Detecting Hate Speech in Multimodal Memes

Abhishek Das, Japsimar Wahi, Siyao Li

11777 F20 Group Project

Carnegie Mellon University

Introduction

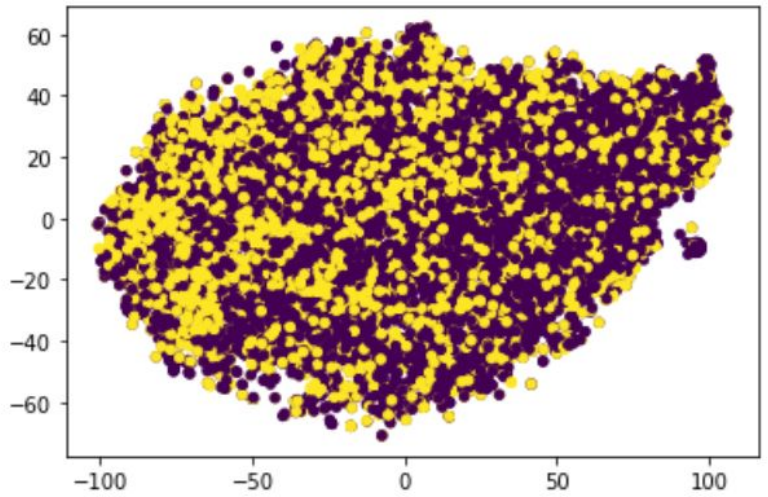
- Detecting Hate-Speech in Multimodal Memes.
- Classify Memes as Hateful or Benign.
- Interpret reasoning behind Images and Caption

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

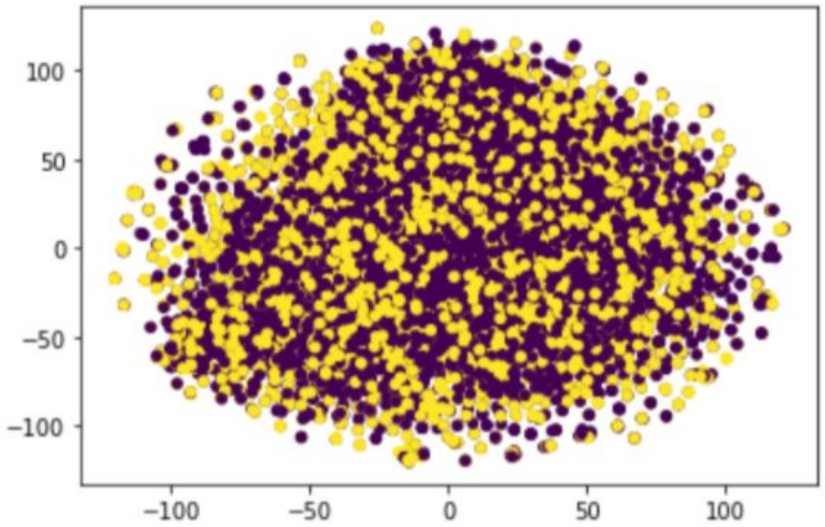
Challenges

- 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

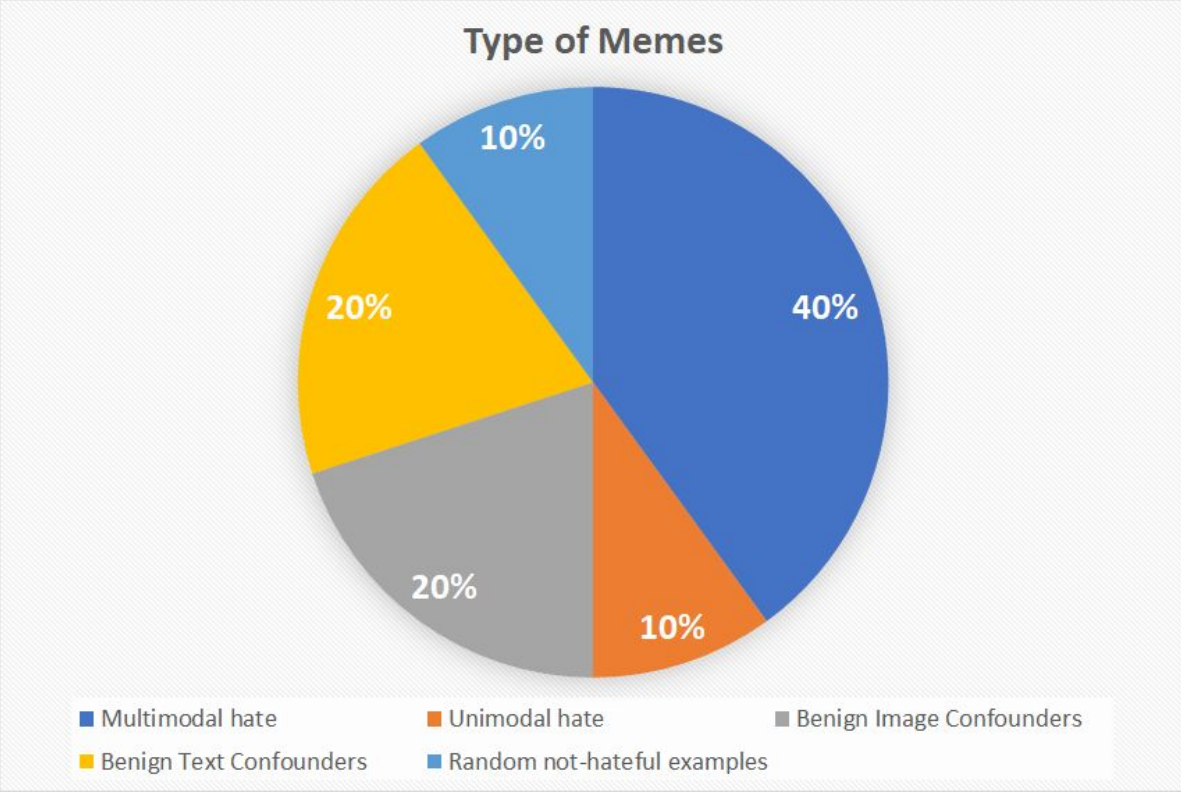


T-SNE on Visual modality



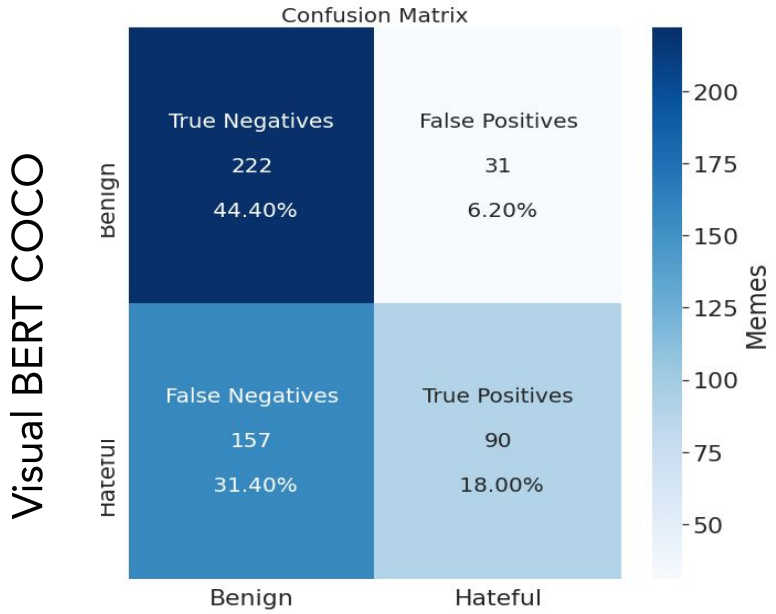
Dataset and Evaluation

- 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

Data analysis



Precision: 74.38 Recall: 36.44 F1 Score: 48.91

Accuracy: 62.4 AUROC:0.62

Annotated label: Benign Predicted label: Hateful Probability: 0.9995

False Positive

Annotated label: Hateful Predicted label: Benign Probability: 0.0002

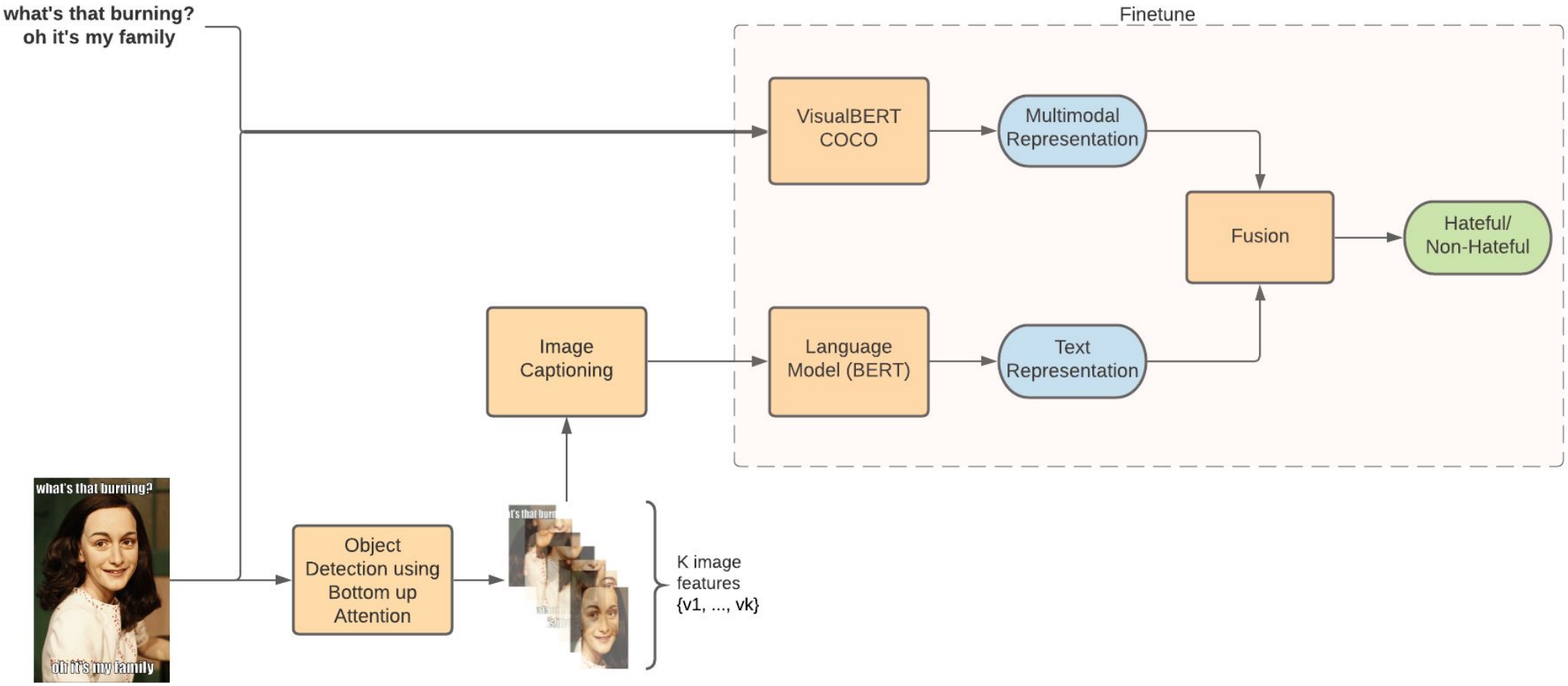
False Negative

Why Captioning?

Hateful Meme Benign Text Confounder

Hateful Meme Benign Text Confounder

Using Object Detection and Image Captioning



Examples and their Explanations - Image Captioning

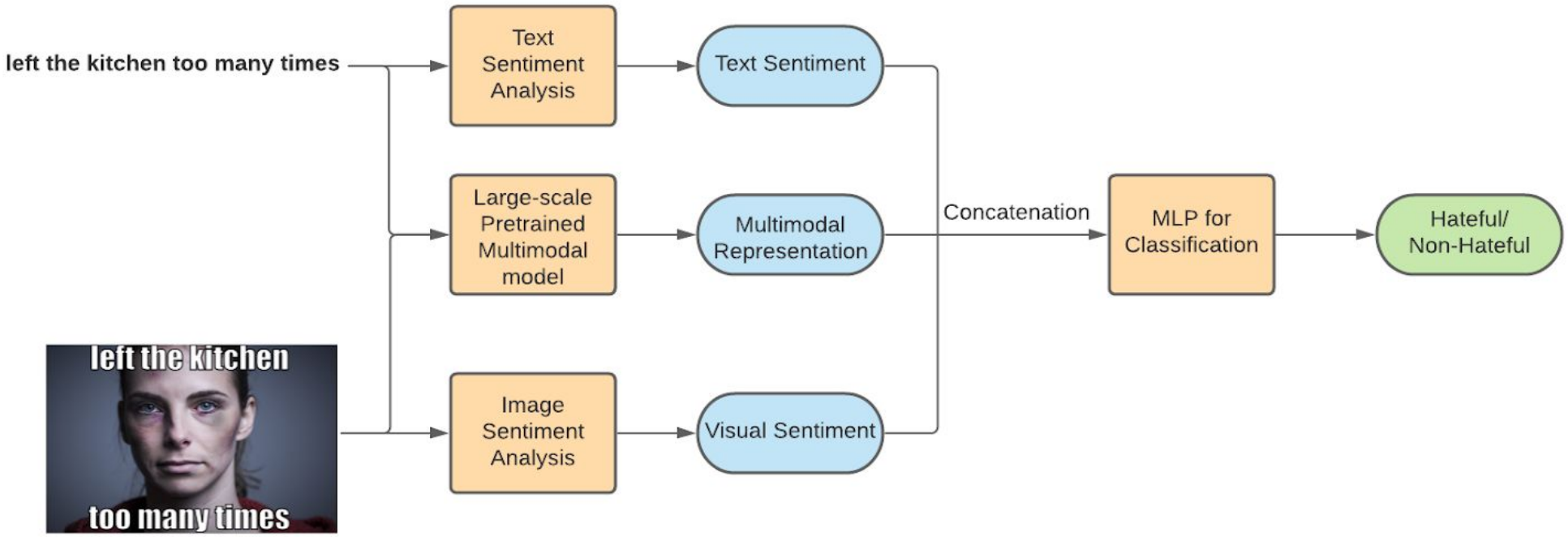
Hateful Meme

Benign Text Confounder

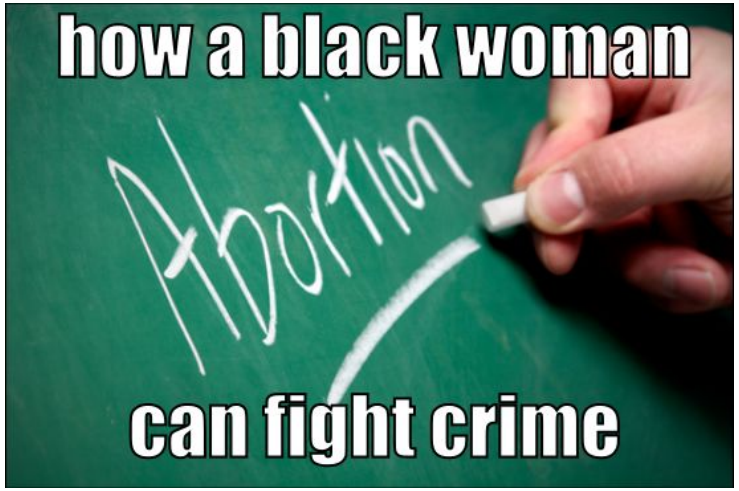
Object Detection Visualization

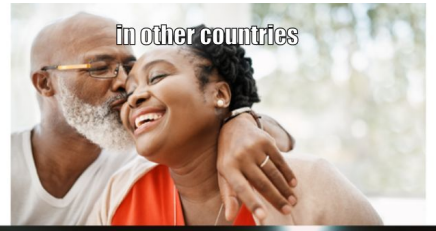
- Original Meme: Hateful
- Benign Confounder (Test Image) : made Non-Hateful by describing the image.
- Object Detection/ Image Captioning: Helps in finding important objects and captioning it.
- Visual Bert COCO Prediction: Hateful
- Our Model - Labels it as Benign with the help of Image Captioning.

Using Sentiment Analysis



Examples and their Explanations - Sentiment Analysis





- Benefits:

  Detect Irony

  Confirm Benign Memes

- Problems:

  The accuracy of sentiment prediction is low

  Doesn't work well in some complicated cases (when both sentiment are negative

Results and Experiments (on the Competition Leaderboard)