FEATURES OR SPURIOUS ARTIFACTS?

Data-centric baselines for fair and robust hate speech detection

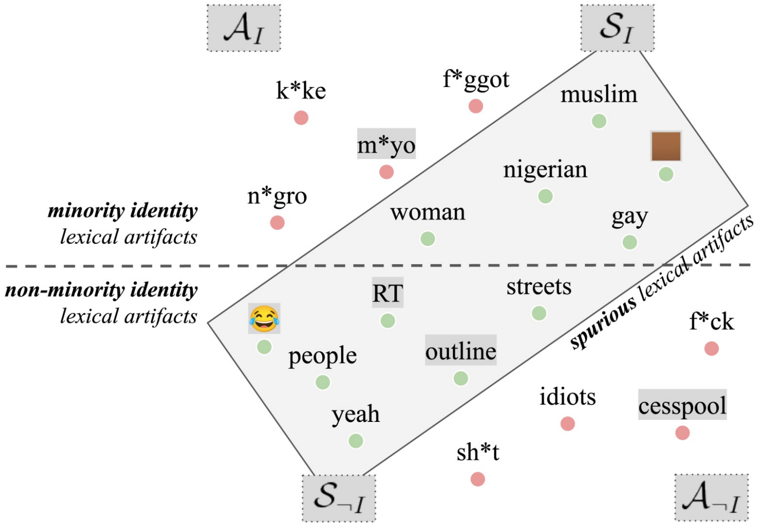
Alan Ramponi, Sara Tonelli *– Fondazione Bruno Kessler, Italy*

FONDAZIONE BRUNO KESSLER

LEXICAL ARTIFACTS

*Hate speech detection models overly rely on lexical artifacts, affecting **fairness** & **robustness***

ARE LEXICAL ARTIFACTS *ALL THE SAME*?



LEXICAL ARTIFACTS *ACROSS* PLATFORMS

Hate speech corpora from:

Reddit Gab Twitter Stormfront

lang: ***English***
mod: ***written***

Cross-distribution PMI *& artifacts annotation*

DATA-CENTRIC LEXICAL DEBIASING

Use *<u>identity</u>* *($\mathcal{S}_I$) and* *<u>non-identity</u>* *($\mathcal{S}_{\neg I}$)* ***spurious artifacts*** *for debiasing models*

- ***Fairness:*** *MASK $\mathcal{S}_I$ strong baseline*
- ***Robustness:*** *Hard to achieve with artifacts or generic solutions only*

Fairness & robustness ***should be studied together*** *in future research*

ARTIFACTS STATEMENT

Documentation *of lexical artifacts to contribute to diagnosis & mitigation of pre-existing biases:*

Template:

I. *TOP LEXICAL ARTIFACTS*
II. *CLASS DEFINITIONS*
III. *METHODS & RESOURCES*

RESOURCES

```
github.com/dhfbk/
hate-speech-artifacts
```