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

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

Trigger Warning Assignment as a Multi-Label Document Classification Problem

github.com/webis-de/ACL-23

zenodo doi.org/10.5281/zenodo.7976807

A trigger is a topic or situation in a piece of content that evokes imagery reminiscent of past discomfort, distress, or trauma.

The first bankday is Laika Day, on the third of November, which celebrates the first animal flight in space and the death of the dog Laika. On this day, [...]

evokes → Memories of the death of the readers dog.

triggers → Feelings of loss and grief.

Trigger warning: A warning about a possible trigger for the audience, displayed before the content.

Originally used in trauma therapy, trigger warnings have been adopted and extensively expanded by online communities.

Can trigger warnings be assigned automatically?

Contributions

1 A corpus of 7.9 million fan fiction works with trigger warnings from Archive of Our Own (AO3).

2 A curated 36-label trigger warning taxonomy based on guidelines from 8 universities.

3 A distant supervision labeling scheme to map freeform tags to trigger warnings.

4 An experimental evaluation of classification difficulty.

- On a 1.1 million document dataset with dense labels.
- Across 4 common multi-label models.

1 A Corpus of Fan Fiction

We constructed a corpus of 7.9 million fan fiction documens by downloading all works (up to 2021) and their metadata from AO3.

Data	
Words	58 billion; 7.4K mean (2.2K median) per work.
Languages	91 (90.5% English).
Genre	Amateur narrative fiction.

Metadata	
Fandom	Characters, Relationships.
Statistics	Hits, Kudos, Comments, ...
Archive Warnings	3 coarse and specific warnings: `Rape/Non-Con, Graphic Violence, Character Death`
Additional Tags	9.7 million unique freeform content descriptors. We identified 240,000 of those as warnings.

2 Trigger Warning Taxonomy

We curated content guidelines from 8 international universities to create a 36-label trigger warning taxonomy to annotate the corpus.

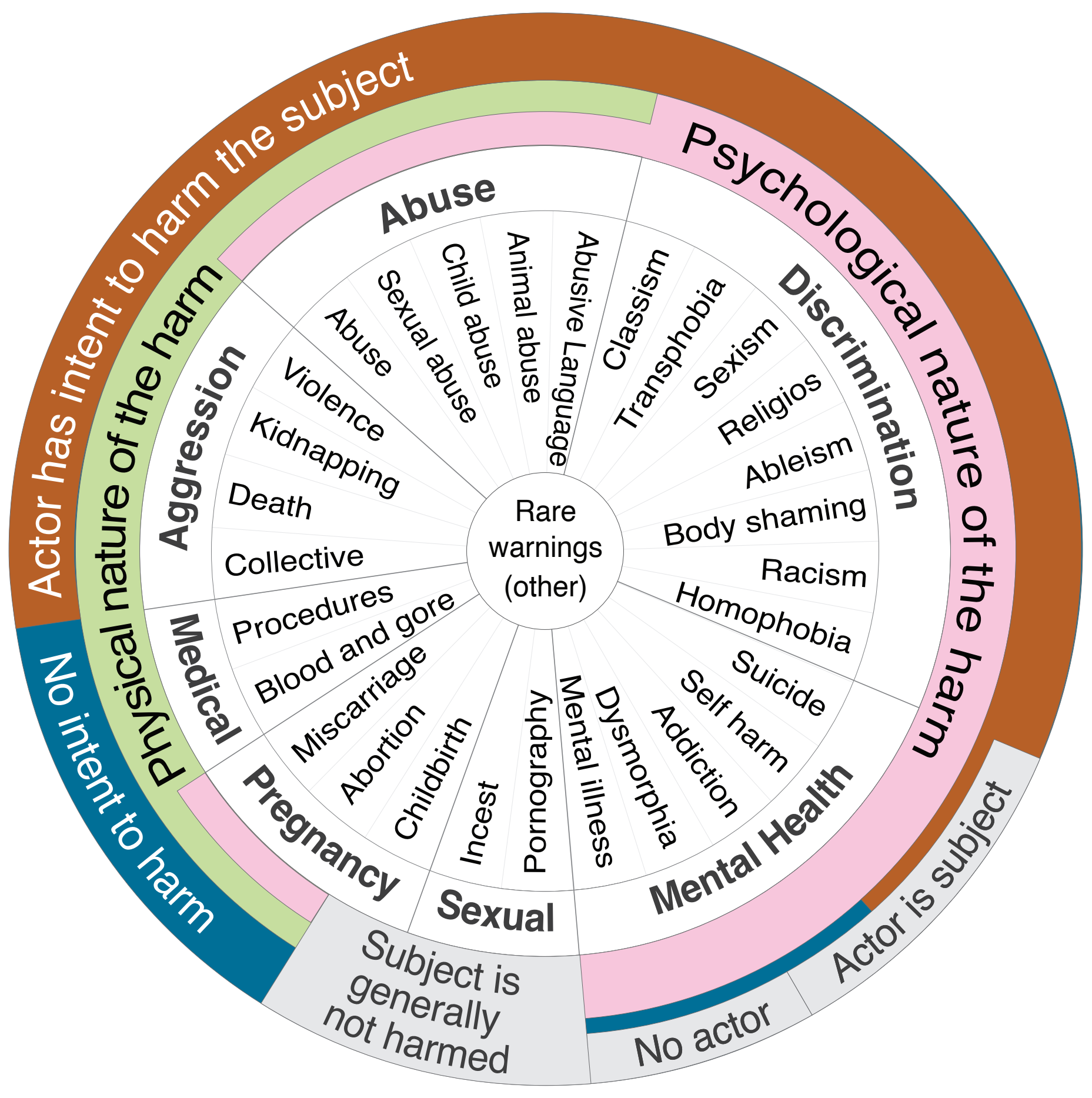
The taxonomy consists of two hierarchical levels:

- 29 fine-grained warning labels with closed-set semantics.
- 7 coarse-grained warning labels with open-set semantics.

The labels are characterized by:

- The nature of the harm depiced in the document.
- The relation of the subject, actor, and intent to harm.

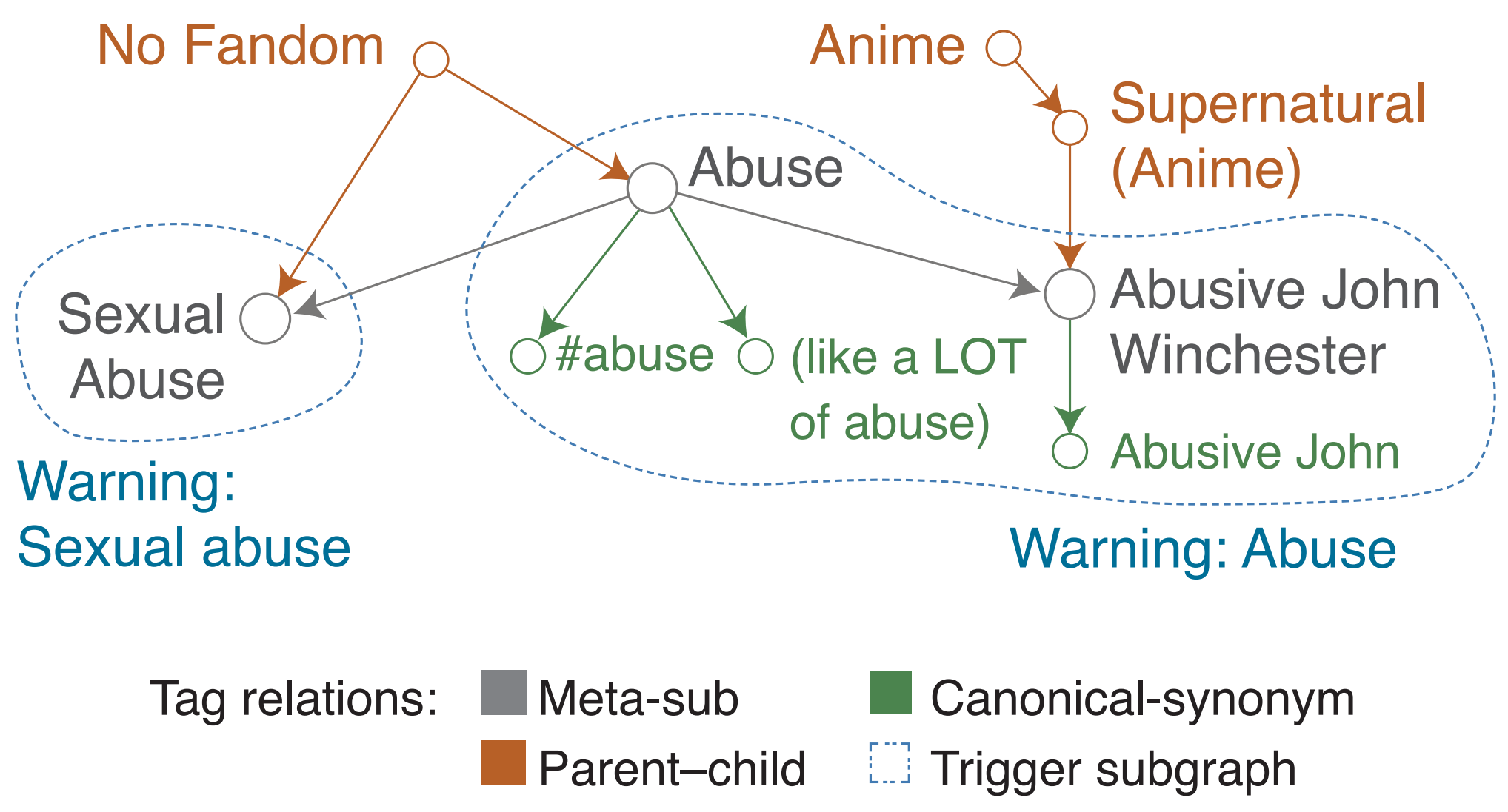
The labels were extracted and grouped in a structured way. The characterization was created in tandem with our annotation guidelines and represent the semantic label interpretation.



3 Labeling Scheme

We assigned the warnings from the taxonomy to the documents in the corpus by annotating the additional tags.

- This is more efficient than annotating the documents.
- It assigns the warnings intended by the author.
- As to not annotate all 9.7 million additional tags, we used the tag relations to identify ca. 6.500 central tags, annotated them manually, and inferred the remainder through the tag graph.
- We manually annotated the 2,000 most common tags.



- The tag graph relates tags with 3+ occurences in an acyclic digraph with 3 relation types.
- Relations are added by community experts (tag wranglers).
- Our evaluation of the labeling scheme against 3,000 manually annotated tags shows an F_1 of 0.94.

Sample	Tag occurrences	Unique Tags
Top 2,000 tags	27.6M (52%)	2,000 (0.02%)
Tag graph	41.0M (77%)	2 M (20%)
All tags	53.1M	9.7M

4 Experimental Evaluation

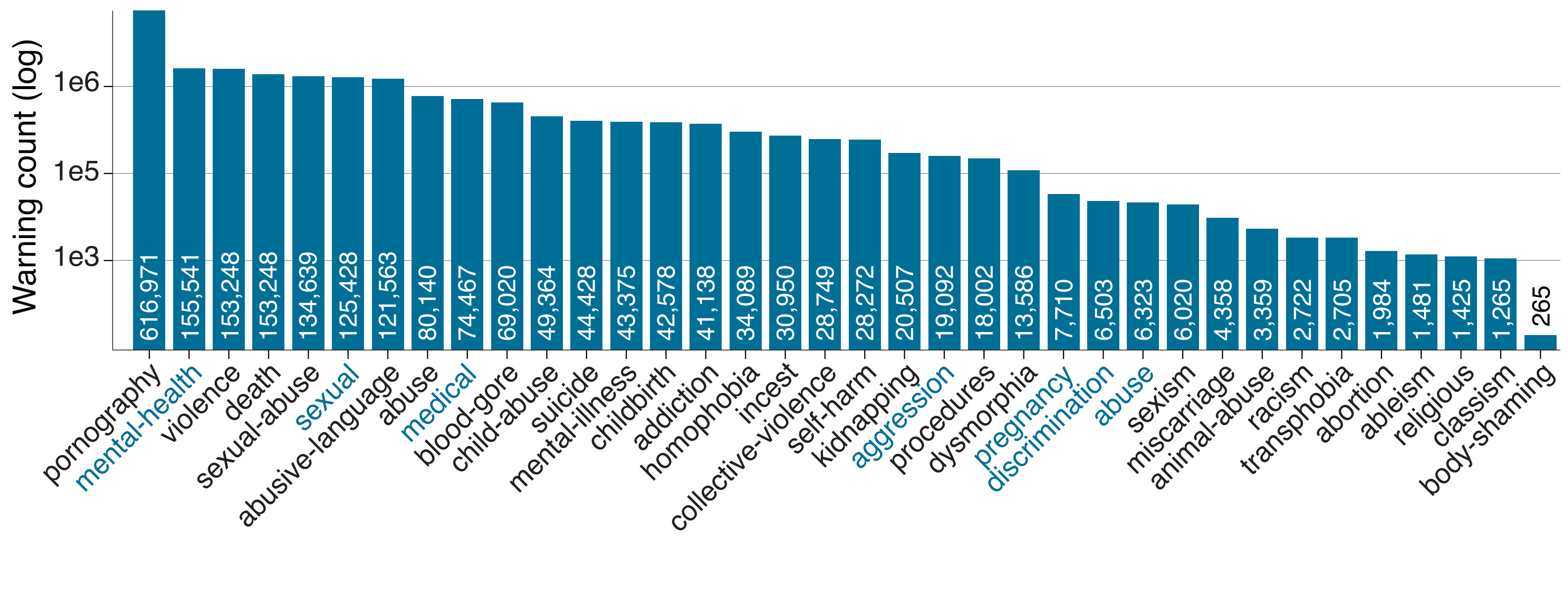
We curated an evaluation dataset with 1.1 million works. Each work has at least 1 warning (mean: 1.5). Most documents are very long (7,986 mean words; median: 3,096).

Curation Criteria

Language	English
Recency	Published after 2009
Length	50-93,000 words

Tag confidence	3-66 additional tags
Popularity conf	>100 hits and >5 kudos
Remove near-duplicates	
Remove works w/ non-annotated tags	

Distribution of warning labels in the dataset



- Labels with open and closed-set semantics are equally difficult.
- Learning on full-text representations is essential.
  Transformer classifiers are very good on short documents.
  Input truncation substantially reduces effectiveness.
- Recall is a key issue.
  Trigger warning assignment is a high-recall task.
  False negatives (missed warnings) cause more harm than false positives.
- Poor effectiveness on rare labels (common for MLC problems).
- Predicting coarse-grained labels (7) is easier (+0.2 F_1).
  Predicting fine-grained labels (36) is much more desirable.

Selection of evaluation results; 2 models on 36-label MLC.

	Micro-average			Macro-average		
	Precision	Recall	F_1	Precision	Recall	F_1
`XGBoost`	0.72	0.40	**0.52**	0.44	0.25	**0.30**
`BERT`	0.56	0.37	0.45	0.36	0.19	0.23