An Interpretable Approach to Hateful Meme Detection

Tanvi Deshpande (Irvington High School) and Nitya Mani (MIT)

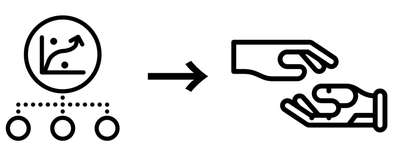
The Problem

- **Memes** = powerful tool for spreading **online hate** on fringe + mainstream sites
- **Poor performance** of humans (84.50 auROC)
- **Transformers** perform well; not a resolution
- **Multimodal** + rely on **cultural nuances** → traditional hate speech detection methods are **ineffective**

Confounders: different image/text → different meanings; reflected in data

Motivation

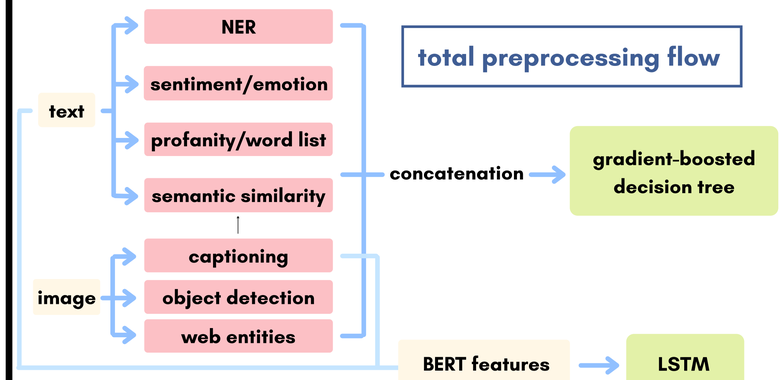
- Past work on hateful memes: fine-tuning large-scale transformers with **little to no** data preprocessing
  - **good** performance, **bad** interpretability, **don't help** human classifiers
- Our approach: **interpretable** models
  - based on human cognition → **reasoning** for final human classifiers
  - **simpler models** (less computationally expensive)



Methodology

- Experimented with textual/image features based on human insights
- Concatenation + embeddings, two classes of models (decision tree + LSTM)

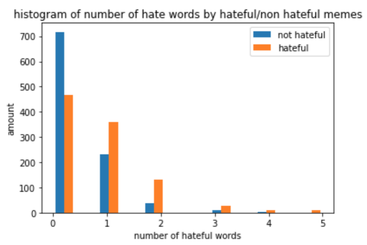
{Preprocessing}



"cadbury's halal creme eggs now available in special edition"
↓ ORG
flagged (Urban Dictionary list)

→ emotion: happiness
text hate: 1
sentiment: (0.379, 0.486)
entities: 1 (cadbury_ORG)

← **text preprocessing** flow example



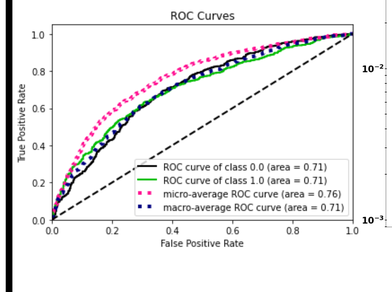
Feature selection: motivation for Urban Dictionary list of "dogwhistle" words

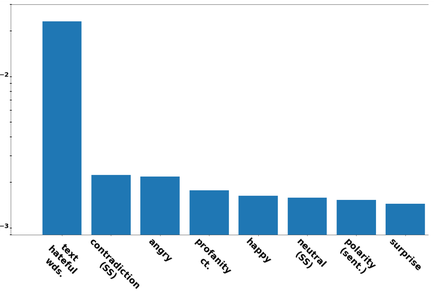
Caption: A woman in a white shirt and black tie.
Web Entities: Ilhan Omar Politics United States Member of Congress

{Models}

- Concatenation + embeddings
  - tf-idf → decision tree w/ XGBoost
  - DistilBERT → LSTM

Results





Source	Model	AUROC Val.	AUROC Test
	Human	-	82.65
Non-transformer Baselines	Text-only BERT	64.65	65.08
	Late Fusion	65.97	64.75
Transformer Baselines	ViLBERT CC	70.07	70.03
	VisualBERT COCO	73.97	71.41
Our Models	GBDT	71.67	70.90
	LSTM	73.78	72.72

- **Matches transformer baselines**
- **Feature importances** = insights on co-occurring features conveying hatefulness
- Differentiates b/w **hateful memes and confounders** + identifies nontrivial hateful memes (see paper)

Takeaways

- There may be promise in a joint human-AI approach
- AI **flags memes** and **provides reasoning**
- How can we practically facilitate **human-AI collaboration?**
- How can we minimize **human bias** in hateful meme detection?

tanvi.md@gmail.com
nmani@mit.edu

[1] Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2021. The Hateful Memes Challenge: Detecting Hate Speech in Multimodal Memes. arXiv:2005.04790 [cs.AI]
[2] Ron Zhu. 2020. Enhance Multimodal Transformer With External Label And In-Domain Pretrain: Hateful Meme Challenge Winning Solution. arXiv:2012.08290 [cs.CL]