Facebook Hateful Meme Challenge

Abhishek Das, Japsimar Wahi, Siyao Li

11777 F20 Group Project

Carnegie Mellon University

Introduction

- Detecting Hate-Speech in Multimodal Memes.
- Classify Memes as Hateful or Benign.
- Interpret reasoning behind Images and Caption

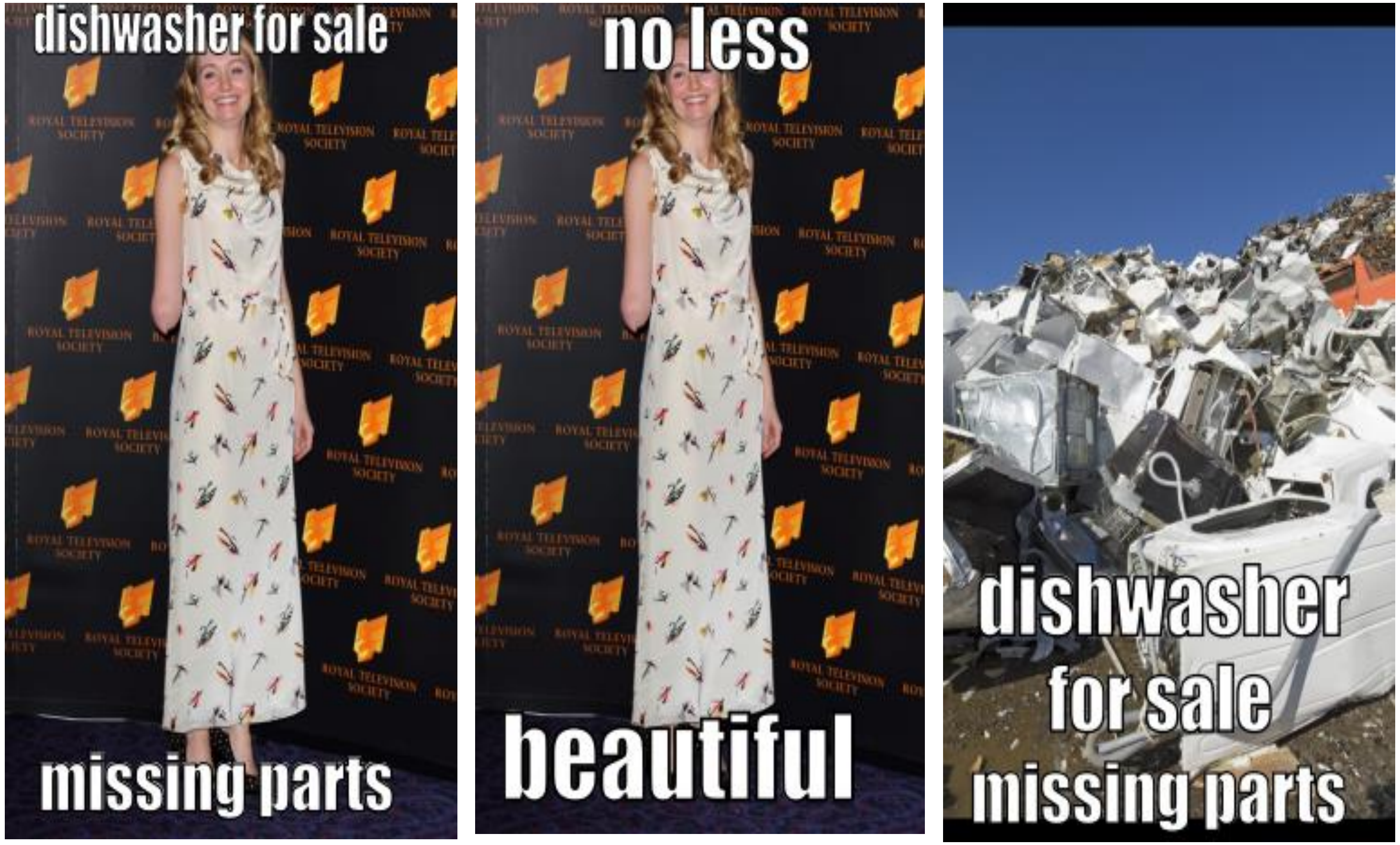


Figure 1: Multimodal "mean" meme and Benign confounders.
Mean meme (left), Benign image confounder (middle) and Benign text confounder (right)

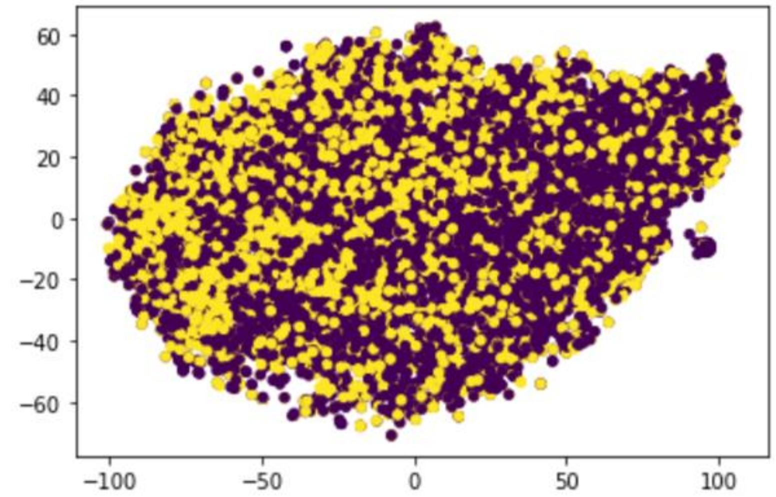
Challenges

- Dataset is designed such that such that models exploiting Unimodal priors fail

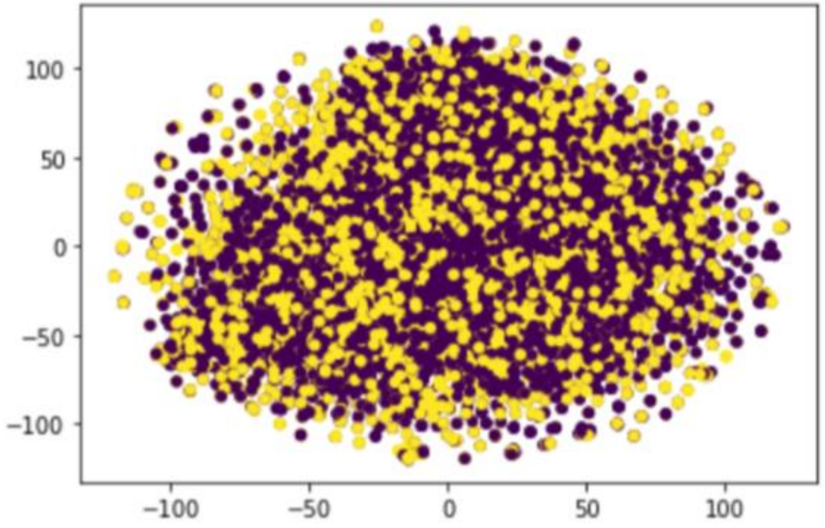
- Benign confounders flip the label from hateful to benign

- A same image/caption can be used to create both hateful and benign meme

T-SNE on Language Modality

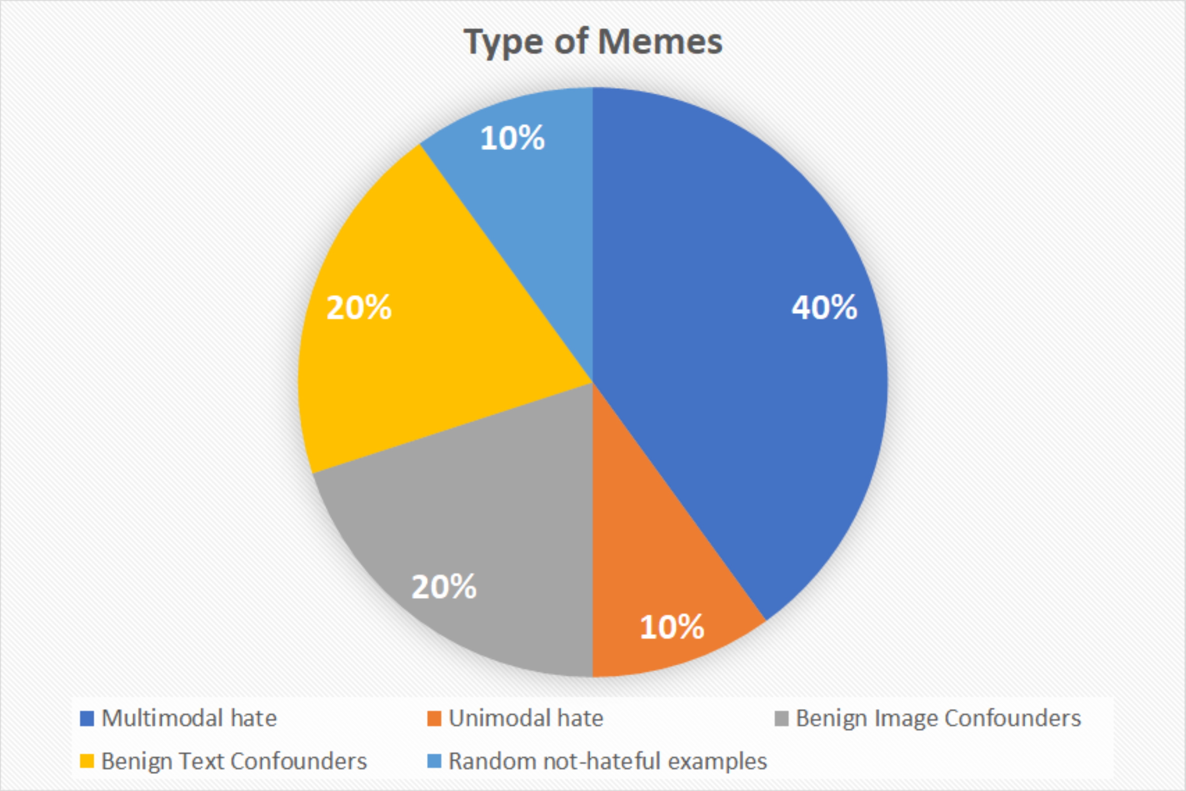


T-SNE on Visual modality



Dataset and Evaluation

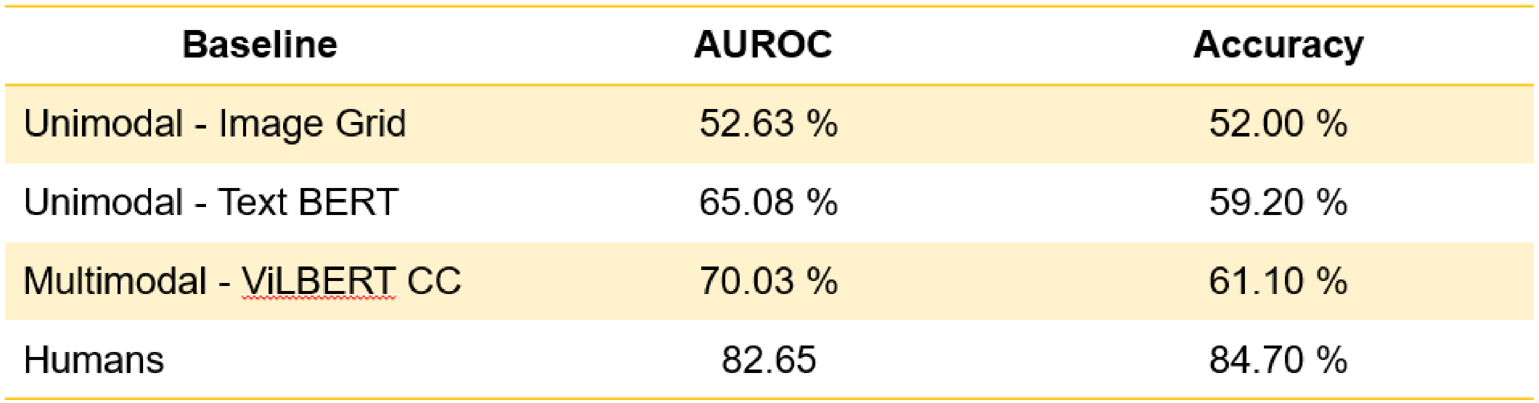
- Facebook Hateful Meme Challenge set of 10k Memes
- Designed by annotators trained for Hate-Speech
- Fully Balanced Training, Validation and Test set



- Metrics
  - Area under the Receiver Operating Characteristics (ROC AUC)
  - Classification Accuracy on Test set

Related Work

- Selecting features and bootstrapping for small cyberbullying dataset.
- Pre-Training on Large-scale Hate-Speech Detection dataset like MMHS150K created from Twitter
- Augmenting Text with Image embedding information followed by attention fusion methods.



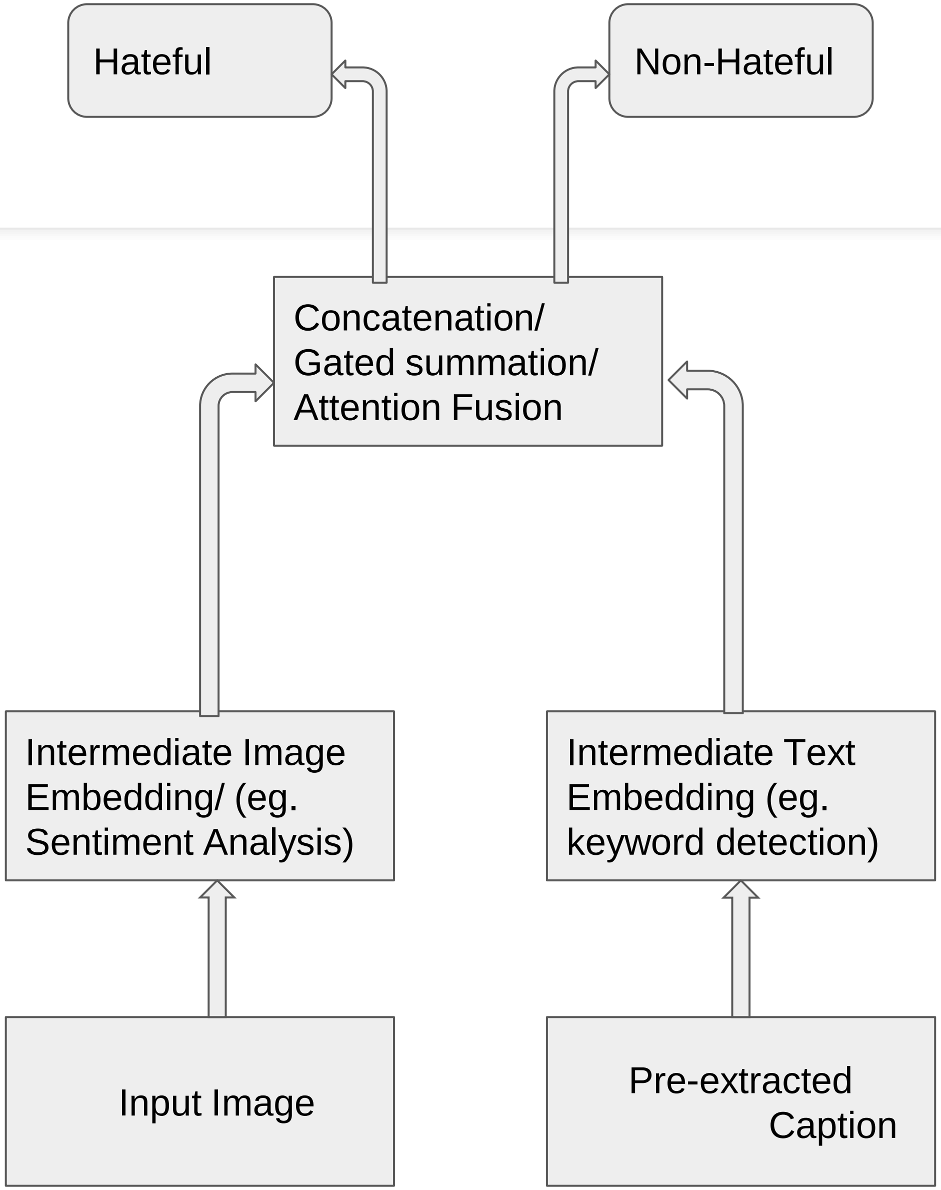
Baseline	AUROC	Accuracy
Unimodal - Image Grid	52.63 %	52.00 %
Unimodal - Text BERT	65.08 %	59.20 %
Multimodal - ViLBERT CC	70.03 %	61.10 %
Humans	82.65	84.70 %

Idea 1

- Finding Intermediate embeddings of both text and image modality to find useful information.

- Fusion of these unimodal important information with techniques like concatenation, gated summation.

- Applying attention or co-attention fusion methods on the system.



Hateful
Non-Hateful
Concatenation/ Gated summation/ Attention Fusion
Intermediate Image Embedding/ (eg. Sentiment Analysis)
Intermediate Text Embedding (eg. keyword detection)
Input Image
Pre-extracted Caption

Other Ideas

- Extending and fine-tuning of Bilinear Attention Models (Popularly being used in VQA systems) for our use case.
- Majority of the datasets have text on the top and on the bottom (separate analysis creating two separate embeddings).
- To learn how different words in text or different objects in images have a possibility of coming together. Ex. small girl with a gun, girl with no arm (similarly in text).
- Pre-Training on similar dataset including both unimodal and multimodal training.

Thank You!