Who's Hated: Detecting and Analyzing the Entities Targeted by Hateful Memes

Zhaoxun Liu
Department of Computer Science
University of Toronto
lorenz@cs.toronto.edu

Abstract

Memes have proliferated rapidly online in recent years. Among them, however, hateful memes pose a significant threat to the well-being of online communities. Therefore, developing automated systems for the detection and analysis of hateful memes is crucial to mitigate their adverse impact; nonetheless, it is an intrinsically difficult and open problem: memes convey messages using both images and texts and, hence, require multimodal reasoning. While previous research has examined similar problems, they are quite limited; a holistic approach is lacking, particularly in terms of reasoning about the target entities. Moreover, there is little analysis that clarifies why certain entities are more susceptible, and no suggested measures have been put forth to specifically curb the dissemination of hateful memes. In this study, we aim to address these issues. Our contributions can be enumerated as (i) presenting a framework to detect and reason about entities targeted by hateful memes; (ii) providing insight into why certain groups are more susceptible to becoming targets of hateful memes; and (iii) proposing a specific preventive measure to curb the spread of hateful memes.

Disclaimer
This paper contains offensive or discriminatory contents from third parties that may be disturbing to some readers.

1 Introduction

The pervasive influence of social media platforms has given rise to a unique and potent form of multimodal expression: memes. Embodying ideas, reflections, or styles transmitted through cultural imitation, these succinct combinations of images and texts have become integral to online communication, spreading rapidly and widely, particularly on social media.

Unfortunately, while have emerged as a novel means of expressing benign or sarcastic humor, memes now suffer from widespread misuse, including the propagation of radical hatred and hostility (Brooke, 2019; Joksimovic et al., 2019; Zannettou et al., 2018). This misuse, termed "hateful memes," targets various groups based on attributes like race, religion, and gender, causing harm at individual and societal levels (Williams et al., 2016; Drakett et al., 2018; Sharma et al., 2022b). Recent reports highlight a significant surge in the circulation of hateful memes, contributing to the alarming trend of online hostility, affecting 41% of American adults (Duggan, 2017).

REASONINGS:
"mocks the asians for their small eyes and having to open them bigger to see better."

Figure 1: A sample of hateful memes that contains both visual and textual information; a ground-truth reason for the hateful nature is annotated.

Every hateful meme contains three main components, just like Figure 1 does: target entities, textual and/or visual messages, and underlying reasons for the hateful nature. Hence, the imperative, and also our research focus, lies in these questions: (i) How to develop an automated system to detect and analyze the entities targeted by hateful memes? (ii) How to find the underlying reasons why these entities are more susceptible to such targeting? (iii) How to formulate a preventive measure to alleviate the adverse consequences associated with hateful memes and foster the well-being of online communities?

To operationalize the research questions, we fine-tuned an encoder-decoder pre-trained language model (PLM), T5 (Raffel et al., 2023), using the HatReD dataset (Hee et al., 2023) to detect targeted entities in a given set of hateful memes and generate reasons for the hateful nature; this model was assessed with a user study. We then mapped entities to their corresponding frequencies of being targeted and extracted the entities most frequently targeted by normalizing the frequencies, obtaining a rank order of entities. Subsequently, we conducted a detailed analysis of the word frequencies within the generated reasons associated with the ranked entities, which is aimed at understanding the root reasons behind the entities' likelihood of being targeted. This comprehensive approach provides insights into the entities most susceptible to being targeted in hateful memes and, combined with our presented framework, we propose a specific preventive measure to mitigate the propagation of hateful memes.

2 Related Work

The intricate nature and often cryptic meanings of memes pose a significant challenge for analysis (Sabat et al., 2019). Facebook's "Hateful Memes Challenge" aimed to classify hateful memes, resulting in diverse multimodal deep learning approaches (Yang et al., 2022; Lee et al., 2021; Lippe et al., 2020); however, these are only binary classifiers inferring whether a meme is hateful. Studies also contributed datasets about hateful memes to support model training (Suryawanshi et al., 2020; Gasparini et al., 2022; Pramanick et al., 2021a; Sharma et al., 2022b, 2023).

While research has focused on classifying hateful memes, explaining predictions is crucial (Kiela et al., 2021). Recent analyses categorized attack types at a finer granularity (Mathias et al., 2021; Zia et al., 2021) and inferred targets (Pramanick et al., 2021a; Sharma et al., 2022b,a; Pramanick et al., 2021b). However, specifics (e.g., targeted ethnicities) are often overlooked. Elsherief et al. (2021) (ElSherief et al., 2021) curated a dataset with implied statements for content moderator understanding. Hee et al. (2023) introduced HatReD, a multimodal dataset with annotated contextual reasons (Hee et al., 2023).

There are also previous work done to prevent the spread of online hatred or cyberbullying (Chaudhary et al., 2021; Windisch et al., 2021, 2022; Tekiroglu et al., 2020; Cassidy et al., 2018) that can serve as valuable references for devising a comprehensive strategy to curb the rampancy of hateful memes.

3 Entities Detecting and Reasoning

3.1 Dataset

We constructed the HatRed dataset following the instructions by Hee et al. (2023) (Hee et al., 2023). The main idea of this dataset is to address the challenge of annotating explanations or reasons for hateful memes by leveraging the Google Web Detect API to extract web entities, providing annotators with sociocultural context from external knowledge bases, such as Wikipedia[1] and Hatebase[2], fostering a deeper understanding of cultural backgrounds and societal prejudices through iterative annotation.

Our constructed HatReD dataset comprises 3,304 annotated reasons corresponding to 3,228 hateful memes. Some memes may have multiple annotated reasons, as they target multiple entities. The minimum length of explanations is 5, the average explanation length is 13.62, and the maximum length is 31. A sample of the constructed dataset can be found at Appendix A. By incorporating annotated ground-truth reasons, a feature unprecedented in previous datasets, this dataset enables us to train models for generating explanations elucidating why a meme should be deemed hateful. Consequently, we can enhance our analysis of the associated entities.

3.2 Model Framework

We framed the task of reasoning hateful memes as a conditional generation task that relies on the meme content. Specifically, with a dataset containing pairs of hateful memes and their explanations, our objective was to learn the generation of a coherent and pertinent rationale. Formally, given textual information x^T and visual information x^V extracted from a hateful meme, we aimed to generate reasons, denoted as a sequence of tokens $r = r^1, \ldots, r^\ell$, where we pad the tokens to a maximal length ℓ.

Algorithm 1 Cross-Entropy Loss for the Hateful Memes Reasoning Task

```
L = 0
for i = 1 to N do
    for j = 1 to ℓ do
        t = log p_θ(r_ij | x_i^T, x_i^V, r_i^1, ..., r_i^j)
        L = L − t
    end for
end for
return L
```

Algorithm 1 is the cross-entropy loss function for the hateful memes reasoning task, where L is the loss, N is the number of hateful memes in the dataset, ℓ is the maximum length of the reasons, x_i^T and x_i^V are the textual and visual information of the i-th meme, r_{ij} is the j-th token of the reason for the i-th meme, p_θ is the probability function of the generative model with parameters θ; overall, $p_\theta(r_{ij}|x_i^T, x_i^V, r_i^1, \ldots, r_i^j)$ is the

[1] Wikipedia, https://en.wikipedia.org/

[2] Hatebase, https://hatebase.org/

probability of generating the j-th token of the i-th reason, given the textual and visual information of the i-th meme and the previous tokens of the i-th reason. It measures the difference between the probability distribution of the generated reasons and the ground truth reasons. A lower cross-entropy loss means that the generated reasons are more similar to the ground truth reasons.

In the realm of conditional generation tasks, a prevalent model architecture is the encoder-decoder PLM. This architecture employs an encoder model to translate inputs into a sequence of continuous representations, subsequently utilized by the decoder to generate the output sequence. Our idea was to convert visual information to textual information so that we could simplify our model to single modality, focusing solely on text. Therefore, we selected T5 (Raffel et al., 2023) to fine-tune because it has proven its significance in similar text-only tasks (Wei et al., 2022; Pilault et al., 2022; Liu et al., 2021; Hee et al., 2023).

As for data pre-processing, the acquisition of text information x^T involves tokenizing the text overlaying the meme image. Notably, the distinct input requirements of the two encoder-decoder PLM types necessitate varied pre-processing for visual information x^V. To grasp the textual context, we employed ClipCap (Mokady et al., 2021) to extract the image caption of the meme, converting visual information into textual format. Furthermore, we leveraged the Google Vision Web Entity Detection API and FairFace classifier (Kärkkäinen and Joo, 2019) to extract the meme's entities and demographic information, respectively. Additionally, inspired by the work on vision-language PLMs by Hee et al. (Hee et al., 2023), we also implemented a comprehensive pre-processing strategy using advanced tools and methodologies to our text-only setup. To extract object regions and bounding boxes from the meme's image, we employed Detectron2 (Wu et al., 2019) integrated with bottom-up attention (Anderson et al., 2018).

During the inference stage of our model, we leveraged both the meme's textual information x^T and visual information x^V to generate candidate token sequences using two decoding strategies. Firstly, the greedy decoding strategy constructs a sequence by selecting the most probable token at each time step, prioritizing immediate probability maximization; secondly, employing beam search, the model generates the most likely N token sequences at each time step and subsequently selects the sequence with the highest cumulative probability. The token sequence with the highest overall score is chosen as the final output.

3.3 Model Evaluation

As previously mentioned in Section 3.2, given the established state-of-the-art performance of T5 in similar text-only tasks, we could bypass automated evaluations, such as N-gram matching, and proceeded directly to human evaluation.

Similar to the evaluation conducted by Hee et al. (Hee et al., 2023) on the explanation annotations for the HatRed dataset, we tasked our participants to rate the generated reasons with these two subjective metrics:

1. *Fluency*: Evaluate the structural and grammatical accuracy of the statements using a 5-point Likert scale. Rate a score of 1 to denote unreadable statements, and a score of 5 for well-articulated statements.
2. *Relevance*: Assess the relevance of the statements using a 5-point Likert scale. Rate a score of 1 if the statements completely distort the hateful meaning, and a score of 5 if the statements precisely capture the essence.

Before the evaluation, we recruited 20 student participants (10 males, 10 females) with an average age of 21.76 ($\sigma^2 = 1.92$). All participants had received undergraduate or higher education, were capable of reading English without difficulty, and did not have any disabilities.

The user study was designed as follows: When a reason is generated in real-time for a randomly chosen meme from the test set of our constructed HatRed dataset, the participant is then asked to rate each reason for its *Fluency* and *Relevance*. This process loops for 20 times. You can find the script of our questionnaire at Appendix B.

The result of this user study [$t(19) = 0.23, p > 0.05$] is reported in Table 1, accepting the null hypothesis that there is no significant difference between fluency and relevance scores ($\mu_{fluency} = \mu_{relevance}$). Both mean ratings were greater than 4, indicating that the generated reasons are sufficiently credible. Therefore, we could collect reasons data using the automatic generation capabilities of our fine-tuned T5 model for later analysis. Examples of reasons generated with high and low ratings can be located in Table 2.

Metric	**Mean**	**Paired-Sample T-Test**
Fluency	4.15	$t(19) = 0.23, p = 0.82$
Relevance	4.05	

Table 1: Results of the user study.

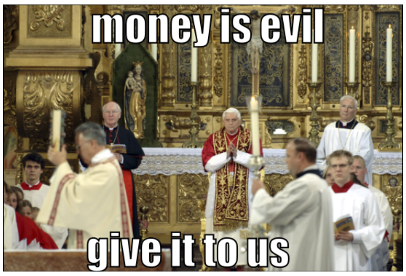
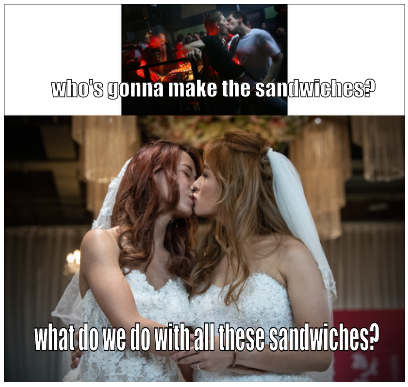
Meme		
Generated Reason	Mocks Christians who collect money despite preaching that money is sinful.	Mocks people within the LGBTQ community their sexual orientations.
Ground Truth	mocks the christians for collecting money when they preach the belief that money is evil.	insults the females by degrading them by playing on a stereotype that women belong in the kitchen and should be subservient to men.
Fluency	5	3
Relevance	5	1

Table 2: Generated reasons for two memes with high and low *Fluency* and *Relevance* ratings.

4 Analysis of Targeted Entities

4.1 Data Collection and Preparation

We utilized our fine-tuned T5 model to generate reasons for all memes in the Facebook Hateful Memes Challenge dataset[3] and created a mapping of targeted entities to their frequencies and generated reasons.

Let E represent the set of targeted entities, F denote the set of corresponding frequencies, and R signify the set of generated reasons. The fine-tuned model provides a mapping M:

$$M : E \to (F, R) \tag{1}$$

To identify the entities most frequently targeted, we normalized the frequencies to scale and standardize numerical values, ensuring that they fall within a consistent and comparable range, and obtained a rank-ordering denoted by $\text{Rank}(E)$; Table 3 shows the top 5 ranks. For the i-th entity, the corresponding $\text{Rank}(E_i)$ can be expressed as:

$$\text{Rank}(E_i) = \text{Normalize}(F_i) = \frac{f_i}{\sum_{j=1}^{n} f_j} \tag{2}$$

Subsequently, we generated word clouds that illustrate the frequencies of keywords in the reasons associated with the ranked entities. Let $W(R_E)$ denote the word cloud tied to the entity E. Examples of $W(R_E)$ can be found in Figure 2.

[3]Facebook Hateful Memes Challenge, https://hatefulmemeschallenge.com/

4.2 Findings

Upon analyzing the $\text{Rank}(E)$, we discovered that the entities with the highest susceptibility tend to cluster in three domains: **race, religion, and politics**. Within the race domain, susceptibility is most pronounced among Jews, Middle Easterners, African Americans, and Asians. In the realm of religion, Muslims and Christians exhibit the highest susceptibility. Concerning politics, susceptibility is notably elevated among Feminists, Republicans, Democrats, and Communists. Upon examining $W(R_E)$, we summarize that the root cause of why these entities are more susceptible to

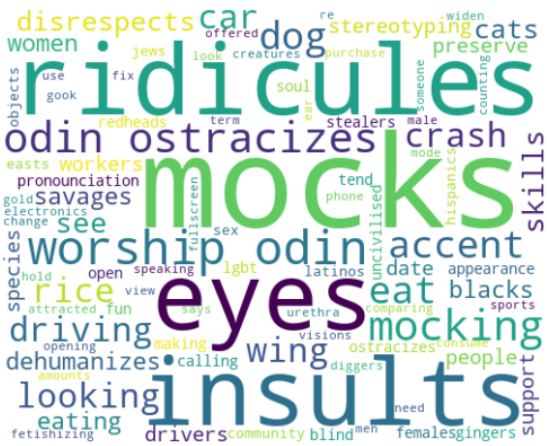


Figure 2: A word cloud of the reasoning keywords with "East Asians" as the targeted entity.

Rank	0.143	0.138	0.101	0.097	0.093
Entity	Jews	Middle Easterners	Muslims	Feminists	African Americans

Table 3: Top 5 detected entities with the highest ranks.

becoming targets of hateful memes lies within these three domains:

Historical Context: Entity hatred's roots lie in the lasting impact of colonialism, slavery, and imperialism, with arbitrary classifications during colonial expansion setting the stage for enduring biases. For example, the transatlantic slave trade and colonial era (Rawley and Behrendt, 2005) entrenched stereotypes of inferiority for African descent communities.

Socio-economic Influences: Economic disparities, limited resource access, and institutionalized discrimination shape perceptions of certain entities. Historical redlining in the U.S. created enduring wealth disparities for African Americans (Thompson and Suarez, 2019), reinforcing stereotypes linking them to poverty and crime. Discriminatory policies like the 19th-century Chinese Exclusion Act targeted Chinese individuals (Chinn, 2016), perpetuating negative stereotypes and marginalization.

Media Influence: Media, a powerful influencer, shapes perceptions and reinforces stereotypes for certain entities. Historical casting practices in Hollywood confined certain racial groups to stereotypical roles (Yuen, 2019), contributing to harmful biases and shaping public perceptions.

5 Prevention of Hateful Memes

In recent studies addressing the prevention of online hatred and cyberbullying, the majority of efforts are directed toward generating counter-narratives to combat hate speech (Tekiroglu et al., 2020). Additionally, some research conducts a comprehensive evaluation of online interventions and their effectiveness (Windisch et al., 2021, 2022; Chaudhary et al., 2021). However, there hasn't been a specifically proposed measure to effectively reduce the spread of hateful memes. As mentioned before, unlike text-based hate speech and verbal cyberbullying, hateful memes are multimodal. Consequently, conventional measures designed for preventing text-based content prove relatively ineffective in addressing the issue of hateful memes, hence their continued rampancy.

We could address this issue by utilizing our fine-tuned T5 model to an automated censoring filter to mitigate the dissemination of hateful memes. Initially, a binary classifier, such as the winning models from the Facebook Hateful Memes Challenge (Zhu, 2020; Muennighoff, 2020; Velioglu and Rose, 2020), is employed to ascertain whether a given meme image conveys hateful content. Subsequently, in positive cases, a detailed analysis of the meme is conducted to elucidate the targeted entities and reasons underlying its hateful nature with our T5 model. This analysis aims to facilitate a more informed response, encouraging prompt content modification.

To evaluate the efficacy of the filter, we utilized Zhu's binary classifier as outlined in their work (Zhu, 2020) and used the MemeCap dataset (Hwang and Shwartz, 2023). The MemeCap dataset, originally created for meme captioning, encompasses a collection of both hateful and non-hateful memes. We conducted this evaluation on the 5,828 meme images in MemeCap; the results yielded a false negative ratio of 0.28% and a false positive ratio of 13.21%, which is a rather secure and conservative strategy to prevent the dissemination of hateful memes.

6 Conclusion

We presented a multimodal framework for detecting and analyzing the entities targeted by hateful memes. We demonstrated the effectiveness of our framework in identifying targeted entities and generating the underlying reasons for the hateful nature; then, we provided insights into the root causal reasons why certain entities are more susceptible of being targeted. We also proposed a preventive measure to curb the rampancy of hateful memes using an automated censoring filter. We hope that our work can contribute to the development of more advanced and comprehensive systems for combating hateful memes and fostering a healthier online environment.

Our framework also has some limitations that need to be addressed in future work. First, our T5 model was trained on a relatively small dataset of hateful memes and their annotated reasons, which may limit its generalization ability and diversity of outputs. Second, our analysis of targeted entities and reasons was based on word frequencies, which may not capture the nuances and subtleties of the hateful messages. Third, our censoring filter relied on a binary classifier that may not be robust to adversarial attacks or new forms of hateful memes.

Acknowledgments

This work is the final project of CSC2552 - Topics in Computational Social Science: AI, Data, and Society, Fall 2023, University of Toronto.

References

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

Sian Brooke. 2019. "condescending, rude, assholes": Framing gender and hostility on Stack Overflow. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 172–180, Florence, Italy. Association for Computational Linguistics.

Wanda Cassidy, Chantal Faucher, and Margaret Jackson. 2018. What parents can do to prevent cyberbullying: Students' and educators' perspectives. *Social Sciences*, 7(12).

Mudit Chaudhary, Chandni Saxena, and Helen Meng. 2021. Countering online hate speech: An nlp perspective.

Stuart Chinn. 2016. Trump and chinese exclusion: Contemporary parallels with legislative debates over the chinese exclusion act of 1882. *Tenn. L. Rev.*, 84:681.

Jessica Drakett, Bridgette Rickett, Katy Day, and Kate Milnes. 2018. Old jokes, new media – online sexism and constructions of gender in internet memes. *Feminism & Psychology*, 28(1):109–127.

Maeve Duggan. 2017. Men, women experience and view online harassment differently. Accessed: November 12, 2023.

Mai ElSherief, Caleb Ziems, David Muchlinski, Vaishnavi Anupindi, Jordyn Seybolt, Munmun De Choudhury, and Diyi Yang. 2021. 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. Association for Computational Linguistics.

Francesca Gasparini, Giulia Rizzi, Aurora Saibene, and Elisabetta Fersini. 2022. Benchmark dataset of memes with text transcriptions for automatic detection of multimodal misogynistic content. *Data in Brief*, 44:108526.

Ming Shan Hee, Wen-Haw Chong, and Roy Ka-Wei Lee. 2023. Decoding the underlying meaning of multimodal hateful memes.

EunJeong Hwang and Vered Shwartz. 2023. Memecap: A dataset for captioning and interpreting memes.

Srecko Joksimovic, Ryan S. Baker, Jaclyn Ocumpaugh, Juan Miguel L. Andres, Ivan Tot, Elle Yuan Wang, and Shane Dawson. 2019. Automated identification of verbally abusive behaviors in online discussions. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 36–45, Florence, Italy. Association for Computational Linguistics.

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. 2021. The hateful memes challenge: Competition report. In *Proceedings of the NeurIPS 2020 Competition and Demonstration Track*, volume 133 of *Proceedings of Machine Learning Research*, pages 344–360. PMLR.

Kimmo Kärkkäinen and Jungseock Joo. 2019. Fairface: Face attribute dataset for balanced race, gender, and age.

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

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.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing.

Lambert Mathias, Shaoliang Nie, Aida Mostafazadeh Davani, Douwe Kiela, Vinodkumar Prabhakaran, Bertie Vidgen, and Zeerak Waseem. 2021. 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. Association for Computational Linguistics.

Ron Mokady, Amir Hertz, and Amit H. Bermano. 2021. Clipcap: Clip prefix for image captioning.

Niklas Muennighoff. 2020. Vilio: State-of-the-art visiolinguistic models applied to hateful memes.

Jonathan Pilault, Amine Elhattami, and Christopher Pal. 2022. Conditionally adaptive multi-task learning: Improving transfer learning in nlp using fewer parameters less data.

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

Shraman Pramanick, Shivam Sharma, Dimitar Dimitrov, Md Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. 2021b. Momenta: A multimodal framework for detecting harmful memes and their targets.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2023. Exploring the limits of transfer learning with a unified text-to-text transformer.

James A Rawley and Stephen D Behrendt. 2005. *The transatlantic slave trade: a history*. U of Nebraska Press.

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

Shivam Sharma, Md. Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. 2022a. Disarm: Detecting the victims targeted by harmful memes.

Shivam Sharma, Firoj Alam, Md. Shad Akhtar, Dimitar Dimitrov, Giovanni Da San Martino, Hamed Firooz, Alon Halevy, Fabrizio Silvestri, Preslav Nakov, and Tanmoy Chakraborty. 2022b. Detecting and understanding harmful memes: A survey.

Shivam Sharma, Atharva Kulkarni, Tharun Suresh, Himanshi Mathur, Preslav Nakov, Md. Shad Akhtar, and Tanmoy Chakraborty. 2023. Characterizing the entities in harmful memes: Who is the hero, the villain, the victim?

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*, pages 32–41, Marseille, France. European Language Resources Association (ELRA).

Serra Sinem Tekiroglu, Yi-Ling Chung, and Marco Guerini. 2020. Generating counter narratives against online hate speech: Data and strategies.

Jeffrey P Thompson and Gustavo Suarez. 2019. Accounting for racial wealth disparities in the united states.

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

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022. Finetuned language models are zero-shot learners.

Amanda Williams, Clio Oliver, Katherine Aumer, and Chanel Meyers. 2016. Racial microaggressions and perceptions of internet memes. *Comput. Hum. Behav.*, 63(C):424–432.

Steven Windisch, Susann Wiedlitzka, and Ajima Olaghere. 2021. Protocol: Online interventions for reducing hate speech and cyberhate: A systematic review. *Campbell Systematic Reviews*, 17(1).

Steven Windisch, Susann Wiedlitzka, Ajima Olaghere, and Elizabeth Jenaway. 2022. Online interventions for reducing hate speech and cyberhate: A systematic review. *Campbell systematic reviews*, 18(2):e1243.

Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. 2019. Detectron2. `https://github.com/facebookresearch/detectron2`.

Chuanpeng Yang, Fuqing Zhu, Guihua Liu, Jizhong Han, and Songlin Hu. 2022. 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. Association for Computing Machinery.

Nancy Wang Yuen. 2019. *Reel inequality: Hollywood actors and racism*. Rutgers University Press.

Savvas Zannettou, Tristan Caulfield, Jeremy Blackburn, Emiliano De Cristofaro, Michael Sirivianos, Gianluca Stringhini, and Guillermo Suarez-Tangil. 2018. On the origins of memes by means of fringe web communities.

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

Haris Bin Zia, Ignacio Castro, and Gareth Tyson. 2021. Racist or sexist meme? classifying memes beyond hateful. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 215–219, Online. Association for Computational Linguistics.

A Appendix

A sample of the constructed HatRed dataset, where *02169.png* is exactly the image shown in Figure 1:

id	2169
img	02169.png
target	the asians
reasonings	mocks the asians for their small eyes and having to open them bigger to see better.
race	East Asian Male
entity	east asian man shorthair healthy
text	to see better, asians sometimes switch to fullscreen view
gold_hate	hateful
gold_pc	[race]
gold_attack	[inferiority]
pc	[[race], [race], [race]]
attack	[[inferiority], [inferiority], [inferiority]]

B Appendix

Each participant was first presented with the current meme image along with the corresponding ground-truth reason annotated in the test dataset, then two questions were asked:

- **Q1**: How would you rate the *Fluency* of this generated reason? From 1 for unreadable to 5 for well-articulated.
- **Q2**: How would you rate the *Relevance* of this generated reason? From 1 for a complete distortion to 5 for a precise capture of the essence.