The Hateful Memes Challenge: Competition Report

Douwe Kiela†, Hamed Firooz†, Aravind Mohan†, Vedanuj Goswami†, Amanpreet Singh†, Casey A. Fitzpatrick†, Peter Bull†, Greg Lipstein†, Tony Nelli†
† Facebook AI
‡ DrivenData Inc.

Ron Zhu*, Niklas Muennighoff§, Riza Velioglu¶, Jewgeni Rose∥, Phillip Lippe†, Nithin Holla†, Shantanu Chandra†, Santhosh Rajamanickam††, Georgios Antoniou†, Ekaterina Shutova†, Helen Yannakoudakis†, Vlad Sandulescu‡‡
Winning competition participants
* Alfred System, § Peking University, ¶ Bielefeld University, ∥ University of Bonn, † University of Amsterdam, †† Slimmer AI, ‡ King’s College London, ‡‡ Wunderman Thompson

Umut Ozertem†, Patrick Pantel†, Lucia Specia‡, Devi Parikh§§
Advisory committee
† Facebook AI, § Georgia Tech, ‡ Imperial College London

Editors: Hugo Jair Escalante and Katja Hofmann

Abstract

Machine learning and artificial intelligence play an ever more crucial role in mitigating important societal problems, such as the prevalence of hate speech. We describe the Hateful Memes Challenge competition, held at NeurIPS 2020, focusing on multimodal hate speech. The aim of the challenge is to facilitate further research into multimodal reasoning and understanding.

Keywords: multimodal, vision and language, hate speech

1. Introduction

At the sheer scale of the internet, malicious content cannot be tackled by having humans inspect every data point. Consequently, machine learning and artificial intelligence play an ever more important role in mitigating important societal problems, such as the prevalence of hate speech. Hate speech is understood to mean “any communication that disparages a target group of people based on some characteristic such as race, colour, ethnicity, gender, sexual orientation, nationality, religion, or other characteristic” (Nockleby, 2000).

Detecting hate speech is a difficult problem, as it often relies heavily on context, requires world knowledge, and can be rather subtle. It is also an important problem, in how it has the potential to affect everyone in our society. One particularly challenging type of hate speech is found in multimodal internet memes—narrowly defined, images overlaid with text, designed to spread from person to person via social networks, often for (perceived) humorous purposes. For this competition, we proposed a new challenge task and dataset: detecting hatefulness in multimodal memes. The long-term hope for the challenge and corresponding
datasets is to facilitate breakthroughs in multimodal methods that can be applied to a very broad set of problems, going far beyond hate speech.

Memes pose an interesting multimodal fusion problem, i.e., their understanding requires a very specific combination of information from different modalities (the text and the image). Consider, as an illustration, a sentence like “you smell nice today” paired with an image of a skunk, or “look how many people love you” with a picture of a tumbleweed in the desert. Unimodally, these examples are boring and harmless, but when the modalities are combined the meaning changes and they suddenly become mean—which is easy for humans to detect, but (so far) challenging to AI systems.

A primary motivation for the challenge, in addition to the obvious importance of tackling hate speech, is that we believe there is room for vision-and-language tasks to extend beyond the popular tasks of visual question answering (Antol et al., 2015; Johnson et al., 2017) and image captioning (Chen et al., 2015; Young et al., 2014; Krishna et al., 2017). While these tasks are very important and have contributed immensely to the progress of the field, one could argue that they are different from many of the problems in industry, using real-world internet data, where the goal might be to classify a tweet, post or comment.

A crucial characteristic of the challenge is that we include so-called “benign confounders” to counter the possibility of models exploiting unimodal priors: for every hateful meme, we find alternative images or captions that make the label flip to not-hateful. Using the examples above, for example, if we replaced the skunk and tumbleweed images with pictures of roses or people, the memes become harmless again. Similarly, we can flip the label by keeping the original images but changing the text to “look how many people hate you” or “skunks have a very particular smell”. Thus, the challenge is designed such that it should only be solvable by models that are successful at sophisticated multimodal reasoning and understanding.

The Hateful Memes Challenge task has obvious direct real-world applicability, and cannot be solved by only looking at the image or the text, instead requiring sophisticated multimodal fusion. It is difficult and requires subtle reasoning, yet is easy to evaluate as a binary classification task. The challenge can thus be said to serve the dual purpose of measuring progress on multimodal understanding and reasoning, while at the same time facilitating progress in a real-world application of hate speech detection.¹

2. Related Work

**Hate speech** There has been a lot of work in recent years on detecting hate speech in network science (Ribeiro et al., 2018) and natural language processing (Waseem et al., 2017; Schmidt and Wiegand, 2017; Fortuna and Nunes, 2018). Several text-only hate speech datasets have been released, mostly based on Twitter (Waseem, 2016; Waseem and Hovy, 2016; Davidson et al., 2017; Golbeck et al., 2017; Founta et al., 2018), and various architectures have been proposed for classifiers (Kumar et al., 2018; Malmasi and Zampieri, 2018, 2017). Hate speech detection has proven to be difficult, and for instance subject to unwanted bias (Dixon et al., 2018; Sap et al., 2019; Davidson et al., 2019). One issue is that not all of these works have agreed on what defines hate speech, and different terminology

¹ The dataset is available at https://hatefulmemeschallenge.com, which also hosts the leaderboard.
The Hateful Memes Challenge

has been used, ranging from offensive or abusive language, to online harassment or aggression, to cyberbullying, to harmful speech, to hate speech (Waseem et al., 2017). Here, we focus exclusively on hate speech in a narrowly defined context (see Section 3.1).

Multimodal hate speech There has been surprisingly little work related to multimodal hate speech, with only a few papers including both images and text. Yang et al. (Yang et al., 2019) report that augmenting text with image embedding information immediately boosts performance in hate speech detection. Hosseinmardi et al. (Hosseinmardi et al., 2015) collect a dataset of Instagram images and their associated comments, which they then label with the help of Crowdflower workers. They asked workers two questions: 1) does the example constitute cyberaggression; and 2) does it constitute cyberbullying. Where the former is defined as “using digital media to intentionally harm another person” and the latter is a subset of cyber-aggression, defined as “intentionally aggressive behavior that is repeatedly carried out in an online context against a person who cannot easily defend him or herself” (Hosseinmardi et al., 2015). They show that including the image features improves classification performance. The dataset consisted of 998 examples, of which 90% was found to have high-confidence ratings, of which 52% was classified as bullying. Singh et al. (Singh et al., 2017) conduct a detailed study, using the same dataset, of the types of features that matter for cyber-bullying detection in this task. Similarly, Zhong et al. (Zhong et al., 2016) collected a dataset of Instagram posts and comments, consisting of 3000 examples. They asked Mechanical Turk workers two questions: 1) do the comments include any bullying; and 2) if so, is the bullying due to the content of the image. 560 examples were found to be bullying. They experiment with different kinds of features and simple classifiers for automatically detecting whether something constitutes bullying.

Our work differs from these works in various ways: our dataset is larger and explicitly designed to be difficult for unimodal architectures; we only include examples with high-confidence ratings from trained annotators and carefully balance the dataset to include different kinds of multimodal fusion problems; we focus on hate speech, rather than the more loosely defined cyberbullying; and finally we test more sophisticated models on this problem. Vijayaraghavan et al. (Vijayaraghavan et al., 2019) propose methods for interpreting multimodal hatespeech detection models, where the modalities consist of text and socio-cultural information rather than images. Concurrently, Gomez et al. (Gomez et al., 2020) introduced a larger (and arguably noisier) dataset for multimodal hate speech detection based on Twitter data, which also contains memes and which would probably be useful as pretraining data for our task.

Vision and language tasks Multimodal hate speech detection is a vision and language task. Vision and language problems have gained a lot of traction in recent years (see Mogadala et al. (Mogadala et al., 2019) for a survey), with great progress on important problems such as visual question answering (Antol et al., 2015; Goyal et al., 2017) and image caption generation and retrieval (Chen et al., 2015; Young et al., 2014; Krishna et al., 2017; Sidorov et al., 2020; Gurari et al., 2020), with offshoot tasks focusing specifically on visual reasoning (Johnson et al., 2017), referring expressions (Kazemzadeh et al., 2014), visual storytelling (Park and Kim, 2015; Huang et al., 2016), visual dialogue (Das et al., 2017; De Vries et al., 2017), multimodal machine translation (Elliott et al., 2016; Specia et al., 2016), visual reasoning (Suhr et al., 2018; Hudson and Manning, 2019; Singh et al., 2019;
The Hateful Memes Challenge

Xie et al., 2019; Gurari et al., 2018), visual common sense reasoning (Zellers et al., 2019) and many others.

A large subset of these tasks focus on (autoregressive) text generation or retrieval objectives. One of the two modalities is usually dominant. They often rely on bounding boxes or similar features for maximum performance, and are not always easy to evaluate (Vedantam et al., 2015). While these tasks are of great interest to the community, they are different from the kinds of real-world multimodal classification problems one might see in industry—a company like Facebook or Twitter, for example, needs to classify a lot of multimodal posts, ads, comments, etc for a wide variety of class labels. These use cases often involve large-scale, text-dominant multimodal classification similar to what is proposed in this task.

Related multimodal classification tasks exist; for instance, there has been extensive research in multimodal sentiment (Soleymani et al., 2017), but there is no agreed-upon standard dataset or benchmark task. Other datasets using internet data include Food101 (Wang et al., 2015), where the goal is to predict the dish of recipes and images; various versions of Yelp reviews (Ma et al., 2018); Walmart and Ferramenta product classification (Zahavy et al., 2016; Gallo et al., 2017); social media name tagging (Twitter and Snapchat) (Lu et al., 2018); social media target-oriented sentiment (Yu and Jiang, 2019); social media crisis handling (Alam et al., 2018); various multimodal news classification datasets (Ramisa, 2017; Shu et al., 2017); multimodal document intent in Instagram posts (Kruk et al., 2019); and predicting tags for Flickr images (Thomee et al., 2015; Joulin et al., 2016). Other datasets include grounded entailment, which exploits the fact that one of the large-scale natural language inference datasets was constructed using captions as premises, yielding a image, premise, hypothesis triplet with associated entailment label (Vu et al., 2018); as well as MM-IMDB, where the aim is to predict genres from posters and plots (Arevalo et al., 2017); and obtaining a deeper understanding of multimodal advertisements, which requires similarly subtle reasoning (Hussain et al., 2017; Zhang et al., 2018). Sabat et al. (Sabat et al., 2019) recently found in a preliminary study that the visual modality can be more informative for detecting hate speech in memes than the text. The quality of these datasets varies substantially, and their data is not always readily available to different organizations. Consequently, there has been a practice where authors opt to simply “roll their own” dataset, leading to a fragmented status quo. We believe that our dataset fills up an important gap in the space of multimodal classification datasets.

3. The Competition

The original Hateful Memes dataset was proposed by Kiela et al. (2020) as a means to measure progress in research on multi-modal reasoning and understanding. It incorporates benign confounders in an attempt to tease apart differences between models that are only superficially multimodal and models that can truly conduct sophisticated multimodal fusion. The dataset is hoped to be particularly useful for evaluating large scale models pre-trained on other data. The original paper describes the collection and annotation procedure, the various splits, and the performance of state of the art vision-and-language systems as baselines. For this competition, an “unseen” test set was constructed specifically for the purpose of evaluating solutions using new source material, ensuring that competition participants would be evaluated on the actual task (which would in the real world include completely
novel unseen examples) and also to mitigate the risk of participants exploiting inadvertent biases. The competition underwent two “phases”: the first phase using the seen test set, which lasted from May to October with one submission allowed per day; and the second phase using the unseen test set, lasting for the month of October, with three submissions allowed in total.

3.1. Task Formulation
Hate speech, in the context of this paper and the challenge set, is strictly defined as follows:

A direct or indirect attack on people based on characteristics, including ethnicity, race, nationality, immigration status, religion, caste, sex, gender identity, sexual orientation, and disability or disease. We define attack as violent or dehumanizing (comparing people to non-human things, e.g. animals) speech, statements of inferiority, and calls for exclusion or segregation. Mocking hate crime is also considered hate speech.

There are some notable but subtle exceptions in this definition, i.e., attacking individuals/famous people is allowed if the attack is not based on any of the protected characteristics. Also, attacking groups perpetrating hate (e.g. terrorist groups) is not considered hate. The definition resembles (but is a very simplified version of) community standards on hate speech employed by e.g. Facebook.

The task is to classify a meme—i.e., an image and some text (the text is pre-extracted from the image in lieu of having to do optical character recognition)—based on whether it is hateful according to the above definition, or not.

3.2. Metrics
The primary metric for the competition, and the metric we encourage the community to use, is the area under the receiver operating characteristic curve (AUROC; Bradley, 1997). We also encourage the community to report the accuracy as a secondary metric, since it is easily interpretable and the dev and test sets are not extremely unbalanced, so accuracy gives a reasonable (though imperfect) signal of model performance. The competition winners were decided based on AUROC, which gives a fine-grained sense of classifier performance.

3.3. Data
The dataset construction procedure is discussed in detail in Kiela et al. (2020). In summary, it consisted of four phases: 1) data filtering; 2) meme reconstruction; 3) hatefulness ratings; 4) benign confounder construction. In order to ensure that the dataset would be freely distributable for research purposes, we partnered with Getty Images for sourcing the images, synthetically constructing memes using those source images as the background upon which text was overlaid.

There are different types of memes in the dataset. Hateful examples can be multimodal in nature, meaning that the classification relies on both modalities, or unimodal, meaning

---

3. Note that in real-world data, the prevalence of hate speech would be much lower.
that one modality is enough to obtain the correct classification label. In addition, the dataset contains “confounders” that are constructed such that minimal changes to one of the modalities cause the label to flip. To balance out the data, we also include random benign examples, which in practice is by far the most common meme category in the wild.

During the early stages of the competition, discrepancies were found in the hatefulness annotations, mostly as a result of noisy examples from the original reconstruction procedure and annotator confusion. This was addressed by having the entire dataset reannotated with better training and stricter guidelines for the “second phase” of the competition.

### 3.4. Splits

Table 1 shows how the dataset breaks down into various categories. In phase 1 of the competition, dev “seen” and test “seen” were used. Unlike the train set, which is dominated by unimodal violating contents, dev “seen” and test “seen” set are dominated by multimodal contents. Moreover, the labels distribution is balanced. For phase 2 (the prize winning phase) “unseen” dev and “Unseen” test set were constructed. There are no unimodal violating contents in these two new sets. This encouraged the competitors to push accuracy of their multimodal models.

<table>
<thead>
<tr>
<th>Type</th>
<th>Model</th>
<th><strong>Unseen Dev</strong></th>
<th><strong>Unseen Test</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Acc.</td>
<td>AUROC</td>
</tr>
<tr>
<td>Unimodal</td>
<td>Image-Region</td>
<td>61.48</td>
<td>53.54</td>
</tr>
<tr>
<td></td>
<td>Text BERT</td>
<td>60.37</td>
<td>60.88</td>
</tr>
<tr>
<td>Multimodal (Unimodal Pretraining)</td>
<td>Late Fusion</td>
<td>61.11</td>
<td>61.00</td>
</tr>
<tr>
<td></td>
<td>Concat BERT</td>
<td>64.81</td>
<td>65.42</td>
</tr>
<tr>
<td></td>
<td>MMBT-Grid</td>
<td>67.78</td>
<td>65.47</td>
</tr>
<tr>
<td></td>
<td>MMBT-Region</td>
<td>70.04</td>
<td>71.54</td>
</tr>
<tr>
<td></td>
<td>ViLBERT</td>
<td>69.26</td>
<td>72.73</td>
</tr>
<tr>
<td></td>
<td>Visual BERT</td>
<td>69.67</td>
<td>71.10</td>
</tr>
<tr>
<td>Multimodal (Multimodal Pretraining)</td>
<td>ViLBERT CC</td>
<td>70.37</td>
<td>70.78</td>
</tr>
<tr>
<td></td>
<td>Visual BERT COCO</td>
<td>70.77</td>
<td>73.70</td>
</tr>
</tbody>
</table>

Table 2: Unseen dev and test set performance for baseline models (see Kiela et al. (2020) for baseline model performance on the “seen” dev and test sets).
### Baselines and starter kit

Baseline scores for various unimodal and several state-of-the-art multimodal models on this task were established at the start of the competition. Starter kit code was provided to all participants, and is available as a part of the MMF multi-modal framework at: 

### Results & Analysis

#### 4.1. Participation

Table 3 shows the overall competition statistics. The competition had a large number of participants, which narrowed down towards the later stages due to stricter submission constraints.

An issue that emerged during the competition was that some teams considered “pseudo labelling” on the test set to be a valid approach. Pseudo labelling can in some cases be perfectly legitimate, e.g. for semi-supervised learning where a small set of supervised data can be used to impute labels for unsupervised data that may be used in subsequent training, but for obvious reasons this approach should not be applied to test set examples. Some contestants also exploited knowledge about the dataset construction process, basing test set example predictions on other test set example labels, which also obviously violates test set integrity. Both approaches were actively discouraged, but these issues constitute a weakness in the dataset that is important to explicitly acknowledge: the construction process led to “triplets” of one hateful memes and two similar non-hateful confounder memes. This knowledge can be trivially exploited by comparing a given example to other examples in the test set, effectively classifying the most-probably-hateful meme as hateful and the others as automatically wholly not-hateful. Obviously, this approach defeats the purpose of the test set (for measuring generalization to novel examples) and violates standard machine learning practice. Solutions that employed this approach were disqualified from the competition.

#### 4.2. Winning solutions

The competition had a prize pool of 100k USD, divided over the top 5 winning teams. As a requirement for prize eligibility, the winning teams were asked to open-source all their code and write an academic paper outlining how to reproduce their results. We hope others across the AI research community will build on their work and be able to improve their own systems.
# The Hateful Memes Challenge

<table>
<thead>
<tr>
<th>#</th>
<th>Team</th>
<th>AUROC</th>
<th>Acc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Ron Zhu</td>
<td>0.844977</td>
<td>0.7320</td>
</tr>
<tr>
<td>2</td>
<td>Niklas Muennighoff</td>
<td>0.831037</td>
<td>0.6950</td>
</tr>
<tr>
<td>3</td>
<td>Team HateDetectron</td>
<td>0.810845</td>
<td>0.7650</td>
</tr>
<tr>
<td>4</td>
<td>Team Kingsterdam</td>
<td>0.805254</td>
<td>0.7385</td>
</tr>
<tr>
<td>5</td>
<td>Vlad Sandulescu</td>
<td>0.794321</td>
<td>0.7430</td>
</tr>
</tbody>
</table>

Table 4: Competition winners

**#1 Ron Zhu: Enhancing Multimodal Transformers with External Labels And In-Domain Pretraining** Zhu (2020) won first prize.\(^4\) The solution stood apart for a number of reasons. It employed a diverse ensemble of VL-BERT (Su et al., 2019), UNITER-ITM (Chen et al., 2019), VILLA-ITM (Gan et al., 2020), and ERNIE-Vil (Yu et al., 2020) models. In addition to the text and image inputs, the models were given entity, race and gender classifications. Entity labels were obtained via the Google Cloud vision API’s web detection tool\(^5\). Race and gender labels were obtained by extracting faces using Mask-RCN (Ren et al., 2015) and classifying them.

**#2 Niklas Muennighoff: State-of-the-art Visio-Linguistic Models applied to Hateful Memes** Muennighoff (2020) won second prize.\(^6\) The implementation fits vision- and language models into a uniform framework and adds specific enhancements. Specifically, masked pre-training helps the models adapt to the Hateful Memes dataset before being trained on classification. A visual token type is added to ease differentiation between text and visual content. Stochastic Weight Averaging (Izmailov et al., 2018) is used to stabilize training and make performance seed-independent. ERNIE-Vil (Yu et al., 2020), UNITER (Chen et al., 2019), OSCAR (Li et al., 2020) and VisualBERT (Li et al., 2019) models are ensembled in a loop to produce the final score.

**#3 Team HateDetectron: Detecting Hate Speech in Memes Using Multimodal Deep Learning Approaches** Velioglu and Rose (2020) won third prize.\(^7\) The solution has a lower complexity compared to other solutions as it only uses a single model—VisualBERT (Li et al., 2019). Singh et al. (2020b) showed that the source domain of the pre-training dataset highly impacts the model’s capability. For this reason, VisualBERT is pre-trained on Conceptual Captions (Sharma et al., 2018), which are similar to this competition’s memes in the sense of multimodality between text and image, and fine-tuned on an aggregated dataset where a part of the Memotion dataset (Sharma et al., 2020) was added to the Hateful Memes dataset. As a result of hyper-parameter tuning, an ensemble of 27 models are used to classify memes using majority voting technique.

**#4 Team Kingsterdam: A Multimodal Framework for the Detection of Hateful Memes** Lippe et al. (2020) won fourth prize.\(^8\) The solution combined UNITER (Chen

---

4. [https://github.com/HimariO/HatefulMemesChallenge](https://github.com/HimariO/HatefulMemesChallenge)
5. [https://cloud.google.com/vision/docs/internet-detection](https://cloud.google.com/vision/docs/internet-detection)
6. [https://github.com/Muennighoff/vilio](https://github.com/Muennighoff/vilio)
7. [https://github.com/rizavelioglu/hateful_memes-hate_detectron](https://github.com/rizavelioglu/hateful_memes-hate_detectron)
8. [https://github.com/Nithin-Holla/meme_challenge](https://github.com/Nithin-Holla/meme_challenge)
et al., 2019) with a number of techniques for improved learning. The text confounders in the dataset were upsampled during training to help improve the model’s multimodal reasoning capabilities, while a loss re-weighting strategy was applied to favour the minority class. This was followed by an ensemble of 15 UNITER models trained on different splits of the data such that subsets from the development set were included in the training folds. This ensured that the high percentage of truly multimodal examples in the development set was utilised during training. The final predictions were obtained via a weighted linear combination of the ensemble predictions, optimised using an evolutionary algorithm on the development set.

#5 Vlad Sandulescu: Detecting Hateful Memes Using a Multimodal Deep Ensemble

Sandulescu (2020) won fifth prize. They experiment with both single-stream Transformer architectures: VL-BERT (Su et al., 2019), VLP (Zhou et al., 2019) and UNITER (Chen et al., 2019) as well as dual-stream models such as LXMERT (Tan and Bansal, 2019), showing single-stream models outperform the two-stream ones on this task. These large architectures are chosen such that by ensembling them one could exploit the fact they are pre-trained on a wide spectrum of datasets from different domains. The highest scoring solution involves an ensemble of UNITER models, each including an extra bidirectional cross-attention mechanism to couple inferred caption information using the Show and Tell model from (Vinyals et al., 2016) to the already supplied meme text. Finally, deep ensembles (Lakshminarayanan et al., 2017), a simple yet very powerful trick, improve on single model predictions by a significant margin.

4.3. Take-aways

In our assessment, the competition was a huge success. We had a large number of participants, a lively competition community and interesting novel solutions to this important problem as prize winners. Here, we list some take-aways from the competition and winning solutions.

Frameworks matter

We provided an easy starter kit codebase using MMF Singh et al. (2020a), so that participants would not have to worry about implementational details and could immediately focus on innovating on e.g. the architecture. Most participants used this codebase, but interestingly not all winning teams did so. Muennighoff (2020) for example built a framework from scratch. Zhu (2020) manually ported ERNIE-Vil from PaddlePaddle10 to PyTorch, a herculean effort that was credited at the competition event as one of the reasons behind their success. Overall, solutions were engineering heavy and the easy availability (or not) of particular methods made a clear difference in giving some participants an edge over the otherwise relatively level playing field.

Pretrained models

Bugliarello et al. (2020) recently described intriguing results showing that differences between various “Vision and Language BERTs” are mostly due to training data and hyperparameters. The winning solutions used a wide variety of such models. Some participants argued that specific architectures were better than others—notably UNITER, VILLA and ERNIE-Vil—but this remains mostly speculative. Similarly, there did appear

10. https://github.com/PaddlePaddle/Paddle
to be a recency bonus for models, where newly released models (even ones released when the competition was long underway) gave participants an upper hand, like in the case of ERNIE-ViL.

**Ensembles** As is usually the case in competition, all winning solutions employed ensembles. Interestingly, the ensembles were not necessarily of different model architectures. This does raise issues for deploying solutions in production, which has heavy computational constraints.

**Entities, Faces and External knowledge** Understanding memes often requires subtle world knowledge, which many participants tried to exploit. The winning solutions’ reliance on a concept detection pipeline is illustrative of this, and we speculate that incorporating rich conceptual knowledge (e.g. not only knowing that the object is a “car” but that it’s a “Volkswagen Beetle Type 2”) will be very helpful. Given the nature of the dataset, having explicit knowledge of hate speech target features (like race and gender) also helped, however incorporating such features in practice raises important ethical dilemmas.

### 5. Conclusion & Outlook

We described the Hateful Memes Challenge competition, the newly collected “unseen” datasets and the winning solutions. Open competitions around important common problems are some of the AI research community’s most effective tools for accelerating progress. Hate speech remains a crucially important challenge, and multimodal hate speech in particular continues to be an especially difficult machine learning problem. The Hateful Memes Challenge competition is over, but the real challenge is far from solved: A lot of work remains to be done in multimodal AI research, and we hope that this work can play an important role in evaluating new solutions that the field comes up with. The dataset design makes it a good candidate for evaluating the power of next-generation multimodal pretrained models, as well as currently still unimagined advances in the field. We hope that this task and these datasets will continue to inform new approaches and methods going forward.

**Acknowledgments**

We thank the many participants who took part in the competition. We also thank our colleagues at Facebook who helped make this dataset and competition possible. We’re thankful to Getty Images for making this competition possible.

**References**


The Hateful Memes Challenge


