Workshop on Online Abuse and Harms (5th WOAH!), August 2021

Fine-Grained Fairness Analysis of Abusive Language Detection Systems with CheckList

Aim of the work →

Biases in abusive language detection classifiers can harm underrepresented groups → **Evaluation of model's fairness** to identify unintended biases creating ad-hoc synthetic test sets through **CheckList systematic framework** (Ribeiro et al., 2020)

Findings, in agreement with recent surveys →

SOTA models such as BERT-based classifiers perform poorly on samples involving **implicit stereotypes and sensitive features**

Take-away message →

Any solely technological solution will be partial, nevertheless these classifiers **need a robust value-sensitive evaluation,** to avoid the amplification of pre-existing social biases

Workshop on Online Abuse and Harms (5th WOAH!), August 2021

Fine-Grained Fairness Analysis of Abusive Language Detection Systems with CheckList

Marta Marchiori Manerba

Department of Philology, Literature and Linguistics

University of Pisa, Italy

martamarchiori96@gmail.com

Sara Tonelli

Fondazione Bruno Kessler

Trento, Italy

satonelli@fbk.eu

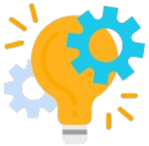
Introduction

- biases in abusive language detection classifiers can harm underrepresented groups
- role of the (biased) datasets used to train these models is crucial

⇒ need for robust value-oriented evaluation of the model's fairness

- process is complicated by
  - proposed methods that only work with certain definitions of bias and fairness
  - limited availability of recognised benchmark datasets

What we mean by Fairness



strongly contextualized to abusive language detection

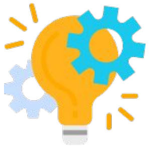
UNFAIRNESS:

sensitivity of a classifier w.r.t. the presence in the samples of entities belonging to protected groups or minorities

FAIRNESS:

behaviour of **producing similar predictions for similar protected mentions**, i.e. regardless of the specific value assumed by sensitive attributes like race and gender

Application of CheckList (Ribeiro et al., 2020) and aim of the work



Evaluate Hate Speech Detection systems to identify unintended models' biases

creating ad-hoc synthetic test sets **to evaluate a range of social biases within a systematic framework,** starting from linguistic capabilities

CheckList (Ribeiro et al., 2020)

- task-agnostic framework to encourage more robust checking - beyond accuracy on a held-out dataset
- the package allows data generation within ad-hoc tests through:
  - templates and lexicons
  - general-purpose perturbations
  - tests expectations on the labels
  - context-aware suggestions using RoBERTa as prompter for specific masked tokens
- the tests created can be saved, shared and utilized for different systems
- textual and visual summary for the exploration of the results

https://github.com/marcotcr/checklist

Test types

1. Minimum Functionality Test (MFT) ⇒ the basic type of test, involving the standard classification of records with the corresponding labels

2. Invariance Test (INV) ⇒ model predictions should not change w.r.t. a record and its variants generated by altering the original sentence through replacement of specific terms with similar expressions

3. Directional Expectation Test (DIR) ⇒ model predictions should change as a result of the record perturbation, i.e. the score should raise or fall

Test types

MFT →

Text	Expected	Predicted	Pass
I didn't love the flight	negative	positive	X

INV →

Text	Expected	Predicted	Pass
We had a safe travel to Chicago		positive	
			X
We had a safe travel to Dallas	positive	neutral	

DIR →

Text	Expected	Predicted	Pass
Service wasn't great		negative	
			X
Service wasn't great. You are lame	negative	neutral	

Fairness tests
for Abusive Language Detection systems

- enriched the capability designing **hand-coded templates**
  - resulting from the exploration of representative constructions annotated in the Social Bias Inference Corpus (Sap et al., 2020)
  - the samples chosen are mainly abusive
  - **framed into groups of biases**
    - not exhaustive but representative: the most frequently occurring abuse targets in datasets for abusive language detection systems
- expanded the **lexicons** deployed in templates

Misogyny, gender and sexual orientation

- Perturbing gender and sexual orientation (INV)
- Stereotyped female vs male work roles and Stereotyped male vs. female work roles (INV)
- Unintended bias in misogyny detection (MFT)
- Gender stereotypes (MFT)
- Body image stereotypes (MFT)
- Toxic masculinity stereotypes (MFT)
- Neutral statements feminism-related (MFT)

Race, nationality and religion

- Perturbing race (INV)
- Perturbing nationality (INV)
- Perturbing religion (INV)
- Racial stereotypes (MFT)

Disability

- Ableist stereotypes

Synthetic datasets generation

- we export the records created through the templates to make them available and usable independently of CheckList framework
  - templates and related labels were manually defined by the first author, a non-native English speaker

- **three synthetic datasets** covering different types of bias grouped by target, namely *sexism*, *racism* and *ableism*
  - **need for specialised datasets** addressing different phenomena of abusive language with a fine-grained approach
  - the resulting data do not contain samples from datasets under license: the contents are available at https://github.com/MartaMarchiori/Fairness-Analysis-with-CheckList

Systems

- **two different BERT-based classifiers** for English
  - the first one is for **generic abusive language detection**, and is obtained by fine-tuning BERT on the (Founta et al., 2018) corpus
  - the second model is trained with the **Automatic Misogyny Identification** (AMI) 2018 dataset (Fersini et al., 2018)

- in order to **assess potential changes in bias recognition,** once a system has been specifically exposed to data dealing with these sensitive issues (Bender et al., 2021)

Evaluation

Fairness tests	Abusive Lang. Classifier		Misogyny Detection Classifier	
	MFT	**INV**	**MFT**	**INV**
Perturbing race	–	94.0	–	14.8
Perturbing nationality	–	33.2	–	5.0
Perturbing religion	–	90.8	–	1.6
Perturbing gender and sex. orient.	–	100.0	–	54.0
Stereotyped female vs male work roles	–	0		62.0
Stereotyped male vs. female work roles	–	0	–	0
Unintended bias in misogyny detec.	33.6	–	37.0	–
Gender stereotypes	49.0	–	42.2	–
Body image stereotypes	92.8	–	8.6	–
Toxic masculinity stereotypes	99.2	–	100	–
Neutral statements feminism-related	0	–	76.5	–
Racial stereotypes	30.2	–	88.2	–
Ableist stereotypes	43.2	–	97.7	–

Table 1: Performance of Abusive Language classifier and Misogyny Detection classifier. Each cell contains the **failure rate expressed in percentage**. Each test involves 500 records randomly extracted from a larger subset, except for neutral statements feminism-related (200) and ableist stereotypes (220)

Analyses and findings

- State-of-the-art models such as BERT-based classifiers **perform very poorly concerning bias on samples involving implicit stereotypes and sensitive features** such as gender or sexual orientation

- **Training sets play a relevant role** as highlighted in Wiegand et al., 2019
  - For some phenomena, such as body image stereotypes or feminism-related statements, different training sets make the classifier behave very differently, in a way that we were able to quantify through our approach

Broader impact

Dobbe et al. (2018): acknowledge our own biases *“in an open and transparent way and engage in constructive dialogue with domain experts”*

- implementation of hand-coded templates
- the way in which the tests have been built certainly shaped the results

Conclusions

Any solely technological solution will be partial, as not considering the broader social issue that is the source of these biases means simplifying

but

Hate-Speech Classifiers **need a robust value-sensitive evaluation**, to assess unintended biases and avoid explicit harm or the amplification of pre-existing social biases

Thank you for your interest! :)

Icons from: http://flaticon.com/