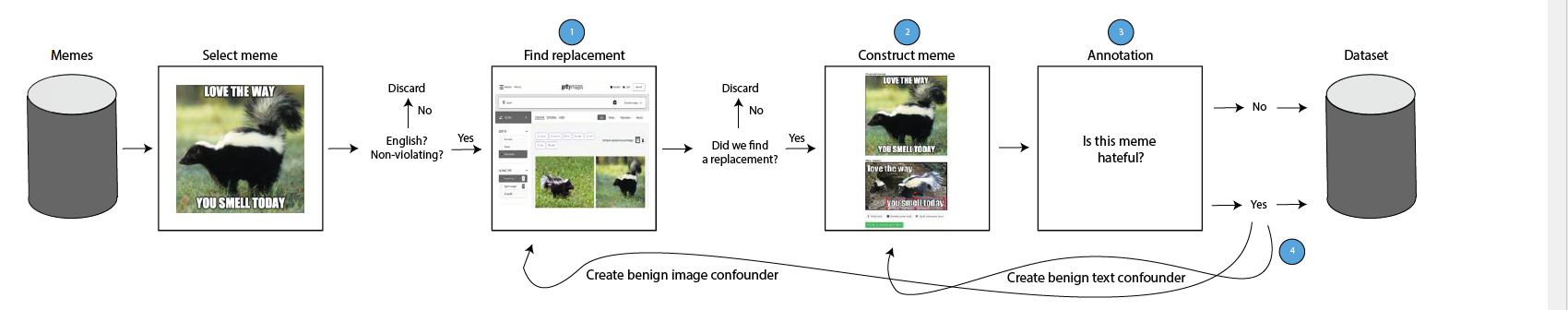
**Introduction**

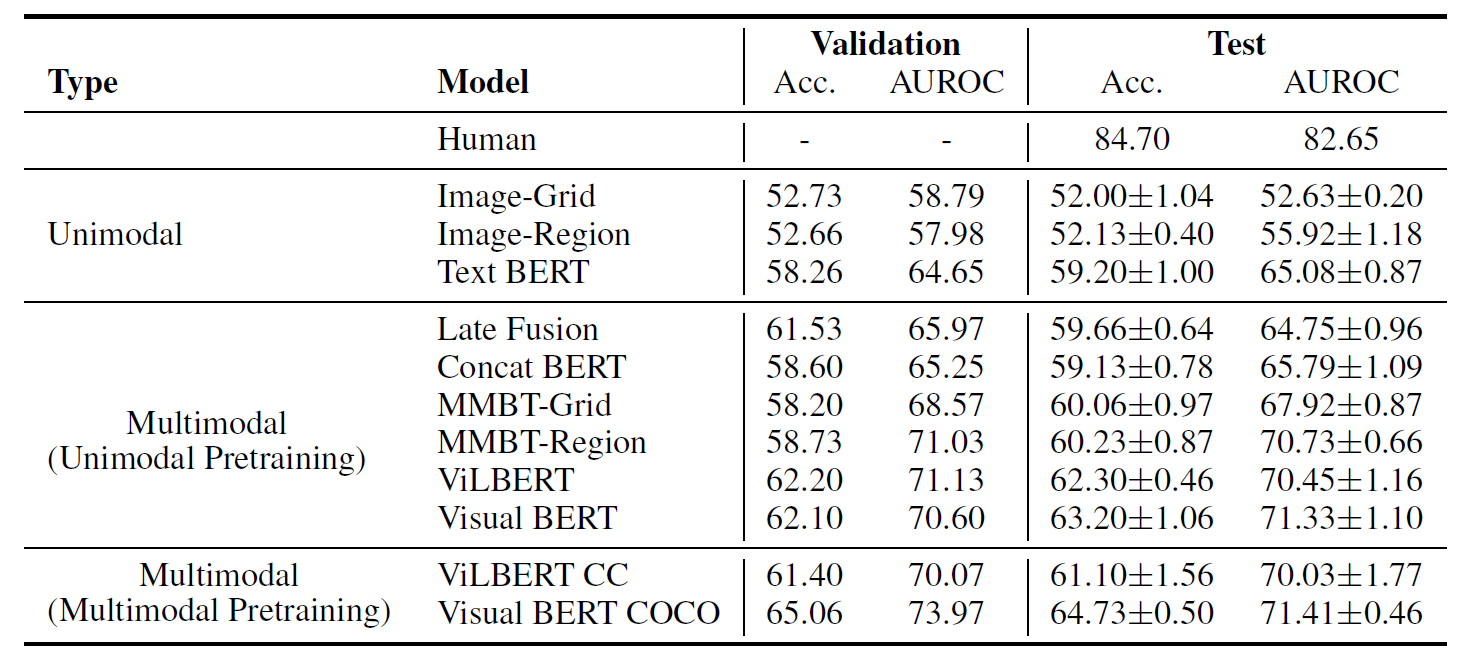
1. Multimodal classification 🡪 multimodal hate memes.
2. Dataset: include difficult examples, hard to rely on unimodal signals.
3. Evaluate it as a binary classification problem.
4. Baseline: (1) unimodal model; (2) multimodal model with sophistication
5. Accuracy: SOTA: 64.73%; human: 84.7%
6. Meme 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**

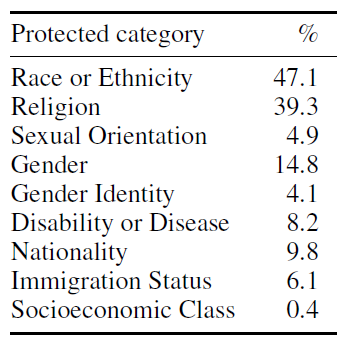
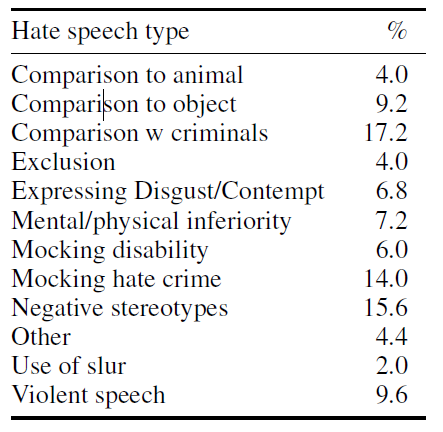
1. **Hateful memes dataset**: not to train model from scratch, but rather to finetune and test large scale pre-trained multimodal model
2. **Hatefulness definition**: A direct or indirect attack on people based on characteristics, including ethnicity, race, nationality, immigration status, religion, 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**:



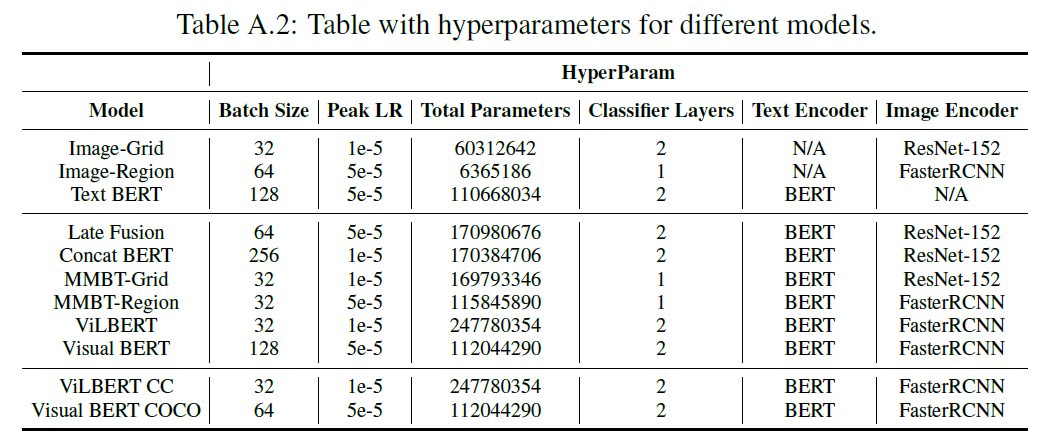
1. **Benign confounder**: make the dataset more challenging, have to use multimodal to solve this task.
2. **Five Data types**: (1) multimodal hate, benign confounders are found for both modalities; (2) unimodal hate, one or both modalities are already hateful on their own; (3) benign image confounders; (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

Challenge homepage: <https://www.drivendata.org/competitions/64/hateful-memes/page/205/>

Dataset: <https://www.drivendata.org/competitions/64/hateful-memes/data/>

a binary classification problem with multimodal input data consisting of the meme image itself (the image mode) and a string representing the text in the meme image (the text mode).

Given a meme id, meme image file, and a string representing the text in the meme image, your trained model should output the probability that the meme is hateful.

**Dataset features**

Separate meme images and text extractions. Use meme id to match them.

Meme images are in the data/img folder. Different .json files contains the information of (id, image number, label, text).

train.jsonl: training set

dev\_seen.jsonl/dev\_unseen.jsonl: validation set

test\_seen.jsonl/test\_unseen.jsonl: test set

visualize the distribution of hateful memes (label 1) and non-hateful memes (label 0)

**performance metric**

1. AUC ROC: Area Under the Curve of the Receiver Operating Characteristic. The metric measures how well your binary classifier discriminates between the classes as its decision threshold is varied.

[sklearn.metrics.roc\_auc\_score](https://scikit-learn.org/stable/modules/generated/sklearn.metrics.roc_auc_score.html).

<https://medium.com/greyatom/lets-learn-about-auc-roc-curve-4a94b4d88152>

1. Accuracy: given by the ratio of correct predictions to the total number of predictions made