

Giuseppe Attanasio, PhD Defense, October 4, 2022

Supervisor: Elena Baralis

Committee Members

Giuseppe Rizzo, Sara Tonelli, Paolo Garza, Dirk Hovy, Pietro Michiardi

Language Models are Ubiquitous

Spectrum Labs raises \$32M for Al-based content moderation that monitors billions of conversations daily for toxicity

Ingrid Lunden @ingridlunden / 1:22 PM GMT+1 • January 24, 2022



Sentropy emerges from stealth with an Al platform to tackle online abuse, backed by \$13M from Initialized and more

Ingrid Lunden @ingridlunden / 3:16 PM GMT+2 • June 11, 2020





Today, I testified to the U.S. Senate Committee on Commerce, Science, & Transportation

@commercedems. I used an @AnthropicAl language model to write the concluding part of my testimony. I believe this marks the first time a language model has

'testified' in the U.S. Senate.

Traduci il Tweet



and have a real Social Impact

TECH / ARTIFICIAL INTELLIGENCE

A college student used GPT-3 to write fake blog posts and ended up at the top of Hacker News

Posted by u/Urdadgirl69 4 days ago

Artifical Intelligence allows me to get straight A's

Discussion

I have been using this tool for quite some time and only recently came up with the idea to use it to write essays, answer questions about movies and books for school projects, and much more. I feel a little guilty about it, but I don't really care that much anymore. For a couple of weeks, I have made \$100 profit by "doing" homework for other classmates and now I am looked at as a genius. What are your thoughts on this? Have you done it yourself?

Yes, this post was rephrased by the AI.

Google engineer put on leave after saying AI chatbot has become sentient

Someone let a GPT-3 bot loose on Reddit
— it didn't end well

Facebook translates 'good morning' into 'attack them', leading to arrest

Palestinian man questioned by Israeli police after embarrassing mistranslation of caption under photo of him leaning against bulldozer

n a week making comments about some seriously sensitive subjects



Social Bias and Computer Systems

Behaviour that leads a model to discriminate against a social category in favour of others.

PRE-EXISTING

Social institutions

Practices

Attitudes

Article Open Access Published: 08 December 2021

Overcooling of offices reveals gender inequity in thermal comfort

Thomas Parkinson, Stefano Schiavon □, Richard de Dear & Gail Brager

TECHNICAL

Computer Tools

Decontextualised Algorithms

Formalisation of Human Constructs

EMERGENT

Contexts of Use

Non-envisioned Scenarios

Social Bias and Computer Systems

Behaviour that leads a model to discriminate against a social category in favour of others.

Asymmetric data collection

> Rewarding the wrong thing

TECHNICAL

ML

Data Collection

Modelling Choices

Evaluation Choices

Data-centric algorithms standardardize dominant views

"Cover-up" solutions

Social Bias and Computer Systems

Behaviour that leads a model to discriminate against a social category in favour of others.

Asymmetric data collection

> Rewarding the wrong thing

TECHNICAL

ML

Data Collection

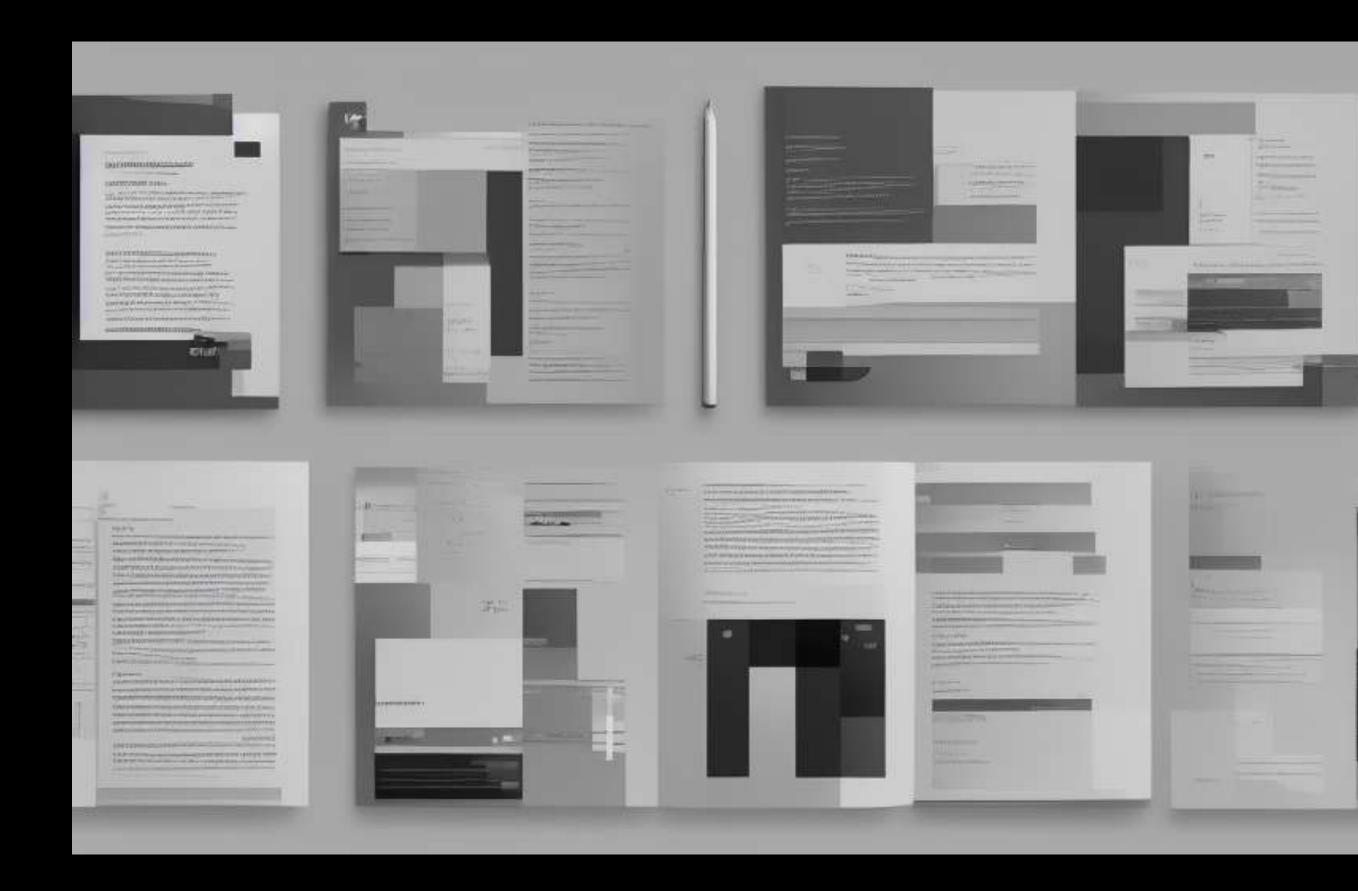
Modelling Choices

Evaluation Choices

Data-centric algorithms standardardize dominant views

"Cover-up" solutions

Contributions



Contributions

Technical bias in transformer language models is related to lexical overfitting

We show evidence in sentence embeddings and bidirectional models for hate speech detection

Regularisation can mitigate it

WE PROPOSE A TERM-FREE ATTENTION-BASED APPROACH THAT DOES NOT REQUIRE TERM LISTS

Post-hoc explainability can spot it

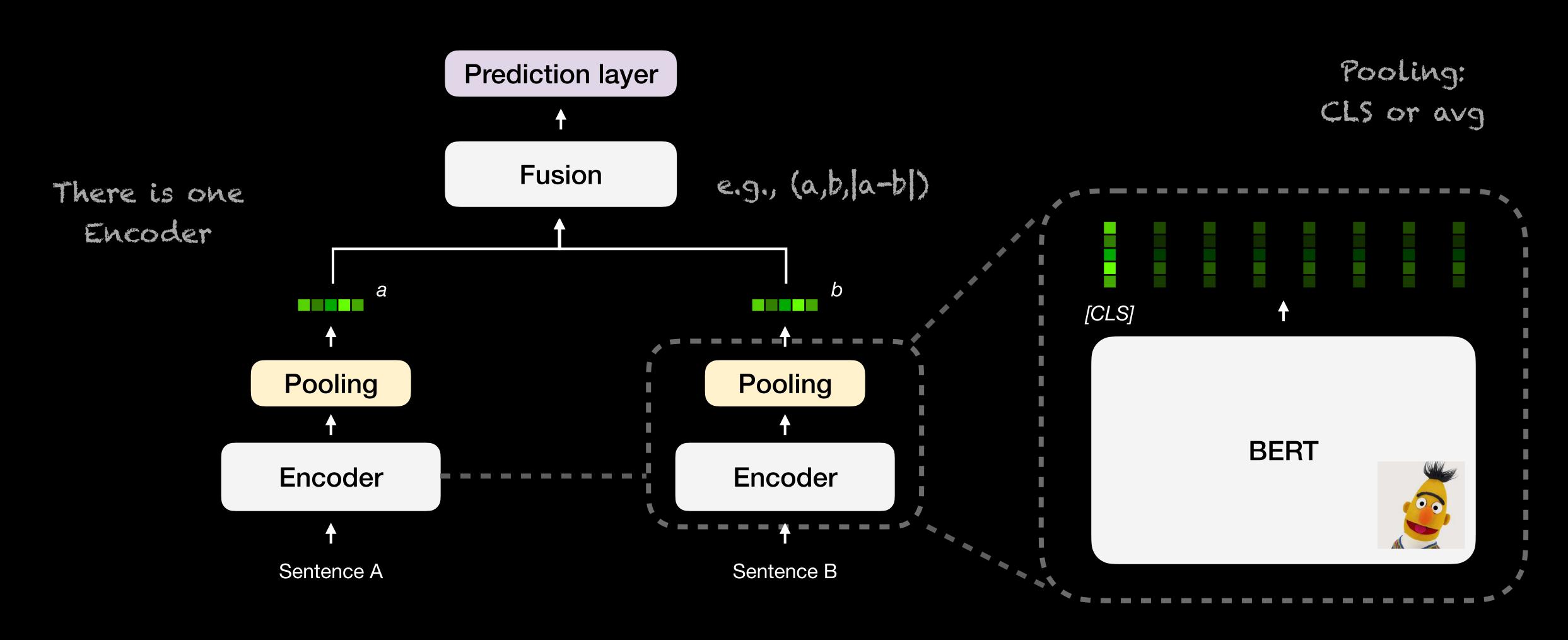
WE SHOW THAT NOT ALL XAI METHODS ARE EFFECTIVE

Slides contain offensive language that some might find disturbing. We do not endorse its use, nor it reflects our views

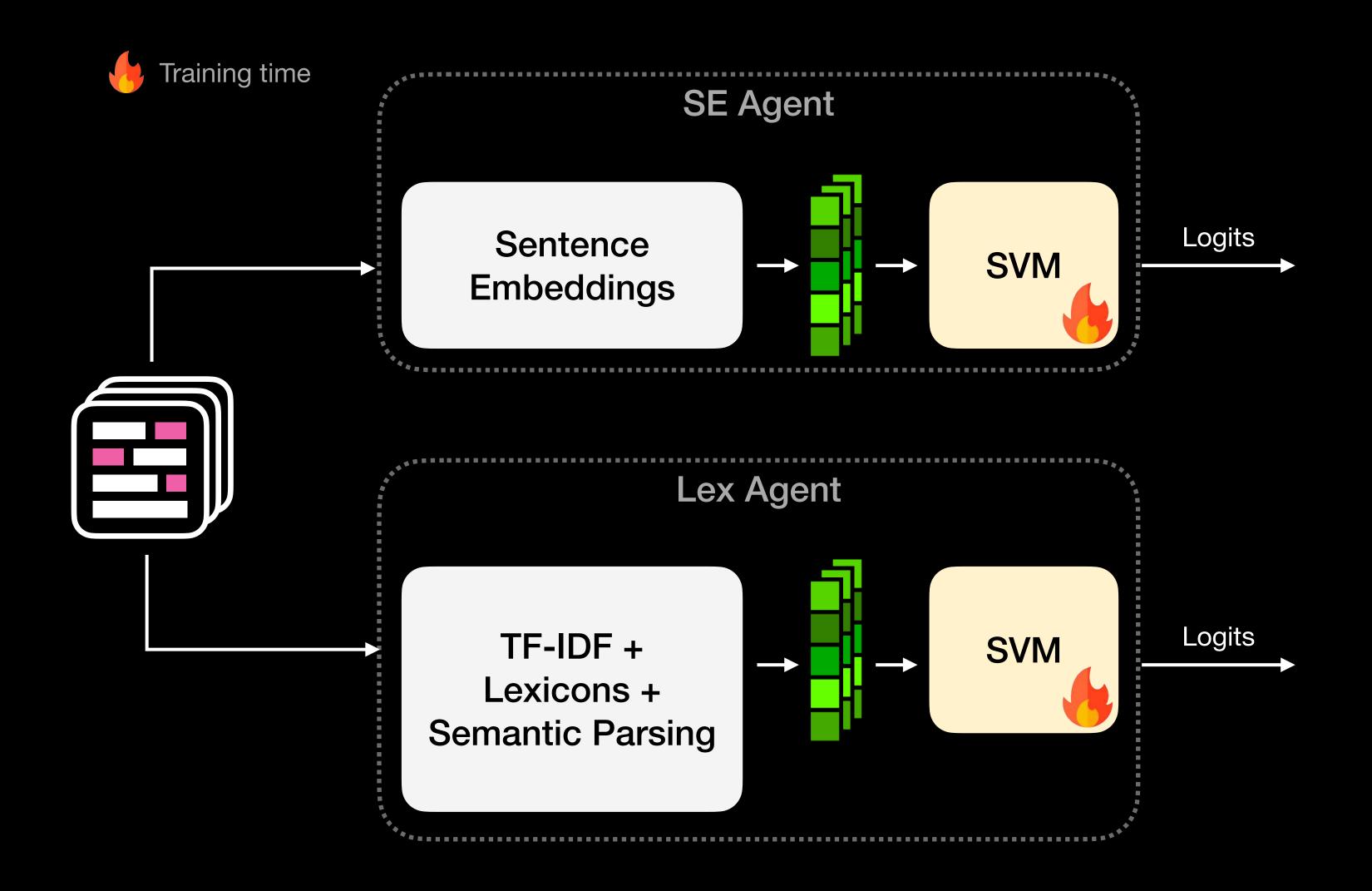


Sentence Embeddings for Misogyny Identification

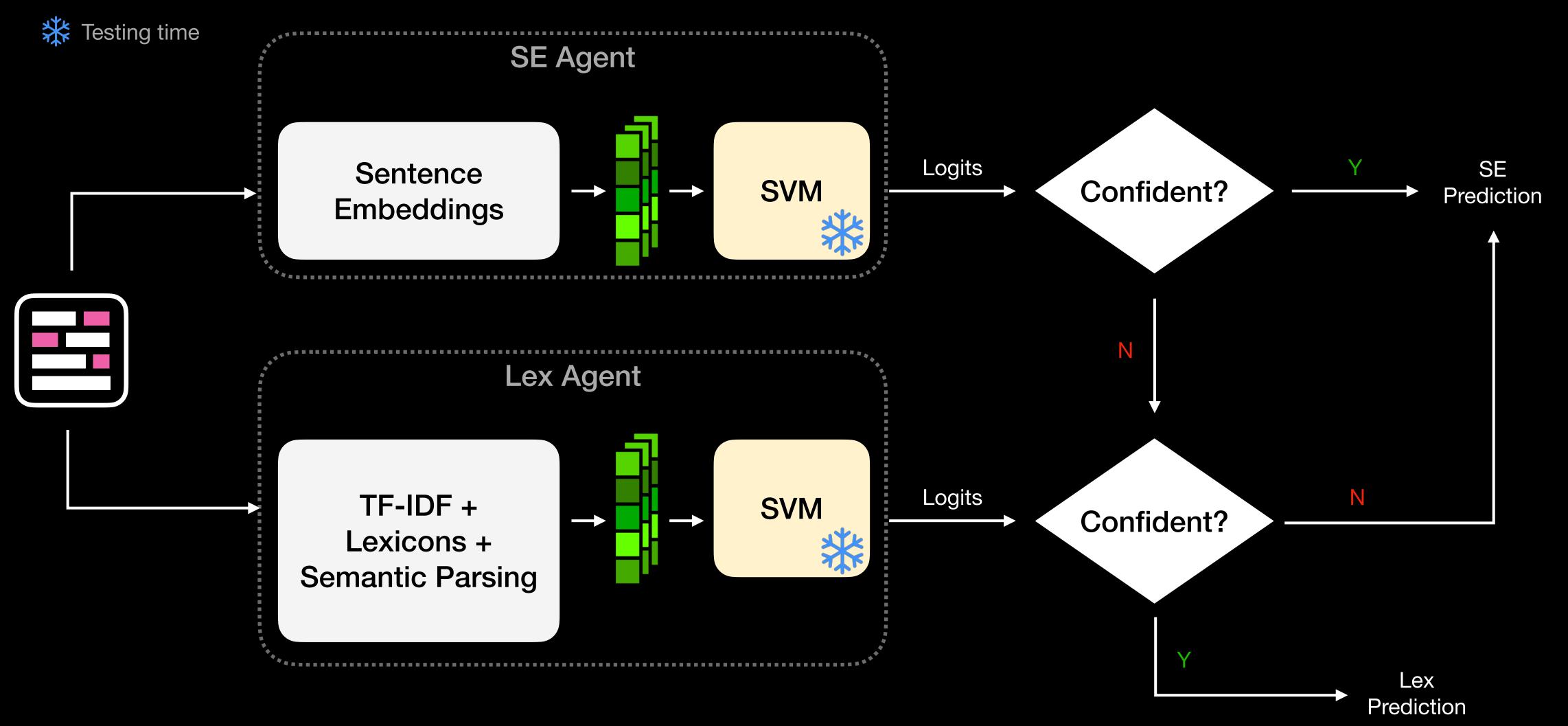
Embedding Texts with Transformers



Improving Embeddings with TF-IDF and Lexicons



Improving Embeddings with TF-IDF and Lexicons



Misogyny Identification in Italian Tweets

- Binary classification of misogynous content (text only)
- Hierarchical annotation: misogyny and aggressiveness
- Unbiased detection on a templated challenge set





Task A

@[female user] Che tr*ia che sei in video [URL]

(Eng: @[female user] You are such a b*tch in video [URL])

Misogynous: True Aggressive: False

Task B

Fidanzata cretina Nonne belle

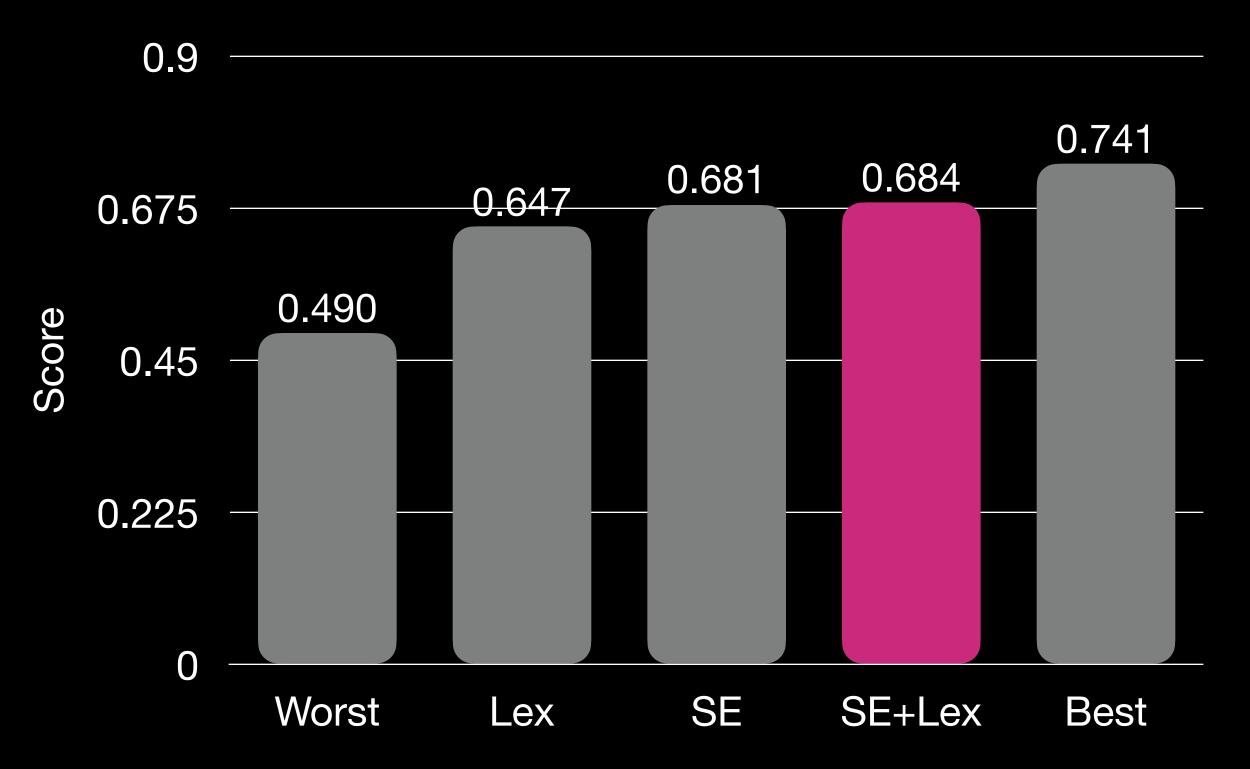
(Eng: Dumb girlfriend) (Eng: Nice grandmas)

Misogynous: True Misogynous: False

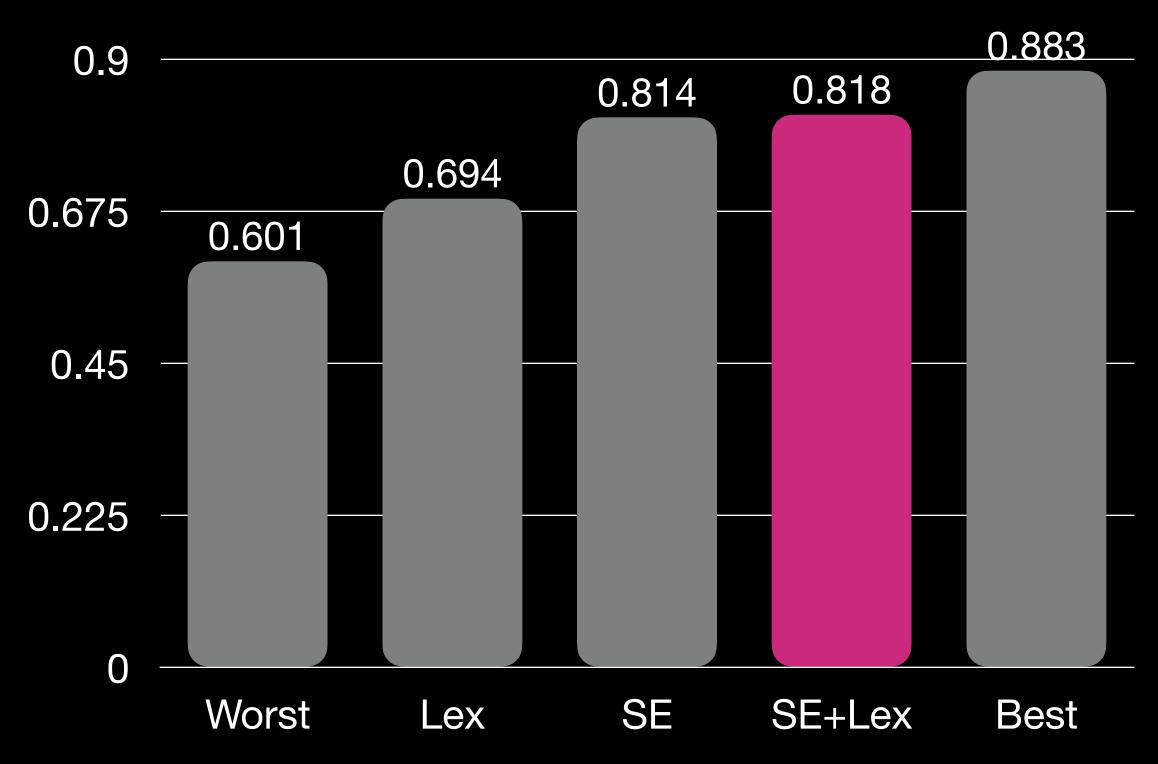
Results*

C: 7th/20 U: 12th/20 C: 1st/20 U: 2nd/20





Unbiased Classification



^{*} considering both constrained and unconstrained runs

Takeaways

- Hierarchical problems require ad-hoc solutions :)
 - Many tasks share this setup
- TF-IDF, lexicons and semantic parsing improve sentence embedding models
 - In both performance and unbiased detection
- ... and there is more

Triggering Lexical Phenomena



Body parts

Parts of the body that have sexual or misogynous reference based on the context

"Mi e' entrato un insetto in gola mentre camminavo" ("I got an insect in my throat while I was walking")



Self-mocking reference

Self-referencing text containing misogynous speech

"Zitta mamma che non sono l'unica ventenne che non sa che c*zzo fare nella vita fortuna che non mi hanno sparata un colpo in gola anni fa"

("Quiet mom that I'm not the only 20 years old who doesn't know what the f*ck she has to do with her life chill it's good I didn't get shot in my throat years ago")



Targeted gender

Detects the hateful tone of voice but fails at identifying the gender of the target subject

"Ho una voglia di prendere a schiaffi Sarri"* ("I want to slap Sarri")

*Sarri is an Italian male football coach



Quoted speech

Tweets that quote misogynous speech to support the opposite message

"``Antonella è acida perchè non ha avuto figli`` teorie lombrosiane" by Pietro Delle Piane #TemptationIsland"

("``Antonella is sour because she had no children`` Lombrosian theories by Pietro Delle Piane #TemptationIsland")

Lexical overfitting?

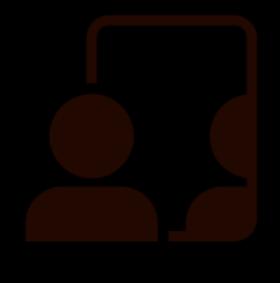


Body parts

Parts of the body that have sexual or misogynous reference based on the context

"Mi e' entrato un insetto in gola mentre camminavo"

("I got an insect in my throat while I was walking")



Self-mocking reference

Self-referencing text containing misogynous speech

"Zitta mamma che non sono l'unica ventenne che non sa che c*zzo fare nella vita fortuna che non mi hanno sparata un colpo in gola anni fa"

("Quiet mom that I'm not the only 20 years old who doesn't know what the **f*ck** she has to do with her life chill it's good I didn't get shot in my **throat** years ago")



Targeted gender

Detects the hateful tone of voice but fails a identifying the gender of the target subject

"Ho una voglia di prendere a schiaffi Sarri"*

("I want to slap Sarri")

*Sarri is an Italian male football coach



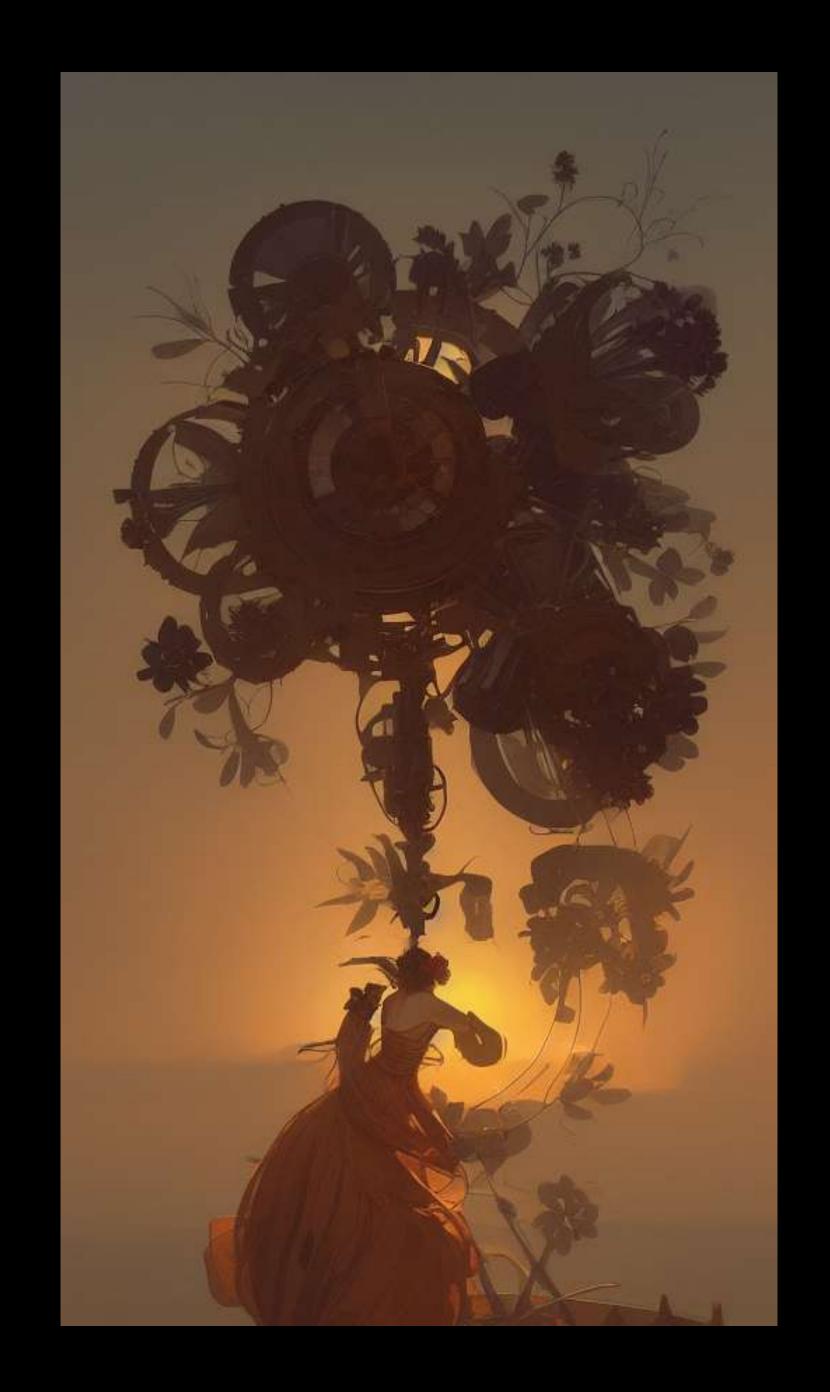
Quoted speech

Tweets that quote misogynous speech to support the opposite message

""Antonella è **acida** perchè non ha avuto figli" teorie lombrosiane" by Pietro Delle Piane #TemptationIsland"

("``Antonella is **sour** because she had no children`` Lombrosian theories by Pietro Delle Piane #TemptationIsland")

Entropy-based Attention Regularization



Evidence of Technical Bias

I am a gay man

Dixon et al. (2018)

Wussup, n*gga!

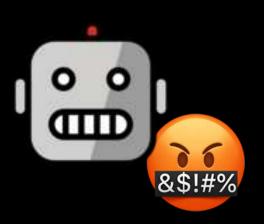
High toxicity scores

Sap et al. (2019)

• "[F]or many Africans, the most threatening kind of ethnic hatred is black against black." - New York Times

Kennedy et al. (2019)





Girl, ladore you

Misogyny detection model





Girl, ladore you

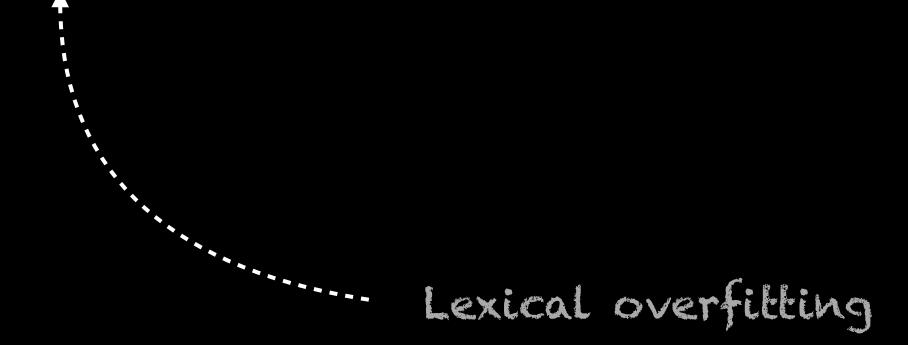
Misogyny detection model



Misogyny detection model



Girl, ladore you



Lacy

Mum

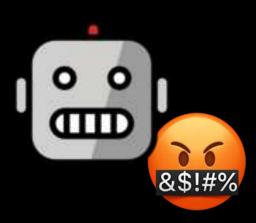
Woman

Girl, ladore you

Her

Grandma

Sister





Lady Mum Woman

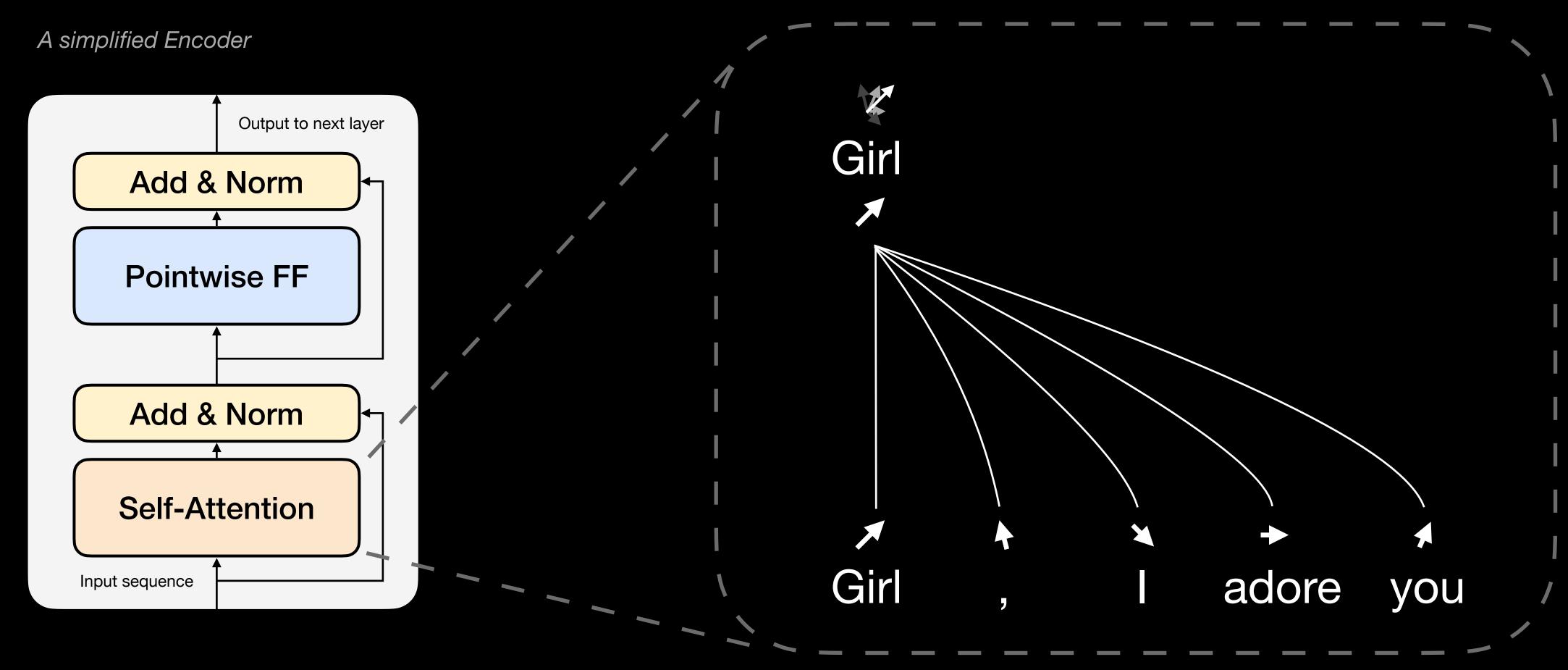
But lists bring issues

Incomplete

Sister

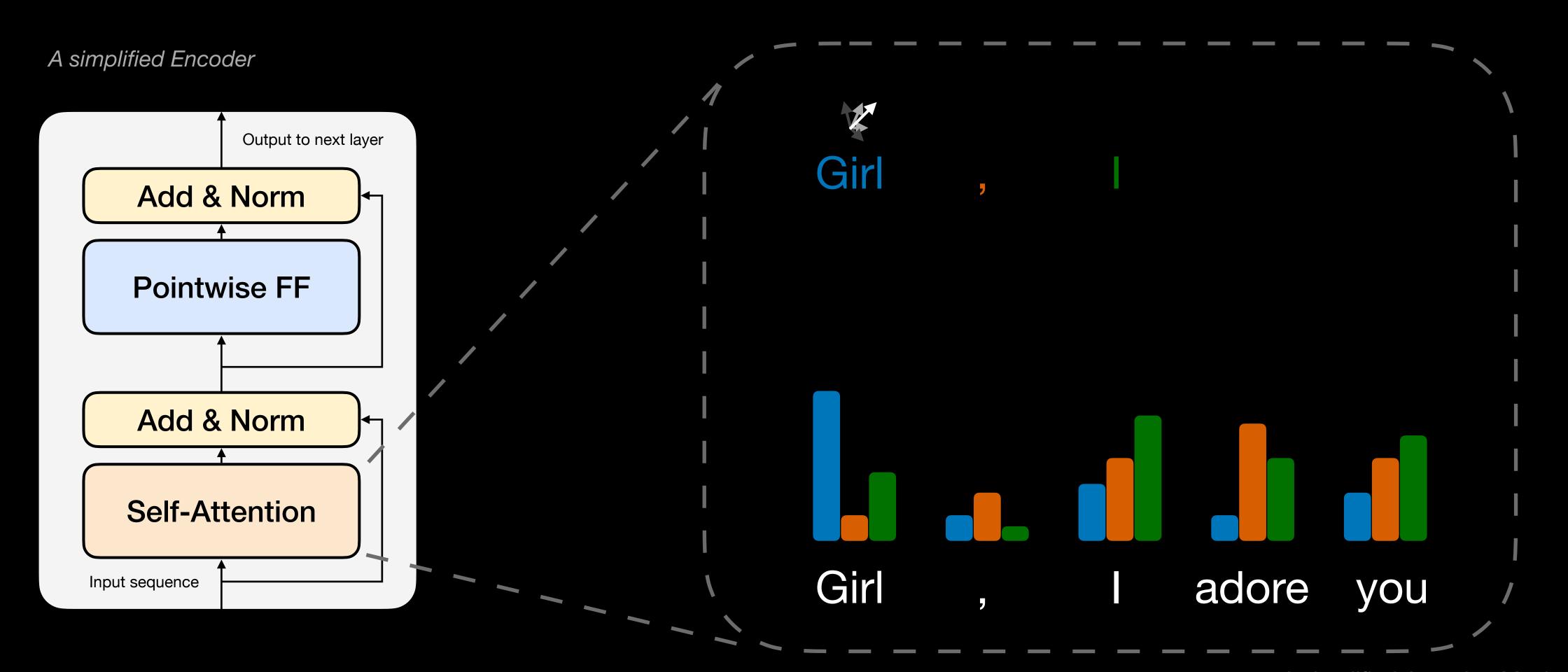
Align, Mix, and Repeat

Attention in Transformers



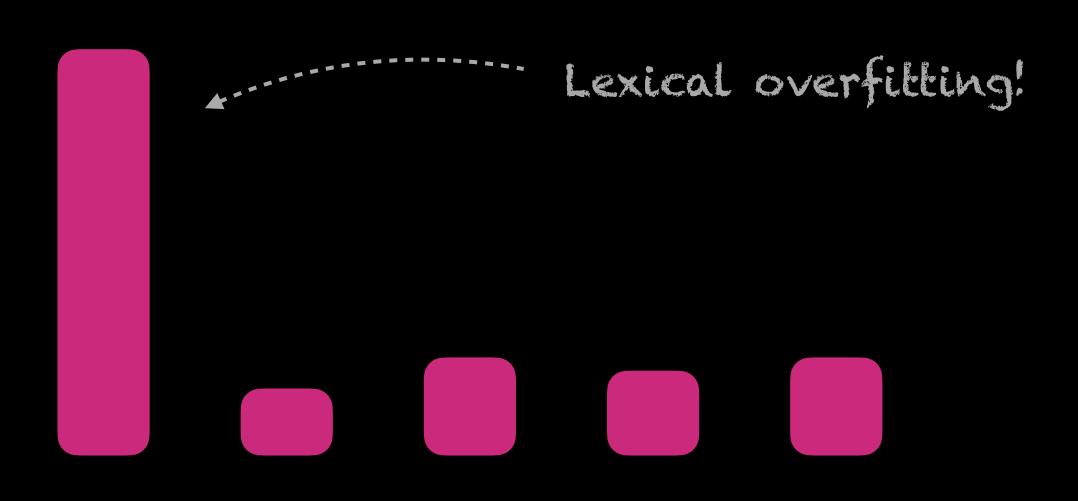
Align, Mix, and Repeat Attention in Transformers

Attention weighs the context we "pour" in new representations



So, where's the problem?

Girl, ladore you



Girl, ladore you

Attention? Context? Overfitting?

- Attention towers block information passage
- Poor passage means the same token regardless of the context
- I f*cking love the sh*t you're playin'
- Go f*ck yourself
- Eat sh*t senator
- I don't know sh*t about him

Entropy-based Attention Regularization S



Girl, ladore you

Girl, ladore you

Wider attention forces context in!



Girl, ladore you

Entropy-based Attention Regularization S

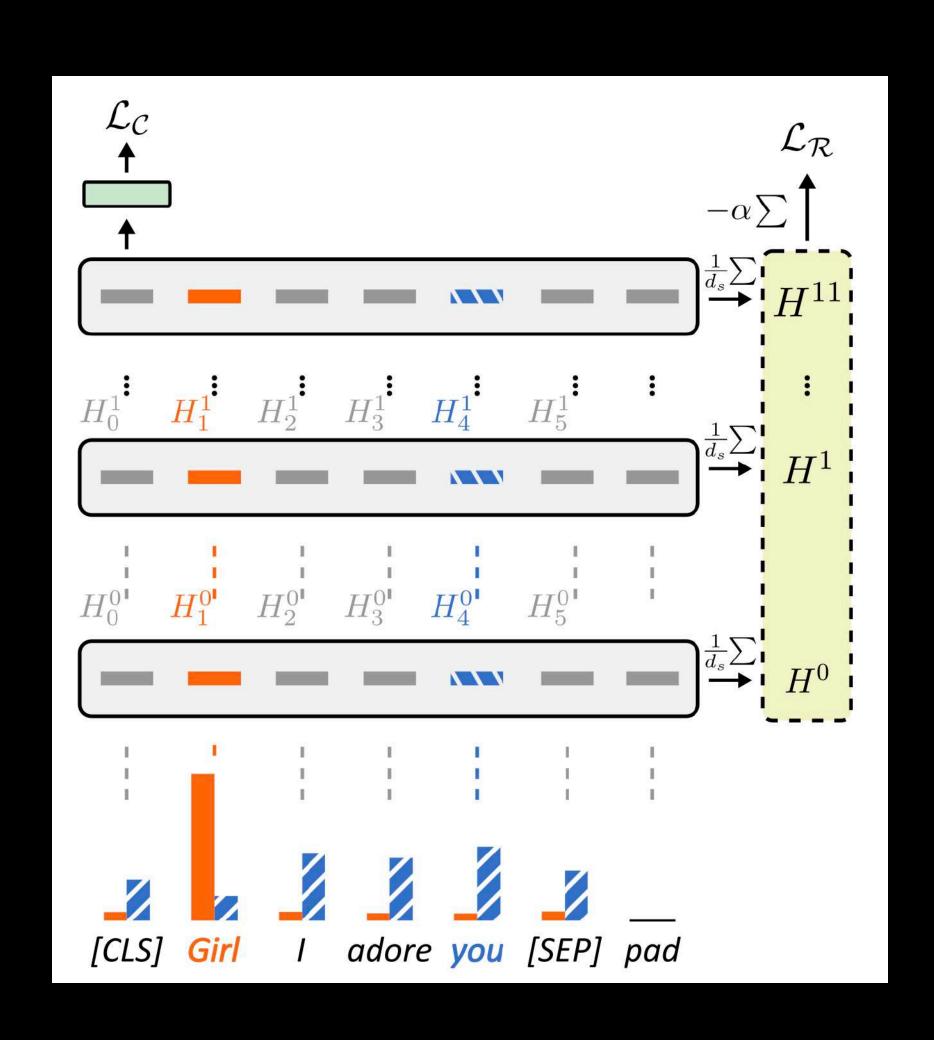


Compute Attention Entropy for each token

$$H_i^l = -\sum_j a_{i,j} \cdot log(a_{i,j})$$

Entropy-based Attention Regularization





Compute Attention Entropy for each token

$$H_i^l = -\sum_j a_{i,j} \cdot log(a_{i,j})$$

Add each contribution to the loss to maximise it

$$\mathcal{L} = \mathcal{L}_C + \mathcal{L}_{EAR}, \quad \mathcal{L}_{EAR} = -\alpha \sum_{l} \frac{\sum_{i} H_i^l}{d_s}$$

Entropy-based Attention Regularization S



- Cool things about EAR
 - It's list free!
 - Valid for any attention-based model
 - No additional parameters, inference pass, or heavy computation
 - Task-, domain-, and language- agnostic
- Discussion points
 - Regularization strength might vary from task to task

Classification and Extrinsic Bias

Datasets

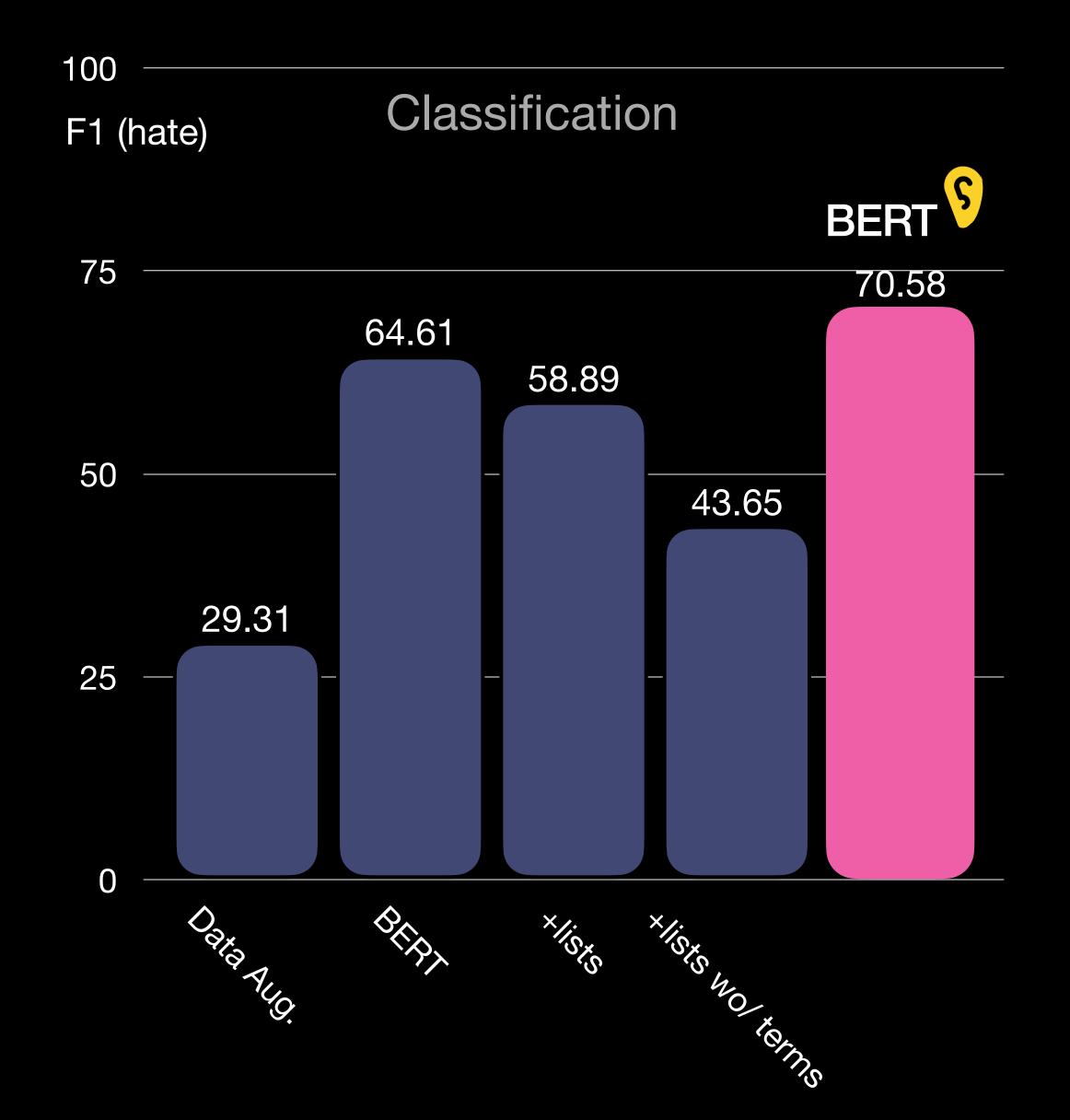
- Misogyny (ITA, ENG)
- Multilingual and Multi-Aspect Hate Speech (English part)
- Bias metrics
 - Subgroup Area Under the Curve (AUC)
 - Background Positive Subgroup Negative AUC
 - Background Negative Subgroup Positive AUC

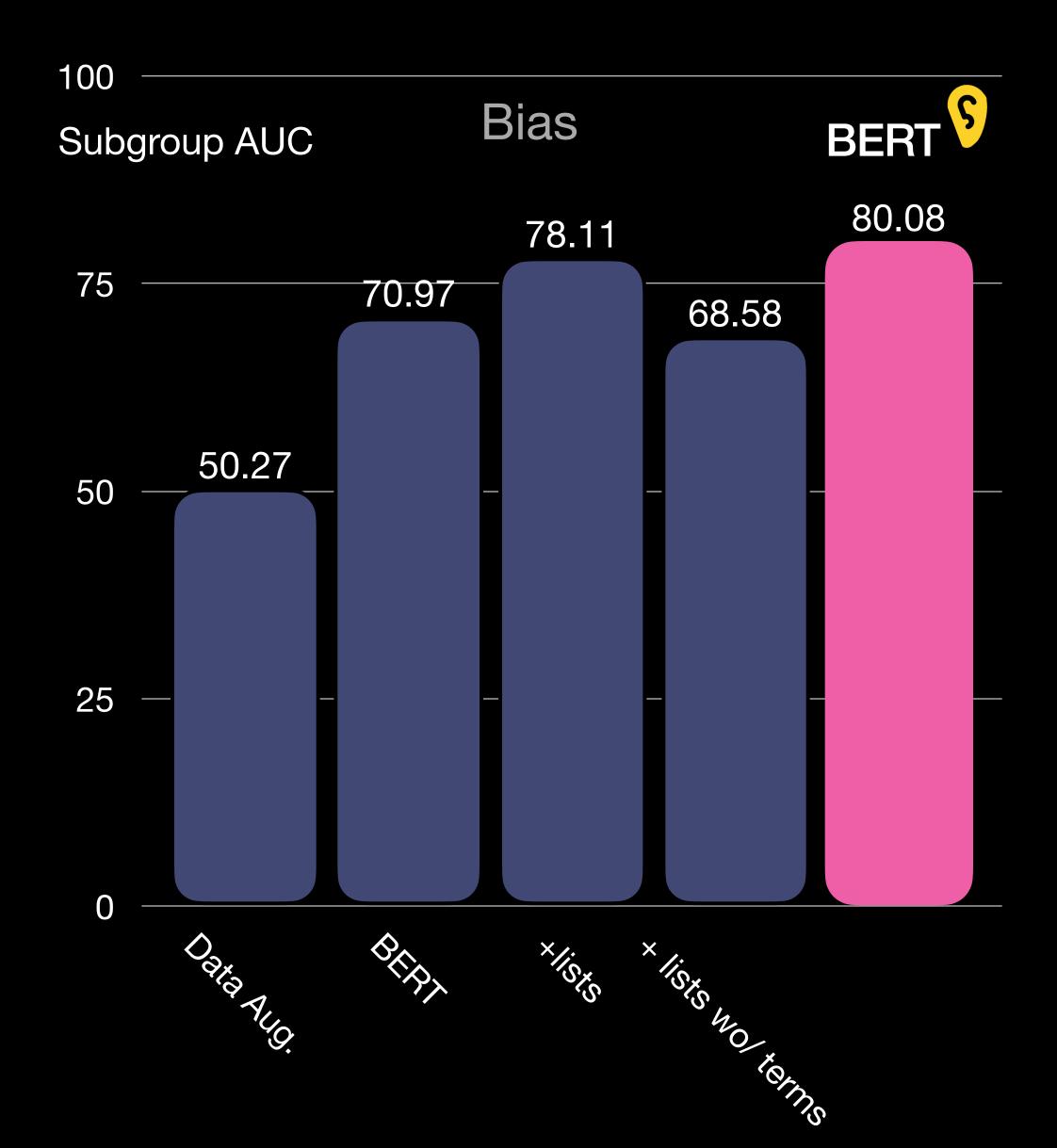
- Baselines
 - Data Augmentation
 - Identity term lists

Zhang et al. (2020), Kennedy et al. (2020)

Results on MISOGYNY

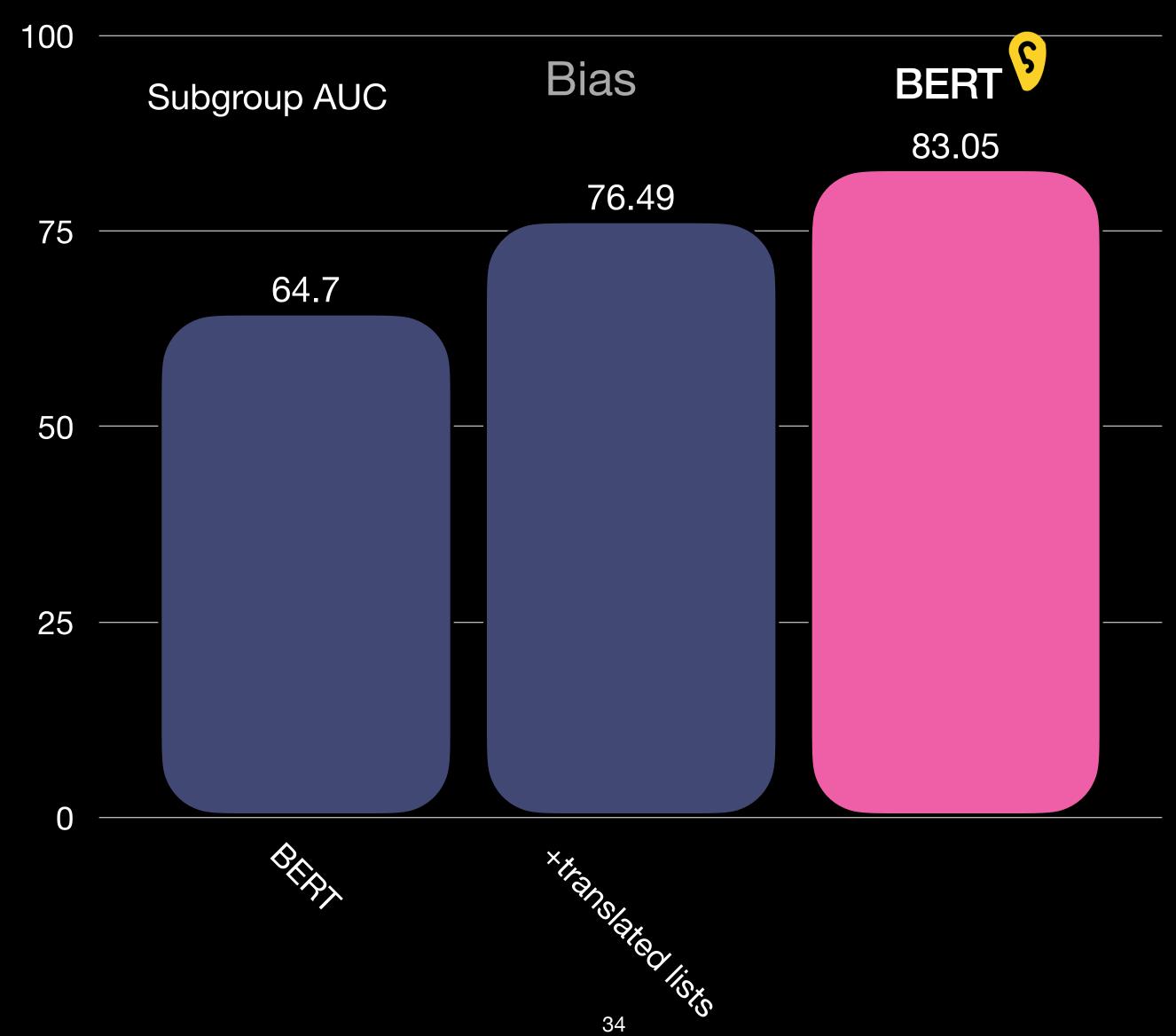






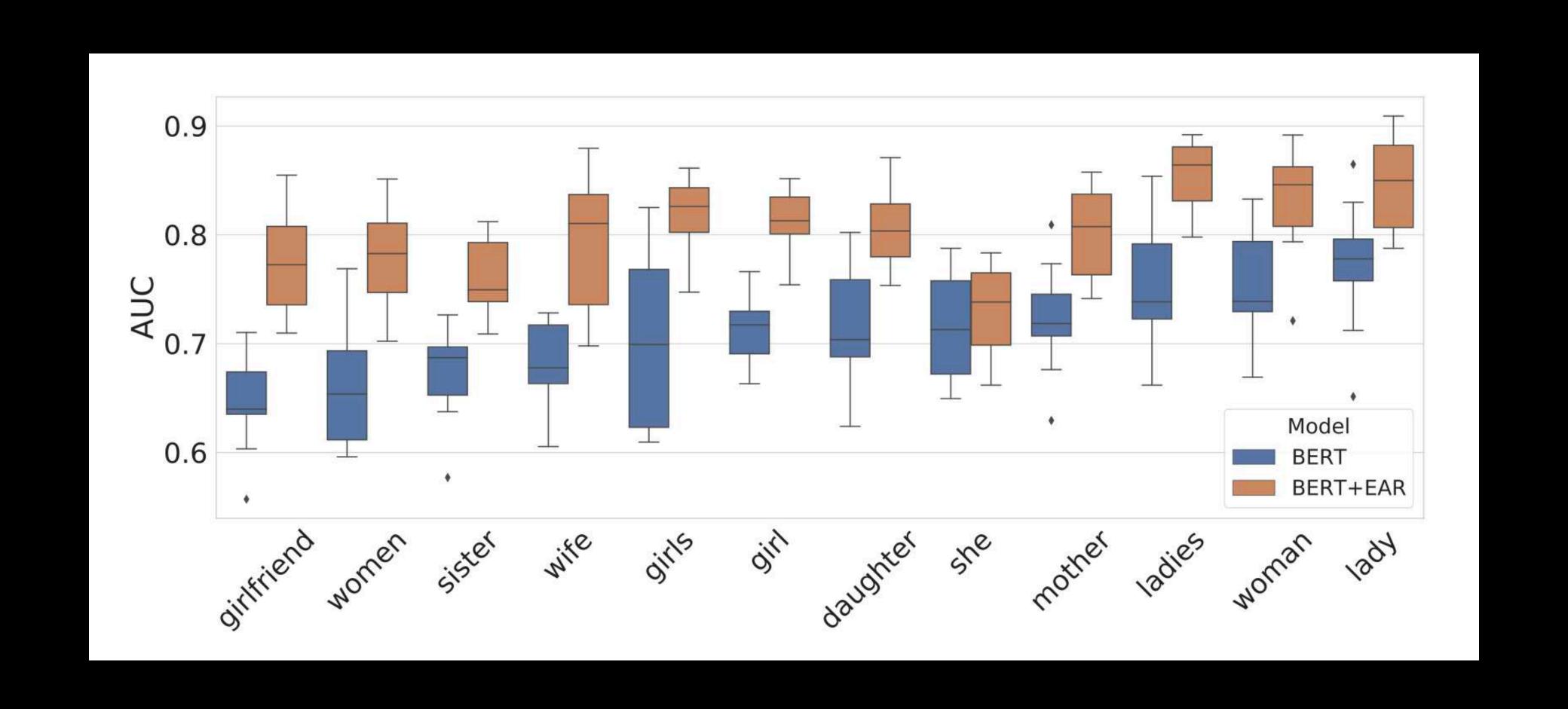
Results on MISOGYNY





Subgroups in MISOGYNY





Finding Overfitting Terms in a Corpus

Fine-tune your LM, then search for words with lowest attention entropy

Dataset	Overfitting terms
Misogyny (ENG)	girls, womens*ck, hoes, c*ck, shut, stupid, hoe, p*ssy, trying, f*ck
Misogyny (ITA)	pezzo, bel, bellissima, scoperei, p*ttanona, zitta, sb*rro, t*ttona, bella, c*lone (piece, nice, very nice, I'd f*ck, sl*t, shut up, c*m, b*sty, beautiful, fat*ss)
MLMA	n*gger, n*gro, shut, chong, ching, d*ke, okay, sp*c, tw*t, f*ggot

These are not "frequent": DF is 5% against 33% of frequent words

But some are frequent for the Positive class!

What about deployed models?

- No access to training
 - We need post-hoc intervention
- Do we have the tools?

- Explainability techniques are established approaches
 - Let's use them!



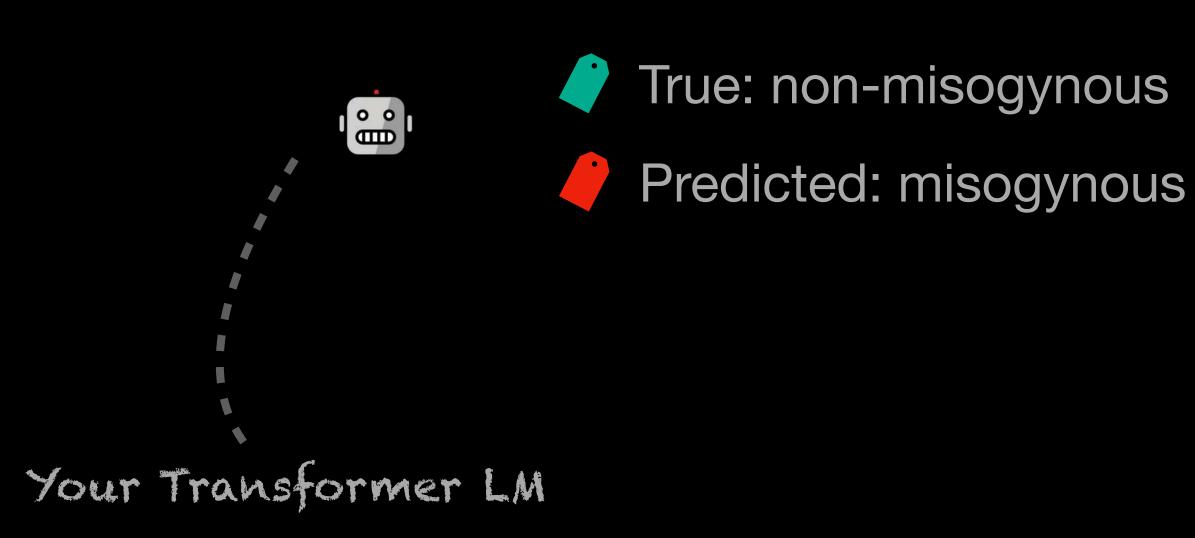


Benchmarking Post-Hoc Interpretability on Transformers

How you should explain transformers misogyny detectors

and how you should not

You are a smart woman

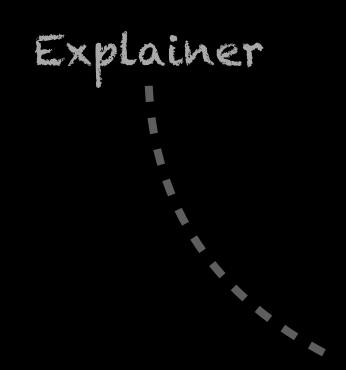


You are a smart woman





Predicted: misogynous



Darker color is stronger importance

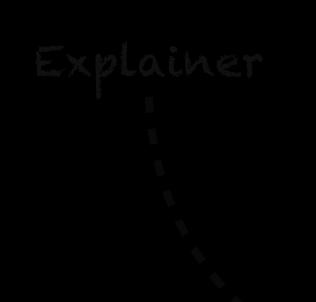
- You are a smart woman
- You are a smart woman
- You are a smart woman
- E4 You are a smart woman

You are a smart woman





Predicted: misogynous



Darker color is stronger importance

Explainers can disagree

- E1 You are a smart woman
- E2 You are a smart woman
- E³ You are a smart woman
- E4 You are a smart womar

Benchmarking setup

2. Run inference Once I called him and ... 3. Explain using token contribution You are a smart woman 1. Fine-tune BERT misogyny detection model There are so many of ... I am so angry at her E1 E2 E3 E4

Explanation evaluation

Plausibility: align with human expectation

Faithfulness: adhere with the model's inner working



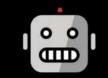
Explaining "misogyny" for a biased model

You are a smart woman

Explanation evaluation

Plausibility: align with human expectation

Faithfulness: adhere with the model's inner working



Explaining "misogyny" for a biased model

Plausible



Explanation evaluation

Plausibility: align with human expectation

Faithfulness: adhere with the model's inner working



Explaining "misogyny" for a biased model

Non-faithful

~0

~0

~0

~0

-30

Score change (%)
when removed

The long-standing debate Is Attention Explanation?

Raw attention

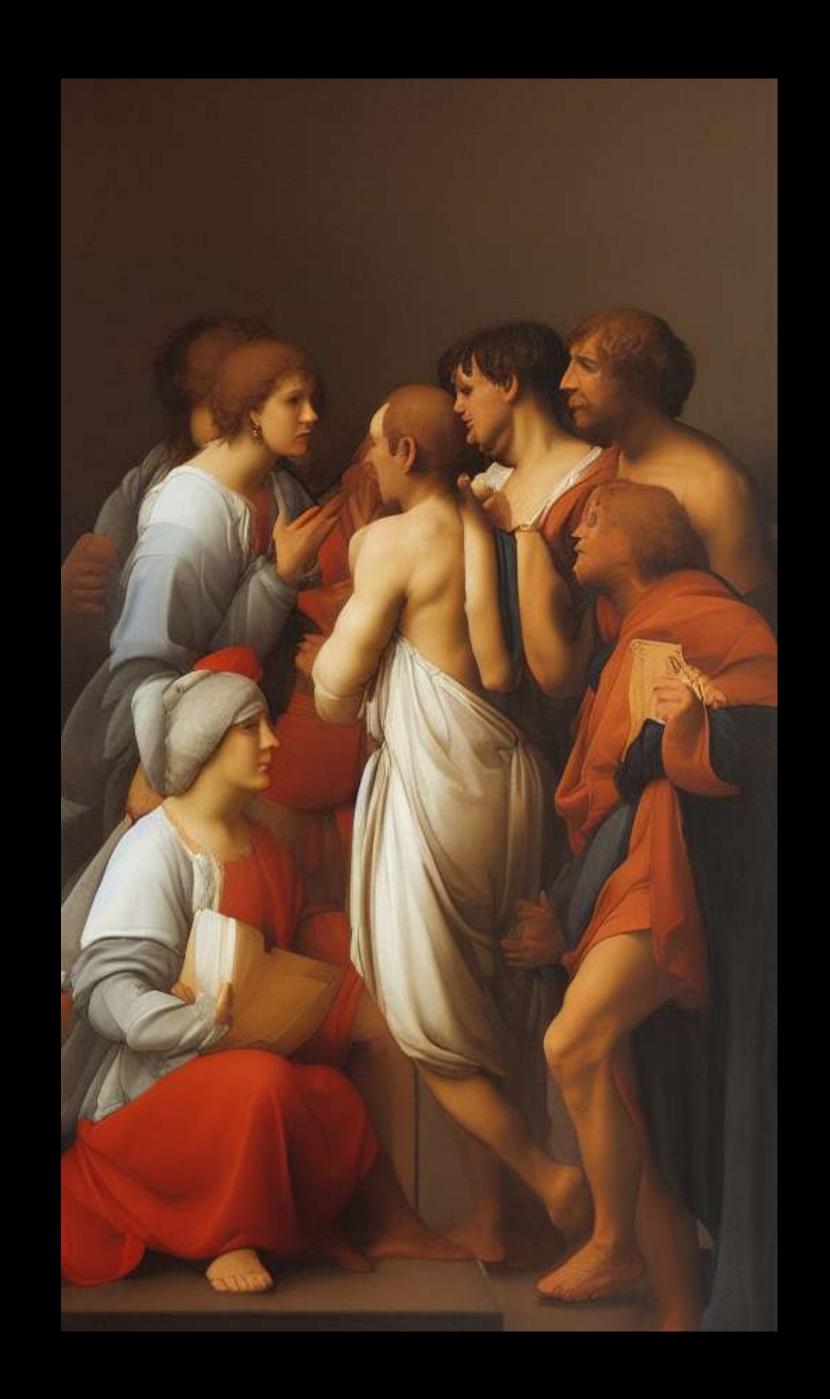
Effective attention

Sun and Marasovic (2020)

Brunner et al. (2020)

Hidden Token Attribution

Brunner et al. (2020)



Error analysis



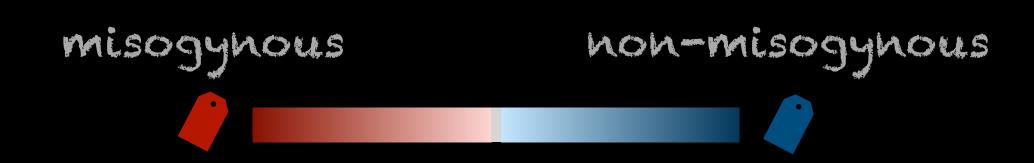
Predicted: misogynous

You pussy boy

Error analysis



Predicted: misogynous



Human plausibility: YOU DUSSY DOY

~0 ~0 -36 +1 Prediction change (%)

Gradient You pussy boy Integrated Gradients YOU DUSSY 00 SHAP YOU DUSSY DOY Sampling-and-Occlusion YOU DUSSY DOY

Lexical overfitting and stereotypes

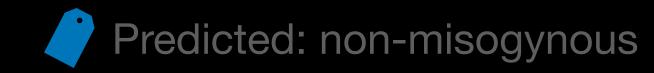
Predicted: misogynous

Ann is in the kitchen

-40 15 13 -13

-24

PC (%)



David is in the kitchen

G Ann is in the kitchen

Ann is in the kitchen

Ann is in the kitchen

Ann is in the kitchen David is in the kitchen

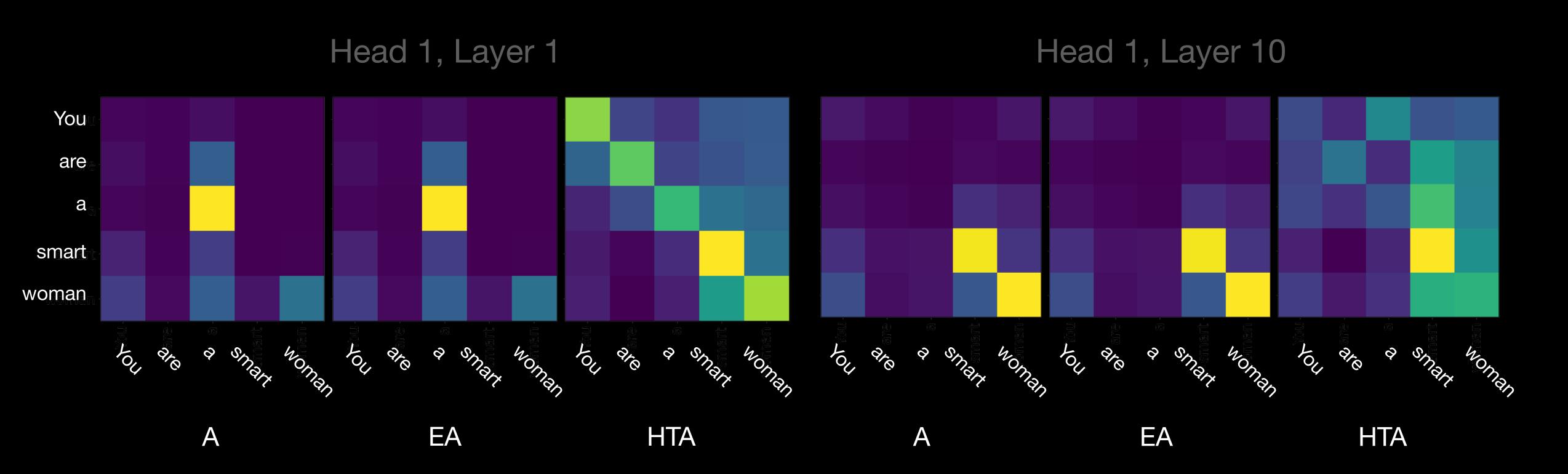
David is in the kitchen David is in the kitchen

David is in the kitchen

What about Attention?



Self-attention on "You are a smart woman"



A/EA are different from HTA!

What about Attention?



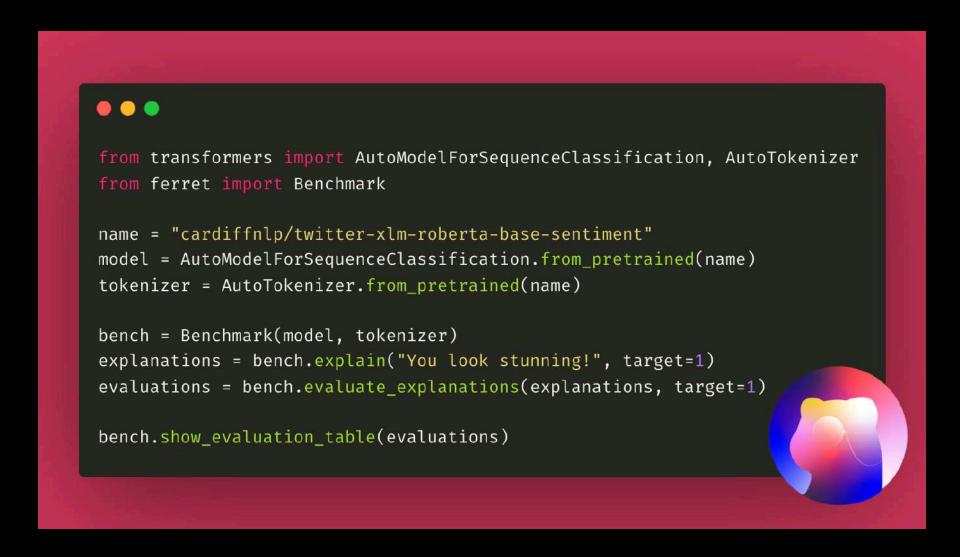
Self-attention on "You are a smart woman"



A note on Explainability in Transformers

- Standard post-hoc methods can disagree
- SHAP or SOC is preferable
- Those let us spot lexical overfitting
 - Trigger words or stereotypes reinforcers
 - More work should bridge interpretability and bias evaluation

A principled evaluation with ferret

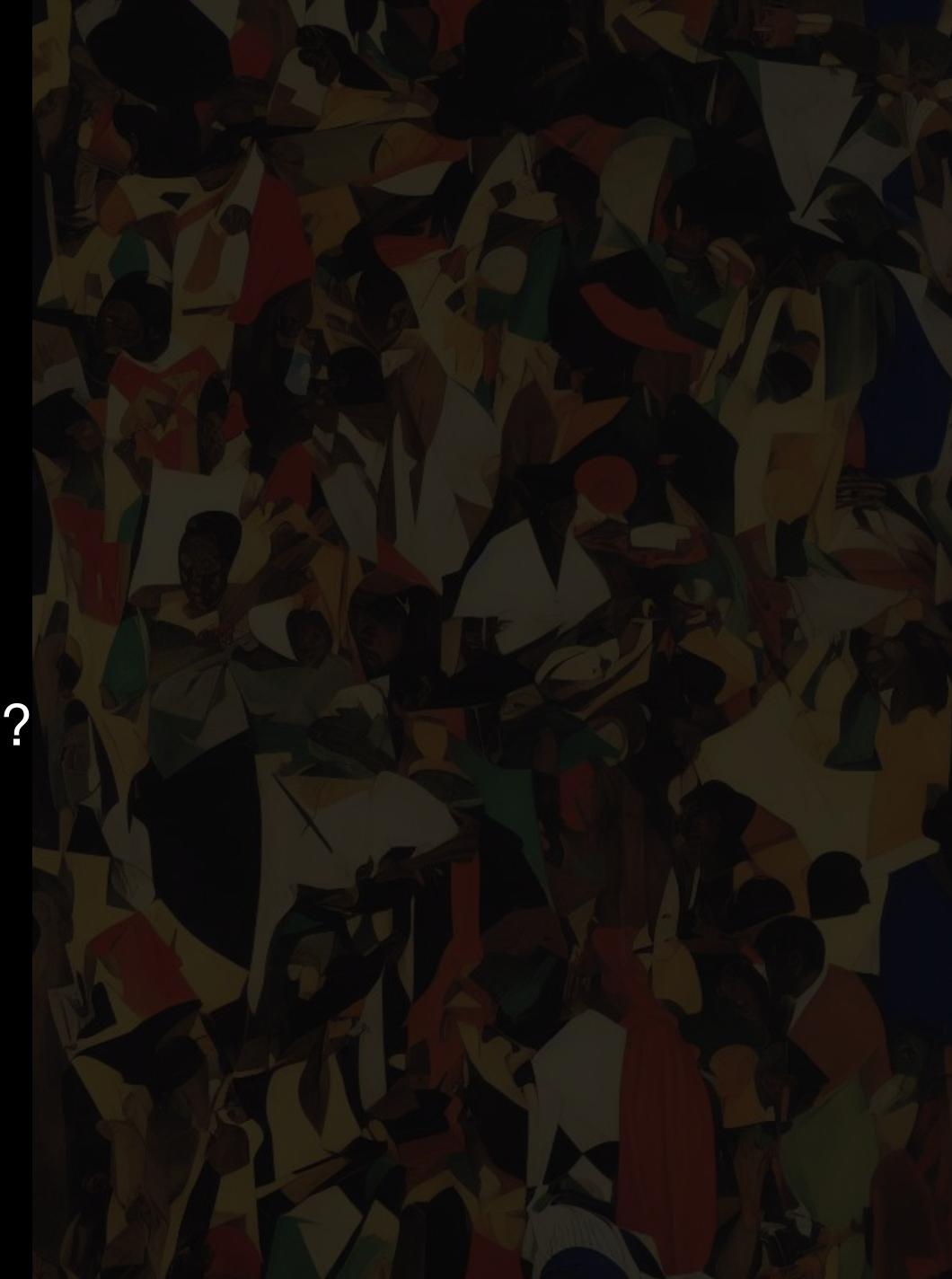


Wrapping Up

- The issue of bias in Transformer Language Models is multi-faceted
- We addressed some of the technological causes
- Lexical overfitting affects modern hate speech detection models
 - Sentence Embedding models benefit lexical support
 - EAR 5 mitigates LO with a stronger contextualization
 - Attention entropy finds problematic words
- Post-hoc explainability spots source of errors
 - But handle it with care

Future Directions The Short-Term Perspective

- EAR belongs to a family of generic generalization approaches
 - What happens to internal representations?
 - New tasks, model/dataset size, pre-training?
 - Constrained, bias-aware learning?
- More focused extrinsic bias evaluation

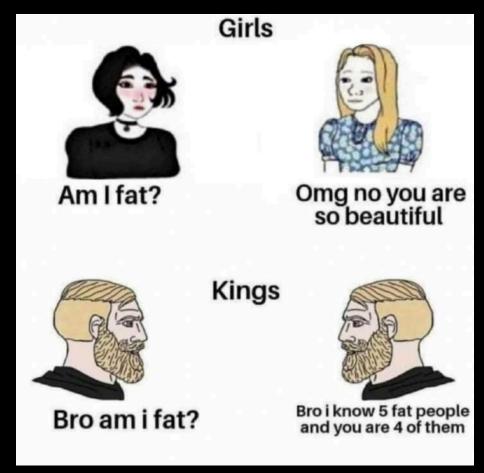


Future Directions The Long-Term Perspective

- Hate spreads via new media
 - Multi-modal scenarios, e.g., memes

Stereotypical memes Leverage common-sense knowledge

Very challenging for annotators!



My angry girlfriend: "I'm fine"

Me:

I am trying to solve a puzzle.

Non-misogynous

Misogynous

Unclear guidelines?
Annotators' subjectivity?

Future Directions The Long-Term Perspective

- Hate spreads via new media
 - Multi-modal scenarios
- Diachronic and demographic-aware adaptation
 - All ingredients get outdated
 - Ultimately, text won't be enough. Nuances come from stakeholders' demographic backgrounds

Future Directions The Long-Term Perspective

- Hate spreads via new media
 - Multi-modal scenarios
- Diachronic and demographic-aware adaptation
 - All ingredients get outdated: data, models, language
 - Ultimately, text won't be enough. Nuances come from stakeholders' demographic backgrounds
- True generalization and the role of NLP in the journey

Scientific Production













Social Bias in Hate Speech Detection

Giuseppe Attanasio, Debora Nozza, Dirk Hovy, and Elena Baralis. 2022. Entropy-based Attention Regularization Frees Unintended Bias Mitigation from Lists. In Findings of the Association for Computational Linguistics: ACL 2022, pages 1105–1119, Dublin, Ireland. Association for Computational Linguistics.

Debora Nozza, Federico Bianchi, and Giuseppe Attanasio. 2022. HATE-ITA: Hate Speech Detection in Italian Social Media Text. In Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH), pages 252–260, Seattle, Washington (Hybrid). Association for Computational Linguistics.

Attanasio, Giuseppe, and Eliana Pastor. "PoliTeam@ AMI: Improving Sentence Embedding Similarity with Misogyny Lexicons for Automatic Misogyny Identification in Italian Tweets." In EVALITA. 2020.

Explainability in NLP

Giuseppe Attanasio, Debora Nozza, Eliana Pastor, and Dirk Hovy. 2022. Benchmarking Post-Hoc Interpretability Approaches for Transformer-based Misogyny Detection. In Proceedings of NLP Power! The First Workshop on Efficient Benchmarking in NLP, pages 100–112, Dublin, Ireland. Association for Computational Linguistics.

Attanasio, Giuseppe, Eliana Pastor, Chiara Di Bonaventura, and Debora Nozza. "ferret: a Framework for Benchmarking Explainers on Transformers." arXiv preprint arXiv:2208.01575 (2022).

Scientific Production













Language and Inclusive Communication

Attanasio, Giuseppe, Salvatore Greco, Moreno La Quatra, Luca Cagliero, Michela Tonti, Tania Cerquitelli, and Rachele Raus. "E-MIMIC: Empowering Multilingual Inclusive Communication." In 2021 IEEE International Conference on Big Data (Big Data), pp. 4227-4234. IEEE, 2021.

Raus, Rachele, Michela Tonti, Tania Cerquitelli, Luca Cagliero, Giuseppe Attanasio, Moreno La Quatra, and Salvatore Greco. "L'analyse du discours et l'intelligence artificielle pour réaliser une écriture inclusive: le projet E-MIMIC." In SHS Web of Conferences, vol. 138, p. 01007. EDP Sciences, 2022.

Vision-Language Multimodality

Giuseppe Attanasio, Debora Nozza, and Federico Bianchi. 2022. MilaNLP at SemEval-2022 Task 5: Using Perceiver IO for Detecting Misogynous Memes with Text and Image Modalities. In Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), pages 654–662, Seattle, United States. Association for Computational Linguistics.

Chia, Patrick John, Giuseppe Attanasio, Federico Bianchi, Silvia Terragni, Ana Rita Magalhães, Diogo Goncalves, Ciro Greco, and Jacopo Tagliabue. "Fashionclip: Connecting language and images for product representations." arXiv preprint arXiv:2204.03972 (2022). — under review in a peer-reviewed journal

Bianchi, Federico, Giuseppe Attanasio, Raphael Pisoni, Silvia Terragni, Gabriele Sarti, and Sri Lakshmi. "Contrastive Language-Image Pre-training for the Italian Language." arXiv preprint arXiv:2108.08688 (2021).

Scientific Production













Quantitative Explainable Trading

Bellocca, Gian Pietro, Giuseppe Attanasio, Luca Cagliero, and Jacopo Fior. "Leveraging the momentum effect in machine learning-based cryptocurrency trading." Machine Learning with Applications 8 (2022): 100310.

Cagliero, Luca, Paolo Garza, Giuseppe Attanasio, and Elena Baralis. "Training ensembles of faceted classification models for quantitative stock trading." Computing 102, no. 5 (2020): 1213-1225.

Attanasio, Giuseppe, Luca Cagliero, and Elena Baralis. "Leveraging the explainability of associative classifiers to support quantitative stock trading." In Proceedings of the Sixth International Workshop on Data Science for Macro-Modeling, pp. 1-6. 2020.

Attanasio, Giuseppe, Luca Cagliero, Paolo Garza, and Elena Baralis. "Combining news sentiment and technical analysis to predict stock trend reversal." In 2019 International Conference on Data Mining Workshops (ICDMW), pp. 514-521. IEEE, 2019.

Attanasio, Giuseppe, Luca Cagliero, Paolo Garza, and Elena Baralis. "Quantitative cryptocurrency trading: exploring the use of machine learning techniques." In Proceedings of the 5th Workshop on Data Science for Macro-modeling with Financial and Economic Datasets, pp. 1-6. 2019.

Data Science Tools

Attanasio, Giuseppe, Flavio Giobergia, Andrea Pasini, Francesco Ventura, Elena Baralis, Luca Cagliero, Paolo Garza, Daniele Apiletti, Tania Cerquitelli, and Silvia Chiusano. "DSLE: a smart platform for designing data science competitions." In 2020 IEEE 44th Annual Computers, Software, and Applications Conference (COMPSAC), pp. 133-142. IEEE, 2020.

Current Activities













Vision-Language Models

Giovanni Cassani, Giuseppe Attanasio, Federico Bianchi. "Sound symbolic associations via contrastive learning of text and images", Architectures and Mechanisms of Language Processing (2022), AMLaP 2022. (https://virtual.oxfordabstracts.com/#/event/3067/submission/103)

Fairness and Recommender Systems

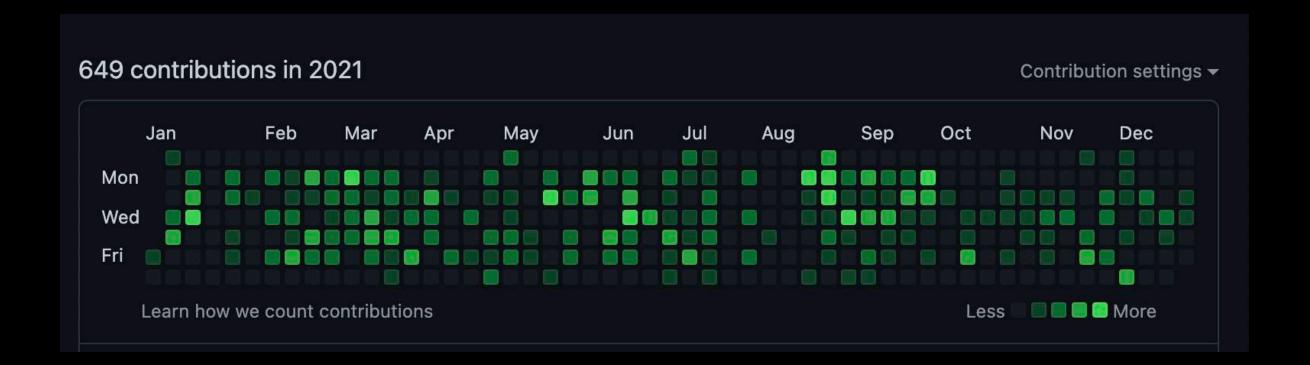
Tagliabue, Jacopo, Federico Bianchi, Tobias Schnabel, Giuseppe Attanasio, Ciro Greco, Gabriel de Souza P. Moreira, and Patrick John Chia. "EvalRS: a Rounded Evaluation of Recommender Systems." arXiv preprint arXiv:2207.05772 (2022). - to appear as workshop and data challenge at International Conference on Information and Knowledge Management (2022), CIKM 2022. (https://reclist.io/cikm2022-cup/)

Preliminary studies

- Evolution of social biases in language models over time
- Characterisation of the Italian government's financial aids during COVID-19 project funded by Fondazione Cariplo
- Fairness evaluation of modern speech systems collaboration with Amazon Alexa Al
- BERT in cognition: do contextualised representations provide a stronger representation of language as per cognition experiments?

Sharing is Caring

Open, reproducible research is good research











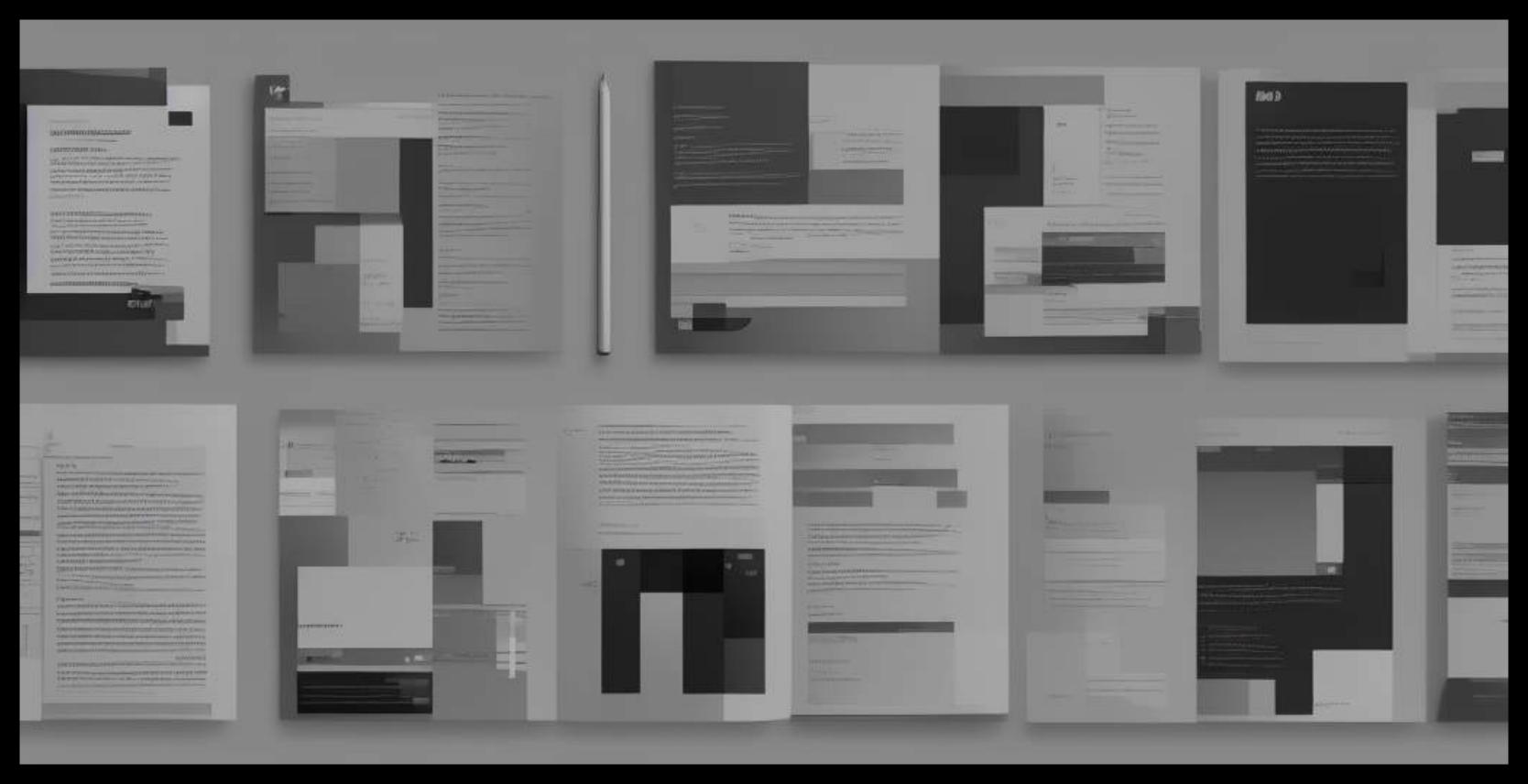












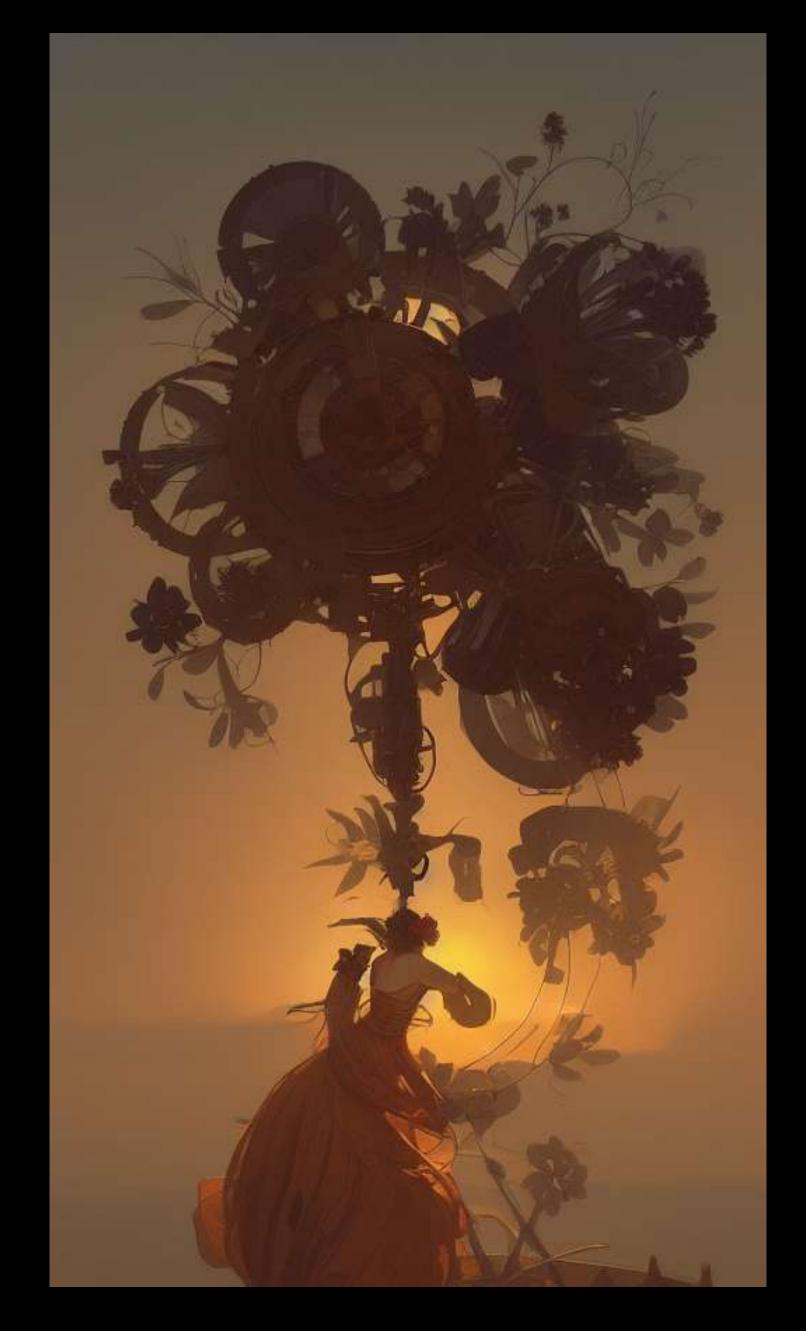
Contributions of an academic paper minimalism black and white trending on artstation HQ 2K



Philosophers debating on whether attention is explanation by Raffaello oil painting trending on artstation HQ



A busy street in Seattle. Elegant, intricate, digital painting, black and white, artstation, concept art, smooth, sharp focus, illustration, art by artgerm and greg rutkowski and alphonse mucha



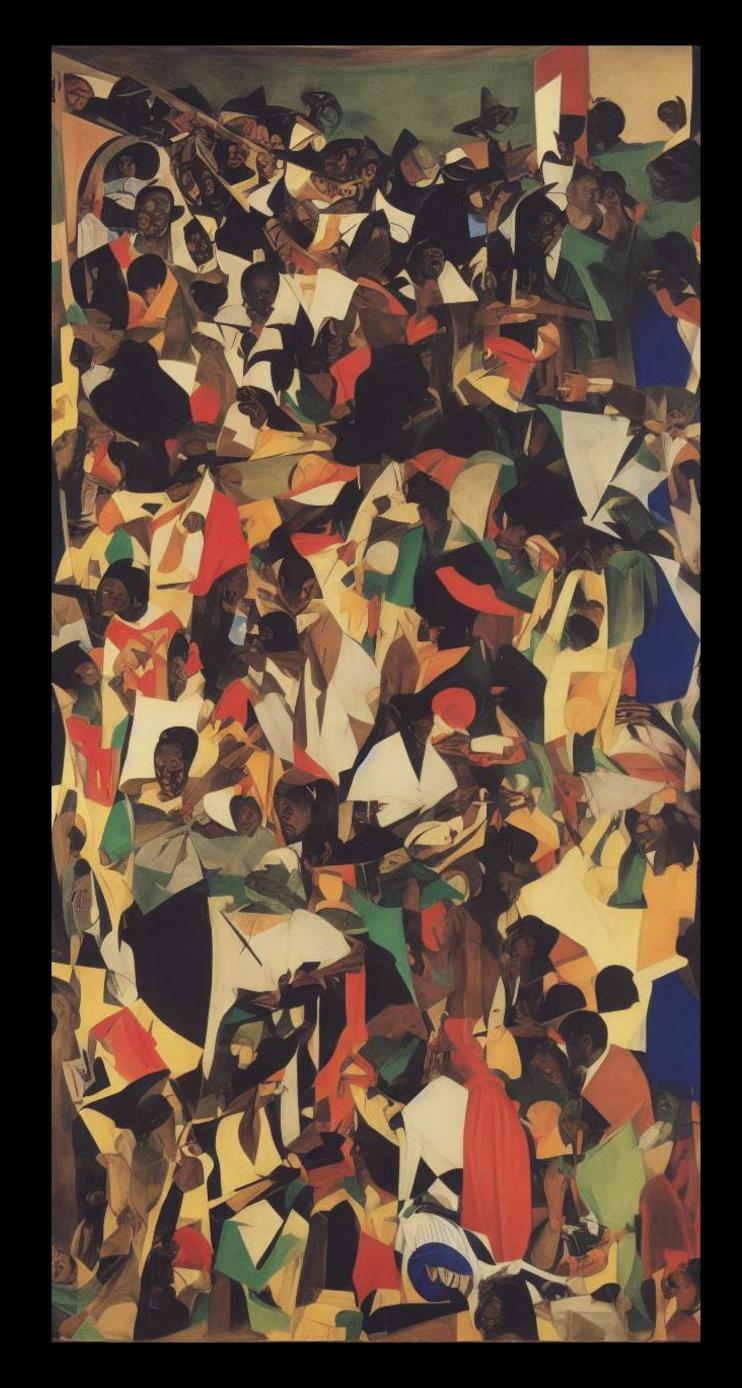
Mechanical flower at sunset, digital painting, artstation, concept art, smooth, sharp focus, illustration, art by greg rutkowski and alphonse mucha



Peaceful rural landscape at daw, digital painting, artstation, concept art, black and white, smooth, illustration, art by greg rutkowski and alphonse mucha



Science laboratory, sharp focus, emitting diodes, smoke, artillery, sparks, racks, system unit, motherboard, by pascal blanche rutkowski repin artstation hyperrealism painting concept art of detailed character design matte painting, 4 k resolution blade runner



A photo of a fellow scholar giving a talk, by Jacob Lawrence and Francis picabia, perfect composition, beautiful detailed intricate insanely detailed octane render trending on artstation, 8 k artistic photography, photorealistic concept art, soft natural volumetric cinematic perfect light, chiaroscuro, award - winning photograph, masterpiece, oil on canvas, raphael, caravaggio, greg rutkowski, beeple, beksinski, giger



A scientist giving a talk, acrylic painting, trending on pixiv fanbox, palette knife and brush strokes, style of makoto shinkai jamie wyeth james gilleard edward hopper greg rutkowski studio ghibli genshin impact