Introducing Computational Research on Trigger Warnings*

Matti Wiegmann[1] **Magdalena Wolska**[1] **Benno Stein**[1] **Martin Potthast**[2]

[1]Bauhaus-Universität Weimar [2]Leipzig University and ScaDS.AI

A trigger warning is used to warn people about potentially harmful content. Any user-generated content can be intentionally harmful and there is a clear approach to moderating it. It must be found and removed. However, online content can also be harmful by addressing topics and situations that may cause mild to severe discomfort or stress in some people, depending on their individual histories, but where removal is not appropriate. To help these people decide whether to consume or avoid such content, many communities have started adding so-called content or trigger warnings, for example at the beginning of YouTube videos, using Mastodon's spoiler feature, or by assigning appropriate search tags on Archive of our Own (AO3).

Trigger warnings were originally used to help patients with post-traumatic stress disorder (Knox, 2017). However, the short list of trauma triggers has been informally expanded to include many more, such as abuse, aggression, discrimination, eating disorders, hate, pornography, or suicide (Charles et al., 2022). Trigger warnings are orthogonal to other harmful content taxonomies, such as violence, hate speech, or toxicity (Wulczyn et al., 2017). Some labels overlap but differ in structure and meaning (Banko et al., 2020).

Wolska et al. (2022) and Wiegmann et al. (2023) introduce computational research on assigning trigger warnings as a text classification task. Based on (academic) guidelines, a new taxonomy of trigger warnings for written content is developed (Figure 1) and a new corpus of about 1 million labeled fan fiction documents from AO3. We examined fan fiction because trigger warnings for (fan) fiction are often desired and because its harmful content is spoken, narrative, and latent.

However, text classification is too coarse-grained for many desired applications such as microblog classification, paragraph detection for detoxification, or assigning warnings to triggering parts of a text. Therefore, we are currently developing new approaches for this task. The challenge is to distinguish between triggering and non-harmful text. Generally, a text is triggering if it evokes an image

*Based on Wolska et al. (2022) and Wiegmann et al. (2023).

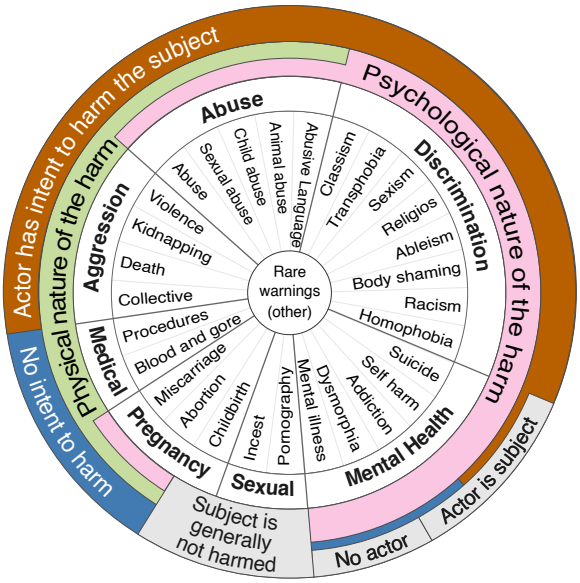

Figure 1: Our taxonomy of trigger warnings for written content (Wiegmann et al., 2023). The white inner rings show 36 trigger warnings, the colored rings show the intent relation and the nature of the harm.

that evokes negative experiences or memories. The more vivid the image, the more likely a warning is needed. Consider the following examples:

- Warning: Death. The disfigurement of each hapless undead slave, some missing limbs, covered in blood and ooze, some naked, some with their skin missing, and more assaulted one's eyes.
- No Warning. I got a few "extra damage towards undead" enchants, and since we were facing nothing but undead for a long time, I made a few.

Key open challenges in computational trigger warning research include among others: (i) Personalizing the label decision based on the intensity of the harmful content and the sensitivity of the reader. (ii) Integrating user models (e.g., a personal exclusion list) to extend the taxonomy, especially in the context of generative AI. (iii) Addressing the scarce and noisy data problem of applying trigger warning technology to other platforms.

References

Michele Banko, Brendon MacKeen, and Laurie Ray. 2020. A unified taxonomy of harmful content. In *Proceedings of the Fourth Workshop on Online Abuse and Harms, WOAH 2020, Online, November 20, 2020*, pages 125–137. Association for Computational Linguistics.

Ashleigh Charles, Laurie Hare-Duke, Hannah Nudds, Donna Franklin, Joy Llewellyn-Beardsley, Stefan Rennick-Egglestone, Onni Gust, Fiona Ng, Elizabeth Evans, Emily Knox, et al. 2022. Typology of content warnings and trigger warnings: Systematic review. *PloS one*, 17(5):e0266722.

Emily Knox. 2017. *Trigger Warnings: History, Theory, Context*. Rowman & Littlefield.

Matti Wiegmann, Magdalena Wolska, Christopher Schröder, Ole Borchardt, Benno Stein, and Martin Potthast. 2023. Trigger Warning Assignment as a Multi-Label Document Classification Problem. In *Proceedings of the 61th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12113–12134, Toronto, Canada. Association for Computational Linguistics.

Magdalena Wolska, Christopher Schröder, Ole Borchardt, Benno Stein, and Martin Potthast. 2022. Trigger warnings: Bootstrapping a violence detector for fanfiction. *CoRR*, abs/2209.04409.

Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017. Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th International Conference on World Wide Web*, WWW '17, pages 1391–1399, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.