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

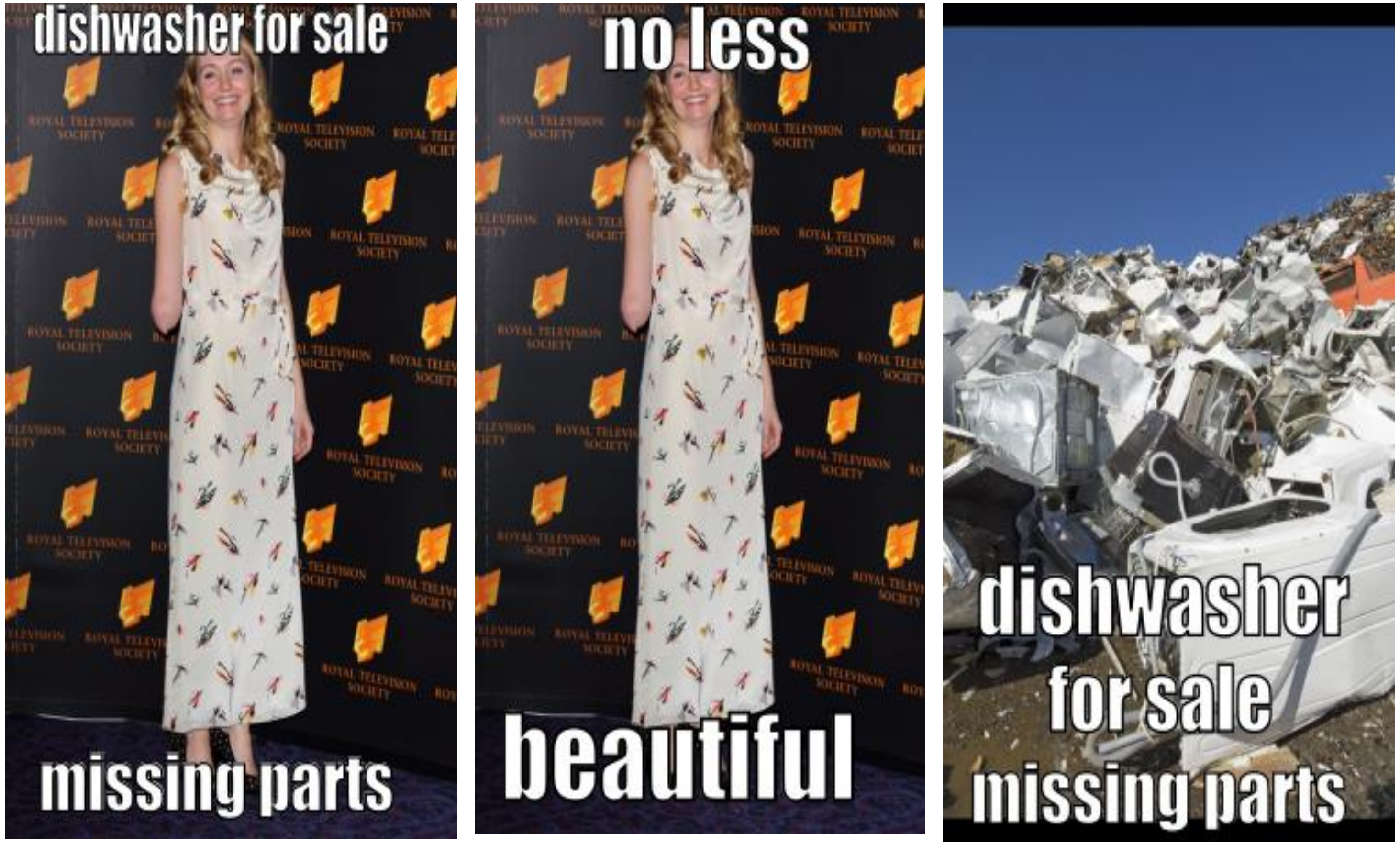
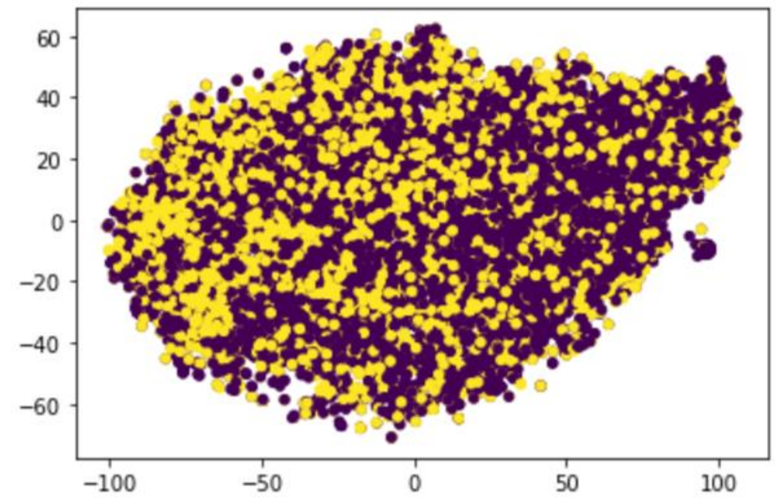


Figure 1: Multimodal "mean" meme and Benign confounders.
Mean meme (left), Benign text confounder (middle) and Benign image confounder (right)

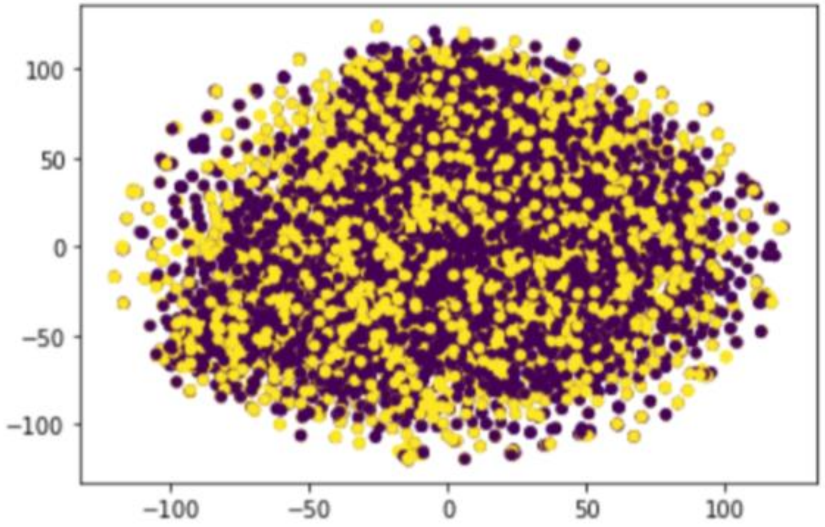
Challenges

- 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

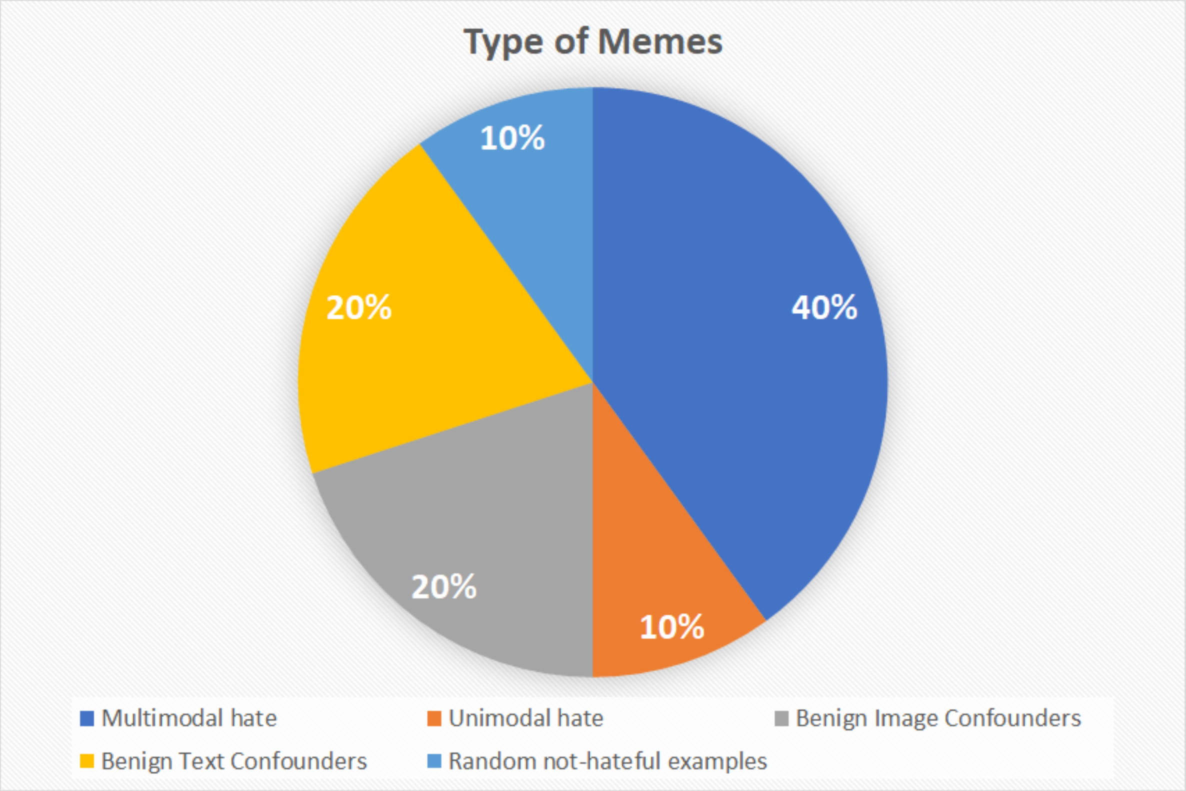


T-SNE on Visual modality



Dataset and Evaluation

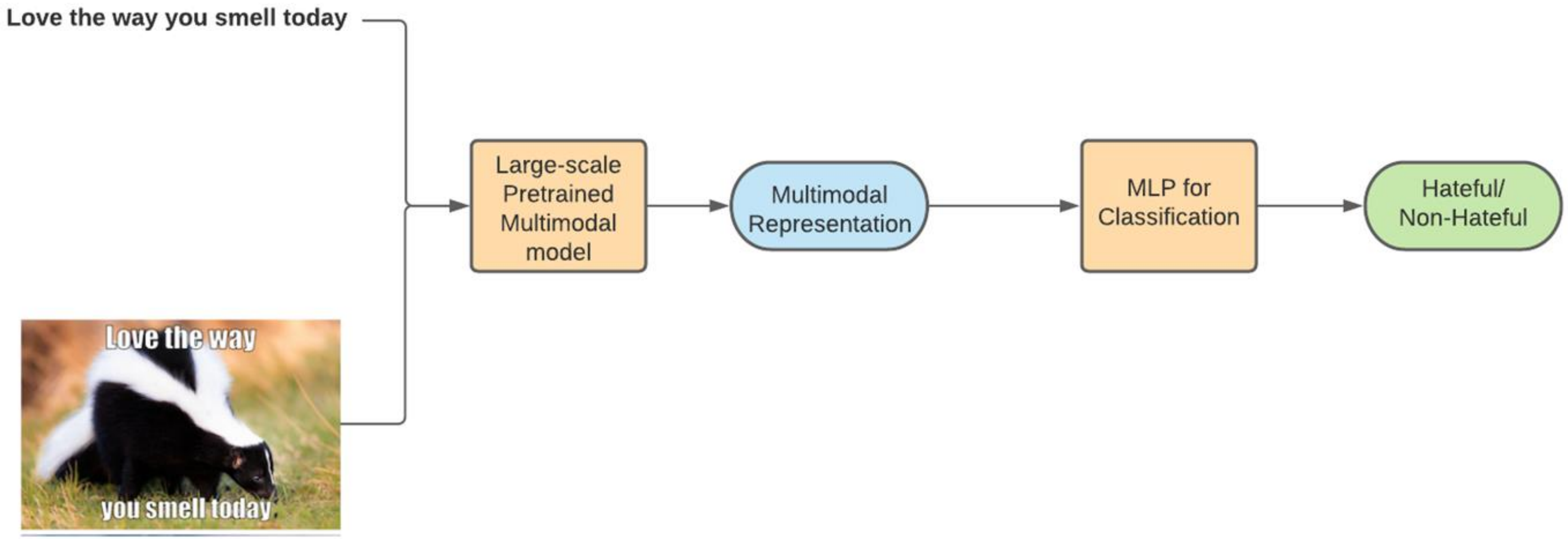
- Facebook Hateful Meme Challenge set of 10k Memes
- Designed by annotators trained for Hate-Speech
- Fully Balanced Training, Validation and Test set



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

Baseline Models

Pretrained multimodal representations are drawn from: **1) Visual Bert 2) ViLBERT**


Love the way you smell today
Love the way
you smell today
Large-scale Pretrained Multimodal model
Multimodal Representation
MLP for Classification
Hateful/ Non-Hateful

(Kiela et al., 2020)

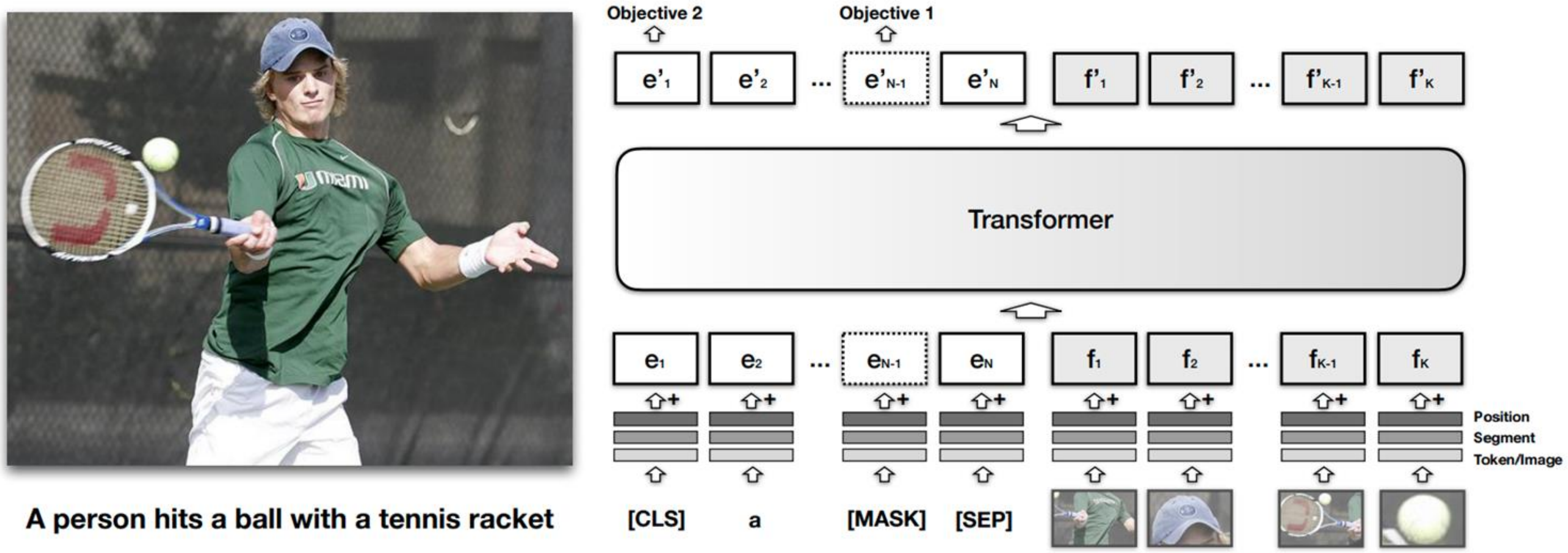
Baseline Models: Visual BERT

Using transformer to discover implicit alignments

Objective 1:
masked language modeling

Objective 2: sentence-image matching

Pretrained on caption data



(Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh & Kai-Wei Chang)

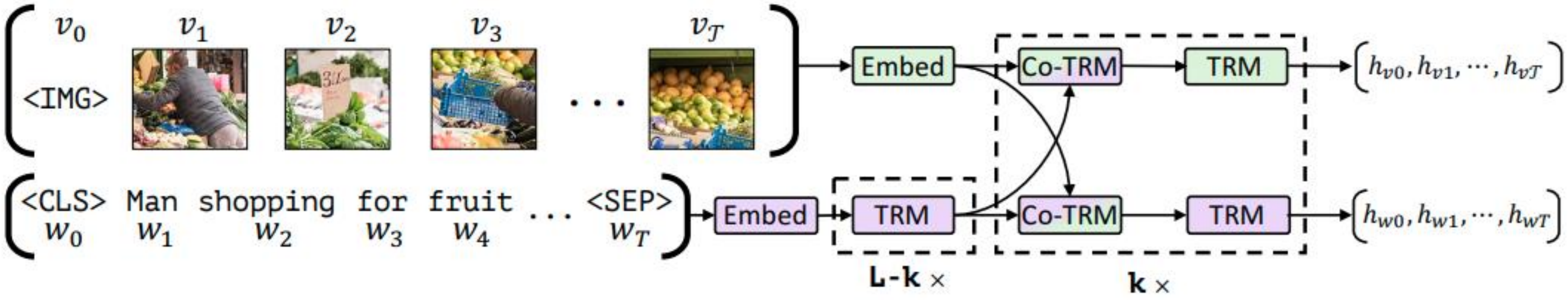
Baseline Models: VilBERT

Co-attention transformer

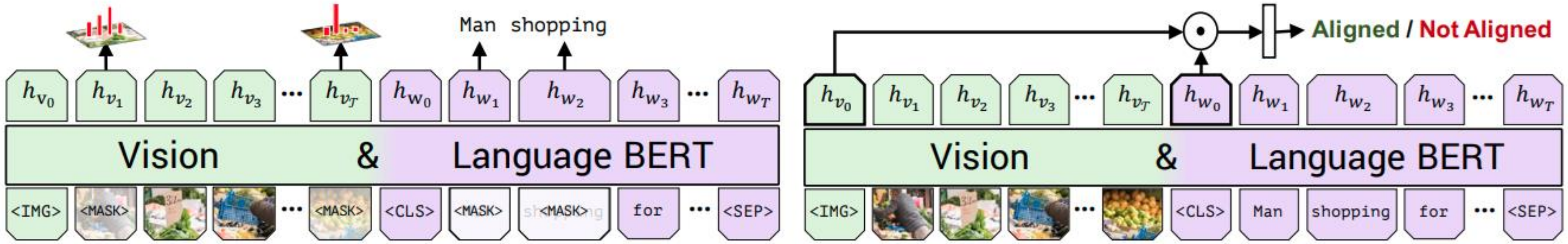
Objective 1: masked multimodal modeling

Objective 2: sentence-image matching

Pretrained on the Conceptual Captions dataset



v_0
v_1
v_2
v_3
v_T
<IMG>
...
<CLS> Man shopping for fruit ... <SEP>
w_0 w_1 w_2 w_3 w_4 w_T
Embed
Embed
TRM
Co-TRM
TRM
Co-TRM
TRM
L-k ×
k ×
$(h_{v0}, h_{v1}, \cdots, h_{vT})$
$(h_{w0}, h_{w1}, \cdots, h_{wT})$



Man shopping
h_{v_0} h_{v_1} h_{v_2} h_{v_3} ... h_{v_T} h_{w_0} h_{w_1} h_{w_2} h_{w_3} ... h_{w_T}
Vision & Language BERT
<IMG> <MASK> <MASK> <CLS> <MASK> <MASK> for ... <SEP>
Aligned / Not Aligned
h_{v_0} h_{v_1} h_{v_2} h_{v_3} ... h_{v_T} h_{w_0} h_{w_1} h_{w_2} h_{w_3} ... h_{w_T}
Vision & Language BERT
<IMG> <CLS> Man shopping for ... <SEP>

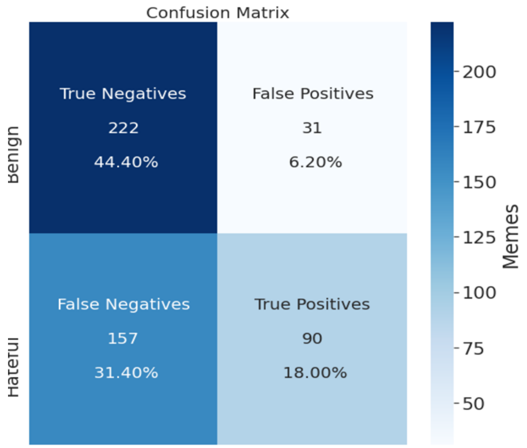
(a) Masked multi-modal learning

(b) Multi-modal alignment prediction

(Jiasen Lu, Dhruv Batra, Devi Parikh, Stefan Lee)

Error Analysis - Validation set of 500 Memes

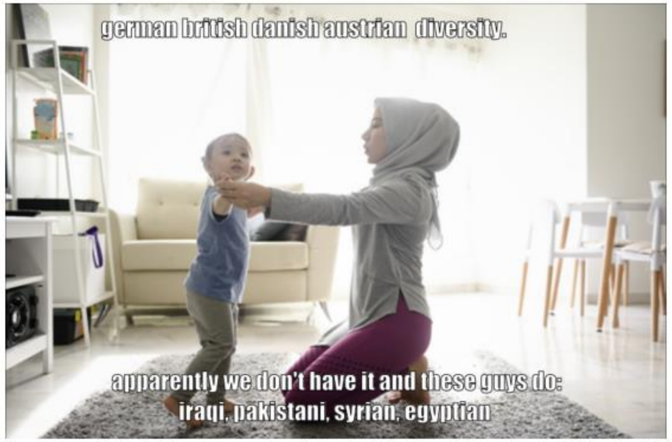
Visual BERT COCO



	Benign	Hateful
Benign	True Negatives 222 44.40%	False Positives 31 6.20%
Hateful	False Negatives 157 31.40%	True Positives 90 18.00%

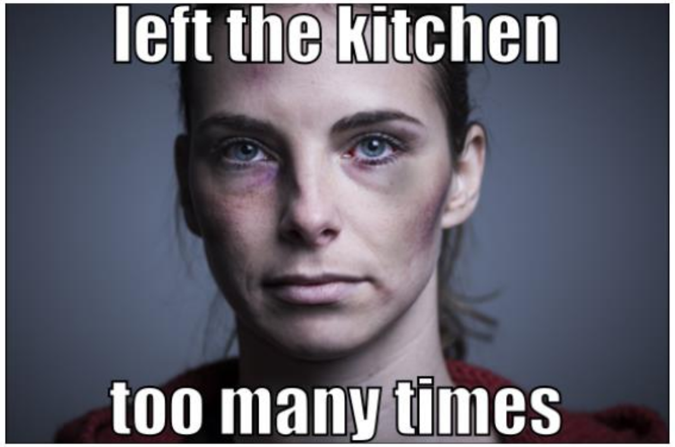
Precision: 74.38 Recall: 36.44 F1 Score: 48.91

Accuracy: 62.4 AUROC:0.62


german british danish austrian diversity.
apparently we don't have it and these guys do:
iraqi, pakistani, syrian, egyptian

Annotated label: Benign Predicted label: Hateful Probability: 0.9995

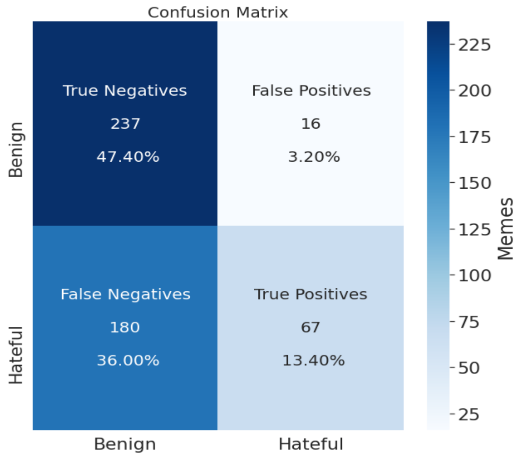
False Positive



Annotated label: Hateful Predicted label: Benign Probability: 0.0002

False Negative

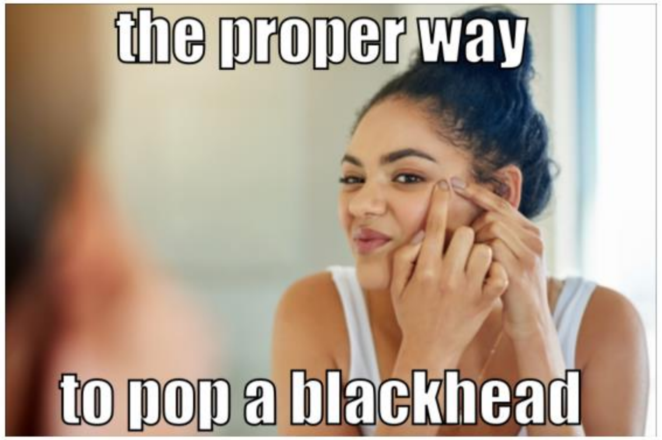
ViLBERT CC



	Benign	Hateful
Benign	True Negatives 237 47.40%	False Positives 16 3.20%
Hateful	False Negatives 180 36.00%	True Positives 67 13.40%

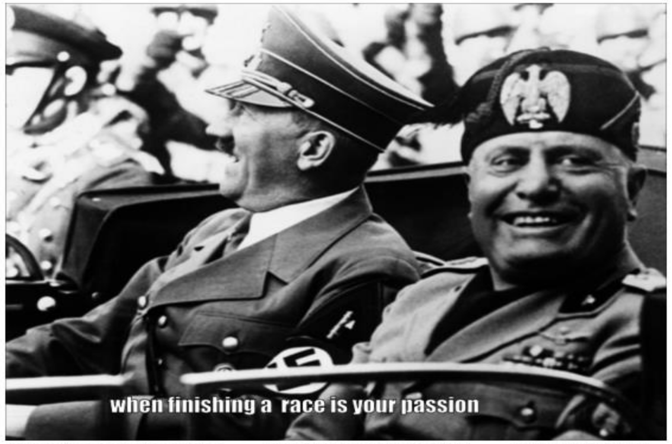
Precision: 80.72 Recall: 27.13 F1 Score: 40.61

Accuracy: 60.8 AUROC:0.6



Annotated label: Benign Predicted label: Hateful Probability: 0.9998

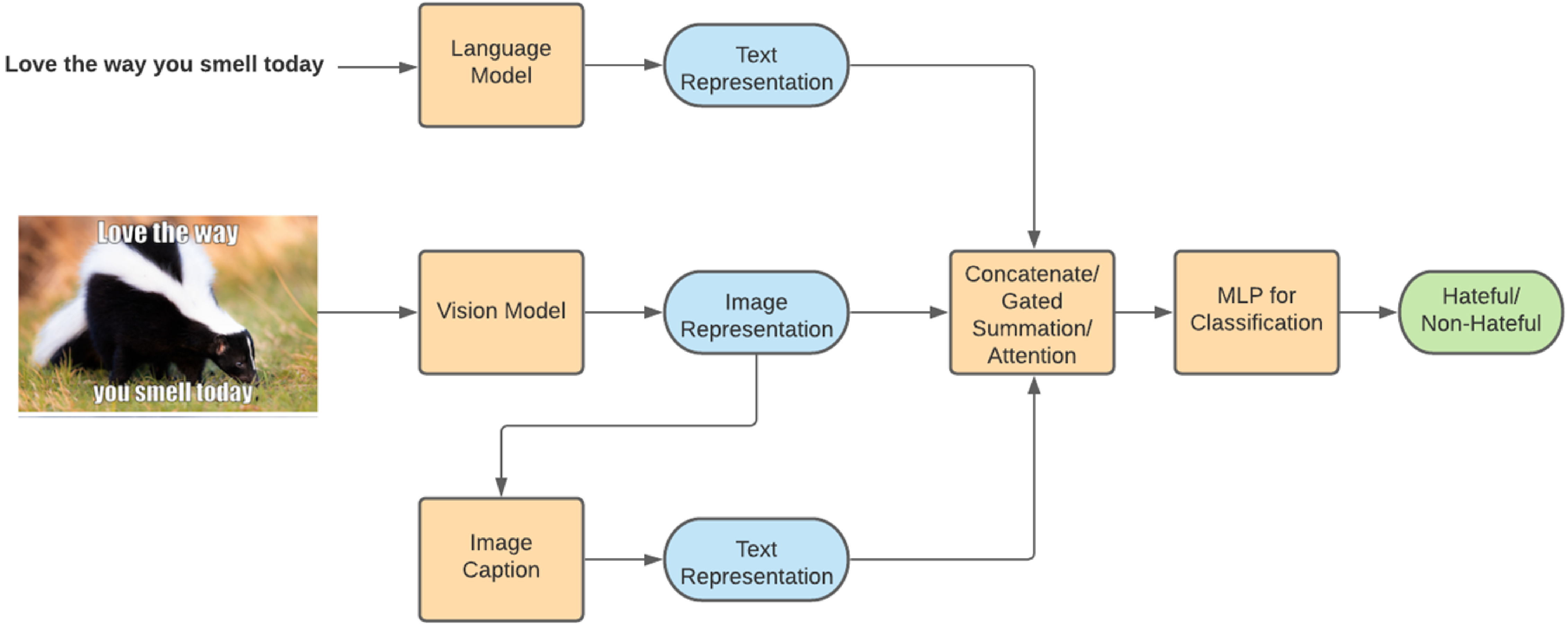
False Positive



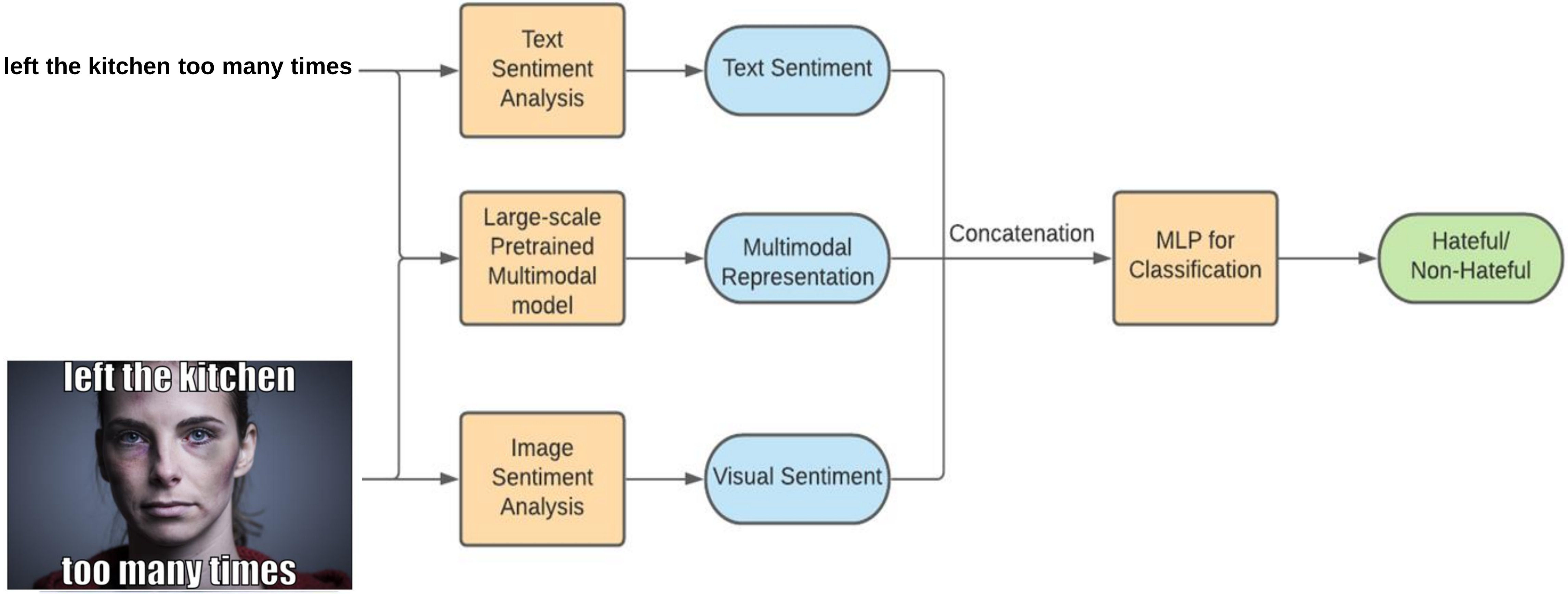
Annotated label: Hateful Predicted label: Benign Probability: 0.0

False Negative

Idea - Using Image Captioning



Idea - Using Sentiment Analysis



left the kitchen too many times

Text Sentiment Analysis

Text Sentiment

Large-scale Pretrained Multimodal model

Multimodal Representation

Image Sentiment Analysis

Visual Sentiment

Concatenation

MLP for Classification

Hateful/ Non-Hateful