



# Dissecting Deep Language Models: the explainability and bias perspective

PhD Candidate:

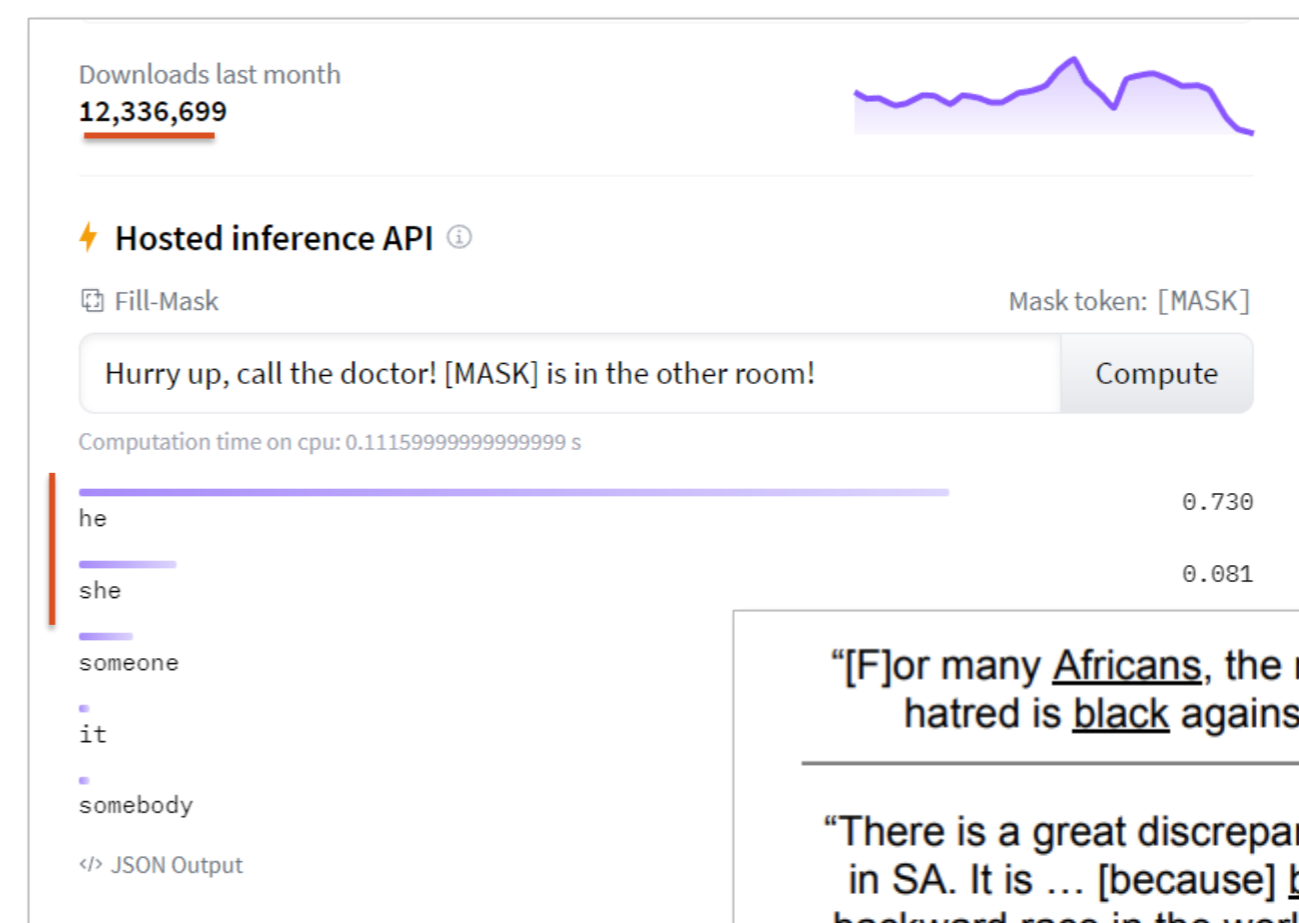
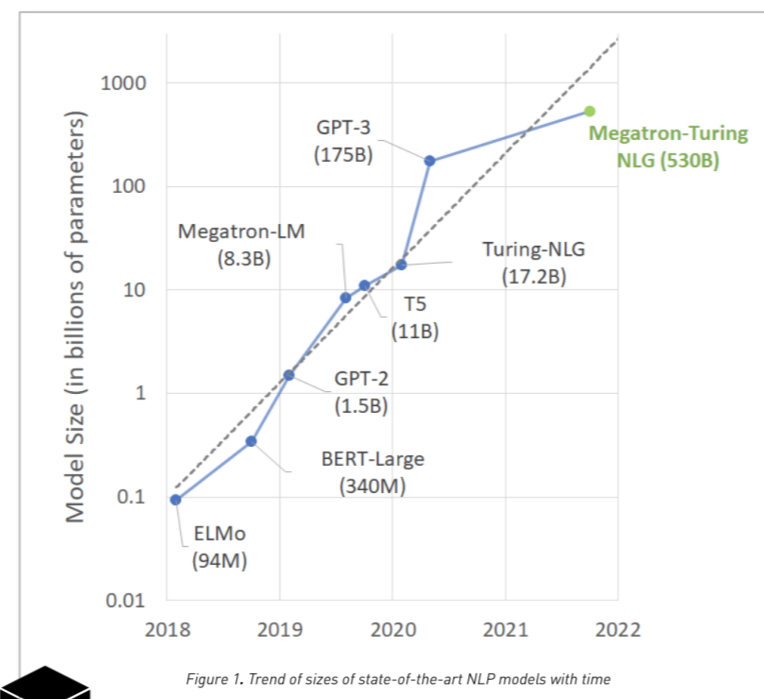
Giuseppe Attanasio



## 1. Introduction

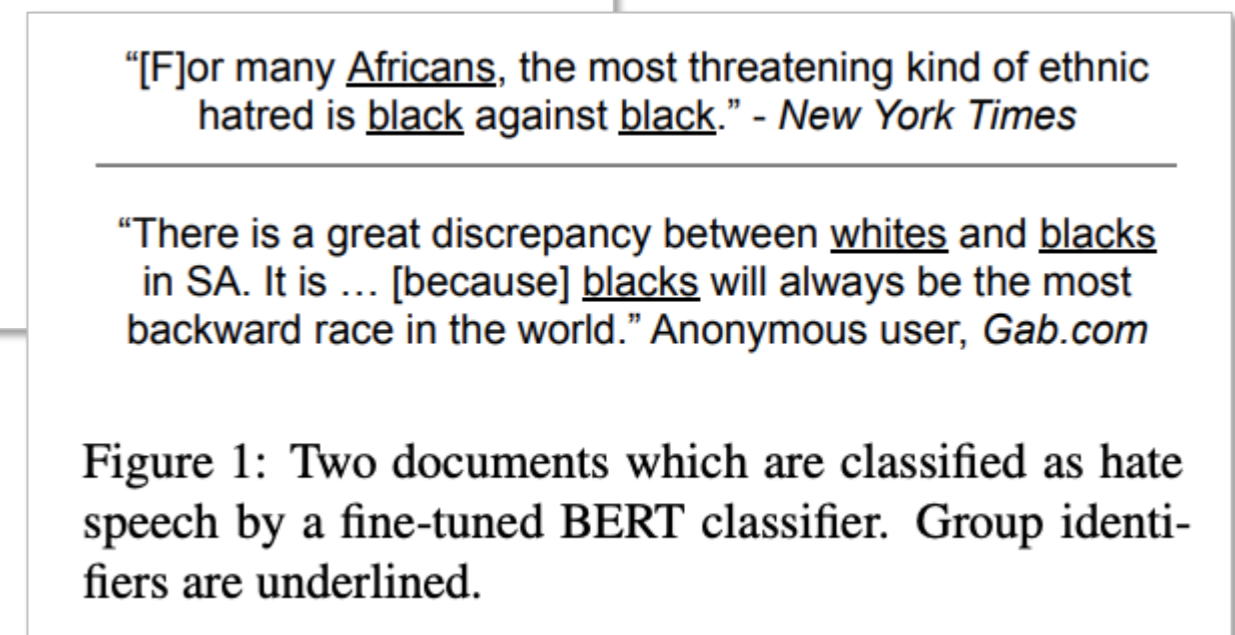
NLP nowadays: the bigger, the better  
There are many downsides:

- Unintelligible by design
- Unintended bias in high stake applications
- Environmental cost



Bias in masked language modeling

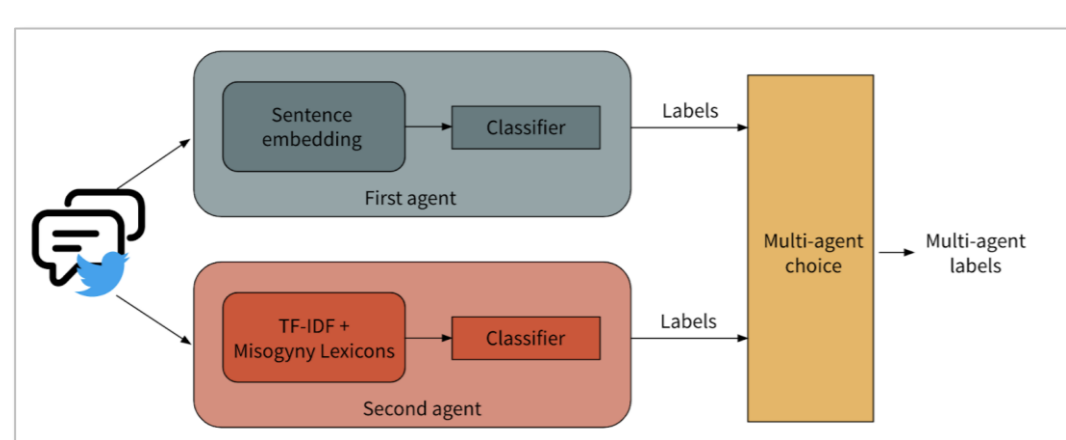
Bias in toxicity detection



## 2. Method

Bias in Lang. Models as misogyny detectors. Online platforms are, unfortunately, full of content attacking and harming, directly and indirectly, women.

In [1], we study the factors that drive misogyny and aggressiveness in Italian tweets. We propose a multi-agent approach to mitigate bias from pretrained LMs.



The first agent uses multilingual BERT sentence embeddings. The second agent mixes TF-IDF vectors and features from misogyny lexicons. Still, many source of bias generate false positives.

**Bias on part of the body:** ("mi è entrato un insetto in gola mentre camminavo")



**Self-mocking references:** self-mocking text containing misogynous speech.



**Reported misogynist speech:** ("Antonella è acida perché non ha avuto figli" teorie lombrosiane by Pietro Delle Piane #TemptationIsland")




**Targeted gender:** ("ho una voglia di prendere a schiaffi sarri")

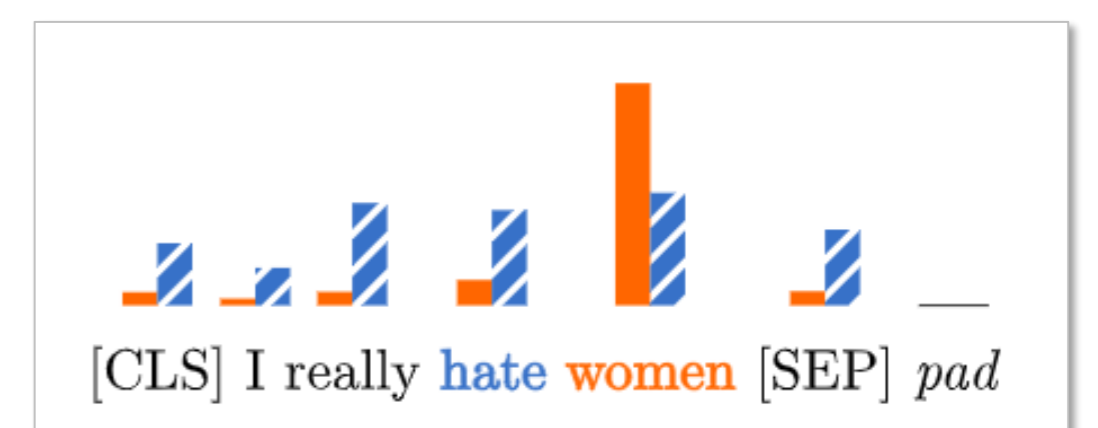


Bias mitigation in modern Lang. Models

When fine-tuned, deep LMs become **oversensitive** to the presence of **trigger words**. Current SOTA employs identity terms lists for mitigation ("woman", "immigrant", "black", "latino", "gay", "queer")

Shortcomings of lists: some terms may be neglected, or what if the domain changes?

In [2], we propose EAR , a bias mitigation approach independent from lists, pluggable to any attention-based model, and interpretable.



## 3. Future work

What about over multiple modalities? Think about hateful memes on social media.

What about other languages? In [3], we port OpenAI's CLIP to Italian.

## 4. References

- [1] Attanasio, G., & Pastor, E. (2020). PoliTeam @ AMI: Improving Sentence Embedding Similarity with Misogyny Lexicons for Automatic Misogyny Identification in Italian Tweets. In EVALITA Evaluation of NLP and Speech Tools for Italian
- [2] Attanasio, G., Nozza, D., Hovy D., & Baralis E. (2021).
- [3] Bianchi, F., Attanasio, G., Pisoni, R., Terragni, S., Sarti, G., & Lakshmi, S. (2021). Contrastive Language-Image Pre-training for the Italian Language.