TECHNISCHE UNIVERSITÄT WIEN

DATA SCIENCE INSIGHTS FOR SOCIETY

Universität Zürich[UZH]

universität wien

AustroTox

A Dataset for Target-Based Austrian German Offensive Language Detection

Pia Pachinger, Janis Goldzycher, Anna Maria Planitzer,
Wojciech Kusa, Allan Hanbury, Julia Neidhardt

paper, data, and contact →

www.pia.wien

Offensive / Toxic
Includes derogatory remarks or incites hatred or violence

Motivation

Offensive

"Bei Vielen ist der Schädel gut mit Gehirn gefüllt... Nur der BIMAZ, der hat noch viel Platz"

target: individual

best secretary of the interior of all times

🇬🇧 *"Many people's skulls are well filled with brains... Only the BIMAZ still has plenty of room"*

→ Need for country-specific toxicity detection

Offensive

"27-year-old [Nationality]. Stopped reading there"

target: group

→ Need for target-aware toxicity detection

Not Offensive

*"f*ck. das fliegt uns jetzt um die ohren"*

vulgarity

🇬🇧 *"F*ck. This is going to blow up in our faces."*

→ Need for vulgarity-aware toxicity detection

Offensive

*"Geh bitte HOITS Z*M!!"*

vulgarity

🇬🇧 *"Please, sh*t up!"*

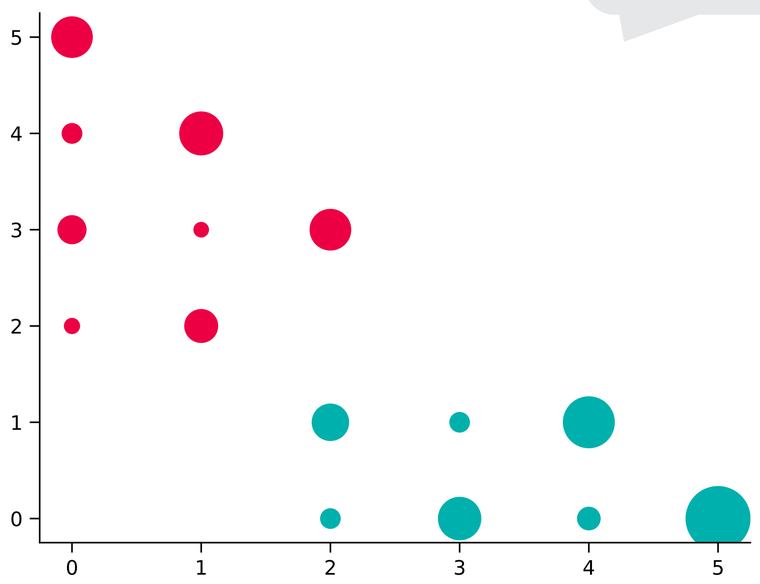
→ Need for toxicity detection aware of country-specific vulgarities

Data source

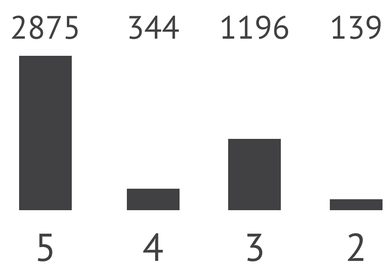
derStandard.at

newspaper online forum

A sensitive approach to data aggregation



Annotators per comment



Dataset Creation

Annotation strategy

Size of AustroTox

Experiments Using AustroTox

F1 scores

		# Params	Offensive	Vulgarity	Target
fine-tuned	BERT en	110M	.64	**.69**	.16
	BERT de	110M	.69	.68	**.22**
	BERT db	110M	.71	**.69**	**.23**
	Gelectra	110M	.50	**.69**	**.24**
	GBERT	110M	.72	**.69**	**.23**
	Gelectra	335M	.52	.56	.19
	GBERT	337M	.73	**.71**	.21
	# Shots				
prompted	LeoLM 0	7B	.61	-	-
	5		.52		
	Mistral 0	7.24B	.30	-	-
	5		.55		
	Llama 3 0	8B	.67	-	-
	5		.59		
	GPT 3.5 0	-	.68	.40	.17
	5		.72	.43	.20
	GPT 4 0	-	.70	.36	.20
	5		**.76**	.41	**.22**

System prompt for generative models

You receive a user comment from a newspaper article. Your task is to answer the followig questions about the comment:

1. Is the comment toxic? (1 = toxic, 0 = non-toxic)
Definition: ...

2. Who or what is the target of toxicity?...

3. Mark vulgarities with "Vulgarity"...

Respond in JSON format with the following fields:
...