Trigger Warning Assignment as a Multi-Label Document Classification Problem

Matti Wiegmann[1] **Magdalena Wolska**[1] **Christoper Schröder**[2]
Ole Borchardt[2] **Benno Stein**[1] **Martin Potthast**[2,3]

[1]Bauhaus-Universität Weimar [2]Leipzig University [3]ScaDS.AI

Abstract

A trigger warning is used to warn people about potentially disturbing content. We introduce trigger warning assignment as a multi-label classification task and create the Webis Trigger Warning Corpus 2022, and with it the first dataset of 1 million fanfiction works from Archive of our Own with up to 36 different warnings per document. To provide a reliable catalog of trigger warnings, we organized 41 millions of free-form tags assigned by fanfiction authors into the first comprehensive taxonomy of trigger warnings by mapping them to the 36 institutionally recommended warnings. To determine the best operationalization of trigger warnings, we explore state-of-the-art multi-label models, examining the trade-off between assigning coarse- and fine-grained warnings, open- and closed-set classification, document length, and label confidence. Our models achieve micro F_1 scores of about 0.5, which reveals the difficulty of the task. Tailored representations and long input sequences might improve and recall rare warning prediction.[1,2]

1 Introduction

Media of any kind can address topics and situations that trigger discomfort or stress in some people. To help these people decide in advance whether they want to consume such media, so-called content warnings or trigger warnings can be added to them. Trigger warnings were originally used to help patients with post-traumatic stress disorder. But after being picked up by various internet communities, the set of known trauma triggers has grown to include many more, such as abuse, aggression, discrimination, eating disorders, hate, pornography, or suicide, to also warn people tending to be "emotionally triggered" by a topic (e.g., to cry). Today, the two terms are often used interchangeably, with "trigger" referring to the semantic cause.

[1]Code: https://github.com/webis-de/ACL-23
[2]Data: https://doi.org/10.5281/zenodo.7976807

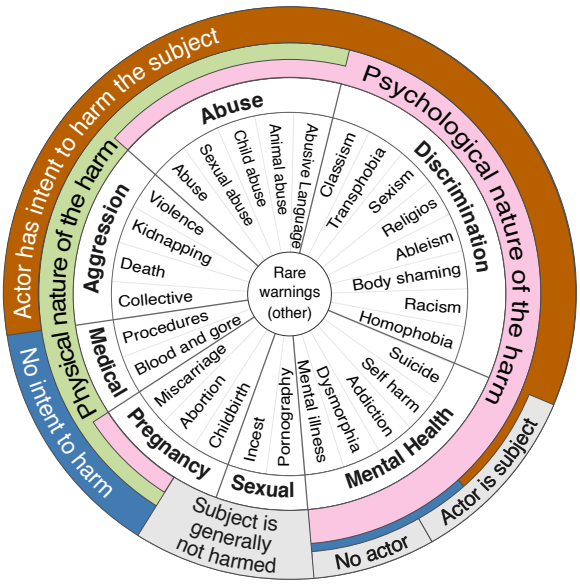

Figure 1: Taxonomy of trigger warnings. The three outer rings are alternative groupings of the inner trigger categories. The inner white ring groups 29 triggers into 7 coarse categories, the inner colored ring by relation between actor, subject, and intent, and the outer colored ring by the nature of the harm. The center represents the long tail of the rare triggers, where each of the 7 coarse categories (white ring) has its own "other" class.

Fiction in particular can make its readers susceptible to triggers. Many readers "lose themselves" in fictional works, identify with their protagonists, and experience their fate with particular intensity. This may partly explain why the community of the fanfiction website Archive of our Own (AO3)[3] is one of the few where trigger warnings are used proactively and as a matter of course: About 50% of the 7.8 million AO3 works have author-assigned warnings. The other half, however, do not, and neither the AO3 moderators nor the readership seem willing or able to fill that gap.

[3]https://archiveofourown.org, where fans write and share stories based on existing characters and worlds from popular media, such as books, movies, or video games ("fanfiction").

In this paper, we introduce the task of trigger warning assignment as multi-label document classification (MLC). Our first contribution is the Webis Trigger Warning Corpus 2022 (Webis-Trigger-22) of 8 million fanfiction works (with 58 billion words and 53 million author-assigned free-form tags; Section 3). Our second contribution is a taxonomy of trigger warnings in texts grounded in the everyday use of warnings, as shown in Figure 1 and Table 7 (Section 4). The taxonomy combines two authoritative sources: Its two top tiers unify eight official lists of trigger warnings from as many institutions, synthesizing them into a hierarchy of 36 semantic categories. To ground the taxonomy in the real-world assignment of warning, we mapped 41 million of the 53 million free-form tags to these 36 categories. From the corpus, we compiled a dataset of 1 million documents from the corpus, densely populated with taxonomy labels (Section 5). As our third contribution, we study the warning assignment effectiveness of an SVM, XGBoost, RoBERTa, and a Longformer, depending on category granularity, open- vs. closed-set classification, document length, and label confidence (Section 6), as a first baseline.

We see low recall (false negatives cause more harm than false positives), low effectiveness for rare categories (especially for *Discriminates*), and representing very long documents as key challenges. Assigning the fine-grained warning categories of the taxonomy's second tier is more desirable but also more difficult than assigning the coarse-grained categories of its first tier, so the key to improving future approaches may lie in specializing in particular categories (Section 6.2).

2 Related Work

Pioneering work on automatic trigger warning assignment is Stratta et al.'s (2020) user study with a browser plugin (DeText) on generic websites. The authors conclude that client-side warnings are feasible and that users respond positively. However, this work is very limited in that *Sexual assault* is the only warning given using a naive dictionary-based approach. In Wolska et al. (2022), we conduct a pilot study on binary document classification for the *Violence* trigger category. This study includes only works labeled with one of the three predefined AO3 warnings (i.e., *Graphic violence*), ignoring the millions free-form tags. No other works have addressed trigger warning assignment until now.

Charles et al. (2022) recently proposed the Narrative Experiences Online (NEON) taxonomy of multi-media trigger warnings. Its two tiers are synthesized like in our first tier from 136 guidelines on the web, consisting of 14 top tier categories (versus our 7) and 76 subcategories (versus our 36). However, unlike ours, NEON's subcategories are not explicitly grounded in warnings that are used on a daily basis by millions of people. Moreover, its categories are non-disjoint, not clearly semantically motivated classes with blurred definitions: For instance, compare category "*4. Disturbing content: Content contains imagery, sounds, or effects that may frighten, disgust or scare*" with category "*9. Parental guidance: Content may not be appropriate for children*". Since our two teams worked in parallel, the synthesis of our complementary taxonomies is a fruitful direction for future work.

Trigger warnings can be seen as orthogonal to other harmful content taxonomies, e.g., for violence, hate speech, or toxicity, where some labels overlap but differ in structure and entailment. Mollas et al. (2020) also study the detection of violence and present the ETHOS dataset of YouTube and Reddit comments with crowdsourced multi-label annotations about verbal violence and its target. Banko et al. (2020) presents a comprehensive taxonomy of harmful online content that has notable overlap with our taxonomy but focuses on online speech. Triggering content, however, can be narrative and does not require an intent to harm to evoke disturbing images. Based on Wulczyn et al.'s (2017) work, the Toxic Comment Classification Challenge (Adams et al., 2017) dataset covers different content moderation topics. It contains 223,000 Wikipedia comments (sentence to paragraph level) annotated with six toxicity subtypes.

Our multi-label classification (MLC) task has (comparably) few labels overall, few labels per document, but features long documents. The main difference to other MLC datasets is the document genre (fanfiction) and the label domain (trigger warnings). The most similar MLC datasets (with mostly shorter documents) are Reuters RCV1 (Lewis et al., 2004) with 80,000 news articles and 103 topic labels, its predecessor Reuters-21578 with 11,000 news articles and 90 labels, and the Arxiv Academic Paper Dataset (AAPD) (Yang et al., 2018) with 56,000 abstracts from computer science and 54 labels. Recent meta-studies on long document classification (Park et al., 2022; Dai et al.,

Corpus size	
Words	58B
Total works	7.9M
- w/ closed-set warnings	2.8M
- w/ open-set warnings	281K
- w/o warnings	4.7M

Filter criteria for Section 5	
Non-English language	751K
Publication pre-2009	246K

Filter criteria for Section 5	
More than 100 chapters	3K
More than 93k words (top 1%)	79K
Less than 50 words (bott. 1%)	122K
More than 66 tags (top 1%)	8K
More than 10% unclean tags	4.7M
Less than 3 tags (conf. thresh.)	2.3M
Less than 5 kudos (popularity	632K
Less than 100 hits threshold)	751K
Duplicates	8K

Table 1: Selection of corpus statistics of the Webis Trigger Warning Corpus 2022. See Appendix C for details.

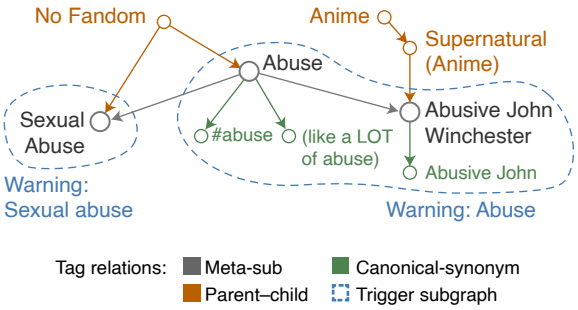

Figure 2: Excerpt of AO3's tag graph. The edges are the three relations added by tag wranglers, connecting three subgraphs of free-form tags to gray canonical tags.

2022) find that sparse-attention transformers, hierarchical models, and input selection methods have little difference in effectiveness to input truncation. Galke and Scherp (2022) compare graph and "bag of words" (BoW) methods with transformers, noting that BoW methods are (often) not far behind.

Further MLC datasets cover tasks with very large label sets: EUR-Lex (Mencía and Fürnkranz, 2008) with 15,000 law documents and 4,000 labels, its successor EURLEX57K (Chalkidis et al., 2019) with 57,000 law documents and 4,300 EUROVOC labels, MIMIC-III (Johnson et al., 2016) with 112,000 clinical reports and 11,600 ICD-9 codes as labels, and the Extreme Labels (Bhatia et al., 2016) collection of datasets for product and Wikipedia article classification. Recent work on large label set MLC addresses label-dependent document representations (Xiao et al., 2019), loss functions for long-tailed label distributions (Huang et al., 2021), prompt-based few-shot learning for rare labels (Yang et al., 2022), and sequence labeling with an attention encoder–decoder LSTM for many-label document MLC (Yang et al., 2018). Transformer–encoder classifiers are common baselines (Chalkidis et al., 2019).

3 The Webis Trigger Warning Corpus 2022

Our inspiration for operationalizing trigger warnings is based on finding "hidden in plain sight" a large collection of fictional works with millions of manually assigned warnings that have accumulated for years on the widely known fanfiction website Archive of our Own (AO3), which to our knowledge have not previously been used as a basis for automating a task. We therefore first compiled a near-complete corpus of AO3 fanfiction (i.e., fanfics, documents) and its metadata, namely language, length, comments, hits (i.e., reads), kudos (likes), (chapter) publication date(s), and, notably the *Additional* (free-form) *tags*: The Webis Trigger Warning Corpus 2022 (Webis-Trigger-22 for short) contains about 8 million works totaling 58 billion words. Table 1 shows selected corpus statistics.

The corpus also reconstructs the tag graph induced by the author-assigned free-form tags. Illustrated in Figure 2, the tag graph defines three relations between tags: canonical-synonym, parent-child (i.e., fandom and media-type relations), and meta-sub relations which form a hierarchy of meanings. All relations form acyclic digraphs where canonical tags from a controlled subset connect the free-form tag subgraphs. Tag relations are manually created and maintained by volunteer community experts (the so-called "tag wranglers") following specific guidelines (The Organization for Transformative Works, 2023). We consider them a highly reliable basis for our subsequent distant-supervision annotation of trigger warnings.

Scraping strategy We scraped all public works from AO3 using their unique URLs which are based on a work's permanent and unique ID. First, we systematically enumerated all publicly available work IDs between August 13, 2008, and August 09, 2021. The most active day yielded about 10,000 works. We then archived the web pages' HTML in WARC files using ChatNoir Resiliparse (Bevendorff et al., 2018). Last, the pages' HTML was parsed using Scrapy,[4] extracting each work's text as a list of chapters along with their metadata.

In addition to the works, we also scraped the relevant section of the tag graph by archiving and parsing the web page of each tag that was used in one of the works. A tag's page lists all relations of that tag so that the relevant section of the tag graph can be reconstructed from our scrape.

[4] https://scrapy.org

Deduplication We removed 8,011 full and near duplicates from the crawl. The 4,249 full duplicates were identified using SHA-256 fingerprinting. Near-duplicates include pairs of works whose text differs only to a very small extent so that neither the meaning and especially not the relevant warning labels change. We identified them by applying MinHash (Shrivastava and Li, 2014) with 8 buckets and considered resulting pairs as near-duplicates if their Jaccard similarity exceeded 0.6 or if their cosine similarity exceeded 0.875. This approach favors precision over recall and ultimately identified 3,762 near-duplicates.

4 A Taxonomy of Trigger Warnings

A manual examination of a sample of the free-form tags on AO3 showed that a considerable fraction are trigger warnings. Authors often append qualifiers to their warnings, which may indicate the nature of a trigger or its connection to the narrative of their work. These tags are manually associated with a controlled subset chosen by the tag wranglers. However, many canonical tags exhibit semantic redundancy, while the subset is too extensive and too sparsely populated with works for operationalization. We therefore first synthesized an authoritative hierarchy of 36 trigger labels based on guidelines from relevant institutions and then embarked on a semi-automatic mapping of the myriad of free-form tags to this condensed set. The outcome is a two-tier taxonomy, which firmly grounded in real-world trigger warning assignment.

4.1 Curating the Trigger Warning Repertoire

While the notion of "content warning" in digital media has been around for a decade, none but one recent attempt has been made to propose a "standardized set" (Charles et al., 2022) due to the open-ended nature of the issue. Most warning labels stem from internet communities, such as social media, gaming, and online-content readers and writers. Not surprisingly, such *community-supplied* labels have all the properties of user-generated content, in particular, heterogeneity and lack of linguistic uniformity, which makes them hardly usable as a set of classes for training classifiers. However, since the arousal of a debate on the use of content warnings in educational settings, many universities issued explicit guidelines on their use. We take eight such *institutionally-recommended* guidelines and frequently referenced lists of warnings as authoritative trigger warning sources and consolidate their label sets in a principled way.

Figure 1 shows the resulting 36-label taxonomy (`TW-36`), consisting of 29 narrowly-defined (closed-set) categories for frequent warnings and 7 more general, higher-level (open-set) labels. The 29 closed-set labels have clear semantics, which is advantageous for classification and practical from the point of view of usability. The 7 open-set labels also match documents that are related to but do not match any of the closed-set labels. This open-set semantics is essential for trigger warnings since traumatic imagery can be evoked by a variety of individually-rare topics (hence the large dimensionality of user-generated warnings). The 7 open-set labels, e.g. *Sexual*, constitute a level of abstraction for the closed-set labels, e.g. *Incest* and *Pornography*; a coarse variant of the label set.

Sources of trigger warnings We collected guideline documents on trigger warning assignments from eight universities from the English-speaking world: Cambridge, Manchester, Michigan, Nottingham, Reading, Stanford, Toronto, and York. Table 7 (Appendix A) illustrates the guidelines, processing, and references. We identified these documents by, first, compiling a list of the top 30 universities according to Times Higher Education (THE, 2023), QS World University Rankings (2023), and the Russel Group (2023) members and, second, searched those universities' domains for combinations of 'trigger', '(sensitive) content', 'warning', 'guide', and 'recommendation'.

The structured set of warning labels Since all guidelines follow a different structure (from paragraphs to term lists) and granularity, we manually processed the documents to (i) extract and segment the warnings, (ii) align and merge warnings that are closely synonymous (e.g., *Transphobia* with *Transphobia and trans misogyny*) across documents to create the 29 closed-set labels, and (iii) group related warnings to form the 7 open-set label groups.

We extracted two units: triggering content concepts and concept groups. Concepts are all terms (*homophobia*) or phrases (*death or dying*) that refer to a singular semantic field. Concept groups are (structural) groupings of related concepts with a dedicated group name (*Discrimination (sexism, racism, homophobia, transphobia)*, where *Discrimination* is the group name). We extracted concepts from the groups and added them to the list of all

Sample	Prec	Rec	F_1	Acc
Fine-grained				
0-2k	0.94	0.94	0.94	0.94
10-11k	0.96	0.96	0.96	0.96
Coarse-grained				
0-2k	0.95	0.95	0.95	0.95
10-11k	0.96	0.96	0.96	0.96

Verbatim warnings	Tag occur.	Unique tags
Total	62,316	27,694
Classified as warning	34,806	9,595
- of all wrangled	0.86	0.79
- of all free-form	0.56	0.35

Table 2: Effectiveness of the distantly supervised classification on two manually annotated tag sets (left). Number of verbatim warnings (e.g., *warning, tw:, ...*) annotated as a warning by our method (right).

Sample	Nr. tags in set (% of all)		Warnings (% of set)			
	Tag occurrence	Unique tags	Closed		Open	
0-2k	27.6M (51.98)	2K (0.02)	538	(26.71)	82	(4.07)
10-11k	0.3M (0.56)	1K (0.01)	127	(12.70)	19	(1.90)
Tag graph	41.0M (77.18)	2M (20.17)	241K	(12.30)	33K	(1.68)
All tags	53.1M	9.7M	–		–	

Table 3: Number of AO3 free-form tags that can be annotated with a trigger warning by different methods. The samples 0-2k and 10-11k contain manually and Tag graph distantly supervised annotations.

concepts. Items of structured lists (same bullet point) or concepts in coordinating conjunctions were not segmented, assuming they belong to the semantic field that defines the warning.

We generally grouped concepts that were mentioned together in a concept group and used this group's name to determine the open-set label. Concepts were split if a term in a concept did not match the group's intention, e.g. *body-shaming* was split from *Eating disorders and body shaming* and grouped with *discrimination*. We created the *sexual* and *childbirth* groups and then assigned the remaining concepts to the most closely related group. Since we were looking for labels with support ("consensus") across different sources, we ignored concepts with singular occurrences.

Properties of the warning labels Four major observations can be made: First, the granularity of triggers is not uniform (e.g., both *abuse* and the more specific *child abuse* are included). Second, the set comprises subsets of related concepts which lend themselves to semantic abstraction (e.g., *sexism*, *classism* and other *-isms* and *-phobias*). Third, the guidelines are not exhaustive (as they point out themselves) due to the open-set nature of traumatic events and triggering imagery. For this reason, we consider the 7 (coarse-grained) categories as a part of the whole set (instead of just a hierarchy tier): they add the needed open-set semantics (e.g., *bullying* is discrimination but would not be covered by the closed-set categories). Fourth, the (lexical) semantic field of the labels is not precise enough to be the sole base for document annotation. We developed sharper definitions based on the annotation procedure in Section 4.2, which are shown in Table 6 (Appendix A). Figure 1 also shows an additional abstraction of the label definitions in two dimensions: the nature of the harm done in the content (physical/psychological) and the interaction between the actor and the subject.

4.2 Taxonomizing the Free-form Tags

We taxonomize all works with free-form warnings by mapping each tags to every semantically matching warning category from our taxonomy. A tag is discarded if no such mapping is possible. The resulting mapping table of free-form tags to trigger warning categories was created by (i) manually annotating the 2,000 most common tags, (ii) effectively identifying substructures of the tag graph that imply a trigger warning so that each of their nodes is automatically mapped to that trigger (distant supervision), and (iii) merging both results, giving priority to manual annotations.

Manual annotation We manually annotated two samples of free-form tags: first, the 2,000 most frequent tags (0-2k), which cover just over 50% of tag occurrences, and second, the 10,000-11,000 most frequent tags (10-11k), which are reasonably common and used to evaluate our distant supervision approach. The annotation process had three stages: (i) Two annotators individually annotated each tag by assigning it a trigger warning from our taxonomy. (ii) Both annotators discussed and resolved every disagreement and updated the annotation guide. (iii) Annotator 1 annotated each tag again using the updated guide. The first annotated sample 0-2k contains 538 tags annotated with one of the 29 closed-set trigger warnings and another 82 open-set ('other') warnings. The ratio of tag-to-warning assignments reduces by about half for less frequent tags and stabilizes at 9–16%. The resulting label definitions and example tags for each label can be found in Table 6 (Appendix A). A summary of our annotation guide can be found in Appendix A.1.

Distant-supervision annotation We annotated 2.0 million unique free-form via distant supervision by identifying rooted subgraphs (trigger graphs) in the tag graph. All tags in a trigger graph indicate a related concept that warrants the same trigger warn-

ing as the only source node (i.e., its root; see Figure 2, *Abuse*). The sources were annotated manually and the respective warnings were also assigned to all successors of the source. Trigger graphs were identified with a 5-stage process: (i) Grouping of all tags via the synonym relation and identification of the canonical tag. One tag per synonym set is marked as canonical by wranglers, all other synonyms are direct successors of the canonical tag and have no other outgoing edges. (ii) Identification of meta-sources: canonical tags that are source nodes in the meta-sub graph. Meta-sub relations indicate a directed lexical entailment between canonical tags and have a typical depth of 2–4. (iii) Identification of candidate sources of trigger graphs: meta-sources that are also direct successors of the *No Fandom* node in the parent–child graph. Sinks in this graph are canonical tags and all predecessors are either a fandom, media type, or *No Fandom*. The latter is added as a parent to tags that apply to many fandoms, including content warnings but also, for example, holidays and languages. This yields about 5,000 candidate sources. (iv) Identification of trigger graph sources: manual annotation of all candidate sources, discarding the nodes without a trigger warning. (v) Identification of all trigger graphs: manual traversal of the tag graph (depth-first) along the meta-sub relation, starting from each trigger graph source. If a successor does not match the trigger warning assigned to its predecessor, the connecting edge is removed, the successor added as a new trigger graph source, and annotated with a new trigger warning.

Annotation evaluation First, we evaluate how effectively our distant-supervision approach annotates the free-form tags by comparing the inferred annotations with the two manually annotated tag sets `0-2k` and `10-11k` across the four different trigger warning sets. As shown in Table 2 (left), our approach scores well above 0.9 in accuracy and weighted average F_1. There is little difference between evaluating the fine-grained labels and their coarse-grained equivalent.

Second, we evaluate how complete the set of all free-form tags can be annotated by our method. As shown in Table 3, due to the long-tailed distribution of the *additional tags*, we can annotate 52% of all occurrences manually with high reliability and another 25% with an accuracy of ca. 0.95. Our method can completely annotate all tags of more than half of all works in the corpus. The other half of the works are only partially annotated since our method only annotates ca. 20% of the unique tags. Tags are only wrangled (i.e. added to the tag graph) if they occur thrice. Our method will miss the 89.9% single occurrence unique *additional tags*.

Dataset Properties	
Mean no. words	8K
Median no. words	3K
90pct no. words	21K
Mean no. chapters	3.0
Median no. chapters	1
Fine warnings	2.1M
Coarse warnings	1.7M

Dataset Properties	
Works w/ < 512 words	56K
Works w/ < 4,096 words	645K
Works w/ only closed warnings	728K
Works w/ only open warnings	94K
Works w/ open and closed warn.	271K
Total Works	1.1M

Table 4: Descriptive statistics of the compiled dataset.

Third, we evaluate how many free-form tags that contain a verbatim *warning* are annotated with a warning from our taxonomy. Table 2 (right) shows that about 80% of verbatim warnings (that are part of the tag graph and can hence be annotated by our method) are also annotated with a taxonomy category. The other 20% are almost exclusively warnings that do not match any category, such as *Politics, Fluff, Police, ...*. This ratio is lower for rare free-form tags which are not wrangled and thus not part of the tag graph. A verbatim tag contains one of the tokens *tw(:)*, *cw(:)*, or *trigger(s)*.

5 Sampling the Evaluation Dataset

As a basis for the computational study of trigger warning assignment and our evaluation, we sampled a densely-annotated (excluding works without labels) dataset with 1,092,322 works from the previously constructed corpus. The sampling has two stages: First, filtering works from the corpus that match reliability criteria (see Table 1). Second, creating stratified standard splits (that preserve label balance). Table 4 shows descriptive statistics and the data statement in Appendix C provides details.

5.1 Sampling Method

The first sampling stage excludes about 4.7M works without trigger warnings as well as all (i) non-English works (although a multi-lingual dataset would be feasible in a few-shot scenario); (ii) works published before AO3's release in 2009 (they have mostly been migrated from other archives and we consider their tags as unreliable); (iii) atypically-sized works and outliers, which include works with more than 100 chapters, more than 93,000 words (the top percentile), less than 50 words (which are usually placeholders for links

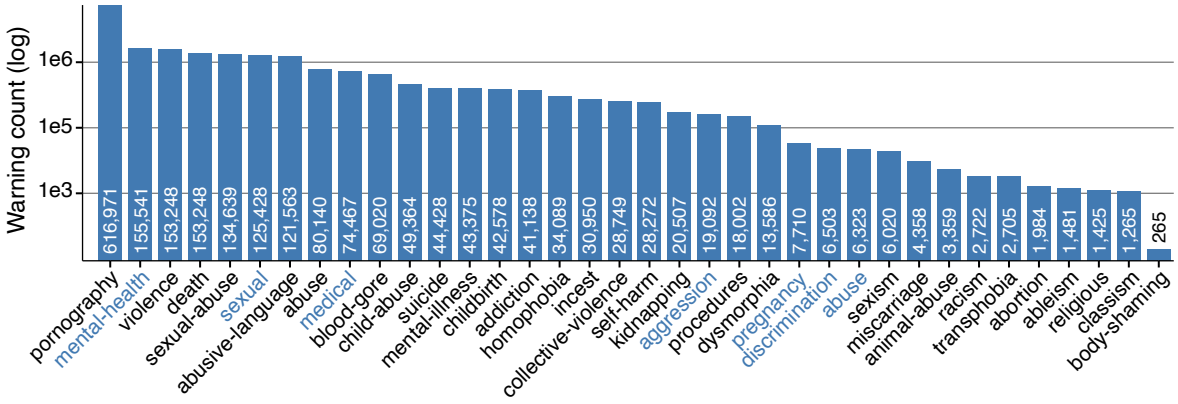

Figure 3: Distribution of the fine-grained warnings over works in the dataset. Open-set warnings are highlighted.

or non-text media), and more than 66 tags (the top percentile); (iv) works with less than 3 tags (warnings are atypically uncommon within these works and we aimed at reducing label noise); (v) unpopular works with less than 5 kudos (i.e., likes) and less than 100 hits (i.e., reads), which are usually low-quality writing; and (vi) works with less than 90% annotated tags. The last criterion filters works whose tags could not be annotated with our methods (i.e., we do not know if the tags indicate a trigger warning, which risks false negatives). However, we allow 10% of the tags to be non-annotated, since the number of works with rare warnings almost doubles while adding only about 70,000 works overall.

In the second stage, we created a standard split of 90:5:5 (training, validation, test). The balance of warning labels was preserved by iterating works with certain warnings from the least to the most common, adding a random work into either test or validation until they contained the targeted number of works with that label and then adding the remaining works into the training set.

5.2 Properties of the Datasets

We analyze five properties of the dataset to characterize trigger warnings in fanfiction and as foundation for the evaluation.

Warning label distribution Figure 3 shows that warnings follow a long-tailed distribution, which is common in multi-label settings: *Pornography* warnings are extremely common since sexual exploration is a relevant part of fanfiction. The open-set *Mental-health* warning is also common since it collects topics of strong anxiety and depression. Conversely, *Discrimination-related* warnings are rare. The number of works with rare labels is sufficient to train standard classification models.

Document length Table 4 and Figure 4 show most works to be short (median about 3,000 words) and that longer works are often split into short chapters (90th percentile chapter length about 5,000 words). This exceeds BERT's input length (512 tokens) but comes close to that of a small Longformer (4,096 tokens). The label distribution is largely robust across document length, except for short documents which cover more *Sexual* content.

Warnings per work Figure 5 shows an exponential decay of documents over number of warnings. A single warning is assigned to about half the works, while more than 10,000 have five or more labels, even in the coarse-grained 7 labels setting.

Support per warning Figures 6 and 7 (Appendix D) show that most warnings have a median support of one free-form tag (mean 1.2–1.5). Most labels have rarely more than one, except for *Incest, Childbirth*, *Sexual-abuse*, and *Mental-health*. Again, *Pornography* is an outlier with a median of three and mean of four supporting tags. Authors tag sexual practices, kinks, and toys in great detail.

Co-occurrences between warnings Figure 8 (Appendix D) shows that warning co-occurrences are common with frequent tags, so that most labels co-occur with *Pornography* 20–40% of the time and 10-30% with *Violence*, *Mental-health*, *Abuse*, and *Death*. Furthermore, labels from the same group tend to weakly (about 10%) co-occur more with each other (especially in *Medical* and *Pregnancy*). Besides, some labels co-occur more frequently: *Pregnancy*, *Sexual-abuse*, and *Sexism* co-occurs with *Pornography* about 60% of the time. *Religious* co-occurs with *Racism* about 30% of the time, as does each, *Body-shaming* and *Transphobia*, with *Dysmorphia* since the latter includes eating disorders and (gender) dysphoria.

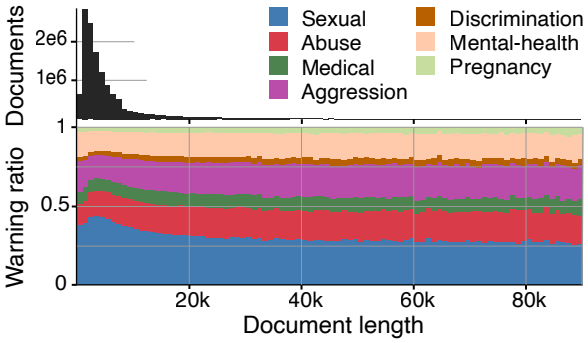

Figure 4: Distribution of document length in the dataset (top, log-scale) and distribution of all coarse-grained warnings split by text length (bottom).

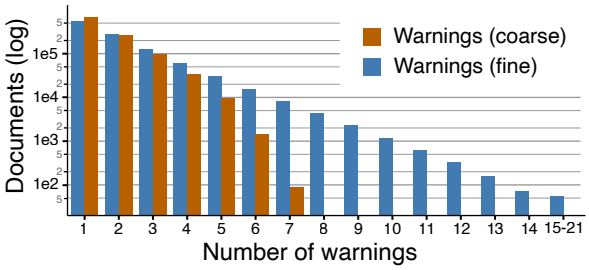

Figure 5: Distribution of the number of documents that have a certain number of fine- and coarse-grained warning labels assigned. Document count is log-scaled.

6 Experimental Evaluation

To rate the impact of label granularity, open-endedness, document length, and support on trigger assignment models, we compared the performance of 4 models over the dataset described in Section 5.

6.1 Models

For the experimental evaluation, we selected four models based on use in recent comparative studies on long-document classification (Dai et al., 2022; Park et al., 2022; Galke and Scherp, 2022): SVM, XGBoost (Chen and Guestrin, 2016) (XGB), RoBERTa (Liu et al., 2019) (BERT), and Longformer (Beltagy et al., 2020) (LF). We trained each model once on the 36-label fine-grained warning set and once on the 7-label coarse-grained label set with identical input documents.

The SVM is a well-established traditional baseline in text classification (Joachims, 1998) which is computationally cheap and serves as a good point of reference. XGBoost, as opposed to the linear SVM, expresses non-linear partitioning of the feature space. Engineered feature spaces are (still) competitive in long-document classification since positional information is less significant than the input size limitation of transformer models. According to the experiments of Dai et al. (2022) and Park et al. (2022), RoBERTa and Longformer with truncation are about as efficient as SotA models.

Model configuration For SVM, we used a linear SVM in one-vs-rest mode from scikit-learn (Pedregosa et al., 2011). As features, we used TF-IDF document vectors of the word 1–3-grams with a minimum document frequency of 5, tokenized by the `bert-base-uncased` tokenizer from Huggingface. For XGB, we used the histogram-optimized tree construction from the XGBoost library (Chen and Guestrin, 2016) with the same features as for SVM. For BERT, we used the `roberta-base` checkpoint from the Hugging Face Model Hub with input padding and truncation to 512 tokens. For LF, we used the `allenai/longformer-base-4096` checkpoint from the Hugging Face Model Hub with input padding and truncation to 4,096 tokens. All models used the same preprocessing: Removing all HTML formatting, removing all non-alphanumeric symbols except `.,!?"'`, and lowercasing.

Appendix B shows detailed experimental settings; Table 8 shows the final model configuration.

6.2 Results

Table 5 shows the total effectiveness (micro- and macro-averaged) of the 4 models, trained once for a 36-label and once for a 7-label problem. The best model has a micro F_1 of 0.52 on the fine-grained dataset, which is lower than the scores of comparable datasets reported on `paperswithcode.com`: 0.91 (Huang et al., 2021) on Reuters-21578 and 72.8 (Liu et al., 2020) on AAPD.

The overall most effective model is XGB with 0.30 macro and 0.52 micro F_1 on the fine-grained label set, followed by SVM and BERT. Precision is generally higher than recall by ca. 0.2–0.3. Micro-averaged scores are higher than macro-averaged scores by ca 0.2 (fine-grained), which is not uncommon for strong label imbalance. The label-wise analysis (cf. Table 9, Appendix D) shows that the models are most effective on the very common warnings (ca. 0.88 on *Pornography*) and least effective on the rare warnings (0.0–0.2). These rare warnings are often *Discrimination*-related. XGB is often more effective for rare labels than the others (ca. +0.25 on *Abortion* and *Transphobia*). BERT is more effective on 7 of the more frequent labels but is ca. 0.1 less effective on the most frequent labels

Set	Fine (36 labels)						Coarse (7 labels)					
	Macro-avg.			Micro-avg.			Macro-avg.			Micro-avg.		
	Prec	Rec	F_1	Prec	Rec	F_1	Prec	Rec	F_1	Prec	Rec	F_1
SVM	**0.47**	0.18	0.25	**0.75**	0.37	0.49	0.59	**0.54**	**0.56**	0.71	**0.61**	**0.66**
XGB	0.44	**0.25**	**0.30**	0.72	**0.40**	**0.52**	**0.65**	0.51	**0.56**	**0.77**	0.58	**0.66**
BERT	0.36	0.19	0.23	0.56	0.37	0.45	0.45	0.52	0.46	0.53	0.54	0.53
LF	0.26	0.23	0.21	0.45	0.30	0.36	0.44	0.48	0.43	0.50	0.47	0.49

Table 5: Results of SVM, XGBoost (XGB), RoBERTa (BERT), and Longformer (LF) on the test dataset.

(*Pornography, Violence, Mental-health*), resulting in reduced total effectiveness. LF failed to generalize to the test data and is weaker than BERT; on the validation data, LF outperforms BERT by ca. 0.1.

Granularity Table 5 also shows the difference between predicting coarse (7) and fine-granular (36) labels. The models are consistently more effective on the coarse-grained label set: recall is higher by ca. 0.2–0.3 and precision by up to 0.2. The macro-average effectiveness improves more than the micro-average since coarse labels are more frequent and the rare *Discrimination* labels are combined, which reduces their impact on the average. Consequently, the difference between the macro and micro-average is also lower (from ca. 0.2 to 0.1). The difference between precision and recall is also lower (from ca. 0.25 to 0.1) since recall improves more than precision. The micro-averaged precision is largely independent of granularity.

Open-endedness Table 10 (Appendix D) shows the average effectiveness of the open and closed-set (fine-grained) warnings. The difference in macro F_1 is negligible, however, the closed-set labels are more effective by 0.1–0.3 in micro F_1 since it is strongly affected by the high scores of *Pornography*. Table 9 (Appendix D) shows no notable difference between open and closed-set labels.

Document Length Table 10 (Appendix D) also shows assignment effectiveness depending on the works' length. The neural models are more effective for works that are shorter than their input length limit. BERT is the most effective model on works with less than 512 tokens by 0.1 macro and 0.2 micro F_1 over XGB. However, BERT becomes less effective the longer the documents are (XGB is more effective by 0.15 for works with more than 16,000 tokens). Longformer behaves the same.

Support Table 10 (Appendix D) also shows the effectiveness on works that have at least two free-form tags supporting each annotated warning label. The support has no impact on macro F_1 but the micro F_1 is higher for the set of works with a minimum support of 2, most likely because *Pornography* is very often supported by multiple free-form tags, hence impacts the micro-average strongly.

6.3 Discussion

We make five key observations from the results: First, there is no notable difference in effectiveness between labels with open and closed-set semantics, which speaks for the inclusion of open-set warnings in the future. Second, learning and predicting from the full text (as opposed to truncation) is essential and more important for trigger warnings than for other MLC datasets. Third, recall is (substantially) worse than precision, which is a key issue. Trigger warning assignment is a high-recall task since false negatives (missed warnings) cause more harm than false positives (superfluous warnings). Fourth, the poor performance on rare labels, common for MLC problems, is another key issue. Fifth, models are more effective on coarse-grained labels. However, predicting fine-grained labels with high reliability can greatly reduce the number of documents that a reader may want or need to skip to be safe. Future work should focus on improving the fine-grained prediction performance.

7 Conclusion

In this paper, we model the problem of automatically assigning trigger warnings to documents as a multi-label classification task. With the Webis Trigger Warning Corpus 2022, we contribute a rich novel resource for this task by scraping 7.9 million fanfiction documents from Archive of our Own. We devised a new taxonomy of trigger warnings from eight authoritative sources and condensed them into a warning set that incorporates two tiers of granularity and the open-set semantic of trigger warnings ("everything can be a trigger") while being sufficiently structured for text classification. Furthermore, the majority the millions of author-assigned free-form tags have been heuristically mapped into the taxonomy, thus grounding our taxonomy in into the real-world assignment of trigger warnings. From the corpus, we sampled 1 million works and explored the assignment effectiveness of four baseline algorithms (SVM, XGBoost, RoBERTa, and Longformer) depending on open-endedness, granularity, document length, and label support, revealing a lot of room for future improvement.

Limitations

It should be noted that our contributions are limited to fanfiction documents. Models trained on our datasets might not transfer to other online content like news articles, websites, or social media posts. Particularly social-media texts are shorter and contain fewer descriptions and more verbal expressions, which is a substantial-enough shift to warrant models explicitly trained in the genre. Similarly, the conclusions of our experiments are limited by the models we used, as well as the genre of the text. Furthermore, the trigger warning scheme we used is a simple structure. Further research should investigate more detailed trigger (warning) typologies with a more rich semantics.

Impact Statement

We hypothesize that an automatic assignment of trigger warnings can help reduce the impact of distressing content on vulnerable groups. They would solve the problem that most social media providers are unwilling[5] or unable to integrate trigger warnings into their platforms, as users could have them automatically assigned by their respective devices before they see disturbing content.

Another potential positive impact of analyzing trigger warnings, such as those voluntarily used by social media users, is that this data can partially if not completely relieve the burden on human content moderators who are otherwise constantly confronted with extreme content. This is especially relevant to the recent news that OpenAI has outsourced content moderation for ChatGPT's output to Kenyan workers.[6] This news follows earlier reports that major social media platforms have done or are still doing the same thing to Filipino workers.[7] Any technology that helps make this type of manual moderation obsolete is very welcome. The labels obtained from manual moderation by these workers will of course be used by OpenAI and the social media providers to develop specific moderation models for ChatGPT or their platforms. We are not currently in a position to analyze whether a domain transfer from fanfiction to these moderation tasks is possible, nor do we know whether web data labeled with trigger warnings are already being used for these purposes in the aforementioned companies, but found insufficient for their purposes. Nor are fanfiction sites likely to cover all aspects of distressing content generated by large language models or found on social media. Nor does any of this absolve companies of their currently largely neglected duty to take responsibility for the welfare of their (external) workers.

As a side note to the ongoing discussion about whether trigger warnings are useful to the warned social media user or not, the above example shows that frequent exposure, even to text-only distressing content by ChatGPT, seems to have a trigger effect on the workers. However, this does not allow any conclusions to be drawn about comparably infrequent exposures that social media users may expect. Of course, moderation workers for distressing content may not benefit from trigger warnings, as they are hired to rate that content at scale.

Regarding potential negative impacts of this work, first, the presented data contains annotated, potentially distressing content, like violence or rape, in sufficient quantities to train generative models. This calls for taking measures to ensure one's personal health of body and mind when conducting manual data analyses with a focus on such distressing content, as exemplified by the above moderation example. Second, some content on AO3 might border on legality in some countries, and dependent on who owns it for what purposes, in particular regarding descriptions of underage sexuality and pedophilia, where what is considered underage differs from country to country. Some works might have meanwhile been removed from the platform but are still included in our dataset. As a precaution, we do not release the works' text in our datasets. Instead, we release only work IDs and utilities to scrape the text from AO3. We further maintain an archived version for reproducibility and ongoing research. Third, some of the stories are written about real, living humans and may include details about them. Additionally, some stories might contain information about the author. Since we did not sanitize the dataset, we will not share the trained models and only provide the code to train them as artifacts. Lastly, we used the data only partially compliant with its intended use: The AO3 tags are intended as trigger warnings, and the fanfiction stories are intended to be read.

[5] E.g., for fear of possible backlash from the community: https://www.lbc.co.uk/news/universities-backlash-trigger-warnings-on-english-literature-texts/

[6] https://time.com/6247678/openai-chatgpt-kenya-workers/

[7] https://www.washingtonpost.com/technology/2019/07/25/social-media-companies-are-outsourcing-their-dirty-work-philippines-generation-workers-is-paying-price/

References

C. J. Adams, Jeffrey Sorensen, Julia Elliott, Lucas Dixon, Mark McDonald, nithum, and Will Cukierski. 2017. Toxic comment classification challenge.

Archive of Our Own. AO3 Census Masterpost [online]. 2013. https://archiveofourown.org/works/17019228. Last accessed: October 10, 2022.

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

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *CoRR*, abs/2004.05150.

Emily M. Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Janek Bevendorff, Benno Stein, Matthias Hagen, and Martin Potthast. 2018. Elastic ChatNoir: Search Engine for the ClueWeb and the Common Crawl. In *Advances in Information Retrieval. 40th European Conference on IR Research (ECIR 2018)*, Lecture Notes in Computer Science, Berlin Heidelberg New York. Springer.

K. Bhatia, K. Dahiya, H. Jain, P. Kar, A. Mittal, Y. Prabhu, and M. Varma. 2016. The extreme classification repository: Multi-label datasets and code.

Cambridge Centre for Teaching and Learning CCTL. 2023. When to use content notes. https://www.cctl.cam.ac.uk/content-notes/how-use/when-use. Last accessed: May 10, 2023.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, and Ion Androutsopoulos. 2019. Large-scale multi-label text classification on EU legislation. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy*, pages 6314–6322. Association for Computational Linguistics.

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

Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, New York, NY, USA. ACM.

Claire Childs, Department of Language, and Linguistic Science. 2021. LLS Departmental Guidance on Content Warnings. https://www.york.ac.uk/media/abouttheuniversity/equality/documents/LLS-Departmental-Guidance-on-Content-Warnings-2021.pdf. Last accessed: May 10, 2023.

University of Toronto, Centre for Teaching and Learning CTL. 2021. Teaching Sensitive Materials. https://hive.utsc.utoronto.ca/public/dean/academic%20administrators/DCD%202021-22/Teaching%20Sensitive%20Materials.pdf. Last accessed: May 10, 2023.

Xiang Dai, Ilias Chalkidis, Sune Darkner, and Desmond Elliott. 2022. Revisiting transformer-based models for long document classification. In *Findings of the Association for Computational Linguistics: EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 7212–7230. Association for Computational Linguistics.

Lukas Galke and Ansgar Scherp. 2022. Bag-of-words vs. graph vs. sequence in text classification: Questioning the necessity of text-graphs and the surprising strength of a wide MLP. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 4038–4051. Association for Computational Linguistics.

Yi Huang, Buse Giledereli, Abdullatif Köksal, Arzucan Özgür, and Elif Ozkirimli. 2021. Balancing methods for multi-label text classification with long-tailed class distribution. In *EMNLP (1)*, pages 8153–8161. Association for Computational Linguistics.

Thorsten Joachims. 1998. Text categorization with support vector machines: Learning with many relevant features. In *Proceedings of the 10th European Conference on Machine Learning*, ECML'98, pages 137–142, Berlin, Heidelberg. Springer-Verlag.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9.

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

David D. Lewis, Yiming Yang, Tony G. Rose, and Fan Li. 2004. RCV1: A new benchmark collection for text categorization research. *J. Mach. Learn. Res.*, 5:361–397.

Han Liu, Caixia Yuan, and Xiaojie Wang. 2020. Label-wise document pre-training for multi-label text classification. In *Natural Language Processing and Chinese Computing - 9th CCF International Conference, NLPCC 2020, Zhengzhou, China, October 14-18, 2020, Proceedings, Part I*, volume 12430 of *Lecture Notes in Computer Science*, pages 641–653. Springer.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *arXiv*, abs/1907.11692.

University of Michigan, College of Literature, Science, and the Arts LSA. 2023. An Introduction to Content Warnings and Trigger Warnings. https://sites.lsa.umich.edu/inclusive-teaching-sandbox/wp-content/uploads/sites/853/2021/02/An-Introduction-to-Content-Warnings-and-Trigger-Warnings-Draft.pdf. Last accessed: May 10, 2023.

University of Manchester, Institute of Teaching and Learning Man. 2023. Content Notes for Programmes, Course Units and Specific Activities and Resources. https://www.staffnet.manchester.ac.uk/umitl/resources/inclusivity/content-notes-in-teaching/. Last accessed: May 10, 2023.

Eneldo Loza Mencía and Johannes Fürnkranz. 2008. Efficient pairwise multilabel classification for large-scale problems in the legal domain. In *ECML/PKDD (2)*, volume 5212 of *Lecture Notes in Computer Science*, pages 50–65. Springer.

Ioannis Mollas, Zoe Chrysopoulou, Stamatis Karlos, and Grigorios Tsoumakas. 2020. ETHOS: an online hate speech detection dataset. *CoRR*, abs/2006.08328.

University of Nottingham Nott. 2021. Content notes policy, 2021-22. https://www.nottingham.ac.uk/educational-excellence/documents/content-notes-policy-2122.pdf. Last accessed: May 10, 2023.

Hyunji Hayley Park, Yogarshi Vyas, and Kashif Shah. 2022. Efficient classification of long documents using transformers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 702–709. Association for Computational Linguistics.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Quacquarelli Symonds Limited QS. 2023. QS World University Rankings 2023: Top global universities. https://www.topuniversities.com/university-rankings/world-university-rankings/2023. Last accessed: May 25, 2023.

University of Reading Read. 2023. Guide to policy and procedures for teaching and learning; Guidance on content warnings on course content ('trigger' warnings). https://www.reading.ac.uk/cqsd/-/media/project/functions/cqsd/documents/qap/trigger-warnings.pdf. Last accessed: May 10, 2023.

Russel Group. 2023. Russel Group: Our universities. https://russellgroup.ac.uk/about/our-universities. Last accessed: May 25, 2023.

Anshumali Shrivastava and Ping Li. 2014. In defense of minhash over simhash. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics, AISTATS 2014, Reykjavik, Iceland, April 22-25, 2014*, volume 33 of *JMLR Workshop and Conference Proceedings*, pages 886–894. JMLR.org.

Manuka Stratta, Julia Park, and Cooper deNicola. 2020. Automated content warnings for sensitive posts. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems, CHI 2020, Honolulu, HI, USA, April 25-30, 2020*, pages 1–8. ACM.

Times Higher Education THE. 2023. World University Rankings 2023. https://www.timeshighereducation.com/world-university-rankings/2023/world-ranking. Last accessed: May 25, 2023.

OTW The Organization for Transformative Works. 2023. Wrangling guidelines.

Stanford Teaching and Learning Hub TLHUB. 2022. Writing Content Notices for Sensitive Content. https://www.reading.ac.uk/cqsd/-/media/project/functions/cqsd/documents/qap/trigger-warnings.pdf. Last accessed: May 10, 2023.

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

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

Lin Xiao, Xin Huang, Boli Chen, and Liping Jing. 2019. Label-specific document representation for multi-label text classification. In *EMNLP/IJCNLP (1)*, pages 466–475. Association for Computational Linguistics.

Pengcheng Yang, Xu Sun, Wei Li, Shuming Ma, Wei Wu, and Houfeng Wang. 2018. SGM: sequence generation model for multi-label classification. In *COLING*, pages 3915–3926. Association for Computational Linguistics.

Zhichao Yang, Shufan Wang, Bhanu Pratap Singh Rawat, Avijit Mitra, and Hong Yu. 2022. Knowledge injected prompt based fine-tuning for multi-label few-shot ICD coding. *CoRR*, abs/2210.03304.

A Sources of Trigger Warnings and the Structured Set

Trigger warnings	Definition and Example Tags
Aggression-related	
Violence	Physical violence and destruction. *Manhandling, Slapping, Vandalism, Torture*
Kidnapping	Kidnapping, abduction, and it's consequences. *Captivity, Hostage situations, Stockholm syndrom*
Death	Graphic death, murder, and dying characters. *Drowning, Decapitation, Corpses*
Collective-violence	Organized violence by groups. *Terrorism, Civil war, Gang violence*
Other-aggression	*Violent thoughts, Slavery, Cannibalism*
Abuse-related	
Abuse	General abusive treatment. *Domestic Abuse, Bullying, Compulsion, Humiliation*
Sexual-abuse	Abuse and assault with sexual intent. *Rape, Sexual harassment, Voyeurism*
Child-abuse	Abuse of a child. *Child neglect, Pedophilia, Grooming, Child marriage*
Animal-abuse	Mistreatment and death of animals. *Animal Sacrifice, Harm to animals*
Abusive-language	Verbal abuse and strong language. *Threats of rape/violence, Insults, Hate speech*
Other-abuse	*Extortion, Intimidation*
Discrimination-related	
Classism	Discrimination based on social class. *Rich/Poor, Caste divide, Social hierarchies*
Transphobia	Discrimination against transgender persons. *Misgendering, Deadnaming, Transmisogyny*
Sexism	Discrimination based on gender stereotypes. *Misogyny, Patriarchy, Slut shaming*
Religious	Discrimination based on religion. *Islamophobia, Antisemitism, Anti-Catholicism*
Ableism	Discrimination against disabled persons. *Ableist slurs, Ableist language*
Body-shaming	Discrimination based on body properties. *Fat-shaming*
Racism	Discrimination based on race. *Racist Language, Segregation, Xenophobia*
Homophobia	Discrimination against homosexuality. *Homophobic Language, Heteronormativity, Gay Panic*
Other-discrimination	Discrimination against other or general groups. *Stereotypes, Bigotry, Cultural appropriation*
Mental Health-related	
Mental-illness	Severe mental illness with consistent or institutional treatment. *Insanity, OCD, Psychosis*
Dysmorphia	Body dissociation and consequential action. *Dysmorphia, Dysphoria, Eating disorder*
Addiction	Substance or gambling addiction and abuse. *Drug abuse, Withdrawal, Drinking to cope*
Self-harm	Self-destructive acts or behavior. *Cutting, Self-destruction*
Suicide	Suicide attempt, ideation, conduct, and aftermath. *Suicide*
Other-mental-health	Psychological issues that require help. *Depression, Trauma, Survivor guilt, Anxiety disorder*
Sexual-related	
Pornography	Graphic display of sex, plays, toys, kinks, technique descriptions.
Incest	Sex between family members. *Sibling Incest, Twincest*
Other-sexual	Non-graphic mentions of/ discussions about sex. *Sex shop, Sex education, Nudity*
Pregnancy-related	
Miscarriage	Death of the unborn and unplanned termination of pregnancy. *Miscarriage, Stillbirth*
Abortion	Planned termination of pregnancy. *Abortion*
Childbirth	Being pregnant and giving birth. *Pregnancy, Childbirth*
Other-pregnancy	Fertility, recovering from pregnancy, and issues with newborn. *Fertility Issues, Lactation*
Medical-related	
Blood and gore	Display of gore. *Blood, Open wounds, Organs*
Procedures	Medical procedures. *Amputation, Stitches, Surgery*
Other-medical	Illnesses and injuries. *Cancer, Hanahaki disease*
Other-content-warning	Crime, Police, Weapons, Needles, Prisons, Fluff, Politics, ...

Table 6: The complete set of 36 trigger warning labels. The *examples* are chosen from the manually classified canonical tags. Since trigger warnings are an inherently open-set problem (concerning every imagery relating to traumatizing experiences), there are other potentially triggering concepts in AO3 which are not part of our taxonomy. The examples are verbatim warnings which are not classified as a warning (cf. Table 2). Consider the annotation note summary in Appendix A.1 for further clarification of the label scope.

Cambridge (CCTL, 2023)	**York** (Childs et al., 2021)	**Stanford** (TLHUB, 2022)	**Reading** (Read., 2023)	**Michigan** (LSA, 2023)	**Manchester** (Man., 2023)	**Toronto** (CTL, 2021)	**Nottingham** (Nott., 2021)	**Merged**	**Labels**
		Discrimination (sexism, racism, homophobia, transphobia)						discrimination	**Discrimination**
		Stereotypes (gender, race, national origin, age)						stereotypes	
		hateful language or behaviour (e.g. racist, sexist, homophobic or transphobic language/behaviour)							
discrimination / bigotry (racism, misogyny, homophobia, transphobia, ableism, anti-Semitism, Islamophobia)								bigotry	
homophobia	*homophobic* lang./behaviour	*homophobia*	Homophobia and heterosexism	Homophobia and heterosexism	Homophobia and Homophobia	homophobia	Homophobia, biphobia, or heterosexism	Homophobia and heterosexism	Homophobia and homophobia
–	–	–	–	–	Heterosexism	–	Heterosexism	Heterosexism	
transphobia	*transphobic* lang./behaviour	*transphobia*	Transphobia and trans misogyny	Transphobia and trans misogyny	Transphobia	–	Transphobia	Transphobia and trans misogyny	transphobia
–	–	–	Classism	Classism	Classism	–	Classism	Classism	classism
–	*sexist* lang./behaviour	*sexism*	Sexism and misogyny	Sexism and misogyny	Sexism	–	Sexism and misogyny	Sexism and misogyny	sexism
misogyny	–	–	–	–	Misogyny	–	–	Misogyny	
–	–	–	*Stereotypes gender*	–	–	–	–	–	
racism	*racist* lang./behaviour	*racism*	Racism and racial slurs	Racism and racial slurs	Racism and racist slurs	depictions of racism or oppression	Racism and racial slurs	Racism and racial slurs	racism
–	–	*Stereotypes race and national origin*	–	–	–	–	–	–	
								Race stereotypes	
–	–	–	Hateful language at religious groups	Hateful language at religious groups	–	–	Hateful language at religious groups (e.g. Islamophobia, antisemitism)	Hateful language at religious groups	religious
Islamophobia	–	–	–	–	–	Islamophobia	*Islamophobia*	islamophobia	
anti-Semitism	–	–	–	–	–	–	*antisemitism*	anti-semitism	
ableism			Mental illness and ableism	Mental illness and ableism	Ableism		Ableism	ableism	ableism
–				*fat phobia*	*fat phobia*			fat phobia	body-shaming
–		*body shaming*						body shaming	
–		*age*						age	–
									Mental-health
disordered eating	–	Eating-disordered behavior or body shaming	Eating disorders and body hatred	Eating disorders, body hatred, and fat phobia	Eating disorders, body hatred, and fat phobia	Eating disorders	Eating disorders and body image	Eating disorder, body hate	dysmorphia
…									

Table 7: Creation of the trigger warning set (Discrimination warnings only) from the institutional lists. Shown are the verbatim statements from the lists, segmented into concepts (one per row). We segmented the source list by splitting the text at term-level into triggering concepts but keeping list-items and non-listing conjunctions intact (as those indicate shared semantics). **Concept groups** (top row), which bracket and name multiple concepts, were split completely. The name of the concept group informed the open-set labels. Terms in a multi-term concept that do not match the grouping were removed and *re-inserted* as new concept (Example: splitting Mental illness from Ableism; splitting body-shaming from body hatred and re-inserting into Discrimination). The **Merged** list contains all concepts with a unified name. The **Labels** group semantically similar concepts. The complete procedure is published in the repository Code: https://github.com/webis-de/ACL-23.

A.1 Annotation Note Summary

While resolving the diverging annotations (cf. Section 4.2), we created the label descriptions in Table 6, the descriptive dimensions (nature of the harm and actor-subject-intent), and the annotation guides that all label annotations must adhere to. We list the essential principles with some examples below.

General Principles The general principles take effect unless there is a label-specific exception declared.

- Exclude general indications of triggers without further specification of the topic. *Triggers, Additional Warnings In Author's Note, Additional Warnings Apply, Other: See Story Notes, Other Additional Tags to Be Added, Warnings May Change, Graphic Description, You Have Been Warned, Disturbing Themes*:
- If a tag or its synonyms is ambiguous and used to indicate triggering and non-triggering content, exclude it: *stuffing, Hardcore, Kinky, Crazy, Coping*.
- If a tag or its synonyms is ambiguous and used to indicate different triggering content, annotate it with all options: *Asphyxiation as sexual and death*.
- Exclude tropes: *Whump, Hurt-comfort, ...*.
- Exclude tags that describe the setting of the work, even if that setting refers to relevant content: *Post-World War 2*.
- Only annotate warnings that are indicated directly. Do not annotate warnings that only could be implied or associated: *Weapons or Safehouses do not indicate violence*
- Annotate fantasy adaptations of real concepts like the real concepts. *Alien or male pregnancy like pregnancy, Hanahaki disease like real disease, species dysphoria like gender dysphoria.*

Aggression

- Violence refers to physical harm. Violence that is mostly psychological is annotated as abuse. *Psychological violence is abuse, Threats of violence is abusive-language*
- Exception to a general principle: Execution devices imply violence *Guillotine, electric chairs*.
- Exception to a general principle: Weapons indicate violence if the violence is mentioned in the term: *Gun violence*.
- If a tag indicates both violence and death, annotate death. *Guillotines*.
- Annotate *Loss and Grief* as *Mental-health*, even if death is implied (cf. general rule on implications).
- Annotate dying and potentially dying as death, even if it does not ultimately lead to death: *Possible character death*.
- Annotate deadly violence as death *Murder, Assassination*. If violence is directly indicated, annotate death and violence: *Fight to the death*.
- Exclude tags where the death is a descriptor of the setting: *Dead Link*.
- Exclude 'Death' as a character.
- *Thought of violence* or other violent intent is aggression unless it is a graphic description in the narrative.
- Annotate acts of war, organized crime, drug-related crime, and organized violence as collective-violence. Mentions of military imagery are excluded (see the general rule about implications).
- Annotate all human trafficking as kidnapping.

Abuse

- Annotate 'forcing others to act' as abuse, including fantasy concepts: *Slavery, Mind-control, Compulsion*.
- If forced action is sexual in nature, annotate as sexual-abuse: *Non-consensual ...*.
- Annotate preferably the more specific abuse label (sexual, child, animal) instead of the more general 'abuse'.
- Annotate stalking, voyeurism, and rape as sexual-abuse.
- Annotate sexual abuse of children as child-abuse.
- Annotate hate-speech, threads, and intimidation as abusive-language. If the hate speech is targeted towards a group, annotate abusive-language and the discrimination-related label. *Racist slurs are racism and abusive-language*.

Mental-health

- Annotate mental-illness if the affliction requires (permanent) stationary treatment: *Schizophrenia, Psychosis, Insanity*.
- Annotate mental-health if the affliction (usually) requires help and causes the afflicted suffering if ignored. *Depression, Anxiety attacks*.
- Exclude stress, angst, or anxiety.

- Annotate substance abuse as addiction. Exclude recreational substance use (of weed/psychedelic drugs, tobacco, alcohol) if addiction or abuse is not indicated.
- Exception to a general principle: always annotate highly addictive drugs as addiction (even if no use/abuse is implied).
- Exclude medical drug use, unless 'self-medication' is implied.
- Annotate (sex/gender/species) dysphoria and eating-disorder as dysmorphia.

Sexual

- Annotate all tags as pornography if they indicate a sex act without intent to harm.
- Exception to a general principle: Annotate sex toys as pornography.
- Exception to a general principle: Annotate sexual position preference (*Top, Bottom*) as pornography.
- Exception to a general principle: Annotate sexual preferences/kinks as pornography if the kink is impossible to practice without any form of sex.
- Annotate kinks that do not (necessarily) require a sexual act as sexual: *Size kink, Praise kink, Plushophilia*

Pregnancy

- Annotate lactation/fertility (issues) and interactions/issues with newborns as pregnancy.

Medical

- Annotate medical if the action has no (direct) intent to harm with that action. Acts of harmful mutilation by others are aggression or abuse, even if they are medical procedures.
- Annotate (chronic) injuries and illnesses as medical, but exclude equipment (*Band-Aids, Needles*) and mild affections (*Allergies) (see general principles).*
- *Annotate wounds and open injuries as blood-gore.*

B Experimental Setting

We tested the effectiveness by undersampling the training dataset to 3 different label thresholds, testing 4 different feature sets for SVM and XGB, and testing all common model parameters. All models were trained once with 36 target labels (fine-grained) and once with 7 target labels (coarse-grained), where both variants were ablated individually. All ablation was done via grid search.

Model	Labels	Sample	Features	Parameters
SVM	fine	10k	1–3-grams, χ^2	$C = 2$
	coarse	10k	1–3-grams, χ^2	$C = 2$
XGB	fine	10k	1-grams	max_depth $= 4$, lr $= 0.25$
	coarse	10k	1–3-grams, χ^2	max_depth $= 4$, lr $= 0.25$
BERT	fine	69k	–	epochs $= 10$, lr $= 2e-5$
	coarse	69k	–	epochs $= 5$, lr $= 2e-5$
LF	fine	10k	–	epochs $= 2$, lr $= 2e-5$
	coarse	69k	–	epochs $= 3$, lr $= 2e-5$

Table 8: The best parameter configuration for SVM, XGBoost (XGB), RoBERTa (BERT) and Longformer (LF) according to macro averaged F_1 on the validation split.

The best configuration was selected by macro F_1 on the validation dataset. Model training was done on a single A100 GPU. The final parameter configurations are shown in Figure 8.

Dataset Sampling Since the training dataset (cf. Section 5) is very large and skewed towards a few very common labels, we undersampled the training dataset in 3 versions:

1. to the 25% quartile (10,000 works/label)
2. to the 50% quartile (28,000 works/label)
3. to the 75% quartile (69,000 works/label)

Our sampling strategy started with the rarest label and randomly added works with this label until, either, the threshold was reached, or, all documents with that label were added. Previously added documents (with multiple labels) counted towards the threshold. We ignored the occasional over-drawing of labels (when a high-frequency label was already sampled over the threshold by sampling the lower-frequency labels alone) since this behavior is difficult to avoid for multi-label datasets and did not very occur often. All models were ablated on all three input data samples, except for LF with fine-grained labels and XGB which were not trained on the 69,000 works sample due to resource limitations.

SVM and XGBoost Features All feature sets used tf-idf vectors of token n-grams (using the `bert-base-uncased` tokenizer) with a minimum document frequency of 5. We ablated the four feature sets:

1. token 1-grams,
2. token 3-grams,
3. token 1–3-grams and χ^2-feature selection, and
4. token 1–5-grams and χ^2-feature selection.

Warning	SVM	XGB	BERT	LF
coarse-grained (7 labels)				
sexual-content	0.87	**0.88**	0.71	0.63
aggression	**0.55**	0.53	0.53	0.52
abuse	**0.47**	0.42	0.40	0.36
mental-health	**0.58**	0.52	0.52	0.47
medical	**0.53**	0.52	0.40	0.39
pregnancy	0.61	**0.67**	0.43	0.42
discrimination	0.33	**0.37**	0.24	0.21
fine-grained (36 labels)				
pornography	0.86	**0.88**	0.76	0.66
violence	0.30	**0.33**	0.27	0.23
mental-health	0.34	**0.35**	0.29	0.33
death	0.24	0.26	**0.27**	0.25
sexual	0.09	0.12	**0.25**	0.07
sexual-abuse	0.33	**0.39**	0.34	0.25
abuse	0.23	**0.26**	0.24	0.23
medical	0.32	0.37	**0.41**	0.33
blood-gore	0.28	**0.34**	0.32	0.25
abusive-language	0.09	0.11	**0.21**	0.12
suicide	0.26	0.32	**0.34**	0.27
child-abuse	0.22	0.25	**0.31**	0.27
childbirth	0.55	**0.63**	0.47	0.44
mental-illness	0.11	**0.16**	**0.16**	0.15
addiction	0.22	**0.33**	0.26	0.27
incest	0.52	**0.53**	0.50	0.37
homophobia	0.31	**0.39**	0.27	0.21
self-harm	0.37	**0.41**	0.33	0.29
kidnapping	0.26	**0.36**	0.25	0.23
aggression	0.33	**0.38**	0.31	0.26
collective-violence	0.35	**0.36**	0.32	0.20
procedures	0.26	**0.30**	0.17	0.17
dysmorphia	0.41	**0.44**	0.34	0.23
pregnancy	0.37	**0.44**	0.21	0.23
abuse	0.20	**0.21**	0.11	0.08
sexism	**0.14**	**0.14**	0.01	0.05
discrimination	0.06	0.06	0.00	0.05
racism	0.10	**0.17**	0.06	0.12
miscarriage	0.18	**0.35**	0.18	0.16
animal-abuse	0.08	**0.17**	0.11	0.14
transphobia	0.14	**0.34**	0.17	0.20
abortion	0.17	**0.32**	0.02	0.18
ableism	0.00	0.06	0.00	**0.07**
religious-discrimination	0.10	**0.12**	0.04	0.09
classism	**0.10**	0.05	0.00	0.04
body-shaming	0.00	0.00	0.00	0.00

Table 9: Classification effectiveness of SVM, XGBoost (XGB), RoBERTa (BERT), and Longformer (LF) on the test dataset. Shown are the micro F_1 scores for each label individually.

For SVM, we selected the best 50,000 features. For XGB, we selected the 20,000 best features. Preprocessing and tokenization were identical for all approaches, as described in Section 6.

Model Parameters For SVM, we ablated the regularization parameter $C \in \{0.1, 0.2, 0.5, 1.0, 2.0\}$. For XGB, we ablated the tree depth max_depth $\in \{2, 3, 4\}$ and the learning rate $\in \{0.25, 0.5, 0.75\}$ with 100 estimators and early stopping at 10 rounds. For BERT, we ablated the number of epochs $\in \{3, 5, 10\}$ and the learning rate $\in \{1e-4, 5e-5, 2e-5, 1e-5\}$ with a batch size of 32. For LF, we ablated the number of epochs $\in \{2, 3, 5\}$ and the learning rate $\in \{1e-4, 5e-5, 2e-5, 1e-5\}$ with a batch size of 4.

C Data Statement

Following Bender and Friedman (2018) we provide a data statement to document the construction of the violence trigger warnings corpus.

C.1 Curation Rationale

The goal is to extract a trigger warning corpus out of an existing resource with imperfect labels. We use the free-form tags that authors assign to their own works to infer trigger warnings, which might introduce biases through the author (ambiguous use, thoughtless use, misuse of labels, differing understanding), our method of inference (true positive/false positives), or our interpretation of the labels during manual annotation steps. However, our assumption is that the authors still know best how to tag their own works.

We curated works from the corpus to be included in the dataset to be used in machine learning experiments. Some of these curation actions were done to mitigate the aforementioned issues of label reliability: we excluded works with non-canonical tags (since our method can't guarantee that the non-canonical tags are no warnings) and with very few hits and kudos (i.e likes) as a form of community-based noise filtering.

Other curation actions were taken to prevent abnormal algorithmic behavior on works with outlier characteristics: we removed the top percentile of works by length, tag count, and chapter count. We also exclude works without warning labels in the dataset to reduce the size and sharpen relevance to the research question. This might bias algorithmic models to misbehave on works without any triggering characteristic.

C.2 Language Variety

Our corpus of Archive of our Own (AO3) includes fanfiction in 91 languages. However, we only considered English documents. English is with 7.1 million documents the predominant language on the platform. The second most common language is Chinese with ca. 370,000 works. There are only between 1 and a few hundred works written in all other languages.

C.3 Speaker Demographic

AO3 hosts fanfiction works from a variety of authors which is why the true demographic is unknown. The only information available to date is a Census taken in 2013, where a survey was

conducted (Archive of Our Own, 2013) to which 10,005 users (not authors but overlap is possible) replied. We summarize key points from that survey: The average user age at that time was 25 years. Most users identified themselves as female (80%), with genderqueer being second (6%), and male third (4%); other options choices were Transgender, Agender, Androgynous, Trans, Neutrois, or Other (2% or less each). Regarding ethnicity, the majority of users identified as White (78%), followed by Asian (7%), Hispanic (5%), Mixed/Multiple (5%), Black (2%), Native American (1%), Pacific Islander (1%), and Other (1%). Only 6% of users stated that they used AO3 for languages other than English. The AO3 Census evaluation states that this survey is not representative and has its limitations but also that the survey should not be dismissed as "[these limitations] do not make the survey useless". Unfortunately, there has not been another Census since then.

C.4 Annotator Demographic

All annotation (and data curation) work was done by the authors of this work, which have an extensive background in computational linguistics, computer science, and communication science. All annotators are Caucasian, aged from 25–50, male and female gender, and of diverse sexualities. Native languages spoken are German and Polish.

C.5 Speech Situation

All of the texts are written works that are or were available online at some point. Fanfiction is often written spontaneously and with little editing, although some authors follow longer planning and editing cycles. Many popular works are edited by community members.

Each work has a publication date attribute which, however, might just reflect the upload date instead of the date of writing since some works were also posted on other sites before. However, it can also be correctly backdated. Most works are recent creations and were created after the launch of AO3 in 2009, with a linear growth in yearly new submissions. AO3 also systematically archives older fanfiction works, the earliest works originate from ca. 1970. However, those are comparably rare.

C.6 Text Characteristics

Almost all texts in this corpus belong to the fanfiction genre. Many fanfiction works revolve (non-exhaustively) around fictional characters from books, cartoons, anime, manga, music, and movies or non-fictional characters such as celebrities. Works will often use specialized vocabulary exclusive to the Fandom they write about and common terms will be used with a fandom-specific meaning. Fanfiction also has a domain-specific vocabulary to describe fanfiction content in general. We frequently used `urbandictionary.com` as well as various fanfiction wikis for explanations of the domain-specific vocabulary.

D Additional Analysis Results

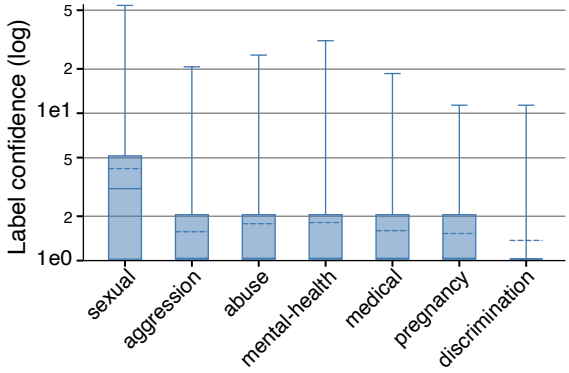

Figure 6: Distribution of the label confidence for each (coarse-grained) label. The label confidence for a warning of one work is the number of free-form tags assigned to that work, which are annotated with the respective warning. Dashed lines indicate the mean.

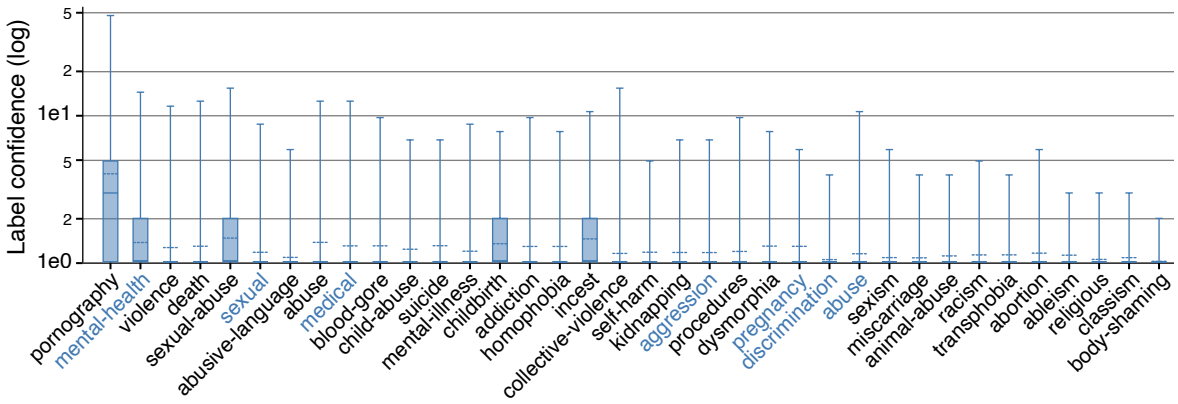

Figure 7: Distribution of the label confidence for each (fine-grained) label. The label confidence for a warning of one work is the number of free-form tags assigned to that work, which are annotated with the respective warning. Dashed lines indicate the mean.

Set	Macro F_1								Micro F_1										
	Total	**Length**				**Open-endedness**		**Confid.**	**Total**	**Length**				**Open-endedness**		**Confid.**	**Section**		
		512	4k	16k	16k+	Open	Closed			512	4k	16k	16k+	Open	Closed		Top	Mid	Bot
fine-grained																			
SVM	0.25	0.21	0.27	0.24	0.18	0.24	0.25	0.28	0.49	0.40	**0.53**	0.49	0.39	0.26	0.54	**0.82**	**0.43**	**0.55**	0.50
XGB	**0.30**	0.19	**0.29**	**0.31**	**0.30**	**0.28**	**0.30**	**0.30**	**0.52**	0.36	**0.53**	**0.53**	**0.49**	**0.29**	**0.56**	**0.82**	0.42	0.52	**0.55**
BERT	0.23	**0.30**	0.27	0.20	0.15	0.23	0.23	0.23	0.45	**0.53**	0.49	0.41	0.34	**0.29**	0.48	0.61	0.37	0.54	0.45
LF	0.21	0.29	0.24	0.17	0.14	0.19	0.21	0.15	0.36	0.46	0.41	0.31	0.26	0.24	0.38	0.52	0.26	0.48	0.46
coarse-grained																			
SVM	**0.56**	0.51	**0.57**	0.56	0.53	–	–	0.57	**0.66**	**0.59**	**0.68**	0.66	0.59	–	–	0.78	–	–	–
XGB	**0.56**	0.37	0.54	**0.57**	**0.59**	–	–	**0.60**	**0.66**	0.48	0.66	**0.67**	**0.66**	–	–	**0.81**	–	–	–
BERT	0.46	**0.52**	0.48	0.43	0.43	–	–	0.40	0.53	**0.59**	0.57	0.50	0.48	–	–	0.57	–	–	–
LF	0.43	**0.52**	0.45	0.39	0.39	–	–	0.37	0.49	0.58	0.52	0.44	0.43	–	–	0.5	–	–	–

Table 10: Classification effectiveness of SVM, XGBoost (XGB), RoBERTa (BERT), and Longformer (LF) on the test dataset, split by various characteristics. **Total** indicates the overall F_1 scores. **Length** indicates the scores on documents in the length (of tokens) intervals 50—512, 512–4,096, 4,096–16,000, and 16,000–93,000 (16k+). **Open-endedness** indicates the scores on the open or closed classes exclusively. Label confidence (**Confid.**) indicates the scores on all works that have at least 2 free-form tags as support for each assigned warning. **Section** indicates the average scores of only the 12 most common tags (top 33%), and equivalently the middle and bottom third.

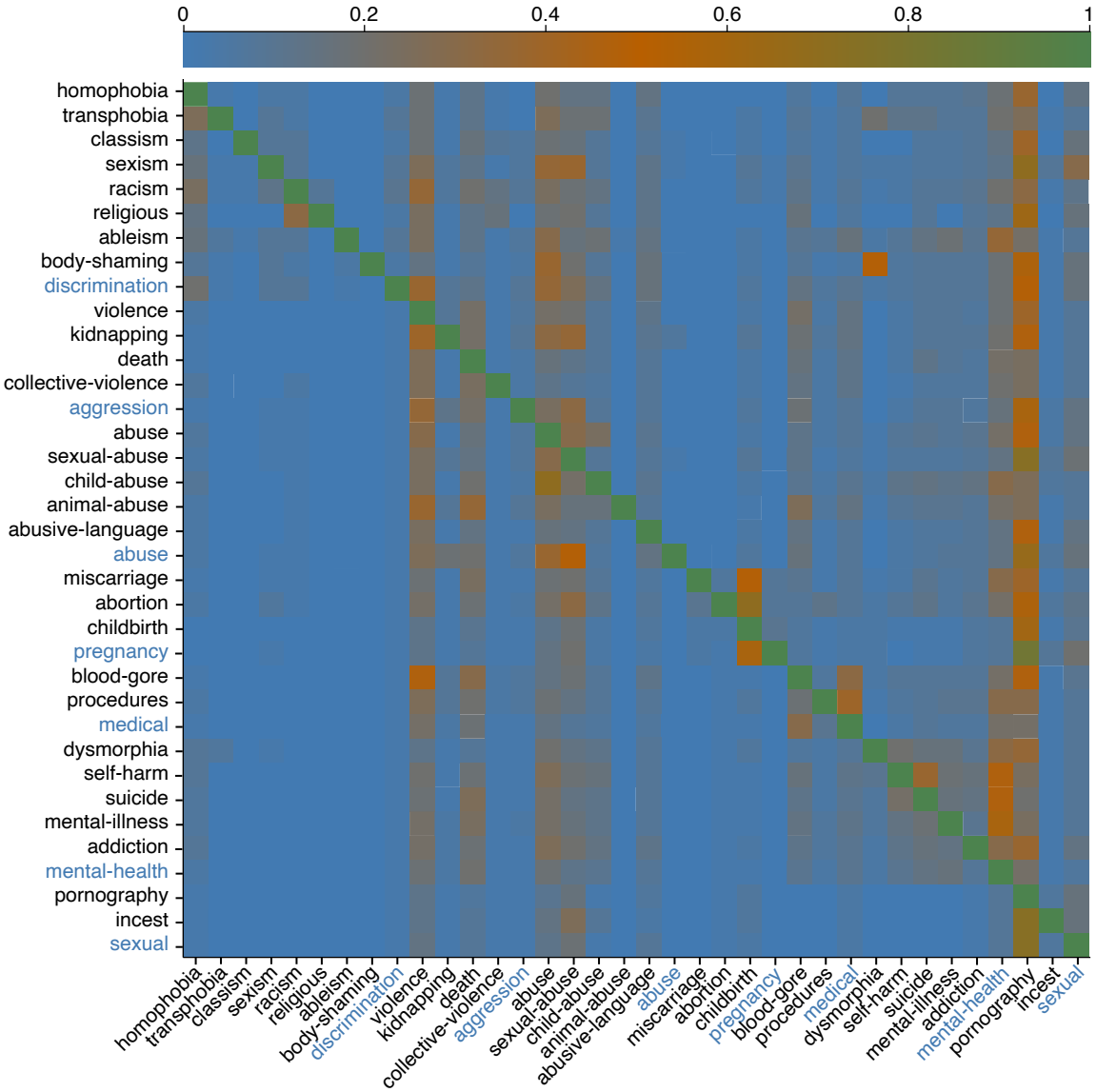

Figure 8: Co-occurrences between labels. Fields show which fraction of the row label also occurs with the column label. Labels are ordered by label group (as in the taxonomy visualization).