AustroTox

A Dataset for Target-Based Austrian German Offensive Language Detection

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

Findings of ACL 2024

paper, data, and contact → www.pia.wien

Offensive / Toxic: *derogatory remarks, or incites to hate or violence*

Motivation

Offensive

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

best secretary of the interior of all times → target: individual

🇬🇧 *"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

Dataset Creation

Data sources

🇦🇹 derStandard.at
newspaper online forum

🇬🇧 *Jigsaw Toxic Comment Classification Challenge*

Annotation strategy

"@!.?? x..@%@ //§.@" vulgarity

article title: **Europacup-Nachlese: Rapid verliert in Zagreb 1:3, Sturm punktet bei Sociedad, LASK fix weiter**

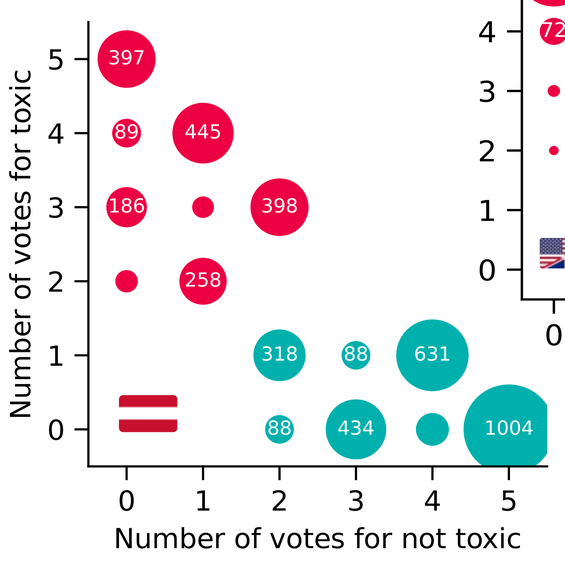
Offensive: *"@!.?? x..@%@ //§.@"* vulgarity

Not Offensive: *"@!.?? x..@%@ //§.@"* vulgarity

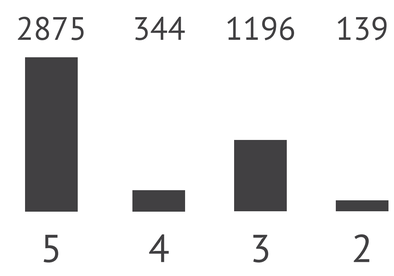
Offensive: *"@!.?? x..@%@ //§.@"* target: group vulgarity

target: group / target: individual / target: other

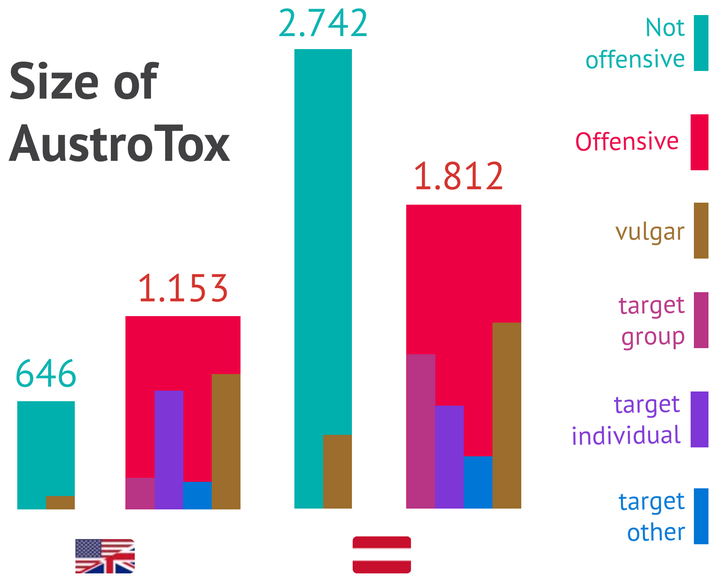
A sensitive approach to data aggregation



Annotators per comment



Size of AustroTox



Experiments Using AustroTox

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:
...

F1 scores

		# Shots	# Params	Offensive 🇦🇹	Offensive 🇬🇧	Vulgarity 🇦🇹	Vulgarity 🇬🇧	Target 🇦🇹	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	-
	Electra		335M	-	**.88**	-	.64	-	.08
	Roberta		335M	-	**.90**	-	**.77**	-	**.27**
not fine-tuned	LeoLM	0	7B	.61	-	-	-	-	-
		5		.52					
	Mistral	0	7.24B	.30	-	-	-	-	-
		5		.55					
	Llama 3	0	8B	.67	.78	-	-	-	-
		5		.59	.82				
	GPT 3.5	0	-	.68	**.89**	.40	.46	.17	.16
		5		.72	**.89**	.43	.47	.20	.18
	GPT 4	0	-	.70	.87	.36	.41	.20	.15
		5		**.76**	**.89**	.41	.43	**.22**	.18