From the Detection of Toxic Spans in Online Discussions to the Analysis of Toxic-to-Civil Transfer

John Pavlopoulos [1,2] Léo Laugier [3] Alexandros Xenos [2] Jeffrey Sorensen [4] Ion Androutsopoulos [2]

[1]Stockholm University [2]Athens University of Economics & Business [3]Télécom Paris, Institut Polytechnique de Paris [4]Google

Motivation: Assist human moderation of online discussions

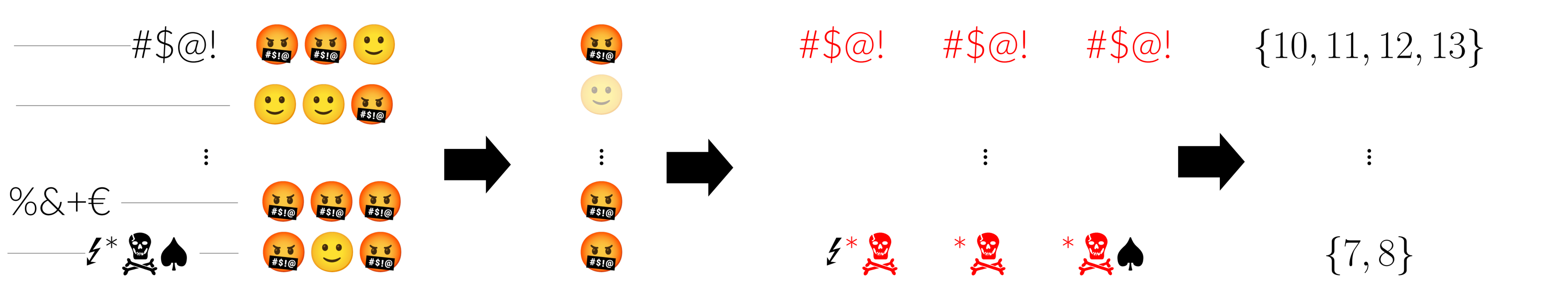
In social media and online fora, **toxic content** can be defined as rude, disrespectful, or unreasonable posts that would make users want to leave the conversation. Although several toxicity detection datasets and models exist, **most of them classify whole posts**, without identifying the specific **spans that make a text toxic**. But highlighting such toxic spans can assist human moderators who often deal with lengthy comments, and who prefer attribution instead of a system-generated unexplained toxicity score per post. **Locating toxic spans** within a text is thus a major step towards successful semi-automated moderation and healthier online discussions.

"Survival of the fittest would not have produced you. You are alive because your weak blood is supported by welfare and food stamps. Please don't reference Darwin in your icon. Loser".

*"Survival of the fittest would not have produced you. You are alive because your **weak blood** is supported by welfare and food stamps. Please don't reference Darwin in your icon. **Loser**".*

A new dataset of toxic posts from the Civil Comments [1] dataset annotated at the span level.

Civil Comments previously labeled by multiple annotators	Posts with majority toxic annotation	Crowsourcing annotation (Appen) on a random subset; 3 annotators per post	ToxicSpans with ground-truth made of the majority offset characters labeled as toxic
$1.2M$	$30K$	$11K$	$11K$

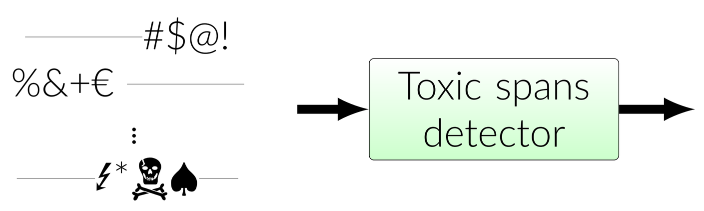


Evaluation with an appropriate $\bar{F}_1$ score

Ground truth: n posts, each associated with a set Y_i of character offsets.
Prediction: System returning a set of character offsets $\hat{Y}_i$ for the i^{th} post.

$$\bar{F}_1 = \frac{1}{n}\sum_{i=1}^{n} F_1^i \quad \text{with per-post } F_1 \text{ score: } F_1^i = \frac{2 \cdot P^i \cdot R^i}{P^i + R^i}$$

$$\text{Precision: } P^i = \frac{|\hat{Y}_i \cap Y_i|}{|\hat{Y}_i|} \qquad \text{Recall: } R^i = \frac{|\hat{Y}_i \cap Y_i|}{|Y_i|}$$



$\hat{Y}_i$	Y_i	P^i	R^i	F_1^i
{10, 11, 12, 13}	{10, 11, 12, 13}	1	1	1
{0, 1, 2}	{0, 1, 2, 3}	1	0.75	0.86
⋮	⋮	⋮	⋮	⋮
{6, 7, 8, 9}	{7, 8}	0.5	1	0.67

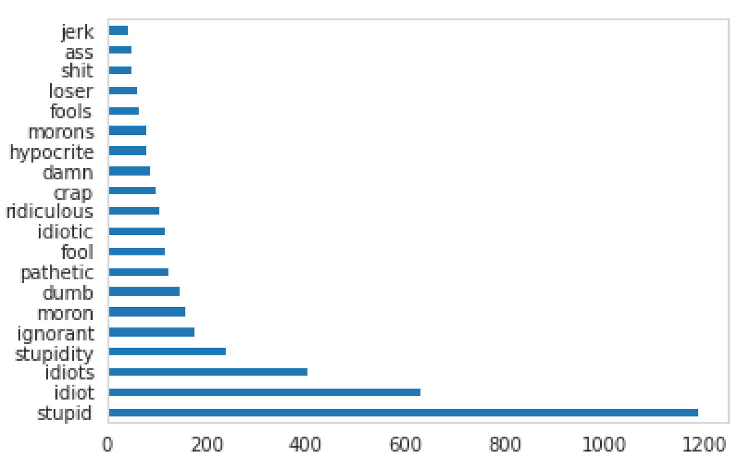
ToxicSpans analysis

Inter-annotator agreement: computed with Cohen's κ
87 randomly selected posts, labeled by 5 (instead of 3) workers: $\boldsymbol{\kappa = 0.48}$
↳ On posts (51) found toxic by *a majority* of annotators: $\kappa = 0.55$
↳ On posts (31) found toxic by *all* annotators: $\kappa = 0.65$

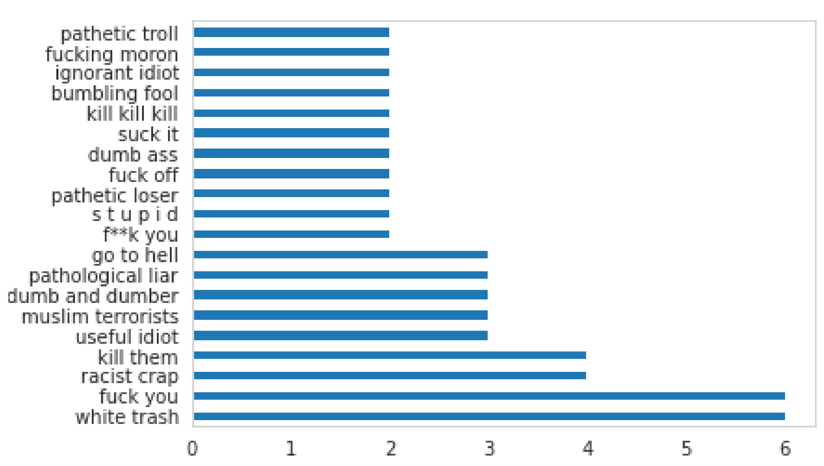
Moderate agreement ⟶ Highly **subjective task**

Exploratory analysis
$5K/11K$ posts have empty ground truth toxic span ⟶ Toxicity does not imply it is "*localized*"
Most posts with toxic spans include a **single** "*dense span*".

Most frequent toxic spans



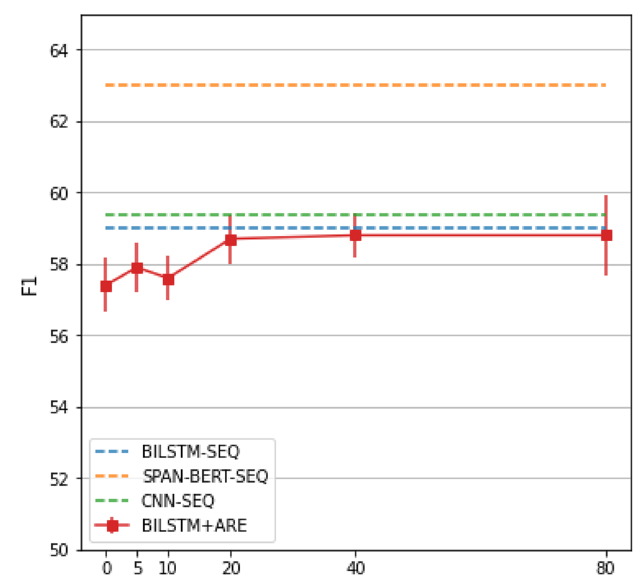
Most frequent multi-word toxic spans

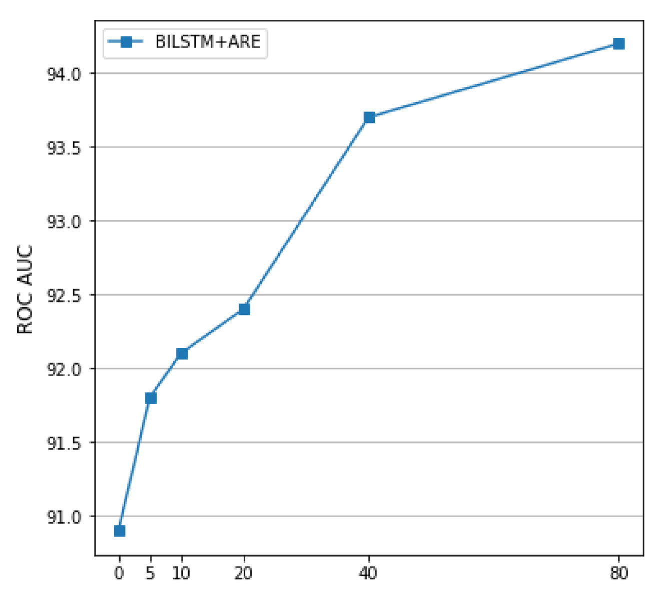


ToxicSpans Systems

		F_1 (%)	P (%)	R (%)	(%)
Baselines	rand	7.3	5.3	25.4	N/A
	train-match	41.0	39.1	48.7	N/A
	hate-match	10.6	7.1	43.7	N/A
Strong supervision	bilstm-seq	58.9	59.8	58.9	N/A
	cnn-seq	59.3	60.7	59.0	N/A
	bert-seq	59.7	60.7	60.0	N/A
	span-bert-seq	**63.0**	**63.8**	**62.8**	N/A
Weak supervision	bilstm+are	57.7	58.4	57.3	90.9
	bert+are	49.1	49.4	49.5	**96.1**

Additional training data for **weakly supervised** (attention-based rationale extraction) systems



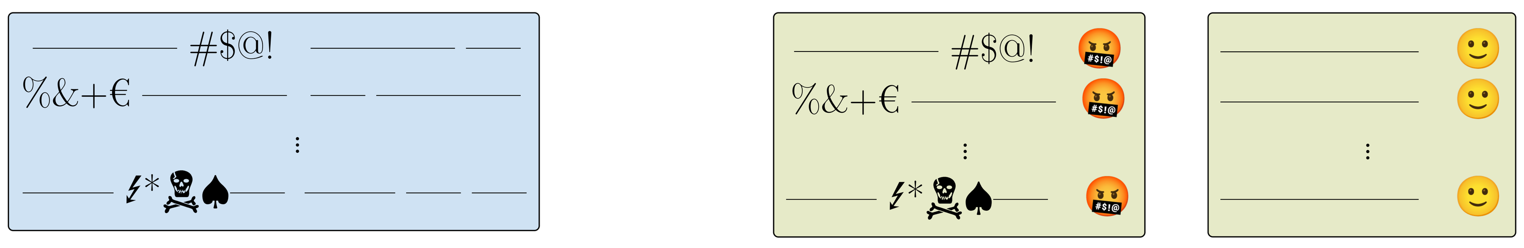


Analysis of Toxic-to-Civil transfer 🤬 → 🙂

Systems fine-tuning a pre-trained transformer

Strongly **S**upervised **E**ncoder **D**ecoder T5 (**SED-T5**) trained with a **parallel** (**P**) dataset made of $\sim 2K$ manually produced toxic-to-civil pairs [2]

Self-supervised **C**onditional **A**uto **E**ncoder **T5** (**CAE-T5**) [3] trained with a **non-parallel** (**NP**) dataset made of respectively $\sim 0.1M$ and $\sim 6M$ unpaired toxic and civil posts [1]



Toxic-to-Civil Transfer scrutinized with ToxicSpan dataset and systems

Evaluation Dataset	Metric	CAE-T5	SED-T5
Non-Parallel (NP)	ACC ↑	**75.0%**	52.2%
	ACC2 ↑	**83.4%**	67.3%
	PPL ↓	**5.2**	11.8
	self-SIM ↑	70.0%	**87.9%**
	GM (self) ↑	**0.466**	0.338
	ACC3 ↑	**86.7%**	64.1%
	ACC4 ↑	**83.2%**	59.5%
Parallel (P)	ACC ↑	94.3%	94.3%
	ACC2 ↑	**94.7%**	94.3%
	PPL ↓	**9.1**	38.3
	ref-SIM ↑	27.6%	**65.3%**
	self-SIM ↑	32.6%	**65.6%**
	GM (ref) ↑	**0.306**	0.252
	GM (self) ↑	**0.323**	0.252
	ACC3 ↑	**98.8%**	94.3%
	ACC4 ↑	**94.7%**	91.9%
ToxicSpans	ACC ↑	**92.9%**	65.6%
	ACC2 ↑	**92.5%**	63.7%
	PPL ↓	**7.2**	24.9
	self-SIM ↑	34.5%	**82.1%**
	GM (self) ↑	**0.355**	0.279
	ACC3 ↑	**96.9%**	62.0%
	ACC4 ↑	**92.0%**	54.7%

Can ToxicSpans data and toxic span detectors be used to **assess the mitigation** of explicit toxicity in Toxic-to-Civil transfer?

↳ Evaluation of toxic spans transfer in **system-detoxified** posts
↳ Study of remaining toxic spans in **human-detoxified** posts

Takeaways:
↳ The models often successfully **detect toxic spans** and try to **rephrase** them
↳ **Humans did rephrase** almost all cases of explicit toxicity in the toxic posts they were given

References

[1] Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. Nuanced metrics for measuring unintended bias with real data for text classification. In *WWW*, pages 491–500, San Francisco, USA, 2019.

[2] Daryna Dementieva, Sergey Ustyantsev, David Dale, Olga Kozlova, Nikita Semenov, Alexander Panchenko, and Varvara Logacheva. Crowdsourcing of parallel corpora: the case of style transfer for detoxification. In *Proceedings of the 2nd Crowd Science Workshop: Trust, Ethics, and Excellence in Crowdsourced Data Management at Scale co-located with 47th International Conference on Very Large Data Bases (VLDB 2021 (https://vldb.org/2021/))*, pages 35–49, Copenhagen, Denmark, 2021. CEUR Workshop Proceedings.

[3] Léo Laugier, John Pavlopoulos, Jeffrey Sorensen, and Lucas Dixon. Civil rephrases of toxic texts with self-supervised transformers. In *EACL*, pages 1442–1461, Online, 2021.

annis@aueb.gr