we conducted human evaluation similar to our annotation trial on 1,200 hateful memes in our HatReD dataset. The annotators are tasked to evaluate the annotated reasons written by others. This translates into having three human evaluation results for each annotated reason. Table 2 shows the average *fluency* and *relevance* of the human evaluation on the hateful reasons for 1,200 memes. We observe a high average score of 4.97 and 4.81 for *fluency* and *relevance*, respectively, suggesting that the annotated reasons fluently capture the hateful speeches in the memes. Furthermore, we observe that the three evaluators have a unanimous agreement for *fluency* ratings in 93.9% of the evaluated annotated reasons, i.e., the evaluators rated the same score in 93.9% of the evaluated reasons. Similarly, the evaluators unanimously agree for their *relevance* ratings in 81.2% of the evaluated annotated reasons.

Table 3 illustrates the distribution of the social targets and hateful reasons found in the HatReD dataset. We observe significant variations in the number of social targets per social category. For example, there are 673 memes targeting the *Sex* social group, comprising three social targets (i.e., *LGBT*, *Female*, and *Males*). In contrast, the *Nationality* social group has 328 memes are attacking 34 unique social targets. Therefore, we expect more diverse and sparsely annotated reasons for hateful memes targeting *Nationality* social targets.

4 Hateful Memes Explanation

The availability of HatReD enables the exploration of a new task, *hateful meme explanation*. Specifically, we propose a natural language generation task where trained models generate the underlying reasons to explain hateful memes. Such generated reasons can help content moderators better understand the severity and nature of automatically-flagged hateful memes. Similar to [ElSherief *et al.*, 2021], our work can alert the users when they intend to share a particular meme flagged as "*hateful*" and explain the underlying reasons. This enables users to recognize the severity of the hateful meme and possibly reconsider their decision to post the meme.

4.1 Task Definition.

We formulate the hateful meme explanation task as a conditional generation task dependent on meme content. Formally, given a dataset of paired hateful memes and reasons

	train	test	total
#Hateful Memes	2,982	246	3,228

Table 4: HatReD Train-Test Split

$\{x^i, r^i\}_{i=1}^{N}$, the goal is to learn the generation of a fluent and relevant reason conditioned on the text information x_T^i and visual information x_V^i extracted from the hateful meme. We can refer to the reasons as a sequence of tokens $r^i = r_1^i, \cdots, r_\ell^i$, where we pad the tokens to a maximal length l. The training objective is defined as follows:

$$\max_\theta \sum_{i=1}^{N} \sum_{j=1}^{\ell} \log p_\theta(r_j^i | x_T^i, x_V^i, r_1^i, \cdots, r_{j-1}^i) \quad (1)$$

where θ denotes the model's trainable parameters.

4.2 Generative Models

A common model architecture used for conditional generation tasks is the encoder-decoder PLMs. Encoder-decoder PLM uses an encoder model to map the inputs to a sequence of continuous representations, which is then passed to the decoder to generate the output sequence. We train two types of PLMs in our experiments: (a) text-only PLMs that only accept text inputs; and (b) vision-language (VL) PLMs that accept text and visual inputs.

For data pre-processing, we obtain the text information x_T by tokenizing the text that overlays on the meme image. The input differences in the two types of encoder-decoder PLMs require the visual information x_V to be pre-processed differently. For text-only PLMs, we extracted the the meme's image caption using ClipCap [Mokady *et al.*, 2021]. In addition, we applied Google Vision Web Entity Detection API and Fairface classifier [Kärkkäinen and Joo, 2019] to extract the meme's entities and demographic information, respectively. Finally, we concatenate the image caption, extracted entities, and demographic information to represent the visual information. For VL PLMs, we used Detectron2 [Wu *et al.*, 2019] with bottom-up attention [Anderson *et al.*, 2018][4] to extract object regions and bounding boxes from the meme's image.

Model Training. We trained the PLMs on the annotated reasons in the HatReD dataset. As the hateful memes in HatReD may contain multiple reasons, we randomly sampled one reason for each hateful meme during training. The training objective is to minimize the cross-entropy loss:

$$L_{CE} = -\sum_{i=1}^{N} \sum_{j=1}^{\ell} \log p_\theta(r_j^i | x_T^i, x_V^i, r_1^i, \cdots, r_{j-1}^i) \quad (2)$$

Model Inference. Conditioned on the meme's text information x_T and visual information x_V, we generate three token sequences via two following decoding strategies: (i) **greedy decoding**, which generates a sequence by greedily selecting the most probable token at each time step; and (ii) **beam search**, which generates the most likely N token sequences at each time step and selecting the token sequence with the overall highest probability. We choose the token sequence with the highest overall score.

---

[4]https://github.com/airsplay/py-bottom-up-attention

Models		N-gram matching			Embedding-based		
Encoder	Decoder	BLEU	ROUGE-L	H. Mean	BERT-P	BERT-R	BERT-F
Text-Only	RoBERTa^{base}	GPT2^{base}	0.068	0.222	0.104	0.112	0.327	0.218
	RoBERTa^{base}	RoBERTa^{base}	0.177	0.389	0.243	**0.508**	0.453	**0.480**
	T5^{large}	T5^{large}	**0.190**	**0.392**	**0.256**	0.485	**0.473**	0.479
Vision-Language	VisualBERT	GPT2^{base}	0.065	0.219	0.100	0.100	0.342	0.219
	VisualBERT	RoBERTa^{base}	0.179	0.391	0.246	0.499	0.449	0.474
	VL-T5	VL-T5	0.180	0.378	0.244	0.472	0.409	0.446

Table 5: Automatic evaluation results of the PLMs' generated reasons on HatReD's test set. All metrics favor higher score and have a cap of 1. All results have a standard deviation of ≤ 0.03

Models		Metrics (Avg.)	
Encoder	Decoder	Fluency	Relevance
RoBERTa^{base}	GPT2^{base}	4.667	2.681
RoBERTa^{base}	RoBERTa^{base}	**4.874**	2.720
T5^{large}	T5^{large}	4.630	**3.112**
VisualBERT	GPT2^{base}	4.870	2.283
VisualBERT	RoBERTa^{base}	4.344	2.931
VL-T5	VL-T5	4.626	2.602
ground truth reasons		*4.937*	*4.352*

Table 6: Human evaluation results of the PLMs' generated reasons on HatReD's test set.

Models		Metrics (Avg.)	
Encoder	Decoder	Fluency	Relevance
T5^{large}	T5^{large}	**4.540**	1.850
VisualBERT	RoBERTa^{base}	3.990	**2.040**

Table 7: Human evaluation results of the PLMs' generated reasons on 50 hateful memes from MAMI.

5 Experiments

5.1 Experiment Settings

Baselines. To understand the challenges of our proposed hateful meme explanation generation task, we fine-tune and evaluate encoder-decoder PLMs using the HatReD dataset. For text-only PLMs, we use T5 [Raffel *et al.*, 2020], RoBERTa [Liu *et al.*, 2019], and GPT2 [Radford *et al.*, 2019]. As GPT2 is a decoder-only architecture, we adopt RoBERTa as its encoder. For VL PLMs, we benchmark [Cho *et al.*, 2021] and VisualBERT [Li *et al.*, 2019]. As VisualBERT is an encoder-only architecture, we utilized RoBERTa and GPT2 as its decoder in two different settings. The evaluation of these PLMs establishes the baselines for this new hateful meme explanation task.

Feature Space Alignment. To align feature spaces between encoders and decoders with different architectures, we place the models into a sequence-to-sequence model architecture with randomly initialized cross-attention layers added to each decoder block [Rothe *et al.*, 2020]. The training error, back-propagated through the cross-attention layer, fine-tunes the weights and aligns the models' feature spaces.

Evaluation Metrics. We perform both automatic and human evaluations on the baselines. We adopt metrics commonly used in natural language generation tasks for automatic evaluation: (1) N-gram matching for word similarity; and (2) embedding-based metric for semantic similarity. For n-gram matching metrics, we compute the average BLEU [Papineni *et al.*, 2002] and ROUGE-L [Lin, 2004]

scores. We also compute the harmonic mean of these two metrics. We compute the precision, recall, and F1 of the BERTScore [Zhang *et al.*, 2019] for embedding-based metric. For human evaluation, we recruit human evaluators to assess the generated reasons on two aspects: *fluency* and *relevance*. The human evaluators are tasked to rate the generated reasons on the Likert scales described in Section 3.1. We also mitigate positional bias by presenting the generated reasons in a scrambled order.

5.2 Experiments on HatReD

We fine-tune the baselines over ten random seeds using the HatReD training set and evaluate the baselines' ability to generate fluent and relevant reasons for the hateful memes in the HatReD test set. Table 4 shows the distribution of the train-test split.

Table 5 shows the average automatic evaluation of the generated reasons for hateful memes in HatReD. We observed that T5 outperforms other PLMs across most evaluation metrics. Nevertheless, the best-performing model still performed badly with low N-gram matching scores (i.e., H. Mean = 0.256) and moderate BERTScore. The results suggest that the generated reasons differ substantially from the ground-truth reasons. However, the generated reasons could still convey similar meanings. Therefore, we perform human evaluations to assess the acceptability of the generated reasons.

Table 6 shows the human evaluation results of the generated reasons and human-written reasons for hateful memes in HatReD. The results indicate that while most PLMs can generate fluent reasons, the generated reasons scored poorly in terms of *relevance*. Specifically, T5, which has the highest average *relevance* score, still had a significantly lower score compared to human-written ground-truth reasons. The superior performance of T5, a text-only model, over multimodal models can be attributed to its larger model size and the abun-

Hateful Meme	(a)	(b)	
Image Caption	woman lying on a bed with her hands on her head.	portrait of a senior man.	portrait of a senior **[white]** man.
T5	dehumanizes the females as less capable humans that are only good for household chores such as dishwashing as well as fulfilling the sexual needs of men	vilifies the muslim by suggesting that they are terrorists	vilifies the white by suggesting that they are terrorists
VisualBERT-RoBERTa	dehumanizes the females as less capable humans suited for household chores like dishwashing and dishwashing	vilifies the immigrants by suggesting that they are terrorists	
Ground Truth	dehumanizes the females as sexual objects as well as less capable beings only good for dishwashing.	ridicules the whites as terrorists by mocking the fact that the majority of shooters in the us are the whites.	

Table 8: Hateful memes from HatReD dataset with reasons generated by VisualBERT-RoBERTa and T5 models. The bracketed **[word]** in the image caption is a manual correction that explores the impact of having accurate and detailed image explanations in text-only models. The highlighted green and red words in the generated explanations outline the correct implications of hate and hallucinations (i.e. misinformation) present in the hateful memes, respectively.
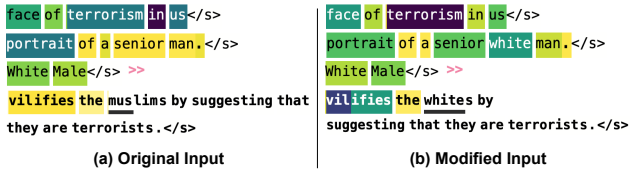

Figure 2: Input Saliency of T5 model on Meme 8b

dance of text data used for pretraining. T5 has approximately three times more parameters and is trained on 40 times more data than other models [Roberts *et al.*, 2020]. Nevertheless, the baseline PLMs' performance suggests the difficulty of the hateful meme reason generation task. We hope that the availability of HatReD enables researchers to design novel and better reason generation models for the task.

5.3 Experiment on Unseen Dataset

To further evaluate the usefulness of HatReD in explaining hateful memes, we conduct a domain adaption experiment where the baselines are trained on HatReD and tested on an unseen dataset. Specifically, we applied two top-performance baselines, namely, VisualBERT-RoBERTa and T5, to generate the reasons for hateful memes in the Multimedia Automatic Misogyny Identification (**MAMI**)[5] datasets [Fersini *et al.*, 2022]. The MAMI dataset contains hateful memes that discriminate against females. As there are no ground truth

reasons for the MAMI dataset, we perform a human evaluation of the generated reasons for 50 randomly sampled MAMI hateful memes.

Table 7 shows the human evaluation results on the generated reasons for the hateful memes in MAMI. Similarly, we observe that the baselines are able to generate fluent reasons. We also noted that the generated reasons have much lower *relevance* scores, which again suggests the difficulty of the hateful meme reason generation task. Nevertheless, we observe that 20% of the hateful misogynous memes are rated highly relevant (i.e., *relevancy* ≥ 4). The promising result demonstrates the possibility of performing domain adaption, where generative models are trained on HatReD to generate the reasons for hateful memes in unseen domains.

5.4 Case Studies

Besides performing quantitative evaluations on the hateful meme reason generation task, we also perform empirical analysis on the best performing PLM's (i.e. T5 and VisualBERT-RoBERTa) generated reasons for hateful memes in the HatReD and MAMI datasets. Specifically, for this analysis, we examine generated reasons with either high (i.e., ≥ 4) or low (< 2) relevance scores.

Table 8 presents examples of generated reasons for hateful memes in the HatReD. We found that both models generate reasons that accurately reflect the implied hate speech for the meme 8a. Notably, while both generated reasons reveal the intention to undermine and objectify women, only the reason generated by the text-only T5 model captures the intention to sexualize women. On the other hand, the

Hateful Meme			
	(a)	(b)	(c)
VisualBERT-RoBERTa	dehumanizes the females as less capable humans suited for household chores like cooking	dehumanizes the females by implying that their only purpose is to cook for men	mocks the females by suggesting they are inferior towards the white women
T5	dehumanizes the females as less capable humans that are only good for cooking, cleaning and making sandwiches	mocks the females by implying that they are lesser people who are only good for making food	disrespects the lgbt community by mocking transgender women and suggesting they are only good for sex

Table 9: Hateful memes from MAMI datasets with reasons generated by VisualBERT-RoBERTa and T5 models. The highlighted green and red words in the generated explanations outline the correct implications of hate and hallucinations (i.e. misinformation) present in the hateful memes, respectively.

T5 and VisualBERT-RoBERTa models generate reasons that misidentify the social target as Muslims and Immigrants for the meme 8b respectively. Examining the text-only T5 model that relies on extracted visual information, we observe that the image caption captures the essential visual information of a woman lying on a bed in meme 8a but fails to capture the crucial demographic information of the white man in meme 8b. The absence of demographic information in the image caption might be critical in associating the face of terrorism with white people, which led to the inaccurate identification of the social target in the generated reason for meme 8b. To examine this possibility, we make manual correction to the meme 8b's image caption and explore the input saliency for the generated social target. We found that manually correcting the image caption helped to generate a new reason that accurately identifies the targeted social target, shown in Figure 8. Additionally, in Figure 2, we explore and show the input saliency of the original and modified inputs via Integrated Gradients [Sundararajan *et al.*, 2017]. The results demonstrate that the manual correction caused the model to focus more on the words "face," "terrorism," "senior," "white," and "man," which are critical word associations required to identify the social target. These observations suggest the significance of having reliable visual information extractors to capture accurate visual information. As for the VisualBERT-RoBERTa model, the generation error is likely due to the model's inability to associate information from different modalities or understand detailed visual information such as the fact that the senior man is a white person. Nonetheless, these hallucinations demonstrate the limitation of state-of-the-art generation models and the potential for future improvement.

Table 9 showcases examples of generated reasons for hateful memes in the MAMI dataset. We observed that the generated reasons are often accurate for memes containing anti-feminism or patriarchal messages, as shown in meme 9a and 9b. This can be attributed to the high percentage of anti-feminism and patriarchy hateful memes in the training dataset, where approximately 26% of the HatReD's hateful memes in the female social category express anti-feminism and patriarchy messages. However, the generated reasons are found to be inaccurate when dealing with memes in new domains. For example, T5 model hallucinates that the meme 9c implies transgender women are sexual objects, despite no indication of this in the textual or visual modalities.

6 Conclusion

In this paper, we introduced HatReD, a new multimodal hateful memes dataset annotated with the underlying hateful contextual reasons. To the best of our knowledge, this is the first hateful meme dataset with written explanations. The availability of HatReD dataset opens the possibilities of training generative models to generate reasons for hateful memes, which can aid content moderators in understanding the severity and nature of flagged content. We defined a new conditional generation task to automatically explain hateful memes, and conducted extensive experiments using state-of-the-art PLMs to establish task baselines. Nevertheless, the quantitative and qualitative evaluations highlighted the difficulty of the hateful meme reason generation task. We hope that HatReD and our benchmark study will encourage more researchers to develop better models to generate fluent and relevant reasons for hateful memes. For future works, we aim to expand HatReD further to cover more domains of hateful memes. We will also explore different strategies to improve the existing reason generation model, such as using retrieval augmentation to incorporate explicit knowledge or improving the utilization of implicit knowledge in PLMs.

Ethical Statement

Research indicates that annotating hateful or offensive content can have negative effects. To protect our annotators, we establish three guidelines: 1) ensuring their acknowledgment of viewing potentially hateful content, 2) limiting weekly annotations and encouraging a lighter daily workload, and 3) advising them to stop if they feel overwhelmed. Finally, we regularly check in with annotators to ensure their well-being.

Another consideration is the usage of Facebook's hateful memes; users will have to agree with Facebook's usage agreement to gain access to the memes. The usage of Facebook's hateful memes in this study is in accordance with its usage agreement. Respecting Facebook's licenses on the memes, the HatReD dataset only contains the annotated reasons for the Facebook memes, but not the hateful memes; users will have to download the memes from the Facebook Hateful Meme challenge separately.

One of HatReD's goals is to train AI systems to provide detailed warnings that explains the hateful nature of the meme content, raise users' awareness and discourages its dissemination. Nevertheless, we acknowledge the potential for malicious users to reverse-engineer and create memes that go undetected (or misunderstood) by the HatReD-trained AI systems. This is *strongly discouraged*. In our paper, HatReD is utilized as training signals for post-hoc explanation generation (i.e., after the meme is flagged as hateful), not hateful meme detection. Researchers and platform providers should be cautious about including HatReD as training signals for hateful meme detection.

References

[Anderson *et al.*, 2018] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*, 2018.

[Cao *et al.*, 2022] Rui Cao, Roy Ka-Wei Lee, Wen-Haw Chong, and Jing Jiang. Prompting for multimodal hateful meme classification. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 321–332, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

[Cho *et al.*, 2021] Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. Unifying vision-and-language tasks via text generation. In *ICML*, 2021.

[ElSherief *et al.*, 2021] Mai ElSherief, Caleb Ziems, David Muchlinski, Vaishnavi Anupindi, Jordyn Seybolt, Munmun De Choudhury, and Diyi Yang. Latent hatred: A benchmark for understanding implicit hate speech. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 345–363, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[Fersini *et al.*, 2022] Elisabetta Fersini, Francesca Gasparini, Giulia Rizzi, Aurora Saibene, Berta Chulvi, Paolo Rosso, Alyssa Lees, and Jeffrey Sorensen. Semeval-2022 task 5: Multimedia automatic misogyny identification. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022). Association for Computational Linguistics*, 2022.

[Gasparini *et al.*, 2021] Francesca Gasparini, Giulia Rizzi, Aurora Saibene, and Elisabetta Fersini. Benchmark dataset of memes with text transcriptions for automatic detection of multi-modal misogynistic content. *arXiv preprint arXiv:2106.08409*, 2021.

[Hee *et al.*, 2022] Ming Shan Hee, Roy Ka-Wei Lee, and Wen-Haw Chong. On explaining multimodal hateful meme detection models. In *Proceedings of the ACM Web Conference 2022*, WWW '22, page 3651–3655, New York, NY, USA, 2022. Association for Computing Machinery.

[Kärkkäinen and Joo, 2019] Kimmo Kärkkäinen and Jungseock Joo. Fairface: Face attribute dataset for balanced race, gender, and age. *arXiv preprint arXiv:1908.04913*, 2019.

[Kiela *et al.*, 2020] Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. The hateful memes challenge: Detecting hate speech in multimodal memes. *Advances in Neural Information Processing Systems*, 33:2611–2624, 2020.

[Kiela *et al.*, 2021] Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Casey A. Fitzpatrick, Peter Bull, Greg Lipstein, Tony Nelli, Ron Zhu, Niklas Muennighoff, Riza Velioglu, Jewgeni Rose, Phillip Lippe, Nithin Holla, Shantanu Chandra, Santhosh Rajamanickam, Georgios Antoniou, Ekaterina Shutova, Helen Yannakoudakis, Vlad Sandulescu, Umut Ozertem, Patrick Pantel, Lucia Specia, and Devi Parikh. The hateful memes challenge: Competition report. In Hugo Jair Escalante and Katja Hofmann, editors, *Proceedings of the NeurIPS 2020 Competition and Demonstration Track*, volume 133 of *Proceedings of Machine Learning Research*, pages 344–360. PMLR, 06–12 Dec 2021.

[Lee *et al.*, 2021] Roy Ka-Wei Lee, Rui Cao, Ziqing Fan, Jing Jiang, and Wen-Haw Chong. Disentangling hate in online memes. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 5138–5147, 2021.

[Li *et al.*, 2019] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019.

[Lin, 2004] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics.

[Lippe *et al.*, 2020] Phillip Lippe, Nithin Holla, Shantanu Chandra, Santhosh Rajamanickam, Georgios Antoniou, Ekaterina Shutova, and Helen Yannakoudakis. A multimodal framework for the detection of hateful memes. *arXiv preprint arXiv:2012.12871*, 2020.

[Liu *et al.*, 2019] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[Mathias *et al.*, 2021] Lambert Mathias, Shaoliang Nie, Aida Mostafazadeh Davani, Douwe Kiela, Vinodkumar Prabhakaran, Bertie Vidgen, and Zeerak Waseem. Findings of the WOAH 5 shared task on fine grained hateful memes detection. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 201–206, Online, August 2021. Association for Computational Linguistics.

[Mokady *et al.*, 2021] Ron Mokady, Amir Hertz, and Amit H. Bermano. Clipcap: CLIP prefix for image captioning. *CoRR*, 2021.

[Muennighoff, 2020] Niklas Muennighoff. Vilio: State-of-the-art visio-linguistic models applied to hateful memes. *CoRR*, 2020.

[Papineni *et al.*, 2002] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics.

[Pramanick *et al.*, 2021a] Shraman Pramanick, Dimitar Dimitrov, Rituparna Mukherjee, Shivam Sharma, Md. Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. Detecting harmful memes and their targets. In *Findings of the Association for Computational Linguistics: ACL/IJCNLP*, pages 2783–2796, 2021.

[Pramanick *et al.*, 2021b] Shraman Pramanick, Shivam Sharma, Dimitar Dimitrov, Md. Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. MOMENTA: A multimodal framework for detecting harmful memes and their targets. In *Findings of the Association for Computational Linguistics: EMNLP*, pages 4439–4455, 2021.

[Radford *et al.*, 2019] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[Raffel *et al.*, 2020] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020.

[Roberts *et al.*, 2020] Adam Roberts, Colin Raffel, and Noam Shazeer. How much knowledge can you pack into the parameters of a language model? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5418–5426, Online, November 2020. Association for Computational Linguistics.

[Rothe *et al.*, 2020] Sascha Rothe, Shashi Narayan, and Aliaksei Severyn. Leveraging pre-trained checkpoints for sequence generation tasks. *Transactions of the Association for Computational Linguistics*, 8:264–280, 2020.

[Sap *et al.*, 2020] Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. Social bias frames: Reasoning about social and power implications of language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5477–5490, Online, July 2020. Association for Computational Linguistics.

[Sharma *et al.*, 2022a] Shivam Sharma, Md Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. Disarm: Detecting the victims targeted by harmful memes. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1572–1588, 2022.

[Sharma *et al.*, 2022b] Shivam Sharma, Firoj Alam, Md Akhtar, Dimitar Dimitrov, Giovanni Da San Martino, Hamed Firooz, Alon Halevy, Fabrizio Silvestri, Preslav Nakov, Tanmoy Chakraborty, et al. Detecting and understanding harmful memes: A survey. *arXiv preprint arXiv:2205.04274*, 2022.

[Sharma *et al.*, 2023] Shivam Sharma, Atharva Kulkarni, Tharun Suresh, Himanshi Mathur, Preslav Nakov, Md Akhtar, Tanmoy Chakraborty, et al. Characterizing the entities in harmful memes: Who is the hero, the villain, the victim? *arXiv preprint arXiv:2301.11219*, 2023.

[Sundararajan *et al.*, 2017] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR, 2017.

[Suryawanshi *et al.*, 2020] Shardul Suryawanshi, Bharathi Raja Chakravarthi, Mihael Arcan, and Paul Buitelaar. Multimodal meme dataset (multioff) for identifying offensive content in image and text. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 32–41, 2020.

[Wu *et al.*, 2019] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019. (accessed February 28, 2023).

[Yang *et al.*, 2022] Chuanpeng Yang, Fuqing Zhu, Guihua Liu, Jizhong Han, and Songlin Hu. Multimodal hate speech detection via cross-domain knowledge transfer. In *Proceedings of the 30th ACM International Conference on Multimedia*, MM '22, page 4505–4514, New York, NY, USA, 2022. Association for Computing Machinery.

[Zhang *et al.*, 2019] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019.