Leipzig University
Institute of Computer Science
Degree Programme Data Science, M.Sc.

Trigger Warnings in Fanfiction: An Analysis of Usage Consistency and the Effect of Prescriptive Annotation Guidelines

Master's Thesis

Sebastian Heineking

1. Referee: Prof. Dr. Martin Potthast
2. Referee: Dr. Harrisen Scells

Submission date: August 12, 2024

Declaration

Unless otherwise indicated in the text or references, this thesis is entirely the product of my own scholarly work. In particular, literal or analogous quotations are marked as such. I am aware that any infringement can also lead to the subsequent withdrawal of the degree. I confirm that the electronic copy corresponds to the printed copies.

Leipzig, Germany, August 12, 2024

. . . .
Sebastian Heineking

Abstract

The assignment of trigger warnings to a piece of content is a subjective task. People with different experiences and sensitivities are thus likely to disagree if a given text sample deserves a warning or not. This thesis explores the subjectivity in trigger warning assignment from two perspectives. The first set of experiments analyzes the warning tags that authors assign to their works on the fanfiction website Archive of Our Own. As the warnings are not assigned by a central authority, they are subject to the individual judgement of each author. To test if the tags are a reliable indicator for the presence of potentially distressing content in a document, we test if a vocabulary of terms related to that content occurs significantly more often in tagged documents than in a comparable baseline. Our analyses reveal that documents tagged for different categories of abuse indeed contain terms related to their categories significantly more often than documents tagged for other forms of abuse. The second set of experiments tests if explicitly discouraging subjectivity in the annotation process for trigger warnings can increase the agreement between annotators. The effectiveness of employing this prescriptive annotation paradigm is tested using sociodemographic prompting of a large language model. The annotation study shows a highly significant increase in pairwise annotator agreements when using a prescriptive in contrast to a descriptive prompt that does not discourage subjectivity.

Contents

List of Acronyms

AO3	Archive of Our Own	2
CDF	Cumulative Distribution Function	52
FGD	Functional Generative Description	14
FWER	Family-Wise Error Rate	22
LLM	Large Language Model	3
NLP	Natural Language Processing	6
PDF	Probability Density Function	43
POS	Part-Of-Speech	20
PTSD	Post-Traumatic Stress Disorder	1
TDS	Term Domain Specificity	14
WTWC-22	Webis Trigger Warning Corpus 2022	2

Chapter 1

Introduction

Trigger warnings are used by publishers of different types of media to inform their audiences about potentially distressing content. The origin of these warnings, also referred to as content warnings, lies in online communities whose members wanted to protect each other from being involuntarily exposed to descriptions or depictions of sexual assault. The behavior to warn other people about potentially distressing content is based on evidence from psychological research that reminders of traumatic experiences can cause painful recollections for people with post-traumatic stress disorder (PTSD). Since their inception, trigger warnings have become more widespread and are applied to texts, images, and videos in a range of different contexts. In addition to that, the types of content that are covered by trigger warnings have expanded to include topics outside of the canonical definitions of trauma such as *Discrimination*, *Pornography* or *Abuse*.

Traumatic experiences that fall under the canonical definition of trauma as well as what can trigger people to have recollections of them, are highly subjective phenomena. The same is true for the colloquial application of trigger warnings as different people might have different opinions on what type of content warrants a warning based on their prior experiences and sensitivities. Mismatches between the warnings assigned to a piece of media and the opinions of people that consume it are undesirable. Applying too many warnings risks devaluing their effectiveness while applying too few warnings potentially exposes people to content they want to avoid. This raises the following question:

Can we trust trigger warnings?

Towards answering that question, we investigate for the first time the labeling-consistency of trigger warnings. The experiments in this thesis are conducted on a dataset of documents already annotated for a taxonomy of 36 different warnings. The Webis Trigger Warning Corpus 2022 (WTWC-22), created by Wiegmann et al. [2023], is a collection of works from the fanfiction website Archive of Our Own (AO3). The authors on AO3 provide a list of tags for their works that can both summarize the story's content as well as inform other users about potentially distressing content through warning tags. As the authors can write their own tags in a freeform field, the tags reflect no single understanding of a given warning. Instead, each author uses their own judgement to decide if their work requires a warning tag or not. Hence, the first set of experiments are guided by the following research question:

> **RQ1**: Do the authors on Archive of Our Own apply warning tags in a way that is consistent with the vocabulary used in their works?

Consistency in this context means that documents tagged for a specific warning contain terms that describe warning-related content with a higher frequency than a comparable baseline. The warning *Physical Abuse*, for instance, suggests that the tagged documents contain scenes of one person inflicting physical harm on another. Examples of terms implied with that warning are *punch*, *bruise* or *injure*. The first contribution of this thesis is a methodology for testing the consistency between tags used to convey a warning and the type of vocabulary that is associated with that warning (Chapter 3). The methodology combines term-specific frequency tests with a distribution test over the entire expected vocabulary. This allows to test two things: First, the term-specific frequency tests reveal which terms are significantly more frequent for documents with a warning. Second, the distribution test assesses if the vocabulary as a whole occurs more frequent in documents with a warning tag than in others. Our experiments reveal that between 38 and 55 % of the expected terms are significantly more frequent in documents tagged for their respective warning category. In addition to that, the distribution tests show a significantly higher average term frequency of the whole vocabulary for all tested categories.

Another problem that arises from the subjectivity of trigger warnings is the difficulty to obtain training data for classifiers. In a recent study, Wiegmann et al. [2024] found the task of annotating pieces of texts for automated trigger warning assignment to cause notable disagreements among annotators. The task shares this challenge with other subjective annotation tasks such as hate speech detection or stance classification. As a way to reduce annotator disagreements, Rottger et al. [2022] proposed the prescriptive annotation paradigm for subjective annotation tasks. This paradigm explicitly discourages subjectivity in the annotation guidelines and provides examples for how different types of content should be annotated.

The second set of experiments thus aims to answer the following research question:

> **RQ2**: Can prescriptive annotation guidelines increase the annotator agreement on the task of labeling text for automated trigger warning assignment?

The second contribution of this thesis is a methodology for testing the effect of prescriptive annotation guidelines (Chapter 4). It compares the effect of three different annotation prompts on the pairwise agreement between annotators. The first prompt follows the descriptive paradigm proposed by Rottger et al. [2022] and specifically asks for subjective judgements. In addition to that, it is not specific by asking for one of the seven open-set labels in the taxonomy of Wiegmann et al. [2023] that cover broader warnings such as *Abuse* or *Discrimination*. The second prompt follows the prescriptive paradigm by discouraging subjectivity and providing a list of examples that meet the annotation criteria. It is also more specific by asking for categories of a warning like *Emotional Abuse* or *Physical Abuse*. The third and final prompt follows the descriptive paradigm but asks for the specific warning categories to control for the effect of higher specificity. In our annotation experiments, the prescriptive annotation prompt results in a significantly higher average agreement between annotations than any of the two descriptive prompts. Using the specific warning category in the descriptive prompt also leads to a significant increase in average agreement in comparison with the less specific prompt that asks for one of the open-set labels.

The experimental setup for both methodologies is outlined in Chapter 5. All experiments in this thesis were conducted on documents tagged for some form of *Abuse*. The specific categories that were used both in consistency testing and the annotation task, were *Emotional Abuse*, *Physical Abuse*, and *Sexual Abuse*. Lacking the resources to conduct a large-scale annotation study with human annotators, the annotation experiments built on the findings of Beck et al. [2024] and Wan et al. [2023] that sociodemographic prompting of large language models (LLMs) can be used to predict which samples are likely to cause disagreement among human annotators. The results are presented in Chapter 6 and discussed in Chapter 7.

Chapter 2

Background & Related Work

This chapter lays the theoretical foundation for this thesis by giving an overview on relevant research. Section 2.1 summarizes research on trigger warnings in clinical psychological and recent advances in automated trigger warning assignment. Section 2.2 is focused on the topic of subjectivity in natural language processing and the research on how to address it. Section 2.3 concludes the chapter with related work in (computational) linguistics as the theoretical basis for the statistical tests on vocabulary consistency.

2.1 Trigger Warnings

Trigger warnings were developed as a way to help people with PTSD in avoiding content that might cause them to have recollections of their traumas. Several studies in the field of clinical psychology have been conducted to test if trigger warnings have the intended positive effects like stress reduction. Section 2.1.1 gives an overview on these studies and also briefly outlines the history of trigger warnings. Section 2.1.2 summarizes recent research on the task of automatically assigning trigger or content warnings to documents.

2.1.1 A Brief History of Trigger Warnings

Trigger warnings inform audiences of different types of media such as texts, videos or images about potentially disturbing content while often providing a description of said content (Boysen [2017], Bridgland et al. [2019]). They were originally created in online communities to help individuals with PTSD avoid reminders of their trauma, specifically sexual assault (Jones et al. [2020]). The reasoning behind trigger warnings is based on evidence that reminders of traumatic experiences can cause people with PTSD to have painful recollections of the event (American Psychiatric Association [2013]). In the recent past, the use of trigger warnings has expanded to other fields such as education or social media and covers warnings that go beyond canonical traumatic events (Bridgland et al. [2023]). Trigger warnings received broader attention in the United States in 2014, when university students advocated for their application in lectures (Wyatt [2016]). This sparked a debate with opponents arguing that trigger warnings constitute a threat to academic freedom while proponents saw them as a way to express solidarity with marginalized groups (Bridgland et al. [2023], Dickman-Burnett and Geaman [2019]).

In addition to the public discourse, trigger warnings are a contentious topic in clinical psychology. Given their origin in online communities rather than clinical studies, trigger warnings have developed independent of the scientific evaluation typically applied to trauma interventions (Jones et al. [2020]). Sanson et al. [2019] conducted a set of six experiments to study how trigger warnings changed the symptoms of distress in college students and crowd sourced workers after being exposed to media with negative content. The authors found only minor differences between people who saw a warning before the exposure and those who did not and concluded that "trigger warnings are at best trivially helpful". These findings are supported by a meta study by Wahlsdorf et al. [2024] of 14 papers that showed primarily no effect of trigger warnings and more negative than positive effects if any occurred. The negative effects are mostly related to an increase in negative anticipatory reactions, meaning that the individual expects stimuli associated with their trauma and might focus specifically on them (Shafir and Sheppes [2020]). Potential positive effects of trigger warnings are a reduction of the distressing effect of averse stimuli when experienced in a foreseeable manner (Grupe and Nitschke [2013]), and avoidance of disturbing content, but Wahlsdorf et al. [2024] found no or even contrary evidence for these effects in the analyzed studies. The authors conclude that trigger warnings in their current form seem not to have the intended effects on people with PTSD. In addition to that, they state that trigger warnings could advocate avoidance as an appropriate response to traumatic experiences and criticize this from a therapeutic standpoint. As a potential alternative, Wahlsdorf et al. [2024] refer to the film rating system used by the Motion Picture Association (e.g. PG-13) that was found to induce less physiological excitation than trigger warnings (Bruce et al. [2021]). The authors advocate for ways to inform audiences about upcoming content without the use of 'warning' or 'trigger', with the latter invoking associations with traumatic experiences.

As this thesis approaches the topic of trigger warnings from a computer science angle, it is not primarily concerned with their usefulness from a psychotherapeutic standpoint. Instead, the focus of this thesis is to analyze if the way that trigger warnings are applied in communities such as AO3 is consistent with the content of the documents they are applied to. Nonetheless, the findings from research in clinical psychology are important and need to inform implementations of automated trigger warning assignment. This is especially true for the question of how best design the human-computer interaction when informing audiences about potentially disturbing content.

2.1.2 Automated Assignment of Trigger Warnings

The automated detection and assignment of trigger or content warnings is a relatively new field in natural language processing (NLP). It shares some similarities with automated content moderation in that both aim to reduce harm by identifying potentially distressing content (Grimmelmann [2015]). In addition to that, both settings can benefit from automation by (1) reducing the exposure of human moderators to harmful material (Stratta et al. [2020]) and (2) scaling the scope to larger sets of content (Horta Ribeiro et al. [2023]). A central difference, however, is that content moderation focuses on enforcing guidelines by an organisation or community and often results in the removal of identified content. In contrast to that, in trigger warning assignment, the focus lies on informing people about content that might only be distressing to a subset of individuals with specific experiences. Furthermore, the content is generally not removed.

An early study related to content warnings was conducted by Stratta et al. [2020], who developed a system called DeText to automatically identify sexual violence in texts of web pages. DeText uses keyword identification and sentiment analysis to check for explicit and implicit keywords as well as the polarity of sentences. If certain thresholds are met, a page is classified as containing sexual violence. To evaluate the effectiveness of DeText, the authors conducted tests on a dataset of 50 web pages, equally divided into a positive (containing sexual violence) and a negative class (no sexual violence), reporting an F1-score of 0.8940. In addition to that, they conducted a user study with DeText as a Google Chrome extension that blurred out any web pages that met classification thresholds and provided users with a warning. The ten participating students provided positive feedback on the design and usability of the extension and generally agreed with its classifications with an F1-score of 0.9325 on a total of 231 web pages (23.1 per student).

The first work on automated assignment of trigger warnings on a comprehensive set of warning categories was conducted by Wiegmann et al. [2023]. The authors built the WTWC-22, a dataset of 7.8 million works from the fanfiction website AO3. The website allows users to write and read stories with connections to popular media such as books, movies, TV shows or video games. The authors on AO3 assign tags to their stories to both facilitate the search for their work as well as give an overview of the content. In addition to that, the tags are used to convey warnings about things like *Violence*, *Abuse* or *Death*, with roughly 50% of the works having author-assigned warnings.

As starting point for the classification task, Wiegmann et al. [2023] created a taxonomy of 29 closed-set categories that define a specific concept such as *Racism* or *Classicism* and seven open-set warnings that abstract from the closed-set categories and cover broader areas like *Discrimination*. This taxonomy is based on a synthesis of guidelines issued by eight universities from the U.S., the United Kingdom, and Canada. The authors then mapped each of the 53 million free-form tags in their dataset to the 36 warning categories using a combination of manual annotation and distant supervision based on the graph of relations between tags constructed by the AO3 community.

Wiegmann et al. [2023] used the collections of 36 fine-grained and 7 coarse-grained labels to conduct a set of experiments with four models for long-document classification: a support vector machine (`SVM`), `XGBoost` (Chen and Guestrin [2016]), `RoBERTa` (Liu et al. [2019]), and `Longformer` (Beltagy et al. [2020]). They found `XGBoost` to be the most effective model with a a micro-F_1 of 0.52 on the fine-grained label set. Only on the subset of documents with fewer than 512 tokens was `RoBERTa` able to achieve higher effectiveness than `XGBoost`. This is consistent with the neural models generally being more effective on the experiments when classifying texts within their context length. The authors observed precision to be higher than recall by about 0.2-0.3 and argue that emphasis should be put on improving recall given that trigger warning assignment should prioritize the reduction of false negatives over false positives to not miss content that requires a warning.

Given that trigger warnings on AO3 are assigned to long documents and do not specify where the triggering content occurs, Wiegmann et al. [2024] conducted a follow-up study. In this study, they explored how reliable individual passages can be (1) annotated by human annotators and (2) automatically classified by a range of models. For the annotation task, the authors collected passages of five consecutive sentences from the WTWC-22 using dictionary-based retrieval for eight closed-set categories; four each from the two most frequently assigned warnings *Aggression* and *Discrimination*. The retrieval keywords for each category were collected by prompting `GPT-3.5-turbo-0301`, manually cleaning the results and dividing them into in- and out-of-distribution keywords. For each category, the first annotator labeled passages in a binary fashion for warnings until 50 positive labels were recorded to ensure label balance across categories. Subsequently, two other annotators were given the same passages for a total of three annotations per sample. In their evaluation, Wiegmann et al. [2024] found no general consensus among annotators. While 55% of all passages were labeled unanimously negative, only 5% were labeled unanimously positive. These consistently positive samples often contained heavy slurs or very graphic language. The authors concluded that the observed variance stems from differences among annotators with regards to sensitivity and opinions about what type of content requires a warning. The observed subjectivity in annotations is not uncommon in NLP tasks that deal with harmful content as is discussed in the following section.

2.2 Subjective NLP Tasks

A central challenge in collecting annotations for subjective NLP tasks is that people with different experiences and sociodemographic backgrounds might perceive the same text differently and assign different labels as a consequence. This phenomenon has been observed for a range of tasks such as classification of toxicity (Sap et al. [2019, 2022]), hate speech (Salminen et al. [2019], Waseem [2016]) or stance detection (ALDayel and Magdy [2021], Luo et al. [2020]). Given the results by Wiegmann et al. [2024], the annotation task for triggering warnings appears to be similarly subjective. In comparison with the aforementioned examples, trigger warning assignment comes with additional annotation challenges. First, whereas a lot of people might have an opinion on and understanding of hatefulness or toxicity on social media, comparatively fewer people have had traumatic experiences and would be able to judge if a piece of content causes them to have a recollection of that. Second, even if people with past trauma would be asked to annotate text passages, which would be ethically questionable, they might still disagree based on their individual experience and what specifically causes them to remember the events (Wahlsdorf et al. [2024]). Third and finally, if people without past trauma make annotations based on what they *assume* could trigger other people, the annotations might differ not only based on the textual content but also the beliefs about people with PTSD.

2.2.1 Prescriptive & Descriptive Annotation Guidelines

Given the inherent subjectivity of certain NLP tasks, Rottger et al. [2022] suggest that dataset creators should (1) be conscious about the intended use case of their dataset and (2) decide whether annotator subjectivity is helpful or detrimental to that use case. Towards this goal, the authors propose a framework of two opposite annotation paradigms. While the *descriptive* paradigm encourages annotator subjectivity to be able to study individual beliefs, the *prescriptive* paradigm discourages it to get consistent annotations. As a consequence, the datasets created using these two paradigms are useful for different types of tasks.

Descriptive Paradigm The descriptive paradigm functions similar to a survey and leads to datasets that allow researchers to analyze how people with different backgrounds perceive certain texts. Among the examples given by the authors are that young adults or people that identify as LGBTQ+ are more likely to rate a given social media comment as toxic (Kumar et al. [2021]) as well as differences in hate speech detection that correlate with sociodemographic characteristics (Waseem [2016]). Salminen et al. [2019] found that people tend to agree about the extreme cases in annotation tasks. Hence, studying differences between groups of annotators can help pinpoint what exactly causes the disagreements. In addition to that, encoding different beliefs in a dataset can be used in model training to develop multi-belief architectures that make predictions in an ensemble approach (Akhtar et al. [2020]) or measure biases in a dataset (Al Kuwatly et al. [2020]). Yet, the descriptive paradigm is unsuited for the creation of datasets for classic NLP tasks that require a single gold standard answer for each sample. Related to that, unknowingly using descriptive annotation guidelines and "resolving" disagreements through majority voting can conceal valid disagreements (Basile et al. [2021], Leonardelli et al. [2021]).

Prescriptive Paradigm The prescriptive paradigm, on the other hand, specifically aims at creating datasets that reflect a single, consistent understanding of the NLP task. Consequently, this understanding or belief needs to be decided upon before collecting the annotations and tends to limit the perspectives on the task to a narrower corridor than the descriptive paradigm. As an example, Rottger et al. [2022] cite the data collection for automated content moderation on social media platforms. While different people may have varied, yet valid beliefs about what type of content should not be present on social media, the platform operators have specific content policies in place. Hence, for their automated enforcement, the platform operators need datasets that reflect said policies and not the subjective opinions of annotators. A central advantage in applying the prescriptive paradigm is that the complexity of dealing with annotator disagreements can be reduced to one of two scenarios: Either the annotation guidelines are ambigous and require improvements, or the annotators made mistakes in their application of the guidelines. Without the conscious decision for the prescriptive paradigm, annotator disagreements pose a larger challenge as they could also be the consequence of annotator subjectivity. As another advantage, Rottger et al. [2022] cite similarities between prescriptive annotation guidelines and data statements in that the guidelines provide users of the datasets with a more detailed understanding of how they were created. Creating the annotation guidelines for the prescriptive paradigm comes with a range of challenges. First, dataset creators need to decide which belief should be annotated for, possibly discarding other valid beliefs. Second, the guidelines need to be developed with an understanding of both the task and the data. This can entail questions such as "What is the legal definition of hate speech?", "What kinds of samples will annotators be presented with?" or "Which criteria can help annotators decide on difficult cases?".

As an illustration of the two paradigms, Rottger et al. [2022] conducted an experiment with 60 annotators, uniformly split into three groups, and asked them to label 200 Twitter posts for being hateful or not. The first group received a descriptive prompt that asked for their personal opinion on whether a given post was hateful. The second group received a prescriptive prompt that explicitly discouraged subjective judgements and asked annotators to check if the post met criteria for hate speech from an extensive list that was provided as a separate link. The third and final group received a prompt that, while also asking annotators to decide if the criteria for hate speech were met, only provided a short list of examples and did not explicitly discourage subjective judgements. This prompt was added to control for the differences in length and complexity between the descriptive and prescriptive prompt. The authors found significantly higher annotator agreements, as expressed by Fleiss' κ, for the prescriptive (0.78) than for the descriptive (0.20) and control prompt (0.15), concluding that prescriptive annotation guidelines help annotators in recording a specific belief.

Given the subjectivity of trigger warnings discussed above, prescriptive guidelines could help increase annotator agreement. This requires to reduce the room for subjective judgements by clearly defining the belief that should be annotated. One way to reduce subjectivity in trigger warning annotations is to not ask annotators what they *think* could cause people with trauma to have painful recollections of their experience, but instead provide lists of examples or criteria for content that is commonly associated with a given warning category. Asking annotators to apply these criteria for their annotations avoids putting the burden of judgement on them to decide if a given passage could cause trauma recollections in other people.

2.2.2 Sociodemographic Prompting of LLMs

A different perspective on the subjectivity of certain NLP tasks was taken by Beck et al. [2024]. The authors explored how sociodemographic prompting, the process of asking an LLM to generate responses as if given by people with that background, impacts task effectiveness. While this paper is not the first to explore sociodemographic prompting (Deshpande et al. [2023], Santurkar et al. [2023], Wan et al. [2023]), it is the largest study to date and covers seven datasets, four different tasks, and six model families.

In their sensitivity analyses, the authors found instruction-tuned models based on T5 (Raffel et al. [2020]) to be the most affected: For Flan-T5 with 3 billion and 11 billion parameters, the predictions changed on average in more than 40% of cases across all datasets when doing zero-shot prompting with and without sociodemographic profiles. Furthermore, the choice of model seemed to be more influential than properties of the text as the authors found no samples with consistently changed predictions for all models. Beck et al. [2024] observed small positive effects of sociodemographic prompting when trying to reproduce the annotations of a specific annotator with the same profile. The predictions, however, were still incorrect for more than half of all samples. Hence, current LLMs seem not to be able to consistently predict how a person with a certain set of sociodemographic attributes might annotate a given piece of text.

While the models are not suited to make annotations, the authors found that some models perform well on the task of predicting if a sample is likely to cause disagreement among human annotators. This application of sociodemographic prompting was first suggested by Wan et al. [2023] and scaled to more models and datasets by Beck et al. [2024]. The prediction of disagreement was done in a ensemble-like fashion. A sample is given to an LLM with different sociodemographic profiles to obtain multiple responses. If at least one response is different from those given with the other profiles, the sample is said to cause disagreement in sociodemographic prompting. These results are compared with the annotations by human annotators to create a binary classification setting: If a sample causes disagreement both among human annotators and among different sociodemographic prompts, it is treated as a true positive. Cases of unanimous agreements in both settings, on the other hand, are true negatives. For this setting, the 11B parameter version of Flan-T5 achieved an average F1-score of 0.62. The best scores were recorded for sentiment analysis (0.82), stance prediction (0.69 and 0.78), and one of the two toxicity datasets (0.73), while hate speech classification appeared to be more difficult (0.41 and 0.44).

2.3 Linguistic Background

As this thesis is concerned with studying language and performing statistical tests on corpora of text documents, it builds on theory in (corpus) linguistics. Therefore, this section will outline the measures of corpus linguistics used to test the hypotheses throughout this thesis (Section 2.3.1) as well as the linguistic framework of the Functional Generative Description (Section 2.3.2).

2.3.1 Corpus Linguistic Measures

Testing the usage consistency of warning tags by authors on AO3 will be done using measures from corpus linguistics. Very generally, corpus linguistics encompasses methodologies to empirically study language on a large scale using one or more corpora of naturally spoken or written text (Meyer [2002]). Wallis and Nelson [2001] proposed a high-level categorization of corpus linguistics into annotation, abstraction, and analysis. The methods applied in this thesis fall into the last category that is concerned with testing hypotheses on corpora by using statistical methods.

Significant Differences in Term Frequencies The first method is taken from the field of digital humanities. In a methodological paper, Lijffijt et al. [2014] compared several statistical tests on the task of estimating the significance of differences in word frequencies between two corpora. Traditionally, this task has been performed using either log-likelihood tests or χ^2 tests (Dunning [1993], Rayson and Garside [2000]), both of which are based on the bag-of-words model. Both tests compare the observed number of occurrences O_t for a given term t in one of the two corpora with an expected number E_t calculated based on combining both corpora. If the observed number of occurrences strongly deviates from the expected one, the term is concluded to occur with a significantly different frequency in the two corpora. The difference between the tests lies in the computation of the test statistic from O_t and E_t. A central assumption in the bag-of-words model is that all terms in a corpus are statistically independent. Lijffijt et al. [2014] challenge this assumption, arguing that terms in the same document are not independent of each other but that the occurrence of one term influences the probability of occurrence in another.

As alternatives, Lijffijt et al. [2014] propose to use other tests, namely Welch's t-test, the Mann-Whitney U-test (also called Wilcoxon rank-sum test), and their own test called Inter-Arrival Time test. Instead of assuming independence on term level, the authors treat the documents in both corpora as independent samples and apply the tests to their term frequencies. For the t-test and the Mann-Whitney U-test, the first step is to compute the document-level term frequencies for a given term. This distribution of frequencies can then be used to perform significance testing. For a given term, Welch's t-test calculates the mean frequency on document level and corresponding standard deviation for both corpora. It then tests if the means of the two distributions are significantly different from each other. In contrast to the Student's t-test, it does not require equal variance between the two distributions but still assumes the means to be normally distributed.

The Mann-Whitney U-test is less limited as it does not require assumptions about the distribution of the mean. The test statistic is calculated by ranking the documents from both corpora by their term frequencies and counting the number of pairs for which one corpus has a lower term frequency than the other. If one corpus has a lower term frequency for a sufficient number of document-pairs, the term is concluded to occur significantly less for that corpus.

The Inter-Arrival Time test is different from the former two. For a term of interest t, it counts how many other terms occur between two instances of t. The inter-arrival time is this number of terms between instances plus one. For two corpora D and $\hat{D}$ the test is performed as follows: First, create an empirical distribution of inter-arrival times for $\hat{D}$ by recording the inter-arrival times for all documents in the corpus. Second, sample inter-arrival times uniformly at randomly from this distribution to create a random corpus of the same size as D. For an example, lets assume that the distribution of inter-arrival times in $\hat{D}$ is the following and that D is of size 42:

$$\text{Inter-Arrival Times} = \{10, 18, 5, 22, 27\}$$

Now, we sample the times 5, 27, and 10 to have the same number of terms as D and three instances of the term t. The test is performed by comparing the frequency of term t in D to the frequency in the randomly created corpus that (1) has the same size as D and (2) is based on the inter-arrival times of $\hat{D}$.

In their experiments, the authors found that the two bag-of-words approaches are prone to overestimate the significance of differences in term frequencies while the other tests lead to fewer type I errors. For this thesis, the term frequencies will be tested using the Mann-Whitney U-test as the test requires fewer assumptions than the t-test and is easier to apply for large corpora than the Inter-Arrival Time test.

An information that the Mann-Whitney U-test does not give, however, is a useful effect size for the difference in term frequencies. For the setting of ranking documents by their term frequencies, the *common language* effect size yields the share of document pairs in which a document from one corpus has a higher frequency. This does not give an indication for *how much* more frequent the term is in the overall corpus. In an extreme case, a high effect size could result from a lot of documents in one corpus having only a negligibly higher term frequency than the documents in the other corpus. An alternative effect size is presented in the following paragraph.

Effect Size Another field, besides digital humanities, that is concerned with differences in term frequencies is the medical domain. In their paper, Schlatt et al. [2022] suggested different "contrastive termhood scores" to measure how health-related a given phrase in a web document is. One of these measures is the term domain specificity (TDS) that unifies similar approaches by Ahmad et al. [1999], Park et al. [2008], and Wong et al. [2007]. For a term of interest, the measure calculates the logarithm of the ratio of the term frequency in one corpus to that in another corpus. By calculating the ratio on corpus level, the TDS provides a measure of effect size for each term that indicates *how much* more frequent the term is in one corpus relative to the other. A different, more literal name for the same approach is the log ratio as suggested by Hardie [2014] for the field of digital humanities. In the following, the measure will be referred to as log ratio as it is the clearer name.

2.3.2 Functional Generative Description

Another part of related work in linguistics is the Functional Generative Description (FGD) developed by Petr Sgall and his colleagues (Sgall et al. [1986]). It is a linguistic framework rooted in the principles of the Prague School of linguistics that focuses on functionality in language (Luelsdorff [1994]). One of the core requirements introduced by Sgall [1967] is that the FGD automatically assigns "structural characteristics" to sentences in a way that is consistent with the understanding by speakers of the studied language (Lockwood [1971]). Towards this end, the FGD distinguishes five layers of language description: The phonetic, the phonological, the morphemic, the surface syntactic, and the deep syntactic layer (UFAL [2014]). The last layer, also called the tectogrammatical layer, is the central component of the FGD and is concerned with describing the meaning of a sentence (Hajicova [2006], Sgall et al. [1986]). The meaning is encoded in a dependency tree with individual nodes representing the meaning units of the sentence that are connected by edges that represent syntactic relations. Central to the framework is a differentiation between linguistic meaning and extra-linguistic content. The meaning of units in the dependency tree is inferred directly from linguistic understanding without modification through external knowledge. In other words, the meaning is taken *at face value.*

One of the two research questions of this thesis is if authors on AO3 apply warnings tags in a way that is consistent with the vocabulary of their works. Towards answering this question, the FGD will be used to semantically analyze the meaning units of a tag and use them for categorization. Following the approach described above, the meaning of units in the tags will be inferred only from linguistic understanding. By not including external knowledge, the tags can be analyzed purely based on how English speakers would understand them. This is supposed to reflect what kind of content a reader will expect in a document based *only* on the information given by the tags. To use a specific example, the tag *Implied Physical Abuse* conveys that the warning category *Physical Abuse* applies to the document's content. In addition to that, the category is qualitatively characterized using the word *implied*. Hence, the expectation changes from *Physical Abuse* explicitly happening in the document's story to it only being implied. To reflect this qualification effect, meaning units that provide additional characterization of a warning category will be referred to as a *qualifiers* throughout this thesis.[1] Qualifiers can be derived from different levels. For the scope of this thesis, these levels are restricted to the content level, a semantic phenomenon, and the level of pragmatic import.

Content Qualifiers The content qualifiers can be taken directly from word semantics. The first example are extent qualifiers such as *A little Abuse* or *A lot of Abuse*. These units characterize the warning by stating *how much* of it occurs in the document. The next group of content qualifiers are temporal qualifiers like *Brief Instance of Abuse*. The unit *brief* relates to time and suggests that the *Abuse* taking place in the story will only occur for a small amount of time. The final group consists of descriptive qualifiers that provide supplemental information to the warning. The tag *Light Physical Abuse*, as an example, suggests that the instances of *Physical Abuse* in the story will not be as intense as in documents tagged for *Physical Abuse* without this qualifier.

Pragmatic Import Pragmatic import is necessary for units of tags whose meaning goes beyond mere word semantics. The pragmatic understanding of meaning units is specific to the domain that is studied. One example in the context of warning tags are references to the discourse about a warning. The example *Implied Physical Abuse* from above falls into this category. Readers can infer that *Physical Abuse* takes place in the story but that it will not be explicit. Another example of pragmatic import for warning tags is hedging. In these cases, the authors use tags such as *Maybe Gaslighting* or *Abuse, I guess*. The hedging qualifiers convey that the author is uncertain if (1) the content can be characterized by the warning tag or (2) there is enough related content to warrant a warning. The final group of pragmatic covers units that explicitly state a warning. Instead of tagging a document for *Sexual Abuse*, some authors choose to use tags like *Trigger Warning for Sexual Abuse* or *CW: Sexual Abuse*. These qualifiers make it explicit that content may be sensitive to some readers.

[1] In linguistic terminology, these units are called modifiers.

Chapter 3

Testing the Consistency of Tags and Vocabulary

This chapter describes the methodological approach used to test how consistent a set of descriptors, commonly referred to as tags, is with the vocabulary of the documents they are describing. Consistency in this context means that the documents contain terms associated with the common language understanding of their tags with a higher frequency than a comparable set of baseline documents. This will be further elaborated on in Section 3.1. Building on the problem statement, Section 3.2 describes how warning tags can be categorized according to the warnings they convey (Section 3.2.1) and the qualifiers of that warning (Section 3.2.2). Following the same categorization, Section 3.3 outlines how to create a category vocabulary, calculate the term frequencies for each term from the vocabulary, and use the term frequencies to perform a range of statistical tests on the consistency between tags and expected vocabulary.

3.1 Problem Statement

Let D be a set of documents d. Each document d contains terms t from a vocabulary T and is assigned a set X_d of descriptors x, called tags, that give an overview on the content of d. The union of all documents tags X_d forms X, the set of all tags assigned to documents in D.

Depending on the goal and domain of the tag analysis, a set C of different categories c_i can be defined. Each category represents a different concept in the analysis domain. Based on these categories, the tags $x \in X$ can be assigned to possibly overlapping sets $X_{c_1}, X_{c_2}, \dots$ with $X_{c_i} \subseteq X$. A tag x is assigned to a tag set X_{c_i} if x is (semantically) related to the category c_i. Following the same categorization, let $T_{c_i} \subseteq T$ be the set of terms $t \in T$ that are semantically related to category c_i. These sets T_{c_i} reflect the vocabulary that is expected for documents with a tag x from X_{c_i}. Based on the categories c_i, the documents in D can be split into two disjoint sets D_{c_i} and $\hat{D}_{c_i}$:

$$D_{c_i} = \{d \in D | X_d \cap X_{c_i} \neq \emptyset\} \tag{3.1}$$

$$\hat{D}_{c_i} = \{\hat{d} \in D | X_{\hat{d}} \cap X_{c_i} = \emptyset\} \tag{3.2}$$

The set D_{c_i} contains all documents tagged for any of the category tags $x \in X_{c_i}$. Correspondingly, the documents in $\hat{D}_{c_i}$ do not have any of the tags in X_{c_i}. In the following, D_{c_i} will be referred to as the category corpus and $\hat{D}_{c_i}$ as the baseline corpus for category c_i. If required by the analysis, the baseline corpus $\hat{D}_{c_i}$ can be restricted further by choosing an additional set of tags X_{c_j}. The set X_{c_j} contains tags that, while not being directly related to the category c_i, might introduce noise into the experiments. An example for that is the warning category *Neglect*, that can be both a form of *Physical* as well as *Emotional Abuse*. Hence, the baseline corpus $\hat{D}_{c_i}$ for the category $c_i =$ *Emotional Abuse* contains documents that are tagged neither for any of the tags in $x \in X_{c_i}$, nor for any of the tags $x \in X_{c_j}$ related to $c_j =$ *Neglect*.

$$\hat{D}_{c_i} = \{\hat{d} \in D | X_{\hat{d}} \cap (X_{c_i} \cup X_{c_j}) = \emptyset\} \tag{3.3}$$

The tags $x \in X_{c_i}$ are assigned in a lexically consistent fashion if the probability $P(t|d)$ of observing a term $t \in T_{c_i}$ is significantly higher for category documents $d \in D_{c_i}$ than $P(t|\hat{d})$ for baseline documents $\hat{d} \in \hat{D}_{c_i}$. To summarize, the following steps are required to test for lexical consistency:

1. Specify a set C of categories c_i
2. Assign tags $x \in X$ to category sets X_{c_i}
3. Create pairs of document sets D_{c_i} and $\hat{D}_{c_i}$ based on the tag sets X_{c_i}
4. Define expected vocabularies T_{c_i} for each category $c_i \in C$
5. Apply statistical tests to the term frequencies for $t \in T_{c_i}$ in D_{c_i} and $\hat{D}_{c_i}$

3.2 Tag Categorization

The first step towards analyzing the consistency between category tags $x \in X_{c_i}$ and a vocabulary of category terms $t \in T_{c_i}$ is to define the set of categories C. The process of defining the categories can differ based on the domain of the documents and the goal of the analysis. For some experiments, it might be required to define a sufficient number of categories c_i as to cover all available tags $x \in X$, while other settings only focus on a subset of domain-specific tags $\tilde{X} \subseteq X$. To use a specific examples, the set X of all tags on AO3 includes a lot of fandom-related tags like characters mentioned in the story. For the analysis of trigger warnings, only the subset $\tilde{X} \subseteq X$ of tags that indicate a warning are relevant. Once the categories are specified, as many of the domain-specific tags as possible need to be identified and mapped to their corresponding categories. Both low recall as well as low precision in mapping tags to categories will introduce noise into downstream analyses by mixing documents tagged for a category with those that are not.

3.2.1 Warning Categories

This section describes the specific process for assigning tags to a set of warning categories used in the dataset WTWC-22 of fanfiction works. As discussed in Section 2.1.2, the authors on AO3 assign warnings using a freeform field when publishing their stories on the platform. As this introduces a lot of lexical diversity in the available tags, volunteers called tag wranglers create relations between synonymous tags or those that define a sub concept of another tag. This creates a graph of tags with canonical root nodes that define broad concepts such as *Humor*, *Sexual Content* or *Friendship*.[1]

Wiegmann et al. [2023] built on the relations created by the tag wranglers to construct a graph of tags related to their taxonomy of trigger warnings (see Section 2.1.2). Using a warning as the root node and traversing the graph along all edges that indicate synonymy or a sub concept results in a set of tags related to that warning. For the root node *Abuse*, the traversal returns tags such as *Abuse of Authority*, *Discussion of Sexual Abuse* or *Evidence of Physical Abuse*.

Depending on the granularity of the experiment, the goal can be either to test the consistency between different warnings or within categories of a single warning. In the former case, the traversal results can be used directly to construct the category sets X_{c_i}. For example, one can collect all tags related to the warnings and root nodes $c_1 =$ *Death*, $c_2 =$ *Discrimination*, and $c_3 =$ *Violence* to construct the respective tag sets X_{c_1}, X_{c_2}, and X_{c_3}. For the warning *Death*, the category corpus D_{c_1} would then contain all documents tagged for any of tags $x \in X_{c_1}$ and the baseline corpus $\hat{D}_{c_1}$ would consist of documents tagged for either *Discrimination* or *Violence*. This setup can then be used to test if documents tagged for *Death* use a death-related vocabulary significantly more often than documents tagged for one of the other two warnings.

In this thesis, the experiments are conducted on the more granular level of testing the categories within a warning. Hence, the warning tags collected

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

with the graph traversal are the domain-specific tags $\tilde{X}$. These tags are then divided manually into the category sets X_{c_i} within the warning. To use another example, $\tilde{X}$ is formed by collecting all tags related to the root node and warning *Abuse*. The tags $x \in \tilde{X}$ are then manually assigned to categories of the warning *Abuse* such as $c_1 =$ *Emotional Abuse* and $c_2 =$ *Physical Abuse*. In this setup, the category corpus D_{c_1} would be analogous to the one in the previous example and contain all documents tagged for *Emotional Abuse*. The baseline corpus $\hat{D}_{c_1}$, however, would not only contain documents tagged for *Physical Abuse* but also all documents tagged for any of the tags $x \in \tilde{X}$ related to *Abuse*.

3.2.2 Qualifier Categories

In addition to the warning categories, tags can be categorized according to meaning units in the tags (see Section 2.3.2). For the context of tags that express a warning, the focus lies on meaning units that qualify the warning. These *qualifiers* can be derived either directly from word semantics (content qualifiers) or through pragmatic import based on meaning units that occur in a lot of tags but are not covered by word semantics alone. While content qualifiers can be taken directly from the theory of the Functional Generative Description, the level of pragmatic import is informed by the data. The unit *trigger warning for*, for instance, is not common in other contexts but a relevant meaning unit when studying tags that convey a warning. It is thus included with pragmatic import.

The qualifier categories are formed by manually annotating the set of all domain-specific tags $\tilde{X}$ for the qualifiers they contain. As the qualifiers on the level of pragmatic import are informed by the data, these qualifiers are created during the annotation process.

The qualifiers do not change which warning category a tag is related to. The tags *Physical Abuse*, *Implied Physical Abuse*, and *TW: Physical Abuse* all belong to the same warning category $c_i =$ *Physical Abuse*. In addition to that category, *Implied Physical Abuse* and *TW: Physical Abuse* are also part of different subcategories based on the qualifiers they contain. More formally, the tags that contain a specific qualifier are assigned to a subcategory $c_{i,q}$, where c_i refers to the warning category and q to the qualifier. In the example above, q is *implied* for the tag *Implied Physical Abuse* and *warning* for the tag *TW: Physical Abuse*.

3.3 Vocabulary Testing

This section describes the methodology for testing the consistency of the warning tags with the expected vocabulary for the corresponding warning. The first step consists of collecting a vocabulary of expected terms (Section 3.3.1). Then, the document- and corpus-level term frequencies are calculated for these terms (Section 3.3.2). Finally, the term frequencies are used to perform significance testing (Section 3.3.3), calculate distributions of log ratios (Section 3.3.4), and analyze how the log ratios change when restricting the category corpus D_{c_i} to documents whose warning tags are modified by a qualifier (Section 3.3.5). Section 3.3.6 describes an approach using the information content equation from the DPH retrieval model that we discarded after an initial pilot study.

3.3.1 Vocabulary Collection

After specifying the set C of categories c_i, the category-specific vocabularies T_{c_i} can be defined. The starting point for that process is a set of seed terms, ideally from an authoritative source or manual collection if no source is available. Building on this set of seed terms, the vocabulary is expanded by collecting synonyms for each term $t \in T_{c_i}$. This can be done both manually using thesauri or semi-automatically by prompting an LLM for synonym generation and filtering the results. The set is finalized by expanding it with versions of all terms $t \in T_{c_i}$ that correspond to different syntactic categories with the same word stem, such as the verb *abuse* and the adjectives *abused* and *abusive* for the noun *abuse*. For greater specificity, each term can be accompanied by a part-of-speech (POS) tag.

3.3.2 Term Frequencies

The next step consists of calculating the (normalized) term frequencies. For a given category c_i, this requires the creation of a category corpus D_{c_i} that contains documents with category-related tags and a baseline corpus $\hat{D}_{c_i}$ with other tags. For each term t in the category vocabulary T_{c_i}, we then calculate a document-level term frequency $\mathrm{tf}(t, d)$ and corpus-level term frequency $\mathrm{TF}(t, D)$ for both the category corpus and the baseline corpus. The document-level term frequency $\mathrm{tf}(t, d)$ will be used for significance testing and the corpus-level term frequency $\mathrm{TF}(t, D)$ will used to calculate the log ratio. The two measures are calculated as follows,

$$\mathrm{tf}(t, d) = \frac{\mathrm{f}_{t,d}}{\sum_{t' \in d} \mathrm{f}_{t',d}} \tag{3.4}$$

$$\mathrm{TF}(t, D) = \frac{\sum_{d \in D} \mathrm{f}_{t,d}}{\sum_{d \in D} \sum_{t' \in d} \mathrm{f}_{t',d}} \tag{3.5}$$

where D is a set of documents and $t \in T_{c_i}$ is a term in the category vocabulary with raw count $\mathrm{f}_{t,d}$ in a document $d \in D$. The sum $\sum_{t' \in d} \mathrm{f}_{t',d}$ is equal to the total number of terms in d, or the document length, sometimes denoted as $|d|$. In order to calculate the term frequencies, each document $d \in D_{c_i}$ and $\hat{d} \in \hat{D}_{c_i}$ is tokenized, POS-tagged and lemmatized. Lemmatization is necessary to treat all forms of a word with the same stem and POS-tag as one term t. POS-tagging can be used to restrict analyses to only specific categories like verbs or adjectives.

3.3.3 Significant Differences in Term Frequencies

The first method is concerned with testing the statistical significance of differences in term frequencies between the category corpus D_{c_i} and the baseline corpus $\hat{D}_{c_i}$. For each of the terms t in the category vocabulary T_{c_i} and documents $d \in D_{c_i}$ and $\hat{d} \in \hat{D}_{c_i}$, calculate the document-level term frequencies $\mathrm{tf}(t, d)$ and $\mathrm{tf}(t, \hat{d})$. Then, apply the Mann-Whitney U-test as outlined by Lijffijt et al. [2014]: For a given term t, jointly rank all documents in ascending order of their term frequencies. The test statistic U is then calculated as follows.

$$U_1 = n_1 n_2 + \frac{n_1(n_1 + 1)}{2} - R_1 \tag{3.6}$$

$$U_2 = n_1 n_2 + \frac{n_2(n_2 + 1)}{2} - R_2 \tag{3.7}$$

$$U = min(U_1, U_2) \tag{3.8}$$

The variable $n_1 = |D_{c_i}|$ is the number of documents in D_{c_i} and R_1 is the sum of ranks for all documents $d \in D_{c_i}$. The variables n_2 and R_2 are the corresponding values for the baseline corpus $\hat{D}_{c_i}$. The metric U_1 is the number of pairs between documents $d \in D_{c_i}$ and $\hat{d} \in \hat{D}_{c_i}$ for which $\mathrm{tf}(t, d) < \mathrm{tf}(t, \hat{d})$. For sample sizes of $n > 25$, the distribution of U is well approximated by a Gaussian distribution. The corresponding z-score is calculated as follows:

$$\mu_U = \frac{n_1 n_2}{2} \tag{3.9}$$

$$\sigma_U = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}} \tag{3.10}$$

$$z = \frac{U - \mu_U}{\sigma_U} \tag{3.11}$$

Tied ranks in the Mann-Whitney U-test are handled by assigning the mean rank to all samples with the same value. If for example, three documents with the same term frequency would be assigned to the ranks 3, 4, 5, and 6, they all receive the rank $\frac{3+4+5+6}{4} = 4.5$ and the subsequent sample receives rank 7. In cases with tied ranks, the equation for standard deviation is adjusted as follows

$$\sigma_U = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12} - \frac{n_1 n_2 \sum_{r=1}^{R} ({l_r}^3 - l_r)}{12n(n-1)}} \tag{3.12}$$

where $n = n_1 + n_2$ is the total number of documents, $r \in [1, R]$ is a unique rank with ties, and l_r is the number of ties for rank r. Accounting for tied ranks is especially relevant in the setting at hand as for large corpora D and rare terms t, the term frequency $\text{tf}(t, d)$ will be 0 for a lot of documents $d \in D$. By adjusting the standard deviation accordingly, the z-score for rare terms will not be underestimated. From the z-scores, we then calculate p-values to test the following hypotheses for each term $t \in T_{c_i}$ in the category vocabulary.

$$H_0 : F_{c_i} = \hat{F}_{c_i}, \quad H_A : F_{c_i} \neq \hat{F}_{c_i}$$

where F_{c_i} is the distribution of term frequencies $\text{tf}(t, d)$ for $d \in D_{c_i}$ and $\hat{F}_{c_i}$ is the distribution of $\text{tf}(t, \hat{d})$ for $\hat{d} \in \hat{D}_{c_i}$. As we perform multiple tests, one for each term t, we need to account for the risk of incorrectly rejecting an individual null hypothesis through multiplicity. This is done by adjusting the p-values using the Bonferroni correction. After calculating p-values from the z-scores, they are multiplied by $|T_{c_i}|$, the number of terms t in the category vocabulary. For a given level of significance α, only rejecting H_0 for a term t if the corrected p-value $p*$ is below α, ensures that the family-wise error rate (FWER) is controlled by α:

$$p^* = p * |T_{c_i}| \leq \alpha$$

The approach described above identifies those terms $t \in T_{c_i}$ that occur significantly more often in the category corpus D_{c_i} than in the baseline corpus $\hat{D}_{c_i}$. However, significance does not answer the question of *how much* more frequent a term t is in D_{c_i} than in $\hat{D}_{c_i}$. Calculating the common language effect size $\frac{U_2}{n_1 n_2}$ only gives the share of pairs between documents for which $\text{tf}(t, d \in D_{c_i}) > \text{tf}(t, \hat{d} \in \hat{D}_{c_i})$. In order to also test the magnitude of differences in term frequencies between the two corpora, we will use the distribution of log ratios described in the following subsection.

3.3.4 Distribution of Log Ratios

The second measure that is calculated for each term $t \in T_{c_i}$ is the log ratio $\mathrm{lr}(t)$ of the corpus-level term frequencies $\mathrm{TF}(t, D)$ (See Section 3.3.2). For a given term t, it is calculated as follows

$$\mathrm{lr}(t) = \log_2 \frac{\mathrm{TF}(t, D_{c_i})}{\mathrm{TF}(t, \hat{D}_{c_i})} \tag{3.13}$$

with D_{c_i} and $\hat{D}_{c_i}$ being the category and baseline corpora, respectively. It is important to note that the log ratio cannot be used to estimate significance on the level of individual terms t. Instead, the distribution of log ratios for the entire category vocabulary T_{c_i} will be tested. As the values for $\mathrm{TF}(t, D)$ lie on the interval $[0, 1]$, their ratios follow a heavily right-tailed distribution on the interval $[0, \infty)$. By applying a logarithmic transformation, the distribution of ratios can be made symmetric. Using the binary logarithm makes it easier to interpret the log ratios as every full unit increase in $\mathrm{lr}(t)$ is equal to a doubling of the ratio. In case of similar vocabularies and term frequencies between the two corpora D_{c_i} and $\hat{D}_{c_i}$, many terms t will have a similar frequency in both, leading their ratio to be close to 1 and their log ratio to be close to 0. For a vocabulary of terms that have higher frequencies in the category corpus D_{c_i}, however, the mean of log ratios will be greater than 0. By calculating the log ratios $\mathrm{lr}(t)$ for all $t \in T_{c_i}$, we obtain a distribution of log ratios LR_{c_i} for the category vocabulary. Following from the reasoning above, we expect the mean of that distribution to be different from 0. This can be formalized in the following hypotheses:

$$H_0 : \overline{\mathrm{lr}}_{c_i} = 0, \quad H_A : \overline{\mathrm{lr}}_{c_i} \neq 0$$

$$\overline{\mathrm{lr}}_{c_i} = \frac{1}{|T_{c_i}|} \sum_{t \in T_{c_i}} \mathrm{lr}(t) \tag{3.14}$$

Given that the distribution of log ratios will be normally distributed under the null hypothesis, a one sample t-test can be used. This test on the mean of log ratios $\overline{\mathrm{lr}}_{c_i}$ for the entire category vocabulary T_{c_i} complements the term-specific significance testing described in the previous subsection with an effect size measure. By combining both, the following two questions can be answered:

1. Do the individual terms t from the category vocabulary T_{c_i} occur significantly more frequent in the category corpus D_{c_i} than in the baseline corpus $\hat{D}_{c_i}$?

2. Is the mean increase in frequency over the entire vocabulary is significantly different from 0?

3.3.5 Effect of Qualifiers

Section 3.2.2 described how to assign tags to subcategories $c_{i,q}$ according to the qualifiers they contain. As these subcategories do not change which warning category c_i a tag belongs to, documents with tags related to any of the qualifier subcategories $c_{i,q}$ are tested for the same vocabulary T_{c_i}. The goal of the analysis is not to test consistency with a qualifier-specific vocabulary. Instead, tests will explore if the restriction of the set of category tags X_{c_i} to a subset $X_{c_{i,q}}$ of tags that contain the warning category c_i modified by the qualifier q leads to similar effects across the different categories. To use a specific example, the extent qualifier *a lot* suggests that more of the category-specific vocabulary is present in a document d than without the qualifier. The tests will explore if documents with tagged for *A lot of Emotional Abuse*, *A lot of Physical Abuse* and *A lot of Sexual Abuse* contain vocabulary terms for the respective categories with a higher frequency than all documents tagged for *Emotional Abuse*, *Physical Abuse*, and *Sexual Abuse* without the restriction to the qualifier *a lot*.

The hypotheses will be tested by comparing the mean log ratio $\overline{\text{lr}}$ for category documents $d \in D_{c_i}$ without a qualifier restriction to the one observed for documents $d \in D_{c_{i,q}}$ tagged for a warning category c_i modified by a qualifier q. This leads to the following hypotheses:

$$H_0 : \overline{\text{lr}}_{c_{i,q}} = \overline{\text{lr}}_{c_i}, \quad H_A : \overline{\text{lr}}_{c_{i,q}} \neq \overline{\text{lr}}_{c_i}$$

Depending on the expected change in vocabulary prevalence, the alternative hypothesis can be $H_A : \overline{\text{lr}}_{c_{i,q}} > \overline{\text{lr}}_{c_i}$ for an increase and $H_A : \overline{\text{lr}}_{c_{i,q}} < \overline{\text{lr}}_{c_i}$ for a decrease in log ratio. The baseline corpus $\hat{D}_{c_i}$ for calculating the log ratios is the same for both D_{c_i} and $D_{c_{i,q}}$. Similar to the term significance tests described in Section 3.3.2, we need to account for multiplicity as the significance is tested for multiple qualifiers. Hence, the Bonferroni-corrected p^* is calculated by multiplying the p-value with the number of qualifiers that were tested.

The differences in mean log ratio indicate if the vocabulary terms are observed *on average* more or less frequent when restricting the documents to a qualifier subcategory $c_{i,q}$. The tests do not indicate if this is due to the same or due to different terms. In order to answer this question, the vocabulary terms are ranked by their log ratio for both the category corpus D_{c_i} and the qualifier subcategory corpus $D_{c_{i,q}}$. This allows us to compute Kendall's τ, a measure for rank correlation. It is computed by looking at all pairs of terms and counting how often their relative position to each other is the same for both rankings.

The number of concordant pairs P_C states how many pairs are ranked equally to each other in both rankings and the number of disconcordant pairs P_D states how many pairs are ranked differently. This leads to the following calculation of τ:

$$\tau = \frac{\mathrm{P}_C - \mathrm{P}_D}{\mathrm{P}_C + \mathrm{P}_D} \tag{3.15}$$

The value will be $\tau = 1$ for perfectly aligned rankings and $\tau = -1$ for inverse rankings. A value close to 0 indicates no correlation. For rankings with a sufficient number of shared terms k, at least $k > 10$, between the rankings, a z-score can be calculated as follows:

$$z = \frac{3\tau\sqrt{k(k-1)}}{\sqrt{2(2k+5)}} \tag{3.16}$$

Using this z-score, we can test the following hypotheses:

$$H_0 : \tau = 0, \quad H_A : \tau \neq 0$$

Rejecting the null hypothesis in this setting means that the correlation of term rankings between the category corpus D_{c_i} and the qualifier subcategory corpus $D_{c_{i,q}}$ is significantly different from 0. If the *same* terms are expected to have a high log ratio for the subcategory, the alternative hypothesis becomes $H_A : \tau > 0$ as we expect a positive rank correlation. For the opposite case of an inverse ranking of the terms, we expect a negative correlation and the alternative hypothesis becomes $H_A : \tau < 0$.

3.3.6 Information Content

In a pilot study, we explored using the information content of the DPH retrieval model by Amati [2006] as a measure for identifying warning-specific terms. Based on the results of the study, we discarded this method but wanted to include it in the thesis for the sake of completeness. The DPH model is a probabilistic retrieval model that estimates the probability of a document d being relevant to a query by calculating the information content of each query term t for d. Abstractly speaking, the information content of a term t is high for a document d if t is relatively rare for the set of all documents D and relatively common for d. For large document collections, the information content $\mathrm{Inf}(\mathrm{f}_{t,d}||d)$ of t for d can be calculated using a binomial distribution:

$$B(|d|, \mathrm{f}_{t,d}, \mathrm{TF}(t, D)) = \binom{|d|}{\mathrm{f}_{t,d}} \mathrm{TF}(t, D)^{\mathrm{f}_{t,d}} (1 - \mathrm{TF}(t, D))^{|d| - \mathrm{f}_{t,d}} \tag{3.17}$$

$$\mathrm{Inf}(\mathrm{f}_{t,d}||d) = -\log_2(B(|d|, \mathrm{f}_{t,d}, \mathrm{TF}(t, D))) \tag{3.18}$$

In the original model, $\mathrm{TF}(t, D)$ is the corpus-level term frequency of t in the corpus D and $|d|$ is the number of terms in d. For our use case, we adapted the equation by using the baseline corpus $\hat{D}_{c_i}$ to calculate the term frequency $\mathrm{TF}(t, \hat{D}_{c_i})$. In addition to that, we replaced the individual document d in the equation with the category corpus D_{c_i}. Hence, instead of comparing the observed occurrences of a term t in a document d with the expected occurrences based on the corpus of all documents D, we compared the observed occurrences of t in the category corpus with the expected occurrences based on the baseline corpus:

$$\mathrm{Inf}(\mathrm{f}_{t,D_{c_i}}||D_{c_i}) = -\log_2(B(|D_{c_i}|, \mathrm{f}_{t,D_{c_i}}, \mathrm{TF}(t, \hat{D}_{c_i}))) \tag{3.19}$$

The rationale was that terms that are rare for the baseline corpus $\hat{D}_{c_i}$ and frequent for the category corpus D_{c_i}, in other words have a high information value, would be specific to the category c_i. In our pilot study, however, we found the approach to return noisy results such as terms associated with fandoms like characters or places. Even after creating stopword lists that removed fandom-specific terms, the results remained noisy.

Chapter 4

Testing the Effect of Prescriptive Annotation Guidelines

This chapter describes the methodology for testing the effect of prescriptive annotation guidelines on annotator agreement in the task of annotating text passages for trigger warnings. In this thesis, the effect is not evaluated in a study with human annotators but by using the sociodemographic prompting of LLMs to predict disagreements between annotators as conducted by Wan et al. [2023] and Beck et al. [2024]. However, the process, barring Section 4.2, can be largely transferred to a study with human annotators. Section 4.1 describes how to select passages for annotation using different groups of terms from the consistency tests for dictionary-based retrieval. Section 4.2 outlines how an LLM can be prompted to predict the annotation decision by people with different sociodemographic profiles. These profiles are integrated into the three annotation prompts presented in Section 4.3. The prompts can be distinguished by both the paradigm they apply (descriptive vs. prescriptive) and the specificity of their annotation question (warning vs. warning category). Section 4.4 concludes the chapter by describing how to perform hypothesis testing on the distributions of annotator agreements.

4.1 Passage Retrieval

In some annotation settings for NLP tasks, the unit of annotation is the full document. Examples are annotating social media posts for hate speech (Kennedy et al. [2022]) or ratings on an eCommerce platform for their sentiment (Mohammad et al. [2016]). For longer documents tagged for trigger warnings, however, only a few sentences might be responsible for a warning tag, while the rest of the text discusses other topics (Wiegmann et al. [2024]).

Providing annotators with the full document risks receiving lower quality annotations by providing unnecessarily long annotation samples (Goyal et al. [2022]). Instead, passages for annotation are selected from the documents by building on the methodology described in Chapter 3. After defining the category vocabularies T_{c_i}, performing significance tests on the term frequencies $\text{tf}(t, d)$ for terms t in these vocabularies and calculating their log ratios $\text{lr}(t)$, the results can be used to identify promising passages for the category of interest c_i. Similar to the approach by Wiegmann et al. [2024], passages will be selected using dictionary-based retrieval. Specifically, we will use the following three groups of terms terms with a high z-score or a high log ratio for category-specific passage retrieval:

1. **Group A**: Top k terms $t \in T_{c_i}$ with the highest z-score
2. **Group B**: Top k terms $t \in T_{c_i}$ with the highest log ratio $\text{lr}(t)$, restricted to terms with a significantly higher term frequency ($p^* < \alpha$)
3. **Group C**: Top k terms $t \notin T_{c_i}$ with the highest z-score that occur in at least 5,000 documents of the category corpus D_{c_i}

The terms in group A and B can be derived directly from the results of the methodology of Chapter 3. Group C requires performing the Mann-Whitney U-test on all terms $t \in T$ that occur in the documents $d \in D_{c_i}$. This is an explorative approach to collect more terms for passage retrieval as opposed to doing significance testing on an expected vocabulary. The rationale behind the different groups is as follows: The first group contains more common vocabulary terms as the z-score will be higher for those that occur in a lot of documents. The second group is expected to contain terms that are more specific to the category than those from group A as their log ratio, the effect size for the relative difference in term frequency on corpus level, is the highest of all vocabulary terms. The third and final group is similar to the first but not restricted to the category vocabulary. It is expected to yield passages that are harder to annotate by containing terms not commonly associated with the category, yet more frequently used in documents from the category corpus.

The initial set of passages S_{c_i} for a category c_i is formed by (1) retrieving all documents d in the category corpus D_{c_i} that contain a term from one of the three groups, (2) identifying the sentences that contain a query term, (3) adding context to each sentence by extending the passage to a number of up to three neighboring sentences in both directions, and (4) merging overlapping passages into one. For the actual annotation, S_{c_i} is reduced to a number s of passages by sampling uniformly for each term in groups A-C.

Figure 4.1: Example for an Underspecified Prompt. It first defines the sociodemographic profile, then asks to imagine having had a warning-related experience, and finally poses the annotation question.

> 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 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 abuse ('yes' or 'no'):

4.2 Sociodemographic Prompting

The specific setting of using LLMs to predict disagreement between human annotators requires the creation of sociodemographic profiles. One part of the experiments conducted by Beck et al. [2024] was to analyze which attributes were the most influential in changing the labels when compared to zero-shot predictions without a sociodemographic profile. The authors found that, in general, a combination of attributes leads to the most label changes (63% across models and datasets), with race and political affiliation being the most influential individual attributes. Hence, in order to increase the diversity of annotations by different prompts, the set of sociodemographic profiles P should be diverse with respect to all attributes but most importantly these two.

Each profile $p \in P$ can then be used to construct a sociodemographic prompt. As Beck et al. [2024] found a negative interaction effect between the additional prompt length and sociodemographic prompting, the prompt is designed to be as short as possible. An example prompt can be found in Figure 4.1.

4.3 Annotation Prompts

Following the examples given by Rottger et al. [2022] in their study on the effectiveness of prescriptive annotation guidelines, three different annotation prompts are created:

1. **Underspecified** (Descriptive): This prompt asks to make the subjective decision if a given passage could lead to associations with previous trauma. It is underspecified in the sense that it asks for broad warnings such as *Abuse* or *Violence*.

2. **Category-Specific** (Descriptive): This prompt is the same as the first with the exception of asking for specific categories of a warning like *Emotional Abuse* or *War-related Violence*.

3. **Prescriptive**: This prompt does not ask whether a given passage might cause associations with trauma but if acts or consequences of a category of a warning are contained in the passage. It also provides examples for a category.

The primary comparison is between the underspecified prompt and the prescriptive prompt. The category-specific prompt functions analogously to the control prompt used by Rottger et al. [2022]. It is used to control for the effects of asking for a detailed category as opposed to a broader warning. An example for the underspecified prompt is given in Figure 4.1 and all prompts are illustrated in Table 5.4a in the experimental setup.

4.4 Testing the Effect on Annotator Agreement

By combining the annotation prompts with each of the sociodemographic profiles P, a total of $3 \cdot |P|$ prompt-combinations is created. These full prompts are used to get annotations on each of the s passages sampled as described in Section 4.1. The annotations are generated by prompting `Flan-T5 11B` as Beck et al. [2024] found this model to perform best at the task of predicting disagreement between human annotators. Compared to experiments with human annotators, the only difference lies in not providing them with the part of the prompt related to the sociodemographic profile. Otherwise, the process can be applied in the same way.

After recording the annotations, a total of $\frac{|P|(|P|-1)}{2}$ pairwise annotator agreements are calculated for each of the three annotation prompts using Cohen's κ. This creates three distribution of annotator agreements K_U, K_C, and K_P for the underspecified, category-specific, and prescriptive prompt, respectively. The hypotheses testing on these distributions will be done using their means $\overline{\kappa}_U$, $\overline{\kappa}_C$, and $\overline{\kappa}_P$:

$$H_0 : \overline{\kappa}_P = \overline{\kappa}_U, \quad H_A : \overline{\kappa}_P \neq \overline{\kappa}_U$$

$$H_0 : \overline{\kappa}_P = \overline{\kappa}_C, \quad H_A : \overline{\kappa}_P \neq \overline{\kappa}_C$$

As our experiments showed the distributions K to be roughly normally distributed, a two sample t-test is applied to test the hypotheses. If the null hypotheses can be rejected in both cases, then there are significant differences in annotator agreements between the prescriptive prompt on the one side and the underspecified as well as the category-specific prompt on the other side.

Chapter 5

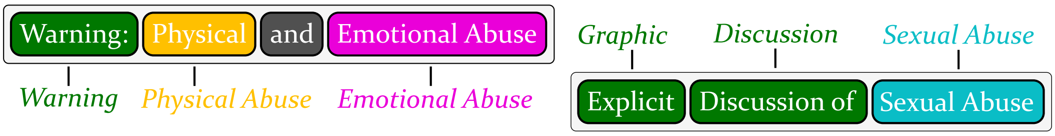
Dataset & Experimental Setup

This chapter gives an overview on the experiments conducted in this thesis as well as the dataset that they were conducted on. Section 5.1 describes the dataset of chapters from AO3 on which all experiments were performed. The following section is divided into three parts: Section 5.2.1 describes the experimental setup for testing the consistency between tags and vocabulary. It presents the warning categories that were tested, how the vocabulary for each category was collected, and concludes by stating the hypotheses. Section 5.2.2 gives an overview on how the effects of qualifiers were tested. It outlines the tag annotation process and states the hypotheses for each qualifier. Section 5.2.3, finally, details the setup for testing the effect of prescriptive annotation guidelines by describing the selection of passages for annotation, the annotation prompts, and stating the hypotheses.

5.1 Dataset

The dataset used in all experiments of this thesis is the WTWC-22, created by Wiegmann et al. [2023]. It consists of a total of 7.8 million fanfiction works from AO3 divided into 21.8 million chapters. The vast majority, 76 % of all works, consist of only one chapter, while 6 % comprise ten or more chapters. The median number of words per work is 3,096 with 49 % of all works having between 1,000 and 5,000 words. English is by far the most frequent language with 7.1 million works, followed by Mandarin with 0.4 million and Russian with 0.1 million. As described in Section 2.1.2, the authors derived a taxonomy of 36 warning categories and assigned freeform tags given by authors of the fanfiction works to these categories.

Figure 5.1: Examples for the Tag Annotation. The units *warning*, *explicit* and *discussion of* are qualifiers. The other highlighted parts indicate the warning category.



5.2 Experimental Setup

We conducted two experiments on the dataset: First, we tested the consistency between vocabulary and warning tags by applying the method from Chapter 3 to the subset of all chapters in the WTWC-22 that are tagged for some form of *Abuse*. This is discussed in detail in Sections 5.2.1 and 5.2.2. For the creation of the chapter subset, the tag *Abuse* was used as the root node in a tag graph to collect a set $\tilde{X}$ of 5,654 synonymous tags and subtags (see Section 3.2.1). The 2,965,898 chapters with at least one of these tags constituted the full set of documents D. Second, from that same subset, we retrieved passages and obtained annotations as described in Chapter 4 to test the effect of prescriptive annotation guidelines on annotator agreement. The setup for the annotation experiments is described in Section 5.2.3.

5.2.1 Consistency between Tags and Vocabulary

Warning Categories The set of categories C consisted of *Emotional Abuse*, *Physical Abuse*, and *Sexual Abuse*. To be able to construct the respective category corpora D_{c_i}, each of the $5,654$ tags was manually annotated for these three categories as well as for *Neglect*. The annotation process was informed by information pages offered by two governmental institutions in the US and UK, the Washington State Department of Social and Health Services[1] and Social Care Institute for Excellence.[2] Both institutions provide an overview of different types of *Abuse* as well as signs and indicators to identify them. Most tags, however, use the literal name of the category (e.g. *Emotional Abuse*) instead of forms of that *Abuse* category (e.g. *Gaslighting*) and were thus easy to annotate.

Example annotations can be found in Figure 5.1. In addition to the warning categories *Emotional Abuse*, *Physical Abuse*, and *Sexual Abuse*, the figure also highlights the qualifier annotation discussed in Section 5.2.2. The first example shows that one tag can contain multiple warning categories at the same time.

[1] https://www.dshs.wa.gov

[2] https://www.scie.org.uk

Table 5.1: Number of Tags and Chapters per Warning Category

Category c_i	**# Tags**	**# Chapters in D_{c_i}**	**# Chapters in $\hat{D}_{c_i}$**
Emotional Abuse	515	49,989	2,872,109
Physical Abuse	694	125,408	2,795,274
Sexual Abuse	618	107,433	2,858,465

Documents tagged with this specific tag are assigned to both the category corpus D_{c_i} for *Emotional Abuse* as well as the one for *Physical Abuse*.

The category *Neglect* was included in the annotations as it can constitute both *Emotional* and *Physical Abuse* and could introduce noise into the experiment results. Consequently, the baseline corpora $\hat{D}_{c_i}$ for *Emotional* and *Physical Abuse* were constructed from documents with tags from neither their respective category set X_{c_i} nor from the set of tags X_{c_j} related to *Neglect*. Chapters tagged for both *Emotional* or *Physical Abuse* as well as *Neglect* were added to the respective category corpus D_{c_i}. In total, 515 tags were labeled to indicate *Emotional Abuse*, 694 *Physical Abuse*, and 618 *Sexual Abuse*. The filtering category *Neglect* was assigned to 310 tags. Table 5.1 gives an overview on the number of chapters in the category and baseline corpus for each category. The most chapters were tagged for *Physical Abuse*. *Emotional Abuse* was noticeably less common than the other two.

Vocabulary Collection The term vocabularies T_{c_i} for the three categories were created as outlined in Section 3.3.1.[3] The seed terms were derived from the same information pages on signs and indicators of *Abuse* mentioned in the previous paragraph. The set of seed terms was then expanded with both synonyms from the Merriam Webster thesaurus[4] as well as manually filtered suggestions by `GPT-4`. The vocabulary of category terms T_{c_i} was finalized by adding different syntactic categories of the terms already in the vocabulary. This resulted in a total of 400 terms for both *Emotional Abuse* and *Physical Abuse*, and 250 for *Sexual Abuse*.

Hypotheses In order to perform the significance tests both on the individual term frequency (see Section 3.3.3) as well as for the distribution of log ratios (see Section 3.3.4), the term frequencies were calculated both on document- and corpus-level for each of the categories. As the tags indicate the presence of a particular warning category, chapters tagged for them were expected to contain terms from the respective category vocabularies with a higher frequency

[3] The vocabularies are available in the code to this thesis or the internal repository.

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

than comparable baseline chapters. More formal, for each category c_i and each term $t \in T_{c_i}$, the distribution F_{c_i} of term frequencies $\mathrm{tf}(t, d)$ for $d \in D_{c_i}$ was expected to be greater than the distribution $\hat{F}_{c_i}$ of term frequencies for $\hat{d} \in \hat{D}_{c_i}$:

$$\forall c_i \in C : \forall t \in T_{c_i} : H_A : F_{c_i} > \hat{F}_{c_i}$$

In addition to that, the mean log ratio $\overline{\mathrm{lr}}$ for terms from each category vocabulary T_{c_i} was expected to be greater than 0:

$$\forall c_i \in C : H_A : \overline{\mathrm{lr}}_{c_i} > 0$$

5.2.2 Effect of Qualifiers

In addition to testing the consistency between tags and vocabulary, another set of experiments was concerned with analyzing the effects of qualifiers based on the principles of the Functional Generative Description described in Section 2.3.2.

Tag Annotation In order to test the effect that qualifiers have on the distribution of log ratios, each of the $5,654$ tags was also manually annotated for a set of 13 qualifiers. An overview of all tested qualifiers with examples is given in Table 5.2 and annotation examples are shown in Figure 5.1. Many of the qualifiers were introduced based on patterns in the tags used by authors on AO3 and were thus on the level of pragmatic import. Especially common were qualifiers that reference the discourse about a warning as an entity such as *description*, *discussion* or *mention*. Other qualifiers mediate the type of warning by either amplifying it with qualifiers such as *extreme* or *graphic* or reducing it with qualifiers such as *brief* and *light/mild* or through *hedging*.

Hypotheses After annotating the tags for qualifiers, each category corpus D_{c_i} was restricted to subcategory corpora $D_{c_{i,q}}$ that are tagged for a warning category c_i modified by a qualifier q. For each subcategory $c_{i,q}$, the corpora D_{c_i} and $D_{c_{i,q}}$ were used to test the two hypotheses outlined in Section 3.3.5. The expected effect for each hypothesis is given in the last two columns of Table 5.2. The first set of hypotheses, $H_A(\overline{lr})$, refers to the mean log ratios for both corpora and whether the qualifier restriction was expected to lead to an increase or decrease when compared to the log ratio of the base category c_i. The expected effect for all references to a discourse (*discussion*, *mention*, ...) was a decrease in mean log ratio as they indicate that the warning is not explicit in the document but something that is discussed or mentioned, for instance by characters in the story. A decrease in mean log ratio was also expected for

Table 5.2: Qualifier Hypotheses. The examples are meaning units of tags that were annotated for the qualifier. The values in the two hypotheses columns $H_A(\overline{\text{lr}})$ and $H_A(\tau)$ are coloured based on the expected effect. Expected decreases are marked in red, increases in green, and no significant effects in yellow.

Qualifier	**Example**	$\boldsymbol{H_A(\overline{\text{lr}})}$	$\boldsymbol{H_A(\tau)}$
A Little	*a bit of* abuse	$\overline{\text{lr}}_{c_{i,j}} < \overline{\text{lr}}_{c_i}$	$\tau > 0$
A Lot	*lots of* gaslighting	$\overline{\text{lr}}_{c_{i,j}} > \overline{\text{lr}}_{c_i}$	$\tau > 0$
Brief	*brief* description of abuse	$\overline{\text{lr}}_{c_{i,j}} < \overline{\text{lr}}_{c_i}$	$\tau > 0$
Description	*descriptions of* physical abuse	$\overline{\text{lr}}_{c_{i,j}} < \overline{\text{lr}}_{c_i}$	$\tau = 0$
Discussion	child abuse *discussed*	$\overline{\text{lr}}_{c_{i,j}} < \overline{\text{lr}}_{c_i}$	$\tau = 0$
Extreme	*extreme* emotional abuse	$\overline{\text{lr}}_{c_{i,j}} > \overline{\text{lr}}_{c_i}$	$\tau = 0$
Graphic	*graphic* sexual abuse	$\overline{\text{lr}}_{c_{i,j}} > \overline{\text{lr}}_{c_i}$	$\tau = 0$
Hedging	abuse, *I guess*	$\overline{\text{lr}}_{c_{i,j}} < \overline{\text{lr}}_{c_i}$	$\tau = 0$
Implied	abusive family *implied*	$\overline{\text{lr}}_{c_{i,j}} < \overline{\text{lr}}_{c_i}$	$\tau = 0$
Light/Mild	*mild* animal abuse	$\overline{\text{lr}}_{c_{i,j}} < \overline{\text{lr}}_{c_i}$	$\tau = 0$
Mention	*mention of* emotional abuse	$\overline{\text{lr}}_{c_{i,j}} < \overline{\text{lr}}_{c_i}$	$\tau = 0$
Non-Graphic	*non-graphic* [...] physical abuse	$\overline{\text{lr}}_{c_{i,j}} < \overline{\text{lr}}_{c_i}$	$\tau = 0$
Warning	*CW:* sexual abuse	$\overline{\text{lr}}_{c_{i,j}} > \overline{\text{lr}}_{c_i}$	$\tau > 0$

a little, *brief*, *light/mild*, *hedging*, and *non-graphic* as these qualifiers indicate fewer or less intense occurrences of the warning category. Analogous to that reasoning, the qualifiers *a lot*, *extreme*, *graphic*, and *warning* were all expected to lead to an increase in mean log ratio as they indicate more frequent or more intense occurrences of the warning category, or draw attention to the fact the content can be distressing.

The second set of hypotheses, $H_A(\tau)$, refers to the expected rank correlation of log ratios between subcategory $c_{i,q}$ and base category c_i. The restriction to a subcategory was expected to lead to a similar ranking of terms in only a few cases. This is indicated in the table by stating the alternative hypothesis as $H_A : \tau > 0$. A positive correlation was expected for the extent qualifiers *a little* and *a lot* as well as the temporal qualifier *brief*. All these qualifiers indicate a change in *how much* of warning-related content is present or *how long* it lasts. The warning itself is not changed. In addition to these qualifiers, a positive rank correlation was expected for *warning*. The qualifier draws attention to the fact that the content can be distressing to some audiences as opposed to restricting the warning to a specific form like the qualifiers *graphic* and *extreme*.

The rank correlation was not expected to be significantly different from 0 for the other qualifiers, meaning that the ranking of terms for the subcategory and base category are (largely) independent of each other. For warning categories qualified with a reference to a discourse, the expectation was that the

Table 5.3: Example Terms used for Passage Retrieval

	Emotional Abuse		**Physical Abuse**		**Sexual Abuse**	
Group	**Term**	**POS**	**Term**	**POS**	**Term**	**POS**
A	hurt	Verb	bruise	Noun	rape	Verb
	gaslighting	Noun	scared	Adj.	sexual	Adj.
	fear	Noun	beat	Verb	scared	Adj.
B	blaming	Adj.	thump	Verb	violating	Adj.
	ridiculing	Adj.	overmedication	Noun	hiv	Noun
	manipulated	Adj.	scalding	Adj.	molestation	Noun
C	sick	Adj.	tear	Noun	abuse	Noun
	flinch	Verb	abuse	Noun	hurt	Verb
	lie	Noun	hurt	Verb	tear	Noun

document only refers to past events or implies the presence of warning-related content. This is expected to be done with a different vocabulary than a direct depiction of the warning-related content. Furthermore, the qualifiers *extreme* and *graphic* indicate the use of a specific, more intense subset of the vocabulary. Similarly, the qualifiers *light/mild*, *hedging*, and *non-graphic* suggest the use of a less intense subset of the vocabulary than the base category. Hence, the ranking of the vocabulary terms is expected to be uncorrelated for these qualifiers as well.

5.2.3 Effect of Prescriptive Annotation Guidelines

The effect of using prescriptive rather than descriptive annotation guidelines was evaluated on each of the three *Abuse* categories separately. Following the same structure as Chapter 4, their effect on pairwise annotator agreement was tested as described below.

Passage Selection For each of the three term groups A, B, and C, the top $k = 20$ terms with the highest z-score or log ratio were selected based on the results of the consistency experiments (see Section 4.1). As examples, Table 5.3 shows the three terms with the highest measure in their respective term group (z-score for groups A and C, and log ratio for group B). Interestingly, the three non-vocabulary terms with the highest z-score (group C) are identical for *Physical Abuse* and *Sexual Abuse*, albeit in a different order. The full lists of terms for each category are available in the appendix in Tables B.3, B.4, and B.5. A chapter was retrieved from the document collection D if it was tagged for one of the category tags in X_{c_i} and contained a term from one of the three groups. Subsequently, all sentences with a retrieval term were extracted

Table 5.4: Annotation Prompts and Lists of Examples

(a) Annotation Prompts. Highlighting indicates the sociodemographic profile, the instruction to imagine relevant experiences, and placeholders for the category and examples.

Type	**Instruction**	**Question**
Underspecified	Imagine you are a person of [*Profile*] that has experienced 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 abuse ('yes' or 'no'):
Category-Specific	Imagine you are a person of [*Profile*] that has experienced [*Category*] 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 [*Category*] ('yes' or 'no'):
Prescriptive	Imagine you are a person of [Profile]. Does the following text contain acts/consequences of [*Category*]? Examples are [*Examples*]	Contains acts/ consequences of [*Category*] ('yes' or 'no'):

(b) Examples for the Prescriptive Prompt of each Category

Category	**Examples**
Emotional Abuse	1) intimidation, harassment, humiliation, and other verbal abuse 2) gaslighting, lying and other forms of manipulation 3) socially isolating a person or preventing them from engaging in meaningful activities
Physical Abuse	1) hitting, kicking, burning and other physical violence 2) withholding food or medication 3) unlawful restraint or confinement of a person
Sexual Abuse	1) rape, attempted rape or sexual assault 2) inappropriate touching, looking, or sexual teasing 3) forced use of pornography or indecent exposure ('flashing')

from the chapters. After constructing all passages by adding the context of three neighboring sentences, a total of $s = 10,000$ passages with between 500 and $1,000$ words was sampled uniformly for the 60 terms used in retrieval. The length restriction was applied as a middle ground between (a) long passages as Beck et al. [2024] found a positive effect between sociodemographic prompting and sample length and (b) sufficiently short passages to avoid "confusing" `Flan-T5 11B` with samples that are too long.

Table 5.5: Sociodemographic Profiles

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

Prompting The annotation prompts for the passages were constructed as follows: The underspecified prompt asked to return "yes" for a sample if it requires a warning for *Abuse* and "no" otherwise. The category-specific prompt was similar to the first but asked for a warning for the specific categories *Emotional*, *Physical* or *Sexual Abuse* depending on the category corpus D_{c_i} the passage was selected from. The prescriptive prompt, finally, asked for these same categories but phrased the classification question as "Does this text contain acts/consequences of [*warning category*]". It also provided a list of examples for the respective category, derived from the same information pages as mentioned in the vocabulary collection part of Section 5.2.1. The prompts are illustrated in Table 5.4a and the lists of examples for the prescriptive prompts are shown in Table 5.4b. For the sociodemographic prompting, ten profiles were generated with emphasis on variance in the attributes race and political affiliation to produce greater annotation diversity (see Section 4.2). Table 5.5 lists all profiles with their respective attributes.

Hypotheses After generating a total of $s \cdot 3 \cdot |P| = 300,000$ annotations for each category, the $\frac{|P|(|P|-1)}{2} = 45$ pairwise annotator agreements were calculated for each of the 3 annotation prompts. This created three distributions K_U, K_C, and K_P of Cohen's κ for the underspecified, category-specific, and prescriptive prompt. The expected effect was for the mean pairwise annotator agreements to be higher for the prescriptive prompt than for both the underspecified and the category-specific prompt:

$$\forall c_i \in C : H_A : \overline{\kappa}_P > \overline{\kappa}_U$$

$$\forall c_i \in C : H_A : \overline{\kappa}_P > \overline{\kappa}_C$$

Chapter 6

Experimental Results

This chapter summarizes the results of the experiments described in Chapter 5. It is divided according to the two main methodologies of this thesis with Section 6.1 presenting the results of testing the consistency between tags and vocabulary and Section 6.2 outlining the effects of using prescriptive guidelines in the task of annotating text passages for trigger warnings. In detail, Sections 6.1.1 and 6.1.2 present the results of the significance tests on document-level term frequency and mean log ratio for the entire vocabulary. Section 6.1.3 combines the results from both previous sections to jointly analyze the z-scores from the Mann-Whitney U-test and log ratios on term level. Section 6.1.4 concludes the results of the consistency tests by presenting how the log ratio is affected by qualifiers. The effect of prescriptive annotations guidelines is first explored in Section 6.2.1 by presenting the results of the hypotheses tests on the distributions of annotator agreements. Section 6.2.2 then presents results on the annotation consistency of sociodemographic prompting with respect to which profiles had the highest agreements with one another. The chapter is concluded by Section 6.2.3 that provides more detailed insights into the annotation decisions by looking at the distributions of positive annotations over all passages, exploring the effect of the different term groups (see Section 4.1), and illustrating the annotation behavior with example passages.

6.1 Consistency between Tags and Vocabulary

The first set of experiments tested the consistency of tags applied by authors on AO3 that indicate some form of warning for *Abuse* with the vocabulary expected for the categories *Emotional Abuse*, *Physical Abuse*, and *Sexual Abuse*. As described in Section 5.2.1, the size of the vocabulary was 400 terms for both *Emotional* and *Physical Abuse*, and 250 terms for *Sexual Abuse*. Not all of these terms occurred in the chapters on AO3. Instead, 343 of those in the

Table 6.1: Terms with the Highest and Lowest z-Scores

(a) Highest z-Scores

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) Lowest z-Scores

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

Emotional Abuse vocabulary were found, 346 of those in the *Physical Abuse* vocabulary, and 232 of those in the *Sexual Abuse* vocabulary.

6.1.1 Significant Differences in Term Frequencies

After obtaining z-scores and corresponding p-values for each individual Mann-Whitney U-test, the p-values were multiplied with the number of terms found in each category corpus D_{c_i} to account for multiplicity with a Bonferroni correction. At a significance level $\alpha = 0.05$, the number of vocabulary terms that occurred significantly more often in chapters tagged for their category, was 190 for *Emotional Abuse*, 130 for *Physical Abuse*, and 124 for *Sexual Abuse*. The null hypothesis was rejected for theses terms. An overview on the ten terms with the highest z-scores is given in Table 6.1a. As shown in the table, there was some cross-category overlap between the highly significant terms. The noun *fear* and the verb *force* occurred in the top ten of both *Emotional* and *Sexual Abuse*. *Emotional* and *Physical Abuse* had the noun *anxiety* in common and *Physical* and *Sexual Abuse* shared the noun *bruise* and the adjective *scared*.

The null hypothesis was also rejected in a few cases for terms that occurred significantly *less* frequent for their category corpora. This applied to 1 term for *Emotional*, 20 terms for *Physical*, and 7 terms for *Sexual Abuse*. For illustration purposes, Table 6.1b shows the ten terms with the lowest z-scores for each category. The remaining terms had document-level term frequencies that were not non-significantly different between the two corpora.

As already indicated by the two tables, both *Physical* and *Sexual Abuse* had longer tails of terms with very high or low z-scores. The two outliers in the right tail are *bruise* for *Physical Abuse* and *rape* for the *Sexual Abuse*. The z-scores in the left tail also take on lower values than those for *Emotional Abuse* but not with distinguishable outliers. The z-scores for the *Emotional Abuse* terms, in contrast to that, are less dispersed on both sides of the distribution. This is also illustrated in the distribution plots in Figure 6.1a. *Emotional Abuse* has the highest mean z-score of all categories. This can largely be attributed to the left tail of the distribution being a lot narrower than for the other two categories. The distribution for *Physical Abuse* is a bit wider than the one for *Emotional Abuse* and its mean z-score is closer to 0. Finally, the distribution of z-scores for *Sexual Abuse* is the widest, with the largest share of terms with $z \geq 10$ of all three categories.

6.1.2 Distribution of Log Ratios

The second measure calculated for each term t was the log ratio $\mathrm{lr}(t)$ of corpus-level term frequencies. Instead of testing each term individually, the overall distribution LR_{c_i} for the category vocabulary T_{c_i} was used to test if the mean log ratio $\overline{\mathrm{lr}}$ was significantly different from 0. A summary of the test results can be found in Table 6.2. As the p-value for the one sample t-test was far below the significance level $\alpha = 0.05$ for all three categories, the null hypothesis was rejected in all cases. While all mean differences were statistically significant, the effect size for *Physical Abuse* was noticeably smaller at roughly 50 % of that found for *Emotional* and *Sexual Abuse*. Directly related to that, the mean log ratio $\overline{\mathrm{lr}}$ was also noticeably lower for *Physical Abuse*.

The distributions LR_{c_i} of log ratios are a lot more similar between categories than those for the z-scores from Section 6.1.1. As shown in Figure 6.1b, all three means are slightly above 0, with the visible part of the left tail ending

Table 6.2: Results for the One Sample t-Test on Mean Log Ratio. The columns show the mean log ratio and standard deviation, p-value, and Cohen's d as effect size.

Category	$\overline{\mathbf{lr}}$	$\sigma_{\mathbf{lr}}$	p	d	$\lvert T_{c_i} \rvert$
Emotional Abuse	0.2636	0.6101	1.93e−14	0.3182	343
Physical Abuse	0.1282	0.5071	3.70e−6	0.1617	346
Sexual Abuse	0.2853	0.6140	1.74e−11	0.3439	232

Table 6.3: Terms with the Highest and Lowest Log Ratio

(a) Highest Log Ratios

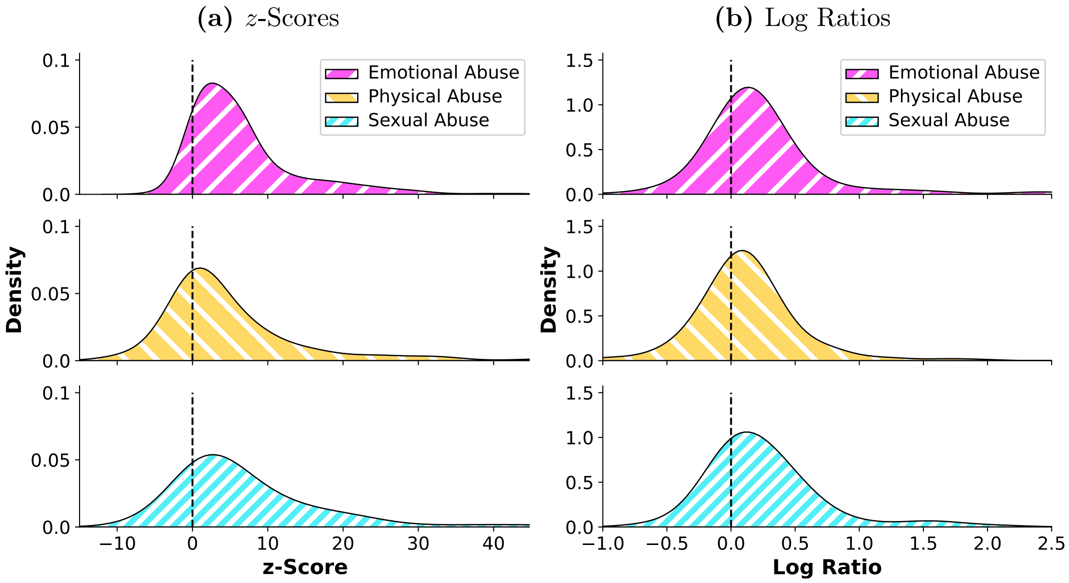
Emotional Abuse		**Physical Abuse**		**Sexual Abuse**	
Term (POS)	lr	Term (POS)	lr	Term (POS)	lr
gaslighting (N)	3.87	thump (V)	4.07	hiv (N)	4.65
ridiculing (A)	3.33	overmedication (N)	3.10	violating (A)	4.65
blaming (A)	3.33	scalding (A)	2.01	depredate (V)	2.23
gaslight (V)	3.15	gnawed (A)	1.84	molestation (N)	1.96
manipulated (A)	3.02	hitting (A)	1.70	gonorrhea (N)	1.94
blamed (A)	2.78	overmedicate (V)	1.65	hypersexuality (N)	1.67
shaming (A)	2.48	slapped (A)	1.43	pedophile (N)	1.65
sniveling (A)	2.43	restraining (A)	1.26	rape (V)	1.62
prohibited (A)	2.43	gashed (A)	1.24	syphilis (N)	1.62
gaslighting (A)	2.33	pushed (A)	1.17	molest (V)	1.57

(b) Lowest Log Ratios

Emotional Abuse		**Physical Abuse**		**Sexual Abuse**	
Term (POS)	lr	Term (POS)	lr	Term (POS)	lr
infringement (N)	-1.78	disjoin (V)	-2.48	subjugated (A)	-1.50
tyrannize (V)	-1.09	tied (A)	-1.82	exhibitionism (N)	-0.74
silenced (A)	-0.84	overfed (A)	-1.23	molested (A)	-0.67
deceiving (A)	-0.70	infecting (A)	-1.00	pussy (N)	-0.53
harasser (N)	-0.66	burning (A)	-0.94	exploited (A)	-0.52
disoriented (A)	-0.56	trapping (A)	-0.93	groped (A)	-0.43
abase (V)	-0.54	cowed (A)	-0.91	stalker (N)	-0.34
whimpering (A)	-0.54	spank (N)	-0.76	obscene (A)	-0.33
fuming (A)	-0.53	thrashed (A)	-0.70	grope (N)	-0.32
coerced (A)	-0.52	biff (N)	-0.60	indecency (N)	-0.30

around -1. Yet, the right tail of the distributions extend further for *Emotional* and *Sexual Abuse*, yielding further illustration to the differences in mean log ratio between the categories. Another overview is given by the ten terms with

Figure 6.1: Probability Density Functions for Vocabulary Terms



(a) z-Scores **(b)** Log Ratios

the highest log ratios (Table 6.3a). First, we see that *Physical Abuse* has two outlier terms with very high log ratios, the verb *thump* and the noun *overmedication*. *Sexual Abuse* also has two outlier terms, the noun *hiv* and the adjective *violating*, with even higher log ratios. *Emotional Abuse*, in contrast to the other two, has five terms that occurred more than 8-times as frequent in the category corpus than in the baseline corpus. A noticeable pattern is the prominence of the topic *Gaslighting* that had three related terms among the top 10.

As for the ten terms with the lowest log ratios (Table 6.3b), the values for *Physical Abuse* were also consistently lower than for the other two categories. An interesting observation is that, for *Sexual Abuse*, the adjective *molested* was among the terms with the lowest log ratio while the corresponding noun and verb ended up at the opposite side of the distribution.

6.1.3 Joint Analysis of z-Scores and Log Ratios

This section expands upon the hypotheses tests of the two previous sections by looking at the z-score and log ratio of vocabulary terms simultaneously. The probability density functions (PDFs), as estimated by a kernel density estimation, in Figure 6.1 show that the values for both distributions had a broader right than left tail and were shifted (slightly) to the right. For a detailed discussion of the category-specific distributions, refer back to Section 6.1.1 for the z-score and and Section 6.1.2 for the log ratio.

Table 6.4: Correlation between z-Score and Log Ratio. The columns show Pearson's correlation coefficient r, the t-statistic and corresponding p-value.

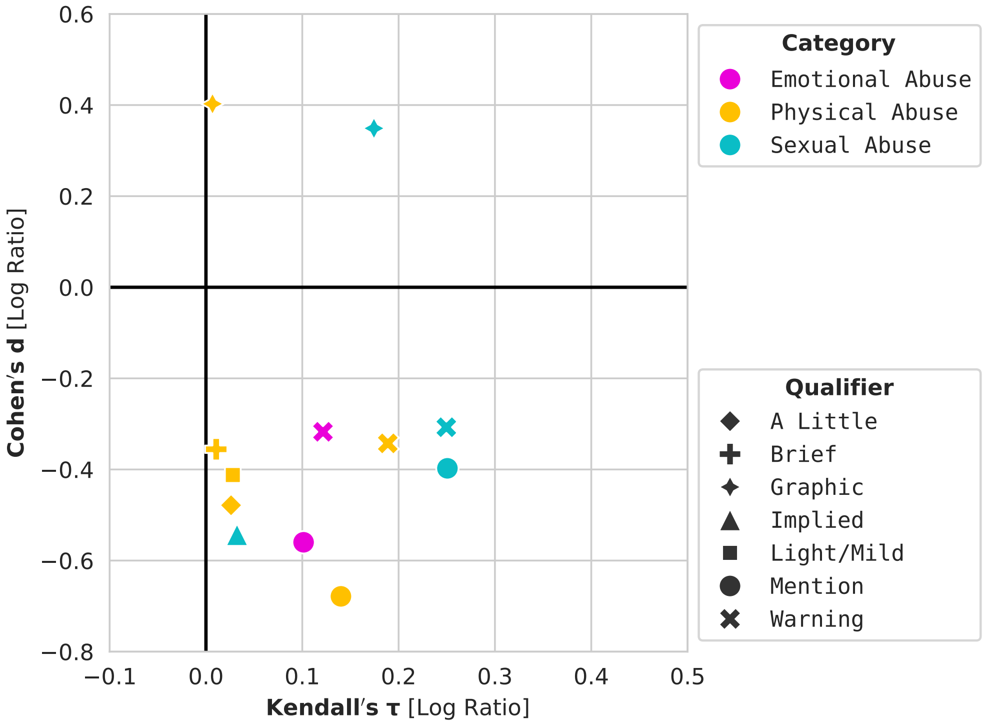
Category	$\boldsymbol{r}$	$\boldsymbol{t}$	$\boldsymbol{p}$	$\lvert \boldsymbol{T_{c_i}} \rvert$
Emotional Abuse	0.17	3.25	1.3e−3	343
Physical Abuse	0.31	6.02	4.5e−9	346
Sexual Abuse	0.32	5.14	5.9e−7	232

Despite the similarly shaped PDFs, z-score and log ratio exhibited only a low to moderate positive correlation on term level. Table 6.4 gives an overview on the Pearson correlation coefficient r and the associated p-value for each category. While both *Physical* and *Sexual Abuse* had a similar correlation of around 0.3, the values for z-score and log ratio for the *Emotional Abuse* vocabulary only had a low correlation of 0.17. Hence, fewer terms from the *Emotional Abuse* vocabulary that occurred significantly more often also had a high log ratio (and vice-versa).

This is visually illustrated in the scatter plots of Figures A.1, A.2, and A.3 in the appendix. For *Emotional Abuse*, only the noun *gaslighting* and the verb *gaslight* had both a high z-score and log ratio. Otherwise, the right tails of both distributions were largely independent, leading to a convex curve in the point cloud. For *Sexual Abuse*, there were a lot more terms with high values on both measures and thus more points in the upper right quadrant of the graph. The point cloud for *Physical Abuse* is more similar to *Emotional Abuse*, with two largely independent tails. Yet, there were more terms with a high z-score that also had a comparably high log ratio, leading to the higher correlation of the two measures for this category.

As an addition to the scatter plots, Table A.1 lists the vocabulary terms for each category that occurred among both the 50 terms with the highest z-score and the highest log ratio. For *Sexual Abuse*, this was the case for 18 terms, often with related word stems or semantic meaning. Terms related to *rape*, *sex*, *molesting*, *violation*, and *genitals* occurred twice each. Both *consent* and *suicide/depression* were present three times each. *Physical Abuse* had ten terms that occurred in both lists. Among them, signs of physical harm were the most common with terms related to *bruising/swelling* occurring five times. For *Emotional Abuse*, only four terms had both a high z-score as well as high log ratio, further illustrating the lower correlation. A concept with especially high values for both was *gaslighting* as already shown in Figure A.1 and Table 6.3a.

Figure 6.2: Qualifier Effects per Category. The origin is defined by the base category without any qualifier restriction. The x-Axis shows the rank correlation of log ratios with the base category. A value close to 1 means that the vocabulary terms, sorted by their log ratios, are in a similar order. A value close to -1 indicates an inverted order. The y-Axis shows the effect size for the difference in mean log ratio $\overline{\text{lr}}$ from base category to qualifier.



6.1.4 Effect of Qualifiers

An additional level of granularity was added to the categorization of tags by separating them according to qualifiers. Figure 6.2 shows the effect of qualifiers q with at least 50 chapters in the subcategory corpus $D_{c_{i,q}}$ and a mean log ratio $\overline{\text{lr}}$ that was significantly different from that of the *base category* c_i without qualifier restriction. Statistical significance was determined after applying a Bonferroni correction for the 13 qualifiers that were tested for each category. As an addition to the figure, detailed summaries of the effect sizes for all qualifiers and their rank correlations are given in Tables 6.5a and 6.5b. Both the figure and the table already show that very few qualifiers had consistent effects across all categories. Specifically, the qualifiers with consistent effects were *mention*, *graphic*, and *warning*. With the exception of *Graphic Emotional Abuse*, these were also qualifiers with a lot of tagged chapters n. In the following, the results for each qualifier will be discussed in alphabetical order.

Table 6.5: Qualifier Effect Sizes and Rank Correlation per Category. The highlighting indicates significantly positive, negative or non-significant results.

(a) Mean Differences in Log Ratio. The columns show the effect size Cohen's d, the corrected p-value p^* and number of chapters n. The row-wise absolute maximum is marked bold.

	Emotional Abuse			**Physical Abuse**			**Sexual Abuse**		
Qualifier	d	p^*	n	d	p^*	n	d	p^*	n
A Little	0.08	1.0	117	**-0.48**	7.7e−8	330	0.28	2.8e−1	34
A Lot	**0.22**	1.8e−1	37	0.02	1.0	85	—	—	—
Brief	**0.95**	6.0e−23	8	-0.36	1.3e−4	349	-0.26	1.1e−1	110
Description	**1.16**	7.8e−19	4	0.03	1.0	286	-0.10	1.0	192
Discussion	**0.47**	5.0e−6	27	-0.01	1.0	79	0.09	1.0	288
Extreme	0.16	1.0	12	**0.33**	5.9e−4	32	0.19	8.1e−1	39
Graphic	**1.07**	2.1e−20	3	0.40	1.8e−5	67	0.35	3.9e−3	257
Hedging	0.15	8.6e−1	102	0.38	1.1e−4	25	**-0.67**	8.6e−8	38
Implied	-0.06	1.0	179	-0.20	1.9e−1	461	**-0.54**	6.1e−7	645
Light/Mild	0.14	1.0	66	**-0.41**	6.1e−6	595	-0.07	1.0	167
Mention	-0.56	6.3e−11	1,218	**-0.68**	7.0e−16	3,570	-0.40	4.7e−4	5,246
Non-Graphic	**0.43**	1.7e−5	23	0.17	5.0e−1	91	-0.12	1.0	212
Warning	-0.32	1.5e−3	312	**-0.34**	2.8e−4	734	-0.31	2.1e−2	728

(b) Rank Correlation on Log Ratio. The columns show the rank correlation τ, corrected p-value p^*, and number of shared terms k. The row-wise absolute maximum is marked bold.

	Emotional Abuse			**Physical Abuse**			**Sexual Abuse**		
Qualifier	τ	p^*	k	τ	p^*	k	τ	p^*	k
A Little	0.23	1.5e−5	204	0.03	1.0	201	**0.26**	2.9e−1	39
A Lot	**0.21**	2.2e−2	104	0.12	7.5e−1	120	—	—	—
Brief	0.00	1.0	75	0.01	1.0	212	**0.17**	4.7e−2	137
Description	0.01	1.0	22	**0.18**	2.6e−3	193	0.13	2.6e−1	153
Discussion	0.15	2.1e−1	112	0.05	1.0	179	**0.24**	7.2e−5	165
Extreme	0.07	1.0	61	**0.16**	5.7e−2	143	0.02	1.0	95
Graphic	**-0.25**	4.5e−1	35	0.01	1.0	144	0.17	8.2e−3	174
Hedging	**0.25**	5.8e−6	189	0.05	1.0	113	0.17	3.4e−1	82
Implied	0.14	1.7e−2	226	**0.24**	7.9e−6	199	0.03	1.0	171
Light/Mild	**0.23**	7.6e−5	176	0.03	1.0	231	0.14	1.3e−1	145
Mention	0.10	1.9e−1	263	0.14	6.3e−3	279	**0.25**	9.8e−7	208
Non-Graphic	0.20	4.3e−2	97	0.18	2.0e−2	144	**0.21**	1.1e−3	156
Warning	0.12	7.4e−2	234	0.19	2.5e−4	231	**0.25**	9.9e−6	178

A Little This qualifier led to an expected decrease in mean log ratio for *Physical Abuse* on 330 chapters but to no significant changes for the other two categories with fewer chapters. The rank correlation was significantly positive for *Emotional Abuse*, pointing to a similar ranking of the terms in the vocabulary. For the other categories, the rank correlation was not significant.

A Lot The emphasis on *a lot* of a certain warning did not lead to the expected increase in mean log ratio. The qualifier was not used for *Sexual Abuse*. The rank correlation was significantly positive for *Emotional Abuse*.

Brief Chapters tagged for *brief* instances of a warning had an expected lower mean log ratio for *Physical Abuse*. For *Emotional Abuse*, the qualifier led to a much higher log ratio but only on a set of eight chapters. The rank correlation was only significantly positive for *Sexual Abuse*.

Description Similar to the *brief* qualifier, *description* was associated with a significant increase in mean log ratio for *Emotional Abuse* based on a negligible number of chapters. For categories with more chapters, *Physical* and *Sexual Abuse*, no change was observed. The rank correlation was only significantly positive for *Physical Abuse*.

Discussion The *discussion* of a warning led to a significant increase in mean log ratio for *Emotional Abuse* and no noticeable difference otherwise. The rank correlation was significantly positive for *Sexual Abuse* but not for the two other categories.

Extreme Tags that qualify a warning category as *extreme* were rare for all three categories. While the shift in mean log ratio was expectedly positive for all three categories, the difference was only significant for *Physical Abuse*. No significant rank correlation was observed for any of the categories. This suggests that the relative increase in mean log ratio was due to different terms than for the *base category*.

Graphic The *graphic* qualifier was the only one with a significant increase in mean log ratio for all three categories. For *Physical* and *Sexual Abuse*, the similarity in effect size was also illustrated in Figure 6.2. For *Emotional Abuse*, it was only observed in the tags of three chapters. Another observation from the figure was that *Graphic Physical Abuse* appeared to use a different vocabulary than the *base category* as the rank correlation was very close to 0. For *Sexual Abuse*, on the other hand, the rank correlation was significantly positive, suggesting a similar vocabulary.

Hedging Tags that indicate some form of *hedging* on behalf of the author were associated with an expected significant decrease in mean log ratio for *Sexual Abuse* but also with a significant increase for *Physical Abuse* on 25 chapters. The rank correlation was significant for *Emotional Abuse* but not for the other two categories.

Implied If a warning category was only *implied*, the tagged chapters had a lower mean log ratio for all categories. The only significant decrease, however, was observed for *Sexual Abuse* with a total of 645 chapters. The rankings of vocabulary terms were correlated significantly positive for both *Emotional* and *Physical Abuse*. For *Sexual Abuse*, no correlation was observed.

Light/Mild The restriction to tags that indicate a reduced severity with *light/mild* qualifiers led to a significant decrease in mean log ratio for *Physical Abuse*. The effects for *Emotional* and *Sexual Abuse* were non-significant. The rank correlation was significantly positive for *Emotional Abuse*.

Mention For chapters that only *mention* a warning category or refer to past events, the mean log ratio was significantly reduced in comparison with the *base category*. Across all categories, this qualifier was also the most frequently observed with 1,218 chapters for *Emotional*, 3,570 for *Physical*, and 5,246 for *Sexual Abuse*. The strongest decrease was observed for *Physical Abuse*. The rank correlations were not as similar, but all positive and significantly so for *Physical* and *Emotional Abuse*.

Non-Graphic Contrary to its opposite qualifier *graphic*, the *non-graphic*-qualifier did not lead to consistent effects. Unexpectedly, the qualifier even led to a significant increase in mean log ratio for *Emotional Abuse*. The rank correlations were significantly positive for all three categories, suggesting a similar ranking of terms than the *base category*.

Warning The *warning* qualifier was the most consistent across categories when looking at mean log ratio and rank correlation simultaneously (as illustrated by Figure 6.2). Contrary to the hypotheses outlined in Table 5.2, however, the mean log ratios were significantly lower rather than higher. The rank correlations were positive for all categories but only significant for *Physical* and *Sexual Abuse*.

6.2 Effect of Prescriptive Annotation Guidelines

After testing the category consistency both on term- as well as on vocabulary-level, the results were used to sample passages for the terms with the highest z-scores or log ratios as described in Section 5.2.3. Based on the annotations made by `Flan-T5 11B` for the ten different sociodemographic profiles, pairwise annotator agreements for each of the three annotation prompts were calculated.

6.2.1 Pairwise Annotator Agreements

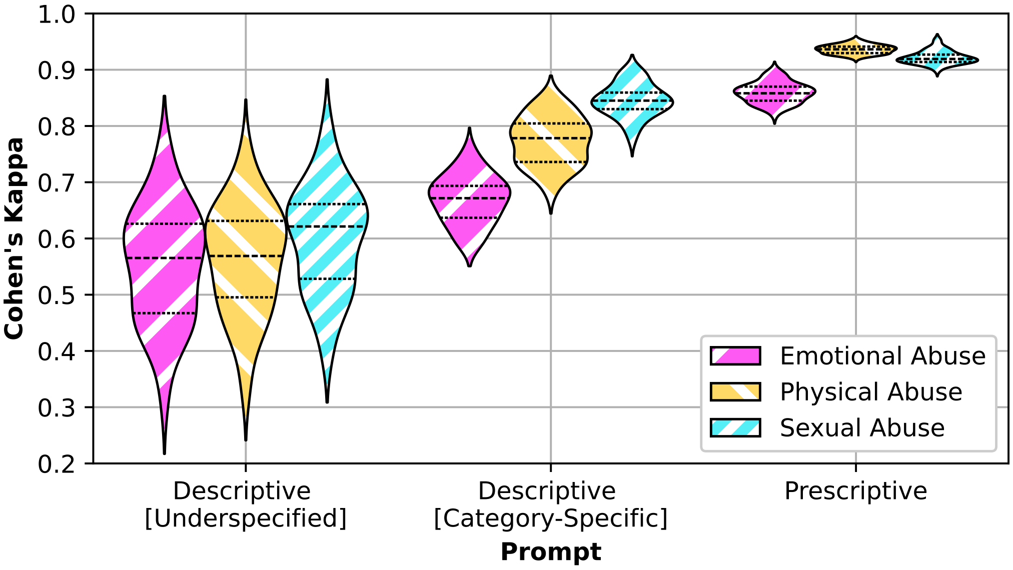
The results of the two sample t-test on the three distributions of Cohen's κ-values can be found in Table 6.3b. The distributions are labeled K_U for the underspecified prompt, K_C for the category-specific prompt, and K_P for the prescriptive prompt. As evident from the table, the differences in mean were significant for all three pairs of distributions across all three categories. In other words, the prescriptive prompt was found to lead to a significantly higher mean pairwise annotator agreement than both other prompts. In addition to that, the category-specific prompt significantly increased the mean agreement in comparison with the underspecified prompt.

For each distribution pair, a different category had the highest effect size. The increase in agreement from the underspecified to the prescriptive prompt was the strongest for *Physical Abuse*. In the comparison of category-specific with prescriptive prompt, the highest effect size was observed for *Emotional Abuse*, suggesting that `Flan-T5 11B` benefited the most from a clarification through examples for this category. Finally, the most improvement from the underspecified to the category-specific prompt was found for *Sexual Abuse*. This indicates that `Flan-T5 11B` predicted less disagreement between simulated annotators for that category even without a list of examples. Overall, *Emotional Abuse* had the lowest agreement scores and the highest standard deviation for all three prompt types. As a side note, the two distribution parameters $\overline{\kappa}$ and σ_κ were very similar between *Emotional Abuse* using the prescriptive prompt and *Sexual Abuse* using the category-specific prompt.

The test results are visually illustrated by Figure 6.3a. The figure shows the clear improvement in annotator agreement as the prompt was made more specific by (1) adding the category instead of asking for the high level warning *Abuse*, and (2) rephrasing the prompt to be prescriptive instead of descriptive. The plot also underlines that *Emotional Abuse* had the most disagreements across all three prompt types.

Figure 6.3: Pairwise Annotator Agreements

(a) Distributions by Prompt and Category. The middle line in each violin plot indicates the mean $\overline{\kappa}$ and the outer lines show the 25th and 75th percentile.



(b) Results for the t-Test. Columns A and B indicate which two distributions are compared. The bold row for each distribution pair shows the lowest p-value and highest effect size as measured by Cohen's d.

A	B	**Category**	$\overline{\kappa}_A(\sigma_{\kappa_A})$	$\overline{\kappa}_B(\sigma_{\kappa_B})$	p	d
K_P	K_U	Emotional Abuse	0.86 (0.02)	0.55 (0.10)	5.26e−34	4.15
		Physical Abuse	0.94 (0.01)	0.56 (0.10)	**6.12e−43**	**5.46**
		Sexual Abuse	0.92 (0.01)	0.60 (0.09)	1.07e−38	4.81
K_P	K_C	Emotional Abuse	0.86 (0.02)	0.67 (0.04)	**6.15e−45**	**5.79**
		Physical Abuse	0.94 (0.01)	0.77 (0.04)	1.10e−41	5.26
		Sexual Abuse	0.92 (0.01)	0.85 (0.03)	2.43e−28	3.44
K_C	K_U	Emotional Abuse	0.67 (0.04)	0.55 (0.10)	5.44e−10	1.47
		Physical Abuse	0.77 (0.04)	0.56 (0.10)	6.63e−23	2.82
		Sexual Abuse	0.85 (0.03)	0.60 (0.09)	**4.17e−29**	**3.53**

6.2.2 Consistency of Sociodemographic Prompting

Beck et al. [2024] found sociodemographic prompting to not be able to reproduce the annotations made by persons with the same profile. In the context of this thesis, this raised the related question if `Flan-T5 11B` was consistent in its predictions about which profiles agreed with each other across prompt types and categories. In order to test this, we looked at each of the ten profiles

Table 6.6: Mean Rank Correlations of Agreement between Annotators

(a) Rank Correlations for the same Category. The abbreviations in brackets refer to the annotation prompts prescriptive, underspecified, and category-specific.

Category	$\overline{\tau}(\sigma_\tau)$ [Pre., Und.]	$\overline{\tau}(\sigma_\tau)$ [Pre., Cat.]	$\overline{\tau}(\sigma_\tau)$ [Cat., Und.]
Emotional Abuse	0.24 (0.23)	0.38 (0.17)	**0.54 (0.15)**
Physical Abuse	0.41 (0.25)	0.25 (0.37)	**0.58 (0.21)**
Sexual Abuse	0.46 (0.27)	0.34 (0.22)	**0.47 (0.22)**

(b) Rank Correlations for the same Annotation Prompt. The acronyms in brackets refer to the categories with Em. for Emotional Abuse, Ph. for Physical Abuse, and Se. for Sexual Abuse.

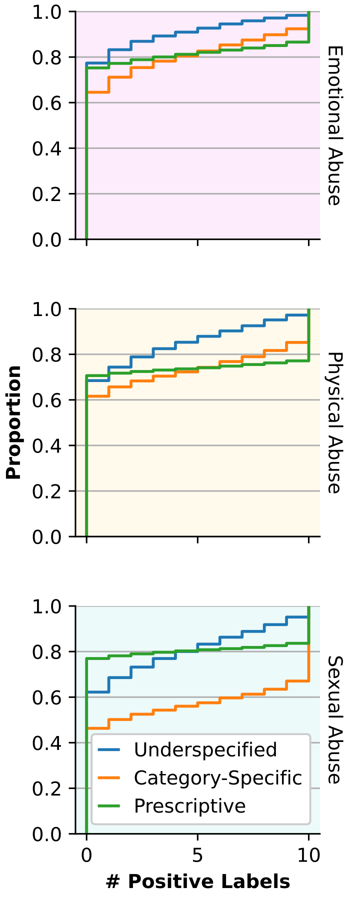
Annotation Prompt	$\overline{\tau}(\sigma_\tau)$ [Em., Ph.]	$\overline{\tau}(\sigma_\tau)$ [Em., Se.]	$\overline{\tau}(\sigma_\tau)$ [Ph., Se.]
Underspecified	0.89 (0.06)	**0.92 (0.08)**	0.89 (0.08)
Category-Specific	0.68 (0.07)	**0.81 (0.17)**	0.70 (0.17)
Prescriptive	0.53 (0.16)	0.46 (0.24)	**0.57 (0.16)**

individually and ranked the nine other annotators by their agreement scores. These rankings were then used to calculate the rank correlation as measured by Kendall's τ. The annotator rankings were compared with respect to both (1) different annotation prompts in the same category as well as (2) different categories with the same annotation prompt. Tables 6.6a and 6.6b present the mean rank correlation $\overline{\tau}$ for both settings alongside the corresponding standard deviation σ_τ. These values were calculated based on the individual rank correlations for each of the ten sociodemographic profiles.

The rank correlations across annotation prompts in the same category were positive but not large. For all three categories, the annotator rankings were the most consistent between the underspecified and the category-specific prompt. The rank correlations for the same annotation prompt across different categories, however, were noticeably higher for both the underspecified as well as the category-specific prompt. In case of the underspecified prompt and the categories *Emotional* and *Sexual Abuse*, $\overline{\tau}$ indicated almost identical rankings. These high rank correlations for the underspecified prompt in combination with the wide range of agreement scores κ for that prompt suggest that the same simulated profiles (dis-)agreed consistently across passages, regardless of the category corpus the passages were sampled from. A detailed overview on the profiles that agreed the most with each other on a given combination of annotation prompt and category can be found in Table B.2 in the appendix.

Figure 6.4: Distribution of Positive Annotations across Passages. Each column of the tables shows the absolute number of passages for a different prompt and the rows indicate the number of positive annotations. The graphs show the same data as empirical cumulative distribution functions (CDFs).

#	Underspec.	Category-Sp.	Prescriptive
0	7,737	6,454	7,524
1	585	662	198
2	368	422	162
3	235	283	120
4	169	228	116
5	176	216	88
6	183	267	98
7	137	216	91
8	124	231	108
9	117	260	152
10	169	761	1,343
0	6,846	6,160	7,065
1	592	410	108
2	448	262	73
3	361	211	64
4	282	185	52
5	260	200	47
6	238	254	74
7	226	214	68
8	257	272	72
9	212	353	90
10	278	1,479	2,287
0	6,221	4,631	7,695
1	636	386	118
2	462	234	87
3	375	179	69
4	310	172	60
5	323	151	52
6	304	213	48
7	254	163	54
8	298	219	78
9	331	358	102
10	486	3,294	1,637



6.2.3 Analysis of Annotations

The prior analyses revealed that prescriptive annotation guidelines increased the average pairwise agreement between simulated annotators. This section expands upon the hypotheses testing by studying the annotations in more detail.

Distribution of Positive Annotations The first analysis of this section is focused on how the number of positive annotations was distributed across all $s = 10,000$ passages per category. An overview on how many passages received no to only positive annotations by the ten annotators is given in the table and graphs of Figure 6.4. For each combination of warning category and annotation prompt, the table shows the absolute number of passages that received a given number of positive annotations. The graphs expand on that by showing the empirical cumulative distribution functions (CDFs) for the same data.

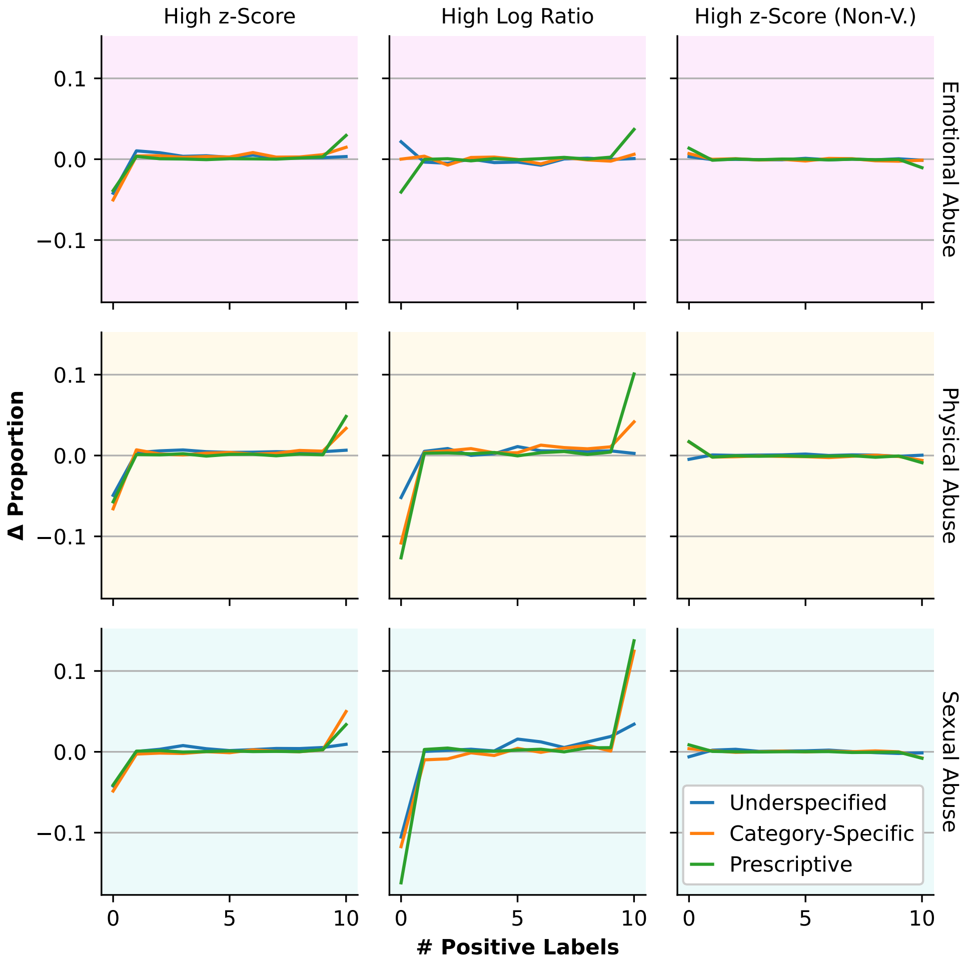
Similar to the results reported by Wiegmann et al. [2024], a much larger share of passages received unanimously negative than unanimously positive annotations. For eight of the nine category-prompt-combinations, the share of passages with no positive annotations was between 62 and 77 %. The only outlier in that regard was the category-specific prompt for *Sexual Abuse*. It led to both the fewest passages with zero (46 %) and the most passages with ten positive annotations (33 %). This is best illustrated by the corresponding empirical CDF being far below the distributions for the other two prompts.

The category-specific prompt for *Sexual Abuse* also had far fewer unanimously negative annotations than the underspecified prompt. In other words, asking for the category *Sexual Abuse* instead of only for *Abuse* resulted in a lot more passages receiving at least one positive annotation. Further specifying *Sexual Abuse* with a list of examples, however, had the opposite effect. The prescriptive prompt led to a lot more passages getting no positive annotations than the underspecified prompt. This suggests that some passages did not fall under the examples given by the prescriptive prompt, an observation that is further discussed in the paragraph on example passages in Section 6.2.3. For both *Emotional* and *Physical Abuse*, the number of unanimously negative annotations was more similar across all three prompts.

Despite these differences, all categories had in common that the category-specific prompt resulted in the fewest unanimously *negative* annotations. Another shared effect was that the underspecified prompt consistently had the lowest number of unanimously *positive* annotations. For *Emotional Abuse*, this was only the case for 169 passages. Furthermore, the prescriptive prompt led to a noticeable increase in the number of unanimously positive annotations for all three categories.

The increase in pairwise agreements when using the prescriptive, instead of the underspecified or category-specific prompt, is also visible in the empirical CDF graphs. The distribution for the prescriptive prompt is very flat for all three categories, underlining that a lot of passages received either unanimously positive or unanimously negative annotations. In detail, the share of passages with 1-9 positive annotations for the prescriptive prompt was 11.3 % for *Emotional Abuse*, 6.5 % for *Physical Abuse*, and 6.7 % for *Sexual Abuse*.

Figure 6.5: Relative Shift in the Empirical CDFs per Term Group. The columns distinguish between the groups of terms used for passage sampling: Vocabulary terms with the highest z-score, vocabulary terms with the highest log ratio, and non-vocabulary terms with the highest z-score.



Effect of Term Groups The passages for the annotation experiments were sampled using three different groups of terms (see Section 4.1): Group A contained the vocabulary terms with the highest z-score. The terms in group B were those with the highest log ratio $\mathrm{lr}(t)$ among all vocabulary terms that also occurred significantly more frequent on document level. Group C, finally, contained the terms with the highest z-score that were not part of of the vocabulary and occurred in at least 5,000 documents.

Figure 6.5 illustrates how a restriction to one of these groups affected the number of positive annotations per passage. Specifically, it depicts by how many percent points the value of the empirical CDF shown in Figure 6.4 increased or decreased when looking only at passages that contained a term from one of the three groups. To illustrate this with one example: The prescriptive prompt resulted in 77 % of all *Sexual Abuse* passages to be annotated unanimously negative. When looking only at the passages that contained a term from the group with the highest log ratios, this share dropped to 61 %. Hence, the middle graph in the last row shows a decrease of 16 percent points in the proportion of passages with zero positive labels. The absolute number of positive annotations for the term groups can be found in the appendix in Table B.1.

The most consistent effect was observed for the vocabulary terms with the highest z-scores. For this group, the number of passages with zero positive annotations was reduced by about 5 percent points across all nine category-prompt-combinations. In addition that, there was also an increase in the number of passages with ten positive annotations. This effect, however, was only pronounced for the prescriptive prompt on *Physical Abuse* passages and the category-specific prompt on *Sexual Abuse* passages. The share of passages with one to nine positive annotations barely fluctuated at all.

The strongest effect resulted from restricting the passages to those that contained a vocabulary term with a high log ratio. For the prescriptive prompt, the share of passages with zero positive annotations decreased by 4.1 percent points for *Emotional*, 12.7 for *Physical*, and 16.2 for *Sexual Abuse*. Furthermore, the share of passages with ten positive annotations increased by 3.7, 10.1, and 13.7 percent points, respectively. For *Sexual Abuse*, similar effects occurred in combination with the category-specific prompt. The share of passages with no positive annotations decreased by 11.7 and the one for passages with only positive annotations increased by 12.5 percent points.

The shifts for the non-vocabulary terms with the highest z-scores were negligible for all category-prompt-combinations. This can in part be explained by this term group having the most passages associated with it (see Table B.1). Group C contained very common terms as the z-scores were generally for higher terms that occurred in a lot of documents (see Section 4.1). As these terms were more common, they also ended up in more passages because overlapping passages with multiple terms were merged into one. Nonetheless, the share of passages with no or only positive annotations was still noticeably different for the terms from the other two groups. This is also reflected in the absolute numbers in Table B.1.

Overall, the results shown in Figure 6.4 were consistent with the motivation behind the term groups described in Section 4.1. The log ratio group featured terms specific to the category and thus received more positive annotations while the non-vocabulary group contained more common terms and consequently received fewer positive annotations.

Table 6.7: Terms with the Highest Share of Unanimously Positive Annotations for the Prescriptive Prompt. The columns show the z-score, log ratio, and the share of unanimously positive annotations for each prompt. The column s shows the number of passages that the term occurred in.

Category	Term (POS)	Group	z	lr	Und.	Cat.	Pre.	s
Emotional Abuse	gaslight (V)	A	26.38	3.15	0.02	0.28	0.80	54
	gaslighting (N)	A	38.71	3.87	0.13	0.44	0.79	131
	infantilize (V)	B	4.41	1.10	0.00	0.50	0.50	2
	manipulated (A)	B	5.63	3.02	0.06	0.17	0.48	204
	shaming (A)	B	2.02	2.48	0.00	0.11	0.33	70
Physical Abuse	caning (N)	B	3.12	0.79	0.10	0.41	0.83	29
	shackled (A)	B	3.12	0.61	0.05	0.35	0.67	172
	welt (N)	B	18.18	0.71	0.05	0.28	0.58	146
	gashed (A)	B	3.66	1.24	0.05	0.30	0.50	40
	bruised (A)	A	28.85	0.75	0.07	0.32	0.48	297
Sexual Abuse	molestation (N)	B	20.06	1.96	0.22	0.83	0.70	92
	rape (V)	A	105.72	1.62	0.25	0.86	0.66	403
	violated (A)	B	4.35	1.48	0.12	0.74	0.64	217
	molest (V)	A	40.85	1.57	0.12	0.76	0.56	147
	pedophile (N)	A	30.76	1.65	0.14	0.79	0.42	174

Table 6.8: Correlation between Unanimous Annotations and Term Measures.

	Underspecified		**Category-Specific**		**Prescriptive**	
Measure	Unan. P.	Unan. N.	Unan. P.	Unan. N.	Unan. P.	Unan. N.
z-Score	0.24	-0.01	**0.34**	-0.14	0.08	0.16
Log Ratio	0.12	-0.33	0.17	-0.26	0.22	**-0.46**

Indicative Terms An analysis of the annotations on term level revealed that all categories had a handful of terms whose presence in a passage was associated with a lot of positive annotations. In the following, these will be referred to as *indicative* terms of the category. A summary of the five terms with the highest share of unanimously positive annotations for the prescriptive prompt can be found in Table 6.7. For *Emotional Abuse*, about 80 % of all passages that contained *gaslight* or *gaslighting* received unanimously positive annotations. This is in stark contrast to only 2 % and 13 % for the underspecified prompt. In the case of *Physical Abuse*, *caning* and *shackled* were the terms with the highest share of unanimously positive annotations, 83 % and 67 %, respectively. Again, these number stood in contrast to the shares for the underspecified prompt with 10 % and 5 %. For *Sexual Abuse*, the indicative terms received consistently high shares of unanimously positive annotations across prompts.

Figure 6.6: Passages with only Positive Annotations. Sampling terms are highlighted based on their category: Emotional abuse, physical abuse, and sexual abuse. The last line gives the number of positive annotations per prompt.

> [...] He was her ex boyfriend, abuser, and convict. He **manipulated** her into needing him for everything, lied to her so she wouldn't see her friends, and almost pushed her out of her own family. [...]
>
> **Underspecified: 10, Category-Specific: 10, Prescriptive: 10**

> [...] And if I didn't do what he wanted, he would **beat** me or cause me **pain** in other ways too horrible to mention. I couldn't find my clothes and even if I had, I doubt he would have let me wear them. [...]
>
> **Underspecified: 10, Category-Specific: 10, Prescriptive: 10**

> [...] Those photos fell out of her pocket, her clothes and hair were disheveled, and she was crying. I believe Mr. Kamoshida used those photos as blackmail to get Eiko into his office so he could **molest** her ... or worse. [...]
>
> **Underspecified: 10, Category-Specific: 10, Prescriptive: 10**

Both *molestation* and *rape* received only positive annotations for over 20 % of their passages with the underspecified, over 80 % for the category-specific, and over 60 % for the prescriptive prompt.

Some of the indicative terms directly occurred in the list of examples given in the prescriptive prompts (see Table 5.4b). The one for *Emotional Abuse* mentioned *gaslighting* and the one for *Sexual Abuse* mentioned *rape*. For *Physical Abuse*, neither *caning* or *shackled* were mentioned explicitly. The closest indirect mentions were *hitting* and *unlawful restraint*.

A more detailed overview, featuring all terms as well as the share of unanimously negative annotations, is given in Tables B.3, B.4, and B.5 in the appendix. Based on these values, Table 6.8 depicts the correlation of the measures z-score and log ratio with the share of unanimous annotations. For the z-score, the correlations were inconclusive across the three prompts. The log ratio of a term, however, was both (weakly) positively correlated with the share of unanimously positive annotations, and negatively correlated with the share of unanimously negative annotations for all prompts.

Example Passages In order to expand upon the quantitative annotation analyses of the previous paragraphs, Figures 6.6 and 6.7 provide some example passages. The first figure shows passages that were annotated unanimously positive for all three annotation prompts. The two examples for *Physical* and

Figure 6.7: Passages with no Disagreement for the Prescriptive Prompt. Sampling terms are highlighted based on their category: Emotional abuse, physical abuse, and sexual abuse. The last line gives the number of positive annotations per prompt.

> "Then he started to use it to **threaten** me. Control me. Didn't work in the end when I figured it out that it didn't make any difference how good I was and I just started acting out. But then he'd just '**punish**' me more." "**Punish** you? It's not your **fault**. You were just a normal boy – a normal teenager." [. . .]
>
> **Underspecified: 4, Category-Specific: 7, Prescriptive: 10**

> She was right there, so close to leaving him forever that he didn't even hesitate to plunge the syringe into her neck. When her legs buckled and her eyes rolled back, he felt **fear** and adrenaline rush through his blood, suddenly panicked that she'd **overdose** and that he'd need to take her to a **hospital**. [. . .]
>
> **Underspecified: 5, Category-Specific: 8, Prescriptive: 10**

> [. . .] I was very much in **pain**, blood was flowing out of my arms, he refused to help me, he grabbed me by the ear and threw me outside. Telling me I couldn't come back if the fight wasn't over yet. The next day, he pressed me down on the **bed**. He would do things to me without **consent**. [. . .]
>
> **Underspecified: 5, Category-Specific: 2, Prescriptive: 0**

Sexual Abuse both use rather graphic descriptions and refer to physical harm. For *Emotional Abuse*, on the other hand, the only vocabulary term is *manipulated* and the descriptions are not as graphic. However, the passage also features the word *abuser*, a possible explanation for the unanimously positive annotations on the underspecified prompt. As contradictory evidence, we found that passages that contained terms with the stem "abus" received fewer unanimously positive annotations for the underspecified than for the other two prompts. For *Emotional Abuse*, 151 of the annotated passages contained this stem. Out of these, the number of passages with ten positive annotations was 27 for the underspecified, 52 for the category-specific, and 84 for the prescriptive prompt. Among the *Physical Abuse* passages, this stem occurred in 358 passages with 51, 103, and 161 cases of ten positive annotations, respectively. For *Sexual Abuse*, this was the case for 552 passages with ten positive annotations for 98, 297, and 207 of them for the three different prompts.

The passages in Figure 6.7 are examples that received unanimous annotations for the prescriptive prompt but caused disagreements for the two other annotation prompts. For *Emotional Abuse*, the example does not contain any direct mentions of physical harm, but clearly mentions *Emotional Abuse* in

Figure 6.8: Passages that Mention Consent. The highlighting indicates sampling terms related to sexual abuse. The last line gives the number of positive annotations per prompt.

> [. . .] "When asked by the police if he had ever forced you into a **sexual** situation, you said it didn't matter because you never technically said no?" "Yes." I nodded. "Did you ever say yes, though? Did he ever explicitly ask for your **consent**?" She asked, looking at me. I started to see her point. "No."
>
> **Underspecified: 2, Category-Specific: 8, Prescriptive: 10**

> [. . .] Namjoon took pictures of every **bruise** I didn't **want**, every scar, every mark that I didn't **consent** to – he said it was for the time I would finally change my mind and report him to the police." "How long?" Taehyung asks. His voice sounds rough, raspy and kind of wet Yoongi couldn't bear the thought of the younger with tears. "How long did you stay with him?"
>
> **Underspecified: 1, Category-Specific: 10, Prescriptive: 0**

the forms of threatening and controlling behaviour.[1] This is a potential explanation for the increase in positive annotations from the underspecified to the two other prompts. The *Physical Abuse* example features the specific form of (violent) misuse of medication. Interestingly, the number of positive annotations increased already when asking for *Physical Abuse* instead of only for *Abuse*, so `Flan-T5` appeared to have some association of misuse of medication with *Physical Abuse* in its embedding space. While the prescriptive prompt mentioned *medication* among its examples, it did so only in the context of withholding rather than misusing it.

The first two examples illustrate how prescriptive prompting helped in getting consistently positive annotations on less explicit or more niche forms of *Abuse*. The example for *Sexual Abuse*, however, shows that the prescriptive prompt also resulted in `Flan-T5` missing indisputable examples of *Abuse*. The last two sentences clearly imply sexually abusive behavior but apparently not sufficiently related to the examples given in the prescriptive prompt. The prompt mentions both *rape* and *sexual assault* as well as *inappropriate touching* among its examples but not the word *consent*. It is unclear why the category-specific prompt, explicitly asking for *Sexual Abuse*, also resulted in fewer positive annotations than the underspecified prompt.

Studying more passages that mentioned some form of consent was not fully conclusive. For illustration purposes, Figure 6.8 shows two example passages that contain the word consent. Both passages have references to *Abuse* and mention consent but the first also explicitly uses the word *sexual*. This passage

[1] The term *punish* can indicate physical harm but this is not made clear in the passage.

received two positive annotations for the underspecified, eight for the category-specific, and ten for the prescriptive prompt. The second passage mentions consent and implies that some form of *Abuse* was inflicted by an unknown third person. Despite the passage not clarifying if the *Abuse* was sexual, the category-specific prompt still led to ten positive annotations. The prescriptive prompt, in contrast to that, resulted in zero positive annotations.

In total, 460 of the *Sexual Abuse* passages contained terms related to consent. On these, the category-specific resulted in ten positive annotations on 214 passages. For the underspecified prompt, only 29 passages were annotated unanimously positive and for the prescriptive prompt, the number was 102. The more than sixfold increase from the underspecified to the category-specific prompt suggests that `Flan-T5` used terms related to consent as an indicator for *Sexual Abuse*. The prescriptive prompt, however, resulted in only half as many unanimously positive annotations as the category-specific prompt. As shown by the second example in Figure 6.8, this can result from the nature of the *Abuse* not being explicitly clarified in the passage. While apparently increasing the annotation precision, the prescriptive prompt also resulted in missed passages that contained *Sexual Abuse* such as the third example in Figure 6.7.

Chapter 7

Discussion

This chapter discusses the results of the experiments presented in Chapter 6 and describes their limitations. A general limitation that applies to all findings, especially those related to the tag-vocabulary consistency, is that this thesis worked with documents of fanfiction. As a consequence, the texts mostly discuss fictional worlds and characters. If and how well the findings of this thesis translate to other text domains needs to be explored in future work. The chapter follows the structure of previous chapters by first discussing the results related to vocabulary consistency (Section 7.1) and then moving on the findings concerning prescriptive prompting (Section 7.2).

7.1 Consistency between Tags and Vocabulary

Before discussing the results of the consistency tests, we want to state their core limitation. We tested the consistency between tags and the documents they are applied to exclusively on the lexical level. The vocabularies of expected terms capture neither semantics on sentence or paragraph-level nor idiosyncratic, metaphorical or indirect usage of language. This can be explored in future work with more compute-intensive approaches such as LLMs.

The central finding of Section 6.1 was that the authors apply warning tags related to *Abuse* in a way that is consistent with the expected vocabulary for the warning categories *Emotional*, *Physical*, and *Sexual Abuse*. The share of significant terms, however, deviated between the categories. For *Emotional* and *Sexual Abuse*, the share of terms with significantly higher frequency was 55 % and 53 %, respectively. For *Physical Abuse*, it was only 38 % of all terms and 6 % even occurred significantly less. In addition to that, both *Physical* and *Sexual Abuse* had longer tails of terms with low z-scores, while the distribution of z-scores for *Emotional Abuse* was a lot more right-skewed. This underlines the importance of deriving the vocabulary from authoritative sources as the choice of terms can strongly influence the results of the consistency tests. As a consequence, the findings of this thesis are also limited to the specific vocabulary chosen for each of the categories.

In line with the significance results on term level, the effect size of the increase in mean log ratio was also the lowest for *Physical Abuse*. This is due to the terms from the *Physical Abuse* vocabulary occurring relatively more frequent in the corresponding baseline corpus than the terms from the other two vocabularies. In other words, the chapters on AO3 tagged for *Abuse* without additional categorization appear to use *Physical Abuse* vocabulary relatively more often than *Emotional* or *Sexual Abuse* vocabulary. Whether the warning *Abuse* is generally applied to chapters that contain *Physical Abuse*, can be explored in future research.

The concept *gaslighting* was very prominent among the highly significant and high log ratio terms for *Emotional Abuse*. Literal mentions of the word *gaslighting*, however, imply discourses about behavior that falls under that term instead of the actual behavior of trying to manipulate a person into questioning their own judgements. The actual behavior is highly context dependent and thus difficult to capture lexically.[1] Consequently, semantic methods like transformers are likely more suited to identify *gaslighting* and similar forms of *Emotional Abuse*. Lexical methods are the first step towards developing more sophisticated approaches, for instance through dictionary-based retrieval of passages for annotation.

The analysis of qualifiers revealed only very few consistent effects across the three categories. A possible explanation is that many qualifier subcategories had a (very) low support of 50 chapters or fewer. Empirical evidence for this hypothesis is given by the qualifiers with a support of 300 chapters or more for all categories. Both *mention* and *warning* showed very consistent effects for the differences in mean log ratio. By extending the analyses of this thesis to more warnings from the taxonomy by Wiegmann et al. [2023], future work can explore if other qualifiers also behave more consistently as the number of related chapters is increased. One way to increase support would be to use the seven open-set warnings of the taxonomy like *Abuse* or *Discrimination* as base categories and study the effect of qualifiers on them instead of dividing them into the warning categories like *Emotional Abuse* or *Homophobia*.

While both *mention* and *warning* led to consistent effects, a decrease in mean log ratio was only expected for the former and not the latter. Qualifying a warning by introducing it with *trigger warning for* or *be warned* was expected to increase the log ratio over that for the *base category* with no qualifier restriction. A potential explanation for the opposite effect is that the qualifier

[1] See also the limitation stated at the beginning of the Section.

is not used to emphasize a warning but to distinguish the tag from other, content-related tags. Many tags of a work on AO3 outline which characters are featured or give non-warning-related descriptions like *fluff* or *relationships*. Consequently, some authors might use the *warning* qualifier to draw attention to the fact that this tag indicates something that certain audiences might want to avoid and not to inform about particularly intense descriptions. An interesting avenue for future research would be to explore if the usage of the *warning* qualifier can be traced to a distinguishable subset of authors on AO3 and if these authors communicate the warnings in their tags exclusively with that qualifier.

A qualifier that seemed to indicate increased intensity was the *graphic qualifier*. It was the only qualifier to increase the mean log ratio for all categories. For *Emotional Abuse*, however, no conclusions should be drawn as it was used on only three chapters.

The analysis of rank correlation between terms ranked according to their log ratios yielded no useful insights. The *non-graphic* qualifier was the only one with significant rank correlations for all categories but only with coefficients τ of around 0.2.

7.2 Effect of Prescriptive Annotation Guidelines

The central finding of Section 6.2 was that prescriptive annotation prompting led to a significant increase in mean pairwise annotator agreement over both the underspecified as well as the category-specific prompt. In addition to that, the category-specific prompt led to a significant increase over the underspecified prompt. From these results we conclude that the prescriptive paradigm can increase annotator agreement on the task of trigger warning annotation. A important limitation of these results is of course that they stem from simulated annotations using sociodemographic prompting. Whether prescriptive guidelines also increase the agreement in human annotators needs to be explored in future work. Regardless of the results of a future study with human annotators, it needs to be reiterated that the prescriptive paradigm should not be considered the strictly superior paradigm. This is also true for the task of trigger warning assignment. As noted by Rottger et al. [2022], it is preferable for the goal of annotating one specific belief and therefore useful for training automated classifiers on the application of that belief. Wahlsdorf et al. [2024] emphasized that traumatic experiences are highly subjective and, as a consequence, so are the triggers that can cause a person to have recollections. Hence, the descriptive paradigm is useful for gathering annotations that reflect this subjectivity to study the perceptions of individuals.

In addition to the abstract view of comparing the mean pairwise agreement scores, we explored how consistent `Flan-T5` predicted which other profiles a simulated annotator most agreed with across annotation prompts and categories. Our evaluations showed that which other annotators a simulated annotator most agreed with was very similar regardless of the passages and prompts being related to *Emotional*, *Physical* or *Sexual Abuse*. The relatively low rank correlation for the prescriptive prompt can be explained by the pairwise agreements between simulated annotators being generally very high. This leaves less room for separate groups of profiles agreeing (only) with each other. Beyond the domain of trigger warnings, studying the consistency of agreement between different sociodemographic profiles might be a way for future research to explore the inherent stereotypes of LLMs.

Similar to the results by Wiegmann et al. [2024], a large share of passages was annotated unanimously negative by all three annotation prompts and ten sociodemographic profiles. For both *Emotional* and *Physical Abuse*, the prescriptive prompt increased the pairwise agreement primarily through more unanimously positive annotations, leaving the share of passages with zero positive annotations similar to that for the underspecified prompt. The examples in Section 6.2.3 illustrate that the number of positive annotations increased for category-specific forms of *Abuse*, such as threatening behaviour for *Emotional Abuse* or non-consensual administration of medication for *Physical Abuse*.

In contrast to the other two categories, the pairwise agreement for *Sexual Abuse* was also increased by a larger share of passages being annotated unanimously negative. Without ground truth labels, it is not possible to make statements about the correctness of these annotations. The examples in Figures 6.7 and 6.8, however, show that the prescriptive prompt led to negative annotations on passages that either clearly contained *Sexual Abuse* or can be understood to do so. As noted by Wiegmann et al. [2024], automated trigger warning assignment is a recall-focused task. The strong increase in unanimously negative annotations appears to go against this goal. A potential remedy is to make the list of examples in the prescriptive prompt more extensive as to not risk missing important concepts such as *consent*. The effects of varying the extensiveness of the list of examples in the prescriptive prompt is something that could be explored in future work.

A final analysis in Section 6.2.3 studied how the annotation behavior was affected by restricting passages to only those that contained a term from one of the three term groups used for passage sampling. For vocabulary terms that both had a high log ratio and occurred significantly more often in the category corpus, the effect was the most pronounced. The passages with one of these terms received both fewer unanimously negative annotations as well as more unanimously positive annotations than the average passage with the prescriptive prompt. On *Physical* and *Sexual Abuse*, the same effect also occurred for the category-specific prompt. This result suggest that combining significance testing on document-level term frequencies with corpus-level log ratios is good at identifying terms for the lexical detection of a warning. Beyond testing the consistency between vocabulary and tags, these terms might be useful for lexical approaches to automated trigger warning assignment and have shown to be a good choice for dictionary-based retrieval of passages for annotation.

Chapter 8

Conclusion

This thesis studied the application of trigger warnings to text documents and passages. What causes people with PTSD to have recollections of canonical traumatic experiences (i.e., is a *trigger*) depends on the individual person. Research on the annotation of text passages for the automated assignment of trigger warnings has found this task to be similarly subjective. This poses a challenge for the development of classifiers that assign warnings to pieces of texts. The first research question of this thesis was thus focused on how reliable author-assigned warning tags are from a lexical perspective:

> **RQ1**: Do the authors on Archive of Our Own apply warning tags in a way that is consistent with the vocabulary used in their works?

To answer this question, documents tagged for the three warning categories *Emotional Abuse*, *Physical Abuse*, and *Sexual Abuse* were compared against a baseline of other documents tagged for the warning *Abuse*. For each of the comparisons, a vocabulary of terms was defined that represents what type of content readers expect in a document tagged for one of the warning categories. The statistical tests on the document-level term frequencies and distribution of corpus-level log ratios showed that authors applied tags in a way that is consistent with these expected vocabularies. Consequently, the first research question can be answered in the affirmative. This means that future research can use *Abuse*-related tags in the dataset WTWC-22 to retrieve and subsequently study documents that contain content associated with that warning. Analogously, readers of works on AO3 can use the *Abuse*-related tags to make informed decisions about which works they want to read and which works they want to avoid. More general answers concerning the reliability of warning tags on AO3 depend on future work extending the methodologies of this thesis to other warnings.

The second research question explored the use of prescriptive annotation guidelines for trigger warning annotation:

> **RQ2**: Can prescriptive annotation guidelines increase the annotator agreement on the task of labeling text for automated trigger warning assignment?

Towards answering this question, a group of ten different annotators was simulated using sociodemographic prompting of `Flan-T5 11B`. The LLM was prompted with all combinations of ten different profiles and three different annotation prompts and asked to annotate text passages retrieved with a dictionary approach. The category-specific prompt was found to lead, on average, to a higher pairwise annotator agreement than the underspecified prompt. The prescriptive prompt led to an even higher average agreement, being significantly higher than those for both of the aforementioned descriptive prompts. Consequently, the answer to the second research question is also *yes*. These results are of course limited to the application in sociodemographic prompting and need to be verified in a study with human annotators. A central challenge we observed for the prescriptive prompt is the choice of examples. The prompt for *Sexual Abuse* lacked an explicit mention of *consent*, which resulted in the false negative annotation of passages that mentioned *Sexual Abuse* exclusively as non-consensual behavior without directly qualifying it as sexual. These results could be an artefact of using an LLM for the annotations but serve as a reminder to be thorough in the construction of the prescriptive annotation guidelines.

Trigger warning assignment is a relatively young task in NLP with an inherent subjectivity. The experiments in this thesis demonstrated that lexical methods are capable of testing the consistency with which authors apply trigger warnings to their texts. In addition to that, the annotation experiments revealed that complementing the significance tests on term frequency with the additional measure of the log ratio helps to identify terms with a strong association with their respective warning category. Passages that contained terms with both a significantly higher term frequency and a high log ratio received overall fewer unanimously negative and more unanimously positive annotations than other passages. Taken together, the two methodologies presented in this thesis are useful to (1) study the language used in documents tagged for a warning, (2) retrieve text passages likely to contain warning-related content, and (3) collect more consistent annotations. With these contributions, we hope to help advance the development of methods for automated trigger warning assignment and support the research on how trigger warnings are applied to textual documents.

Appendix A

Vocabulary Consistency

Figure A.1: Scatterplot of z-Score and Log Ratio on Term-Level (Emotional Abuse)

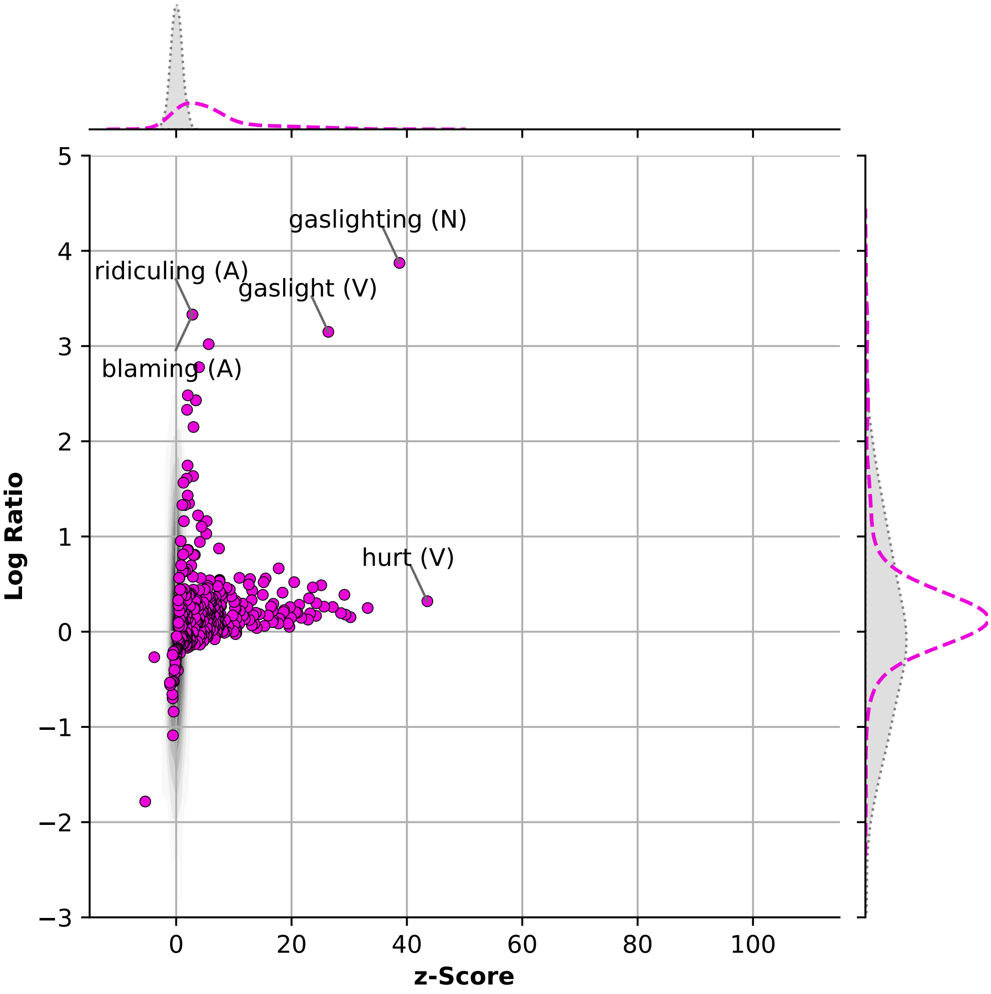


Figure A.2: Scatterplot of z-Score and Log Ratio on Term-Level (Physical Abuse)

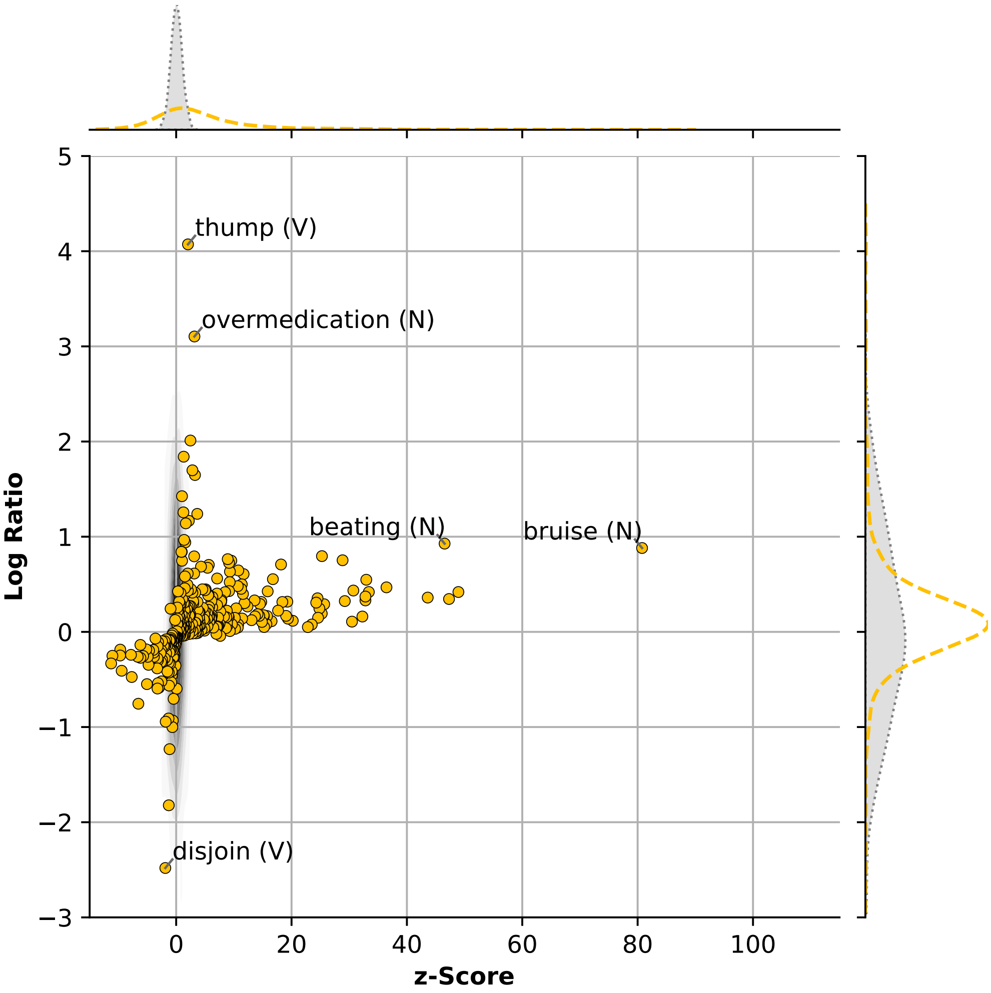


Figure A.3: Scatterplot of z-Score and Log Ratio on Term-Level (Sexual Abuse)

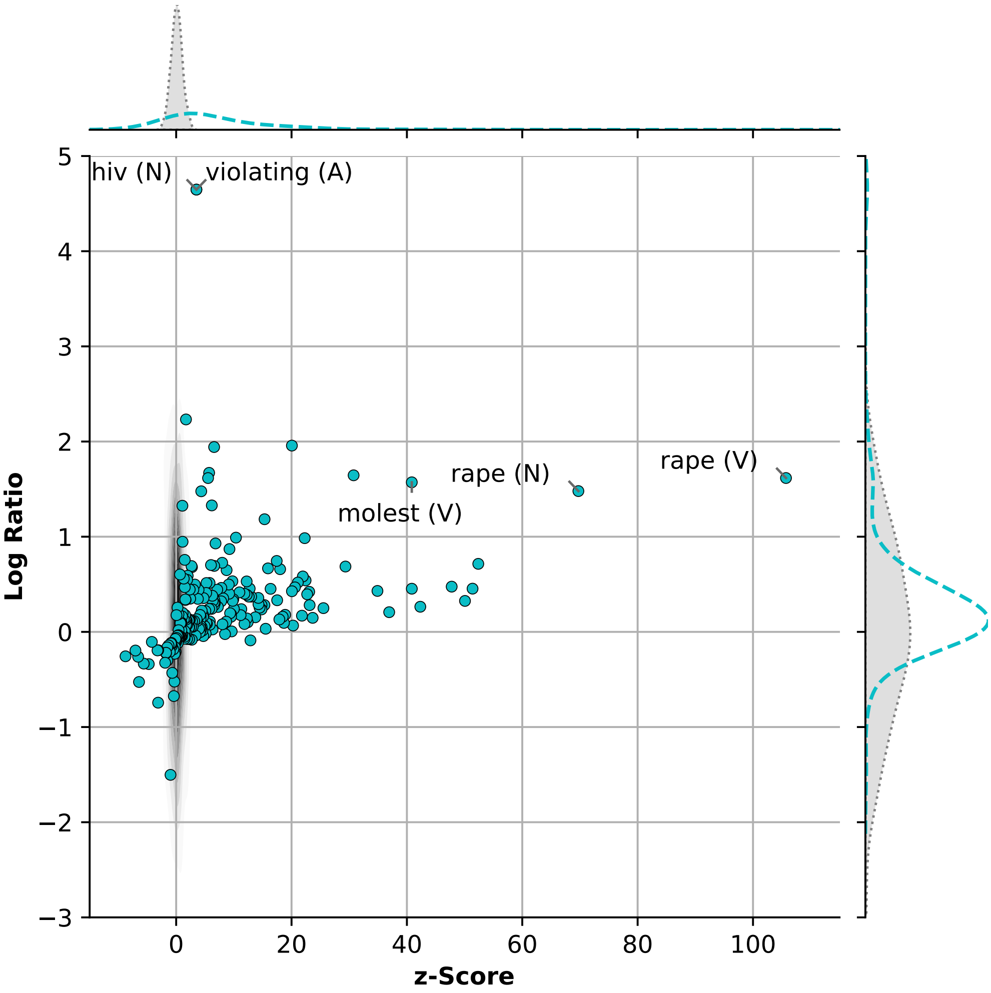


Table A.1: Vocabulary Terms with High z-Score and High Log Ratio. The terms occur among both the 50 terms with the highest z-score and the highest log ratio.

(a) Emotional Abuse

Term	**POS-Tag**	z	**lr**
gaslighting	Noun	38.71	3.87
gaslight	Verb	26.38	3.15
isolation	Noun	17.77	0.67
toxic	Adjective	15.50	0.56

(b) Physical Abuse

Term	**POS-Tag**	z	**lr**
bruise	Noun	80.77	0.88
beating	Noun	46.53	0.93
bruise	Verb	36.46	0.47
cut	Noun	32.96	0.55
punish	Verb	30.72	0.43
bruised	Adjective	28.85	0.75
bruising	Noun	25.27	0.80
welt	Noun	18.18	0.71
swelling	Noun	16.77	0.55
laceration	Noun	11.69	0.61

(c) Sexual Abuse

Term	**POS-Tag**	z	**lr**
rape	Verb	105.72	1.62
rape	Noun	69.73	1.48
sexual	Adjective	52.38	0.71
sex	Noun	47.76	0.48
molest	Verb	40.85	1.57
pedophile	Noun	30.76	1.65
violate	Verb	29.35	0.69
suicide	Noun	22.44	0.54
consensual	Adjective	22.28	0.98
depression	Noun	21.96	0.58
consent	Noun	21.09	0.52
penis	Noun	20.59	0.47
molestation	Noun	20.06	1.96
consent	Verb	18.02	0.66
genital	Noun	17.43	0.75
suicidal	Adjective	15.92	0.67
std	Noun	15.32	1.18
violation	Noun	12.74	0.45

Appendix B

Annotation Experiments

Table B.1: Number of Positive Annotations per Category, Prompt, and Term Group

		High z-Score			**High Log Ratio**			**High z-Score (Non-V.)**		
Category	#	Und.	Cat.	Pre.	Und.	Cat.	Pre.	Und.	Cat.	Pre.
Emotional Abuse	0	4,358	3,544	4,250	1,752	1,422	1,568	5,283	4,436	5,209
	1	410	418	139	121	154	43	394	449	126
	2	267	277	100	70	77	37	249	291	112
	3	161	182	73	52	67	22	156	186	77
	4	126	155	66	28	56	27	110	154	80
	5	119	145	56	31	47	18	127	132	60
	6	139	208	61	24	46	23	119	188	61
	7	91	144	55	31	52	25	94	151	62
	8	84	154	73	30	49	24	79	144	69
	9	81	188	108	25	52	39	83	160	102
	10	120	541	975	39	181	377	106	509	842
Physical Abuse	0	3,532	3,058	3,609	1,622	1,302	1,487	5,085	4,733	5,412
	1	349	267	67	165	115	35	448	292	66
	2	281	156	45	137	82	27	336	185	52
	3	239	121	48	93	76	22	274	154	41
	4	183	119	25	78	56	23	217	130	37
	5	166	132	34	95	60	11	208	139	27
	6	155	151	50	76	98	28	178	171	50
	7	152	134	36	72	80	30	175	153	49
	8	167	186	50	79	91	22	196	206	38
	9	144	226	55	69	118	34	151	255	63
	10	191	1,009	1,540	78	486	845	211	1,061	1,644
Sexual Abuse	0	3,396	2,428	4,266	774	518	910	4,833	3,664	6,105
	1	375	210	74	96	43	22	515	313	97
	2	289	128	60	72	22	20	387	179	69
	3	265	93	39	61	25	12	299	141	53
	4	204	102	37	48	19	10	251	142	48
	5	198	82	37	72	29	11	263	118	42
	6	194	139	30	64	31	12	255	173	40
	7	173	100	35	46	31	8	202	129	36
	8	198	129	47	63	45	19	226	181	57
	9	225	224	75	78	55	23	244	281	77
	10	339	2,221	1,156	124	680	451	371	2,525	1,222

Table B.2: Ranked Agreements between Sociodemographic Profiles. For each profile p, the columns contain a list of other profiles ranked in descending order (left to right) by their agreement scores (Cohen's κ) for a given annotation prompt.

p	**Category**	**Underspecified**	**Category-Specific**	**Prescriptive**
0	Emotional	[6, 5, 8, 4, 9, 2, 7, 3, 1]	[5, 6, 8, 2, 1, 3, 4, 9, 7]	[4, 2, 1, 6, 3, 5, 8, 9, 7]
	Physical	[6, 5, 9, 8, 4, 2, 7, 1, 3]	[8, 6, 5, 9, 1, 2, 3, 4, 7]	[2, 4, 7, 6, 1, 9, 5, 3, 8]
	Sexual	[6, 5, 8, 9, 4, 2, 7, 3, 1]	[5, 8, 6, 1, 2, 3, 9, 4, 7]	[2, 6, 9, 4, 1, 3, 5, 8, 7]
1	Emotional	[3, 5, 7, 9, 6, 2, 8, 4, 0]	[3, 2, 5, 7, 0, 6, 4, 9, 8]	[5, 3, 9, 0, 2, 4, 6, 8, 7]
	Sexual	[3, 7, 5, 9, 2, 6, 4, 8, 0]	[3, 2, 5, 7, 0, 6, 9, 8, 4]	[5, 9, 3, 2, 7, 0, 6, 4, 8]
	Physical	[3, 5, 7, 9, 2, 6, 4, 8, 0]	[3, 2, 7, 6, 5, 9, 8, 0, 4]	[5, 3, 9, 2, 7, 0, 8, 6, 4]
2	Emotional	[7, 3, 1, 4, 6, 5, 9, 0, 8]	[3, 1, 7, 6, 5, 0, 4, 8, 9]	[3, 0, 7, 4, 1, 5, 6, 8, 9]
	Sexual	[7, 3, 1, 4, 6, 5, 9, 0, 8]	[3, 1, 7, 5, 6, 0, 8, 9, 4]	[0, 3, 7, 1, 5, 6, 9, 4, 8]
	Physical	[7, 4, 3, 1, 6, 5, 9, 0, 8]	[3, 1, 7, 6, 5, 9, 8, 0, 4]	[0, 7, 3, 4, 1, 6, 9, 5, 8]
3	Emotional	[1, 7, 5, 6, 2, 9, 8, 4, 0]	[1, 5, 2, 7, 6, 0, 9, 4, 8]	[2, 1, 5, 0, 4, 7, 8, 9, 6]
	Sexual	[1, 7, 5, 2, 6, 9, 4, 8, 0]	[1, 2, 5, 7, 6, 0, 9, 8, 4]	[1, 7, 5, 6, 2, 9, 8, 4, 0]
	Physical	[1, 7, 5, 2, 6, 9, 4, 8, 0]	[1, 2, 7, 5, 6, 9, 4, 0, 8]	[1, 2, 7, 5, 6, 4, 8, 9, 0]
4	Emotional	[2, 6, 0, 7, 5, 1, 3, 9, 8]	[6, 2, 5, 3, 0, 1, 8, 7, 9]	[6, 0, 2, 3, 8, 9, 7, 5, 1]
	Sexual	[2, 6, 0, 7, 1, 5, 3, 9, 8]	[6, 5, 9, 0, 3, 8, 2, 1, 7]	[6, 7, 3, 9, 0, 2, 5, 1, 8]
	Physical	[2, 6, 0, 7, 5, 3, 9, 1, 8]	[6, 5, 1, 3, 2, 8, 0, 9, 7]	[7, 6, 0, 2, 5, 9, 3, 1, 8]
5	Emotional	[9, 6, 1, 3, 8, 0, 7, 2, 4]	[6, 3, 9, 8, 0, 1, 2, 7, 4]	[9, 1, 3, 8, 7, 4, 2, 6, 0]
	Sexual	[9, 6, 1, 8, 3, 0, 7, 2, 4]	[6, 9, 8, 3, 0, 2, 1, 4, 7]	[9, 1, 3, 8, 7, 6, 2, 4, 0]
	Physical	[9, 6, 1, 3, 8, 0, 7, 2, 4]	[9, 6, 8, 0, 1, 3, 7, 2, 4]	[9, 1, 8, 7, 3, 4, 6, 2, 0]
6	Emotional	[5, 9, 7, 3, 0, 1, 2, 4, 8]	[5, 4, 9, 8, 7, 0, 2, 3, 1]	[4, 8, 0, 9, 7, 3, 5, 2, 1]
	Sexual	[5, 9, 0, 7, 3, 1, 8, 4, 2]	[5, 9, 8, 4, 7, 3, 0, 2, 1]	[7, 4, 3, 9, 0, 5, 2, 8, 1]
	Physical	[5, 9, 7, 0, 4, 3, 1, 2, 8]	[5, 9, 8, 4, 7, 1, 0, 3, 2]	[7, 4, 8, 9, 0, 2, 5, 3, 1]
7	Emotional	[2, 3, 1, 6, 5, 9, 4, 0, 8]	[2, 6, 3, 1, 9, 5, 8, 0, 4]	[2, 8, 9, 3, 6, 5, 4, 0, 1]
	Sexual	[2, 3, 1, 6, 9, 5, 4, 0, 8]	[2, 3, 6, 1, 9, 5, 8, 0, 4]	[6, 3, 2, 4, 5, 9, 1, 8, 0]
	Physical	[2, 3, 6, 1, 9, 5, 4, 0, 8]	[3, 1, 2, 6, 9, 5, 8, 4, 0]	[6, 2, 4, 9, 8, 3, 0, 5, 1]
8	Emotional	[9, 5, 6, 0, 1, 3, 7, 2, 4]	[9, 6, 5, 0, 7, 3, 2, 1, 4]	[9, 6, 5, 7, 4, 3, 2, 0, 1]
	Sexual	[9, 5, 6, 0, 3, 1, 7, 2, 4]	[9, 6, 5, 0, 2, 7, 3, 1, 4]	[5, 9, 3, 6, 0, 7, 1, 2, 4]
	Physical	[9, 5, 0, 6, 1, 3, 7, 2, 4]	[9, 6, 5, 0, 2, 1, 7, 3, 4]	[9, 6, 7, 5, 1, 3, 2, 0, 4]
9	Emotional	[5, 6, 8, 1, 3, 0, 7, 2, 4]	[8, 5, 6, 7, 3, 0, 2, 1, 4]	[8, 5, 1, 6, 7, 4, 3, 0, 2]
	Sexual	[5, 6, 8, 1, 3, 0, 7, 2, 4]	[8, 6, 5, 7, 3, 0, 2, 4, 1]	[5, 1, 3, 6, 8, 7, 4, 0, 2]
	Physical	[5, 6, 1, 8, 3, 7, 0, 2, 4]	[5, 8, 6, 7, 1, 0, 3, 2, 4]	[5, 7, 8, 1, 6, 0, 4, 2, 3]

Table B.3: Share of Unanimous Annotations for Emotional Abuse. The columns show the term group, z-score, log ratio, and the share of unanimously positively (Pos.) and negatively (Neg.) annotated passages for each prompt. The terms are sorted by the share of unanimously positive annotations for the prescriptive prompt.

				Undersp.		**Cat.-Spec.**		**Prescr.**		
Term (POS)	**Group**	z	**lr**	Pos.	Neg.	Pos.	Neg.	Pos.	Neg.	s
gaslight (V)	A	26.38	3.15	0.28	0.22	0.02	0.74	0.80	0.11	54
gaslighting (N)	A	38.71	3.87	0.44	0.15	0.13	0.43	0.79	0.09	131
infantilize (V)	B	4.41	1.10	0.50	0.50	0.00	0.50	0.50	0.00	2
manipulated (A)	B	5.63	3.02	0.17	0.44	0.06	0.63	0.48	0.36	204
shaming (A)	B	2.02	2.48	0.11	0.56	0.00	0.76	0.33	0.41	70
ridiculed (A)	B	1.98	1.75	0.17	0.51	0.01	0.77	0.32	0.53	77
manipulate (V)	A	25.16	0.49	0.07	0.58	0.01	0.72	0.30	0.54	201
punish (V)	A	23.04	0.35	0.17	0.39	0.03	0.56	0.29	0.52	247
punishment (N)	A	21.33	0.28	0.12	0.47	0.03	0.62	0.21	0.62	307
deserve (V)	C	30.39	0.30	0.13	0.56	0.03	0.72	0.21	0.67	418
invalidate (V)	B	7.43	0.87	0.03	0.67	0.00	0.91	0.21	0.73	33
berating (A)	B	2.93	1.63	0.13	0.52	0.01	0.79	0.21	0.67	170
sick (A)	C	34.68	0.30	0.10	0.61	0.03	0.71	0.20	0.69	466
sniveling (A)	B	3.42	2.43	0.07	0.54	0.01	0.70	0.20	0.65	71
rage (N)	A	21.03	0.20	0.10	0.51	0.02	0.68	0.18	0.62	365
anger (N)	A	24.24	0.17	0.07	0.60	0.01	0.74	0.18	0.66	592
cry (V)	A	22.86	0.13	0.09	0.53	0.02	0.65	0.17	0.69	545
fault (N)	C	30.10	0.30	0.10	0.57	0.02	0.77	0.16	0.70	585
angry (A)	A	21.69	0.15	0.08	0.62	0.02	0.74	0.16	0.70	587
fear (N)	A	33.21	0.25	0.10	0.55	0.02	0.71	0.16	0.68	653
anorexic (A)	B	5.29	1.16	0.08	0.65	0.00	0.85	0.15	0.77	26
blamed (A)	B	3.97	2.78	0.07	0.64	0.02	0.79	0.15	0.75	207
hurt (V)	A	43.55	0.32	0.10	0.55	0.02	0.70	0.15	0.72	1,196
lie (N)	C	33.97	0.40	0.08	0.59	0.01	0.77	0.15	0.70	485
scared (A)	A	23.22	0.19	0.07	0.64	0.02	0.72	0.15	0.77	567
sob (V)	A	24.29	0.30	0.13	0.49	0.02	0.63	0.15	0.73	320
flinch (V)	C	34.14	0.35	0.10	0.56	0.02	0.70	0.15	0.71	260
awful (A)	C	29.12	0.36	0.10	0.60	0.02	0.75	0.15	0.73	344
hurting (A)	B	2.00	1.43	0.11	0.54	0.02	0.70	0.14	0.74	373
force (V)	A	30.22	0.15	0.08	0.64	0.02	0.69	0.14	0.72	477
wrong (A)	C	31.33	0.21	0.07	0.63	0.01	0.77	0.14	0.74	959
blaming (A)	B	2.83	3.33	0.05	0.68	0.01	0.83	0.14	0.75	198
tear (N)	A	29.35	0.18	0.07	0.56	0.02	0.72	0.14	0.75	347
understand (V)	C	28.29	0.13	0.07	0.65	0.02	0.79	0.14	0.75	733
matter (V)	C	30.04	0.26	0.07	0.63	0.01	0.77	0.14	0.74	785
try (V)	C	31.63	0.13	0.08	0.66	0.02	0.79	0.13	0.76	860
trust (V)	A	28.57	0.20	0.07	0.69	0.01	0.82	0.13	0.78	560
guilt (N)	A	25.64	0.26	0.06	0.64	0.01	0.81	0.13	0.77	394
cold (A)	C	27.81	0.16	0.10	0.62	0.02	0.73	0.13	0.71	678
remember (V)	C	30.38	0.19	0.07	0.66	0.02	0.74	0.12	0.78	674
panic (N)	A	27.15	0.26	0.06	0.65	0.01	0.76	0.11	0.80	422
disregarding (A)	B	3.00	2.15	0.05	0.72	0.00	0.85	0.11	0.78	156
dread (N)	A	23.68	0.47	0.07	0.67	0.00	0.81	0.11	0.78	259
memory (N)	C	29.28	0.26	0.05	0.67	0.01	0.79	0.11	0.80	399
swallow (V)	C	29.51	0.20	0.06	0.61	0.03	0.72	0.10	0.78	280
promise (V)	C	27.71	0.18	0.05	0.68	0.01	0.80	0.10	0.82	494
easy (A)	C	27.37	0.15	0.05	0.73	0.01	0.80	0.10	0.81	509
seclusion (N)	B	5.22	1.03	0.03	0.87	0.02	0.94	0.10	0.82	113
stressing (A)	B	2.07	0.86	0.02	0.85	0.01	0.92	0.09	0.88	171
abandoned (A)	B	3.80	1.22	0.06	0.71	0.02	0.83	0.09	0.79	251
anxiety (N)	A	29.17	0.39	0.05	0.75	0.00	0.88	0.08	0.86	306
blink (V)	C	29.86	0.22	0.06	0.67	0.01	0.78	0.08	0.82	259
catfish (V)	B	1.97	0.86	0.00	0.85	0.00	0.77	0.08	0.85	13
suppose (V)	C	29.58	0.17	0.04	0.76	0.01	0.87	0.08	0.85	319
safe (A)	C	27.88	0.18	0.05	0.69	0.01	0.80	0.07	0.85	588
comfort (N)	C	28.45	0.31	0.06	0.71	0.02	0.82	0.06	0.87	393
ridiculing (A)	B	2.83	3.33	0.00	0.78	0.06	0.83	0.06	0.72	18
prohibited (A)	B	3.42	2.43	0.04	0.81	0.00	0.86	0.04	0.81	69
desertion (N)	B	4.11	0.94	0.00	0.90	0.00	0.90	0.00	0.97	31
invalidation (N)	B	2.24	1.35	0.00	0.50	0.00	1.00	0.00	0.50	2

Table B.4: Share of Unanimous Annotations for Physical Abuse. The columns show the term group, z-score, log ratio, and the share of unanimously positively (Pos.) and negatively (Neg.) annotated passages for each prompt. The terms are sorted by the share of unanimously positive annotations for the prescriptive prompt.

				Undersp.		**Cat.-Spec.**		**Prescr.**		
Term (POS)	**Group**	z	**lr**	Pos.	Neg.	Pos.	Neg.	Pos.	Neg.	s
caning (N)	B	3.12	0.79	0.41	0.03	0.10	0.10	0.83	0.10	29
shackled (A)	B	3.12	0.61	0.35	0.34	0.05	0.40	0.67	0.25	172
welt (N)	B	18.18	0.71	0.28	0.25	0.05	0.46	0.58	0.26	146
gashed (A)	B	3.66	1.24	0.30	0.23	0.05	0.60	0.50	0.35	40
bruised (A)	A	28.85	0.75	0.32	0.26	0.07	0.45	0.48	0.38	297
bruising (N)	A	25.27	0.80	0.33	0.24	0.06	0.45	0.46	0.45	226
bruise (N)	A	80.77	0.88	0.32	0.26	0.04	0.47	0.45	0.41	318
hit (V)	A	32.28	0.16	0.22	0.46	0.04	0.61	0.42	0.50	713
pain (N)	C	64.59	0.37	0.22	0.42	0.04	0.53	0.42	0.51	1,117
abuse (N)	C	79.61	1.19	0.29	0.32	0.15	0.21	0.41	0.50	280
bleed (V)	A	24.83	0.24	0.26	0.40	0.05	0.49	0.41	0.50	239
gash (V)	B	5.66	0.70	0.27	0.30	0.04	0.62	0.40	0.45	198
hitting (A)	B	2.82	1.70	0.23	0.47	0.03	0.63	0.40	0.50	326
beating (N)	A	46.53	0.93	0.27	0.45	0.06	0.52	0.38	0.56	323
beat (V)	A	47.30	0.35	0.27	0.47	0.06	0.54	0.37	0.56	447
punishment (N)	A	33.40	0.42	0.21	0.48	0.05	0.49	0.35	0.57	289
chafed (A)	B	2.01	0.61	0.22	0.52	0.02	0.62	0.34	0.60	98
broken (A)	A	32.81	0.33	0.22	0.50	0.05	0.58	0.33	0.62	578
whipping (N)	B	10.77	0.65	0.21	0.56	0.04	0.63	0.33	0.63	205
laceration (N)	B	11.69	0.61	0.20	0.30	0.02	0.67	0.30	0.55	82
hematoma (N)	B	4.31	0.69	0.15	0.50	0.05	0.55	0.30	0.70	20
cry (V)	C	56.19	0.34	0.21	0.56	0.05	0.54	0.28	0.65	552
hurt (V)	C	75.90	0.43	0.17	0.54	0.04	0.60	0.28	0.65	1,124
cut (N)	A	32.96	0.55	0.18	0.53	0.03	0.61	0.27	0.65	586
wound (N)	A	25.24	0.19	0.21	0.51	0.03	0.68	0.27	0.63	372
numb (A)	A	24.51	0.35	0.17	0.58	0.03	0.62	0.27	0.65	261
scalding (A)	B	2.47	2.01	0.12	0.58	0.02	0.66	0.26	0.65	173
pushed (A)	B	2.19	1.17	0.16	0.58	0.03	0.64	0.26	0.66	581
fear (N)	C	47.05	0.29	0.16	0.59	0.03	0.60	0.26	0.69	571
punish (V)	A	30.72	0.43	0.15	0.54	0.03	0.54	0.25	0.63	213
shiner (N)	B	9.56	0.75	0.13	0.50	0.00	0.71	0.25	0.59	103
father (N)	C	53.29	0.50	0.18	0.62	0.04	0.58	0.24	0.71	665
scar (N)	A	29.23	0.32	0.20	0.54	0.04	0.69	0.23	0.71	284
flinch (V)	A	43.61	0.36	0.10	0.60	0.02	0.63	0.22	0.70	242
malnourishment (N)	B	5.43	0.67	0.14	0.74	0.05	0.81	0.21	0.71	42
sob (V)	C	43.93	0.47	0.17	0.61	0.02	0.57	0.21	0.72	327
thump (V)	B	2.05	4.07	0.12	0.65	0.01	0.72	0.21	0.72	188
door (N)	C	58.44	0.22	0.14	0.66	0.03	0.69	0.20	0.74	1,331
room (N)	C	43.68	0.14	0.13	0.67	0.03	0.70	0.20	0.75	1,514
injury (N)	A	25.66	0.29	0.11	0.66	0.02	0.79	0.20	0.75	250
break (V)	A	30.48	0.11	0.15	0.68	0.02	0.72	0.19	0.76	486
scared (A)	A	48.93	0.42	0.14	0.65	0.03	0.68	0.19	0.76	509
sorry (A)	C	50.40	0.26	0.12	0.67	0.02	0.68	0.19	0.76	848
lock (V)	A	24.60	0.15	0.13	0.72	0.03	0.70	0.19	0.75	297
shake (V)	C	45.16	0.17	0.12	0.68	0.03	0.71	0.18	0.75	363
discoloration (N)	B	9.18	0.72	0.15	0.47	0.02	0.67	0.18	0.70	94
tear (N)	C	82.37	0.47	0.13	0.65	0.03	0.69	0.18	0.76	352
house (N)	C	50.39	0.29	0.12	0.71	0.04	0.71	0.17	0.79	639
hospital (N)	C	44.15	0.47	0.11	0.69	0.03	0.75	0.17	0.79	395
stay (V)	C	44.77	0.18	0.10	0.72	0.02	0.76	0.16	0.80	642
walk (V)	C	47.10	0.19	0.10	0.73	0.02	0.78	0.14	0.82	516
sit (V)	C	47.50	0.16	0.11	0.73	0.02	0.78	0.13	0.83	515
kitchen (N)	C	43.52	0.30	0.09	0.77	0.02	0.81	0.13	0.84	456
okay (A)	C	49.00	0.31	0.09	0.72	0.02	0.77	0.13	0.83	935
car (N)	C	44.16	0.32	0.09	0.76	0.02	0.78	0.12	0.83	520
contusion (N)	B	8.94	0.76	0.12	0.41	0.00	0.82	0.12	0.65	17
overdose (V)	B	9.32	0.63	0.06	0.79	0.01	0.80	0.10	0.85	172
anxiety (N)	A	32.81	0.37	0.03	0.81	0.00	0.86	0.08	0.86	272
overmedicate (V)	B	3.26	1.65	0.00	0.00	0.00	0.00	0.00	0.00	0
overmedication (N)	B	3.17	3.10	0.00	0.00	0.00	0.00	0.00	0.00	0

Table B.5: Share of Unanimous Annotations for Sexual Abuse. The columns show the term group, z-score, log ratio, and the share of unanimously positively (Pos.) and negatively (Neg.) annotated passages for each prompt. The terms are sorted by the share of unanimously positive annotations for the prescriptive prompt.

				Undersp.		**Cat.-Spec.**		**Prescr.**		
Term (POS)	**Group**	z	**lr**	Pos.	Neg.	Pos.	Neg.	Pos.	Neg.	s
molestation (N)	B	20.06	1.96	0.83	0.01	0.22	0.14	0.70	0.17	92
rape (V)	A	105.72	1.62	0.86	0.01	0.25	0.12	0.66	0.18	403
violated (A)	B	4.35	1.48	0.74	0.09	0.12	0.21	0.64	0.24	217
molest (V)	A	40.85	1.57	0.76	0.06	0.12	0.28	0.56	0.33	147
pedophile (N)	A	30.76	1.65	0.79	0.06	0.14	0.26	0.42	0.43	174
sexual (A)	A	52.38	0.71	0.60	0.18	0.13	0.45	0.41	0.49	446
violate (V)	A	29.35	0.69	0.59	0.28	0.10	0.40	0.39	0.51	170
genital (N)	B	17.43	0.75	0.70	0.11	0.10	0.46	0.39	0.53	70
violating (A)	B	3.51	4.65	0.60	0.25	0.13	0.36	0.37	0.49	182
abuse (N)	C	84.87	1.36	0.53	0.25	0.18	0.15	0.37	0.52	435
whore (N)	C	51.37	0.80	0.62	0.13	0.09	0.26	0.35	0.52	322
rectal (A)	B	6.82	0.93	0.79	0.00	0.10	0.38	0.34	0.48	29
assaulted (A)	B	7.98	0.73	0.46	0.32	0.07	0.46	0.34	0.58	225
consensual (A)	A	22.28	0.98	0.57	0.21	0.07	0.49	0.24	0.64	194
sex (N)	A	47.76	0.48	0.49	0.21	0.06	0.57	0.23	0.66	663
pain (N)	C	49.23	0.28	0.39	0.40	0.05	0.48	0.21	0.72	1,050
consent (V)	B	18.02	0.66	0.40	0.35	0.06	0.61	0.20	0.71	271
victimize (V)	B	10.36	0.99	0.30	0.50	0.00	0.60	0.20	0.70	10
touch (V)	A	50.07	0.33	0.43	0.34	0.04	0.57	0.20	0.71	856
bleed (V)	A	25.53	0.25	0.42	0.39	0.09	0.45	0.20	0.72	244
sick (A)	C	54.19	0.47	0.36	0.47	0.06	0.56	0.19	0.73	613
exploitation (N)	B	6.04	0.70	0.27	0.52	0.08	0.56	0.19	0.77	52
drug (N)	C	43.22	0.60	0.31	0.52	0.05	0.66	0.19	0.77	278
force (V)	A	36.93	0.21	0.32	0.48	0.04	0.57	0.18	0.75	534
sob (V)	C	49.97	0.56	0.38	0.39	0.05	0.55	0.17	0.75	394
fear (N)	A	42.33	0.26	0.35	0.43	0.06	0.57	0.17	0.76	713
hurt (V)	C	75.83	0.48	0.33	0.43	0.04	0.54	0.16	0.77	1,271
feel (V)	C	51.49	0.17	0.31	0.48	0.04	0.62	0.16	0.78	1,733
cry (V)	C	60.95	0.39	0.32	0.45	0.05	0.56	0.16	0.77	644
want (V)	C	62.72	0.21	0.34	0.46	0.05	0.62	0.15	0.78	2,321
trauma (N)	C	54.15	1.02	0.29	0.49	0.05	0.66	0.15	0.79	335
bed (N)	C	47.75	0.21	0.34	0.40	0.04	0.63	0.15	0.78	1,160
tear (N)	C	67.94	0.40	0.29	0.49	0.06	0.57	0.14	0.81	414
terrified (A)	A	22.71	0.39	0.28	0.45	0.04	0.57	0.13	0.81	391
stroke (V)	A	21.80	0.17	0.38	0.40	0.04	0.71	0.13	0.81	273
vulnerable (A)	A	23.14	0.28	0.26	0.50	0.02	0.68	0.13	0.82	330
baby (N)	C	43.56	0.40	0.29	0.52	0.04	0.68	0.12	0.82	637
pregnant (A)	A	23.06	0.42	0.20	0.58	0.04	0.73	0.12	0.84	297
bathroom (N)	C	49.02	0.41	0.31	0.46	0.03	0.65	0.12	0.82	495
scared (A)	A	51.38	0.45	0.24	0.54	0.03	0.66	0.10	0.85	654
shake (V)	C	43.24	0.18	0.23	0.53	0.03	0.71	0.09	0.87	426
bruise (N)	A	40.85	0.45	0.29	0.47	0.04	0.42	0.09	0.83	273
chlamydia (N)	B	6.18	1.33	0.09	0.45	0.00	0.64	0.09	0.73	11
suicidal (A)	B	15.92	0.67	0.12	0.74	0.05	0.80	0.08	0.91	199
safe (A)	C	48.36	0.33	0.21	0.57	0.03	0.69	0.08	0.87	736
okay (A)	C	55.29	0.41	0.22	0.57	0.02	0.73	0.08	0.87	1,081
suicide (N)	A	22.44	0.54	0.13	0.78	0.04	0.82	0.08	0.91	240
nightmare (N)	C	43.88	0.46	0.22	0.57	0.03	0.72	0.07	0.88	330
depression (N)	A	21.96	0.58	0.13	0.77	0.04	0.79	0.07	0.90	231
dysfunction (N)	B	6.60	0.69	0.13	0.69	0.02	0.84	0.07	0.91	45
therapy (N)	C	51.59	1.01	0.14	0.65	0.02	0.78	0.07	0.90	303
gonorrhea (N)	B	6.58	1.94	0.44	0.25	0.00	0.75	0.06	0.88	16
syphilis (N)	B	5.53	1.62	0.25	0.38	0.00	0.81	0.06	0.69	16
anxiety (N)	A	34.89	0.43	0.17	0.65	0.01	0.81	0.05	0.91	360
promiscuous (A)	B	9.24	0.87	0.20	0.50	0.03	0.78	0.05	0.84	100
hypersexuality (N)	B	5.71	1.67	0.00	0.00	0.00	0.00	0.00	0.00	0
fingermark (N)	B	2.74	0.68	0.00	0.00	0.00	0.00	0.00	0.00	0
hiv (N)	B	3.51	4.65	0.00	0.00	0.00	0.00	0.00	0.00	0
std (N)	B	15.32	1.18	0.33	0.00	0.00	0.67	0.00	1.00	3
incontinence (N)	B	2.66	0.69	0.29	0.71	0.00	0.86	0.00	1.00	7

Bibliography

Khurshid Ahmad, Lee Gillam, and Lena Tostevin. University of Surrey Participation in TREC8: Weirdness Indexing for Logical Document Extrapolation and Retrieval (WILDER). In Ellen M. Voorhees and Donna K. Harman, editors, *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), 1999. URL `http://trec.nist.gov/pubs/trec8/papers/surrey2.pdf`.

Sohail Akhtar, Valerio Basile, and Viviana Patti. Modeling Annotator Perspective and Polarized Opinions to Improve Hate Speech Detection. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 8(1): 151–154, October 2020. doi: 10.1609/hcomp.v8i1.7473.

Hala Al Kuwatly, Maximilian Wich, and Georg Groh. Identifying and Measuring Annotator Bias Based on Annotators' Demographic Characteristics. In Seyi Akiwowo, Bertie Vidgen, Vinodkumar Prabhakaran, and Zeerak Waseem, editors, *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 184–190, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.alw-1.21.

Abeer ALDayel and Walid Magdy. Stance Detection on Social Media: State of the Art and Trends. *Information Processing & Management*, 58(4):102597, 2021. ISSN 0306-4573. doi: https://doi.org/10.1016/j.ipm.2021.102597.

Giambattista Amati. Frequentist and Bayesian Approach to Information Retrieval. In *Proceedings of the 28th European Conference on Advances in Information Retrieval*, ECIR'06, page 13–24, Berlin, Heidelberg, 2006. Springer-Verlag. ISBN 3540333479. doi: 10.1007/11735106_3. URL `https://doi.org/10.1007/11735106_3`.

American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders: DSM-5*, volume 5. American Psychiatric Association Washington, DC, 2013.

Valerio Basile, Michael Fell, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, Massimo Poesio, and Alexandra Uma. We Need to Consider Disagreement in Evaluation. In Kenneth Church, Mark Liberman, and Valia Kordoni, editors, *Proceedings of the 1st Workshop on Benchmarking: Past, Present and Future*, pages 15–21, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.bppf-1.3. URL `https://aclanthology.org/2021.bppf-1.3`.

Tilman Beck, Hendrik Schuff, Anne Lauscher, and Iryna Gurevych. Sensitivity, Performance, Robustness: Deconstructing the Effect of Sociodemographic Prompting. In Yvette Graham and Matthew Purver, editors, *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, March 2024. Association for Computational Linguistics. doi: 10.48550/arXiv.2309.07034.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. Longformer: The Long-Document Transformer, 2020. URL `https://arxiv.org/abs/2004.05150`.

Guy A. Boysen. Evidence-Based Answers to Questions About Trigger Warnings for Clinically-Based Distress: A Review for Teachers. *Scholarship of Teaching and Learning in Psychology*, 3:163–177, 06 2017. doi: 10.1037/stl0000084.

Victoria M. E. Bridgland, Deanne M. Green, Jacinta M. Oulton, and Melanie K. T. Takarangi. Expecting the Worst: Investigating the Effects of Trigger Warnings on Reactions to Ambiguously Themed Photos. *Journal of Experimental Psychology: Applied*, 25:602–617, 2019. doi: 10.1037/xap0000215.

Victoria M. E. Bridgland, Payton J. Jones, and Benjamin W. Bellet. A Meta-Analysis of the Efficacy of Trigger Warnings, Content Warnings, and Content Notes. *Clinical Psychological Science*, 0, 2023. doi: 10.1177/21677026231186625.

Madeline J. Bruce, Sara M. Stasik-O'Brien, and Heather Hoffmann. Students' Psychophysiological Reactivity to Trigger Warnings. *Current Psychology*, 42:5470–5479, 2021. doi: 10.1007/s12144-021-01895-1.

Tianqi Chen and Carlos Guestrin. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 785–794, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450342322. doi: 10.1145/2939672.2939785. URL `https://doi.org/10.1145/2939672.2939785`.

Ameet Deshpande, Vishvak Murahari, Tanmay Rajpurohit, Ashwin Kalyan, and Karthik Narasimhan. Toxicity in ChatGPT: Analyzing Persona-Assigned Language Models. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1236–1270, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.88.

Victoria L. Dickman-Burnett and Maribeth Geaman. Untangling the Trigger-Warning Debate: Curating a Complete Toolkit for Compassionate Praxis in the Classroom. *Journal of Thought*, 53:35–52, 2019. ISSN 00225231, 2375270X.

Ted Dunning. Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics*, 19(1):61–74, 1993. URL `https://aclanthology.org/J93-1003`.

Tanya Goyal, Junyi Jessy Li, and Greg Durrett. FALTE: A Toolkit for Fine-grained Annotation for Long Text Evaluation. In Wanxiang Che and Ekaterina Shutova, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 351–358, Abu Dhabi, UAE, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-demos.35.

James Grimmelmann. The Virtues of Moderation, 2015. URL `http://hdl.handle.net/20.500.13051/7798`.

Daniel Grupe and Jack Nitschke. Uncertainty and Anticipation in Anxiety: An Integrated Neurobiological and Psychological Perspective. *Nature Reviews Neuroscience*, 14:488–501, 2013. doi: 10.1038/nrn3524.

Eva Hajicova. ACL Lifetime Achievement Award: Old Linguists Never Die, They Only Get Obligatorily Deleted. *Computational Linguistics*, 32(4):457–469, 2006. doi: 10.1162/coli.2006.32.4.457.

Andrew Hardie. Log Ratio - An Informal Introduction, 2014. URL `https://cass.lancs.ac.uk/log-ratio-an-informal-introduction/`. Accessed: 2024-07-05.

Manoel Horta Ribeiro, Justin Cheng, and Robert West. Automated Content Moderation Increases Adherence to Community Guidelines. In *Proceedings of the ACM Web Conference 2023*, WWW '23, page 2666–2676, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9781450394161. doi: 10.1145/3543507.3583275.

Payton J. Jones, Benjamin W. Bellet, and Richard McNally. Helping or Harming? The Effect of Trigger Warnings on Individuals With Trauma Histories. *Clinical Psychological Science*, 8:905–917, 2020. doi: 10.1177/2167702620921341.

Brendan Kennedy, Mohammad Atari, Aida Mostafazadeh Davani, Leigh Yeh, Ali Omrani, Yehsong Kim, Kristopher Coombs, Shreya Havaldar, G J Portillo-Wightman, Elaine Gonzalez, Joe Hoover, Aida Azatian, Alyzeh Hussain, Austin Lara, Gabriel Cardenas, Adam Omary, Christina Park, Xin Wang, Clarisa Wijaya, and Morteza Dehghani. Introducing the Gab Hate Corpus: Defining and Applying Hate-Based Rhetoric to Social Media Posts at Scale. *Language Resources and Evaluation*, 56:1–30, March 2022. doi: 10.1007/s10579-021-09569-x.

Deepak Kumar, Patrick Gage Kelley, Sunny Consolvo, Joshua Mason, Elie Bursztein, Zakir Durumeric, Kurt Thomas, and Michael Bailey. Designing Toxic Content Classification for a Diversity of Perspectives. In *Proceedings of the Seventeenth USENIX Conference on Usable Privacy and Security*, SOUPS'21, USA, 2021. USENIX Association. ISBN 978-1-939133-25-0.

Elisa Leonardelli, Stefano Menini, Alessio Palmero Aprosio, Marco Guerini, and Sara Tonelli. Agreeing to Disagree: Annotating Offensive Language Datasets with Annotators' Disagreement. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10528–10539, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.822. URL `https://aclanthology.org/2021.emnlp-main.822`.

Jefrey Lijffijt, Terttu Nevalainen, Tanja Säily, Panagiotis Papapetrou, Kai Puolamäki, and Heikki Mannila. Significance Testing of Word Frequencies in Corpora. *Digital Scholarship in the Humanities*, 31(2):374–397, 12 2014. ISSN 2055-7671. doi: 10.1093/llc/fqu064.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A Robustly Optimized BERT Pretraining Approach, 2019. URL `https://arxiv.org/abs/1907.11692`.

David G. Lockwood. Review: A Functional Approach to Syntax in Generative Description of Language. *Language*, 47(3):691–700, 1971. URL `http://www.jstor.org/stable/412384`.

Philip A. Luelsdorff. *The Prague School of Structural and Functional Linguistics.* John Benjamins Publishing Company, November 1994. ISBN 9781556192661. doi: 10.1075/llsee.41.

Yiwei Luo, Dallas Card, and Dan Jurafsky. Detecting Stance in Media On Global Warming. In Trevor Cohn, Yulan He, and Yang Liu, editors, *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3296–3315. Association for Computational Linguistics, November 2020. doi: 10.18653/v1/2020.findings-emnlp.296.

Charles F. Meyer. *English Corpus Linguistics*, volume 1. Cambridge University Press, 2002. ISBN 9780511606311. doi: 10.1017/CBO9780511606311.

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. SemEval-2016 Task 6: Detecting Stance in Tweets. In Steven Bethard, Marine Carpuat, Daniel Cer, David Jurgens, Preslav Nakov, and Torsten Zesch, editors, *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 31–41, San Diego, California, June 2016. Association for Computational Linguistics. doi: 10.18653/v1/S16-1003. URL `https://aclanthology.org/S16-1003`.

Youngja Park, Siddharth Patwardhan, Karthik Visweswariah, and Stephen C. Gates. 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, 2008. doi: 10.21437/INTERSPEECH.2008-537.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *J. Mach. Learn. Res.*, 21(1), January 2020. ISSN 1532-4435.

Paul Rayson and Roger Garside. Comparing corpora using frequency profiling. In *The Workshop on Comparing Corpora*, pages 1–6, Hong Kong, China, October 2000. Association for Computational Linguistics. doi: 10.3115/1117729.1117730. URL `https://aclanthology.org/W00-0901`.

Paul Rottger, Bertie Vidgen, Dirk Hovy, and Janet Pierrehumbert. Two Contrasting Data Annotation Paradigms for Subjective NLP Tasks. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz, editors, *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, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.13.

Joni Salminen, Hind Almerekhi, Ahmed Mohamed Kamel, Soon-gyo Jung, and Bernard J. Jansen. Online Hate Ratings Vary by Extremes: A Statistical Analysis. In *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval*, CHIIR '19, page 213–217, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450360258. doi: 10.1145/3295750.3298954.

Mevagh Sanson, Deryn Strange, and Maryanne Garry. Trigger Warnings Are Trivially Helpful at Reducing Negative Affect, Intrusive Thoughts, and Avoidance. *Clinical Psychological Science*, 7(4):778–793, 2019. doi: 10.1177/2167702619827018.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. Whose Opinions do Language Models Reflect? In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023. doi: 10.48550/arXiv.2303.17548.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. The Risk of Racial Bias in Hate Speech Detection. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1163.

Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A. Smith. Annotators with Attitudes: How Annotator Beliefs And Identities Bias Toxic Language Detection. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz, editors, *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, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.431.

Ferdinand Schlatt, Dieter Bettin, Matthias Hagen, Benno Stein, and Martin Potthast. Mining Health-related Cause-Effect Statements with High Precision at Large Scale. In Nicoletta Calzolari, Chu-Ren Huang, Hansaem Kim, James Pustejovsky, Leo Wanner, Key-Sun Choi, Pum-Mo Ryu, Hsin-Hsi Chen, Lucia Donatelli, Heng Ji, Sadao Kurohashi, Patrizia Paggio, Nianwen Xue, Seokhwan Kim, Younggyun Hahm, Zhong He, Tony Kyungil Lee, Enrico Santus, Francis Bond, and Seung-Hoon Na, editors, *29th International Conference on Computational Linguistics (COLING 2022)*, pages 1925–1936. International Committee on Computational Linguistics, October 2022. URL `https://aclanthology.org/2022.coling-1.167`.

Petr Sgall. Functional Sentence Perspective in a Generative Description. In Lubomir Dolezel, editor, *Prague Studies in Mathematical Linguistics*, volume 14, pages 143–172. University of Alabama Press, 1967.

Petr Sgall, Eva Hajicová, and Jarmila Panevová. *The Meaning of the Sentence in its Semantic and Pragmatic Aspects*, volume 1. Springer Dordrecht, 1986. ISBN 9789027718389.

Roni Shafir and Gal Sheppes. How Anticipatory Information Shapes Subsequent Emotion Regulation. *Emotion (Washington, D.C.)*, 20:68–74, 2020. doi: 10.1037/emo0000673.

Manuka Stratta, Julia Park, and Cooper deNicola. Automated Content Warnings for Sensitive Posts. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI EA '20, page 1–8, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450368193. doi: 10.1145/3334480.3383029.

UFAL. Functional Generative Description, 2014. URL `https://ufal.mff.cuni.cz/functional-generative-description`. Accessed: 2024-07-08.

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

Sean Wallis and Gerald Nelson. Knowledge Discovery in Grammatically Analysed Corpora. *Data Mining and Knowledge Discovery*, 5:305–335, 2001. doi: 10.1023/A:1011453128373.

Ruyuan Wan, Jaehyung Kim, and Dongyeop Kang. Everyone's Voice Matters: Quantifying Annotation Disagreement Using Demographic Information. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37 (12):14523–14530, Jun. 2023. doi: 10.1609/aaai.v37i12.26698. URL `https://ojs.aaai.org/index.php/AAAI/article/view/26698`.

Zeerak Waseem. Are You a Racist or Am I Seeing Things? Annotator Influence on Hate Speech Detection on Twitter. In David Bamman, A. Seza Doğruöz, Jacob Eisenstein, Dirk Hovy, David Jurgens, Brendan O'Connor, Alice Oh, Oren Tsur, and Svitlana Volkova, editors, *Proceedings of the First Workshop on NLP and Computational Social Science*, pages 138–142, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/W16-5618.

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

Matti Wiegmann, Jennifer Rakete, Magdalena Wolska, Benno Stein, and Martin Potthast. If there's a Trigger Warning, then where's the Trigger? Investigating Trigger Warnings at the Passage Level, 2024. URL `https://arxiv.org/abs/2404.09615`.

Wilson Wong, Wei Liu, and Mohammed Bennamoun. Determining Termhood for Learning Domain Ontologies using Domain Prevalence and Tendency. In Peter Christen, Paul J. Kennedy, Jiuyong Li, Inna Kolyshkina, and Graham J. Williams, editors, *Sixth Australasian Data Mining Conference (AusDM 2007)*, volume 70 of *CRPIT*, pages 47–54, Gold Coast, Australia, 2007. ACS. URL `https://crpit.scem.westernsydney.edu.au/abstracts/CRPITV70Wong.html`.

Wendy Wyatt. The Ethics of Trigger Warnings. *Teaching Ethics*, 16:17–35, 2016. doi: 10.5840/tej201632427.