# Trigger Warnings are Grounded in a Shared Vocabulary: A Corpus Analysis with User-Generated Labels

**Sebastian Heineking**[1], **Matti Wiegmann**[2], **Magdalena Wolska**[2], **Benno Stein**[2], **Martin Potthast**[1,3]
[1]University of Kassel, [2]Bauhaus-Universität Weimar, [3]hessian.AI and ScaDS.AI
Kassel, Germany; Weimar, Germany
[1]{firstname.lastname}@uni-kassel.de, [2]{firstname.lastname}@uni-weimar.de

**Abstract**

Trigger warnings advise of potentially disturbing content. On that note: *This document discusses abuse*. But can we trust trigger warnings? For a warning to be credible, independent authors must have a shared understanding of the type of content that advises caution. We investigate for the first time whether trigger warnings are aligned with the vocabulary of texts written by uncoordinated authors. To quantify the lexical alignment of trigger warnings, we conduct a series of statistical tests on the texts of fan fiction authors who used warnings relating to emotional, physical, or sexual abuse. We find that the vocabulary of texts with these warnings is aligned with a curated dictionary of terms related to abuse. However, a high frequency of a term in texts with a warning does not necessarily indicate a semantic relation.

**Keywords:** Digital humanities, text analytics, corpus, trigger warnings

## 1. Introduction

Many pieces of writing deal with potentially disturbing topics. To advise of them in advance, trigger or content warnings have emerged as a new type of meta information. They originated in online forums where people shared traumatic experiences (Jones et al., 2020), but have since become more widespread and are used for different media and contexts (Bridgland et al., 2023). Despite their broad application, trigger warnings are by far not standardized and there is no central authority that assigns them (Charles et al., 2022).

Wiegmann et al. (2024) found the task of annotating texts with trigger warnings to lead to noticeable disagreements between annotators. This is not surprising since, unlike in traditional—reliable—coding scenarios, where annotation constructs are defined by annotation guidelines and coded by experts or trained annotators, the very notion of *triggering content* is inherently vague. For the authors of a text, the assessment of what might be triggering is subjective and at least partially based on empathy alone rather than personal exposure to warning-related experiences. From the point of view of the audience, the range of triggering phenomena is open-ended (Knox, 2017) and, crucially, what causes a person to have recollections of negative experiences is highly individual (Wahlsdorf et al., 2024). We therefore ask whether warnings assigned by uncoordinated individuals are credible labels, for example, for the training of large language models (LLMs). In colloquial terms:

Can we trust trigger warnings?

In this paper, we study the vocabulary of texts with trigger warnings by conducting corpus comparisons on a large collection of fan fiction from Archive of Our Own (AO3), tagged with warnings by their authors.[1] In the first part, we test if trigger warnings are lexically aligned with the texts they are applied to. As a test of lexical alignment, we measure if terms related to a given warning are more frequent in texts with that warning than in those without it. For example, one might expect terms like 'bruise' or 'beating' to be more frequent in texts with the warning *Physical Abuse*. Conversely, such terms may constitute lexical hints that help locate the portion of text that justifies the warning. This information would be valuable for readers and researchers, but is often missing as trigger warnings are applied to documents rather than specific passages (Wiegmann et al., 2024). Hence, in the second part, we flip the perspective and test if high frequency terms exhibit a semantic relation to the warning. Our research questions are twofold:

**RQ1** *Are warnings lexically aligned with the vocabulary of their texts?* For a given trigger warning, we (i) define a dictionary of expected terms, (ii) compare documents tagged with a warning with a baseline corpus to determine term frequency differences, and (iii) statistically test if these differences indicate lexical alignment (Section 3). Applied to 2.9 million documents tagged for *Abuse*, our method reveals the warnings *Emotional*, *Physical*, and *Sexual Abuse* to be aligned with vocabularies expected for them (Section 4). This suggests that the author-assigned warnings are credible labels.[2]

[1]Code: https://github.com/webis-de/LREC-26
[2]The dictionaries for all warnings can be found in Tables 8-13 in Appendix A.



*Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026)*, pages 7106–7125
11–16 May 2026. 

**RQ2** *Are highly frequent terms semantically related to the warning?* Given the previous findings, we test if terms with a relatively higher frequency are semantically related to the warning (Section 5). We (i) collect terms with the highest relative frequency in documents with a warning, (ii) construct a dataset of 10,000 text passages that contain these terms, (iii) simulate annotations using sociodemographic prompting of an LLM, and (iv) test if passages with these terms are more likely to be labeled with a warning than random passages. We find this to be the case for 58 % of the highly frequent terms for *Physical Abuse*, 55 % for *Sexual* and 43 % for *Emotional Abuse*.

## 2. Background and Related Work

Below, we review related work on trigger warnings, comparing corpora, and annotator simulation.

**Trigger Warnings in Context** To avoid disturbing content, many refer to the meta information of a medium as if it were a kind of nutrition facts label, but for information (Fuhr et al., 2017). Depending on the type of content, relevant meta information may include the age rating, genre label, title, summary, and reviews. However, these pieces of information usually convey only superficially *why* some content might be disturbing, since their respective purposes differ; namely to pique interest (title, summary), to criticize (reviews), to protect minors (age rating), and to categorize the content (genres).[3] Hence, people often avoid certain content, such as the horror genre or mature-rated content, entirely.

Trigger warnings provide much more specific information about potentially disturbing topics. But unlike content rating systems, they were not developed by a central authority. Instead, they originated in online communities to help individuals with post-traumatic stress disorder to avoid reminders of their trauma, specifically sexual assault (Jones et al., 2020). Eventually, they were adapted to other contexts like social media and education, and expanded to warnings beyond canonical traumatic events (Bridgland et al., 2023).[4] One community that uses trigger warnings extensively is the fan fiction website Archive of Our Own (AO3).[5]

[3]Originally, Plato defined genres to be "primarily [about] evaluating the suitability of literary production for the ideal state and the education of its young [...]" (Miller and Devitt, 2019).

[4]Trigger warnings are also subject to controversial debate; see the "Ethical Considerations" for more background.

[5]https://archiveofourown.org; named after a LiveJournal user's idea in 2007 to create a non-profit platform to avoid the commercial exploitation of fan fiction by third parties.

**Subjectivity of Trigger Warnings** Building on the tags that authors on AO3 assign to their works, Wiegmann et al. (2022) created the Webis Trigger Warning Corpus 2022 (WTWC-22) and released it under CC BY 4.0. It consists of 7.8M works, with an average of three chapters and English as the predominant language (7.1M works). In a follow-up study, Wiegmann et al. (2024) collected annotations from seven annotators for passages from the WTWC-22. The authors reported an inter-annotator agreement of 0.22–0.52 Krippendorff's $\alpha$, relating trigger warning assignment to other subjective tasks with similar agreement scores (Rottger et al., 2022), like annotating toxicity (Sap et al., 2019, 2022) and hate speech (Salminen et al., 2019; Waseem, 2016). Given these annotation results, we take a step back and analyze the underlying data. We raise the question if the texts written by uncoordinated authors have a shared vocabulary that is lexically aligned with the trigger warnings applied to the texts.

**Comparing Corpora** Our first goal is to determine if trigger warnings are lexically aligned with the vocabulary of their texts. Traditionally, a task like this is operationalized using statistical methods for comparing corpora (Kilgarriff, 2001; Kour et al., 2022). We test if certain terms are more frequent in documents with a trigger warning than in documents without it. Past methods employ log-likelihood tests (Rayson and Garside, 2000) or $\chi^2$ tests (Dunning, 1993), but Lijffijt et al. (2014) found these methods to be prone to Type I errors. We follow their recommendation and instead apply the Mann-Whitney U test (MWU) as outlined in their paper, quantifying the *evidence* for corpus differences by ranking documents in them by their term frequencies. We further complement this test with an effect size measure that indicates *how much* more frequent a term is in one corpus relative to another. Schlatt et al. (2022) developed such a measure for medical texts based on similar approaches by Ahmad et al. (1999), Park et al. (2008), and Wong et al. (2007). We adapt their *term domain specificity*, and use the more literal name *log ratio* (Hardie, 2014).

Our second goal is to determine if terms with a relatively higher frequency are semantically related to the warning. In addition to the MWU, we use the Jensen-Shannon Divergence (JSD; Lin (1991)) to identify terms with a relatively higher frequency in one corpus. Given the findings by Lu et al. (2020), we apply the version of the JSD developed by Pechenick et al. (2015) instead of the alternative by Gallagher et al. (2018).[6]

[6]We do not use JSD in the first set of our experiments as it (i) does not allow for significance testing, and (ii) the log ratio is a more intuitive effect size measure.

**Predicting Annotation Subjectivity** To estimate if a given term is semantically related to a warning, we simulate annotations on passages that contain it. Specifically, we prompt an LLM to decide for a passage of text if it requires a warning or not. Given the aforementionned subjectivity of trigger warning annotation, we use sociodemographic prompting to predict if a passage is likely to receive unanimous annotations by annotators or cause disagreement between them. Sociodemographic prompting influences the output of LLMs by asking them to generate responses as if given by individuals with a specific sociodemographic profile (Hwang et al., 2023). A text sample is predicted to cause disagreements between annotators if different profiles lead to different annotations (Wan et al., 2023). In a comprehensive study with multiple model families, Beck et al. (2024) found `Flan-T5` 11B to be the most effective model at this task with an average F1-score of $0.62$ across seven datasets.

## 3. Measuring Lexical Alignment

For a trigger warning to be credible, uncoordinated authors must have a common understanding of the warning. We hypothesize that this common understanding is reflected in a lexical alignment between the trigger warnings and the corpora of texts they are applied to. Hence, documents with a trigger warning can be expected to contain terms related to the warning significantly more often than comparable baseline documents. To use a specific example, a text with a *Sexual Abuse* warning is more likely to contain terms such as 'molest' than documents without such a warning.

### 3.1. Corpus Creation

The authors on AO3 assign tags using a freeform field when publishing their works. A work about Lord of the Rings, for example, may have the tags *Legolas* or *Frodo Baggings*. The number of tags per work is typically around 20. To cope with the lexical diversity resulting from freeform tagging, volunteers, known as tag wranglers, create relations between synonymous tags or those that define a sub-concept of another tag. This results in a graph of tags with canonical root nodes that define broad concepts such as *Humor*, *Sexual Content*, or *Friendship*.[7] Wiegmann et al. (2023) refined these relations to extract the graph of tags that formed the basis for their trigger warning taxonomy.

**Tag Collection** We first collect all tags used on AO3 that are subordinate to *Abuse*. We start at the corresponding root node in the refined tag graph by Wiegmann et al. and traverse it along all edges that indicate synonymy or a sub-concept. This results in a total of 5,654 tags such as *Abuse of Authority*, *TW Physical Harm*, or *Gaslighting*. Next, we retrieve a total of $|D| =$ 2,965,898 documents, i.e. chapters, from the WTWC-22 corpus written in English and annotated by Wiegmann et al. (2022) for *Abuse*.

| **Category** $c_i$ | $\lvert X_{ci}\rvert$ | $\lvert D_{ci}\rvert$ | $\lvert \bar{D}_{ci}\rvert$ |
|---|---|---|---|
| $c_1$ Emotional Abuse | 515 | 49,989 | 2,872,109 |
| $c_2$ Physical Abuse | 694 | 125,408 | 2,795,274 |
| $c_3$ Sexual Abuse | 618 | 107,433 | 2,858,465 |

Table 1: Tags and documents per category

**Tag Annotation** Given the variety of tags referring to *Abuse*, the number of documents for each individual warning tag is insufficient to construct a corpus for statistical analyses. However, reviewing the tags, we notice that the authors often explicitly use three types of abuse, namely, $c_1 =$ *Emotional Abuse*, $c_2 =$ *Physical Abuse*, and $c_3 =$ *Sexual Abuse*. Further inspection revealed that many other tags implicitly refer to them, like *Manipulative Parents* to emotional or *Child Molestation* to sexual abuse.

Based on this observation, we annotated the 5,654 tags for these three warning categories.[8] As most tags contain the literal name of the category (e.g., *Emotional Abuse*), instead of sub-categories (e.g., *Gaslighting*), they are straightforward to annotate. For the remainder, we guided our annotation using information pages offered by two government institutions in the US and UK, namely the Washington State Department of Social and Health Services and the Social Care Institute for Excellence.[9] Both pages provide an overview of different types of *Abuse* and ways to identify them.[10]

For each warning category $c_i$, Table 1 shows size of the respective tag set $X_{ci}$ that contains all tags annotated for $c_i$. Additionally, the table lists the number of documents $D_{ci}, \subset D$ with at least one tag from $X_{ci}$. The most documents are tagged for *Physical Abuse*, while *Emotional Abuse* is noticeably less common. As a reference for measuring alignment, we form a complementary set of documents $\bar{D}_{ci} = D \setminus D_{ci}$ called baseline corpus.[11]

[7] See https://archiveofourown.org/tags/ for an overview.

[8] All annotated tags can be found in our code repository.

[9] https://www.dshs.wa.gov and https://www.scie.org.uk.

[10] As *Neglect* can be both a form of *Emotional* or *Physical Abuse*, we annotated it separately.

[11] To avoid introducing noise, the baseline corpora $\bar{D}_{ci}$ for *Emotional* and *Physical Abuse* contain only documents that are tagged neither with $c_i$ nor *Neglect*. Documents tagged with both *Neglect* as well as $c_1 =$ *Emotional* or $c_2 =$ *Physical Abuse* are added to $D_{ci}$.

### 3.2. Corpus Comparison

To test if author-assigned tags related to $c_i$ are lexically aligned with the vocabulary of their documents, we compare the term frequencies in documents $d \in D_{ci}$ against the frequencies in the baseline corpus $\bar{D}_{ci}$. We (i) define a dictionary $T_{ci}$ of warning-related terms, (ii) calculate the term frequencies for each document, and (iii) perform significance tests on the distribution of term frequencies.

**Category Dictionaries** To collect candidates for each warning-related dictionary, we tap three sources: First, we derive a set of seed terms from the information pages on *Abuse* mentioned above. Then, we expand the seed terms with synonyms from the Merriam Webster thesaurus[12] as well as manually filtered suggestions by `GPT-4o`. The dictionary of category terms $T_{ci}$ is finalized by adding syntactic variations of the terms in the dictionary to allow better disambiguation.[13] After exhausting the first two sources for new terms, we stopped collecting when `GPT-4o` showed diminishing returns, resulting in a total of 400 terms for both *Emotional* and *Physical Abuse*, and 250 for *Sexual Abuse*.

**Term Frequencies** For each term $t$ in $T_{ci}$ and document $d \in D$, we calculate a term frequency $\text{tf}(t, d)$. Additionally, we calculate corpus-level term frequencies $\text{tf}(t, D_{ci})$ and $\text{tf}(t, \bar{D}_{ci})$. Both are normalized by the total number of terms in the respective document or corpus.

**Significant Differences in Frequency** To test if a term occurs significantly more often in $D_{ci}$ compared to $\bar{D}_{ci}$, we apply the Mann-Whitney U test as outlined by Lijffijt et al. (2014). For a term $t$, all documents $d$ in both corpora are ranked jointly by their term frequencies $\text{tf}(t, d)$ in ascending order. This ranking is used to calculate the test statistic $U = min(U_1, U_2)$, where

$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1, \text{and} \quad (1)$$

$$U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2. \quad (2)$$

Here, $n_1 = |D_{ci}|$ and $n_2 = |\bar{D}_{ci}|$ are the total number of documents, and $R_1$ and $R_2$ are the rank sums of all documents in the two corpora. $U_2$ is the number of pairs between documents in which the document from $D_{ci}$ has a higher term frequency than the document from $\bar{D}_{ci}$. As our sample sizes are sufficiently large, the distribution of $U$ is well approximated by a Gaussian distribution and allows for the calculation of a $z$-score (Siegel, 1956; Lehmann, 1975). From the $z$-score, we finally calculate a $p$-value to test the following hypotheses for each term $t \in T_{ci}$:

$$H_0 : F_{t,ci} = G_{t,ci}; \quad H_A : F_{t,ci} \neq G_{t,ci},$$

where $F_{t,ci}$ is the distribution of term frequencies for $t$ in $D_{ci}$, and $G_{t,ci}$ is the distribution of term frequencies in $\bar{D}_{ci}$. As we perform multiple tests, one for each term, we account for the risk of incorrectly rejecting the null hypothesis through multiplicity with a Bonferroni correction. Significance is tested based on the corrected $p$-value $p^* = p \cdot |T_{ci}|$.

**Effect Size** As a measure of effect size for the increase in term frequency, we calculate a log ratio of the corpus-level term frequencies. For a term $t$, the log ratio $\text{lr}(t)$ is calculated as follows:

$$\text{lr}(t) = \log_2 \frac{\text{tf}(t, D_{ci})}{\text{tf}(t, \bar{D}_{ci})}. \quad (3)$$

While the Mann-Whitney U test measures *how much evidence* there is for a difference in term frequency between the two corpora, the log ratio indicates *how pronounced* that difference is. Taking the log of the ratio centers it around zero. Using the binary logarithm facilitates the interpretation, as every full unit increase in log ratio is equal to a doubling of the ratio of frequencies.

## 4. Results of the Alignment Tests

As a first step, we tokenize each of the 2.9 million documents and apply POS-tagging and lemmatization to each token.[14] For each category $c_i$ and each lemma in documents of the category corpus $D_{ci}$, we perform the Mann-Whitney U test. Of the terms in the respective category dictionaries $T_{ci}$, 343 occur for *Emotional*, 346 for *Physical*, and 233 for *Sexual Abuse*.

### 4.1. Significant Differences in Frequency

At a significance level $\alpha = 0.05$, we find 190 dictionary terms to be significantly more frequent in $D_{ci}$ for *Emotional Abuse*, 130 for *Physical*, and 124 for *Sexual Abuse*.[15] Some of the highly significant terms are the same between categories: The noun 'fear' and the verb 'force' are among the ten terms with the highest $z$-score for both *Emotional* and *Sexual Abuse*. *Emotional* and *Physical Abuse* have the noun 'anxiety' in common, and *Physical*

[12] https://www.merriam-webster.com/thesaurus/

[13] Example: For the verb 'insult', we add the adjectives 'insulted' and 'insulting', as well as its noun version 'insult'.

[14] For tokenization, lemmatization, and POS-tagging, we use `en_core_web_sm` of spacy version 3.7.2.

[15] All terms can be found in Table 8-13 in Appendix A.

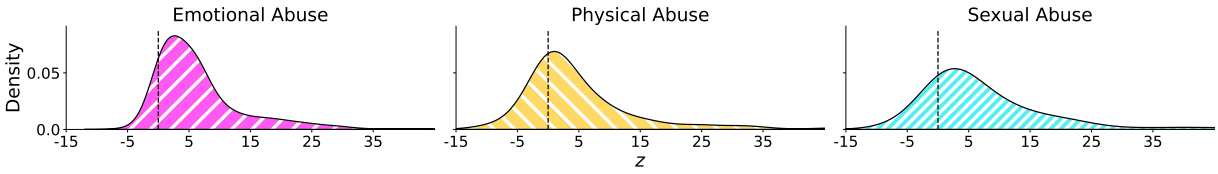



Figure 1: Distribution of the $z$-score over the dictionaries. The plots were created using kernel density estimation.

(a)

| Emotional Abuse | | Physical Abuse | | Sexual Abuse | |
|---|---|---|---|---|---|
| Term (POS) | $z$ | Term (POS) | $z$ | Term (POS) | $z$ |
| hurt (V) | 43.55 | bruise (N) | 80.77 | rape (V) | 105.72 |
| gaslighting (N) | 38.71 | scared (A) | 48.93 | rape (N) | 69.73 |
| fear (N) | 33.21 | beat (V) | 47.30 | sexual (A) | 52.38 |
| force (V) | 30.22 | beating (N) | 46.53 | scared (A) | 51.38 |
| tear (N) | 29.35 | flinch (V) | 43.61 | touch (V) | 50.07 |
| anxiety (N) | 29.17 | bruise (V) | 36.46 | sex (N) | 47.76 |
| trust (V) | 28.57 | punishment (N) | 33.40 | fear (N) | 42.33 |
| panic (N) | 27.15 | cut (N) | 32.96 | bruise (N) | 40.85 |
| gaslight (V) | 26.38 | broken (A) | 32.81 | molest (V) | 40.85 |
| guilt (N) | 25.64 | anxiety (N) | 32.81 | force (V) | 36.93 |

(b)

| Emotional Abuse | | Physical Abuse | | Sexual Abuse | |
|---|---|---|---|---|---|
| Term (POS) | $z$ | Term (POS) | $z$ | Term (POS) | $z$ |
| infringement (N) | -5.36 | trap (N) | -11.27 | stalk (V) | -8.78 |
| curse (V) | -3.80 | bind (V) | -11.05 | thrust (V) | -7.06 |
| whimpering (A) | -1.09 | push (N) | -9.71 | concentration (N) | -6.60 |
| disoriented (A) | -1.03 | capture (V) | -9.69 | pussy (N) | -6.44 |
| abase (V) | -0.88 | spank (V) | -9.45 | obscene (A) | -5.62 |
| harasser (N) | -0.66 | pull (N) | -7.84 | stalker (N) | -4.75 |
| yelled (A) | -0.62 | slapping (N) | -7.70 | flash (V) | -4.22 |
| fuming (A) | -0.60 | denial (N) | -6.66 | leak (V) | -3.23 |
| deceiving (A) | -0.60 | spank (N) | -6.56 | exhibitionism (N) | -3.13 |
| tyrannize (V) | -0.56 | attack (V) | -6.20 | creepy (A) | -3.11 |

Table 2: Selected results of the Mann-Whitney U test. The tables show the ten terms with the (a) highest and (b) lowest $z$-scores for all three categories.

(a)

| **Category** $c_i$ | | $\overline{z}$ | $\sigma_z$ | 95 % CI | $\lvert T_{ci}\rvert$ |
|---|---|---|---|---|---|
| $c_1$ | Emotional Abuse | 6.42 | 7.15 | (5.66, 7.18) | 343 |
| $c_2$ | Physical Abuse | 5.02 | 10.06 | (3.95, 6.08) | 346 |
| $c_3$ | Sexual Abuse | 8.04 | 12.78 | (6.39, 9.69) | 233 |

(b)

| **Category** $c_i$ | | $\overline{\text{lr}}$ | $\sigma_{\text{lr}}$ | 95 % CI | $\lvert T_{ci}\rvert$ |
|---|---|---|---|---|---|
| $c_1$ | Emotional Abuse | 0.26 | 0.61 | (0.20, 0.33) | 343 |
| $c_2$ | Physical Abuse | 0.13 | 0.51 | (0.07, 0.18) | 346 |
| $c_3$ | Sexual Abuse | 0.29 | 0.61 | (0.21, 0.36) | 233 |

Table 3: Aggregated (a) $z$-score and (b) log ratio.

and *Sexual Abuse* share the noun 'bruise' and the adjective 'scared'. The terms are listed in Table 2.

The null hypothesis is also rejected for a few terms that are significantly *less* frequent in their category corpus. This applies to the noun 'infringement' for *Emotional Abuse*, 20 terms for *Physical*, and seven for *Sexual Abuse*. The remaining terms have document-level term frequencies that are not significantly different between the two corpora.

For illustrative purposes, Figure 1 and Table 3a provide an aggregated view on the $z$-scores over the entire dictionaries. One observation is that the *Emotional Abuse* dictionary has the narrowest distribution and the fewest terms with $z \leq 0$. The figure also shows that the *Sexual Abuse* dictionary has the most terms with a lot of evidence for a higher frequency. Despite some outlier terms with very high $z$-scores, we observe the lowest mean $z$-score for *Physical Abuse*.

## 4.2. Effect Sizes

In addition to the significance tests, we quantify the difference in term frequency using the log ratio. The aggregated results for each category are given in Table 3b. Analogously to the $z$-scores, we observe the lowest average log ratio for *Physical Abuse*. For *Emotional* and *Sexual Abuse*, we observe similar values for the mean log ratio and standard deviation. Additionally, the log ratios of significantly more frequent terms are also higher for these two categories than for *Physical Abuse*. *Emotional Abuse* has three terms with particularly high log ratios: the noun 'gaslighting' with 3.87, the verb 'gaslight' with 3.15, and the adjective 'manipulated' with 3.02 (see also Table 4).

## 4.3. Discussion

Our key finding is that documents with warnings indicating *Emotional*, *Physical*, or *Sexual Abuse* also contain terms from the corresponding vocabularies with a higher frequency than documents without these warnings. While 48 % and 50 % of the dictionary terms for *Emotional* and *Sexual Abuse* have significantly higher frequencies, the same is true for only 33 % of the *Physical Abuse* terms. Likewise, the increase in mean log ratio is also the lowest for *Physical Abuse*. A possible explanation is that documents on AO3, which are tagged for *Abuse* without additional categorization, use terms describing *Physical Abuse* relatively more often than terms from the other two vocabularies.

(a)

| Emotional Abuse | | Physical Abuse | | Sexual Abuse | |
|---|---|---|---|---|---|
| Term (POS) | lr | Term (POS) | lr | Term (POS) | lr |
| gaslighting (N) | 3.87 | beating (N) | 0.93 | molestation (N) | 1.96 |
| gaslight (V) | 3.15 | bruise (N) | 0.88 | gonorrhea (N) | 1.94 |
| manipulated (A) | 3.02 | bruising (N) | 0.80 | hypersexuality (N) | 1.67 |
| blamed (A) | 2.78 | contusion (N) | 0.76 | pedophile (N) | 1.65 |
| anorexic (A) | 1.16 | bruised (A) | 0.75 | rape (V) | 1.62 |
| infantilize (V) | 1.10 | shiner (N) | 0.75 | syphilis (N) | 1.62 |
| seclusion (N) | 1.03 | discoloration (N) | 0.72 | molest (V) | 1.57 |
| desertion (N) | 0.94 | welt (N) | 0.71 | rape (N) | 1.48 |
| invalidate (V) | 0.87 | gash (V) | 0.70 | violated (A) | 1.48 |
| isolation (N) | 0.67 | hematoma (N) | 0.69 | chlamydia (N) | 1.33 |

(b)

| Emotional Abuse | | Physical Abuse | | Sexual Abuse | |
|---|---|---|---|---|---|
| Term (POS) | lr | Term (POS) | lr | Term (POS) | lr |
| infringement (N) | -1.78 | spank (V) | -0.76 | pussy (N) | -0.53 |
| — | — | bound (A) | -0.55 | stalker (N) | -0.34 |
| — | — | slapping (N) | -0.47 | obscene (A) | -0.33 |
| — | — | spank (V) | -0.41 | concentration (N) | -0.26 |
| — | — | skeletal (A) | -0.35 | stalk (V) | -0.26 |
| — | — | trap (N) | -0.33 | thrust (V) | -0.20 |
| — | — | spanking (N) | -0.27 | flash (V) | -0.11 |
| — | — | manhandle (V) | -0.26 | — | — |
| — | — | attacker (N) | -0.26 | — | — |
| — | — | denial (N) | -0.26 | — | — |

Table 4: Selected log ratios. The tables show the ten terms with the (a) highest and (b) lowest log ratios, restricted to terms with a significantly higher/lower term frequency.

## 5. Vocabulary of Texts With Warnings

Section 4 revealed that documents with warnings for *Emotional*, *Physical*, and *Sexual Abuse* also use terms from curated dictionaries with a higher frequency. For the second set of experiments, we take the opposite perspective and analyze terms with a relatively higher frequency in texts with a warning. On the one hand, these terms provide further insight into the vocabulary of documents with trigger warnings. On the other hand, the terms could be semantically related to the warning and thus lexical hints for warning-related content in a piece of text. An example of a lexical hint is the verb 'bruise' for the warning *Physical Abuse*.

An effective process for identifying lexical hints would be helpful in advancing the research on trigger warnings by facilitating the collection of annotations. As warnings are applied to (sometimes book-length) documents as a whole, annotators would be required to read long samples of texts, large parts of which are irrelevant to the warning. This can negatively affect annotation quality as each sample requires a lot of time, annotators might disagree about the exact position of warning-related content or get distracted.

Lexical hints can address this issue by extracting passages that are (i) shorter and thus more easy to annotate, and (ii) more likely to be warning-related. Finally, the retrieval of passages with lexical hints scales to large collections of documents in contrast to more compute-intensive methods like LLMs.

### 5.1. Collection of Terms

We collect highly frequent terms for each warning based on (i) the $z$-score from the Mann-Whitney U test on term frequencies, and (ii) the contribution of each term to the Jensen-Shannon Divergence (JSD) between the two corpora. For the contribution JSD$(t)$ of a term $t$, we use the equation by Pechenick et al. (2015):

$$\text{JSD}(t) = -h(m_t) + \frac{1}{2}(h(p_t) + h(q_t)); \quad (4)$$

It calculates the contribution $h(p)$ of $p$ to the Shannon entropy

$$h(p) = p \log_2 p, \quad (5)$$

for the corpus-level term frequencies $p_t = \text{tf}(t, D_{ci})$ and $q_t = \text{tf}(t, \bar{D}_{ci})$ as well as their mean $m_t = \frac{p_t + q_t}{2}$.

We collect 100 terms for each warning category $c_i$. Half of the terms are taken from the category dictionary $T_{ci}$ and the other half are taken from all other terms in the category corpus. Each half consists of (i) the 25 terms with the highest $z$-score, and (ii) 25 additional terms with the highest JSD. To exclude overly niche terms, we only collect terms that occur in at least $n = 100$ documents.

### 5.2. Semantic Relation to the Warning

To test if a term is semantically related to a warning, we collect passages that contain it. Then, we test if these passages are more often labeled as requiring a warning than random passages from texts with the same warning. If yes, we conclude that the term is a lexical indicator for the warning and helps to locate warning-related content.

**Passage Retrieval** For each warning category $c_i$, we form a set of passages $P_{ci}$ by (i) retrieving all documents in the category corpus $D_{ci}$ with one of the 100 terms, (ii) identifying sentences that contain one of these terms, (iii) and adding two neighboring sentences in both directions as context. If passages overlap, we merge them into a single larger passage.

For the actual annotation, we reduce $P_{ci}$ to a total of 10,000 passages by sampling uniformly over the 100 terms.[16] In addition to $P_{ci}$, we sample a set $\hat{P}_{ci}$ of 1,000 random passages from documents in the category corpus $D_{ci}$. While not containing any of the 100 terms, these random passages are still part of documents tagged for the same warning category and thus a suitable baseline.

[16] As the same passage can contain multiple terms, the number of passages per term deviated.

Imagine you are a person of gender 'male', race 'Hispanic', age '25-34', education level 'Bachelor's degree in college (4-year)' and political affiliation 'Independent' that has experienced emotional abuse themselves or of their relatives in the past. Is there anything in the given text that you could associate with this experience? [...]
Requires warning for emotional abuse ('yes' or 'no'):

Figure 2: Example for an annotation prompt. It first defines the sociodemographic profile, then asks to imagine having had a warning-related experience, and finally poses the annotation question.

| # | Gender | Race | Education | Age | Political Affiliation |
|---|---|---|---|---|---|
| 1 | Female | White | College (no degree) | Under 18 | Liberal |
| 2 | Male | Hispanic | Bachelor's degree | 25 - 34 | Independent |
| 3 | Male | Black | College (no degree) | Under 18 | Independent |
| 4 | Male | White | Associate degree | 35 - 44 | Liberal |
| 5 | Nonbinary | White | Master's degree | 25 - 34 | Liberal |
| 6 | Female | White | Associate degree | 25 - 34 | Conservative |
| 7 | Female | Black | Doctoral degree | 25 - 34 | Liberal |
| 8 | Male | American Indian | Master's degree | 55 - 64 | Conservative |
| 9 | Female | Native Hawaiian | College (no degree) | Under 18 | Liberal |
| 10 | Female | White | Bachelor's degree | 35 - 44 | Conservative |

Table 5: Sociodemographic profiles. The attributes and values are based on Beck et al. (2024).

**Simulating Subjective Annotations** As trigger warning annotation causes disagreement among human annotators, we simulate annotations using sociodemographic prompting to reflect this subjectivity. In their experiments, Beck et al. (2024) found `Flan-T5` 11B to be (i) strongly influenced in its generation behavior by sociodemographic prompting, and (ii) the most effective model at predicting disagreements between annotators. Hence, we use this LLM to simulate subjective annotations for the 11,000 passages.

The key idea of sociodemographic prompting is to use profiles with defined attributes. As Beck et al. (2024) found race and political affiliation to be the most influential individual attributes in changing label predictions, and since trigger warnings, too, are discussed controversially from a political perspective, we create ten profiles with emphasis on these attributes. The profiles are summarized in Table 5. The profiles are inserted into the annotation prompt shown in Figure 2 to influence the annotations by `Flan-T5` 11B. The prompt is adapted from the one used by Wiegmann et al. (2024).

**Distribution of Annotations** Each of the 11,000 passages receives either a 'yes' (positive) or 'no' (negative) annotation by each of the ten profiles, resulting in 0-10 positive annotations per passage. For each term $t$, we create a distribution $A_t$ of the number of positive annotations that passages containing $t$ received. We then compare the distribution $A_t$ against the distribution of positive annotations $\hat{A}_{ci}$ on the 1,000 random passages for category $c_i$. If passages containing $t$ receive significantly more positive annotations, we conclude that $t$ is semantically related to $c_i$.

As the positive annotations are not normally distributed, we again test for significance using the Mann-Whitney U test. Instead of ranking documents by their term frequencies, we rank passages by the number of positive annotations they received. Otherwise, the method is the same as the one outlined in Section 3. For each term $t$, we test the following hypotheses:

$$H_0 : A_t = \hat{A}_{ci}; \quad H_A : A_t \neq \hat{A}_{ci}.$$

Rejecting the null hypothesis means that the distribution of positive annotations for term $t$ is significantly different from the corresponding distribution for random passages. We account for multiplicity by multiplying the $p$-value with the number of terms per warning: $p^* = p \cdot 100$.

In addition to the significance tests, we calculate the share of pairs between a passage containing $t$ and a random passage in which the "term passage" had more positive annotations. This *common language* effect size is calculated as follows (Kerby, 2014):

$$f = \frac{U_2}{n_1 n_2}. \tag{6}$$

$U_2$ is the number of pairs for which a passage from $A_t$ had more positive annotations, $n_1$ is the number of passages that contain $t$, and $n_2 = 1,000$ is the number of random passages. The effect size $f$ lies on the interval $[0, 1]$ with 1 meaning that all "term passages" received more positive annotations than all random passages.

## 6. Results of the Vocabulary Analysis

For each of the three warning categories, we collected 100 highly frequent terms, 50 from the dictionaries we used for alignment testing and another 50 based entirely on their relative term frequency.

**Non-Dictionary Terms** The collection of terms based only on relatively higher frequency returns useful additions to the dictionaries from Section 3. Examples include 'sob', 'fault', and 'lie' for *Emotional Abuse*, 'pain', 'grab', and 'hospital' for *Physical Abuse*, and 'whore', 'hurt', and 'bed' for *Sexual Abuse*. It also identifies some general terms like the verbs 'try' or 'tell' or the adjective 'okay'. Some noise occurs across all categories, primarily character names such as 'brock' or 'ramsay'.

**Annotation Distributions** For each term $t$, we tested if the distribution $A_t$ of annotations on its passages is significantly different from the one on random passages. Table 7 summarizes how many terms received more positive annotations than random passages (effect size $f > 0.5$). This is the case for more than 80 % of dictionary terms across all categories. The share of significant results deviates between 54 % for *Emotional* and 76 % for *Physical Abuse*. The shares are lower for the non-dictionary terms with at least 72-76 % of terms receiving more positive annotations but only 32-42 % of the distributions also being significantly different from the random counterpart. We observe the highest effect sizes for the verb 'gaslight' with 88 %, the noun 'bruise' with 81 %, and the verb 'rape' with 86 % of their passages receiving more positive annotations than the random passages for the warnings for *Emotional*, *Physical*, and *Sexual Abuse*, respectively.

**Correlation with Frequency Measures** We correlated the effect size $f$ of a term with its $z$-score and JSD-value to test if an increase in these two measures translates into more positive annotations. Table 6 provides a visualization and correlation scores. The first observation is that there is a positive correlation between the $z$-score and effect size for all warnings and both dictionary and non-dictionary terms. In contrast to that, we observe no consistent correlation between JSD and effect size. In more detail, the $z$-score has a significantly positive correlation for the dictionary terms of *Emotional Abuse*, the non-dictionary terms of *Physical Abuse*, and both sets of terms for *Sexual Abuse*. For JSD, we find two significant correlations: A positive one for the dictionary terms of *Sexual Abuse* and a negative one for the non-dictionary terms of *Emotional Abuse*.

Another observation from Table 6a is that the *Emotional Abuse* terms are concentrated more strongly around $f = 0.6$, while the scatterplots for the other two warnings show a wider dispersion. In detail, only four terms have an effect size of $f \geq 0.7$, while the same is true for 21 *Physical* and 16 *Sexual Abuse* terms. Finally, both *Physical* and *Sexual Abuse* share the verb 'thump' as an outlier term with a very high contribution to the JSD. This suggests that this term is comparatively rare in the baseline corpus of all documents tagged for *Abuse*.

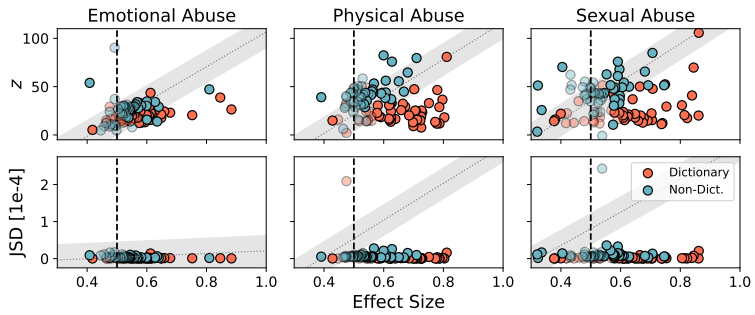
**Discussion** The collection of terms using $z$-score and JSD identifies some useful additions to the dictionaries curated for the alignment tests. Yet, only 32-42 % of the newly identified terms receive significantly more positive annotations than random passages. Hence, a relatively higher frequency does not necessarily indicate a semantic relation between term and warning. Instead, terms collected through frequency differences need to be manually analyzed to remove domain-specific noise like character names or common terms like 'try' that receive a high $z$-score due to occuring in a lot of documents. The effect of manual curation is evident from the higher share of significant results for the dictionary terms. Across all three categories, the share of terms with significantly more positive annotations is 22-36 percentage points above the corresponding value for non-dictionary terms.

Comparing the results between categories, we observed that the share of terms with significantly more positive annotations is consistently higher for *Physical* and *Sexual Abuse* than for *Emotional Abuse*. A possible explanation is that the former two categories are more concrete concepts with defined vocabularies and a clear expression on the lexical level. *Emotional Abuse*, in contrast, is a more abstract concept, requiring contextual information to be identified.

## 7. Conclusion

We investigate the vocabulary of texts with trigger warnings using methods for corpus comparison on 2,965,898 documents tagged for some form of *Abuse*. For the trigger warnings that signal *Emotional*, *Physical*, and *Sexual Abuse*, we found 48 %, 33 %, and 50 % of the warning-related terms from curated dictionaries to be significantly more frequent in texts with the respective warning than in texts with other warnings related to *Abuse*. This indicates lexical alignment between these three warnings and the texts they are applied to, and points to a shared understanding of the warnings among the authors on Archive of Our Own. An avenue for future work is to compare the vocabulary of individual authors for a more detailed analysis.

(a)




(b)

| St. | Cat. | Dictionary | | Non-Dict. | |
|---|---|---|---|---|---|
| | | $r$ | $p$ | $r$ | $p$ |
| $z$ | Emo. | 0.49 | 3.5e-4 | 0.19 | 1.8e-1 |
| | Phy. | 0.23 | 1.1e-1 | 0.58 | 1.2e-5 |
| | Sex. | 0.39 | 4.9e-3 | 0.30 | 3.5e-2 |
| JSD | Emo. | 0.12 | 3.9e-1 | -0.38 | 6.8e-3 |
| | Phy. | -0.23 | 1.1e-1 | 0.27 | 5.5e-2 |
| | Sex. | 0.32 | 2.2e-2 | 0.02 | 8.9e-1 |

Table 6: Relation between the number of positive annotations (effect size) and $z$-score / JSD. (a) Effect sizes per term. Terms to the right of the vertical line received more positive annotations than random. Non-significant results are transparent. The grey corridor indicates perfect correlation based on the minimum and maximum $z$-score / JSD. (b) Correlation between effect size and z-score / JSD (Pearson $r$).

| Category | Dictionary | Non-Dict. | All Terms |
|---|---|---|---|
| Emotional Abuse | 0.86 (*0.54*) | 0.76 (*0.32*) | 0.81 (*0.43*) |
| Physical Abuse | 0.92 (*0.76*) | 0.74 (*0.40*) | 0.83 (*0.58*) |
| Sexual Abuse | 0.84 (*0.68*) | 0.72 (*0.42*) | 0.78 (*0.55*) |

Table 7: Share of terms with more positive annotations than random passages ($f > 0.5$). The share of terms with significant increases is given in parentheses.

The results for *Physical Abuse*, however, also show that the warning-related vocabulary is not exclusive to texts with the respective warning. Instead, two thirds of the terms from the *Physical Abuse* dictionary are also frequent in texts with warnings for *Abuse* without further qualification.

A major obstacle to future work is that authors typically do not assign warnings at the passage level but to full documents. This poses a challenge for the collection of annotations, as annotators are presented with unnecessarily long text samples. We find that high-frequency terms can help in locating warning-related passages, thereby reducing the sample length and facilitating the collection high quality annotations. The frequency of a term alone, however, does not necessarily reflect its semantic relation to a warning. Furthermore, we find that some warnings are more easily identified on the lexical level than others. Specifically, high frequency terms in texts with the more "graphic" warnings *Sexual* and *Physical Abuse* are more often also semantically related to their warning than their counterparts for *Emotional Abuse*.

## 8. Supplementary Material

### 8.1. Limitations

Our contributions are limited to fan fiction documents, especially those with warnings for abuse. Hence, the results of our experiments may be limited to this genre and register of text and the specific warnings (categories) we analyzed. Authors of other text types such as news articles or blog posts might behave differently when assigning warnings to their works. In addition, the vocabulary of other text genres and registers might be different, leading to different relative term frequencies and relationships between the measures we calculated. Whether trigger warnings assigned by authors of these documents are aligned with an expected vocabulary needs to be investigated in future work.

Another limitation is that we test the alignment between warnings and documents at the lexical level only. The dictionaries of expected terms do not capture semantics, nor do they capture idiosyncratic, metaphorical, or indirect language use. Our results suggest some forms of warning-related content are more easily captured at the lexical level, while others are apparent through context and semantics. Testing alignment at the semantic level will be explored in future work.

A limitation of our annotation experiment is that the results come from simulated annotations with sociodemographic prompts. An annotation study with human annotators could lead to different results. However, previous work already established the subjectivity of annotating trigger warnings (Wiegmann et al., 2024). As is the case for all prompt-based research, despite our best efforts to capture a wide range of sociodemographic profiles, different profiles or different prompt formulations may yield different results. Beck et al. (2024) conducted their experiments on toxicity detection, stance detection, hatespeech detection, and sentiment classification. While hatespeech detection is related to (emotional) abuse specifically, and trigger warning assignment in general (Wiegmann et al., 2024), sociodemographic prompting has not been verified against human annotations for this task. This is a limitation of our annotation experiment. To minimize this limitation, we analyzed the annotator agreement between the ten different profiles, veri-

fying that the profiles lead to consistent differences in the annotation behavior, thus reflecting the subjectivity we wanted to simulate. The results can be found in Figure 3 and Table 14 in Appendix B.

### 8.2. Ethical Considerations

Trigger warnings are a controversial topic in societal discourse, as the debate about their use in education exemplifies (Dickman-Burnett and Geaman, 2019). Opponents see them as a threat to academic freedom, an invitation to avoid disturbing content, and unnecessary paternalism towards students. Proponents argue that trigger warnings are a way of showing solidarity with marginalized groups and people with trauma. In addition, studies in clinical psychology have found no or even slightly negative effects of trigger warnings (Sanson et al., 2019; Bruce et al., 2021; Bridgland et al., 2019). At present, they do not seem to be helpful in clinical contexts. The results of Bruce et al. indicate in particular that the phrase 'trigger warning' itself elicits more negative reactions than other forms of communicating the same information.

Outside of clinical psychology, where trigger warnings are directly linked to post-traumatic stress disorder (PTSD), the notion of a trigger warning can be considered a social construct, that is, a construct whose meaning and interpretation is based on mutual agreement by people, members of a society. Trigger warnings emerge as a consequence of the social construction of triggers. Boghossian (2001) points out that "social[] construct[ion] ... emphasize[s] ... dependence on contingent aspects of our social selves. It is to say: This thing could not have existed had we not built it; and we need not have built it at all, at least not in its present form." (see also (Galbin, 2014; Diaz-Leon, 2015) for the socio-cultural notion of social construction). Trigger warnings have socially constructed meanings (Helmhout et al., 2009) without what could be considered a reference or "expert" definition.

In this regard it is not surprising that the findings of clinical psychology do not seem to deter communities, such as the one we studied, from continuing to use warning labels on their content. It points to the need for some kind of more detailed meta-information that allows individuals to make more nuanced decisions about their media consumption. Especially on a platform where such content appears frequently, as in the early web forums where trigger warnings emerged (Jones et al., 2020) or in present fan fiction stories where the fringes of literary taboo may be explored.

On a more practical note, a potential negative impact of our work is inherent to the type of texts we studied. As results of our experiments, we collected lists of terms that are (strongly) related to emotional, physical, and sexual abuse. Although these terms may help in assigning warnings to text, under the assumption of triggering content, they may also be used to collect material that causes distress for vulnerable individuals. Their responsible use involves to take measures to prevent that.

## 9. Acknowledgements

This publication has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101070014 (OpenWebSearch.EU, https://doi.org/10.3030/101070014).

## 10. Bibliographical References

Khurshid Ahmad, Lee Gillam, and Lena Tostevin. 1999. University of Surrey Participation in TREC8: Weirdness Indexing for Logical Document Extrapolation and Retrieval (WILDER). In *Proceedings of The Eighth Text REtrieval Conference, TREC 1999, Gaithersburg, Maryland, USA, November 17-19, 1999*, volume 500-246 of *NIST Special Publication*. National Institute of Standards and Technology (NIST).

Tilman Beck, Hendrik Schuff, Anne Lauscher, and Iryna Gurevych. 2024. Sensitivity, Performance, Robustness: Deconstructing the Effect of Sociodemographic Prompting. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2589–2615, St. Julian's, Malta. Association for Computational Linguistics.

Paul A. Boghossian. 2001. What is social construction? *Times Literary Supplement*.

Victoria M. E. Bridgland, Deanne M. Green, Jacinta M. Oulton, and Melanie K. T. Takarangi. 2019. Expecting the Worst: Investigating the Effects of Trigger Warnings on Reactions to Ambiguously Themed Photos. *Journal of Experimental Psychology: Applied*, 25:602–617.

Victoria M. E. Bridgland, Payton J. Jones, and Benjamin W. Bellet. 2023. A Meta-Analysis of the Efficacy of Trigger Warnings, Content Warnings, and Content Notes. *Clinical Psychological Science*, 0.

Madeline J. Bruce, Sara M. Stasik-O'Brien, and Heather Hoffmann. 2021. Students' Psychophysiological Reactivity to Trigger Warnings. *Current Psychology*, 42:5470–5479.

Ashleigh Charles, Laurie Hare-Duke, Hannah Nudds, Donna Franklin, Joy Llewellyn-Beardsley, Stefan Rennick-Egglestone, Onni Gust, Fiona Ng, Elizabeth Evans, Emily Knox, Ellen Townsend, Caroline Yeo, and Mike Slade. 2022. Typology of Content Warnings and Trigger Warnings: Systematic Review. *PLoS One*, 17(5):e0266722.

Esa Diaz-Leon. 2015. What is social construction? *European Journal of Philosophy*, 23(4).

Victoria L. Dickman-Burnett and Maribeth Geaman. 2019. Untangling the Trigger-Warning Debate: Curating a Complete Toolkit for Compassionate Praxis in the Classroom. *Journal of Thought*, 53:35–52.

Ted Dunning. 1993. Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics*, 19(1):61–74.

Norbert Fuhr, Anastasia Giachanou, Gregory Grefenstette, Iryna Gurevych, Andreas Hanselowski, Kalervo Jarvelin, Rosie Jones, Yiqun Liu, Josiane Mothe, Wolfgang Nejdl, Isabella Peters, and Benno Stein. 2017. An Information Nutritional Label for Online Documents. *SIGIR Forum*, 51(3):44–66.

Alexandra Galbin. 2014. An introduction to social constructionism. *Social research reports*, 6(26):82–92.

Ryan J. Gallagher, Andrew J. Reagan, Christopher M. Danforth, and Peter Sheridan Dodds. 2018. Divergent Discourse Between Protests and Counter-Protests: #BlackLivesMatter and #AllLivesMatter. *PLOS ONE*, 13(4):e0195644.

Andrew Hardie. 2014. Log Ratio - An Informal Introduction. Accessed: 2024-07-05.

Martin Helmhout, René J. Jorna, and Henk W. Gazendam. 2009. The semiotic actor: From signs to socially constructed meaning. *Semiotica*, 2009(175).

EunJeong Hwang, Bodhisattwa Majumder, and Niket Tandon. 2023. Aligning language models to user opinions. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5906–5919, Singapore. Association for Computational Linguistics.

Payton J. Jones, Benjamin W. Bellet, and Richard McNally. 2020. Helping or Harming? The Effect of Trigger Warnings on Individuals With Trauma Histories. *Clinical Psychological Science*, 8:905–917.

Dave S. Kerby. 2014. The simple difference formula: An approach to teaching nonparametric correlation. *Comprehensive Psychology*, 3.

Adam Kilgarriff. 2001. Comparing corpora. *International Journal of Corpus Linguistics*, 6(1):97–133.

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

George Kour, Samuel Ackerman, Eitan Daniel Farchi, Orna Raz, Boaz Carmeli, and Ateret Anaby Tavor. 2022. Measuring the Measuring Tools: An Automatic Evaluation of Semantic Metrics for Text Corpora. In *Proceedings of the 2nd Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pages 405–416, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Erich L. Lehmann. 1975. *Nonparametrics. Statistical Methods Based on Ranks*. McGraw-Hill.

Jefrey Lijffijt, Terttu Nevalainen, Tanja Säily, Panagiotis Papapetrou, Kai Puolamäki, and Heikki Mannila. 2014. Significance Testing of Word Frequencies in Corpora. *Digital Scholarship in the Humanities*, 31(2):374–397.

Jianhua Lin. 1991. Divergence Measures Based on the Shannon Entropy. *IEEE Transactions on Information Theory*, 37(1):145–151.

Jinghui Lu, Maeve Henchion, and Brian Mac Namee. 2020. Diverging Divergences: Examining Variants of Jensen Shannon Divergence for Corpus Comparison Tasks. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6740–6744, Marseille, France. European Language Resources Association.

Carolyn R. Miller and Amy J. Devitt, editors. 2019. *Landmark Essays on Rhetorical Genre Studies*. Routledge.

Youngja Park, Siddharth Patwardhan, Karthik Visweswariah, and Stephen C. Gates. 2008. An Empirical Analysis of Word Error Rate and Keyword Error Rate. In *9th Annual Conference of the International Speech Communication Association, INTERSPEECH 2008, Brisbane, Australia, September 22-26, 2008*, pages 2070–2073. ISCA.

Eitan Adam Pechenick, Christopher M. Danforth, and Peter Sheridan Dodds. 2015. Characterizing the Google Books Corpus: Strong Limits to Inferences of Socio-Cultural and Linguistic Evolution. *PLOS ONE*, 10(10):e0137041.

Paul Rayson and Roger Garside. 2000. Comparing corpora using frequency profiling. In *The Workshop on Comparing Corpora*, pages 1–6, Hong Kong, China. Association for Computational Linguistics.

Paul Rottger, Bertie Vidgen, Dirk Hovy, and Janet Pierrehumbert. 2022. Two Contrasting Data Annotation Paradigms for Subjective NLP Tasks. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 175–190, Seattle, United States. Association for Computational Linguistics.

Joni Salminen, Hind Almerekhi, Ahmed Mohamed Kamel, Soon-gyo Jung, and Bernard J. Jansen. 2019. Online Hate Ratings Vary by Extremes: A Statistical Analysis. In *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval*, CHIIR '19, pages 213–217, New York, NY, USA. Association for Computing Machinery.

Mevagh Sanson, Deryn Strange, and Maryanne Garry. 2019. Trigger Warnings Are Trivially Helpful at Reducing Negative Affect, Intrusive Thoughts, and Avoidance. *Clinical Psychological Science*, 7(4):778–793.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. The Risk of Racial Bias in Hate Speech Detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678, Florence, Italy. Association for Computational Linguistics.

Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A. Smith. 2022. Annotators with Attitudes: How Annotator Beliefs And Identities Bias Toxic Language Detection. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5884–5906, Seattle, United States. Association for Computational Linguistics.

Ferdinand Schlatt, Dieter Bettin, Matthias Hagen, Benno Stein, and Martin Potthast. 2022. Mining Health-related Cause-Effect Statements with High Precision at Large Scale. In *29th International Conference on Computational Linguistics (COLING 2022)*, pages 1925–1936. International Committee on Computational Linguistics.

Sidney Siegel. 1956. *Nonparametric Statistics for the Behavioral Sciences*. McGraw-Hill.

Nathalie Wahlsdorf, Tanja Michael, Johanna Lass-Hennemann, and Roxanne Sopp. 2024. Triggerwarnungen: Hilfreich, Wirkungslos – oder Sogar Schädlich? *Psychotherapeutenjournal*, 1:50–56.

Ruyuan Wan, Jaehyung Kim, and Dongyeop Kang. 2023. Everyone's Voice Matters: Quantifying Annotation Disagreement Using Demographic Information. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(12):14523–14530.

Zeerak Waseem. 2016. Are You a Racist or Am I Seeing Things? Annotator Influence on Hate Speech Detection on Twitter. In *Proceedings of the First Workshop on NLP and Computational Social Science*, pages 138–142, Austin, Texas. Association for Computational Linguistics.

Matti Wiegmann, Jennifer Rakete, Magdalena Wolska, Benno Stein, and Martin Potthast. 2024. If there's a Trigger Warning, then where's the Trigger? Investigating Trigger Warnings at the Passage Level.

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 *61th Annual Meeting of the Association for Computational Linguistics (ACL 2023) (Volume 1: Long Papers)*, pages 12113–12134, Toronto, Canada. Association for Computational Linguistics.

Wilson Wong, Wei Liu, and Mohammed Bennamoun. 2007. Determining Termhood for Learning Domain Ontologies using Domain Prevalence and Tendency. In *Sixth Australasian Data Mining Conference (AusDM 2007)*, volume 70 of *CRPIT*, pages 47–54, Gold Coast, Australia. ACS.

## 11. Language Resource References

Wiegmann, Matti and Wolska, Magdalena and Schröder, Christopher and Borchardt, Ole and Stein, Benno and Potthast, Martin. 2022. *Webis Trigger Warning Corpus 2022*. Distributed via Zenodo. PID https://doi.org/10.5281/zenodo.7976807.

# A. Alignment Between Trigger Warnings and Vocabulary

| Term (POS) | $z$ | lr | $\#d$ | Term (POS) | $z$ | lr | $\#d$ | Term (POS) | $z$ | lr | $\#d$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| abandon (V) | 13.57 | 0.08 | 6,507 | confused (A) | 6.68 | -0.08 | 11,570 | enraged (A) | 1.35 | -0.13 | 577 |
| abandoned (A) | 3.80 | 1.22 | 14 | cry (V) | 22.86 | 0.13 | 22,382 | fatigue (N) | 7.71 | 0.18 | 1,100 |
| abandon (V) | 13.57 | 0.08 | 6,507 | confused (A) | 6.68 | -0.08 | 11,570 | enraged (A) | 1.35 | -0.13 | 577 |
| abandoned (A) | 3.80 | 1.22 | 14 | cry (V) | 22.86 | 0.13 | 22,382 | fatigue (N) | 7.71 | 0.18 | 1,100 |
| abandonment (N) | 7.56 | 0.42 | 392 | crybaby (N) | 2.00 | -0.01 | 155 | fear (N) | 33.21 | 0.25 | 17,642 |
| abase (V) | -0.88 | -0.54 | 6 | curse (V) | -3.80 | -0.27 | 6,556 | fearful (A) | 11.03 | 0.22 | 1,585 |
| abased (A) | 0.01 | -0.21 | 1 | cyberstalking (N) | 0.69 | 0.44 | 1 | forbid (V) | 4.40 | -0.03 | 1,707 |
| abasement (N) | -0.15 | -0.32 | 5 | deceive (V) | 6.72 | 0.22 | 725 | forbiddance (N) | -0.09 | -0.37 | 1 |
| accusation (N) | 18.45 | 0.41 | 1,996 | deceiving (A) | -0.60 | -0.70 | 3 | forbidden (A) | 1.98 | -0.03 | 377 |
| accuse (V) | 16.44 | 0.26 | 3,154 | deception (N) | 10.97 | 0.57 | 591 | force (V) | 30.22 | 0.15 | 22,330 |
| accusing (A) | 1.14 | 0.63 | 4 | defamation (N) | 0.56 | 0.21 | 21 | frantic (A) | 13.82 | 0.20 | 3,166 |
| aggress (V) | -0.09 | -0.30 | 3 | defame (V) | 0.42 | 0.09 | 14 | frustrate (V) | 7.09 | 0.07 | 1,741 |
| aggression (N) | 5.68 | 0.09 | 808 | degradation (N) | 5.78 | 0.36 | 203 | frustrated (A) | 8.70 | 0.00 | 4,004 |
| aggressive (A) | 4.31 | -0.05 | 1,984 | degrade (V) | 3.55 | 0.10 | 342 | frustrating (A) | 4.85 | 0.01 | 1,324 |
| aggressor (N) | 1.46 | -0.07 | 101 | degraded (A) | 1.26 | 0.09 | 52 | frustration (N) | 10.42 | 0.02 | 7,049 |
| agitate (V) | 6.51 | 0.26 | 458 | degrading (A) | 5.29 | 0.20 | 335 | fuming (A) | -0.60 | -0.53 | 7 |
| agitated (A) | 7.21 | 0.09 | 1,275 | demean (V) | 5.59 | 0.38 | 165 | furious (A) | 8.35 | -0.01 | 3,958 |
| aloof (A) | 1.92 | -0.04 | 346 | demeaning (A) | 3.73 | 0.27 | 159 | gaslight (V) | 26.38 | 3.15 | 101 |
| aloofness (N) | 3.32 | 0.32 | 82 | denigrate (V) | -0.14 | -0.20 | 16 | gaslighting (A) | 1.89 | 2.33 | 1 |
| anger (N) | 24.24 | 0.17 | 14,672 | denigration (N) | 0.91 | 0.38 | 5 | gaslighting (N) | 38.71 | 3.87 | 137 |
| anger (V) | 4.86 | 0.00 | 1,123 | deny (V) | 10.41 | -0.02 | 7,356 | guilt (N) | 25.64 | 0.26 | 7,532 |
| angered (A) | -0.03 | -0.23 | 206 | depressed (A) | 5.36 | 0.11 | 900 | harass (V) | 4.03 | -0.04 | 883 |
| angry (A) | 21.69 | 0.15 | 15,585 | depression (N) | 10.32 | 0.32 | 1,093 | harasser (N) | -0.66 | -0.66 | 11 |
| annoy (V) | 3.21 | -0.13 | 4,550 | deprivation (N) | 7.77 | 0.39 | 390 | harassment (N) | 2.42 | 0.10 | 298 |
| annoyed (A) | 4.74 | -0.10 | 3,687 | deprive (V) | 4.20 | 0.02 | 908 | harm (N) | 12.70 | 0.12 | 3,829 |
| annoying (A) | 1.90 | -0.17 | 4,043 | deprived (A) | 1.03 | 0.10 | 22 | harm (V) | 7.09 | -0.01 | 2,756 |
| anorexia (N) | 0.41 | 0.28 | 6 | desert (V) | -0.37 | -0.21 | 412 | helpless (A) | 13.39 | 0.17 | 3,665 |
| anorexic (A) | 5.29 | 1.16 | 45 | deserted (A) | -0.43 | -0.24 | 484 | helplessness (N) | 7.78 | 0.27 | 604 |
| anxiety (N) | 29.17 | 0.39 | 6,577 | deserting (A) | 0.98 | 0.43 | 5 | hinder (V) | 1.86 | -0.08 | 527 |
| anxious (A) | 16.72 | 0.22 | 4,599 | desertion (N) | 4.11 | 0.94 | 54 | hindering (A) | 1.83 | 1.61 | 2 |
| apathetic (A) | 5.08 | 0.22 | 325 | detached (A) | 7.50 | 0.53 | 237 | humiliate (V) | 7.70 | 0.08 | 1,511 |
| apathy (N) | 7.51 | 0.54 | 229 | disable (V) | 2.91 | -0.01 | 416 | humiliated (A) | 0.98 | 0.05 | 40 |
| appetite (N) | 10.30 | 0.24 | 1,358 | disabling (A) | 0.87 | 0.70 | 2 | humiliating (A) | 7.20 | 0.12 | 1,042 |
| ashamed (A) | 10.94 | 0.08 | 3,822 | discomfort (N) | 14.45 | 0.18 | 3,518 | humiliation (N) | 7.35 | 0.05 | 1,625 |
| bawl (V) | 7.95 | 0.33 | 519 | disoriented (A) | -1.03 | -0.56 | 17 | hurt (V) | 43.55 | 0.32 | 29,136 |
| belittle (V) | 7.97 | 0.38 | 383 | disregard (N) | 4.73 | 0.20 | 306 | hurting (A) | 2.00 | 1.43 | 3 |
| belittlement (N) | 1.01 | 0.25 | 10 | disregard (V) | 5.97 | 0.14 | 851 | hysteria (N) | 7.19 | 0.48 | 294 |
| berate (V) | 10.20 | 0.25 | 1,123 | disregarded (A) | 1.72 | 0.40 | 21 | ignorance (N) | 6.83 | 0.16 | 882 |
| berating (A) | 2.93 | 1.63 | 5 | disregarding (A) | 3.00 | 2.15 | 3 | ignore (V) | 19.63 | 0.05 | 20,261 |
| blackmail (N) | 1.76 | -0.10 | 358 | disrespect (N) | 6.12 | 0.31 | 395 | ignored (A) | 1.52 | 1.33 | 2 |
| blackmail (V) | 0.66 | -0.22 | 454 | disrespect (V) | 5.64 | 0.18 | 487 | impel (V) | 0.82 | 0.24 | 19 |
| blame (N) | 13.09 | 0.33 | 1,251 | disrespected (A) | -0.08 | -0.30 | 1 | infantilize (V) | 4.41 | 1.10 | 25 |
| blame (V) | 16.70 | 0.10 | 8,779 | distraught (A) | 4.69 | 0.11 | 530 | infantilizing (A) | 0.10 | -0.05 | 1 |
| blamed (A) | 3.97 | 2.78 | 3 | distress (N) | 11.56 | 0.17 | 2,643 | infringe (V) | 1.91 | 0.18 | 52 |
| blaming (A) | 2.83 | 3.33 | 1 | distress (V) | 6.91 | 0.30 | 430 | infringement (N) | -5.36 | -1.78 | 18 |
| bored (A) | 10.08 | 0.07 | 3,819 | distressed (A) | 8.45 | 0.12 | 1,378 | insecure (A) | 1.75 | -0.04 | 480 |
| boredom (N) | 5.47 | 0.07 | 970 | distressing (A) | 5.26 | 0.37 | 163 | insecurity (N) | 5.33 | 0.02 | 1,113 |
| brainwash (N) | 0.66 | 0.13 | 13 | distrust (N) | 4.80 | 0.15 | 475 | insomnia (N) | 5.97 | 0.31 | 312 |
| brainwash (V) | 7.19 | 0.14 | 595 | distrust (V) | 6.23 | 0.35 | 249 | insomniac (N) | 0.18 | -0.18 | 59 |
| brainwashing (N) | 4.02 | 0.14 | 170 | distrustful (A) | 4.15 | 0.27 | 172 | insult (N) | 4.76 | -0.04 | 2,614 |
| bully (N) | 0.07 | -0.22 | 998 | ditch (V) | 1.94 | -0.13 | 943 | insult (V) | 6.30 | 0.02 | 2,298 |
| bully (V) | -0.43 | -0.23 | 1,139 | dread (N) | 23.68 | 0.47 | 3,615 | insulted (A) | 0.53 | -0.10 | 49 |
| catfish (V) | 1.97 | 0.86 | 5 | dread (V) | 7.29 | 0.06 | 1,816 | insulting (A) | 5.35 | 0.19 | 406 |
| coerce (V) | 3.78 | 0.09 | 361 | eavesdrop (V) | 1.59 | -0.09 | 547 | intimidate (V) | 4.06 | -0.08 | 2,726 |
| coerced (A) | -0.41 | -0.52 | 4 | eavesdropping (N) | 3.38 | 0.12 | 239 | intimidated (A) | 1.17 | 0.65 | 4 |
| coercion (N) | 4.93 | 0.44 | 131 | embarrass (V) | 2.38 | -0.08 | 1,808 | intimidating (A) | 2.41 | -0.06 | 602 |
| coercive (A) | 3.03 | 0.81 | 21 | embarrassed (A) | -0.47 | -0.21 | 5,393 | intimidation (N) | 5.82 | 0.23 | 415 |
| compel (V) | 6.45 | 0.02 | 1,047 | embarrassing (A) | 3.66 | -0.10 | 2,876 | invalidate (V) | 7.43 | 0.87 | 102 |
| compelled (A) | -0.28 | -0.44 | 5 | embarrassment (N) | 0.06 | -0.19 | 4,559 | invalidated (A) | 1.49 | 0.63 | 9 |
| condescend (V) | 13.18 | 0.43 | 963 | emotion (N) | 19.37 | 0.09 | 12,607 | invalidation (N) | 2.24 | 1.35 | 4 |
| condescending (A) | 2.36 | 0.39 | 34 | emotional (A) | 18.70 | 0.23 | 4,647 | isolate (V) | 15.16 | 0.52 | 1,071 |
| condescension (N) | 7.51 | 0.53 | 232 | enrage (V) | 1.80 | -0.07 | 519 | isolated (A) | 7.68 | 0.28 | 629 |

Table 8: Emotional abuse dictionary (Part 1). For each term, the columns show the $z$-score from the Mann-Whitney U test on term frequencies, the log ratio, and the number of documents $d \in D_{ci}$, it occurred in. The shading of the $z$ column indicates if the term is significantly more or less frequent in $D_{ci}$.

| Term (POS) | $z$ | lr | $\#d$ | Term (POS) | $z$ | lr | $\#d$ | Term (POS) | $z$ | lr | $\#d$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| isolation (N) | 17.77 | 0.67 | 914 | restless (A) | 6.67 | 0.04 | 1,787 | swear (V) | 5.25 | -0.06 | 11,278 |
| jealous (A) | 2.26 | -0.16 | 3,780 | restlessness (N) | 4.31 | 0.19 | 273 | swearing (N) | -0.46 | -0.21 | 286 |
| jealousy (N) | 2.75 | -0.09 | 1,847 | restrict (V) | 7.07 | 0.18 | 828 | tear (N) | 29.35 | 0.18 | 22,326 |
| leave (V) | 13.97 | 0.04 | 43,020 | restricted (A) | 0.94 | -0.07 | 127 | tearful (A) | 3.72 | 0.10 | 429 |
| loneliness (N) | 11.49 | 0.29 | 1,322 | restricting (A) | 1.42 | 0.45 | 10 | tearfulness (N) | -0.11 | -0.35 | 1 |
| manipulate (V) | 25.16 | 0.49 | 3,057 | restriction (N) | 5.72 | 0.15 | 471 | teary (A) | 5.99 | 0.09 | 1,013 |
| manipulated (A) | 5.63 | 3.02 | 5 | ridicule (N) | 1.07 | -0.04 | 126 | tense (A) | 16.79 | 0.12 | 7,403 |
| manipulating (A) | 1.08 | 1.33 | 1 | ridicule (V) | 1.13 | -0.12 | 222 | tension (N) | 13.12 | 0.07 | 7,528 |
| meltdown (N) | 4.83 | 0.17 | 386 | ridiculed (A) | 1.98 | 1.75 | 2 | terrified (A) | 16.84 | 0.29 | 2,806 |
| mental (A) | 18.63 | 0.21 | 6,188 | ridiculing (A) | 2.83 | 3.33 | 1 | terrorize (V) | 6.88 | 0.29 | 440 |
| mistreatment (N) | 3.76 | 0.23 | 131 | scapegoat (N) | 4.72 | 0.31 | 144 | terrorized (A) | 0.36 | 0.33 | 1 |
| mock (V) | 9.01 | 0.06 | 3,459 | scapegoat (V) | 1.21 | 0.81 | 3 | threat (N) | 17.89 | 0.14 | 7,597 |
| mocked (A) | -0.32 | -0.40 | 7 | scared (A) | 23.22 | 0.19 | 11,922 | threaten (V) | 19.91 | 0.15 | 9,745 |
| mockery (N) | 10.06 | 0.31 | 831 | scold (V) | 5.46 | -0.02 | 3,656 | threatened (A) | 0.94 | -0.05 | 76 |
| mocking (A) | -0.06 | -0.30 | 1 | scolded (A) | -0.06 | -0.25 | 2 | threatening (A) | 6.80 | 0.09 | 1,322 |
| moodiness (N) | 1.20 | 0.09 | 38 | scolding (A) | 0.80 | 0.95 | 1 | toxic (A) | 15.50 | 0.56 | 988 |
| neglect (N) | 5.80 | 0.33 | 335 | scolding (N) | 3.14 | 0.03 | 395 | troll (N) | 6.02 | 0.09 | 529 |
| neglect (V) | 5.83 | 0.03 | 1,468 | scream (N) | 10.33 | 0.01 | 6,688 | troll (V) | 0.30 | -0.41 | 90 |
| neglected (A) | 2.24 | 0.38 | 31 | scream (V) | 17.73 | 0.09 | 17,750 | trust (N) | 15.94 | 0.17 | 4,380 |
| neglecting (A) | 2.63 | 0.70 | 18 | screamed (A) | 1.03 | 0.18 | 22 | trust (V) | 28.57 | 0.20 | 17,425 |
| negligence (N) | 3.25 | 0.25 | 137 | screaming (A) | 0.97 | 0.06 | 39 | tyrannize (V) | -0.56 | -1.09 | 1 |
| nervous (A) | 10.29 | -0.03 | 10,838 | seclude (V) | 1.27 | -0.06 | 273 | ultimatum (N) | 4.01 | 0.15 | 220 |
| nervousness (N) | 2.50 | -0.12 | 1,477 | secluded (A) | 0.74 | -0.12 | 513 | unattended (A) | 4.77 | 0.27 | 255 |
| overbear (V) | 3.31 | 0.15 | 218 | seclusion (N) | 5.22 | 1.03 | 148 | uncomfortable (A) | 15.30 | 0.06 | 9,823 |
| overbearance (N) | 1.26 | 1.56 | 1 | shame (N) | 20.74 | 0.20 | 7,754 | uncommunicative (A) | 1.61 | 0.38 | 15 |
| overbearing (A) | 4.54 | 0.21 | 255 | shame (V) | 5.38 | 0.19 | 537 | uncooperative (A) | 5.43 | 0.46 | 186 |
| overeat (V) | 1.04 | 0.14 | 15 | shamed (A) | 1.36 | 0.20 | 29 | underweight (A) | 5.82 | 0.54 | 125 |
| overweight (A) | 0.46 | -0.10 | 91 | shaming (A) | 2.02 | 2.48 | 1 | underweight (N) | 3.23 | 0.81 | 25 |
| overweight (N) | 0.30 | -0.08 | 9 | shaming (N) | 1.81 | 0.33 | 25 | unease (N) | 15.02 | 0.39 | 1,533 |
| panic (A) | 1.90 | 0.65 | 10 | silence (V) | 7.02 | 0.01 | 2,916 | uneasy (A) | 10.44 | 0.14 | 2,580 |
| panic (N) | 27.15 | 0.26 | 10,139 | silenced (A) | -0.44 | -0.84 | 1 | unresponsive (A) | 5.11 | 0.16 | 500 |
| panic (V) | 9.96 | 0.00 | 6,204 | slander (N) | 1.13 | -0.03 | 72 | unsettled (A) | 9.56 | 0.37 | 655 |
| panicked (A) | 9.80 | 0.17 | 1,759 | slander (V) | 3.97 | 0.44 | 109 | upset (V) | 12.59 | 0.50 | 745 |
| patronising (A) | 2.54 | 0.40 | 42 | slandered (A) | 0.52 | 0.56 | 1 | upsetting (A) | 9.29 | 0.44 | 484 |
| patronize (V) | 8.73 | 0.46 | 393 | sleepless (A) | 5.56 | 0.21 | 417 | verbal (A) | 5.28 | 0.04 | 1,183 |
| patronizing (A) | 5.45 | 0.48 | 144 | sleeplessness (N) | 1.15 | 0.17 | 35 | voiceless (A) | 1.55 | 0.12 | 42 |
| powerless (A) | 7.41 | 0.17 | 982 | snivel (V) | 6.85 | 0.36 | 297 | weep (V) | 12.35 | 0.26 | 1,993 |
| pressure (V) | 7.19 | 0.17 | 881 | sniveling (A) | 3.42 | 2.43 | 3 | whimper (V) | 5.44 | -0.05 | 4,698 |
| pressured (A) | 1.72 | -0.01 | 144 | snoop (N) | 1.79 | 0.04 | 47 | whimpering (A) | -1.09 | -0.54 | 23 |
| prevent (V) | 4.13 | -0.14 | 4,571 | snoop (V) | 5.72 | 0.19 | 540 | withdraw (V) | 6.73 | -0.06 | 2,938 |
| privacy (N) | 8.19 | 0.02 | 3,358 | sob (V) | 24.29 | 0.30 | 7,357 | withdrawal (N) | 6.50 | 0.33 | 487 |
| prohibit (V) | 0.52 | -0.12 | 173 | sobbing (A) | 0.50 | 0.06 | 8 | withdrawn (A) | 2.98 | 0.44 | 45 |
| prohibited (A) | 3.42 | 2.43 | 3 | solitude (N) | 5.99 | 0.22 | 630 | withheld (A) | -0.01 | -0.25 | 1 |
| psychological (A) | 12.81 | 0.55 | 733 | stalk (V) | 5.23 | -0.03 | 3,509 | withhold (V) | 7.24 | 0.23 | 642 |
| punish (V) | 23.04 | 0.35 | 5,011 | stalker (N) | 2.92 | -0.02 | 555 | withholding (A) | 1.35 | 1.16 | 2 |
| punishment (N) | 21.33 | 0.28 | 4,811 | stalking (A) | 2.89 | 0.58 | 28 | withholding (N) | 1.01 | 0.08 | 34 |
| rage (N) | 21.03 | 0.20 | 7,326 | stalking (N) | 2.38 | 0.27 | 72 | worthless (A) | 20.46 | 0.52 | 1,931 |
| rage (V) | 1.73 | -0.14 | 2,062 | stress (N) | 19.73 | 0.26 | 4,495 | worthlessness (N) | 4.25 | 0.56 | 61 |
| reclusive (A) | 1.98 | 0.13 | 111 | stress (V) | 11.90 | 0.13 | 3,108 | yell (N) | 2.22 | -0.13 | 1,383 |
| reject (V) | 7.45 | 0.05 | 2,491 | stressed (A) | 4.15 | 0.09 | 421 | yell (V) | 1.76 | -0.16 | 12,159 |
| rejection (N) | 6.44 | 0.08 | 1,193 | stressful (A) | 6.82 | 0.13 | 1,098 | yelled (A) | -0.62 | -0.24 | 370 |
| repress (V) | 7.12 | 0.22 | 842 | stressing (A) | 2.07 | 0.86 | 7 | yelling (A) | 3.13 | 0.24 | 111 |
| repressed (A) | 5.53 | 0.26 | 289 | suppress (V) | 8.10 | 0.01 | 3,367 | yelling (N) | 6.53 | 0.17 | 776 |
| repressing (A) | 0.55 | 0.56 | 1 | suppressed (A) | 1.76 | 0.17 | 54 | — | — | — | — |
| repression (N) | 3.52 | 0.33 | 79 | suppression (N) | 0.27 | -0.17 | 80 | — | — | — | — |

Table 9: Emotional abuse dictionary (Part 2). For each term, the columns show the $z$-score from the Mann-Whitney U test on term frequencies, the log ratio, and the number of documents $d \in D_{ci}$, it occurred in. The shading of the $z$ column indicates if the term is significantly more or less frequent in $D_{ci}$.

| Term (POS) | $z$ | lr | $\#d$ | Term (POS) | $z$ | lr | $\#d$ | Term (POS) | $z$ | lr | $\#d$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| abrasion (N) | 5.62 | 0.39 | 334 | chafed (A) | 2.01 | 0.61 | 19 | flinch (N) | 14.53 | 0.29 | 3,797 |
| aggress (V) | 0.75 | 0.41 | 10 | chain (N) | 7.52 | 0.09 | 9,466 | flinch (V) | 43.61 | 0.36 | 25,009 |
| aggression (N) | 4.03 | 0.12 | 1,823 | chain (V) | 13.58 | 0.33 | 3,448 | fracture (N) | 11.38 | 0.50 | 1,003 |
| aggressive (A) | 3.26 | 0.03 | 4,724 | chained (A) | -0.03 | -0.12 | 13 | fracture (V) | 5.58 | 0.23 | 988 |
| aggressor (N) | 0.18 | -0.03 | 222 | chasten (V) | 0.47 | 0.03 | 58 | fractured (A) | 10.68 | 0.48 | 1,091 |
| anxiety (N) | 32.81 | 0.37 | 14,990 | chastened (A) | -3.00 | -0.59 | 68 | frail (A) | 2.27 | 0.27 | 98 |
| assault (N) | 5.43 | 0.06 | 4,432 | chastise (V) | 0.74 | -0.03 | 2,268 | frightened (A) | 19.27 | 0.32 | 6,034 |
| assault (V) | 9.01 | 0.17 | 3,900 | chastised (A) | -1.03 | -0.33 | 30 | gaping (A) | -2.43 | -0.29 | 206 |
| assaulted (A) | 0.86 | 0.07 | 153 | chastisement (N) | 2.95 | 0.32 | 163 | gash (N) | 11.85 | 0.30 | 2,803 |
| assaulter (N) | 1.54 | 0.21 | 85 | choke (V) | 16.54 | 0.11 | 25,539 | gash (V) | 5.66 | 0.70 | 120 |
| attack (N) | 6.97 | -0.02 | 19,110 | choked (A) | 0.78 | 0.01 | 284 | gashed (A) | 3.66 | 1.24 | 19 |
| attack (V) | -6.20 | -0.14 | 15,968 | clobber (V) | -0.57 | -0.19 | 110 | gnaw (V) | 7.28 | 0.15 | 2, 981 |
| attacked (V) | -0.10 | -0.12 | 10 | confine (N) | -3.96 | -0.19 | 1,716 | gnawed (A) | 1.30 | 1.84 | 1 |
| attacker (N) | -3.86 | -0.26 | 2,050 | confine (V) | 1.64 | 0.03 | 1,915 | handcuff (N) | 6.03 | 0.14 | 2, 049 |
| bash (N) | 2.26 | 0.17 | 164 | confinement (N) | 0.32 | -0.05 | 681 | handcuff (V) | 7.10 | 0.28 | 1,044 |
| bash (V) | 5.37 | 0.14 | 1,717 | confining (A) | 0.97 | 0.18 | 54 | handcuffed (A) | 0.10 | -0.05 | 39 |
| bashed (A) | 0.18 | 0.08 | 3 | constrain (V) | -2.62 | -0.32 | 218 | harm (N) | 7.19 | 0.06 | 8,456 |
| basher (N) | -0.11 | -0.36 | 10 | constrained (A) | -1.53 | -0.42 | 41 | harm (V) | 8.81 | 0.10 | 6,695 |
| batter (V) | 12.36 | 0.27 | 3,522 | constraining (A) | 0.32 | 0.43 | 1 | harmer (N) | 1.02 | 0.75 | 7 |
| battered (A) | 8.48 | 0.42 | 760 | contuse (V) | 0.79 | 0.29 | 8 | harmful (A) | 3.39 | 0.09 | 787 |
| batterer (N) | 0.60 | 0.28 | 3 | contusion (N) | 8.94 | 0.76 | 250 | hematoma (N) | 4.31 | 0.69 | 48 |
| battering (N) | -0.73 | -0.04 | 136 | cowed (A) | -1.32 | -0.91 | 5 | hit (A) | -0.44 | -0.21 | 41 |
| beat (V) | 47.30 | 0.35 | 36,837 | crack (V) | 19.37 | 0.14 | 23,910 | hit (V) | 32.28 | 0.16 | 56,422 |
| beaten (A) | 0.22 | 0.14 | 2 | cracked (A) | 2.38 | 0.17 | 268 | hitter (N) | -3.24 | -0.60 | 104 |
| beating (N) | 46.53 | 0.93 | 6,448 | cuff (N) | 3.44 | -0.02 | 4,326 | hitting (A) | 2.82 | 1.70 | 7 |
| biff (N) | 0.08 | -0.60 | 2 | cuff (V) | 6.13 | 0.15 | 2,143 | hungry (A) | 8.64 | 0.04 | 17,149 |
| biff (V) | 0.16 | -0.03 | 8 | cuffed (A) | 0.53 | 0.32 | 4 | imprison (V) | -0.61 | -0.02 | 1,490 |
| bind (N) | -3.19 | -0.30 | 629 | cut (N) | 32.96 | 0.55 | 14,971 | imprisoned (A) | -0.97 | 0.24 | 16 |
| bind (V) | -11.05 | -0.25 | 10,978 | cut (V) | 15.25 | 0.05 | 47,797 | incarcerate (V) | -0.09 | -0.10 | 195 |
| binding (A) | -1.54 | -0.29 | 89 | cutter (N) | 2.26 | 0.15 | 431 | incarcerated (A) | 0.47 | 0.17 | 9 |
| bite (N) | 2.22 | -0.02 | 16,694 | cutting (N) | 4.79 | 0.13 | 1,257 | incarceration (N) | 2.46 | 0.19 | 210 |
| bite (V) | -0.60 | -0.09 | 27,723 | deficient (A) | -1.28 | -0.33 | 70 | infect (V) | 4.94 | 0.12 | 2,304 |
| biting (A) | -2.01 | -0.15 | 994 | denial (N) | -6.66 | -0.26 | 2,597 | infected (A) | 1.18 | 0.04 | 446 |
| bitten (A) | -0.36 | -0.16 | 44 | deny (V) | -2.43 | -0.10 | 16,120 | infecting (A) | -0.66 | -1.00 | 1 |
| bleed (V) | 24.83 | 0.24 | 16,613 | depression (N) | 15.87 | 0.43 | 2,706 | infection (N) | 8.78 | 0.22 | 2,186 |
| bleeding (A) | 2.58 | 0.45 | 62 | deprivation (N) | -2.71 | -0.22 | 596 | injure (V) | 9.67 | 0.08 | 9,919 |
| blood (N) | 23.51 | 0.08 | 49,665 | deprive (V) | -1.15 | -0.09 | 1,936 | injured (A) | 7.84 | 0.33 | 1,445 |
| bone (N) | 13.66 | 0.16 | 22,663 | deprived (A) | 0.71 | 0.07 | 49 | injuring (A) | 0.21 | 0.26 | 1 |
| bound (A) | -5.07 | -0.55 | 188 | detain (V) | -0.83 | -0.10 | 510 | injury (N) | 25.66 | 0.29 | 14,002 |
| break (V) | 30.48 | 0.11 | 72,670 | detainment (N) | -0.30 | -0.07 | 26 | intoxicate (V) | -4.70 | -0.20 | 2,437 |
| broken (A) | 32.81 | 0.33 | 18,904 | diminished (A) | -0.86 | -0.18 | 74 | intoxicated (A) | 5.72 | 0.26 | 729 |
| bruise (N) | 80.77 | 0.88 | 22,468 | disciplinary (A) | -1.08 | -0.18 | 249 | intoxicating (A) | -2.00 | -0.28 | 184 |
| bruise (V) | 36.46 | 0.47 | 12,288 | discipline (V) | 11.26 | 0.45 | 1,207 | intoxication (N) | 1.12 | 0.04 | 349 |
| bruised (A) | 28.85 | 0.75 | 3,119 | discolor (V) | 3.70 | 0.32 | 223 | irritability (N) | 1.34 | 0.08 | 146 |
| bruising (A) | 1.96 | 0.50 | 32 | discoloration (N) | 9.18 | 0.72 | 313 | irritated (A) | 3.41 | 0.00 | 6,734 |
| bruising (N) | 25.27 | 0.80 | 2,081 | discolored (A) | 5.00 | 0.44 | 235 | isolate (V) | 5.65 | 0.16 | 1,934 |
| burn (N) | 9.11 | 0.15 | 8,428 | disjoin (V) | -1.88 | -2.48 | 1 | isolated (A) | -1.07 | -0.09 | 1,131 |
| burn (V) | 10.21 | 0.03 | 40,540 | disjointed (A) | -1.08 | -0.19 | 160 | isolation (N) | 3.97 | 0.15 | 1,431 |
| burned (A) | -0.53 | -0.27 | 23 | dislocate (V) | 9.17 | 0.43 | 1,006 | jab (N) | -3.89 | -0.22 | 1,923 |
| burning (A) | -1.82 | -0.94 | 9 | dislocated (A) | 7.10 | 0.56 | 331 | jab (V) | -3.21 | -0.15 | 2,665 |
| burning (N) | 3.63 | 0.04 | 4,334 | drug (N) | 24.28 | 0.31 | 10,793 | kick (N) | 11.44 | 0.14 | 7,416 |
| cane (N) | 4.97 | 0.21 | 2,150 | drug (V) | 5.75 | 0.11 | 3,412 | kick (V) | 20.21 | 0.12 | 34,949 |
| cane (V) | 4.08 | 0.20 | 193 | drugged (A) | 0.69 | 0.00 | 207 | kicked (A) | 1.67 | 1.14 | 4 |
| caning (N) | 3.12 | 0.79 | 25 | drugging (A) | 0.00 | -0.07 | 4 | kicking (A) | 0.26 | 0.22 | 2 |
| captive (A) | 0.56 | -0.05 | 1,782 | drugging (N) | 2.04 | 0.43 | 32 | kicking (N) | 6.09 | 0.13 | 2,314 |
| captive (N) | -1.86 | -0.14 | 1,027 | emaciated (A) | 1.42 | 0.06 | 177 | knife (N) | 13.07 | 0.12 | 17,052 |
| captivity (N) | 2.77 | 0.08 | 818 | fatten (V) | -2.42 | -0.32 | 210 | lacerate (V) | 1.86 | 0.24 | 94 |
| captor (N) | 8.29 | 0.16 | 2,470 | fattening (N) | -0.77 | -0.41 | 11 | lacerated (A) | 1.37 | 0.16 | 59 |
| capture (V) | -9.69 | -0.19 | 9,354 | fearful (A) | 14.75 | 0.31 | 3,810 | laceration (N) | 11.69 | 0.61 | 779 |
| chafe (V) | -3.15 | -0.30 | 497 | fester (V) | 5.38 | 0.20 | 999 | lesion (N) | 1.13 | 0.13 | 82 |

Table 10: Physical abuse dictionary (Part 1). For each term, the columns show the $z$-score from the Mann-Whitney U test on term frequencies, the log ratio, and the number of documents $d \in D_{ci}$, it occurred in. The shading of the $z$ column indicates if the term is significantly more or less frequent in $D_{ci}$.

| Term (POS) | $z$ | lr | $\#d$ | Term (POS) | $z$ | lr | $\#d$ | Term (POS) | $z$ | lr | $\#d$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| lock (V) | 24.60 | 0.15 | 36,647 | scalded (A) | 1.49 | 0.30 | 42 | strap (V) | -0.21 | -0.05 | 2,876 |
| locked (A) | 4.18 | 0.07 | 2,501 | scalding (A) | 2.47 | 2.01 | 3 | strapped (A) | -1.12 | -0.09 | 1,354 |
| malnourished (A) | 9.28 | 0.52 | 571 | scar (N) | 29.23 | 0.32 | 16,928 | strike (V) | 1.46 | -0.04 | 19,036 |
| malnourishment (N) | 5.43 | 0.67 | 136 | scar (V) | 11.58 | 0.40 | 1,648 | struck (A) | -0.35 | -0.27 | 6 |
| malnutrition (N) | 5.11 | 0.45 | 282 | scared (A) | 48.93 | 0.42 | 31,080 | swell (N) | -5.73 | -0.25 | 2,526 |
| manhandle (V) | -4.93 | -0.26 | 1,334 | scarred (A) | 4.90 | 0.04 | 3,620 | swell (V) | 3.73 | -0.01 | 6,873 |
| manhandled (A) | 0.28 | 0.09 | 8 | scarring (A) | 0.24 | 0.09 | 4 | swelling (A) | 0.98 | 0.84 | 2 |
| mutilate (V) | 4.62 | 0.25 | 637 | scarring (N) | 6.28 | 0.37 | 560 | swelling (N) | 16.77 | 0.55 | 1,647 |
| mutilated (A) | 3.72 | 0.23 | 395 | sedate (V) | 7.79 | 0.32 | 1,067 | swollen (A) | 8.82 | 0.06 | 7,924 |
| mutilation (N) | 1.69 | 0.07 | 185 | sedated (A) | 0.90 | 0.09 | 91 | thrash (N) | -0.48 | -0.10 | 263 |
| nervous (A) | 14.87 | 0.10 | 26,604 | sedation (N) | 2.95 | 0.30 | 227 | thrash (V) | 3.93 | 0.06 | 4,202 |
| numb (A) | 24.51 | 0.35 | 8,637 | shackle (N) | 6.20 | 0.17 | 1,411 | thrashed (A) | -0.45 | -0.70 | 1 |
| numb (V) | 10.41 | 0.25 | 2,951 | shackle (V) | 3.31 | 0.10 | 821 | thrashing (A) | 1.07 | 0.34 | 16 |
| overdose (N) | 6.11 | 0.33 | 474 | shackled (A) | 3.12 | 0.61 | 38 | thrashing (N) | 3.93 | 0.21 | 660 |
| overdose (V) | 9.32 | 0.63 | 524 | shattered (A) | 3.65 | 0.13 | 993 | thump (N) | -2.00 | -0.10 | 2,958 |
| overdosed (A) | 0.63 | 0.15 | 13 | shiner (N) | 9.56 | 0.75 | 279 | thump (V) | 2.05 | 4.07 | 3,692 |
| overfed (A) | -1.15 | -1.23 | 2 | shove (N) | 3.56 | 0.05 | 2,624 | tie (V) | -0.60 | -0.09 | 18,565 |
| overfeed (V) | -0.75 | -0.46 | 12 | shove (V) | 15.65 | 0.08 | 33,602 | tied (A) | -1.28 | -1.82 | 1 |
| overmedicate (V) | 3.26 | 1.65 | 7 | skeletal (A) | -4.79 | -0.35 | 715 | trap (N) | -11.27 | -0.33 | 5,181 |
| overmedication (N) | 3.17 | 3.10 | 2 | slap (N) | 15.17 | 0.17 | 6,788 | trap (V) | -0.86 | -0.08 | 13,658 |
| pierce (V) | 4.92 | 0.01 | 10,725 | slap (V) | 10.68 | 0.05 | 17,961 | trapped (A) | 5.55 | 0.38 | 352 |
| pierced (A) | -0.35 | -0.07 | 201 | slapped (A) | 1.01 | 1.43 | 1 | trapping (A) | -0.59 | -0.93 | 1 |
| piercing (A) | 0.14 | 0.01 | 10 | slapping (N) | -7.70 | -0.47 | 710 | ulcer (N) | -0.18 | 0.12 | 102 |
| prison (N) | 14.41 | 0.17 | 8,057 | slash (N) | -3.59 | -0.07 | 1,586 | underfed (A) | 1.54 | 0.19 | 111 |
| prisoner (N) | 2.85 | 0.01 | 4,915 | slash (V) | -1.75 | -0.11 | 2,727 | underfeed (V) | 1.49 | 0.46 | 22 |
| pull (N) | -7.84 | -0.24 | 4,252 | slashed (A) | 1.31 | 0.30 | 22 | unease (N) | 1.20 | -0.03 | 2,703 |
| pull (V) | 22.83 | 0.05 | 95,016 | slice (N) | 8.70 | 0.15 | 4,719 | uneasy (A) | 10.17 | 0.15 | 5,995 |
| pummel (V) | -0.31 | -0.09 | 947 | slice (V) | 2.87 | -0.02 | 5,352 | unhealed (A) | 2.81 | 0.41 | 81 |
| pummeled (A) | 1.38 | 0.97 | 4 | sliced (A) | 1.77 | 0.12 | 209 | uppercut (N) | -1.74 | -0.22 | 176 |
| punch (N) | 15.76 | 0.17 | 9,687 | slit (V) | 3.65 | 0.10 | 1,592 | violate (V) | 5.70 | 0.16 | 2,451 |
| punch (V) | 19.06 | 0.17 | 18,057 | smack (N) | -0.17 | -0.14 | 2,882 | violated (A) | -0.29 | -0.24 | 6 |
| punched (A) | -0.85 | -0.53 | 11 | smack (V) | 3.75 | -0.01 | 10,543 | violator (N) | 1.45 | 0.23 | 34 |
| punching (N) | 7.81 | 0.32 | 818 | smacked (A) | -1.23 | -0.56 | 14 | violence (N) | 18.41 | 0.31 | 6,681 |
| puncture (N) | -1.15 | -0.22 | 300 | smacking (A) | -0.62 | -0.19 | 26 | violent (A) | 17.67 | 0.22 | 9,093 |
| puncture (V) | 6.24 | 0.26 | 1,056 | smash (V) | 7.33 | 0.06 | 7,345 | welt (N) | 18.18 | 0.71 | 1,429 |
| punctured (A) | 2.50 | 0.27 | 168 | smashed (A) | 0.53 | 0.11 | 18 | welt (V) | 4.54 | 0.45 | 197 |
| punish (V) | 30.72 | 0.43 | 11,963 | snap (V) | 9.31 | 0.01 | 43,958 | whack (N) | -3.25 | -0.28 | 385 |
| punishing (A) | -2.55 | -0.52 | 66 | snapped (A) | 0.66 | 0.24 | 11 | whack (V) | -0.51 | -0.07 | 1,148 |
| punishment (N) | 33.40 | 0.42 | 11,946 | spank (N) | -6.56 | -0.76 | 327 | whip (N) | 7.66 | 0.21 | 3,652 |
| punitive (A) | -1.28 | -0.43 | 33 | spank (V) | -9.45 | -0.41 | 1,431 | whip (V) | 6.45 | 0.05 | 11,913 |
| push (N) | -9.71 | -0.25 | 4,977 | spanking (N) | -6.20 | -0.27 | 615 | whipped (A) | 0.39 | 0.23 | 12 |
| push (V) | 7.64 | -0.04 | 70,706 | splinter (V) | 1.43 | 0.02 | 1,181 | whipping (A) | 2.01 | 0.26 | 104 |
| pushed (A) | 2.19 | 1.17 | 7 | splintered (A) | 0.74 | 0.01 | 323 | whipping (N) | 10.77 | 0.65 | 499 |
| restless (A) | 1.52 | -0.03 | 3,927 | stab (N) | 1.03 | -0.04 | 2,887 | withdraw (V) | -1.76 | -0.08 | 6,381 |
| restlessness (N) | -2.17 | -0.15 | 481 | stab (V) | 7.05 | 0.05 | 9,135 | withdrawal (N) | 7.08 | 0.40 | 1,123 |
| restrain (V) | -0.30 | -0.05 | 5,557 | stabbed (A) | 0.28 | 0.10 | 5 | withheld (A) | 1.58 | 0.94 | 5 |
| restrained (A) | -2.56 | -0.20 | 848 | stabber (N) | 1.53 | 0.58 | 15 | withhold (V) | 0.21 | -0.05 | 1,222 |
| restraining (A) | 1.26 | 1.26 | 2 | stabbing (A) | 3.30 | 0.41 | 112 | wound (N) | 25.24 | 0.19 | 19,800 |
| restraint (N) | -3.13 | -0.11 | 4,644 | stabbing (N) | 2.18 | 0.07 | 525 | wound (V) | -0.91 | -0.06 | 5,470 |
| rip (N) | 0.04 | -0.06 | 1,609 | starvation (N) | 4.86 | 0.18 | 810 | wounded (A) | 2.19 | 0.04 | 1,416 |
| rip (V) | 10.74 | 0.08 | 23,300 | starve (V) | 14.11 | 0.19 | 8,044 | wrest (N) | 0.91 | 0.84 | 2 |
| rope (N) | -5.18 | -0.18 | 6,371 | starved (A) | -3.15 | -0.54 | 87 | wrest (V) | -3.30 | -0.38 | 186 |
| rope (V) | -0.94 | -0.08 | 809 | stomp (N) | 0.70 | -0.01 | 697 | yank (N) | 1.44 | -0.02 | 1,183 |
| roughhouse (V) | 1.25 | 0.08 | 145 | stomp (V) | 4.41 | 0.02 | 6,889 | yank (V) | 15.18 | 0.16 | 15,378 |
| scald (N) | -0.53 | -0.14 | 19 | strangle (V) | -0.04 | -0.07 | 5,665 | — | — | — | — |
| scald (V) | 2.45 | 0.03 | 1,328 | strangled (A) | -1.57 | -0.14 | 1,113 | — | — | — | — |

Table 11: Physical abuse dictionary (Part 2). For each term, the columns show the $z$-score from the Mann-Whitney U test on term frequencies, the log ratio, and the number of documents $d \in D_{ci}$, it occurred in. The shading of the $z$ column indicates if the term is significantly more or less frequent in $D_{ci}$.

| Term (POS) | $z$ | lr | $\#d$ | Term (POS) | $z$ | lr | $\#d$ | Term (POS) | $z$ | lr | $\#d$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| aids (N) | -0.08 | -0.21 | 1 | exploitation (N) | 6.04 | 0.70 | 138 | itch (N) | -0.81 | -0.13 | 1,636 |
| anal (A) | 8.47 | -0.02 | 1,005 | exploited (A) | -0.30 | -0.52 | 1 | itching (N) | 1.48 | 0.01 | 576 |
| anus (N) | 9.79 | 0.33 | 701 | expose (V) | 4.71 | -0.04 | 14,515 | itchy (A) | 3.61 | 0.11 | 1,053 |
| anxiety (N) | 34.89 | 0.43 | 13,299 | exposed (A) | 2.76 | 0.11 | 416 | jerk (V) | 1.45 | -0.07 | 15,245 |
| asphyxiate (V) | 1.55 | 0.18 | 130 | exposure (N) | 0.81 | -0.05 | 1,139 | leak (V) | -3.23 | -0.19 | 7,109 |
| ass (N) | -0.49 | -0.18 | 27,608 | eyeball (V) | 1.82 | 0.12 | 286 | lesion (N) | 2.35 | 0.44 | 80 |
| assault (N) | 17.51 | 0.33 | 4,484 | fap (V) | 1.32 | 0.56 | 11 | libido (N) | -1.75 | -0.23 | 526 |
| assault (V) | 20.07 | 0.43 | 3,937 | fear (N) | 42.33 | 0.26 | 36,486 | loneliness (N) | 4.92 | 0.06 | 2,319 |
| assaulted (A) | 7.98 | 0.73 | 208 | fearful (A) | 14.27 | 0.29 | 3,300 | lonely (A) | 6.31 | 0.03 | 7,306 |
| avoid (V) | 2.25 | -0.08 | 26,172 | fingermark (N) | 2.74 | 0.68 | 22 | lurk (V) | -0.12 | -0.09 | 2,689 |
| avoidance (N) | 1.71 | 0.07 | 455 | flash (V) | -4.22 | -0.11 | 15,988 | masturbate (V) | 5.83 | 0.07 | 1,260 |
| bleed (V) | 25.53 | 0.25 | 14,549 | flirtation (N) | 1.90 | 0.07 | 411 | masturbation (N) | 0.27 | -0.12 | 349 |
| bleeding (N) | 14.46 | 0.25 | 3,860 | force (V) | 36.93 | 0.21 | 45,796 | mistrust (N) | 0.90 | 0.02 | 359 |
| blood (N) | 20.26 | 0.07 | 42,683 | forced (A) | 2.23 | 0.05 | 713 | molest (V) | 40.85 | 1.57 | 1,426 |
| bloodstained (A) | 0.60 | -0.04 | 316 | forgetful (A) | -0.06 | -0.09 | 242 | molestation (N) | 20.06 | 1.96 | 232 |
| boundary (N) | 11.29 | 0.24 | 3,499 | genital (N) | 17.43 | 0.75 | 936 | molested (A) | -0.41 | -0.67 | 1 |
| breast (N) | -1.77 | -0.22 | 8,151 | gonorrhea (N) | 6.58 | 1.94 | 31 | neck (N) | 15.51 | 0.03 | 50,602 |
| bruise (N) | 40.85 | 0.45 | 16,033 | groom (V) | 9.90 | 0.34 | 1,513 | nightmares (N) | 0.42 | 0.02 | 4 |
| bruised (A) | 16.37 | 0.45 | 2,267 | groomed (A) | 0.93 | 0.10 | 51 | nonconsensual (A) | 1.63 | 0.34 | 32 |
| bruising (N) | 11.83 | 0.40 | 1,416 | grooming (N) | 5.25 | 0.51 | 236 | objectify (V) | 3.77 | 0.35 | 133 |
| brutalize (V) | 8.74 | 0.65 | 312 | grope (N) | -1.91 | -0.32 | 177 | obscene (A) | -5.62 | -0.33 | 1,371 |
| catcall (V) | 0.75 | 0.09 | 87 | grope (V) | 5.11 | -0.02 | 2,799 | offender (N) | 5.06 | 0.22 | 464 |
| celibate (A) | 1.71 | 0.11 | 123 | groped (A) | -0.68 | -0.43 | 9 | ooze (V) | -0.04 | -0.11 | 2,260 |
| chlamydia (N) | 6.18 | 1.33 | 36 | groping (A) | -0.01 | -0.11 | 1 | outflow (N) | 0.81 | 0.10 | 13 |
| choke (V) | 17.89 | 0.13 | 22,212 | groping (N) | 3.12 | 0.11 | 398 | overpower (V) | 3.96 | 0.03 | 2,626 |
| coerce (V) | 5.72 | 0.24 | 776 | harass (V) | 4.17 | 0.04 | 1,824 | pedophile (N) | 30.76 | 1.65 | 666 |
| coerced (A) | 2.00 | 0.59 | 17 | harasser (N) | 0.40 | 0.18 | 31 | penetrate (V) | 5.33 | 0.09 | 3,034 |
| coercion (N) | 4.75 | 0.35 | 246 | harassment (N) | 4.69 | 0.17 | 665 | penetration (N) | 7.23 | 0.26 | 926 |
| coercive (A) | 1.60 | 0.34 | 31 | helpless (A) | 13.73 | 0.15 | 7,378 | penis (N) | 20.59 | 0.47 | 2,747 |
| concentration (N) | -6.60 | -0.26 | 3,462 | hematoma (N) | 1.68 | 0.13 | 30 | perverse (A) | 6.72 | 0.31 | 757 |
| consensual (A) | 22.28 | 0.98 | 982 | hiv (N) | 3.51 | 4.65 | 1 | perversion (N) | 7.83 | 0.46 | 385 |
| consent (N) | 21.09 | 0.52 | 3,397 | humiliate (V) | 14.80 | 0.28 | 3,397 | pervert (N) | 12.22 | 0.39 | 1,905 |
| consent (V) | 18.02 | 0.66 | 1,536 | humiliated (A) | 1.61 | 0.14 | 87 | perverted (A) | 6.55 | 0.31 | 615 |
| consenting (A) | 1.49 | 0.76 | 6 | humiliating (A) | 10.00 | 0.24 | 2,199 | predator (N) | 4.55 | 0.01 | 2,590 |
| creepy (A) | -3.11 | -0.19 | 3,592 | humiliation (N) | 14.79 | 0.24 | 3,665 | pregnant (A) | 23.06 | 0.42 | 7,035 |
| defile (V) | 6.14 | 0.24 | 720 | hygiene (N) | 4.43 | 0.22 | 544 | pressure (V) | 12.58 | 0.37 | 1,963 |
| degradation (N) | 6.33 | 0.33 | 396 | hypersexuality (N) | 5.71 | 1.67 | 20 | pressured (A) | 7.13 | 0.44 | 382 |
| degrade (V) | 7.80 | 0.32 | 792 | impregnate (V) | 4.15 | 0.18 | 535 | promiscuity (N) | 3.24 | 0.50 | 77 |
| degraded (A) | 3.88 | 0.44 | 130 | impregnation (N) | 2.44 | 0.35 | 48 | promiscuous (A) | 9.24 | 0.87 | 255 |
| degrading (A) | 8.15 | 0.35 | 723 | inappropriate (A) | 4.42 | 0.05 | 2,507 | pussy (N) | -6.44 | -0.53 | 4,049 |
| depredate (V) | 1.70 | 2.23 | 2 | inattentive (A) | 0.32 | -0.04 | 81 | ram (V) | -0.90 | -0.12 | 2,134 |
| depredation (N) | 0.06 | 0.18 | 9 | incapacity (N) | -0.26 | -0.23 | 20 | rape (N) | 69.73 | 1.48 | 4,679 |
| depressed (A) | 14.25 | 0.36 | 2,184 | incest (N) | 6.26 | 0.38 | 335 | rape (V) | 105.72 | 1.62 | 9,487 |
| depression (N) | 21.96 | 0.58 | 2,623 | incestuous (A) | 2.38 | 0.13 | 161 | raped (A) | 1.11 | 0.95 | 2 |
| discharge (N) | 2.29 | 0.12 | 453 | incontinence (N) | 2.66 | 0.69 | 27 | ravage (V) | 1.02 | -0.03 | 1,118 |
| discolored (A) | 2.18 | 0.12 | 173 | indecency (N) | -1.54 | -0.30 | 140 | ravish (V) | 0.37 | -0.08 | 818 |
| disorganized (A) | -1.11 | -0.21 | 163 | indecent (A) | -0.24 | -0.16 | 69 | ravished (A) | 1.13 | 0.20 | 31 |
| distrust (N) | 0.92 | -0.02 | 849 | infected (A) | 2.88 | 0.05 | 415 | rectal (A) | 6.82 | 0.93 | 88 |
| dominate (V) | 2.28 | 0.05 | 2,523 | inflamed (A) | 1.50 | 0.47 | 19 | rectal (N) | 1.08 | 1.33 | 1 |
| dominated (A) | 0.23 | 0.25 | 1 | insert (V) | 1.02 | -0.05 | 2,407 | rectum (N) | 9.75 | 0.53 | 323 |
| dominating (A) | 1.90 | 0.55 | 19 | inserted (A) | 0.66 | 0.60 | 2 | restless (A) | 1.44 | -0.05 | 3,364 |
| dread (N) | 8.69 | 0.11 | 5,919 | insertion (N) | 1.32 | 0.13 | 137 | restrain (V) | 4.32 | 0.03 | 5,049 |
| dread (V) | 9.51 | 0.16 | 3,821 | insomnia (N) | 4.23 | 0.14 | 575 | restrained (A) | 0.75 | -0.05 | 808 |
| drugged (A) | 5.21 | 0.41 | 235 | intimacy (N) | 13.00 | 0.37 | 2,907 | rub (V) | 18.68 | 0.09 | 43,000 |
| drunk (A) | 18.47 | 0.16 | 12,300 | intimate (A) | 12.26 | 0.14 | 6,526 | salacious (A) | -0.74 | -0.15 | 294 |
| dysfunction (N) | 6.60 | 0.69 | 139 | intoxicated (A) | 4.24 | 0.16 | 599 | scared (A) | 51.38 | 0.45 | 27,487 |
| exhibit (V) | 0.54 | -0.06 | 613 | involuntary (A) | -0.01 | -0.08 | 1,494 | secretion (N) | 2.98 | 0.44 | 108 |
| exhibitionism (N) | -3.13 | -0.74 | 56 | irritated (A) | -1.32 | -0.14 | 5,449 | sensitive (A) | 2.71 | -0.08 | 10,623 |
| explicit (A) | 4.65 | 0.16 | 938 | isolate (V) | 4.19 | 0.06 | 1,626 | sex (N) | 47.76 | 0.48 | 23,067 |
| exploit (V) | 1.75 | -0.01 | 1,109 | isolation (N) | 5.23 | 0.10 | 1,283 | sexual (A) | 52.38 | 0.71 | 10,805 |

Table 12: Sexual abuse dictionary (Part 1). For each term, the columns show the $z$-score from the Mann-Whitney U test on term frequencies, the log ratio, and the number of documents $d \in D_{ci}$, it occurred in. The shading of the $z$ column indicates if the term is significantly more or less frequent in $D_{ci}$.

| Term (POS) | $z$ | lr | $\#d$ | Term (POS) | $z$ | lr | $\#d$ | Term (POS) | $z$ | lr | $\#d$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| shame (N) | 18.89 | 0.18 | 15,254 | suicide (N) | 22.44 | 0.54 | 3,420 | unsolicited (A) | 0.27 | -0.07 | 125 |
| sleepless (A) | 2.60 | 0.05 | 752 | survivor (N) | 9.35 | 0.20 | 2,161 | unwanted (A) | 9.52 | 0.18 | 3,186 |
| soil (V) | 9.18 | 0.40 | 896 | syphilis (N) | 5.53 | 1.62 | 22 | unwelcome (A) | -1.56 | -0.16 | 1,193 |
| solitude (N) | 1.51 | -0.02 | 1,117 | terrified (A) | 22.71 | 0.39 | 5,884 | unwilling (A) | 0.44 | -0.06 | 3,012 |
| stalk (V) | -8.78 | -0.26 | 6,197 | thigh (N) | 12.88 | -0.09 | 23,737 | vagina (N) | 9.12 | 0.50 | 514 |
| stalker (N) | -4.75 | -0.34 | 903 | thrust (V) | -7.06 | -0.20 | 11,410 | vaginal (A) | 6.57 | 0.28 | 303 |
| stalking (A) | 1.02 | 0.16 | 41 | touch (N) | 23.66 | 0.15 | 29,152 | victimize (V) | 10.36 | 0.99 | 197 |
| stare (N) | 0.06 | -0.06 | 13,606 | touch (V) | 50.07 | 0.33 | 49,381 | violate (V) | 29.35 | 0.69 | 3,098 |
| stare (V) | 9.63 | 0.00 | 64,029 | unconscious (A) | 1.88 | -0.08 | 7,342 | violated (A) | 4.35 | 1.48 | 16 |
| std (N) | 15.32 | 1.18 | 402 | underage (A) | 10.96 | 0.42 | 883 | violating (A) | 3.51 | 4.65 | 1 |
| strangle (V) | 4.11 | -0.00 | 5,117 | underclothing (N) | -0.14 | -0.21 | 10 | violation (N) | 12.74 | 0.45 | 1,160 |
| strangulation (N) | 5.79 | 0.51 | 172 | underwear (N) | 11.84 | 0.08 | 7,334 | vulgar (A) | 3.55 | 0.13 | 901 |
| strip (V) | 12.40 | 0.10 | 9,691 | undress (V) | 15.30 | 0.28 | 4,136 | vulnerable (A) | 23.14 | 0.28 | 9,539 |
| stroke (V) | 21.80 | 0.17 | 21,572 | undressed (A) | 6.62 | 0.33 | 561 | wet (V) | 11.17 | 0.18 | 4,513 |
| subjugate (V) | 0.42 | -0.03 | 183 | unease (N) | -0.84 | -0.12 | 2,225 | withdraw (V) | 3.37 | -0.04 | 5,790 |
| subjugated (A) | -0.98 | -1.50 | 1 | uneasy (A) | 8.02 | 0.08 | 5,045 | withdrawal (N) | 12.22 | 0.53 | 1,116 |
| suggestive (A) | -0.14 | -0.07 | 972 | unfocused (A) | 5.85 | 0.11 | 1,919 | — | — | — | — |
| suicidal (A) | 15.92 | 0.67 | 1,313 | unprotected (A) | 0.22 | -0.04 | 517 | — | — | — | — |

Table 13: Sexual abuse dictionary (Part 2). For each term, the columns show the $z$-score from the Mann-Whitney U test on term frequencies, the log ratio, and the number of documents $d \in D_{ci}$, it occurred in. The shading of the $z$ column indicates if the term is significantly more or less frequent in $D_{ci}$.

## B. Vocabulary of Texts With Trigger Warnings

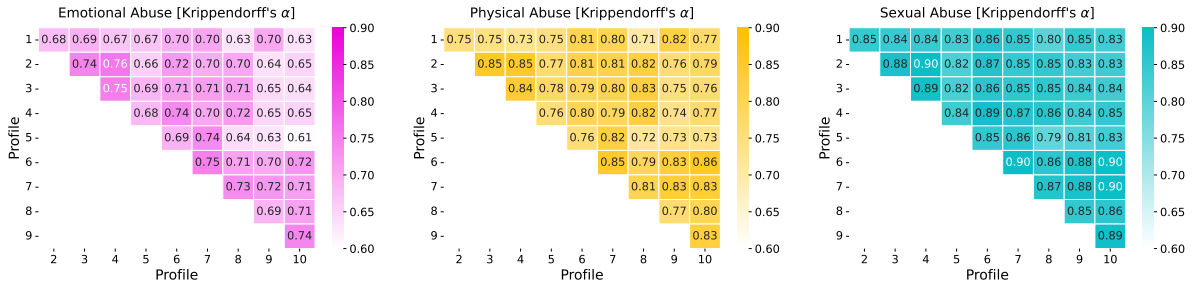



Figure 3: Pairwise agreement between the sociodemographic profiles (Krippendorffs's alpha). The agreement is reported separately for each warning category.

| | | Emotional Abuse | | Physical Abuse | | Sexual Abuse | |
|---|---|---|---|---|---|---|---|
| **Attribute** | **Attribute Value** | $\alpha$ | $\Delta\alpha$ | $\alpha$ | $\Delta\alpha$ | $\alpha$ | $\Delta\alpha$ |
| Gender | Female | 0.709 | 0.016 | 0.822 | 0.033 | 0.874 | 0.020 |
| | Male | 0.732 | 0.040 | 0.837 | 0.048 | 0.873 | 0.020 |
| Race | Black | 0.714 | 0.021 | 0.803 | 0.013 | 0.855 | 0.001 |
| | White | 0.676 | -0.016 | 0.773 | -0.017 | 0.852 | -0.001 |
| Education | College (no degree) | 0.680 | -0.012 | 0.772 | -0.018 | 0.843 | -0.010 |
| | Associate degree | 0.738 | 0.046 | 0.800 | 0.010 | 0.889 | 0.036 |
| | Bachelor's degree | 0.652 | -0.041 | 0.785 | -0.004 | 0.833 | -0.020 |
| | Master's degree | 0.641 | -0.051 | 0.717 | -0.072 | 0.793 | -0.061 |
| Age | Under 18 | 0.680 | -0.012 | 0.772 | -0.018 | 0.843 | -0.010 |
| | 25 - 34 | 0.708 | 0.016 | 0.803 | 0.013 | 0.857 | 0.004 |
| | 35 - 44 | 0.652 | -0.041 | 0.769 | -0.020 | 0.851 | -0.002 |
| Political Affiliation | Conservative | 0.714 | 0.022 | 0.816 | 0.027 | 0.872 | 0.018 |
| | Independent | **0.744** | 0.052 | **0.852** | 0.062 | **0.880** | 0.026 |
| | Liberal | 0.688 | -0.004 | 0.778 | -0.012 | 0.846 | -0.008 |

Table 14: Agreement between sociodemographic profiles with a shared attribute value (Krippendorffs's $\alpha$). For example, profiles 1, 6, 7, 9, and 10 share the value "Female" for the attribute "Gender". The agreement is reported separately for each warning category, with the intragroup agreement $\alpha$ and the difference $\Delta\alpha$ between the intragroup agreement and the agreement across all profiles. The highlighting indicates if the intragroup agreement is higher or lower than the overall agreement (0.692 for Emotional, 0.790 for Physical, and 0.854 for Sexual Abuse). Only attribute values with at least two profiles are reported. The highest agreement for a warning category is marked in boldface.

In our annotation experiment, we used sociodemographic prompting to simulate disagreement between annotators and thereby reflect the subjectivity of trigger warning assignment (see Sections 5 and 6). To verify that different profiles lead to different annotations, we measured the annotator agreement using Krippendorff's $\alpha$. The pairwise agreement scores in Figure 3 show that, on average, the profiles have the lowest agreement for *Emotional* and the highest agreement for *Sexual Abuse*. Across all warnings, we see consistent effects like profile 5 having the lowest agreement with profiles 8-10, and profile 4 having a high agreement with profiles 2 and 3. Table 14 presents the agreement scores for groups of profiles with shared attributes. We find that profiles with the "Independent" political affiliation have the highest intragroup agreement, while profiles with the "Master's degree" education have the lowest agreement. This analysis shows that (i) `Flan-T5` 11B generates different annotations for different sociodemographic profiles, and (ii) the agreement and disagreement between profiles is consistent across warning categories.

.

He was a collected teenager, a loyal companion and good son. An awarded babysitter, as Hizashi and Nemuri would joke. But instead of getting over his trauma, finding reason behind what came about, or simply moving on, it was **isolation** and self hate. It was **guilt**, **emotional** suppression, and turning to murder. The whiplash of losing it all truly damaged his soul, carving Shouta into a dangerous mold. And yet, a kid with immunity to the noble infection had become the one to soften such edges.

**9/10 Positive Annotations for Emotional Abuse**

"I think... I think it's a false choice. **Trying** to make decisions based solely on **emotion** can lead to really **bad** choices, and **trying** to make decisions based solely on logic is impossible without **lying** to yourself. But I do know that learning to master your **emotions** in the moment is a skill that every person needs to learn regardless of species." "Why?" "Think of it this way. How do you think I **felt** when I received the message that Spock was in danger?" "You were afraid. And frustrated."

**0/10 Positive Annotations for Emotional Abuse**

That they would become a family and that would be all they needed. Then controlling who he spent time with. How long and pretty soon he'd developed this **anxiety**. If he was late to a date or getting **home**, Matthew would be furious. Words turned into slaps, slaps turned **punches**, and soon **kicking** and being pushed against walls and **doors**. One time so hard that he'd gotten a concussion. Once he was away from Matthew, away from that apartment that had served to be his own personal hell hole.

**10/10 Positive Annotations for Physical Abuse**

"Of course, sport. We **love** having you **stay** with us." He **smiles** again and, careful to telegraph his movements, reaches out to pat Tim's shoulder. This time, Tim doesn't **flinch** away or tense up too badly. Compared to the **frightened** child who first came to **stay** with them nearly two months ago, he's made a lot of progress. Bruce stands up and ignores the way that his knees **crack**. "Well, I've bothered you two for long enough. I'll let you get back to contemplating the Middle Earth."

**0/10 Positive Annotations for Physical Abuse**

The rest of the team was still huddled there, concern and confusion on all their faces. "Hey, guys...listen up for a moment." The silver-haired **boy** exhaled deeply before continuing. "Hinata experienced...a past **trauma** with a volleyball player. He sometimes has **anxiety** about this still. He's **told** me it's getting better but this training camp might not be good for him. Please, just be conscious of this and don't **pressure** him to do anything he doesn't **want** to do, okay?" The team **nodded** in unison.

**10/10 Positive Annotations for Sexual Abuse**

Midorima was practicing in the front yard and Kagami was still out running. "I don't think you are able to answer the question. Or if you are, I don't even **want** to **ask**. I just wondered how it is to have **consensual sex**." "Con ? woah, yeah, that's kind of a hard question. No, I have never been **raped**, I can only imagine how horrible that must be. You really never had **sex** that you **wanted**? Not even once?" There was a crease between Takao's eyebrows but at least he was able to stand the conversation.

**2/10 Positive Annotations for Sexual Abuse**

Figure 4: Example passages for each of the three trigger warning categories. High frequency terms related to emotional, physical, and sexual abuse are highlighted accordingly.