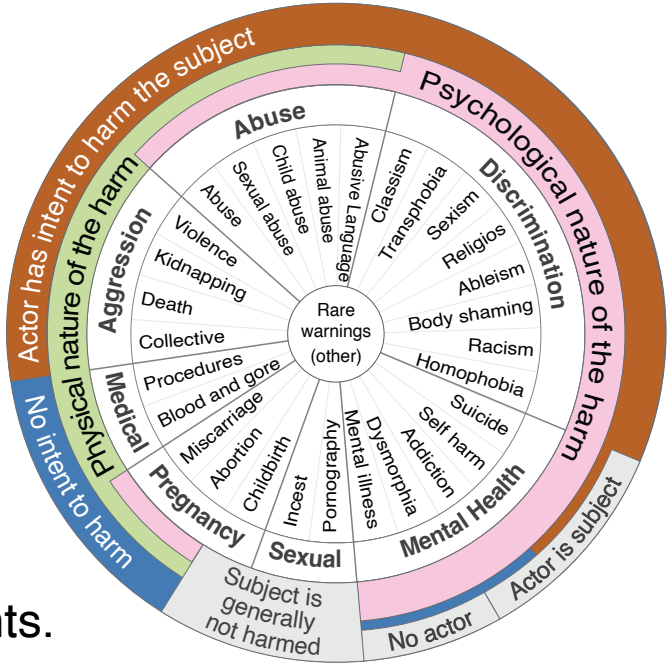
Computational Research on Trigger Warning Assignment

Matti Wiegmann Magdalena Wolska Benno Stein Martin Potthast

Trigger warnings mark online content that is harmful to certain individuals or groups.

- Assumption: The individual decides if the content is harmful. Most people are unaffected, so stronger actions (i.e. deletions) are not warranted.
- Labels are based on the type of content instead of the type of harm done. doxing, hate speech, …
- Research concerns are subjective annotations, model sensitivity, and personalization.



Our contributions: taxonomy, datasets, document classification, passage classification, LLM experiments.