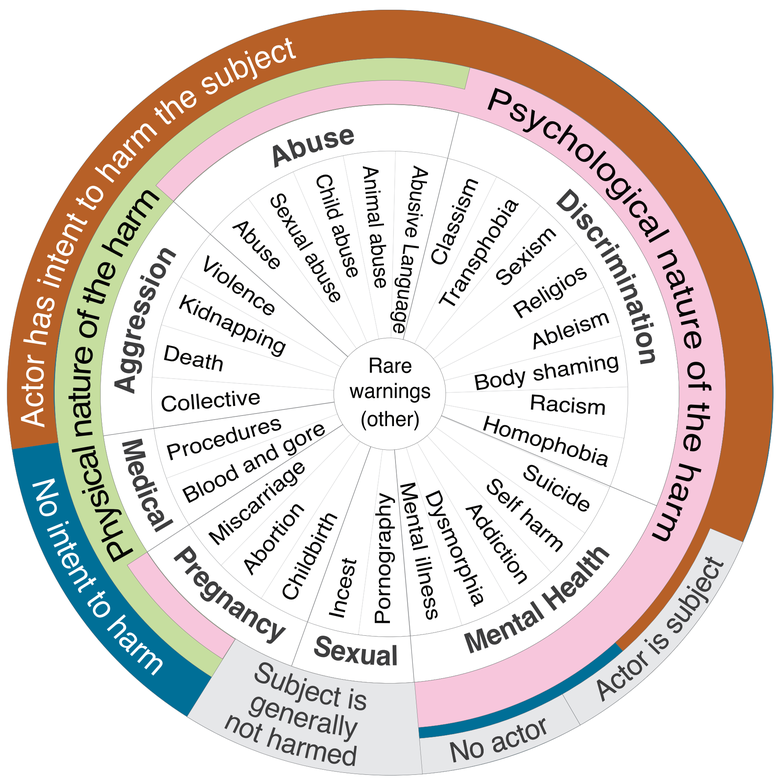
Computational Research on Trigger Warning Assignment

Matti Wiegmann, Magdalena Wolska, Benno Stein
Bauhaus-Universität Weimar

Martin Potthast
Leipzig University and ScaDS.AI

Trigger Warning Taxonomy

36-label trigger warning taxonomy based on 8 curated content guidelines.



ACL 23

Warning Assignment on Documents

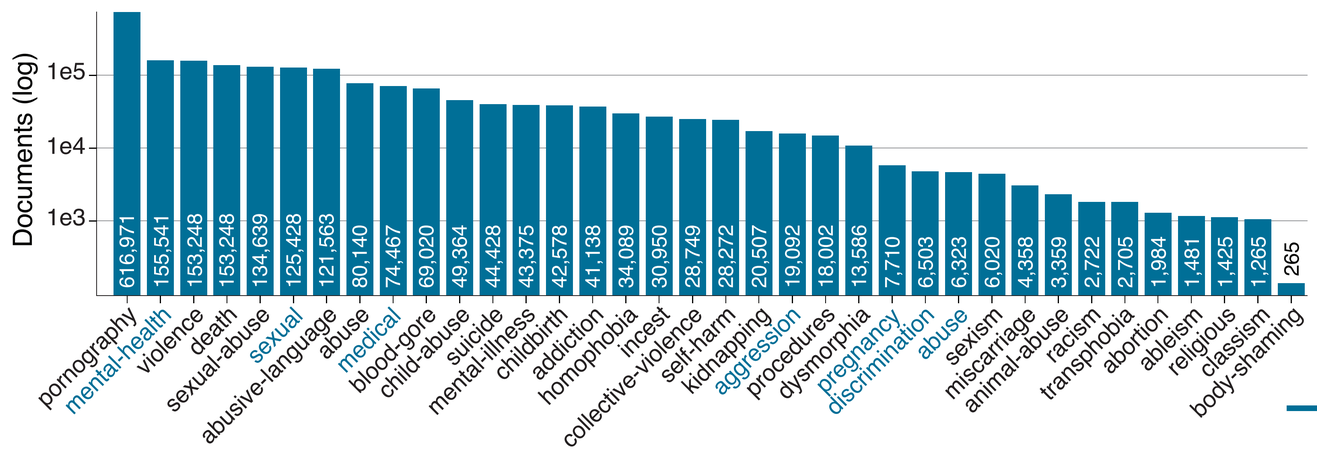
Show a trigger warning before a document if it contains harmful content.

Dataset of 1.1 million fan fiction documents.

- Multi-label (mean: 1.5 labels/document)
- Often long documents (7,986 mean words/document; median: 3,096).
- Warnings are based on information from the authors.

Classification F_1 over 36 warnings (multi-label):

- Micro-average ca. 0.52;
- Macro-average ca. 0.30



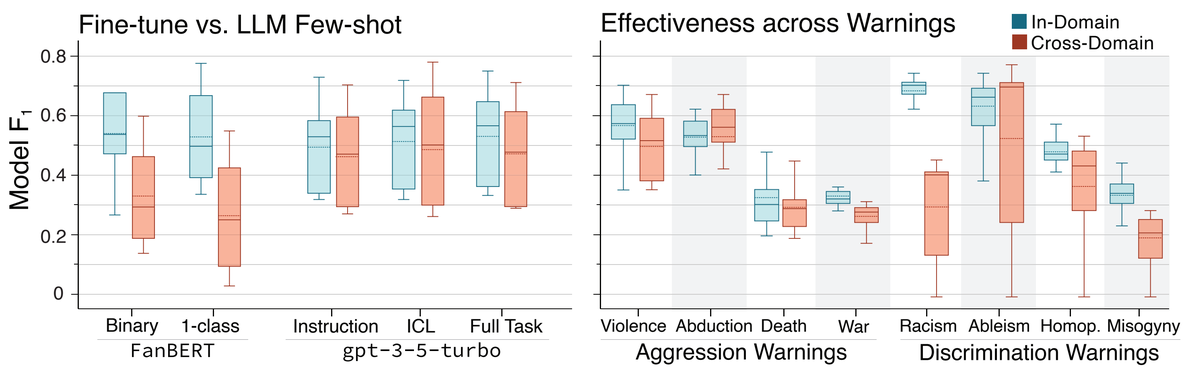
Warning Assignment on Passages

In Review

Show a trigger warning at the point in the text where the harmful content appears.

Classification across 42 expriments:

- Effectiveness varies a lot between warnings.
- LLM prompting is better cross-domain.
- Prompt variations have little effect (unlike for annotators).



Annotation of 4,000 passages (3 votes; 6 trained annotators).

- Subjective; Krippendorff α of .23–.45 (avg. .37).
- Prioritize personalization over definitve decisions?

Structure of the annotation task and prompts:

PERSONA Assume you have suffered through contact with death ...

DEFINITION Death includes graphic or implied descriptions...

INSTRUCTION 1 Mark the following story piece `positive` if contains death ...

INSTRUCTION 2 A Warning is required if the text evokes ...

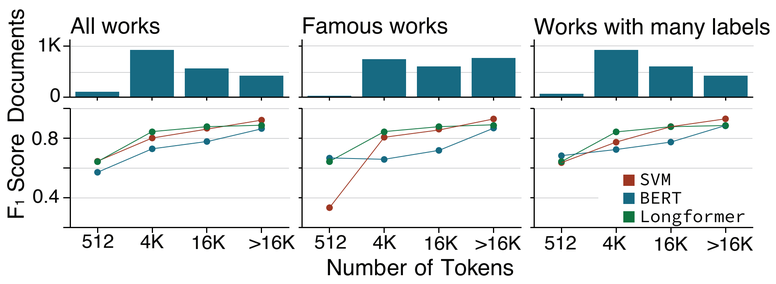
PASSAGE Text: The moans and cries of the ...

Analysis of Warnings

Findings 23

What differentiates violent and non-violent documents?

- Models are more effective on longer documents.
- Finding the violent passages is important.



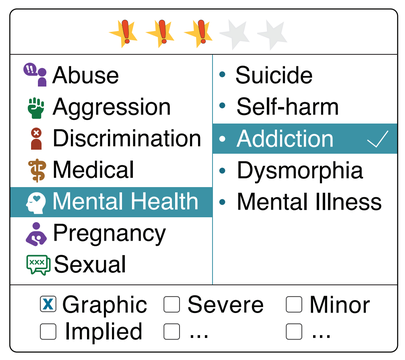
- Domain-independent features are more discriminative.

Discriminative Features				
Least	4.65 blood	2.40 dead	2.37 kill	2.33 scream
Most	-1.67 kiss	-1.07 manage	-1.01 ridiculous	-0.92 admit

Reserach Directions

Annotation Quality

- Let individuals report content that triggered them. As done in clinical research.
- Do intensity ratings or qualifiers help to reduce disagreement and subjectivity?



Data from different Genres

- Online Speech. Mastodon, ...
- Web Text. non-fiction, blogs, ...

Personalization

- Calilbrate classification and generation models to individuals.