Proceedings of Machine Learning Research 133:344–360, 2021 NeurIPS 2020 Competition and Demonstration Track

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.

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 constititute 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 is 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;

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[2].

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[3], 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

2. https://www.facebook.com/communitystandards/hate_speech
3. Note that in real-world data, the prevalence of hate speech would be much lower.

	Total	Not-hate	Hate	MM Hate	UM Hate	Img Conf	Txt Conf	Rand Benign
Train	8500	5481	3019	1100	1919	1530	1530	2421
Dev seen	500	253	247	200	47	100	100	53
Dev uneen	540	340	200	200	0	170	170	0
Test seen	1000	510	490	380	110	190	190	130
Test unseen	2000	1250	750	750	0	625	625	0

Table 1: Dataset splits

		Unseen Dev		**Unseen Test**	
Type	**Model**	Acc.	AUROC	Acc.	AUROC
Unimodal	Image-Region	61.48	53.54	60.28±0.18	54.64±0.80
	Text BERT	60.37	60.88	63.60±0.54	62.65±0.40
Multimodal (Unimodal Pretraining)	Late Fusion	61.11	61.00	64.06±0.02	64.44±1.60
	Concat BERT	64.81	65.42	65.90±0.82	66.28±0.66
	MMBT-Grid	67.78	65.47	66.85±1.61	67.24±2.53
	MMBT-Region	70.04	71.54	70.10±1.39	72.21±0.20
	ViLBERT	69.26	72.73	70.86±0.70	73.39±1.32
	Visual BERT	69.67	71.10	71.30±0.68	73.23±1.04
Multimodal (Multimodal Pretraining)	ViLBERT CC	70.37	70.78	70.03±1.07	72.78±0.50
	Visual BERT COCO	70.77	73.70	69.95±1.06	74.59±1.56

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).

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 wining 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.

Phase	Start	End	Participants	Submissions	Sub. Part.	Constraints
1 - "seen"	May 10th	Oct. 1st	3,532	3,646	510	1 per day
2 - "unseen"	Oct. 1st	Oct. 31th	3,173	276	105	3 total

Table 3: Participation Statistics. "Participants" is the number of people who registered on the competition page. "Sub. Part" is the number of participants who submitted at least one solution.

3.5. 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: https://github.com/facebookresearch/mmf/tree/master/projects/hateful_memes.

4. 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.

#	Team	AUROC	Acc.
1	Ron Zhu	0.844977	0.7320
2	Niklas Muennighoff	0.831037	0.6950
3	Team HateDetectron	0.810845	0.7650
4	Team Kingsterdam	0.805254	0.7385
5	Vlad Sandulescu	0.794321	0.7430

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
5. https://cloud.google.com/vision/docs/internet-detection
6. https://github.com/Muennighoff/vilio
7. https://github.com/rizavelioglu/hateful_memes-hate_detectron
8. 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.[9] 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 PaddlePaddle[10] 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

9. https://github.com/vladsandulescu/hatefulmemes
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

Firoj Alam, Ferda Ofli, and Muhammad Imran. Crisismmd: Multimodal twitter datasets from natural disasters. In *Twelfth International AAAI Conference on Web and Social Media*, 2018.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015.

John Arevalo, Thamar Solorio, Manuel Montes-y Gómez, and Fabio A González. Gated multimodal units for information fusion. In *Proceedings of ICLR*, 2017.

Andrew P Bradley. The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern recognition*, 30(7):1145–1159, 1997.

Emanuele Bugliarello, Ryan Cotterell, Naoaki Okazaki, and Desmond Elliott. Multimodal pretraining unmasked: Unifying the vision and language berts. *arXiv preprint arXiv:2011.15124*, 2020.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Learning universal image-text representations. *arXiv preprint arXiv:1909.11740*, 2019.

A. Das, S. Kottur, K. Gupta, A. Singh, D. Yadav, J. M. Moura, Yadav D., D. Parikh, and D. Batra. Visual dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 326–335, 2017.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. Automated hate speech detection and the problem of offensive language. In *Eleventh international aaai conference on web and social media*, 2017.

Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. Racial bias in hate speech and abusive language detection datasets. In *Proceedings of ACL*, 2019.

H. De Vries, F. Strub, S. Chandar, O. Pietquin, H. Larochelle, and A. Courville. Guesswhat?! visual object discovery through multi-modal dialogue. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5503–5512, 2017.

Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, , and Lucy Vasserman. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 67–73, 2018.

Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. Multi30k: Multilingual english-german image descriptions. *arXiv:1605.00459*, 2016.

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

Antigoni Maria Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. Large scale crowdsourcing and characterization of twitter abusive behavior. In *Twelfth International AAAI Conference on Web and Social Media*, 2018.

I. Gallo, A. Calefati, and S. Nawaz. Multimodal classification fusion in real-world scenarios. In *14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, volume 5, pages 36–41, 2017.

Zhe Gan, Yen-Chun Chen, Linjie Li, Chen Zhu, Yu Cheng, and Jingjing Liu. Large-scale adversarial training for vision-and-language representation learning. *arXiv preprint arXiv:2006.06195*, 2020.

Jennifer Golbeck, Zahra Ashktorab, Rashad O. Banjo, Alexandra Berlinger, Siddharth Bhagwan, Cody Buntain, Paul Cheakalos, Alicia A. Geller, Quint Gergory, Rajesh Kumar Gnanasekaran, Raja Rajan Gunasekaran, Kelly M. Hoffman, Jenny Hottle, Vichita Jienjitlert, Shivika Khare, Ryan Lau, Marianna J. Martindale, Shalmali Naik, Heather L. Nixon, Piyush Ramachandran, Kristine M. Rogers, Lisa Rogers, Meghna Sardana Sarin, Gaurav Shahane, Jayanee Thanki, Priyanka Vengataraman, Zijian Wan, and Derek Michael Wu. A large labeled corpus for online harassment research. In *Proceedings of the 2017 ACM on Web Science Conference*, pages 229–233, 2017.

Raul Gomez, Jaume Gibert, Lluis Gomez, and Dimosthenis Karatzas. Exploring hate speech detection in multimodal publications. In *The IEEE Winter Conference on Applications of Computer Vision*, pages 1470–1478, 2020.

Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6904–6913, 2017.

Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3608–3617, 2018.

Danna Gurari, Yinan Zhao, Meng Zhang, and Nilavra Bhattacharya. Captioning images taken by people who are blind. *arXiv preprint arXiv:2002.08565*, 2020.

Homa Hosseinmardi, Sabrina Arredondo Mattson, Rahat Ibn Rafiq, Richard Han, Qin Lv, and Shivakant Mishra. Detection of cyberbullying incidents on the instagram social network. *arXiv:1503.03909*, 2015.

Ting-Hao Kenneth Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Aishwarya Agrawal, Jacob Devlin, Ross Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, C. Lawrence Zitnick, Devi Parikh, Lucy Vanderwende, Michel Galley, and Margaret Mitchell. Visual storytelling. In *Proceedings of NAACL-HLT*, pages 1233–1239, 2016.

Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. *arXiv preprint arXiv:1902.09506*, 2019.

Z. Hussain, M. Zhang, X. Zhang, K. Ye, C. Thomas, Z. Agha, N. Ong, and A. Kovashka. Automatic understanding of image and video advertisements. pages 1705–1715, 2017.

Pavel Izmailov, Dmitrii Podoprikhin, Timur Garipov, Dmitry Vetrov, and Andrew Gordon Wilson. Averaging weights leads to wider optima and better generalization. *arXiv preprint arXiv:1803.05407*, 2018.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2901–2910, 2017.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*, 2016.

S. Kazemzadeh, V. Ordonez, M. Matten, and T. Berg. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing*, pages 787–798, 2014.

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. *arXiv preprint arXiv:2005.04790*, 2020.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73, 2017.

J. Kruk, J. Lubin, K. Sikka, X. Lin, D. Jurafsky, and A. Divakaran. Integrating text and image: Determining multimodal document intent in instagram posts. In *Proceedings of EMNLP*, 2019.

Ritesh Kumar, Atul Kr Ojha, Shervin Malmasi, and Marcos Zampieri. Benchmarking aggression identification in social media. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, pages 1–11, 2018.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *NIPS*, 2017.

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.

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, pages 121–137. Springer, 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.

Di Lu, Leonardo Neves, Vitor Carvalho, Ning Zhang, and Heng Ji. Visual attention model for name tagging in multimodal social media. In *Proceedings of ACL*, pages 1990–1999, 2018.

Y. Ma, Z. Xiang, Q. Du, and W. Fan. Effects of user-provided photos on hotel review helpfulness: An analytical approach with deep leaning. *International Journal of Hospitality Management*, 71:120–131, 2018.

S Malmasi and M Zampieri. Detecting hate speech in social media. *arXiv:1712.06427*, 2017.

Shervin Malmasi and Marcos Zampieri. Challenges in discriminating profanity from hate speech. *Journal of Experimental & Theoretical Artificial Intelligence*, 30(2):187–202, 2018.

A. Mogadala, M. Kalimuthu, and D. Klakow. Trends in integration of vision and language research: A survey of tasks, datasets, and methods. *arXiv:1907.09358*, 2019.

Niklas Muennighoff. Vilio: State-of-the-art visio-linguistic models applied to hateful memes. *arXiv preprint arXiv:2012.07788*, 2020.

J. T. Nockleby. Hate speech. In *Encyclopedia of the American Constitution*, volume 3, page 1277–79, 2000.

C. C. Park and G. Kim. Expressing an image stream with a sequence of natural sentences. In *Advances in neural information processing systems*, pages 73–81, 2015.

Arnau Ramisa. Multimodal news article analysis. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, pages 5136–5140, 2017.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

M. H. Ribeiro, P. H. Calais, Y. A. Santos, V. A. Almeida, and W. Meira Jr. Characterizing and detecting hateful users on twitter. In *Twelfth International AAAI Conference on Web and Social Media*, 2018.

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

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

Maarten Sap, Dallas Card, Saadia Gabriela, Yejin Choi, and Noah A. Smith. The risk of racial bias in hate speech detection. In *Proceedings of ACL*, pages 1668–1678, 2019.

Anna Schmidt and Michael Wiegand. A survey on hate speech detection using natural language processing. 2017.

Chhavi Sharma, Deepesh Bhageria, William Scott, Srinivas PYKL, Amitava Das, Tanmoy Chakraborty, Viswanath Pulabaigari, and Bjorn Gamback. Semeval-2020 task 8: Memotion analysis–the visuo-lingual metaphor! *arXiv preprint arXiv:2008.03781*, 2020.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, 2018.

Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. Fake news detection on social media: A data mining perspective. *arxiv:1708.01967*, 2017.

Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. Textcaps: a dataset for image captioning with reading comprehension. *arXiv preprint arXiv:2003.12462*, 2020.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8317–8326, 2019.

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

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

Vivek K. Singh, Souvick Ghosh, and Christin Jose. Toward multimodal cyberbullying detection. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, pages 2090–2099, 2017.

M. Soleymani, D. Garcia, B. Jou, B. Schuller, S. F. Chang, and M. Pantic. A survey of multimodal sentiment analysis. *Image and Vision Computing*, 65:3–14, 2017.

L. Specia, S. Frank, K. Sima'an, and D. Elliott. A shared task on multimodal machine translation and crosslingual image description. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, 2016.

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

Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. A corpus for reasoning about natural language grounded in photographs. *arXiv preprint arXiv:1811.00491*, 2018.

Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. In *EMNLP*, 2019.

Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. Yfcc100m: The new data in multimedia research. *arXiv preprint arXiv:1503.01817*, 2015.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

Riza Velioglu and Jewgeni Rose. Detecting hate speech in memes using multimodal deep learning approaches: Prize-winning solution to hateful memes challenge. *arXiv preprint arXiv:2012.12975*, 2020.

Prashanth Vijayaraghavan, Hugo Larochelle, and Deb Roy. Interpretable multi-modal hate speech detection. In *AI for Social Good Workshop at the International Conference on Machine Learning*, 2019.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: Lessons learned from the 2015 mscoco image captioning challenge. In *PAMI*, 2016.

Hoa Trong Vu, Claudio Greco, Aliia Erofeeva, Somayeh Jafaritazehjan, Guido Linders, Marc Tanti, Alberto Testoni, Raffaella Bernardi, and Albert Gatt. Grounded textual entailment. *arXiv preprint arXiv:1806.05645*, 2018.

Xin Wang, Devinder Kumar, Nicolas Thome, Matthieu Cord, and Frederic Precioso. Recipe recognition with large multimodal food dataset. In *2015 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, pages 1–6. IEEE, 2015.

Z. Waseem and D. Hovy. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. pages 88–93, 2016.

Z. Waseem, T. Davidson, D. Warmsley, and I. Weber. Understanding abuse: A typology of abusive language detection subtasks. 2017.

Zeerak Waseem. Are you a racist or am I seeing things? annotator influence on hate speech detection on twitter. In *Proceedings of the First Workshop on NLP and Computational Social Science*, pages 138–142, 2016.

Ning Xie, Farley Lai, Derek Doran, and Asim Kadav. Visual entailment: A novel task for fine-grained image understanding. *arXiv preprint arXiv:1901.06706*, 2019.

Fan Yang, Xiaochang Peng, Gargi Ghosh, Reshef Shilon, Hao Ma, Eider Moore, and Goran Predovic. Exploring deep multimodal fusion of text and photo for hate speech classification. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 11–18, 2019.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014.

Fei Yu, Jiji Tang, Weichong Yin, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. Ernie-vil: Knowledge enhanced vision-language representations through scene graph. *arXiv preprint arXiv:2006.16934*, 2020.

J. Yu and J. Jiang. Adapting bert for target-oriented multimodal sentiment classification. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pages 5408–5414, 2019.

T. Zahavy, A. Magnani, A. Krishnan, and S. Mannor. Is a picture worth a thousand words? a deep multi-modal fusion architecture for product classification in e-commerce. *arXiv:1611.09534*, 2016.

R. Zellers, Y. Bisk, A. Farhadi, and Y. Choi. pages 6720–6731, 2019.

Mingda Zhang, Rebecca Hwa, and Adriana Kovashka. Equal but not the same: Understanding the implicit relationship between persuasive images and text. *arxiv:1807.08205*, 2018.

Haoti Zhong, Hao Li, Anna Squicciarini, Sarah Rajtmajer, Christopher Griffin, David Miller, and Cornelia Caragea. Content-driven detection of cyberbullying on the instagram social network. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, pages 3952–3958, 2016.

Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason J. Corso, and Jianfeng Gao. Unified vision-language pre-training for image captioning and vqa. In *AAAI*, 2019.

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