Hate Speech in Pixels: Detection of Offensive Memes towards Automatic Moderation


Benet Oriol


Cristian Canton


Xavier Giro

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH

2019 AI for Social Good

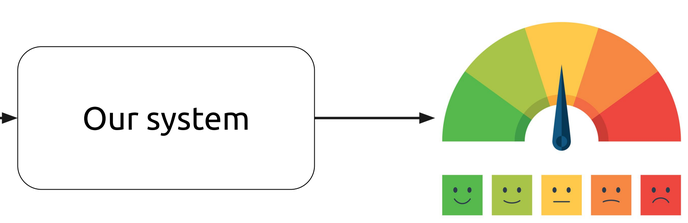
Deep Learning Barcelona Symposium
#DLBCN 2019 http://DeepLearning.Barcelona

Summary

This work addresses the challenge of **hate speech detection in Internet memes**, and attempts **using visual information to automatically detect hate speech**, unlike any previous work of our knowledge. Hate memes spread hate through social networks, so their automatic detection would help reduce their harmful societal impact. Our results indicate that the model can learn to detect some of the memes, but that the task is far from being solved with this simple architecture. While previous work focuses on linguistic hate speech, our experiments indicate how **the visual modality can be much more informative for hate speech detection** than the linguistic one in memes.

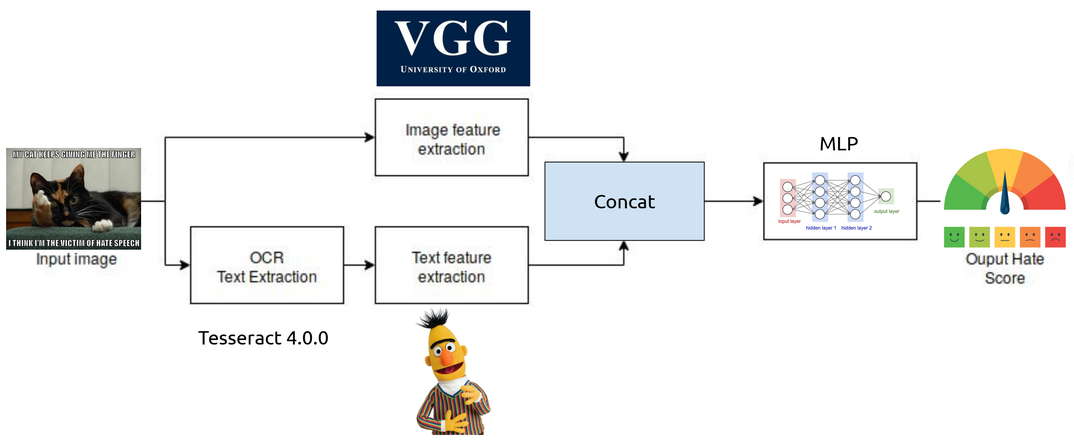
Motivation

Detection of hate speech memes to moderate their spread through social networks.



While hate speech detection has traditionally focused on language, we explore the impact of visual information for this task.

System overview



OCR Extraction	**Pytessteract** (Tesseract 4.0.0)
Text Feature Extraction	**BERT:** bert-base-multilingual-cased. This version has 12 layers, 768 hidden dimensions, 12 attention heads with a total of 110M parameters and is trained on 104 languages. Frozen weights. Feature size = 768.
Visual Feature Extraction	**VGG-16.** Pretrained on ImageNet for input images of 224x224 Frozen weights. Feature size = 4096.
Hate Speech Detector	**Multi Layer Perceptron.** 2 Hidden Layers of 100 dims. ReLU Activation. 1 output neuron to predict hate speech
Training	Adam optimizer and MSE loss function. Converges after 60 epochs

Dataset

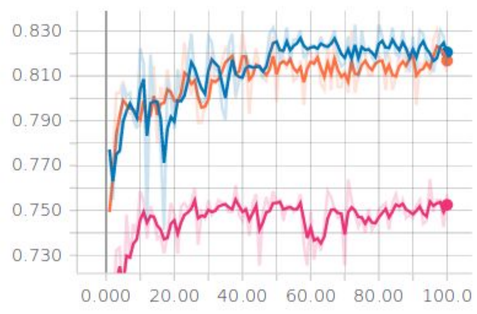
New dataset of 5020 memes built for this work, partition as 85% (4266) for training and 15% (754) for validation.

	😄	😡
Source	Reddit Memes dataset • Assumption that no hate speech memes are present in this public dataset.	Google images queried with: • “racist meme” • “jew meme” • “muslim meme”
Size	3325	1695

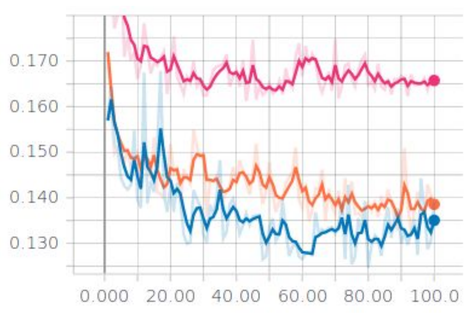
Challenges:
- Diversity of styles.
- High image compression rates make OCR performance drop

Model performance

Language vs Vision vs Multimodal



Accuracy



MSE Loss

Hate Pixels

Model	Max. Accuracy	Smth. Max. Accuracy
Multimodal	0.833	0.823
Image	0.830	0.804
Text	0.761	0.750

- All three configurations perform better than 0.66 random predictions in this biased dataset.
- Visual modality is more important than language.
- Results improve when combining both.

PyTorch

Acknowledgments:

AI for Social Good · NVIDIA · Gobierno de España, Ministerio de Economía y Competitividad · Generalitat de Catalunya