@DocXavi

Hate Speech in Pixels

Detection of Offensive Memes towards Automatic Moderation

AI for Social Good

Benet Oriol

Cristian Canton

Xavier Giro

UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH

telecom BCN

Deep Learning Barcelona Symposium

Photo of Torre Glòries by Barcelona Turisme

Facebook to set up Barcelona global hub to fight fake news

Social media company rents eight floors in Torre Glòries

07 May 2018 01:06 PM by ACN | Barcelona

Facebook has chosen Barcelona as its global hub to tackle fake news. The social media company will set up its new offices in the **Torre Glòries** (formerly known as Torre Agbar), one of the city's most emblematic skyscrapers. According to Spanish newspapers, some 500 people will be working to monitor content shared by more than 2 billion users worldwide.

Facebook reached an agreement with the European firm Competence Call Center over the renting of **9,000 square meters** spanning eight floors, as announced by the real estate company Engel & Völkers on April 26.

Torre Glòries, designed by French architect Jean Nouvel, was the crown jewel of Barcelona's candidacy to host the European Medicines Agency (EMA) in its relocation following Brexit. Yet the Catalan capital was eliminated in the first round, and Amsterdam became the chosen city.

> ***"It's clear now that we didn't do enough to prevent these tools from being used for harm"***
>
> **Mark Zuckerberg** · Facebook CEO

Barcelona's Torre Glòries (by ACN)

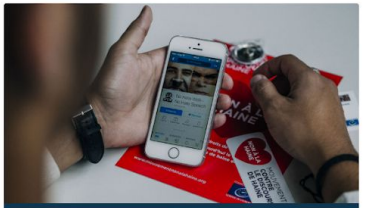
Facebook to set up Barcelona global hub to fight fake news

Reporting Hate Speech

Reporting prohibited speech and hate speech is essential and effective, especially when it constitutes a type of hate crime. Reporting also ensures the safety of all Internet users and the protection of human rights online.

Most European countries have established national reporting mechanisms and support for victims of cyber bullying, hate speech and hate crime, provided by national authorities and NGOs.

Social media platforms offer tips to help protect users from cyber bullying and hate speech, and provide tools for reporting them to the platform administrators or moderators.

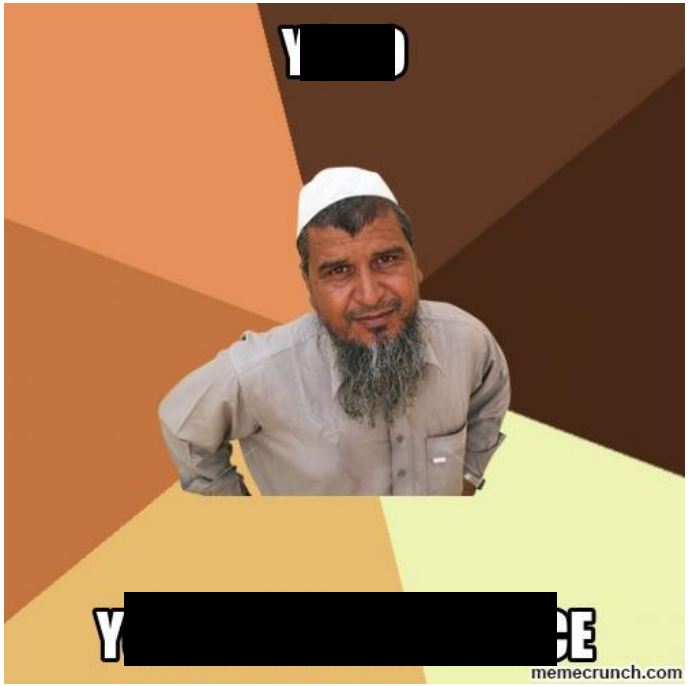


Reporting on Social Media Platforms

Reporting to National Bodies

Hate Memes (racist)

@DocXavi



Classic style meme

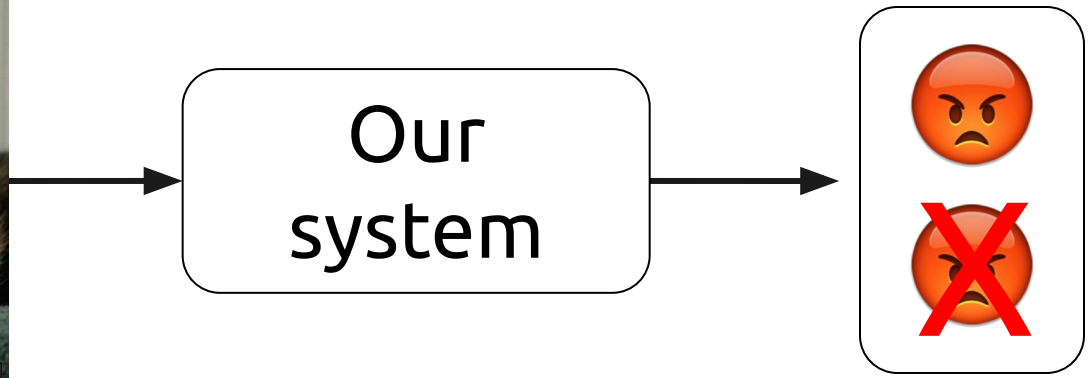
Modern style meme

UN Sustainable Development Goals

@DocXavi

Motivation

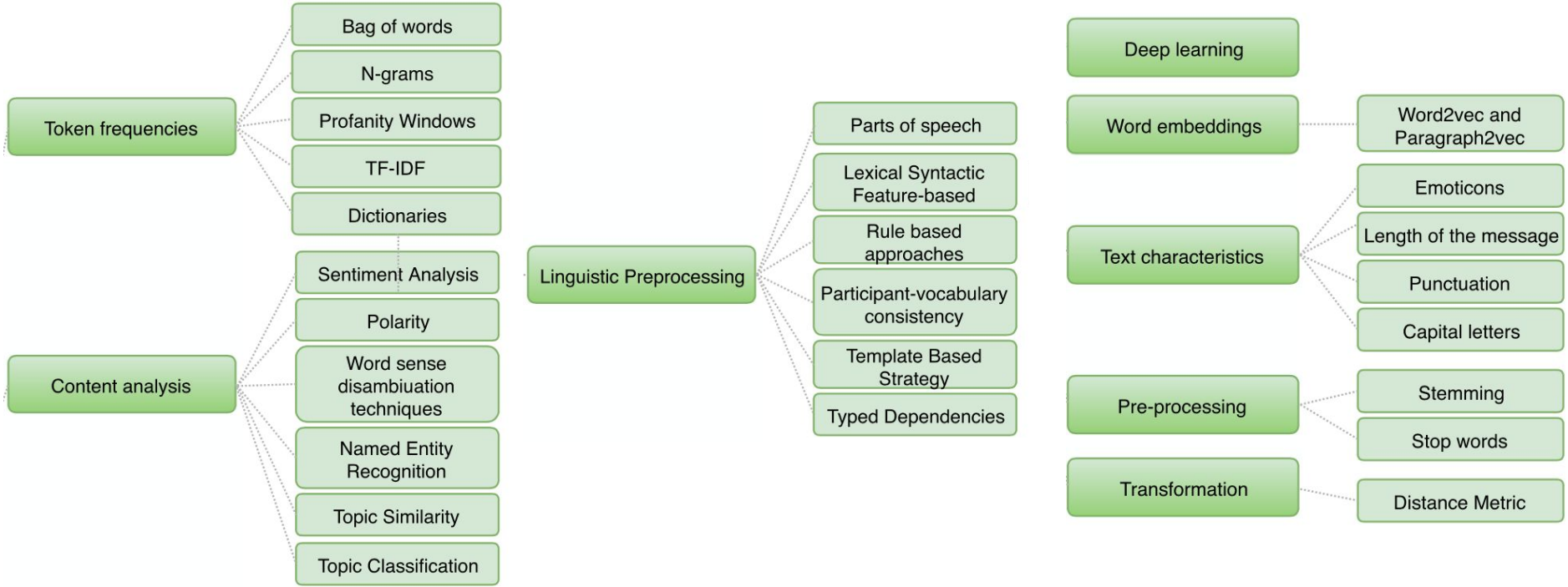

MY CAT KEEPS GIVING ME THE FINGER
I THINK I'M THE VICTIM OF HATE SPEECH


Our system

Hate Speech in Language

@DocXavi

Most (all ?) previous efforts focused on language only.

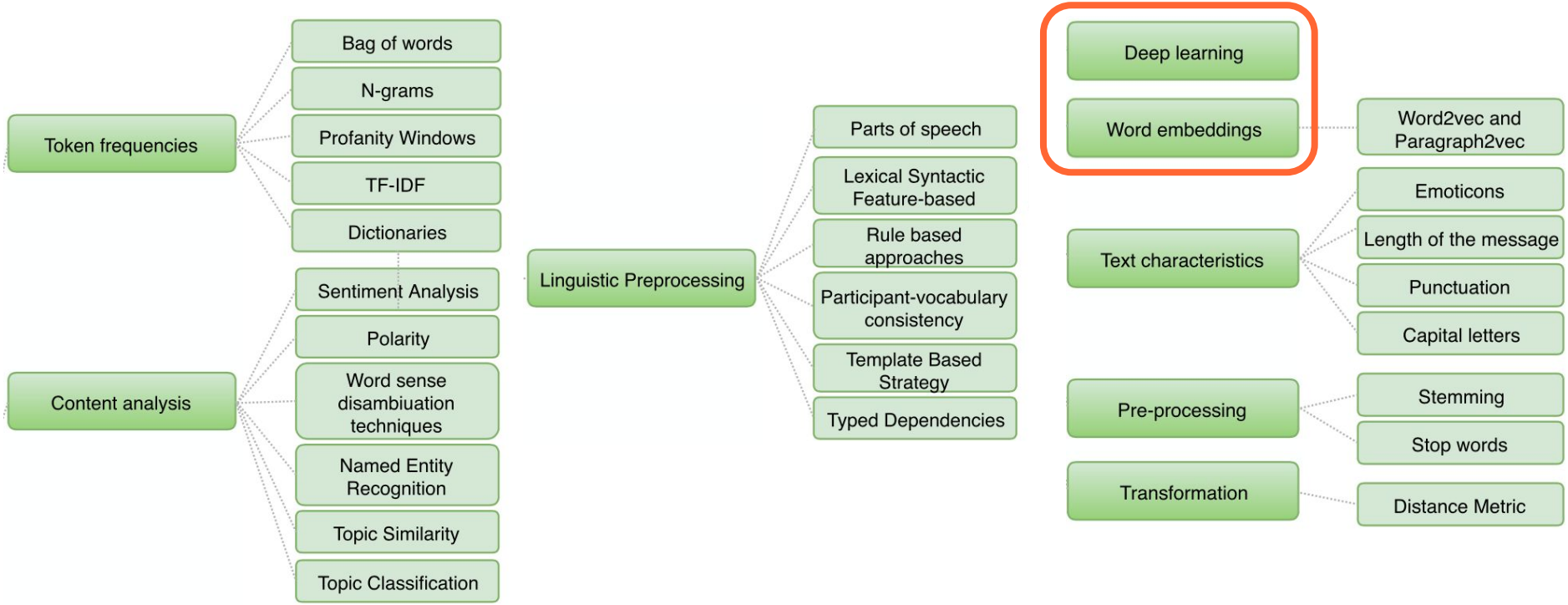


Paula Fortuna, Sergio Nunes, “A Survey on Automatic Detection of Hate Speech in Text”. ACM Comput. Surv 2018.

Hate Speech in Language

@DocXavi

Our system focuses on a deep learning approach.



Paula Fortuna, Sergio Nunes, “A Survey on Automatic Detection of Hate Speech in Text”. ACM Comput. Surv 2018.

Hate Speech in Language

Some language patterns clearly correlate with hate speech:

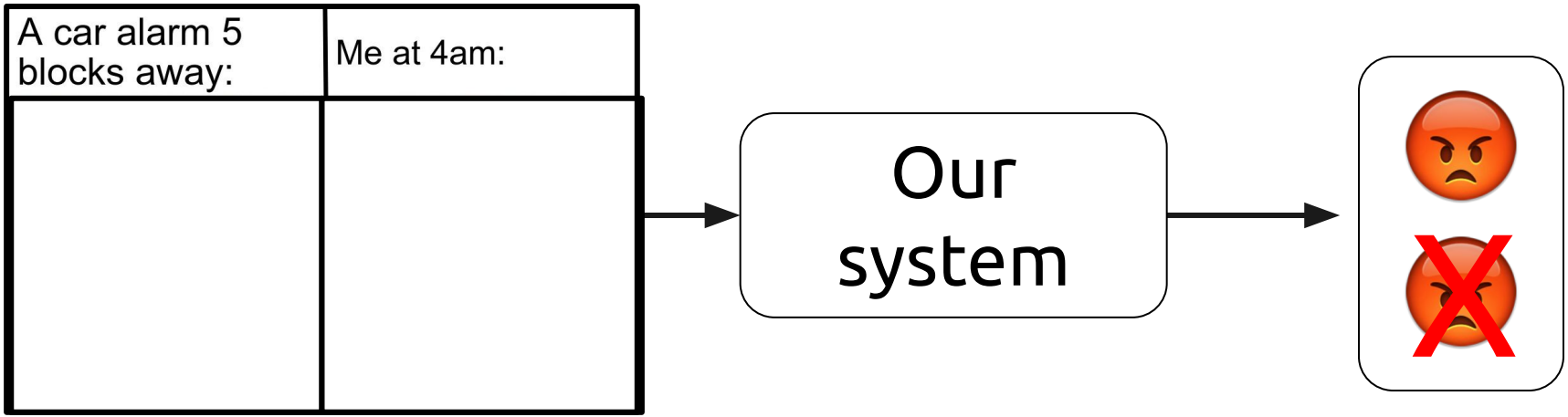
Twitter	% posts	Whisper	% posts
I hate	70.5	I hate	66.4
I can't stand	7.7	I don't like	9.1
I don't like	7.2	I can't stand	7.4
I really hate	4.9	I really hate	3.1
I fucking hate	1.8	I fucking hate	3.0
I'm sick of	0.8	I'm sick of	1.4
I cannot stand	0.7	I'm so sick of	1.0
I fuckin hate	0.6	I just hate	0.9
I just hate	0.6	I really don't like	0.8
I'm so sick of	0.6	I secretly hate	0.7

Table 1: Top ten hate expressions in Twitter and Whisper.

Silva, Leandro, Mainack Mondal, Denzil Correa, Fabrício Benevenuto, and Ingmar Weber. "Analyzing the targets of hate in online social media." In Tenth International AAAI Conference on Web and Social Media. 2016.

Hate Speech in Memes

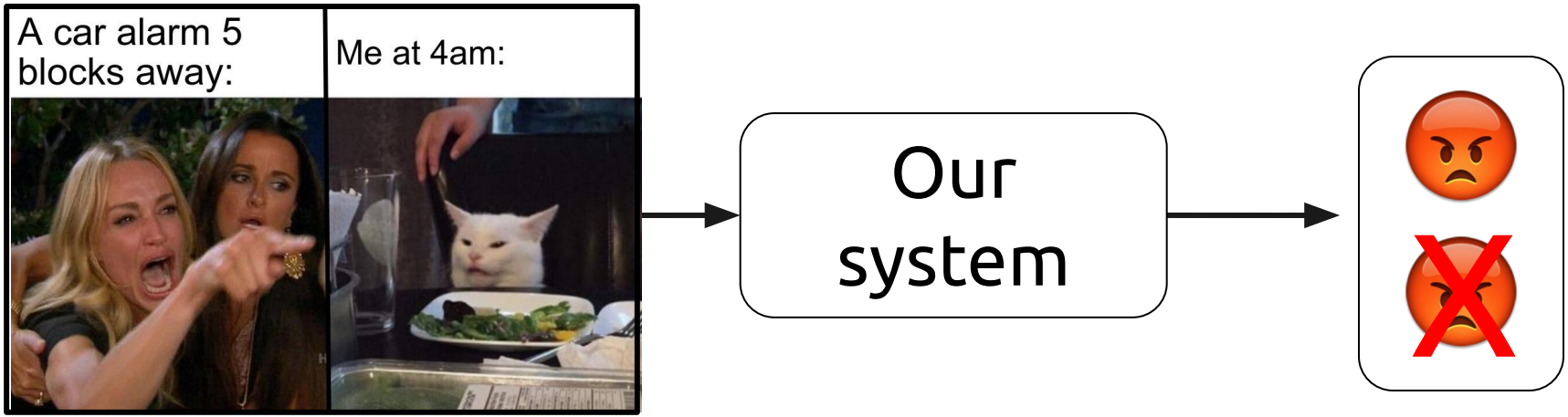
Language is not enough to understand memes.



Hate Speech in Memes

@DocXavi

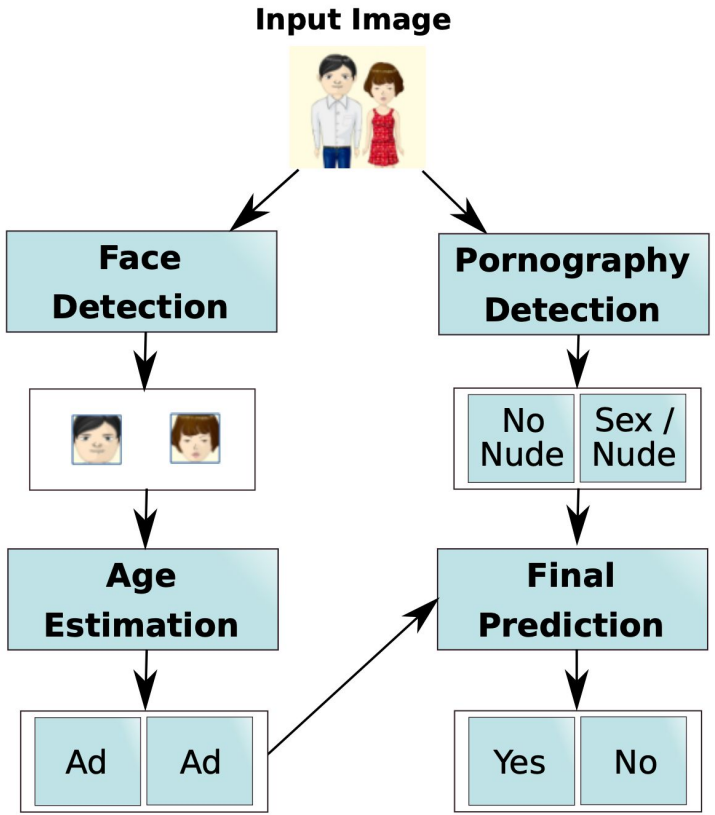
Language is not enough to understand memes.



Context: Pixel Moderation

Computer vision tools are already deployed for content moderation:

- Nudity
- Pornography
- Violence
- ...



João Macedo,, Filipe Costa, Jefersson A. dos Santos, “A Benchmark Methodology for Child Pornography Detection” SINGRAPI 2018.

Context: Study case

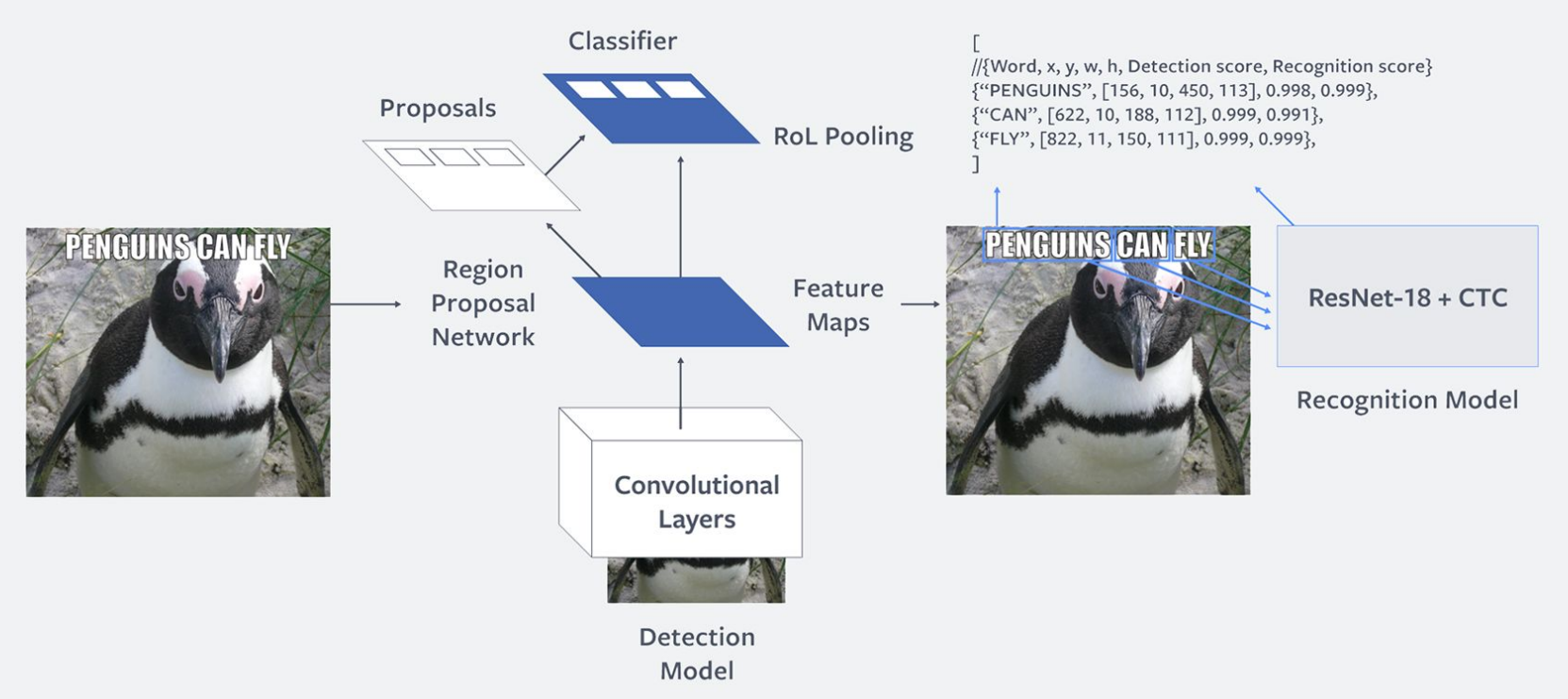
Our experiments focus on racist memes.

Twitter	
Categories	**% posts**
Race	48.73
Behavior	37.05
Physical	3.38
Sexual orientation	1.86
Class	1.08
Ethnicity	0.57
Gender	0.56
Disability	0.19
Religion	0.07
Other	6.50

Hate type	Frequency
General hate speech	26
Racism	18
Sexism	6
Religion	4
Anti-semitism	1
Nationality	1
Other	1
Physical/mental handicap	1
Politics	1
Sectarianism	1
Socio-economical status	1

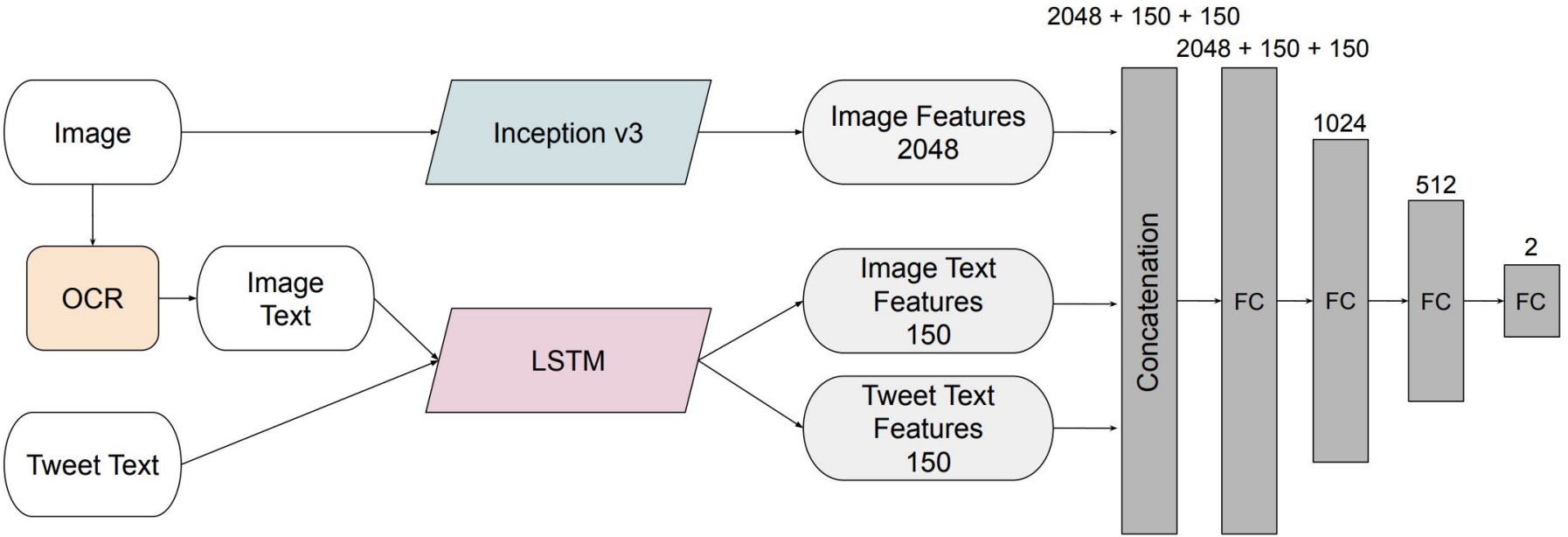
[1] Silva, Leandro, Mainack Mondal, Denzil Correa, Fabrício Benevenuto, and Ingmar Weber. "Analyzing the targets of hate in online social media." In Tenth International AAAI Conference on Web and Social Media. 2016.
[2] Paula Fortuna, Sergio Nunes, "A Survey on Automatic Detection of Hate Speech in Text". ACM Comput. Surv 2018.

Related Work



#Rosetta Fedor Borisyuk, Albert Gordo, Viswanath Sivakumar, "Rosetta: Large scale system for text detection and recognition in images". KDD 2018. [blog]

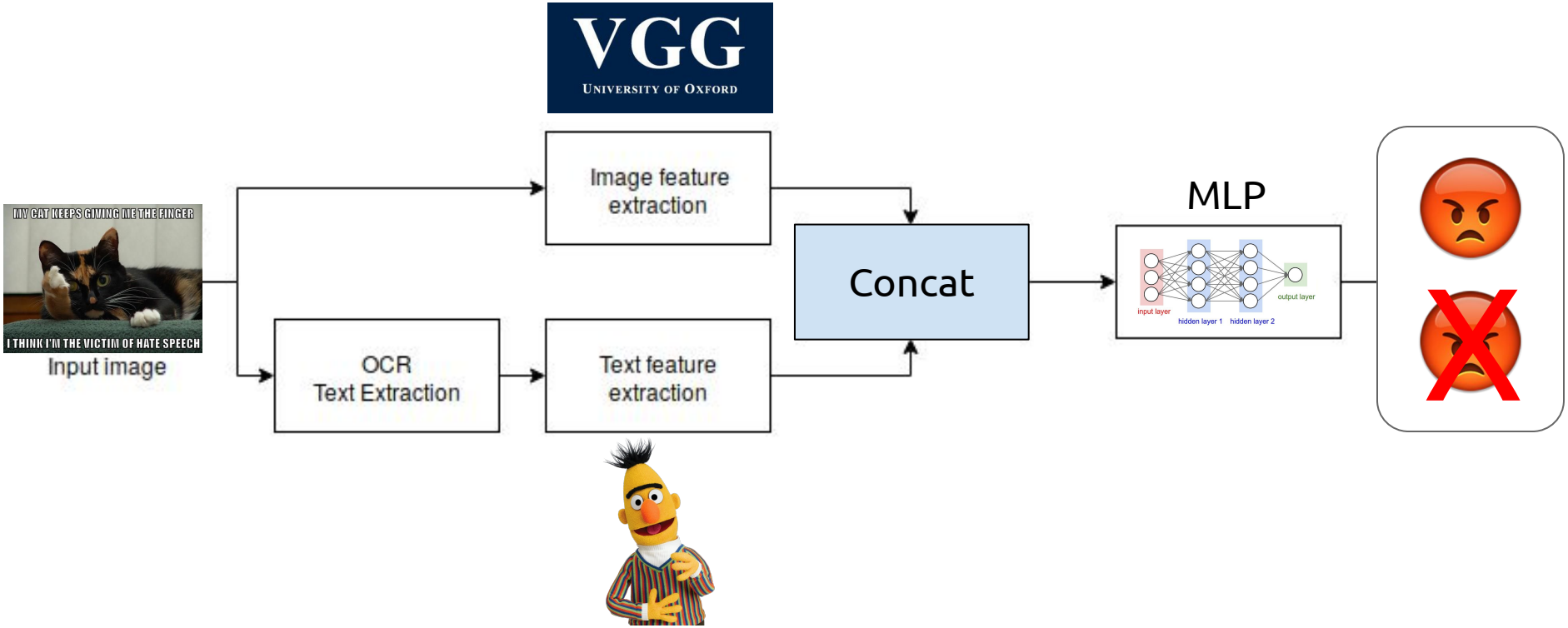
Related Work (concurrent)



#MMHS150K Raul Gomez, Jaume Gibert, Lluis Gomez, Dimosthenis Karatzas, "Exploring Hate Speech Detection in Multimodal Publications". WACV2020 (to be presented)

Deep Neural Architecture

@DocXavi



Deep Neural Architecture

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

Dataset Collection

		
Source	Queries to Google images: • “racist meme” • “jew meme” • “muslim meme”	Reddit Memes dataset [*] Assumption: No hate speech memes in it.

[*] https://www.kaggle.com/sayangoswami/reddit-memes-dataset

Dataset Partition

@DocXavi

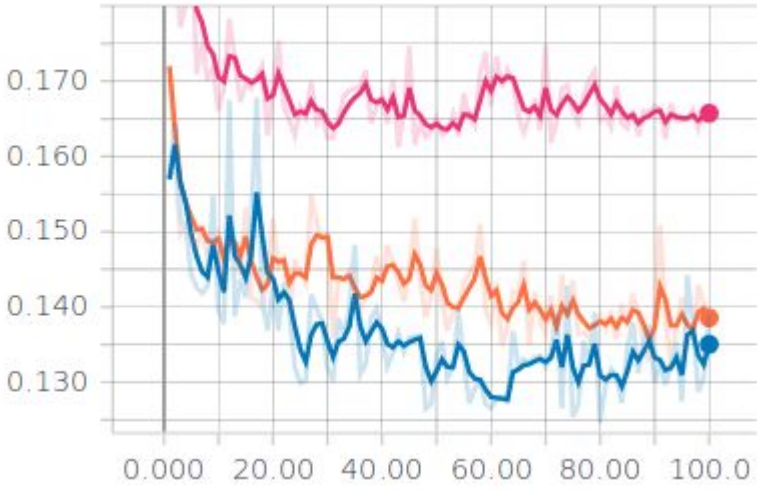
Unbalanced dataset towards

		
Train (85%)	1441	2825
Val (15%)	254	500
Total	1695	3325
	5020	

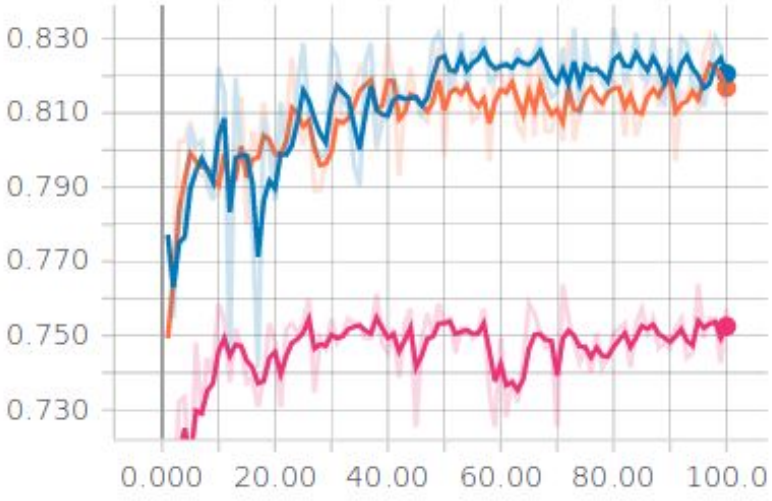
Model optimized with ADAM

@DocXavi

Language vs Vision vs Multimodal



MSE Loss



Accuracy

(All No Hate = 0.66)

Results

Accuracy on the *validation* dataset, not enough data to build a third *test* partition.

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

Results

All three configurations perform better than the 0.66 accuracy of a naive “All no hate” predictions.

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

Results

- Visual modality is more important than language.

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

Results

- Visual modality is more important than language.

Hypothesis #1: Dataset bias in terms of visual style: hate speech memes follow a more classical style.

Results

- Visual modality is more important than language.

Hypothesis #2: Memes often present visual effects and sitortions that damage the OCR performance.

Results

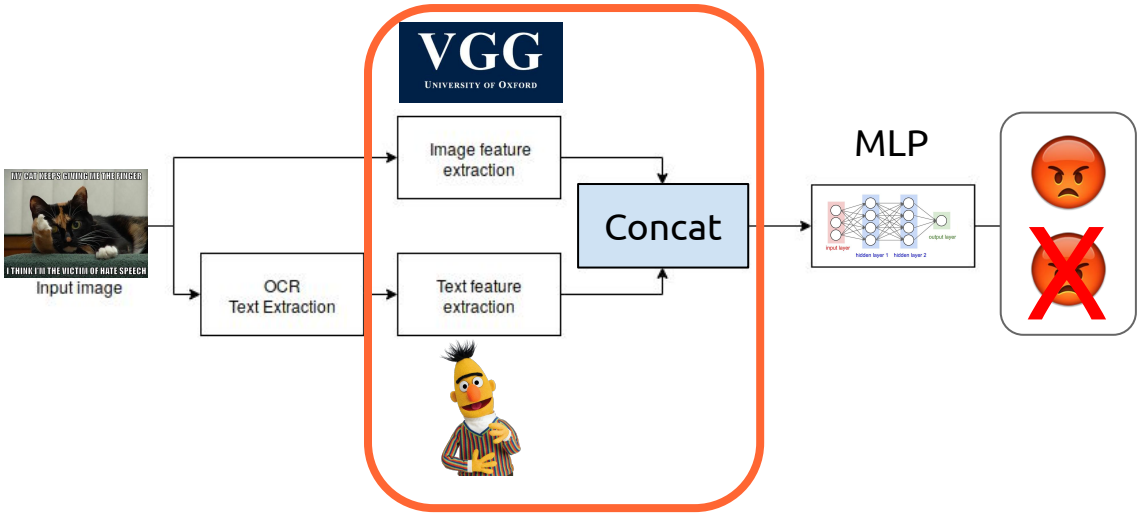
- Visual modality is more important than language.
- Small improvement when combining vision and language.

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

Results

- Visual modality is more important than language.
- Small improvement when combining vision and language.

Hypothesis #3: Because of the larger dimensionality of the representations: 4096 (vision) vs 784 (language).

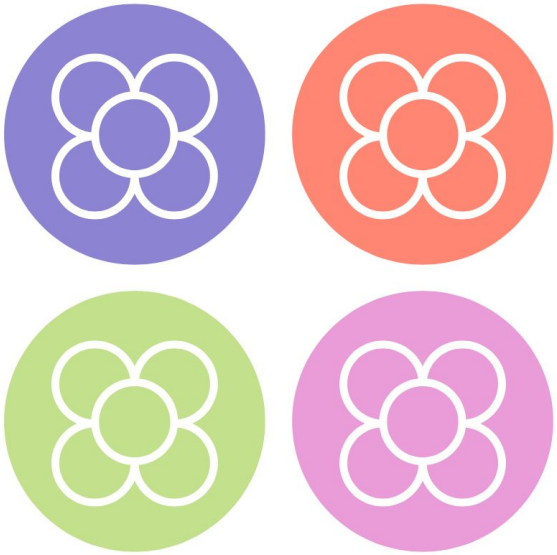


Conclusions

- Hate speech detection with naive ImageNet-Bert-like feature fusion:
  - Can capture some valuable patterns…
    - 83% accuracy improves over the 66% baseline.
  - ...but does not fully solve the challenge.
    - Gain of 17%, but still another 17% before 100%.

Conclusions

- Challenging & attractive scientific task with high societal impact.
- Next steps:
  - Larger & cleaner dataset (MMHS150K ?).
  - Inspiration from other multimodal understanding tasks (eg. visual Q&A, visual dialog…)
  - Knowledge bases (history, breaking news, demographics…)



Deep Learning
Barcelona
Symposium

Cristian
Canton

Benet
Oriol

Raul
Gomez

LLuis
Gómez

http://deeplearning.barcelona

Deep Learning Barcelona Symposium

Oriol Vinyals

Àgata Lapedriza

Marta R. Costa-jussà

Aleix Martínez

http://deeplearning.barcelona

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

Summary

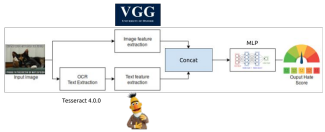
This work addresses the challenge of **hate speech detection in Internet memes**, and attempts **using visual information to automatically detect hate speech**, unlike any previous work of our knowledge. Hate memes spread hate through social networks, so their automatic detection would help reduce their 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

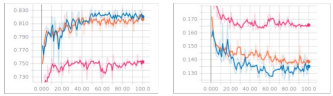
New dataset of 5020 memes built for this work, partition as 85% (4266) for training and 15% (754) for validation.

	😄	😡
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, Generalitat de Catalunya

@DocXavi

bit.ly/hatepixels

Travel grant by:

AI for Social Good

XPRIZE

intel AI

IVADO

Mila

@DocXavi

bit.ly/hatepixels