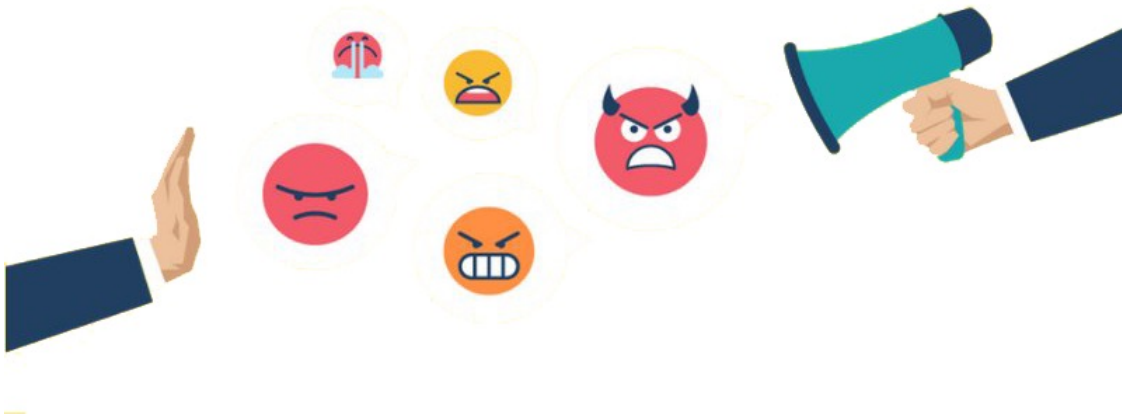
NAACL 2022
Seattle, WA, USA

Features or spurious artifacts?

Data-centric baselines for fair and robust hate speech detection

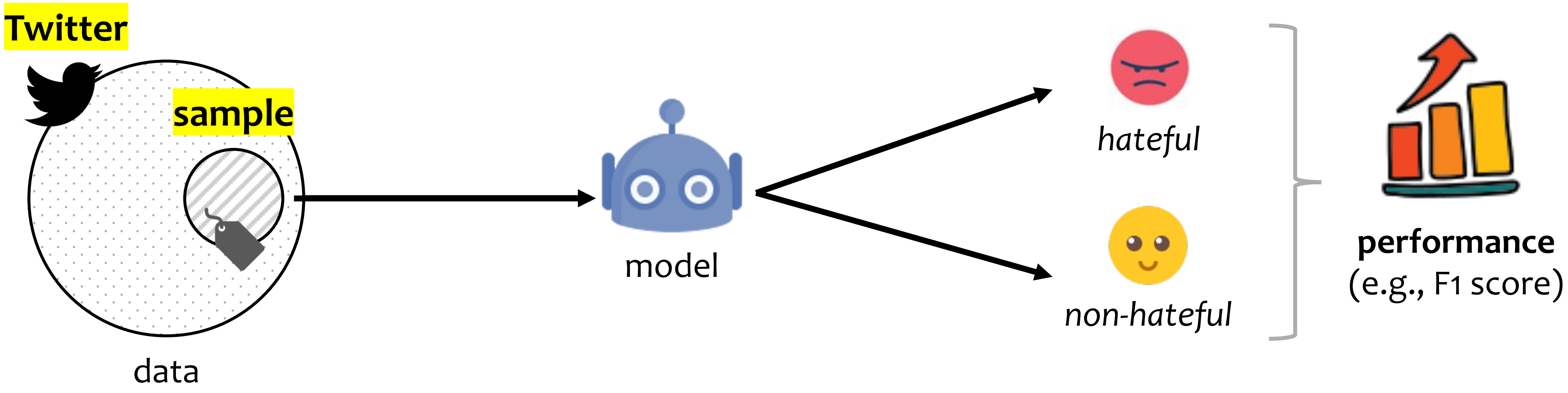
Alan **Ramponi**, Sara **Tonelli**

Fondazione Bruno Kessler, Trento, Italy



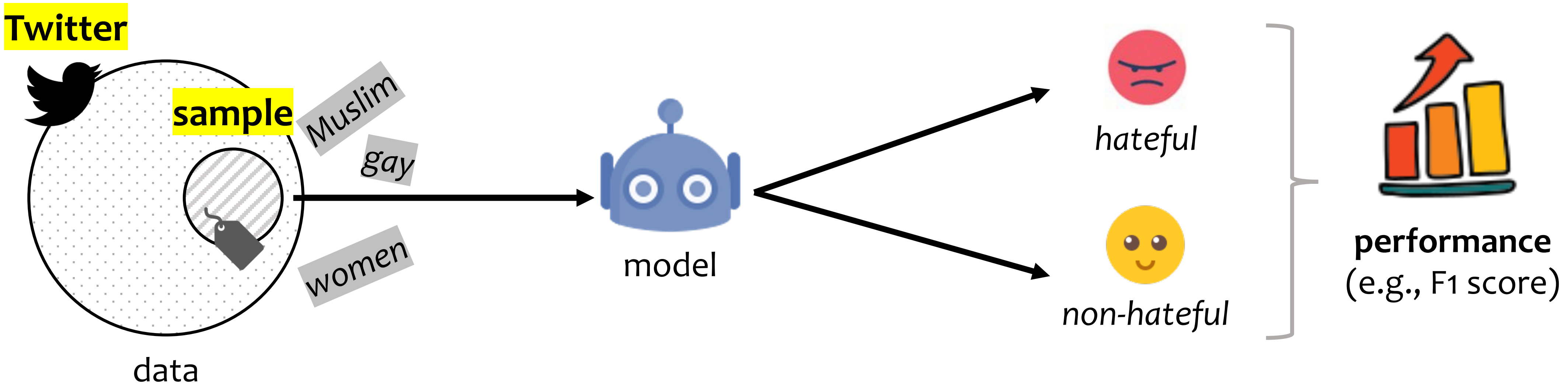
Warning: *This presentation contains content that may be offensive/upsetting*

Bias in hate speech detection



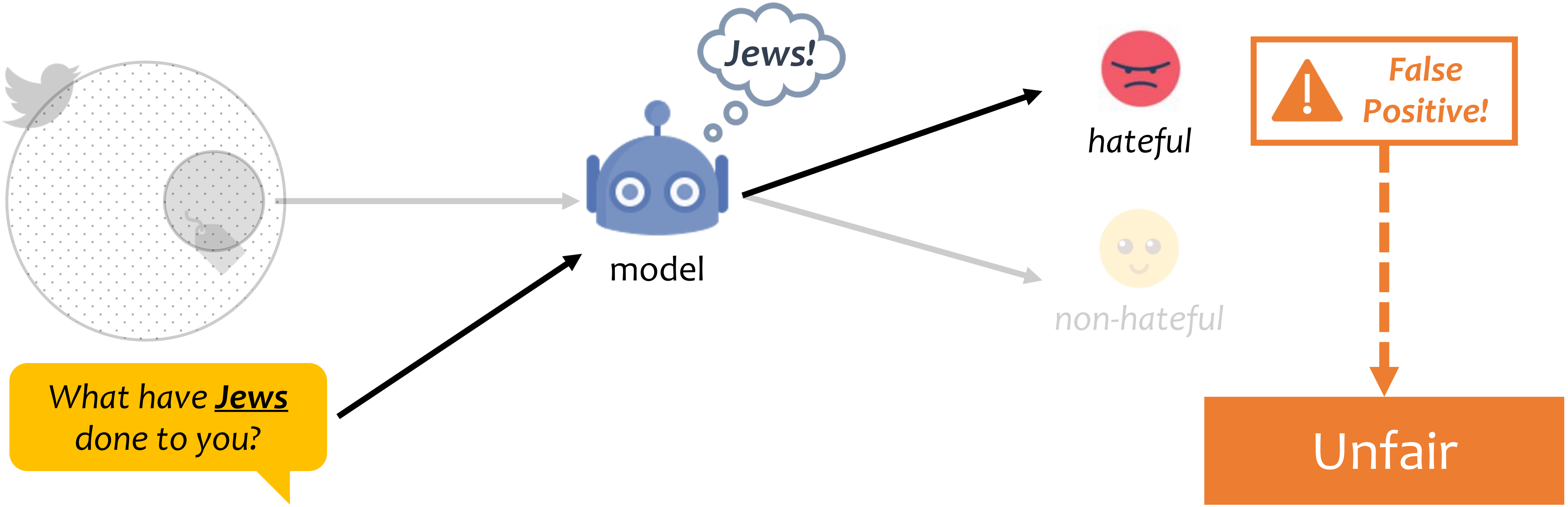
- **Focused sampling** introduces **topic-specific terms** [Wiegand+ 2019; *i.a.*]
- **Platforms**: norms, practices & lang use introduce **platform-specific terms**

Bias in hate speech detection



- **Focused sampling** introduces **topic-specific terms** [Wiegand+ 2019; *i.a.*]
- **Platforms**: norms, practices & lang use introduce **platform-specific terms**
- **Data collection** shapes distribution of hate targets – i.e., **identity terms**

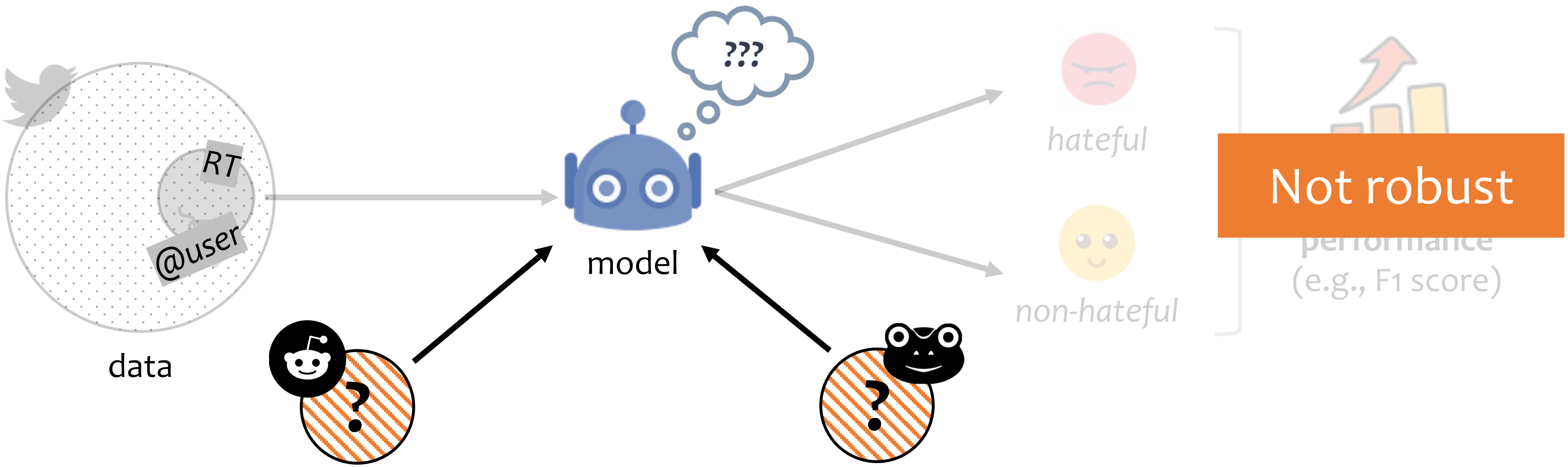
Undesired identity bias



Identity terms as shortcuts for prediction [Zhou+ 2021; Kennedy+ 2020; *i.a.*]

OOD: *out-of-distribution*
ID: *in-distribution*

Weak out-of-distribution robustness



Platform-specific terms as shortcuts for prediction “annotation artifacts” in NLI

Poliak+ 2018

Belinkov+ 2019

Gururangan+ 2018

inter alia

Focus of this work

Lexical artifacts in *hate speech detection*

> *“**Statistical correlations** between surface lexical items and labels in training data, **which models exploit to derive predictions**”*

Contributions

- **Characterization** and **cross-platform study** English
- **Impact on OOD robustness & fairness**
- **Lexical artifacts statement** for diagnosis of pre-existing bias

Characterization of lexical artifacts

***OI**:*	**possibly offensive identity mentions**	*possibly* offensive or stereotyping identity terms *e.g., n*gro, f*ggot*
***OnI**:*	**possibly offensive non-identity mentions**	*possibly* offensive swear words and profanities *e.g., f*ck, idiot*
***nOI**:*	**non-offensive identity mentions**	non-offensive terms describing identities *e.g., Jews, women, gay*

Lexical bias categories in [Zhou+ 2021]

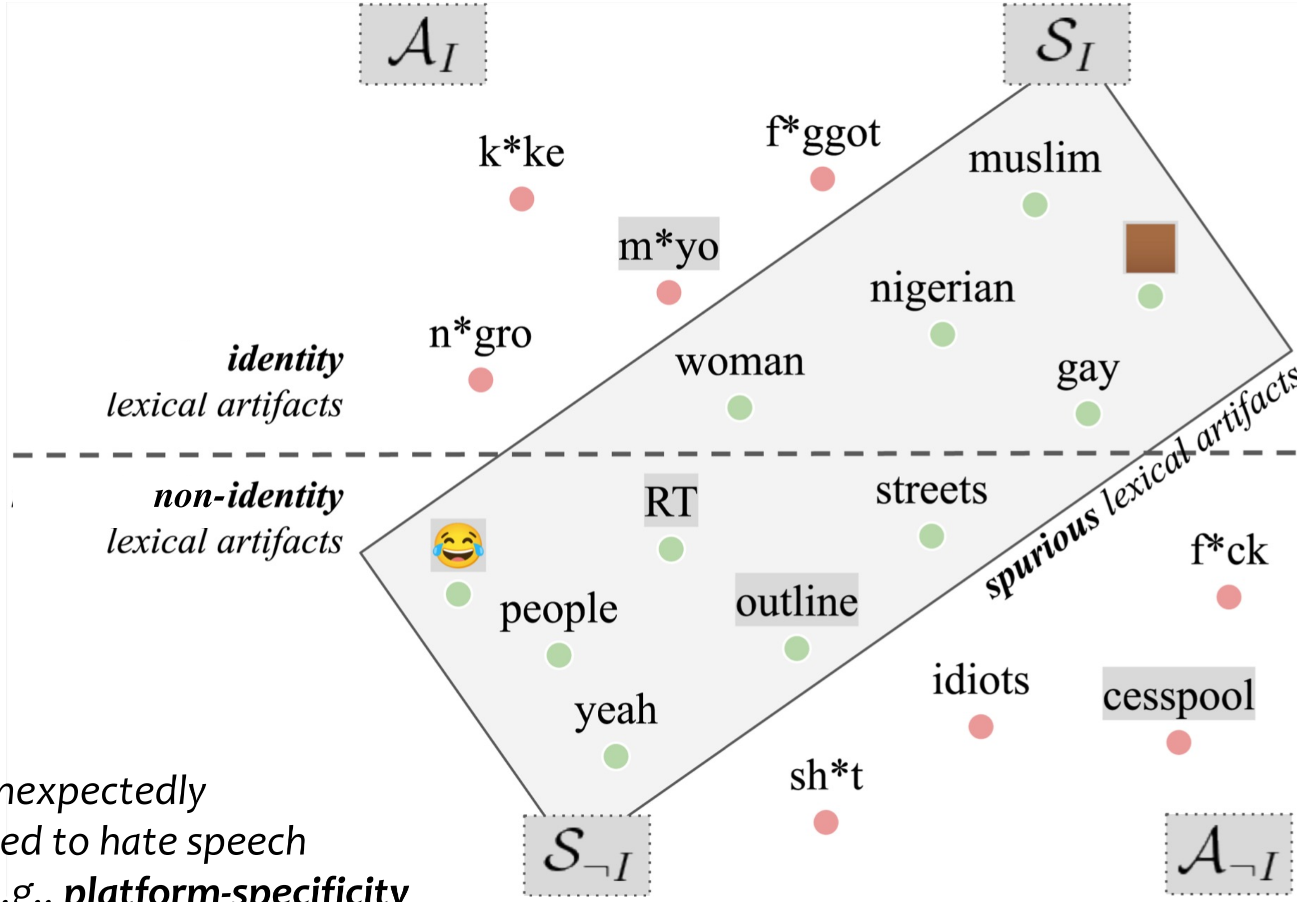
Characterization of lexical artifacts

Authentic
lexical artifacts

- identity-related * $\mathcal{A}_I$
- non identity-related * $\mathcal{A}_{\neg I}$

Spurious
lexical artifacts

- identity-related * $\mathcal{S}_I$
- non identity-related $\mathcal{S}_{\neg I}$



*terms unexpectedly associated to hate speech due to e.g., **platform-specificity***

**OI*, *OnI* and *nOI* in [Zhou+ 2021]

Datasets & unified preprocessing

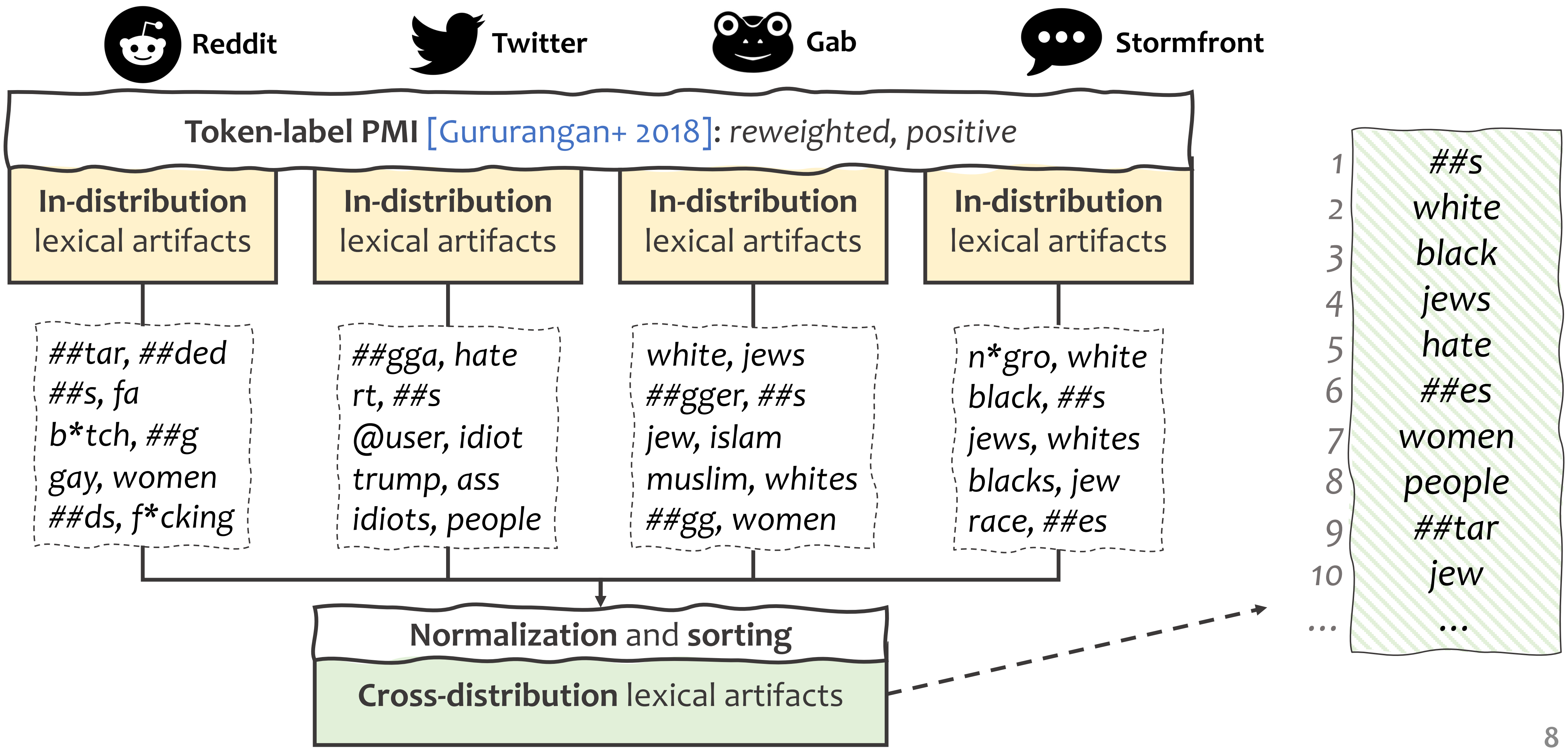
Selection criteria: *(i)* different platforms, *(ii)* minimize topic bias, *(iii)* similar annotation guidelines

Reddit	Twitter	Gab	Stormfront
[Vidgen+ 2021]	[Founta+ 2018]	[Kennedy+ 2020]	[deGibert+ 2018]

- Consistent preprocessing, cleaning, and label binarization
- **Deduplication** – many duplicates for all datasets, reliability of bias studies

WordPiece tokenization
consistent to end model's input
("##" is a subword marker)

Computation of lexical artifacts



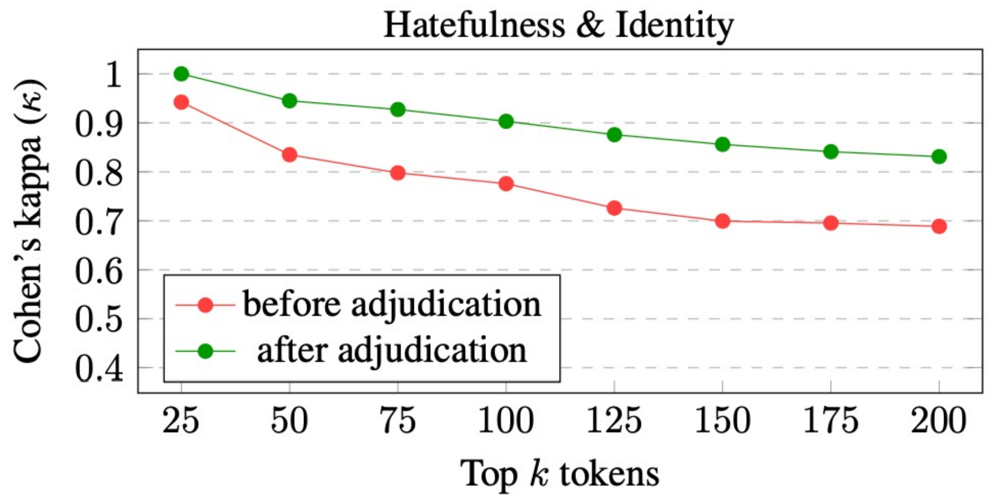
Annotation of lexical artifacts

Task: *"Is the token potentially <u>hateful</u> and/or related to <u>identities</u>?"*

- **Top-k predictive tokens from cross-distribution rank** (k=200)
- **Tokens in context** (randomly sampled posts from multiple platforms)
- **2 annotators** (M&F; fluent in English; background in NLP and linguistics)

Inter-annotator agreement

- Before adjudication: κ = **0.6887**
- After adjudication: κ = **0.8311**



! **Disagreement** correlates with **rank**

Experiments

Investigate the **impact of *spurious* lexical artifacts**

- **ID/OOD experiments:** training & testing on same/different platforms
- **Evaluation:** macro F1 (*performance*); FPR on subset w/ $\mathcal{S}_I$ (*identity bias reduction*)

Baselines and **data-centric methods**

1. **Vanilla**: BERT-base, CE loss w/ balanced class weights

2. **Filtering**: train on 33% most ambiguous instances – *Vanilla's training dynamics*
   - Promotes OOD generalization while preserving ID performance [Swayamdipta+ 2020]

Experiments *(cont'd)*

3. **Removal:** prior to fine-tuning, *<u>remove</u>* spurious lexical artifacts

    3a. **Removal**($\mathcal{S}_I$): commonly employed "fairness" baseline [Kennedy+ 2020]

    3b. **Removal**($\mathcal{S}_{\neg I}$): removal variant for non identity-related lexical artifacts

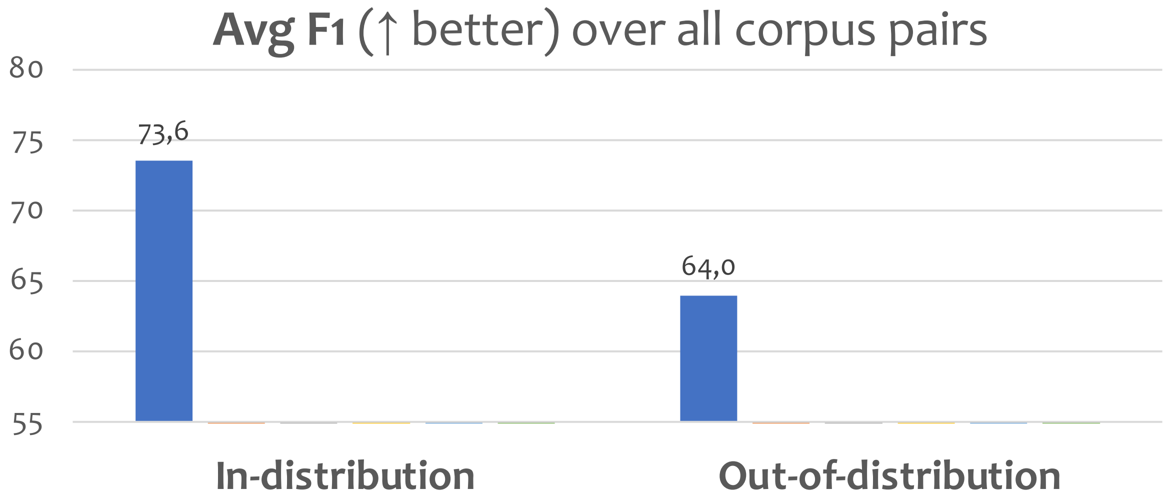
4. **Masking:** prior to fine-tuning, *<u>mask</u>* spurious lexical artifacts

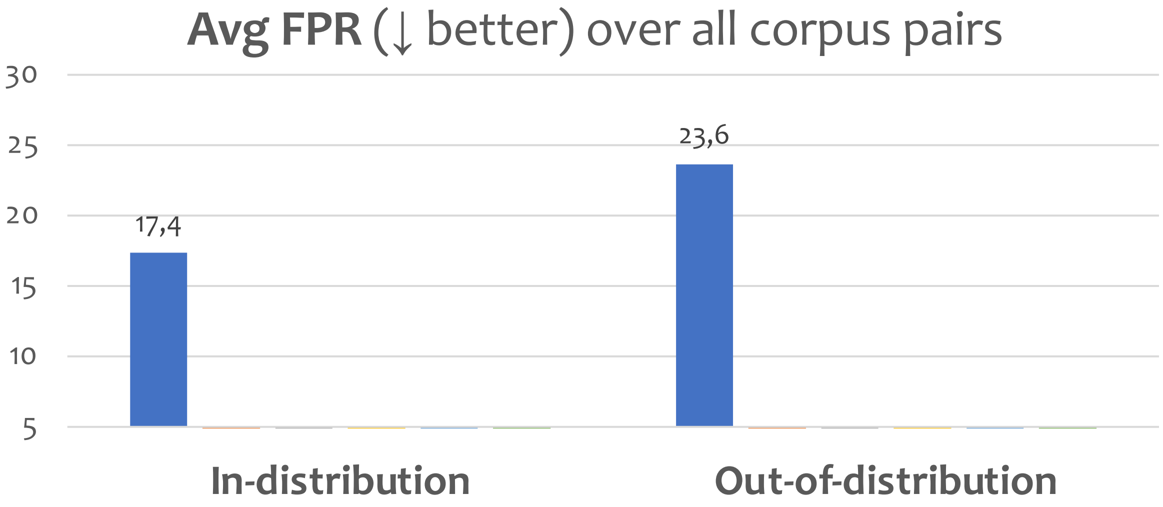
    <u>Hypothesis</u>: *encourages model to blend all lexical artifacts to a single token representation that will never appear during testing*

    4a. **Masking**($\mathcal{S}_I$): mask identity-related lexical artifacts

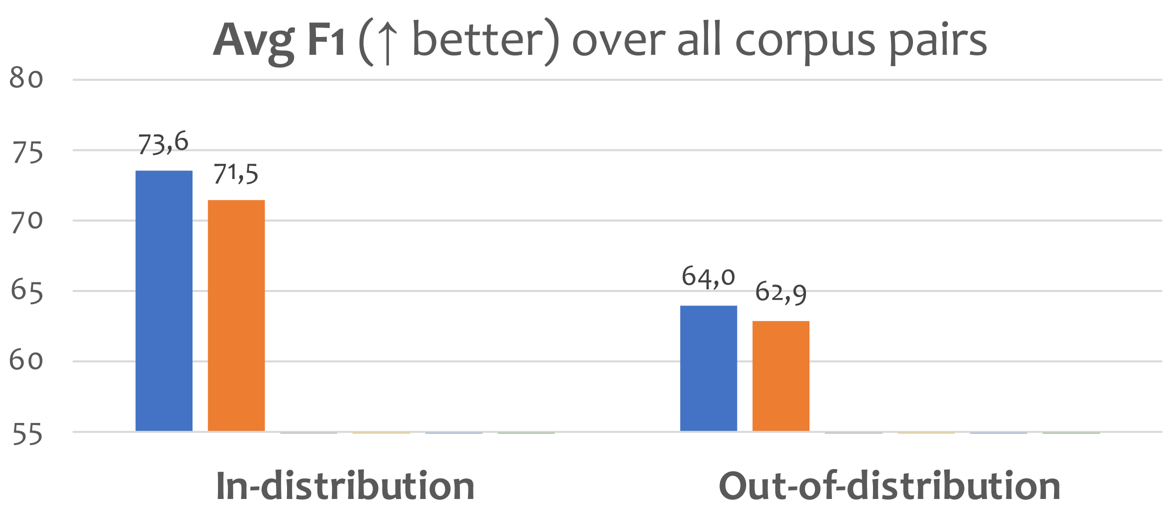
    4b. **Masking**($\mathcal{S}_{\neg I}$): mask non identity-related lexical artifacts

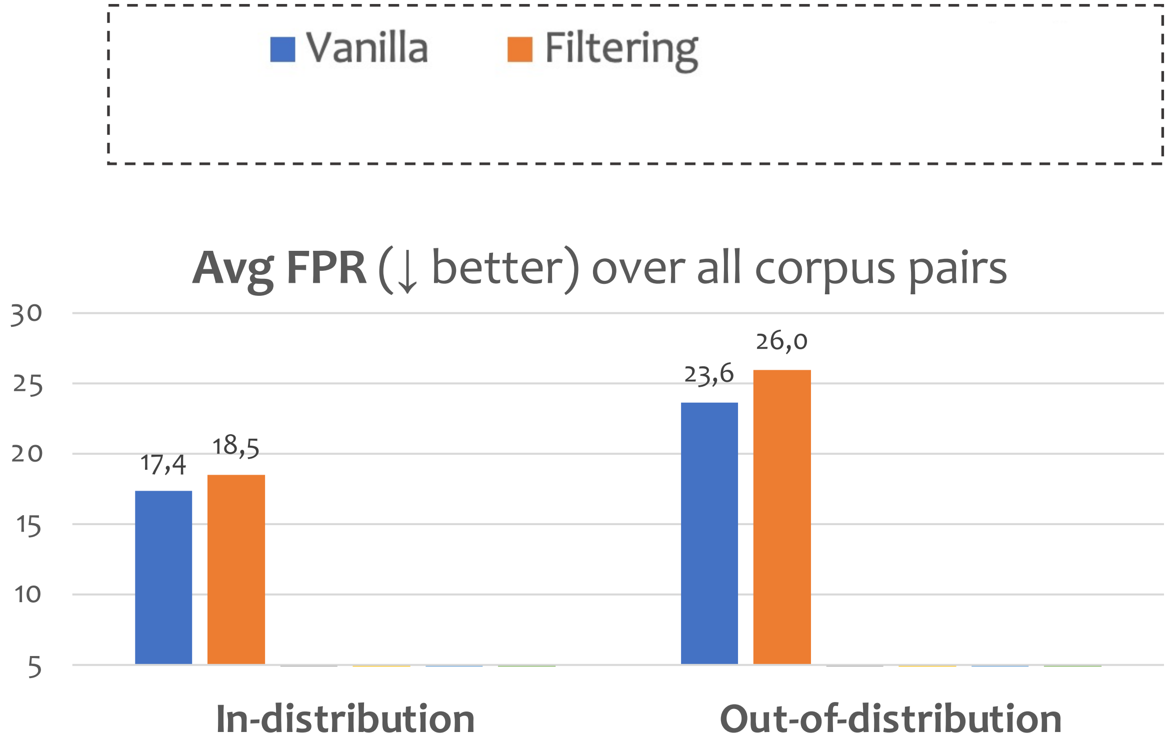
Results and discussion





Results and discussion



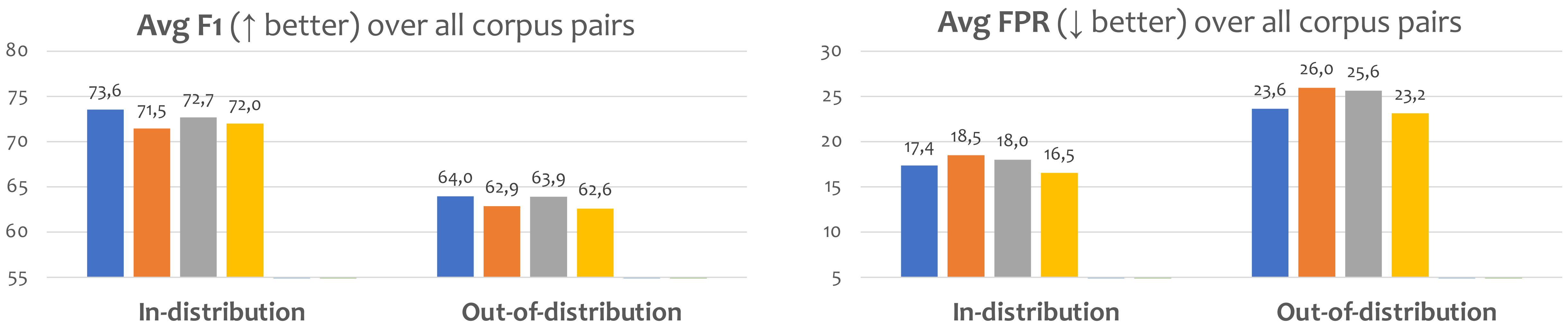


Filtering is not a *one-size-fits-all* solution

- Detrimental effect: hate speech detection requires targeted approaches
- Consistent w/ results on Twitter [Zhou+ 2021], confirmed *across* platforms

*This holds true with more/less aggressive thresholds too

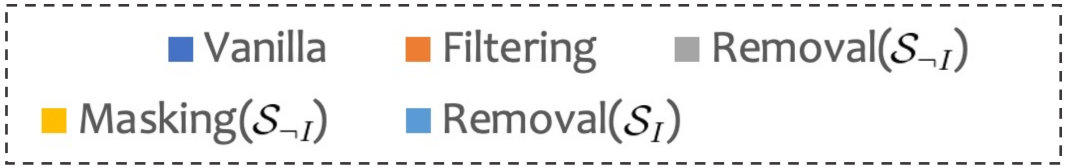
Results and discussion

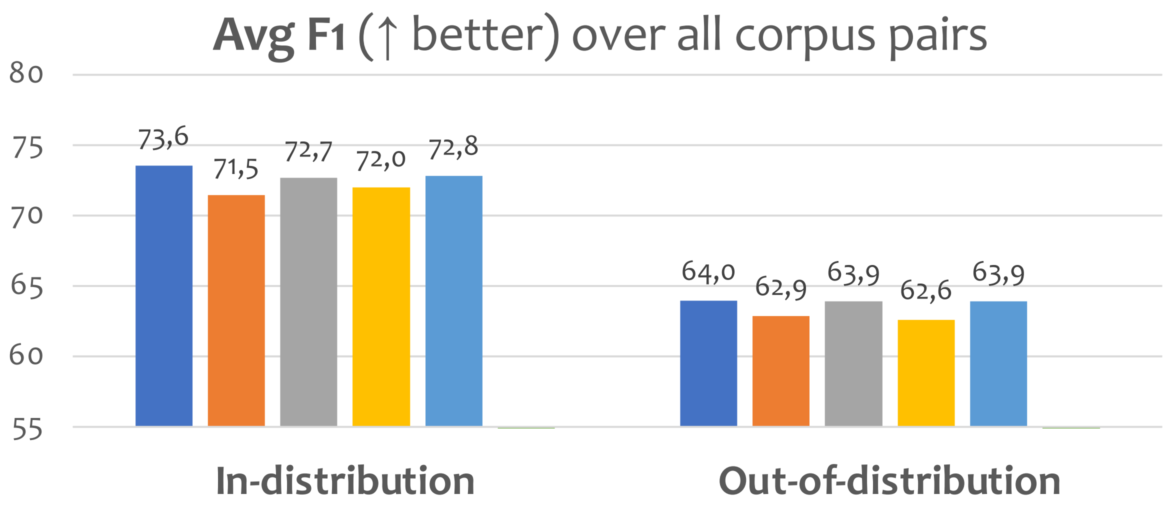


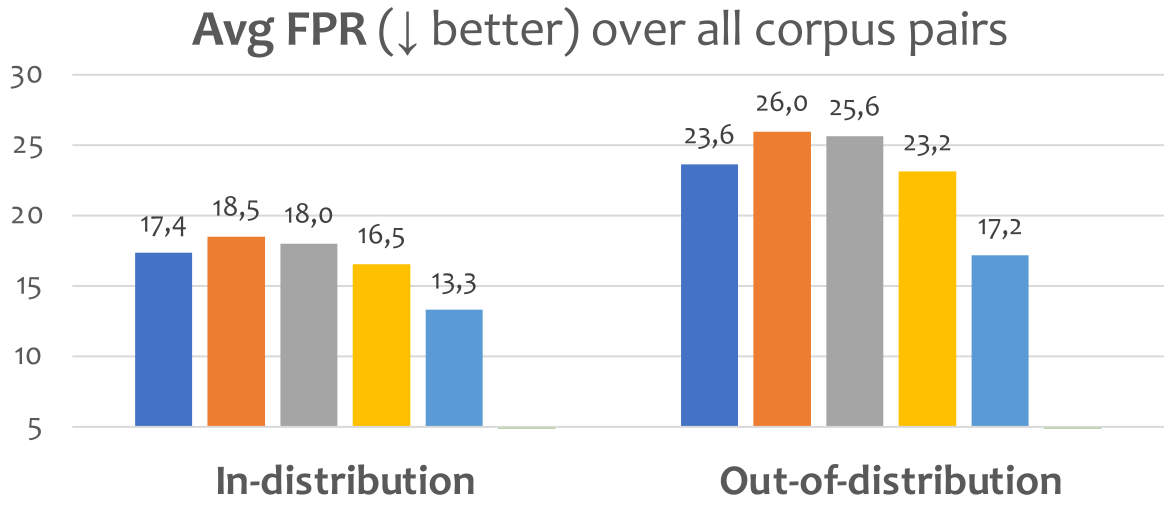
Operating on $\mathcal{S}_{\neg I}$ artifacts does not help

- Removal($\mathcal{S}_{\neg I}$) worsen ID/OOD performance and identity bias reduction
- Masking($\mathcal{S}_{\neg I}$) reduces identity bias only slightly
- Mixed results for both when looking closely at train/test pairs

Results and discussion



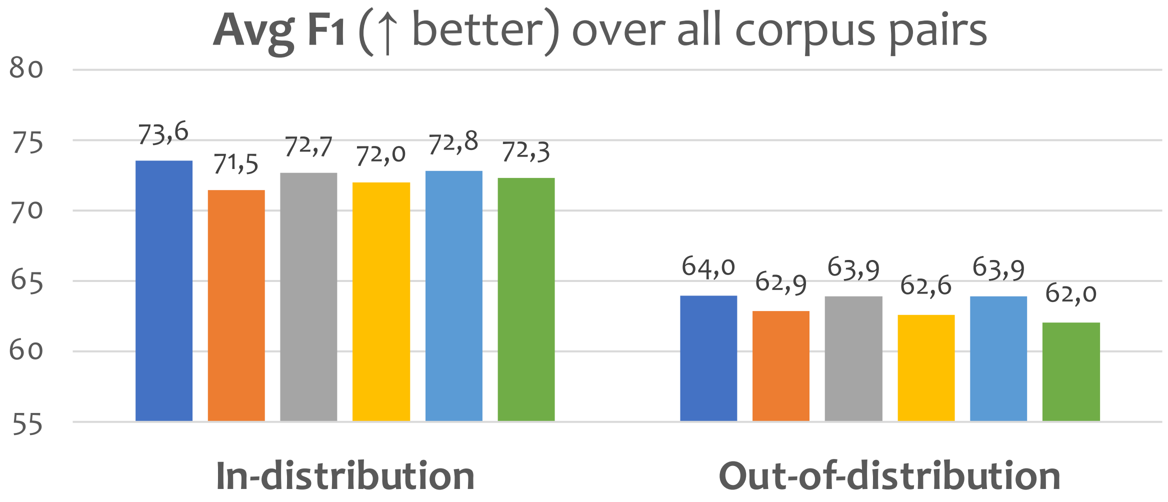


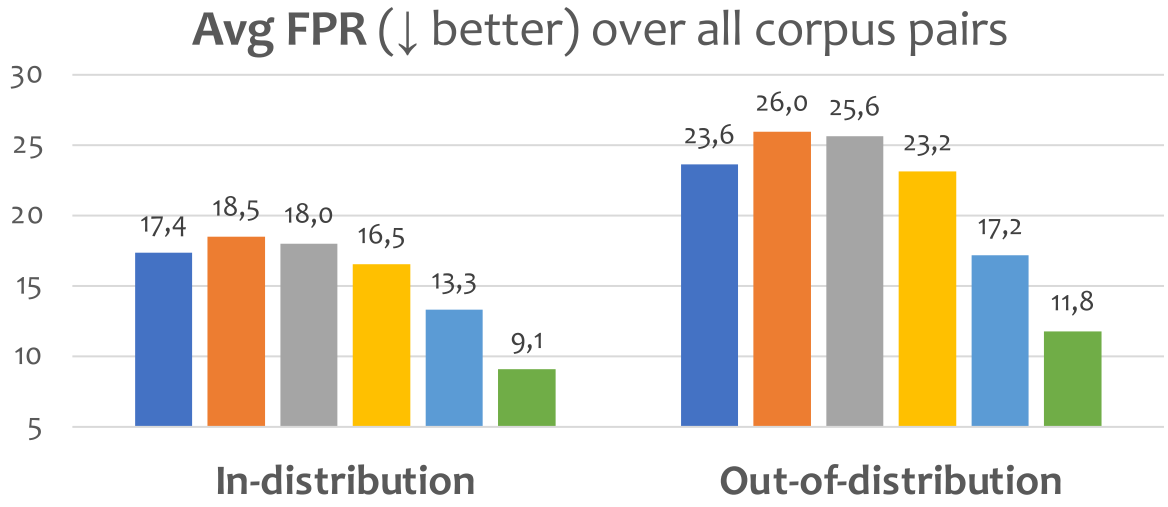


Removal($\mathcal{S}_I$) *mostly* reduces identity bias

- Not on all pairs, so not as strong as it has been previously thought
- ID/OOD performance are only slightly reduced over the Vanilla baseline

Results and discussion





Masking($\mathcal{S}_I$) *consistently* reduces identity bias

- **Large improvement** over *all* approaches, both ID/OOD, on *all* platforms
- Strong baseline for identity bias reduction in future research

F1 scores **reflect more realistically the performance** of a system that **do not rely on identity mentions** when making predictions!

Towards **artifacts documentation**

Inspired by *data statements* [Bender & Friedman 2018]

Lexical artifacts statement to document and early diagnose *lexical* biases when datasets are created/released

I. Top lexical artifacts

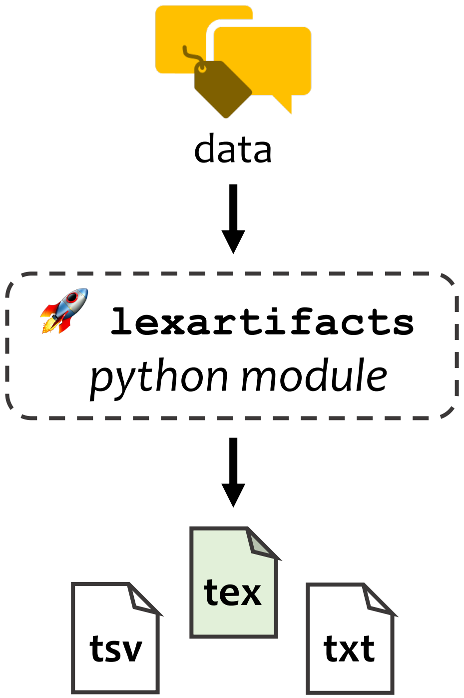
k>=10 most informative tokens to classes of interest w/ scores

II. Class definitions

Explicit definition of target class(es) for lexical artifacts

III. Methods and resources

Method (e.g., PMI), preprocessing, deduplication, and additional resources



Conclusions

- **Cross-platform study** of lexical artifacts
  - More attentive sampling is not enough: platforms do play a central role
- **Impact** of spurious lexical artifacts
  - Masking approach; robustness & identity bias are intertwined aspects
- **Documentation** is first step towards mitigation
  - Lexical artifacts statement for better understanding of lexical biases

Thank you!


Alan Ramponi **Sara Tonelli**

Fondazione Bruno Kessler, Italy

NAACL reproducibility badges

Resources

https://github.com/dhfbk/hate-speech-artifacts

- Source code and documentation
- Lexical artifacts statement template
- *Disaggregated* annotated lexical artifacts
- Fine-tuned language models
- `lexartifacts` package to ease documentation