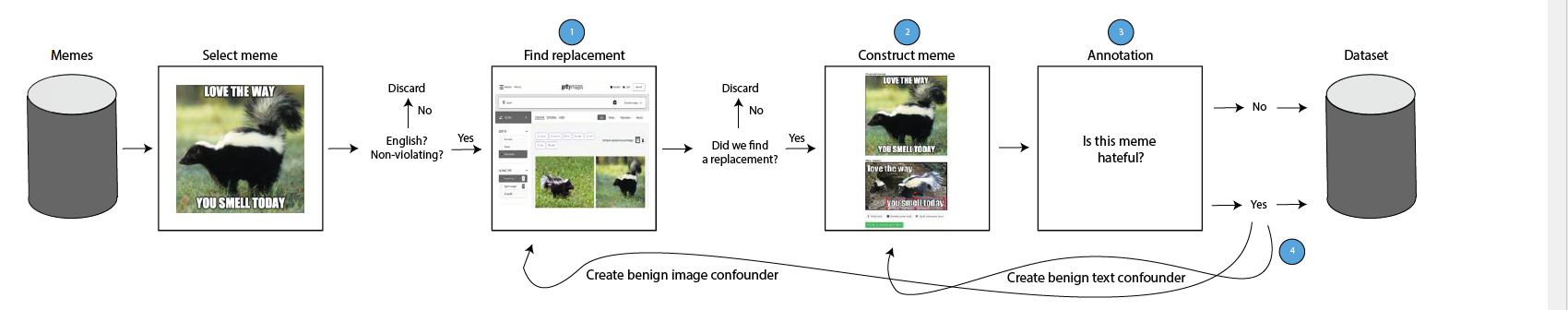
**Introduction**

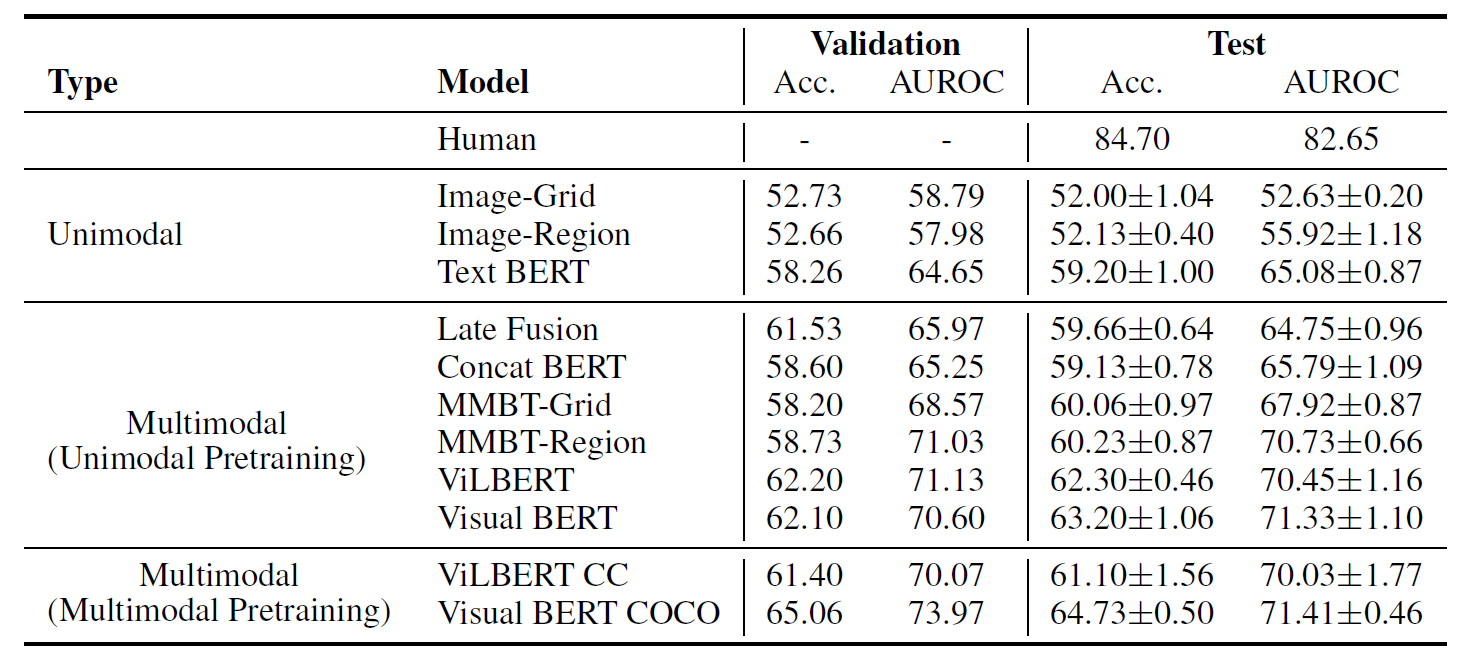
1. Multimodal classification 🡪 multimodal hate memes.
2. Dataset: include difficult examples, hard to rely on unimodal signals.
3. Evaluate it as a binary classification problem.
4. Baseline: (1) unimodal model; (2) multimodal model with sophistication
5. Accuracy: SOTA: 64.73%; human: 84.7%
6. Meme type: (1) mean; (2) benign image confounders; (3) benign text confounders
7. Challenge purpose: (1) measure progress on multimodal understanding and reason; (2) improve hate speech detection.

**Hateful memes dataset-10K**

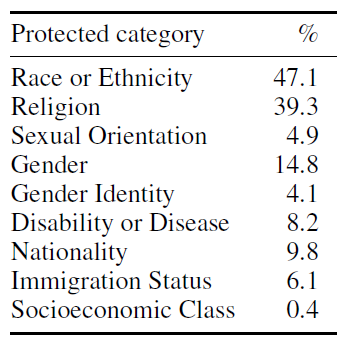
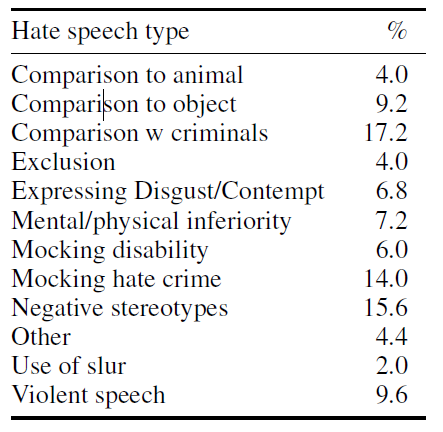
1. **Hateful memes dataset**: not to train model from scratch, but rather to finetune and test large scale pre-trained multimodal model
2. **Hatefulness definition**: A direct or indirect attack on people based on characteristics, including ethnicity, race, nationality, immigration status, religion, 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.
3. **Hatefulness exception**: attacking individuals/famous people is allowed if the attack is not based on any of the protected characteristics listed in the definition. Attacking groups perpetrating hate (e.g. terrorist groups) is also not considered hate. This means that hate speech detection also involves possibly subtle world knowledge.
4. **Annotation process**:



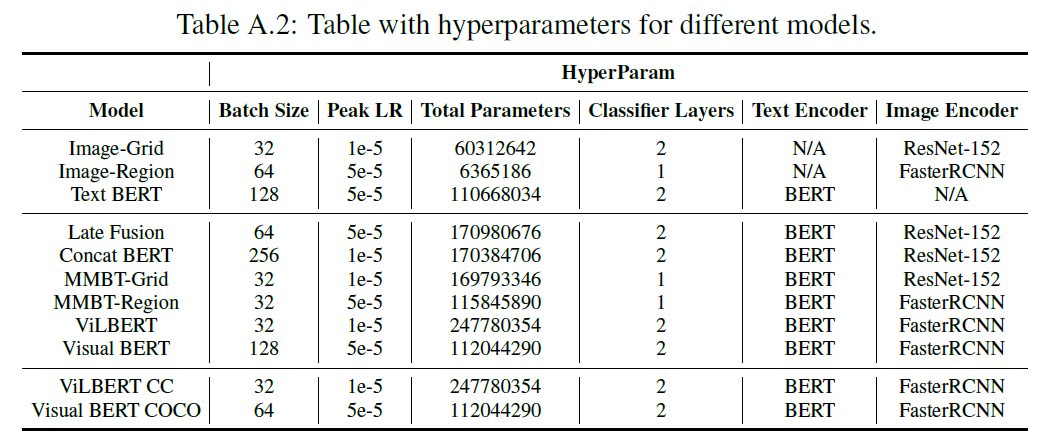
1. **Benign confounder**: make the dataset more challenging, have to use multimodal to solve this task.
2. **Five Data types**: (1) multimodal hate, benign confounders are found for both modalities; (2) unimodal hate, one or both modalities are already hateful on their own; (3) benign image confounders; (4) benign text confounders; (5) random not-hateful.
3. **Dev (5%) and test data (10%) percentage:** balanced; 40% multimodal hate, 10% unimodal hate, 20% benign text confounder, 20% benign image confounder, 10% random non-hateful. Training data: 85%.
4. **Task object:** given an image and pre-extracted text (no OCR needed), classify memes according to their hatefulness.
5. **Evaluation metric:** (1) main metric: the area under the receiver operating characteristic curve (ROC AUC); (2) auxiliary metric: accuracy.
6. **Baseline:** multimodal-unimodal pretraining, separate pretrained text and image model



1. **Hate speech type (classes) and protected category**:



1. **Hyperparameters for different models**:



1. **Starting code**:https://github.com/facebookresearch/mmf/tree/master/projects/hateful\_memes
2. Start from FB MMF framework

https://colab.research.google.com/github/facebookresearch/mmf/blob/notebooks/notebooks/mmf\_hm\_example.ipynb#scrollTo=1mB-z-6XWdBd

1. Since the dataset is designed to tackle the task with multimodal techniques, we will first replicate the baseline multimodal models. Specifically, we will pick the model in multimodal (unimodal pretraining) with the highest accuracy and AUROC, which is Visaual BERT. Then we will pick another model in multimodal (multimodal pretraining), which is Visual BERT COCO. The purpose of this trial is to have a sense of how SOTA behaves.
2. Pick one of the top submission from released github repo

<https://github.com/drivendataorg/hateful-memes/>