Trigger Warnings: Bootstrapping a Violence Detector for Fan Fiction

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

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

Abstract

We present the first dataset and evaluation results on a newly defined task: assigning trigger warnings. We introduce a labeled corpus of narrative fiction from Archive of Our Own (AO3), a popular fan fiction site, and define a document-level classification task to determine whether or not to assign a trigger warning to an English story. We focus on the most commonly assigned trigger type "violence" using the warning labels provided by AO3 authors as ground-truth labels. We trained SVM, BERT, and Longfomer models on three datasets sampled from the corpus and achieve F_1 scores between 0.8 and 0.9, indicating that assigning trigger warnings for violence is feasible.

Warning. This paper shows potentially triggering terms related to the subject of violence.

1 Introduction

> "[The witch] crept up and thrust her head into the oven. Then Grethel gave her a push that drove her far into it, and shut the iron door, and fastened the bolt. Oh! then she began to howl quite horribly, but Grethel ran away, and the godless witch was miserably burnt to death."
>
> *Hansel and Gretel, a fairy tale*[†]

Violence and cruelty are commonplace in literature. Folk tales, especially fairy tales, but also children's and youth literature are full of dark, horror images, such as burning a human being alive in an oven, as in the fairy tale by the Brothers Grimm quoted above. And even if most people will not be deeply shaken by such content, some readers may mentally relive their past traumas evoked by the imagery. To proactively alert readers that a text they are about to read contains potentially disturbing material, so-called "trigger warnings" have been proposed.

Trigger warnings (also referred to as content warnings/notifications/alerts) emerged in online communities (e.g., on Tumblr and LiveJournal) in the early 2000s (Knox, 2017). They are usually presented as short phrases/keywords preceding a text and warn of potentially disturbing content. While there are no universally accepted trigger warnings (anything can be a trigger), many universities meanwhile published guidelines (see, e.g., lists published by the Universities of Reading and Michigan (UR list; UM list)). They include largely overlapping lists of triggers referring to health (eating disorders, mental illness) or sexuality (sexual assault, pornography), verbal violence (hate speech, racial slurs), and physical violence (animal cruelty, blood, suicide), among others.

Surprisingly, assigning trigger warnings is considered a manual task, and, to our knowledge, there is no work in Computer Science in general, and in Natural Language Processing in particular, that addresses content warnings. We lay the foundation to close this gap by introducing the new NLP task of trigger warning assignment, formulated as follows:

> Given a text and a trigger label, assign a warning to the text if it contains a corresponding trigger.

When multiple trigger labels are predefined, this task can be extended from a binary classification problem to a multi-class or multi-label problem and solved by, for example, a set of binary classifiers, one for each trigger. However, the preceding first step is to investigate the feasibility of automatic trigger warning assignment and for this purpose we create the first trigger warning corpus from narratives with and without triggers, using the trigger warnings supplied by the works' authors.

Our contributions are the following: we (1) introduce the new task of automatic trigger warning assignment, (2) introduce the first corpus compiled from a public archive of fan fiction marked with a trigger warning for violence (Section 3), and (3) evaluate models for assigning trigger warnings and analyze their effectiveness (Section 4).[1]

[†]Translation by Margaret Hunt

[1]Code and data: https://github.com/webis-de/EMNLP-23

2 Related Work

Constructs related to "trigger warnings" have been investigated using computational approaches under different terms and have spanned a broad range of phenomena. Recent research employs terms such as "objectionable content", "objectionable material", "harmful content", "harmful text" (Banko et al., 2020; Solorio et al., 2021; Kirk et al., 2022) as broad terms covering diverse types of content that can potentially evoke negative emotions in the recipient of the material (be it verbal or visual), i.e. cause emotional harm at different degrees of severity. The type of content that is often subsumed under those terms includes violence, sexual content, misguided messages, misinformation, verbal aggression, malice, callousness, or social aggression, among others. And while there is also a clear link to sentiment analysis, phenomena subsumed under "objectionable/harmful content" lie only on one end of the sentiment scale (that of negative sentiment), however, have a finer granularity (cf. range of specific types of content, mentioned above, that may evoke harm).

Now, the notion of "triggering" is equally underspecified (open-ended), but even broader. While most of the objectionable types are indeed unobjectionably harmful—in that they can be linked to *intention to harm*—there may exist concept associations that are triggering to some individuals which, objectively speaking, have little to no link to intention to harm; consider, for instance, that a mention of a thunderstorm may be triggering to a victim of a severe lightning injury. Thus, triggering covers also concepts which would normally be understood to lie at the positive end of a sentiment scale, which can, however, evoke negative associations in some individuals due to their specific traumatic past experience related to the concept. A "trigger warning" just gives a nominal label to the signal that is considered triggering. While we are not aware of prior work on automatic trigger warning assignment nor specifically violence warning assignment, below we outline prior work in NLP and computer science that covers most closely related topics.

Identifying Causes of Emotions While affect and emotion recognition in non-fiction text—sentiment analysis more generally—has been long studied in NLP (Alswaidan and Menai, 2020), research into interactions between emotions and their triggering cause events was introduced only about a decade ago (Lee et al., 2010). Cause events here refer to (verb) arguments or events in the text that are highly correlated with a certain emotion, positive or negative. The goal of the emotion cause extraction task is to identify the emotion's stimulus and the computational methods range from rule-based lexico-syntactic approaches through traditional classifiers to recently also deep learning; see Khunteta and Singh (2021) for an overview of the emotion cause extraction area. By contrast the trigger warning assignment task is rather about identifying potentially triggering content which may evoke strongly negative emotions in readers.

Identifying Verbal Violence Interest in broadly understood verbal violence—although not explicitly referred to as such—has a long history in the NLP community. Waseem et al. (2017) and Kogilavani et al. (2021) propose taxonomies of abusive and offensive language, respectively; Kogilavani et al. also survey techniques for offensive language detection. Fortuna and Nunes (2018) and Schmidt and Wiegand (2019) provide an overview on hate speech and Mishra et al. (2019) more generally on abuse detection methods with "abuse" defined as "any expression that is meant to denigrate or offend a particular person or group". While not considered from the point of view of triggering, this definition fits the category 'Hateful language" listed in the institutional guidelines. While most work on verbal violence has been carried out in the context of social media (methods ranging from feature engineering to neural networks) it would be useful to extend those systems to cover a broader range of verbal violence, e.g., literary dialogue, in the context of the trigger warning assignment task.

Identifying Health-related Triggering Content Closest to our research, however, focused on a different trigger type is the work of De Choudhury (2015) investigating behavioral characteristics of the anorexia affected population on Tumblr. Analysis of several thousand posts has shown that the platform contains vast amounts of triggering content which may prompt and/or reinforce anorexia-oriented lifestyle choices. Two sub-groups of the anorexia community were identified—pro-anorexia and pro-recovery—with distinguishing affective, social, cognitive, and linguistic properties. Predictive models based on language features extracted from the posts were able to detect anorexia content at 80% accuracy. Like De Choudhury, we focus on

Sample	Trigger	No. of	Median no. of			
		Works	Words	Kudos	Hits	FF
Corpus	violent	571,525	5,732	40	782	8
	non-violent	4,4 M	1,847	52	758	5
Random	violent	10,000	6,773	51	1,088	8
	non-violent	10,000	1,869	74	1,074	5
Fame	violent	10,000	16,810	238	4,706	11
	non-violent	10,000	2,859	224	3,155	6
Tags	violent	10,000	7,161	60	1,255	9
	non-violent	10,000	2,127	84	1,235	6

Table 1: Descriptive statistics of corpus and sample datasets. Shown are number of works and median numbers of words, kudos, hits, and freeform tags (FF). The median is reported due to the long-tailed nature of the measures; the mean is ca. 2-4 times higher.

a single trigger type, but in fiction texts and with warnings assigned by the authors.

3 The Violence Trigger Warnings Corpus

As data source, we used Archive of Our Own (AO3),[2] a public online anthology of fan fiction, i.e., amateur writings inspired by existing works of fiction: e.g., novels, cartoons, manga. At the time of corpus creation, AO3 hosted about 8 million works. Aside from basic meta-data, such as title, author, language, statistics (number of words, chapters, etc.), reader reactions, ratings, fandoms (original source(s)/inspiration), and relationships (characters involved in romantic/platonic relationship(s)), crucially for this research, works are labeled with *Archive Warnings* and *Additional Tags*.

Archive Warnings AO3 defines a set of six content warnings. Authors must actively assign at least one to each of their works. The labels are: (1) *Major Character Death*, (2) *Underage* (contains sexual activity by characters under 18), (3) *Rape/Non-Con* (non-consensual sexual activity), (4) *Graphic Depictions of Violence* (gory, explicit violence), (5) *Creator Chose Not To Use Archive Warnings* to avoid spoilers, and (6) *No Archive Warnings Apply*, if the work has no triggering content.

Additional Tags AO3 allows authors to define open-set, freeform content descriptors, which are used as keywords for search and browsing, like *romance, slow burn, fluff, and jealousy*, but also to assign additional trigger warnings like *abandonment, monsters, blood drinking*. Additional Tags are heterogeneous, user-generated content but frequently used tags are "canonized" by volunteer "tag wranglers". The use of canonized tags is encouraged and supported by the web interface.[3]

Corpus Acquisition For the purpose of corpus acquisition, the entire AO3 was crawled. Works were identified via AO3-search using the `created_at:`DATE-RANGE query parameter. Individual searches were started for each day since the date of the site's creation in order to distribute the load; AO3's crawling limits were observed. URLs which were not publicly accessible, redirected to external sites or yielded HTTP errors were omitted. Our complete crawl contains 7,866,512 works with 9,705,174 distinct *Additional Tags*. 571,525 works are labeled with *Graphic Depictions Of Violence*.

Dataset Sampling Because AO3 works do not include any annotations below document level—that is, we do not know the extent of violent content nor where in the text it can be found—our goal was to build a corpus with high-confidence examples of texts with and without violence. We apply three sampling strategies with varying reliability criteria: random sampling to represent the corpus, fame-based sampling to exclude low-effort works, and tag-based sampling to exclude works that are not thoroughly tagged so that *Archive Warnings* might be less reliable. Table 1 gives an overview of the corpus and the three sampled datasets.

All sampling strategies randomly select 10,000 violent works (tagged with *Graphic Depictions of Violence*) and 10,000 non-violent works (tagged with *No Archive Warnings Apply* but not with *Graphic Depictions of Violence*). Before selecting the examples, we discarded all works with less than 100 words and works written in a non-English language. The random sample then draws the examples uniformly at random. The fame-based sample first discards all works with $<$1,000 hits and $<$100 kudos and then draws uniformly at random. The tag-based sample discards all works with $<$10 *Additional Tags* (including characters and relationships) and then draws uniformly at random.

Table 1 shows the meta-data of the entire corpus and the three samples, extended by Table 4 in Appendix A. The random and tag-based samples are highly similar to the overall corpus; the fame-based sample diverts by having longer (esp. violent) documents with more freeform tags.

[2] https://archiveofourown.org

Sample	Model	F_1	P	R	Acc.
Random	SVM	**0.864**	**0.860**	0.868	**0.863**
	Longformer	0.862	0.842	**0.882**	0.859
	BERT	0.786	0.751	0.825	0.775
Fame	SVM	**0.893**	**0.881**	**0.905**	**0.892**
	Longformer	0.862	0.823	**0.905**	0.856
	BERT	0.796	0.808	0.784	0.799
Tag-frequency	SVM	**0.864**	**0.876**	0.851	**0.866**
	Longformer	0.848	0.829	0.868	0.844
	BERT	0.789	0.701	**0.901**	0.756

Table 2: Classification effectiveness on the test set for all sample datasets; reported are F_1 score, precision (P), recall (R), and accuracy (Acc.); bold = best result.

4 Assigning Violence Trigger Warnings

We evaluate the four labeled datasets in a text classification setting by building classification models to assign trigger warnings at the document level.

Models We use three long-document classification baselines for our experiments: SVM, BERT, and Longformer. First, we use support vector machines (SVM) (Joachims, 1998) since they are often used for text classification, are easily interpretable, and are not limited by the input sequence length. Second, we use a BERT transformer (Devlin et al., 2019) as the go-to classification baseline; we used the pretrained `bert-base-uncased` checkpoint with 12 layers and 110M parameters, fine-tuned on our classification task. Third, we use a sparse-attention Longformer (Beltagy et al., 2020) as the state-of-the-art in many long document classification tasks (Park et al., 2022). We used the `allenai/longformer-base-4096` pretrained checkpoint, fine-tuned on our classification task.

Text Preprocessing For the SVM, we remove HTML tags, URLs, emojis, numbers, punctuation, and special characters and apply the Porter Stemmer (Porter, 1980). For BERT and Longformer, we only remove HTML tags, URLs, numbers, and special characters, while punctuation is retained. For both neural models, the inputs are truncated at (and padded to) the maximum sequence length.

Classification Setup The preprocessed data are split into 90:10 training and test sets via stratified sampling to maintain the class distribution.

As features for the SMV we use binary, uni- and bigram bag-of-word document vectors obtained from the lowercased preprocessed text; we keep only each dataset's 100,000 most frequent features. Maximum sequence lengths of 512 tokens for BERT and 4,096 tokens for Longformer are used. The training configuration and ablation can be found in Appendix C.

[3] https://archiveofourown.org/wrangling_guidelines/2

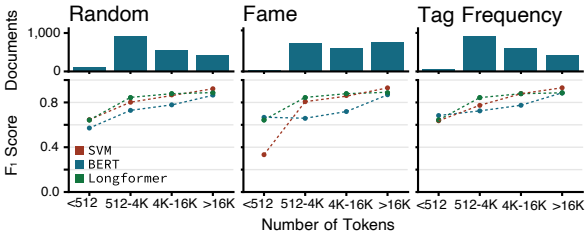

Figure 1: Classification effectiveness in terms of F_1 on the sample datasets over intervals of number of tokens.

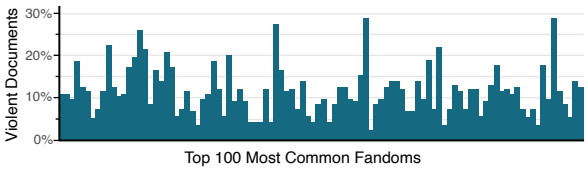

Figure 2: Ratio of violent to non-violent fan fiction for the top 100 most common fandoms in out corpus.

Results For each sample and model, we train a model on the training set and evaluate on the test set, the results of which are reported in Table 2. It can be seen that the SVM reaches overall best scores except for recall. Across the three sample datasets, the models achieve best F_1 on the fame-based sample, followed by the random and the tag-based sample. Recall is higher than precision for most neural models and vice versa for the SVM.

Figure 1 shows the effectiveness of the models on subsets of documents of varying lengths over input length. If the documents are shorter than the model's maximum input length, the SVM almost always performs worse (in terms of F_1) than the neural models and vice versa.

5 Discussion and Limitations

The final result (the SVM beats both neural models) is unexpected and can be (partially) explained by the influence of document length and topic.

Document Length Although the SVM has no contextual semantic information, it covers the tokens of the whole document through the bag-of-words representation, while BERT and Longformer are limited to a fixed input sequence (512/4,096 tokens respectively), which is only a fraction of the documents (cf. Table 1). Our analysis of the relation between text length and effectiveness (cf. Figure 1) reveals that neural models perform better

Random	Fame	Tag
Features indicating violence		
4.65 blood	3.82 blood	4.54 blood
2.40 dead	2.32 screams	2.62 dead
2.37 kill	2.02 scream	2.23 screams
2.33 screams	1.94 dead	2.13 pain
1.99 screamed	1.91 kill	2.03 bloody
1.95 flesh	1.89 pain	1.96 scream
1.89 screaming	1.89 killed	1.93 bleeding
1.86 scream	1.84 bloody	1.93 blade
1.79 pain	1.81 bleeding	1.91 kill
1.77 killed	1.75 blade	1.87 killed
⋮	⋮	⋮
0.91 hannibal (84)	0.55 sith (341)	0.97 hannibal (67)
Features indicating non-violence		
-1.67 kiss	-1.16 kiss	-1.86 kiss
-1.07 managed	-0.96 embarrassing	-1.00 teasing
-1.01 ridiculous	-0.91 halfway	-0.93 spent
-0.92 admit	-0.90 experience	-0.92 demanded
-0.91 teasing	-0.90 surprised	-0.90 hadn
-0.91 shoulders	-0.87 close	-0.89 fin
-0.89 snorted	-0.82 dance	-0.89 flushed
-0.89 curled	-0.81 teasing	-0.87 imagined
-0.88 weekend	-0.80 ridiculous	-0.85 ridiculou
-0.88 surprised	-0.80 kissing	-0.84 carefully

Table 3: Most discriminative SVM features for both classes and all three sample datasets. The upper row group also lists the first topic (fandom-specific) feature, it's score, and position in the list (rank). It should be noted that there are almost no topic features in the top 1000 features which we inspected manually.

than the SVM on documents shorter than their input limit; on longer documents, the violence might not have been part of the truncated input.

Topic Another possible explanation for the SVM's effectiveness is that the classes are separable by topic words (characters, fandom concepts) due to co-occurrence with (non-)violent documents; hence the classifier could not learn the more complex concept of violence. Our analysis (cf. Figure 2) shows that some fandoms are more violent than others (between 5–30% of works) and that about 5% of tagged characters and 2% of freeform tags are strongly associated with violent documents (strongly non-violent ones are rare). Conversely, the top SVM features (cf. Table 3) contain hardly topic words but mostly words clearly associated with violence. We hypothesize that topic impacts our violence classifier, but the evidence is not conclusive, warranting deeper analysis.

Class Distribution We see that the classification seems to be effective with F_1 scores ranging from 0.837 to 0.939. While these results are promising, the task is far from solved. Due to the skewed class distribution in the fan fiction corpus (ca. 13% of works are violent; likely more extreme for other genres), a high precision is crucial for a model to be transferable to real-world applications.

Limitations

We believe to have cast a challenging task which cannot be trivially solved using transformer models due to their length limitation; the proposed corpus contributes to both experimental analysis and detection of violence in long documents. We want to outline some known limitations, lest people prematurely consider the problem "solved" when observing our results: First, we only consider *Graphic Depictions of Violence*, whereas AO3 includes other warnings, e.g., *Major Character Death*. The large set of freeform tags suggests potential for more trigger warnings, but this would require annotations external to AO3. Second, although trigger warnings are usually used for documents, it would be interesting to pin-point the potentially triggering content exactly within a document, i.e., using fine-grained annotations of a defined "violence" construct at sentence or paragraph level. Third, the trigger warnings in our corpus were assigned by fan fiction authors and not via principled annotation. While the authors' assessment of their content and warning assignment certainly can be considered ground-truth, the AO3 definition of violence —"[t]he content contains gory, graphic, explicitly described violence"—leaves room for interpretation. Lastly, it is unclear if our negative class indeed never includes violence-related triggers (cf. our Curation Rationale in Appendix B.1). With a working trigger detection approach, relabeling the data by experts will become feasible.

Impact Statement

Note that any automation of trigger warning assignment can be abused to the opposite than the intended effect of trigger warnings, that is, to identify documents with specific triggering content with the goal to target vulnerable individuals.

We refrain from directly publishing the corpus since we do not have explicit permission from the AO3 authors to republish their work. However, since AO3 is publicly accessible, we will release a file with the IDs of works included in our experimental setup, so the splits can be reproduced.

Acknowledgments

This work was partially supported by the European Commission under grant agreement GA 101070014 (OpenWebSearch.eu)

References

Nourah Alswaidan and Mohamed El Bachir Menai. 2020. A survey of state-of-the-art approaches for emotion recognition in text. *Knowledge and Information Systems*, 62(8):2937–2987.

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 4th Workshop on Online Abuse and Harms*, pages 125–137.

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.

Munmun De Choudhury. 2015. Anorexia on Tumblr: A Characterization Study. In *Proceedings of the 5th International Conference on Digital Health 2015*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)*, 51(4):1–30.

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.

Arunima Khunteta and Pardeep Singh. 2021. Emotion cause extraction—a review of various methods and corpora. In *Proceedings of the 2nd International Conference on Secure Cyber Computing and Communications (ICSCCC)*, pages 314–319. IEEE.

Hannah Kirk, Abeba Birhane, Bertie Vidgen, and Leon Derczynski. 2022. Handling and presenting harmful text in nlp research. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 497–510.

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

S. V. Kogilavani, S. Malliga, K. R. Jaiabinaya, M. Malini, and M. Manisha Kokila. 2021. Characterization and mechanical properties of offensive language taxonomy and detection techniques. *Materials Today: Proceedings*.

Sophia Yat Mei Lee, Ying Chen, and Chu-Ren Huang. 2010. A text-driven rule-based system for emotion cause detection. In *Proceedings of the 2010 NAACL-HLT Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pages 45–53.

Pushkar Mishra, Helen Yannakoudakis, and Ekaterina Shutova. 2019. Tackling online abuse: A survey of automated abuse detection methods. *CoRR*, abs/1908.06024.

Hyunji 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)*, pages 702–709, Dublin, Ireland. Association for Computational Linguistics.

Martin F. Porter. 1980. An algorithm for suffix stripping. *Program*, 14(3):130–137.

Anna Schmidt and Michael Wiegand. 2019. A Survey on Hate Speech Detection using Natural Language Processing. In *Proceedings of the 5th International Workshop on Natural Language Processing for Social Media*, pages 1–10.

Thamar Solorio, Mahsa Shafaei, Christos Smailis, Brad J Bushman, Douglas A Gentile, Erica Scharrer, Laura Stockdale, and Ioannis Kakadiaris. 2021. White paper — objectionable online content: What is harmful, to whom, and why. *arXiv preprint arXiv:2104.03903*.

UM list. University of Michigan, An Introduction to Content Warnings and Trigger Warnings [online]. 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: October 10, 2022.

UR list. University of Reading, Guidance on content warnings on course content ('trigger' warnings) [online]. https://www.reading.ac.uk/cqsd/-/media/project/functions/cqsd/documents/qap/trigger-warnings.pdf. Last accessed: October 19, 2022.

Zeerak Waseem, Thomas Davidson, Dana Warmsley, and Ingmar Weber. 2017. Understanding abuse: A typology of abusive language detection subtasks. In *Proceedings of the First Workshop on Abusive Language Online*, pages 78–84.

	Random	Fame	Tag
Rating			
Δ_{Mature}	0.551	0.492	0.537
$\Delta_{\text{Not Rated}}$	0.140	0.211	0.167
Δ_{Explicit}	0.275	0.206	0.231
$\Delta_{\text{Teen+}}$	-0.047	-0.141	-0.058
$\Delta_{\text{All Audiences}}$	-0.790	-0.840	-0.826
Characters			
$\lvert D_i \rvert$	27,320	22,036	28,974
$\Delta_i > 0.75$	193	346	199
$\Delta_i > 0.25$	946	1,154	993
$\Delta_i < -0.25$	173	205	184
$\Delta_i < -0.75$	26	28	22
Most violent	*Original Characters*		(430)
	Original Female Character(s)		(298)
	Original Male Character(s)		(237)
	Harry Potter		(126)
Least violent	*Katsuki Yuuri*		(-54)
	Victor Nikiforov		(-56)
	Sherlock Holmes		(-143)
	Victor Nikiforov		(-148)
Freeform			
$\lvert D_i \rvert$	64,961	80,364	71,767
$\Delta_i > 0.75$	333	504	357
$\Delta_i > 0.25$	922	1268	961
$\Delta_i < -0.25$	252	299	345
$\Delta_i < -0.75$	30	27	41
Most violent	*Angst*		(976)
	Violence		(967)
	Torture		(554)
	Drama		(534)
Least violent	*Fluff*		(-1174)
	Established Relationship		(-365)
	Drabble		(-184)
	Humor		(-155)

Table 4: Differences in the Meta-data frequency between violent and non-violent documents. Shown are the Δ_i as described in Appendix A as well as the absolute distance for the example tags split by ratings, characters (as indicator of fandom and plot points), and freeform tags as content descriptors.

A Figures and Tables

Meta-data (Tag) Differences Between Classes Table 4 shows the effect of topic on classification effectiveness. We list the relative count difference between all works D_i with an *Additional Tag* i (rating, freeform, characters) between violent v and non-violent nv documents defined as:

$$\Delta_i = \frac{|D_i^{\mathrm{v}}| - |D_i^{\mathrm{nv}}|}{|D_i^{\mathrm{v}} \cup D_i^{\mathrm{nv}}|}.$$

A $\Delta_i = 1$ indicates that all occurrences of the tag were assigned to violent documents and $\Delta_i = -1$ indicates the opposite.

B Data Statement

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

B.1 Curation Rationale

Our goal was to extract a trigger warning corpus from an existing resource with imperfect labels. In the original data, we are dealing with false negative, false positive, and even contradictory labels, where a work is labeled as both "Graphic Depictions of Violence" and "No Archive Warnings Apply." However, the corpus should be clearly separable in terms of positive and negative examples. To address this situation, we relied on the existing labels, but filtered the positive and negative classes using a co-occurrence analysis between each tag and "Graphic Depiction of Violence."

B.2 Language Variety

While Archive of our Own (AO3) includes fan fiction in many languages, we discarded all non-English documents. For language detection we used Resiliparse.[4] This language constraint is only for the purpose of this study, the remaining documents are of course relevant for future research.

B.3 Speaker Demographic

AO3 hosts fan fiction works from a variety of authors whose demographics are 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. In summary, 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 were Transgender, Agender, Androgynous, Trans, Neutrois, and 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 "[these limitations] do not make the survey useless". There has been been another census since then.

[4] https://resiliparse.chatnoir.eu v0.13.5

B.4 Annotator Demographic

We used pre-existing labels from AO3 for this corpus. Trigger warnings are assigned by the authors of the respective works. We do not have any additional information about these groups.

B.5 Speech Situation

All of the texts are written works that are or were available online. Each work has a publication date which might reflect the upload date instead of the date of writing, since some works were also posted on other sites before, but backdating is possible.

B.6 Text Characteristics

Almost all texts in our corpus belong to the fan fiction genre. Many fan fiction works revolve (non-exhaustively) around fictional characters from books, cartoons, anime, manga, music, and movies, or non-fictional characters such as celebrities. Aside from that, AO3 includes meta posts (such as the previously mentioned AO3 Census or placeholders which link to other works). They have been filtered by our tag-based filtering.

C Classification Setup and Ablation

All document vectors are subsequently normalized using the L_2 norm. The cost parameter C is set to 0.5, which is weighted for each class inversely proportional to its occurrence in the training set. For BERT, we use a maximum sequence length of 512. We fine-tune for 10 epochs with a learning rate of $2e^{-5}$ and batches of size 32. For Longformer, we use a maximum sequence length of 4,096. We fine-tune for 20 epochs with a learning rate of $2e^{-5}$ and batches of size 4. Hyperparameters were optimized for all models via an exhaustive search, evaluating possible combinations using cross-validation on the training set.

For the SMV, we evaluated the cost parameters C in $\{0.1, 0.2, 0.5, 1.0, 2.0\}$, feature normalization of $\{\mathrm{L}_1, \mathrm{L}_2, \text{none}\}$, n-grams over the range of $\{1, 2, 3\}$, lowercasing the features $\{\text{yes}, \text{no}\}$, using a per-class C-parameter that is inversely balanced to the class distribution $\{\text{yes}, \text{no}\}$, and keeping $\{25K, 50K, 100K, \text{all}\}$ of the features.

For BERT and Longformer, we evaluated the learning rates $\{1e^{-5}, 2e^{-5}, 5e^{-5}, 1e^{-4}\}$ and epochs $\{5, 10, 15\}$, and inversely weighting the loss based on the class distribution.