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 and Business
[3]Télécom Paris, Institut Polytechnique de Paris
[4]Google

ACL 2022

Contents

Contents

Introduction (1/2): Assist human moderation of 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"

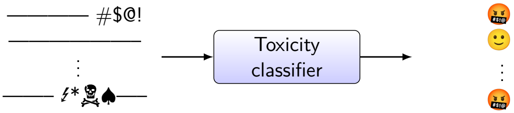
Introduction (1/2): Assist human moderation of online discussions

— "..."

— "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**"

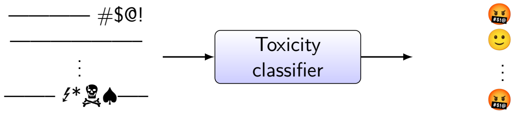
Introduction (2/2): Approaches to semi-automated moderation and healthier online discussions

Classification: Existing; leveraged here

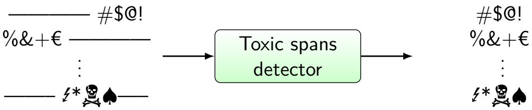


Introduction (2/2): Approaches to semi-automated moderation and healthier online discussions

Classification: Existing; leveraged here

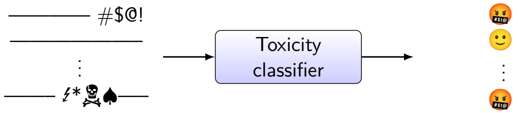


Toxic Span: Introduced here

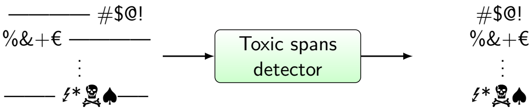


Introduction (2/2): Approaches to semi-automated moderation and healthier online discussions

Classification: Existing; leveraged here



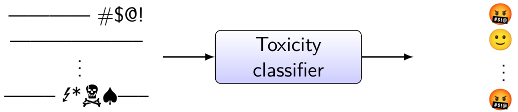
Toxic Span: Introduced here



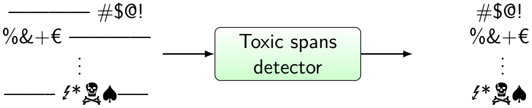
Text transfer: Existing; studied here

Introduction (2/2): Approaches to semi-automated moderation and healthier online discussions

Classification: Existing; leveraged here



Toxic Span: Introduced here



Text transfer: Existing; studied here

Contents

ToxicSpans task (1/3): Dataset annotation

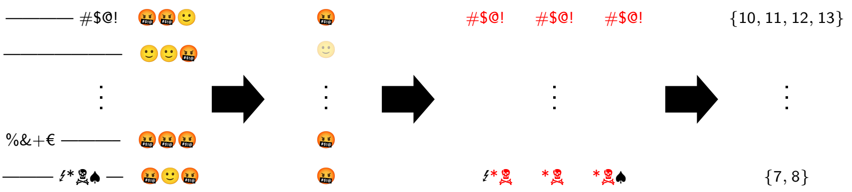
Civil Comments previously labeled by multiple annotators

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



ToxicSpans task (2/3): Analysis

Inter-annotator agreement: computed with Cohen's κ

87 randomly selected posts, labeled by 5 (instead of 3) workers: $\kappa = 0.48$

↳ 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 $\longrightarrow$ Highly **subjective task**

ToxicSpans task (2/3): Analysis

Inter-annotator agreement: computed with Cohen's κ

87 randomly selected posts, labeled by 5 (instead of 3) workers: $\kappa = 0.48$
↳ 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 $\longrightarrow$ Highly **subjective task**

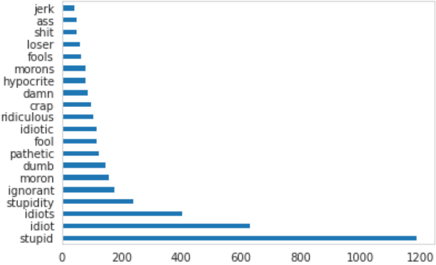
Exploratory analysis

$5K/11K$ posts have empty ground truth toxic span.
$\longrightarrow$ Toxicity does not imply it is "*localized*"

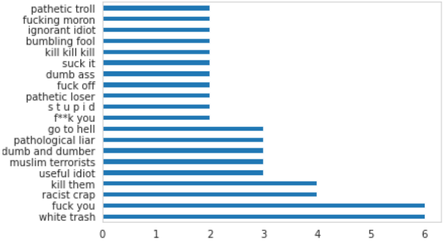
Most posts with toxic spans include a **single** "*dense span*".

❗ Next slide shows explicit language

ToxicSpans task (2/3): Analysis



(a) Most frequent toxic spans



(b) Most frequent multi-word toxic spans

ToxicSpans task (3/3): Evaluation

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.

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

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

X_i → Toxic spans detector →

X_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
⋮	⋮	⋮	⋮	⋮	⋮
—— ϟ*☠♠——	$\{6, 7, 8, 9\}$	$\{7, 8\}$	0.5	1	0.67

Contents

Methods (1/2): Systems

Baselines

Random: RAND

Naive: Lookup methods

- HATE-MATCH from a pre-defined hateful vocabulary [1].
- TRAIN-MATCH from the TOXICSPANS train set.

Methods (1/2): Systems

Baselines

Random: RAND

Naive: Lookup methods

- HATE-MATCH from a pre-defined hateful vocabulary [1].
- TRAIN-MATCH from the TOXICSPANS train set.

Strong supervision: Standard deep learning architectures

RNN: BILSTM-SEQ

CNN: CNN-SEQ

BERT: BERT-SEQ and SPAN-BERT-SEQ [2]

Methods (1/2): Systems

Baselines

Random: RAND

Naive: Lookup methods

- HATE-MATCH from a pre-defined hateful vocabulary [1].
- TRAIN-MATCH from the TOXICSPANS train set.

Strong supervision: Standard deep learning architectures

RNN: BILSTM-SEQ

CNN: CNN-SEQ

BERT: BERT-SEQ and SPAN-BERT-SEQ [2]

Weak (inexact) **supervision**: **A**ttention-based **R**ationale **E**xtraction

RNN: BILSTM+ARE [3]

BERT: BERT+ARE

Methods (2/2): Weakly-supervised systems

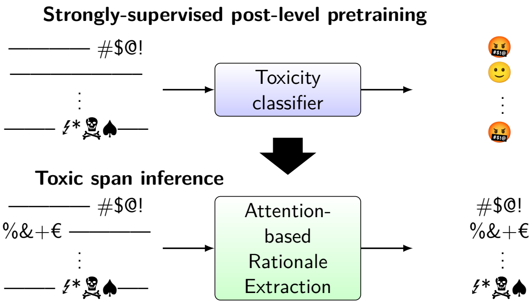
Weak (inexact) **supervision**: **A**ttention-based **R**ationale **E**xtraction

RNN: BILSTM+ARE [3]
BERT: BERT+ARE

Methods (2/2): Weakly-supervised systems

Weak (inexact) **supervision**: **A**ttention-based **R**ationale **E**xtraction

RNN: BILSTM+ARE [3]
BERT: BERT+ARE



Contents

Results (1/3): Quantitative analysis

		F_1 (%)	P (%)	R (%)	ROC AUC[1] (%)
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

[1] of the post-level toxic classifier

Results (2/3): Error analysis

Type I error (**False positives**)

• Not sure if "people are **dumb**" is the best descriptor, but you are correct that we tend to seek out and grasp at anything that supports our beliefs and hopes. Hence the proliferation of "fake news", which feeds those wants.

• They can shuffle the cabinet seven ways from Sunday and it's still a cabal of **losers**.

Results (2/3): Error analysis

Type I error (**False positives**)

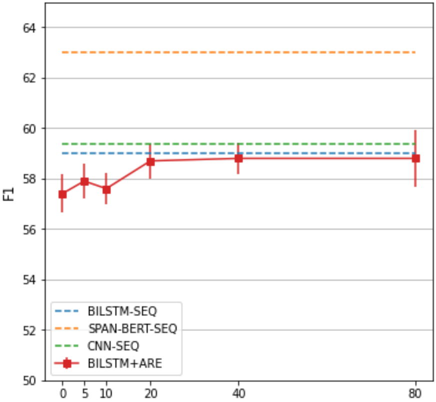
• Not sure if "people are **dumb**" is the best descriptor, but you are correct that we tend to seek out and grasp at anything that supports our beliefs and hopes. Hence the proliferation of "fake news", which feeds those wants.

• They can shuffle the cabinet seven ways from Sunday and it's still a cabal of **losers**.

Type II error (**False negatives**)

• You can stick your d**k up anyone's butt. Why have any laws at all?

Results (3/3): **Additional training data** for weakly supervised (attention-based rationale extraction) systems



Increasing the train size of underlying post-level classifiers improves the toxic-span detectors, almost reaching the performance of strongly-supervised systems.

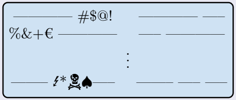
Contents

Toxic-to-Civil Transfer (1/2): Transformer-based systems

Toxic-to-Civil Transfer (1/2): Transformer-based systems

Strongly **S**upervised **E**ncoder **D**ecoder **T5** (SED-T5)

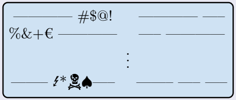
Parallel (P) dataset made of $\sim 2K$ manually produced toxic-to-civil pairs [4]



Toxic-to-Civil Transfer (1/2): Transformer-based systems

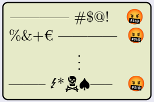
Strongly **S**upervised **E**ncoder **D**ecoder **T5** (SED-T5)

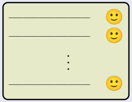
Parallel (P) dataset made of $\sim 2K$ manually produced toxic-to-civil pairs [4]



Self-supervised **C**onditional **A**uto **E**ncoder **T5** (CAE-T5) [5]

Non-parallel (NP) dataset made of respectively $\sim 0.1M$ and $\sim 6M$ unpaired toxic and civil posts [6]





Toxic-to-Civil Transfer (2/2): 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%

- 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

Contents

Conclusion

- ToxicSpan introduces the first large-scale dataset annotated at the span level.
- SPAN-BERT-SEQ achieves best results on this new task.
- Weak supervision + data augmentation catches up with some strongly-supervised span detectors.
- Part of the TOXICSPAN dataset has been used in the SemEval-2021 Task 5.
- ToxicSpan helps to evaluate automatic and human toxic-to-civil transfer.

Future work

- Remove the toxicity assumption by adding a component detecting whether a post is toxic or not
- Leverage weak supervision and apply TOXICSPAN detection in low-resource languages

References I

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee.
Hatexplain: A benchmark dataset for explainable hate speech detection.
In *AAAI*, pages 14867–14875, 2021.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy.
SpanBERT: Improving pre-training by representing and predicting spans.
TACL, 8:64–77, 2020.

References II

John Pavlopoulos, Prodromos Malakasiotis, and Ion Androutsopoulos.
Deep learning for user comment moderation.
In *Proceedings of the First Workshop on Abusive Language Online*, pages 25–35, Vancouver, BC, Canada, 2017. Association for Computational Linguistics.

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.

References III

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.

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.

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 and Business
[3]Télécom Paris, Institut Polytechnique de Paris
[4]Google

ACL 2022