# Findings of the WOAH 5 Shared Task on Fine Grained Hateful Memes Detection

**Lambert Mathias**[†], **Shaoliang Nie**[†], **Aida Davani**[‡],

**Douwe Kiela**[†], **Vinodkumar Prabhakaran**[⋆] **Bertie Vidgen**[§], **Zeerak Waseem**[¶]

[†]Facebook AI Research; [‡]University of Southern California; [⋆]Google AI
[§]The Alan Turing Institute; [¶]University of Sheffield
mathiasl@fb.com

## 1 Abstract

We present the results and main findings of the shared task at WOAH 5 on hateful memes detection. The task include two subtasks relating to distinct challenges in the fine-grained detection of hateful memes: (1) the protected category attacked by the meme and (2) the attack type. 3 teams submitted system description papers. This shared task builds on the hateful memes detection task created by Facebook AI Research in 2020.

## 2 Introduction

The spread and impact of online hate is a growing concern across societies, and increasingly there is consensus that social media companies must do more to counter such content (League, 2020; Vidgen et al., 2021). At the same time, any interventions must be balanced with protecting people's freedom of expression and ability to engage in open discussions. Ensuring that online spaces are both open and safe requires being able to reliably and accurately find, rate and remove harmful content such as hate. Scalable machine learning based solutions offer a powerful way of solving this problem, reducing the burden on human moderators.

To date, detecting online hate has proven remarkably difficult and concerns have been raised about the performance, robustness, generalizability and fairness of even state-of-the-art models (Waseem et al., 2018; Vidgen et al., 2019; Caselli et al., 2020b; Mishra et al., 2019; Davidson et al., 2019). To advance the field, and develop models which can be used in real-world settings, research needs to go beyond simple binary classifications of textual content. To this end, we have used trained professional moderators to reannotate the hateful memes dataset from (Kiela et al., 2020)[1]. It contains two sets of labels, which correspond to our two sub-tasks: the protected category that has been attacked (e.g., women, black people, immigrants) as well as the type of attack (e.g., inciting violence, dehumanizing, mocking the group).

Detecting hateful memes is a particularly challenging task because the content is multi-modal rather than uni-modal, such as text or images alone. When humans look at memes they do not think about the words and photos independently but, instead, combine the two together. In contrast, most AI detection systems analyze text and image separately and do not learn a joint representation. This is inefficient and limits the performance of systems. They are likely to fail when an image that by itself is non-hateful is combined with non-hateful text to produce content that expresses hate through the *interaction* of the image and text. For AI to detect hate communicated through multiple modalities, it must learn to understand content the way that people do: holistically. In this paper we present the results of the WOAH 5 shared task on fine-grained hateful memes detection.

## 3 Dataset

### 3.1 Dataset Size

The dataset we present for the shared task is from phase 1 of the hateful memes challenge Kiela et al. (2020)[2]. Table 1 shows the distribution and data splits associated with the released dataset. We reannotated the hateful memes for the two fine-grained categories (Protected category and Attack type). For the non-hateful memes we assigned a label of 'none' for both categories.

[1] Dataset is available at https://github.com/facebookresearch/fine_grained_hateful_memes

[2] Dataset is available at https://hatefulmemeschallenge.com/

*Proceedings of the Fifth Workshop on Online Abuse and Harms*, pages 201–206
August 6, 2021. ©2021 Association for Computational Linguistics

| label | train | dev_seen | dev_unseen | test_seen |
|---|---|---|---|---|
| not_hateful | 5493 | 254 | 341 | 520 |
| hateful | 3007 | 246 | 199 | 480 |
| Total | 8500 | 500 | 540 | 1000 |

Table 1: Hateful Memes Dataset Statistics

### 3.2 Dataset Labels

Each meme was originally labelled as 'Hateful' or 'Not Hateful' by Kiela et al. (2020). Hate is a contested concept and there is no generally agreed upon definition or taxonomy in the field (Caselli et al., 2020a; Waseem et al., 2017; Zampieri et al., 2019). For the purposes of this work, hate is defined as a direct attack against people based on 'protected characteristics'[3]. Protected characteristics are core aspects of a person's social identity which are generally fixed or immutable. Table 2 provides the set of fine-grained labels for protected classes and attack types.

### 3.3 Annotations

Each hateful meme was annotated by three annotators for the protected characteristic and the attack type (from the set defined in Table 2). If no clear protected group or attack type could be identified the annotator could select "not sure". Annotators were allowed to select multiple labels for both the protected characteristic and attack type.

Since our annotation is multi-label, we computed Krippendorff's $\alpha$, which supports multiple annotators as well as multi-label agreement computation (Krippendorff, 2018). We obtain Krippendorff's $\alpha = 0.77$ for the protected categories, and $\alpha = 0.66$ for attack types, indicating that while there is some uncertainty, it is within usable range i.e $\alpha \geq 0.66$ (Krippendorff, 2004). This indicates 'moderate' to 'strong' agreement (Mchugh, 2012) and compares favourably with other abusive content datasets (Gomez et al., 2020; Fortuna and Nunes, 2018; Wulczyn et al., 2017), especially given that our labels contain five and seven levels respectively. We used a majority voting scheme to decide the final labels from the annotations.

## 4 Shared Task Results & Analysis

### 4.1 Shared Task Setup

For WOAH 5, collocated with ACL, we introduced two hateful meme detection tasks:

**Task A: Protected Category** For each meme, detect the protected category. The protected categories recorded in the dataset are: race, disability, religion, nationality, sex.[4] If the meme is not hateful the protected category is recorded as "pc_empty".

**Task B: Attack Type** For each meme, detect the attack type. The attack types recorded in the dataset are: contempt, mocking, inferiority, slurs, exclusion, dehumanizing, inciting violence. If the meme is not hateful the attack type is recorded as "attack_empty".

Tasks A and B are multi-label because each meme can contain attacks against multiple protected categories and can involve multiple attack types. For evaluating performance on both tasks we use the standard ROC_AUC metric for multi-label classification (Pedregosa et al., 2011).[5].

We used the same splits from the original dataset as described in Table 1. Participants had access to the train, dev_seen and dev_unseen splits for developing and tuning their models. The final evaluation was done on the test_seen split. The ground truth labels were not provided at time of submission and each participant was expected to submit their predictions with model scores. Each participant was limited to a maximum of 2 submissions per task.

### 4.2 System Descriptions

**Majority Baseline** A simple majority decision-rule, applied over the entire dataset. We predict the majority class for all instances, i.e. "pc_empty" for Task A and "attack_empty" for Task B.

**VisualBERT Baseline** A VisualBERT multimodal model (Li et al., 2019) that has been pre-trained on the MS COCO

[3] This aligns with the definition described in https://www.facebook.com/communitystandards/hate_speech

[4] Note that the characterisation and definition of some protected categories, such as race, is highly contested. For further analysis of the concept of 'race' see Omi and Winant (2005)

[5] The evaluation script and fine-grained labels are available at https://github.com/facebookresearch/fine_grained_hateful_memes

| **Protected Category** | **Definition** |
|---|---|
| Religion | A group defined by a shared belief system |
| Race | A group defined by similar, distinct racialised physical characteristics |
| Sex | A group defined by their physical sexual attributes or sexual identifications |
| Nationality | A group defined by the country/region they belong to |
| Disability | A group defined by conditions that generally lead to permanent dependencies (on people, medical treatments or equipment) |
| **Attack Type** | **Definition** |
| Dehumanizing | Explicitly or implicitly describing or presenting a group as subhuman |
| Inferiority | Claiming that a group is inferior, less worthy<br>or less important than either society in general or another group |
| Inciting violence | Explicitly or implicitly calling for harm to be<br>inflicted on a group, including physical attacks |
| Mocking | Making jokes about, undermining, belittling, or disparaging a group |
| Contempt | Expressing intensely negative feelings or emotions about a group |
| Slurs | Using prejudicial terms to refer to, describe or characterise a group |
| Exclusion | Advocating, planning or justifying the exclusion or<br>segregation of a group from all of society or certain parts |

Table 2: Protected Category and Attack Type definitions used for fine-grained annotations.

| **Fine-grained attributes** | | **train** | **dev_unseen** | **dev_seen** | **test_seen** |
|---|---|---|---|---|---|
| Attack type | dehumanizing | 1318 | 104 | 121 | 209 |
| | inferiority | 658 | 35 | 49 | 102 |
| | inciting_violence | 407 | 23 | 26 | 68 |
| | mocking | 378 | 29 | 35 | 84 |
| | contempt | 235 | 6 | 10 | 21 |
| | slurs | 205 | 4 | 6 | 10 |
| | exclusion | 114 | 8 | 13 | 12 |
| Protected category | religion | 1078 | 77 | 95 | 166 |
| | race | 1008 | 63 | 78 | 169 |
| | sex | 746 | 46 | 56 | 82 |
| | nationality | 325 | 20 | 26 | 42 |
| | disability | 255 | 17 | 22 | 63 |

Table 3: Distribution of attack types and protected characteristics on the "hateful" subset of the hateful memes dataset in Table 1

| **System** | **Task A - protected category** | **Task B - attack type** |
|---|---|---|
| Majority Baseline | 0.70 | 0.72 |
| VisualBERT Baseline | 0.864 | 0.873 |
| LTL-UDE1 | 0.912 | - |
| LTL-UDE2 | **0.914** | - |
| QMUL | 0.901 | **0.913** |
| SU1 | 0.876 | 0.881 |
| SU2 | 0.865 | 0.89 |

Table 4: Overall results from the shared task submissions on the blind test set partition

dataset (Lin et al., 2014). We use the setup in MMF (Singh et al., 2020) to pre-train the models. Each task is trained and evaluated independently.[6] VisualBERT was also used in the original hateful memes paper by Kiela et al. (2020), although here we set it up for multilabel detection.

**Duisburg-Essen System 1 (LTL-UDE1)** The solution builds on the multimodal approach used for the winning entry in the hateful memes challenge (Zhu, 2020) - a VLBERT multimodal model with image specific metadata. It was fine-tuned on the fine-grained data. The system was only submitted for Task A.

**Duisburg-Essen System 2 (LTL-UDE2)** An additional emotion tags are added to DE1 which are extracted from the facial expressions of persons objects available in the meme image. The system was only submitted for Task A.

**Queen Mary University London (QMUL)** The submitted system is a multimodal model that uses CLIP (Radford et al., 2021) image encoder to embed the meme images, and CLIP text encoder, LASER (Artetxe and Schwenk, 2019) & LaBSE (Feng et al., 2020) to embed the meme text. All the representations are concatenated, and a multi-label logistic regression classifier is trained, one for each task, to predict the labels.

**Stockholm University System 1 (SU1)** A BERT-base based model that only uses the text of the meme as input. The BERT model was fine-tuned independently for each task.

**Stockholm University System 2 (SU2)** A multimodal model (ImgBERT) which combines SU1 with image embeddings. The image embeddings were extracted using DenseNet-121 convolutional neural networks(CNNs), pretrained on ImageNet (Deng et al., 2009). The input to the multi-label classification layer is the concatenation of the text representation from the [CLS] token of SU1, and the image embedding. The final classifier is an ensemble between the ImgBERT model and the text-only model from SU1. The scores provided by each of the labels were averaged to decide the final label.

### 4.3 Analysis

Table 4 shows the performance on the 2 tasks across all the participants. All the systems used some variant of pre-trained multimodal representations fine-tuned on the shared task datasets. None of the submissions exploited the correlation across all the tasks, and instead trained the systems independently on each of the tasks. The systems from LTL-DE1 and LTL-DE2 were the only ones to exploit image level metadata as an additional signal that was not part of the provided training data that showed best performance on Task A. Moreover, the LTL-DE1 and LTL-DE2 submissions were the only ones to leverage state of the art multimodal representations from VLBERT (Su et al., 2019), while all other submissions encoded the image and text channel independently. Interestingly, SU1, which is a text BERT system finetuned on the tasks performed remarkably strongly, even outperforming their multimodal system and the provided baselines. It is unclear if the model is picking up some unintended biases in the data, considering the relatively small size of the datasets provided for the shared task. QMUL system encoded the text representation using multiple different pre-trained representations concatenated with the image representation, further supporting the evidence that potentially stronger encoding of text might be sufficient to achieve strong performance on this dataset.

## 5 Conclusion

Detecting hate remains technically difficult, with many unaddressed or unsolved challenges and frontiers. Hateful memes are one issue that has received little attention, despite the ubiquity of such media online. The shared task at WOAH 5, with two subtasks for fine-grained detection of the protected category and the attack type, is another step forward in this still-nascent research area.

For future work, we hope to scale the fine-grained annotations to other hate speech datasets, as we think it is important toi develop classifiers that can detect the nuances of hate speech. Meanwhile, the annotated datasets are publicly available and we welcome researchers to make use of them.

[6] See `https://github.com/facebookresearch/mmf/tree/master/projects/hateful_memes/fine_grained` for training configuration

## References

Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

Tommaso Caselli, Valerio Basile, Jelena Mitrović, Inga Kartoziya, and Michael Granitzer. 2020a. I Feel Offended, Don't Be Abusive! Implicit/Explicit Messages in Offensive and Abusive Language. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC)*, pages 6193–6202.

Tommaso Caselli, Valerio Basile, Jelena Mitrović, Inga Kartoziya, and Michael Granitzer. 2020b. I feel offended, don't be abusive! implicit/explicit messages in offensive and abusive language. In *Proceedings of the 12th language resources and evaluation conference*, pages 6193–6202.

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

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2020. Language-agnostic bert sentence embedding. *arXiv preprint arXiv:2007.01852*.

Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *ACM Comput. Surv.*, 51(4).

Raul Gomez, Jaume Gibert, Lluis Gomez, and Dimosthenis Karatzas. 2020. Exploring Hate Speech Detection in Multimodal Publications. In *Proceedings of the Winter Conference on Applications of Computer Vision*, pages 1–8.

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. *Advances in Neural Information Processing Systems*, 33.

Klaus Krippendorff. 2004. Reliability in content analysis: Some common misconceptions and recommendations. *Human communication research*, 30(3):411–433.

Klaus Krippendorff. 2018. *Content analysis: An introduction to its methodology*. Sage publications.

Anti-Defamation League. 2020. Online hate and harassment. the american experience 2021. *Center for Technology and Society. Retrieved from www. adl. org/media/14643/download*.

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

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 *European conference on computer vision*, pages 740–755. Springer.

Mary L Mchugh. 2012. Interrater reliability: the Kappa statistic. *Biochemia Medica*, 22(3):276–282.

Pushkar Mishra, Helen Yannakoudakis, and Ekaterina Shutova. 2019. Tackling online abuse: A survey of automated abuse detection methods. *arXiv preprint arXiv:1908.06024*.

Michael Omi and Howard Winant. 2005. The Theoretical Status of the Concept of Race. In Cameron McCarthy, Warren Crichlow, Greg Dimitriadis, and Nadine Dolby, editors, *Race, Identity and Representation in Education*. Routledge, London.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*.

Amanpreet Singh, Vedanuj Goswami, Vivek Natarajan, Yu Jiang, Xinlei Chen, Meet Shah, Marcus Rohrbach, Dhruv Batra, and Devi Parikh. 2020. Mmf: A multimodal framework for vision and language research. `https://github.com/facebookresearch/mmf`.

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2019. Vl-bert: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530*.

Bertie Vidgen, Alex Harris, Josh Cowls, Ella Guest, and Helen Margetts. 2021. *An agenda for research into online hate*. The Alan Turing Institute, London.

Bertie Vidgen, Alex Harris, Dong Nguyen, Rebekah Tromble, Scott Hale, and Helen Margetts. 2019. Challenges and frontiers in abusive content detection. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 80–93, Florence, Italy. Association for Computational Linguistics.

Zeerak Waseem, Thomas Davidson, Dana Warmsley, and Ingmar Weber. 2017. Understanding Abuse: A Typology of Abusive Language Detection Subtasks. In *Proceedings of the First Workshop on Abusive Language Online*, pages 78–84.

Zeerak Waseem, James Thorne, and Joachim Bingel. 2018. Bridging the gaps: Multi task learning for domain transfer of hate speech detection. In *Online harassment*, pages 29–55. Springer.

Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017. Ex Machina: Personal Attacks Seen at Scale. In *Proceedings of the International World Wide Web Conference*, pages 1391–1399.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. Predicting the type and target of offensive posts in social media. In *Proceedings of NAACL HLT 2019*, volume 1, pages 1415–1420.

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