Overview of the Trigger Detection Task at PAN 2023

Matti Wiegmann[1], Magdalena Wolska[1], Martin Potthast[2,3] and Benno Stein[1]

[1]*Bauhaus-Universität Weimar, Weimar, Germany*
[2]*Leipzig University, Leipzig, Germany*
[3]*ScaDS.AI, Leipzig, Germany*
pan@webis.de *https://pan.webis.de*

Abstract

Trigger warnings are document labels that warn the reader about content that might cause discomfort or distress. These labels are often asked for by online communities, especially by vulnerable groups. Here, we present trigger detection at PAN 2023 as a multi-label document classification task: *Given a fan fiction document, assign all appropriate trigger warnings from the given label set.* We derive a set of 32 trigger warnings based on two widely referenced institutional guidelines on sensitive content. We compile a 341,000 document evaluation resource, fan fiction documents from Archive of our Own (AO3), which we fully annotated with the 32 trigger warnings. Six participants submitted solutions to the task. The submissions cover several different methods; the most effective submissions use hierarchical deep learning with RoBERTa-based encodings. The top approach achieves a macro F_1 of 0.35 and a micro F_1 of 0.75.

1. Introduction

A trigger in psychology is a stimulus that elicits negative emotions or feelings of distress. In general, triggers include a broad range of stimuli—such as smells, tastes, sounds, textures, or sights—which may relate to or evoke memories of possibly distressing acts or events. To proactively apprise the audience that a piece of media (writing, audio, video, etc.) contains potentially distressing material, the use of "trigger warnings"—labels indicating the type of potentially triggering content present—has become common in institutionalized education but also in online communities [1], making it possible for sensitive audience to prepare themselves for the content and better manage their reactions. Particularly online communities have expanded the trauma-related concept of trigger warnings used in psychology to many more types of potentially distressing content like eating disorders, discrimination, suicide, abuse, or pornography.

Fiction in particular can make its readers susceptible to triggers, as it often serves "escaping" reality for a while by identifying oneself with the characters in a story and experiencing their fate with particular intensity. This may partly explain why the community of the fan fiction site Archive of our Own (AO3) 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-provided warnings.

CLEF 2023: Conference and Labs of the Evaluation Forum, September 18–21, 2023, Thessaloniki, Greece

matti.wiegmann@uni-weimar.de (M. Wiegmann); magdalena.wolska@uni-weimar.de (M. Wolska); martin.potthast@uni-leipzig.de (M. Potthast); benno.stein@uni-weimar.de (B. Stein)
0000-0002-3911-0456 (M. Wiegmann); 0000-0003-2451-0665 (M. Potthast); 0000-0001-9033-2217 (B. Stein)

In this pilot edition of the Trigger Detection task at PAN 2023, we establish the computational problem of identifying whether or not a given document contains triggering content. In particular, we formalize trigger detection as a multi-label document classification (MLC) task as follows:

Given a fan fiction document, assign all appropriate trigger warnings from the given label set.

We created a new evaluation resource, *PAN23-trigger-detection*, containing ca. 340,000 fan fiction works from Archive of our Own (AO3) annotated with a 32-label trigger warning set. We rely on user-generated labels (authors assigned warning-like labels) and follow the authors' understanding of triggers and which documents require a warning. The warnings are assigned via AO3's freeform content descriptors ("tags"), a custom, high-dimensional label system. Since tags include also non-warning descriptors, we developed a distant-supervision strategy to detect if a freeform tag corresponds to one of the 32 predefined warnings compiled from institutional content warning guidelines. The task is primarily evaluated with the standard measures for multi-label classification, micro and macro F_1. In total, 6 participants submitted software to Trigger Detection 2023.

This overview paper first details the creation of the evaluation resource (Section 2), in particular the distillation of the warning label set from two institutional content guidelines, the scraping of AO3, the distant-supervision labeling, and the curation of the works. Furthermore, the evaluation procedure (metrics and baselines) are described in Section 3, the 6 participant submissions are described in Section 4, and the results are discussed in Section 5.[1,2]

2. Data

For the trigger detection task, we created a new evaluation resource, the *PAN23-trigger-detection* corpus, consisting of 341,246 fan fiction works downloaded from Archive of our Own and annotated in a multi-label setting with a set of 32 warning labels. An extended version of our annotation method and the evaluation resource is presented by Wiegmann et al. [2].

2.1. Curating a Set of Warning Labels

Since there is no authoritative (closed-set) set of trigger warning labels, we derived these labels for use in our dataset from two guideline documents for labeling sensitive content: the University of Reading list of "themes that require trigger warnings" [3] and the University of Michigan list of content warnings [4]. The two largely overlapping lists comprise, each, 21 categories of triggering concepts, including health-related (*eating disorders, mental illness*), sexually-oriented (*sexual assault, pornography*) as well as verbal (*hate speech, racial slurs*), and physical abuse (*animal cruelty, blood, suicide*). The lists were pre-processed to unfold compound categories into individual elements (e.g. "Animal cruelty or animal death" → "animal cruelty", "animal death") and lower-cased. Table 9 (see Appendix) shows the aligned source labels and the merged set. This merged set of warnings comprises 35 categories; we removed the rarest three labels since there were too few annotated documents with those labels in the final dataset. From them, we derived the 32 label trigger warning set for the PAN 2023 Trigger Detection task (see Table 1).

[1] The baseline and evaluation code used for this task is available at github.com/pan-webis-de/pan-code.

[2] The data used are available at zenodo.org/record/7612628.

Table 1
The complete trigger warning set used in PAN23-trigger-detection grouped into 7 more general, semantically coherent groups. The italic part of the definitions are examples of AO3's freeform tags that match the respective warning label.

Trigger domain	Definition and Example Tags
Discrimination/Prejudice-related	
ableism	Discrimination against disabled persons, *Ableist Language*
classism	Discrimination based on social class, *Rich/Poor Divide, Class Oppression*
homophobia	Discrimination against homosexuals, *Homophobic Language, Gay Panic*
misogyny	Discrimination/hate against women, *Misogynistic Language*
racism	Discrimination based on race (including fantasy races), *Fantastic Prejudice*
sexism	Discrimination based on gender stereotypes, *Misgendering, Deadnaming*
transphobia	Discrimination against transgender persons
Hostile Acts/Aggression-related	
violence	Physical violence, *Manhandling, Torture, Murder*
animal-cruelty	Violence/Harm against animals, *Animal Abuse, Animal Mistreatment*
sexual-assault	Physical, sexual Violence, *Rape, Sexual Abuse, Non-consensual Actions*
abuse	*Domestic Violence, Verbal Abuse, Psychological Abuse, Bullying*
child-abuse	Like abuse, but explicitly directed against children
abduction	Abduction by deception, often non-violent, *Stockholm Syndrom*
kidnapping	Violent kidnapping, hostage situations, *Captivity*
Pregnancy-related	
pregnancy	(Issues of) being pregnant, *Male Pregnancy, Fertility Issues*
miscarriages	(Aftermath of) Miscarriages, *Child Loss*
childbirth	Act of Giving birth
abortion	Termination of pregnancy, including non-voluntary
Anatomy-related	
dissection	Dissection of body parts, *Mutiliation, Body Horror, Surgery, Loosing Limbs*
blood	*Blood, Gore, Wounds*
Death-related	
dying	The process of dying from the subject's perspective, *Drowning, Euthanasia*
death	Death of others *Character death, Killing, Corpses, Coping with Loss or Grief*
animal-death	Death of animals
Mental Health-related	
mental-illness	Severe mental illness, *Hallucinations, Dissociative Identity Disorder, Insanity*
suicide	Suicide attempt, ideation, conduct, and aftermath
eating-disorders	*Anorexia, Bulimia, Self-starvation, Binge Eating*
fat-phobia	*Obesity, Fat-Shaming, Weight (Loss) Issues*
body-hatred	Body and gender dysmorphia
self-harm	Self-destructive acts or behavior, *Self-mutiliation*
Sexuality-related	
incest	Sex between family members, *Sibling Incest, Twincest*
underage	Sex with a minor, consensual and non-consensual, *Pedophilia*
pornographic-content	Graphic display of sex, plays, toys, technique descriptions

Three major observations can be made of the merged university label set (Table 9): 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* can be considered types of prejudice). Third, the prescribed list is not exhaustive, as is also pointed out on both websites.

Table 2
Sizes of the works in the corpus of works with various properties relevant for sampling datasets.

Corpus Size and properties	
Words	58B
Total works	7.9M
with warnings	2.6M
Non-English Language	751K

Corpus Size and properties	
Release Date pre-2009	246K
More than 1 chapter	1.9M
More than 6k words	1.8M
Duplicates	8K

Corpus Size and properties	
Less than 3 tags	2.3M
More than 66 tags	8K
Up to 90% annotated works	3.7M
Less than 10 kudos	1.3M
Less than 1000 hits	4.5M

Table 3
Number of AO3 freeform tags that can be annotated with a trigger warning by different methods. `0-2k` and `10-11k` contain manually annotated tags, `tag graph` contains tags annotated via distant supervision, and `combined` contains `0-2k` and `tag graph`.

Set	**No. tags in set (% of all)**		**TWs (% of set)**	**Mapping evaluation**			
	Tag occurrence	Unique tags		Prec	Rec	F_1	Acc
`0-2k`	27.7M (52.14)	2K (0.02)	463 (22.92)	0.95	0.94	0.94	0.94
`10-11k`	0.3M (0.57)	1K (0.01)	85 (8.47)	0.98	0.97	0.97	0.97
`tag graph`	42.2M (79.51)	1.4M (14.55)	116K (8.19)	–	–	–	–
`combined`	44.3M (83.35)	1.4M (14.55)	116K (8.19)	–	–	–	–
All tags	53.1M (100.00)	9.7M (100.00)	–	–	–	–	–

To obtain labels that abstract over the inconsistent granularity (Table 1), that are orthogonal in terms of semantics, and to better inform later annotation decisions, we grouped the original labels into semantically related subsets. The grouping was done by identifying semantic fields, *trigger domains*, with which the triggering concepts can be associated via some semantic relation, for instance, *is-a* or *results-in*; since the label set is sufficiently small, grouping was done manually. Technically speaking, for warnings formulated as complex nouns, we first identified the semantic content-bearing lexeme and used that as the basis for grouping. For most complex nouns, the head noun was used; the label "pornographic content" is an example of an exception in the case of which the content-bearing adjective was used to identify its semantic domain.

2.2. Acquiring the Source Documents

Table 2 shows the descriptive statistics of our source data: ca. 8 million works of fan fiction from Archive of our Own. We initially downloaded all works released between August 13, 2008 (the platform launch) and August 09, 2021, from `archiveofourown.org` and extracted the document text and metadata (i.e., the freeform tags) from the scraped HTML. To download the HTML page of each work, we scraped the output of the search function to get the work ID and then constructed a direct URL to that work's page. Since the search function was limited to 10,000 works per page, we constructed queries to search for all works released on one particular day, for each day in the release window.

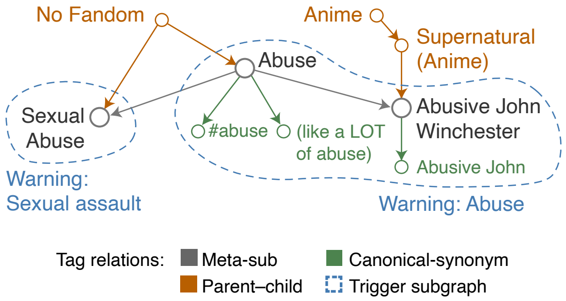

Figure 1: Schematic display of the properties of the tag graph and how they were used to infer trigger labels from the additional tags with minimal manual annotation.

2.3. Assigning Trigger Warnings to Source Documents

We labeled all works in the source data via distant supervision based on the freeform tags assigned to the works by their author(s). The results and evaluation are shown in Table 3. We first identify the freeform tags that also indicate a warning from our 32-label set and, second, we assign this warning to all works labeled with the indicative freeform tag. The underlying mapping table, which maps from freeform tag to trigger warning, was created by (i) manually annotating the 2,000 most common tags, (ii) efficiently identifying sub-structures of the tag graph that indicate a trigger warning, annotating each node in the structure with that warning, and (iii) merging both results, giving priority to the manual annotations.

We manually annotated two sets of freeform 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`) as an evaluation dataset. All tags were annotated by two annotators; diverging annotations were merged by critical discussion. Then, the sub-structures of the tag graph that indicate the same trigger warning across all nodes were identified (cf. Figure 1) by extracting and manually annotating rooted sub-graphs from the tag graph in a 5-stage process:

1. 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 arcs.
2. 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 to 4.
3. 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 the 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 ca. 5,000 tags.
4. Identification of trigger graph sources: manual annotation of all candidate sources, discarding the nodes without a trigger warning.
5. Identification of all trigger graphs: manual depth-first traversal of the tag graph along the meta-sub relation, starting from a trigger graph source. If a successor does not agree with the trigger warning assigned to its predecessor, the arc between them is removed, and the successor added as new trigger graph source to be annotated with a new trigger warning.

Table 4
Descriptive statistics of the training, validation, and test split of the dataset.

Training Dataset		**Validation Dataset**		**Test Dataset**	
Total Works	307,102	Total Works	17,104	Total Works	17,040
< 512 words	15,233	< 512 words	861	< 512 words	813
< 4,096 words	261,156	< 4,096 words	14,571	< 4,096 words	14,555
Mean no. words	2,400	Mean no. words	2,386	Mean no. words	2,388
Median no. words	2,126	Median no. words	2,115	Median no. words	2,101
90pct no. words	4,579	90pct no. words	4,550	90pct no. words	4,558

Table 3 shows that we can annotate 52% of all freeform tag occurrences manually with high reliability. With our method, we can completely annotate more than half of all works in the corpus. The other half of the works are only partially annotated since our method only annotates 15% of the unique tags. Tags are only wrangled (i.e., added to the tag graph) if they occur thrice. Since 89.9% of unique freeform tags occur only once, our method misses them. We evaluate the effectiveness of our annotation approach 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 3, our approach achieves both 0.95 accuracy and F_1.

2.4. Sampling the Evaluation Dataset

From the resulting collection of annotated fan fiction works, we sampled *PAN23-trigger-detection* by discarding all works that had no warning assigned, were originally published pre-2009 (as opposed to posted after that since AO3 also archives works from older fan fiction sites), had freeform tags that could not clearly mapped, was not in English (ca. 8% of the works), had less than 50 or more than 6,000 words (outliers; ease of computation), less than 2 or more than 66 freeform tags (confidence threshold), less than 1,000 hits (views), or, less than 10 kudos (likes; popularity threshold). We also removed all (near) duplicates. The resulting dataset contains 341,246 fan fiction works, which was split with stratified sampling into 90:5:5 training, validation, and test sets; i.e., we kept the label distribution equal across the three splits.

2.5. Properties of the Evaluation Dataset

Table 4 shows the descriptive statistics of the dataset splits. The training dataset with ca. 300,000 works is large enough to train deep neural classifiers. The datasets contain ca. 5% very short documents (<512 words) that can be used by a BERT-based system without truncation and ca. 85% medium-sized documents (<4,096 words) that can be used by a sparse-attention model. Figure 2 shows the distribution of the labels over the test dataset. The most frequent label is *pornography* and occurs in ca. 77% of the documents. Most labels are less common, between ca. 10% for *sexual-assault* and 6e-4% for *animal-cruelty*. Documents have 1–13 labels per document, ca. 71% with a single label, 20% with two, and 6% with three.

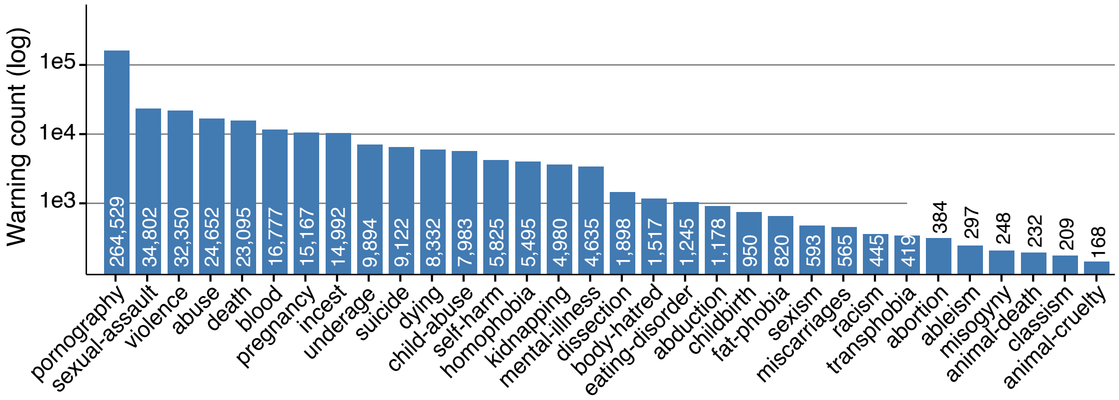

Figure 2: Distribution of the 32 classes in the *PAN23-trigger-detection* dataset.

3. Evaluation and Baselines

We evaluate the submissions primarily through the established multi-label classification metrics F_1 and Accuracy (primary metrics). In addition, we also evaluate the effectiveness of individual labels/label groups (extended metrics) and the effectiveness in relation to document metadata. Lastly, we construct and evaluate voting-based ensembles from the submissions.

As primary metrics, we compare precision, recall, and F_1 at both micro- and macro average, and subset accuracy, which measures accuracy on a per-sample basis (i.e., if all labels of one example are set correctly). In our assessments, we favor the macro over the micro F_1 scores due to the label imbalance. We also favor recall over precision, since we consider trigger warning assignment a high-recall task where false negatives cause more harm than false positives. However, we opted not to modify the metrics or their parameters to reflect this preference.

As extended metrics, we compare precision and recall of *pornography* (due to its frequency), the average effectiveness of the 15 next-most common labels (*sexual-assault–dissection*), and the average effectiveness of the 16 least common labels. We also compute the number of classes with either zero or a very low (<0.1) precision and recall to check for high-frequency label bias.

As metadata-based metrics, we compare micro and macro F_1 for the document subsets that fall within certain metadata thresholds. First, we compare short (<500), medium, and long (>4,000) documents. We assume that short works are easier to classify since models can capitalize more directly on BERT (which has a short input size). Second, we compare works with few (<5), medium, and many (>20) freeform tags. We assume that works with many freeform tags are easier to classify because many tags suggest that authors took greater care with annotating their works and the resulting higher label quality leads to better effectiveness. Third, we compare works with low (<50 comments, <60 bookmarks, >450 kudos, >8,500 hits), medium, and high (>280 comments, >330 bookmarks, >1,850 kudos, >35,000 hits) popularity. We assume that works with high popularity are also easier to classify because authors are more diligent when tagging works that gain much attention. Fourth, we compare works with an archive warning (*Graphic Depictions Of Violence, Major Character Death, Rape/Non-Con, Underage*), without warning (*No Archive Warnings Apply*), and works that do not specify the warnings (*Choose Not*

Table 5
Overview of the submitted methods. Listed is the (dominant) model architecture, the feature representation, the method used to handle long documents, and the sampling strategy used to handle the skewed label distribution.

Participant	Model	Features	Length	Imbalance
Sahin et al. [6]	RoBERTa + LSTM	CLS embedding	Hierarchical cls.	Weighted loss
Su et al. [7]	RoBERTa + CNN	Context embeddings with 1D convolution and mean-pooling	Hierarchical cls.	–
XGBoost baseline	XGBoost	TF·IDF	Document features	Undersampling
Cao et al. [8]	RoBERTa	CLS token	Voting	Over- and under-sampling
Cao et al. [9]	RoBERTa	CLS token	Voting	Over- and under-sampling + separate classifiers
Felser et al. [10]	MLP	Aggregate word emb. + topic model	Document features	Weighted loss
Lakshmaiah et al. [11]	LSTM	GloVE	(LSTM)	–

To Use Archive Warnings). We assume that works with a warning are easier to classify and works without specified warning are the hardest, since authors hide warning tags within spoilers and might therefore less diligently annotate freeform warnings. Fifth, we compare works with an *Explicit* or *Mature* rating to works with neither. We assume that explicit or mature works contain more markers and are thus easier to classify.

Finally, we construct four ensembles from the submitted results, where the assignment of a true label is decided by voting to surpass a threshold τ. The *Top-3* ensemble uses the three best submissions with $\tau = 2$, the other ensembles use all submissions with $\tau = \{3, 5, 7\}$.

As a baseline, we trained an XGBoost [5] classifier based on word-1–3-gram features encoded as TF·IDF document vectors with a minimum document frequency of 5. We used only the top 10,000 features according to a χ^2 feature selection. The dataset was undersampled uniformly at random to 1,000 samples per label. As parameters, we used a `max depth` or 3, a `learning rate` of 0.25, and 300 estimators with 10-round early stopping. The features word-1, 2, and 3-grams and character-3 and 5-grams were evaluated, as well as feature selection (with or without), model parameters, and the thresholds for over- and undersampling via grid search.

4. Submissions

The 6 submissions to the PAN 2023 Trigger Detection task employed a broad set of techniques, from hierarchical transformer structures to strategic feature engineering. Table 5 shows an overview of the different strategies used by the participants. All participants used a form of a neural network as a model, where RoBERTa was most common and most successful as a classifier or pre-trained model to produce a strong input encoding. Most submissions also focused on improving the long document aspect of the task (most documents are longer than the input size of the state-of-the-art classification models) by using hierarchical classifiers (chunks are encoded, and prediction is based on a combination of encodings), or voting-based approaches (chunks are labeled individually, document labels are aggregated over chunk labels). The submissions cope with the label imbalance (the most common label (*pornography*) is an order of magnitude more common than the other labels) through over- and undersampling or by changing class-weights in the loss function, so that misclassifying a rare class increases the error more than a common label.

Sahin et al. [6] submitted a hierarchical transformer architecture that achieved the top macro F_1 score (by a slim margin of 0.002) and came in second in micro F_1 and accuracy, while having a relatively high recall within the top approaches. The approach first segments the document into chunks (200 words with 50 words overlap) and then pre-trains a RoBERTa transformer on the chunks to learn the genre. The architecture then embeds all chunks of a document using the pre-trained transformer, followed by an LSTM for each label (in a one-vs-all setting), predicting the class from a sequence of chunk-embeddings (RoBERTa's `[CLS]` token). To cope with label imbalance, the approach assigns positive weights in the loss function to the rare half of the labels.

Su et al. [7] submitted a siamese transformer that achieves the second-best macro F_1 score (by a slim margin of 0.002) and the top scores in micro F_1 and accuracy, while notably favoring precision over recall. The approach segments the documents into 505-word chunks, encodes the first and last chunk using RoBERTa, mean-pools the contextual embeddings (ignoring the `[CLS]` token), and classifies based on the pooled embeddings using a 1D convolutional neural network.

Cao et al. [8] submitted a voting-based transformer that favors recall over precision. The approach segments the training documents into chunks, assigns each chunk the labels from its source document, and trains a single RoBERTa-based classifier on each chunk. To make predictions, the documents are again chunked, the labels for each chunk are predicted, and a label is assigned to the document if it is assigned to more than half of the chunks. The training data was dynamically over- and undersampled: pornography was undersampled to 5,000 examples and other labels to 2,000 examples. Examples with rare labels were replicated 8-10 times.

Cao et al. [9] also submitted a voting-based transformer that achieved very balanced results, neither favoring macro over micro scores nor precision over recall. The approach chunks and votes similarly to Cao et al. [8] but builds two different models to overcome the data imbalance, one for pornography and one for the other 31 classes. The pornography model was trained on a random selection of 40,000 works with and 40,000 works without the pornography warning. The model for the other labels removes works with only the pornography warning, undersamples frequent classes to 3,000 examples, and oversamples rare labels by replicating works 4-6 times.

Felser et al. [10] submitted a 1-vs-rest multi-layer perceptron based on two features: fasttext-based document embeddings and superclass probabilities. This approach achieved the top micro and macro recall, at the cost of precision on the test dataset. Document embeddings were created by training a fasttext model from the training data, generating the embeddings for each unique word in a document, scaling them by term frequency, and adding and normalizing the scaled word vectors over the document. The superclass probabilities were determined by grouping the 32 labels semantically into 6 superclasses, bootstrapping a seeded LDA with the 50 most relevant bi-grams of each group (determined through a TF·IDF-like approach for n-gram weighting, which downgrades pornographic terms), and training a classifier to predict the superclass based on the topic model outputs, using class probabilities as features. Label imbalance was addressed via class penalties in the loss function, where the MLP-2 variant has a higher penalty.

Lastly, **Lakshmaiah et al.** [11] present an LSTM-based approach using GloVE-embeddings, which is third in micro F_1 with very high precision but rather weak in macro average scores.

Table 6
Participant scores of the trigger detection task at PAN 2023. Shown are the core metrics, sorted by macro F_1. Bold indicates the leading approach for each metric. Scores of the voting-based ensembles are bold when they are better than the leading submission.

Participant	**Macro**			**Micro**			**Acc**
	Prec	Rec	F_1	Prec	Rec	F_1	
Sahin et al. [6]	0.37	0.42	**0.352**	0.73	0.74	0.74	0.59
Su et al. [7]	**0.54**	0.30	0.350	0.80	0.71	**0.75**	**0.62**
XGBoost baseline	0.52	0.25	0.301	0.88	0.57	0.69	0.53
Cao H. et al. [8]	0.24	0.29	0.228	0.43	0.79	0.56	0.18
Cao G. et al. [9]	0.28	0.22	0.225	0.58	0.66	0.62	0.32
Felser et al. [10]	0.11	**0.63**	0.161	0.27	**0.82**	0.40	0.27
Lakshmaiah et al. [11]	0.10	0.04	0.048	0.82	0.50	0.63	0.52
Ensemble (Top 3)	**0.56**	0.30	0.36	0.88	0.68	**0.77**	**0.63**
Ensemble ($\tau = 3$)	0.38	0.42	**0.37**	0.65	0.80	0.72	0.52
Ensemble ($\tau = 5$)	0.55	0.20	0.26	0.88	0.65	0.75	0.60
Ensemble ($\tau = 7$)	0.39	0.07	0.10	**0.97**	0.50	0.66	0.53

Table 7
Participant scores of the trigger detection task at PAN 2023. Shown are the extended metrics: precision and recall for *Pornography*, the more common half of labels excluding pornography (**Mid**) and the rare half (**Bot**) as well as the number of classes where precision and recall is either **zero** or below **0.1**. Participants are sorted by total macro F_1 (cf. Table 6).

Participant	**Porn.**		**Mid**		**Bot**		**Zero P/R**		**<0.1 P/R**	
	Prec	Rec	Prec	Rec	Prec	Rec	Prec	Rec	Prec	Rec
Sahin et al. [6]	0.95	0.96	**0.62**	0.48	0.12	0.51	3	3	10	6
Su et al. [7]	0.90	0.97	0.61	0.43	**0.57**	0.19	6	6	6	9
XGBoost baseline	0.98	0.87	0.62	0.21	0.38	0.24	2	2	2	9
Cao H. et al. [8]	0.86	**0.98**	0.22	0.61	0.16	0.12	5	5	7	13
Cao G. et al. [9]	0.97	0.88	0.29	0.42	0.24	0.09	4	4	8	15
Felser et al. [10] (MLP1)	0.97	0.91	0.18	**0.72**	0.03	**0.64**	7	5	24	**5**
Felser et al. [10] (MLP2)	0.97	0.91	0.26	0.45	0.03	0.31	13	13	22	13
Lakshmaiah et al. [11]	0.93	0.91	0.18	0.04	0.00	0.00	23	24	25	30
Ensemble (Top 3)	0.96	0.96	**0.72**	0.38	0.50	0.25	5	5	5	7
Ensemble ($\tau = 3$)	0.94	0.97	0.43	0.61	0.28	0.36	4	4	4	6
Ensemble ($\tau = 5$)	0.97	0.93	0.68	0.33	0.76	0.10	9	9	9	16
Ensemble ($\tau = 7$)	**0.98**	0.87	0.82	0.08	**0.91**	0.02	18	20	18	25

5. Results

Table 6 shows the evaluation results for the primary metrics as discussed in Section 3, ordered by macro F_1. Here, the hierarchical classifiers are the most effective by a large margin, followed by the XGBoost baseline. The most effective approach by macro F_1 is the one by Sahin et al. with 0.352, a small margin before that of Su et al. with 0.350. The best approach by micro F_1 and subset accuracy is the one by Su et al.. The XGBoost baseline is only beaten by these two top approaches. The models score very differently in precision and recall, depending on the architecture. Four models score generally higher in recall, the other 4 in precision. There is no obvious relationship between effectiveness and preference for precision or recall. The ensembles (top 3 and $\tau = 3$) beat the submissions but by a very small margin of ca. 0.02.

Table 8

Participant scores of the trigger detection task at PAN 2023. Shown are scores of examples with certain properties based on different document lengths, number of freeform tags (tag confidence), popularity confidence (hits, kudos, comments, bookmarks), works with, without, and with unspecified AO3 *archive warning*, and works with or without and explicit or mature rating. Participants are sorted by total macro F_1 (cf. Table 6).

Participant Team	**Length**			**Tag count**			**Popularity**			**AO3 Warning**			**Explicit**	
	Short	Med.	Long	Few	Med.	Many	Low	Med.	High	w/	w/o	Unk.	Yes	No
Macro F_1														
Sahin et al. [6]	0.28	0.35	**0.34**	0.36	**0.35**	**0.30**	**0.30**	**0.35**	**0.35**	**0.35**	**0.35**	**0.31**	**0.31**	0.37
Su et al. [7]	**0.39**	**0.36**	0.27	**0.37**	0.34	0.25	0.22	0.33	**0.35**	**0.35**	**0.35**	0.29	0.28	**0.38**
XGBoost baseline	0.24	0.30	0.29	0.31	0.30	0.22	0.16	0.28	0.30	0.30	0.30	0.25	0.28	0.30
Cao H. et al. [8]	0.23	0.23	0.22	0.21	0.23	0.23	0.19	0.25	0.22	0.24	0.19	0.24	0.21	0.23
Cao G. et al. [9]	0.22	0.23	0.18	0.23	0.22	0.19	0.20	0.25	0.22	0.23	0.21	0.20	0.18	0.25
Felser et al. [10] (MLP1)	0.13	0.16	0.17	0.14	0.17	0.20	0.17	0.16	0.16	0.16	0.14	0.18	0.16	0.15
Felser et al. [10] (MLP2)	0.12	0.15	0.16	0.14	0.16	0.17	0.15	0.15	0.15	0.15	0.14	0.16	0.15	0.10
Lakshmaiah et al. [11]	0.04	0.05	0.04	0.05	0.05	0.04	0.04	0.04	0.05	0.04	0.04	0.05	0.05	0.04
Ensemble (Top 3)	0.32	**0.37**	0.33	**0.41**	0.35	0.25	0.23	**0.36**	0.36	0.36	**0.35**	0.31	0.31	**0.39**
Ensemble ($\tau = 3$)	0.31	**0.37**	**0.36**	0.38	**0.37**	**0.33**	0.25	0.35	**0.37**	**0.37**	**0.35**	**0.35**	**0.34**	0.38
Ensemble ($\tau = 5$)	0.27	0.27	0.21	0.29	0.25	0.18	0.20	0.28	0.26	0.26	0.25	0.21	0.23	0.27
Ensemble ($\tau = 7$)	0.09	0.10	0.07	0.10	0.10	0.07	0.08	0.08	0.10	0.10	0.10	0.08	0.09	0.10
Mirco F_1														
Sahin et al. [6]	0.73	0.74	**0.72**	0.75	0.74	**0.66**	0.76	0.75	0.73	0.74	0.79	0.62	0.79	**0.59**
Su et al. [7]	**0.76**	**0.76**	**0.72**	**0.77**	**0.75**	**0.66**	**0.77**	**0.77**	**0.75**	**0.76**	**0.80**	**0.64**	**0.81**	**0.59**
XGBoost baseline	0.58	0.69	0.70	0.72	0.68	0.59	0.70	0.72	0.68	0.70	0.76	0.52	0.77	0.41
Cao H. et al. [8]	0.52	0.56	0.57	0.54	0.57	0.58	0.63	0.58	0.55	0.55	0.56	0.56	0.60	0.45
Cao G. et al. [9]	0.58	0.62	0.61	0.61	0.62	0.59	0.69	0.65	0.61	0.62	0.63	0.56	0.66	0.47
Felser et al. [10] (MLP1)	0.31	0.40	0.42	0.38	0.41	0.44	0.43	0.43	0.40	0.39	0.42	0.38	0.50	0.25
Felser et al. [10] (MLP2)	0.45	0.54	0.56	0.54	0.54	0.52	0.57	0.57	0.53	0.54	0.59	0.44	0.66	0.31
Lakshmaiah et al. [11]	0.60	0.63	0.61	0.67	0.61	0.50	0.58	0.64	0.62	0.63	0.72	0.39	0.73	0.28
Ensemble (Top 3)	**0.78**	**0.77**	**0.75**	**0.80**	**0.76**	0.66	**0.78**	**0.78**	**0.77**	**0.77**	**0.82**	**0.64**	**0.82**	**0.61**
Ensemble ($\tau = 3$)	0.68	0.72	0.72	0.72	0.72	**0.68**	0.76	0.74	0.71	0.71	0.76	**0.64**	0.77	0.58
Ensemble ($\tau = 5$)	0.74	0.75	0.73	0.78	0.74	0.64	0.76	0.76	0.74	0.75	0.80	0.61	0.81	0.55
Ensemble ($\tau = 7$)	0.62	0.67	0.65	0.71	0.65	0.54	0.64	0.68	0.66	0.67	0.75	0.44	0.75	0.31

Table 7 shows the evaluation results for the extended metrics. Unsurprisingly, all submissions score very high on *pornography* and notably lower on all rare labels, which explains the difference between macro and micro F_1. There is a clear decrease in efficiency with decreasing label frequency. It also becomes more obvious that models tend to be good in either precision or recall with large differences between them. Combining the strength of the high-recall and high-precision approaches is a potential way forward, albeit our basic ensemble exploits that only marginally.

Table 8 shows the evaluation results based on document subsets with common metadata values. Regarding the document length, the macro F_1 scores are mixed: Models that use the complete work as single examples during training (Sahin et al. [6], the baseline, and Felser et al. [10]) are slightly (0.05–0.1) less effective on short texts; models that use only a section of the document (Su et al. [7], Cao, G. et al. [9]) are slightly (0.05–1.0) less effective on long texts. On micro F_1, all models tend to perform worse on shorter texts. This contradicts our assumption (and prior evidence [2]) that models will be generally better on short texts which can fully capitalize on BERTs strength on short inputs. An alternative hypothesis is that shorter documents are simply less clear and have fewer of the markers that the classifier expects to make a positive prediction. Regarding the tag count, the top models are slightly (0–0.1) less effective when there are many freeform tags. There is no difference between the less effective models. This also contradicts our assumption that models with many tags are easier to classify due to higher label reliability. Regarding popularity, there is no notable difference in micro F_1. On macro F_1, models are slightly (0.04–0.14) more efficient on high popularity works than on low popularity works. This agrees with our assumption that labels of popular works are more reliable. Regarding the archive warnings, there is no notable difference between works with or without warnings. However, the most effective models are slightly (ca. 0.05 macro, ca. 0.15 micro) less effective on works with undeclared warnings than on others. This agrees with our assumption that these works are less diligently tagged by their authors (potentially as a spoiler tag). Lastly, regarding the rating, models are more (ca. 0.2–0.3 micro F_1) effective on explicit works, which is likely an artifact from the very effective classification of the *pornography* label. On macro F_1, contrary to the micro score, the submissions are slightly (0–0.1) less effective on explicit works. This also contradicts our assumptions that explicit or mature works are easier to classify.

6. Discussion and Conclusion

We present the first task on trigger detection at PAN 2023, for which we created a 341,000 document evaluation resource of fan fiction works annotated with up to 32 labels in a multi-label classification setting. We extensively evaluate the results of six participant submissions. The most effective submissions score 0.35 on macro F_1 and 0.75 on micro F_1.

We find several factors that impact the effectiveness of the submissions. First, we find that encoding and training on the full documents is important for good scores on long documents and hierarchical models appear to be best in this regard. We assume that it is central to find triggering passages that only appear in some parts of the document and that inform the classification decision, instead of finding the topic or style that is also present in the beginning. Surprisingly, short documents appear to be much harder to classify, so models with a strong encoding for short texts (BERT) are important and document vectors are less effective as features. None of the top models manage to be great at both, short and long-document effectiveness, leaving potential for improvement. The effect sizes on all metadata comparisons are small (ca. 0.05–0.15).

Second, we find that all submissions are much less effective on rare labels and very effective on very common labels. We assume that the triggering concept goes beyond what can be observed from the passages in the training data, hence the models can not connect the triggers in the test data to the learned concept.

Third, we find that the submissions are more effective on popular works and less effective on works with an *Choose Not To Use Archive Warnings* declaration. We assume that authors' diligence in annotating freeform tags varies a lot, so some works are under-tagged (i.e. authors want to avoid spoilers) and authors are more diligent in assigning warnings for popular works. However, we also find that the submissions are less effective on works with many freeform tags, so the reverse assumption (over-tagging decreases label reliability) also has some merit.

References

[1] E. Knox, Trigger Warnings: History, Theory, Context, Rowman & Littlefield, 2017.

[2] M. Wiegmann, M. Wolska, C. Schröder, O. Borchardt, B. Stein, M. Potthast, Trigger Warning Assignment as a Multi-Label Document Classification Problem, in: Proceedings of the 61th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Toronto, Canada, 2023.

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

[4] University of Michigan, An Introduction to Content Warnings and Trigger Warnings, 2023. URL: 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.

[5] T. Chen, C. Guestrin, XGBoost: A scalable tree boosting system, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16, ACM, New York, NY, USA, 2016, pp. 785–794. URL: http://doi.acm.org/10.1145/2939672.2939785. doi:10.1145/2939672.2939785.

[6] U. Sahin, I. E. Kucukkaya, C. Toraman, ARC-NLP at PAN 2023: Hierarchical Long Text Classification for Trigger Detection , in: M. Aliannejadi, G. Faggioli, N. Ferro, M. Vlachos (Eds.), Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum, CEUR-WS.org, 2023.

[7] Y. Su, Y. Han, H. Qi, Siamese Networks in Trigger Detection task , in: M. Aliannejadi, G. Faggioli, N. Ferro, M. Vlachos (Eds.), Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum, CEUR-WS.org, 2023.

[8] H. Cao, Z. Han, G. Cao, R. Zhu, Y. Liang, S. Liu, M. Huang, Trigger Warning Labeling with RoBERTa and Resampling for Distressing Content Detection, in: M. Aliannejadi, G. Faggioli, N. Ferro, M. Vlachos (Eds.), Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum, CEUR-WS.org, 2023.

[9] G. Cao, Z. Han, H. Cao, X. Huang, Z. Zeng, Y. Tan, J. Cai, A dual-model classification method based on RoBERTa for Trigger Detection, in: M. Aliannejadi, G. Faggioli, N. Ferro, M. Vlachos (Eds.), Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum, CEUR-WS.org, 2023.

[10] J. Felser, C. Demus, D. Labudde, M. Spranger, FoSIL at PAN?23: Trigger Detection with a Two Stage Topic Classifier, in: M. Aliannejadi, G. Faggioli, N. Ferro, M. Vlachos (Eds.), Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum, CEUR-WS.org, 2023.

[11] S. H. Lakshmaiah, A. Hegde, F. Balouchzahi, Trigger Detection in Social Media Text , in: M. Aliannejadi, G. Faggioli, N. Ferro, M. Vlachos (Eds.), Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum, CEUR-WS.org, 2023.

A. Tables and Figures

Table 9
Triggering concepts from the Uni Reading and Michigan trigger lists (verbatim). The right column shows the extracted trigger warning labels in a rough semantic grouping.

Uni Reading [3]	Uni Michigan [3]	Unfolded, merged set
Mental illness and ableism	Mental illness and ableism	ableism
Classism	Classism	classism
Homophobia and heterosexism	Homophobia and heterosexism	homophobia
Homophobia and heterosexism	Homophobia and heterosexism	heterosexism
Sexism and misogyny	Sexism and misogyny	misogyny
Racism and racial slurs	Racism and racial slurs	racism
Sexism and misogyny	Sexism and misogyny	sexism
Transphobia and trans misogyny	Transphobia and trans misogyny	transphobia
Transphobia and trans misogyny	Transphobia and trans misogyny	trans-misogyny
Hateful language directed at religious groups (e.g., Islamophobia, antisemitism)	Hateful language direct at religious groups (e.g., Islamophobia, anti-Semitism)	religious-discrimination
Violence	Violence	violence
Animal cruelty or animal death	Animal cruelty or animal death	animal-cruelty
Sexual Assault	Sexual assault	sexual-assault
Abuse	Abuse	abuse
Child abuse/paedophilia/incest	Child abuse/pedophilia/incest	child-abuse
Kidnapping and abduction	Kidnapping and abduction	abduction
Kidnapping and abduction	Kidnapping and abduction	kidnapping
Pregnancy/Childbirth	Pregnancy/childbirth	pregnancy
Miscarriages/Abortion	Miscarriages/abortion	miscarriages
Pregnancy/Childbirth	Pregnancy/childbirth	childbirth
Miscarriages/Abortion	Miscarriages/abortion	abortion
Dissection		dissection
Blood	Blood	blood
Death or dying	Death or dying	dying
Death or dying	Death or dying	death
Animal cruelty or animal death	Animal cruelty or animal death	animal-death
Mental illness and ableism	Mental illness and ableism	mental-illness
Self-harm and suicide	Self-harm and suicide	suicide
Eating disorders and body hatred	Eating disorders, body hatred, and fat phobia	eating-disorders
Eating disorders and body hatred	Eating disorders, body hatred, and fat phobia	fat-phobia
Eating disorders and body hatred	Eating disorders, body hatred, and fat phobia	body-hatred
Self-harm and suicide	Self-harm and suicide	self-harm
Child abuse/paedophilia/incest	Child abuse/pedophilia/incest	incest
Child abuse/paedophilia/incest	Child abuse/pedophilia/incest	underage
Pornographic content	Pornographic content	pornographic-content