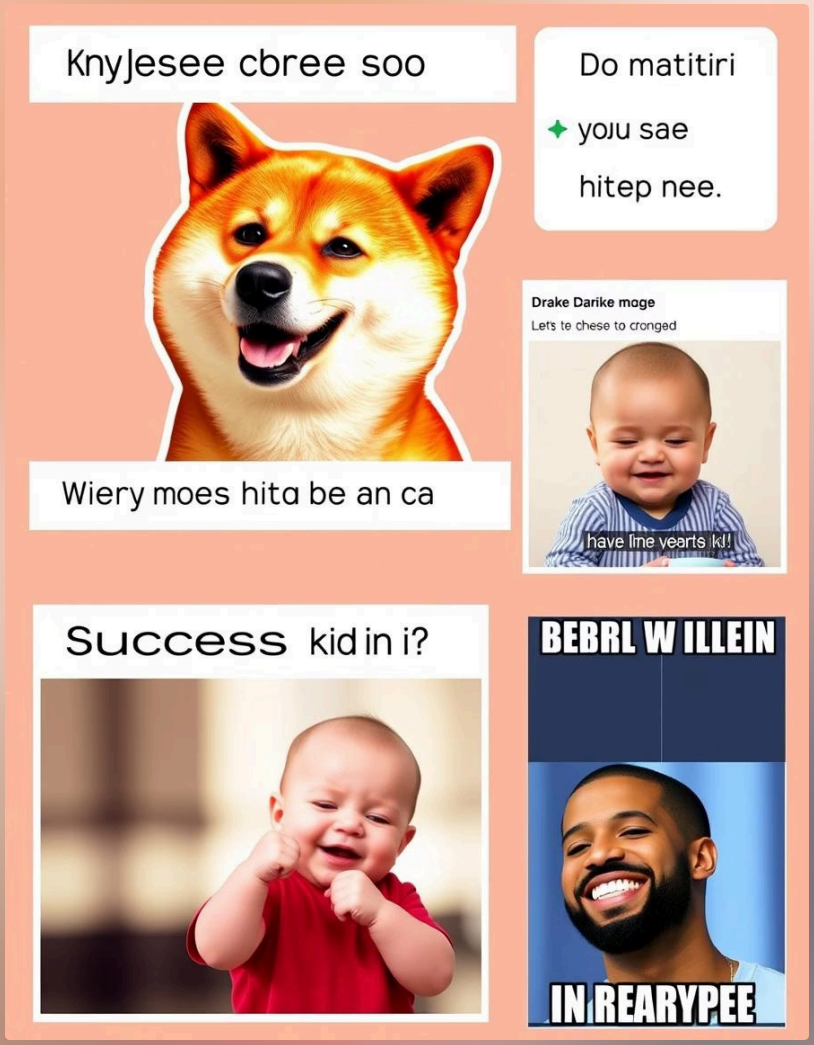


The Hateful Memes Challenge: Detecting Hate Speech in Multimodal Memes

This presentation introduces the Hateful Memes Challenge, a new benchmark for detecting hate speech in multimodal memes. The dataset is designed to be challenging for unimodal models by including "benign confounders" that flip labels, requiring sophisticated multimodal reasoning. The task is a binary classification problem with real-world relevance in combating online hate speech.

We provide baseline results for various unimodal and multimodal models, showing a significant gap between AI and human performance, highlighting the challenge's difficulty and importance.

by Venkata Koushik Nagasarapu

Challenge Set Construction and Hate Speech Definition

Hate Speech Definition

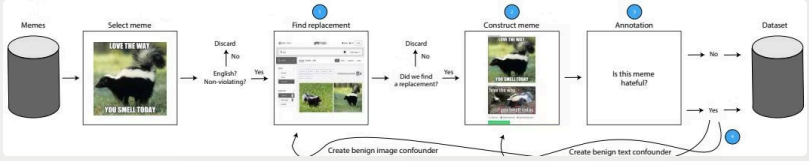
Defined as direct or indirect attacks on people based on protected characteristics like race, gender, religion, or disability. Attacks include violent, dehumanizing, or exclusionary speech.

Dataset Creation

Memes are reconstructed using licensed Getty Images to avoid copyright issues and reduce visual bias, ensuring semantic preservation of original memes.

Annotation Process

Trained annotators rated memes for hatefulness, with multiple phases including filtering, meme reconstruction, rating, and creation of benign confounders.



Annotation and Filtering Process

1

Phase 1: Filtering

From 1 million images, 162k memes were filtered for English text, non-violence, and no slurs, resulting in 46k candidates with suitable replacement images.

2

Phase 2: Meme Construction

Annotators recreated memes using new images and original text, preserving meaning, using a custom tool storing PNG and SVG formats.

3

Phase 3: Hatefulness Rating

Each meme was rated by five annotators on a 1-3 scale, with expert review for disagreements, yielding binary hate labels.

4

Phase 4: Benign Confounders

For hateful memes, alternative images or texts were created to flip labels to non-hateful, ensuring the need for multimodal reasoning.

Dataset Composition and Task Objective

Dataset Types

- Multimodal hate (40%)
- Unimodal hate (10%)
- Benign text confounders (20%)
- Benign image confounders (20%)
- Random non-hateful (10%)

Task Objective

Given an image and pre-extracted text, classify memes as hateful or not. Evaluation uses ROC AUC and accuracy on balanced dev and test sets.

A competition was held with an unseen test set to benchmark progress.

Inter-Annotator Agreement and Dataset Analysis

Inter-Annotator Agreement

Cohen's kappa score of 68.4 indicates moderate agreement, reflecting the complexity of hate speech detection and nuanced definitions.

Hate Categories

Most prevalent categories are race and religion-based hate, with multiple protected categories per meme possible.

Types of Attack

Includes dehumanization, negative stereotypes, mocking hate crimes, and violent speech, with comparison to criminals and objects common.

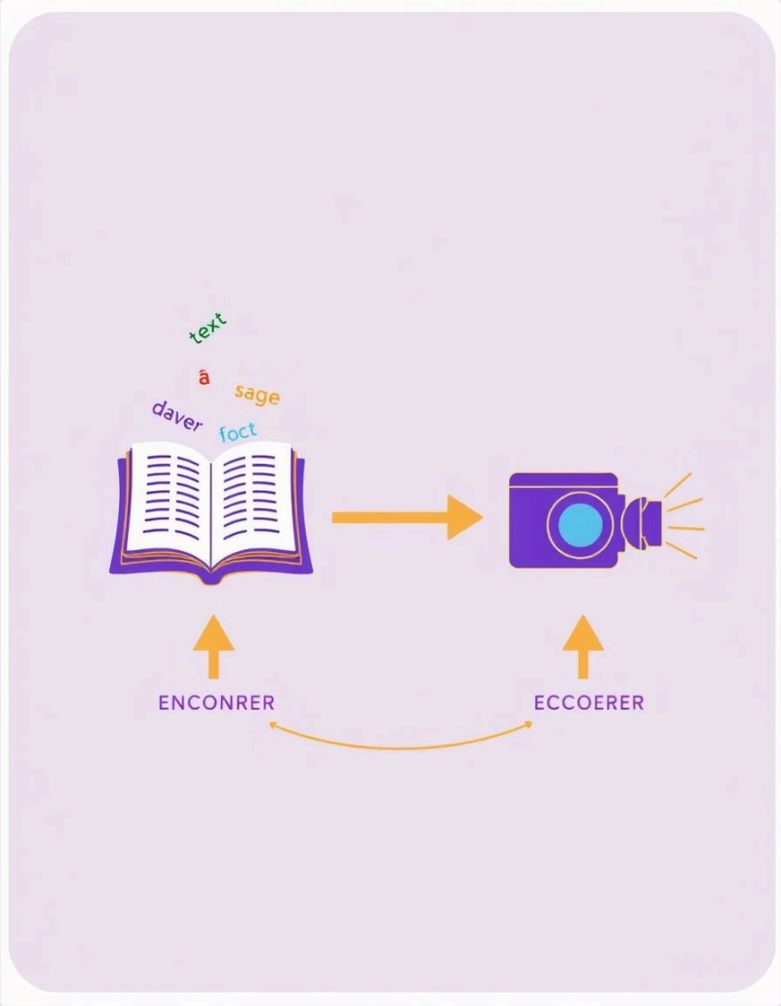
Model Baselines and Performance

Model Types

- Unimodal models (image or text only)
- Multimodal models with unimodal pretraining
- Multimodal models with multimodal pretraining

Performance Highlights

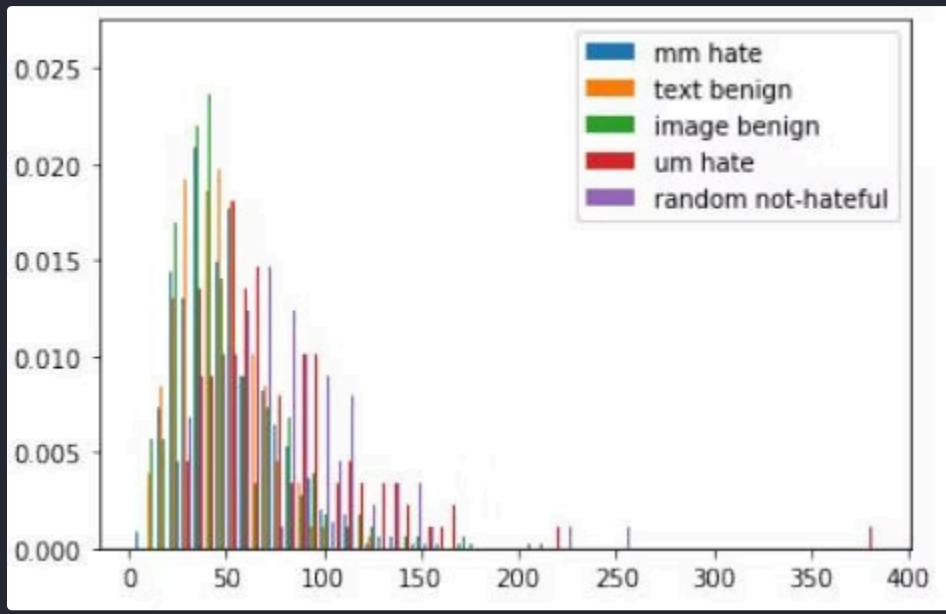
Text-only models outperform image-only. Multimodal models perform better, with early fusion methods leading. Best models still lag behind human accuracy (84.7%).

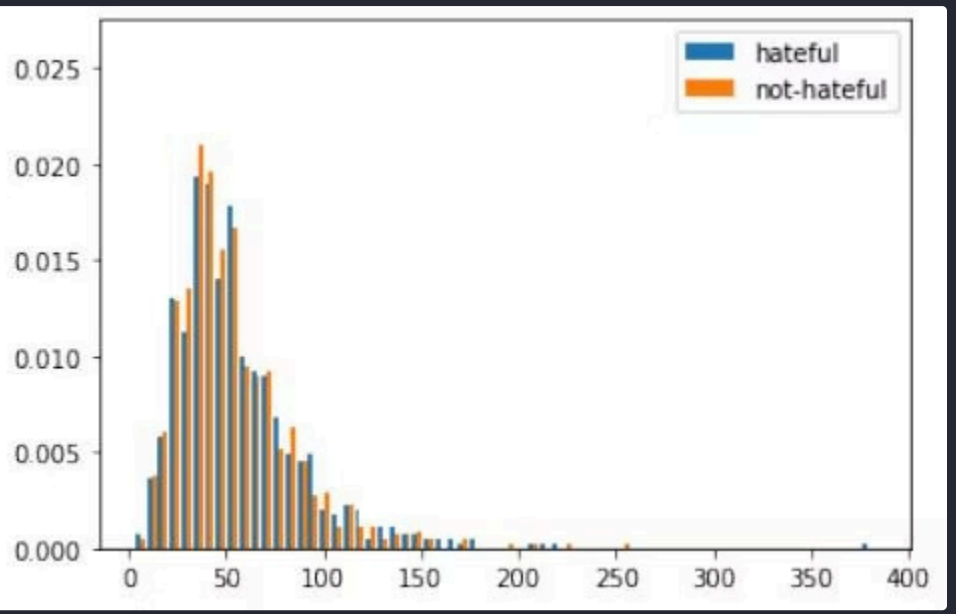


Model Details and Hyperparameters

Model	Batch Size	Learning Rate	Parameters
Image-Grid	32	1e-5	60M
Image-Region	64	5e-5	6M
Text BERT	128	5e-5	110M
Late Fusion	64	5e-5	170M
Concat BERT	256	1e-5	170M
MMBT-Grid	32	1e-5	169M
MMBT-Region	32	5e-5	115M
ViLBERT	32	1e-5	247M
Visual BERT	128	5e-5	112M

Lexical Statistics of Text and Visual Modalities





Textual Analysis

Frequent words include "people," "like," "black," and "white," with stronger language in unimodal hate targeting religious groups. Some words appear in benign confounders, showing complexity.

Visual Analysis

Common objects detected include person, tie, car, dog, and chair. Visual modality is balanced but biased, adding complexity to multimodal classification.

Related Work in Hate Speech and Multimodal Detection

Text-Only Hate Speech

Extensive research exists with various datasets and classifiers, but definitions and biases vary widely.

Multimodal Hate Speech

Few datasets combine text and images; prior work shows image features improve detection. Our dataset is larger, balanced, and designed to challenge unimodal models.

Vision and Language Tasks

Multimodal classification differs from generation tasks and is critical for real-world applications like content moderation on social media.

Conclusion and Broader Impact

The Hateful Memes Challenge introduces a difficult benchmark for multimodal hate speech detection, requiring subtle reasoning beyond unimodal cues. Current models lag behind humans, indicating room for progress.

This work aims to advance multimodal understanding and help address societal issues like online hate speech. While beneficial, improved AI systems may also pose risks such as job automation or misuse, which require careful mitigation.