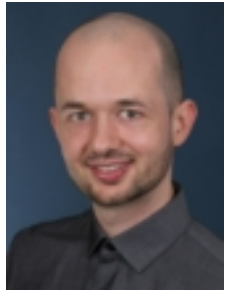
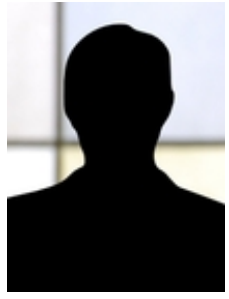
Trigger Warning Assignment as a Multi-Label Document Classification Problem



Matti Wiegmann Magdalena Wolska Christopher Schröder Ole Borchardt Martin Potthast Benno Stein

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

webis.de

Trigger Warnings

Trigger:

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

"*[. . .] 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 Warnings

Trigger:

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

"*[. . .] 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.

Trigger Warnings

Trigger:

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

"*[. . .] 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?

Trigger Warning Corpus

Contribution: A corpus of 7.9 million fan fiction documents with trigger warnings from Archive of Our Own (AO3).

Rating:	Mature
Archive Warnings:	**No Archive Warnings Apply, Major Character Death, Graphic Depictions Of Violence**
Category:	M/M
Fandom:	Harry Potter - J. K. Rowling
Relationships:	Sirius Black/Remus Lupin, Sirius Black & Remus Lupin, James Potter/Lily Evans Potter
Characters:	Remus Lupin, Sirius Black, James Potter, Lily Evans Potter, Peter Pettigrew, Severus Snape, Minerva McGonagall, Bellatrix Black Lestrange, Narcissa Black Malfoy, Albus Dumbledore, Mulciber Sr. (Harry Potter), Horace Slughorn, Mary Macdonald, Marlene McKinnon, Poppy Pomfrey, Walburga Black, Regulus Black, Fenrir Greyback
Additional Tags:	Marauders' Era, Marauders, Marauders Friendship, wolfstar, Get Together, Slow Burn, so slow, it's slow, seriously, Complete, Canon Compliant, Angst, Fluff, Fluff and Angst, Requited Love, Canonical Character Death, First War with Voldemort, First Kiss, Period Typical Attitudes
Language:	English
Series:	Part 1 of All the Young Dudes • Next Work →
Stats:	Published: 2017-03-02 Completed: 2018-11-12 Words: 526,969 Chapters: 188/188 Comments: 30,603 Kudos: 155,026 Bookmarks: 30,939 Hits: 10,623,619

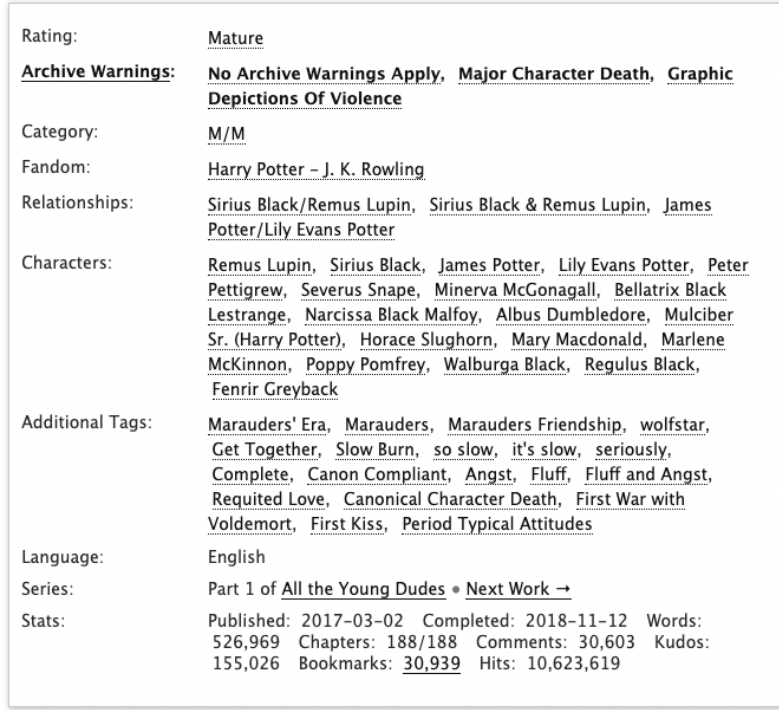
[MsKingBean89, 2018]

Trigger Warning Corpus

Contribution: A corpus of 7.9 million fan fiction documents with trigger warnings from Archive of Our Own (AO3).

Data:

Words	58 billion (7.4K mean; 2,2K median)
Languages	91 (90.5% English)
Genre	Amateur narrative fiction



Rating:	Mature
Archive Warnings:	No Archive Warnings Apply, Major Character Death, Graphic Depictions Of Violence
Category:	M/M
Fandom:	Harry Potter - J. K. Rowling
Relationships:	Sirius Black/Remus Lupin, Sirius Black & Remus Lupin, James Potter/Lily Evans Potter
Characters:	Remus Lupin, Sirius Black, James Potter, Lily Evans Potter, Peter Pettigrew, Severus Snape, Minerva McGonagall, Bellatrix Black Lestrange, Narcissa Black Malfoy, Albus Dumbledore, Mulciber Sr. (Harry Potter), Horace Slughorn, Mary Macdonald, Marlene McKinnon, Poppy Pomfrey, Walburga Black, Regulus Black, Fenrir Greyback
Additional Tags:	Marauders' Era, Marauders, Marauders Friendship, wolfstar, Get Together, Slow Burn, so slow, it's slow, seriously, Complete, Canon Compliant, Angst, Fluff, Fluff and Angst, Requited Love, Canonical Character Death, First War with Voldemort, First Kiss, Period Typical Attitudes
Language:	English
Series:	Part 1 of All the Young Dudes • Next Work →
Stats:	Published: 2017-03-02 Completed: 2018-11-12 Words: 526,969 Chapters: 188/188 Comments: 30,603 Kudos: 155,026 Bookmarks: 30,939 Hits: 10,623,619

[MsKingBean89, 2018]

Trigger Warning Corpus

Contribution: A corpus of 7.9 million fan fiction documents with trigger warnings from Archive of Our Own (AO3).

Data:

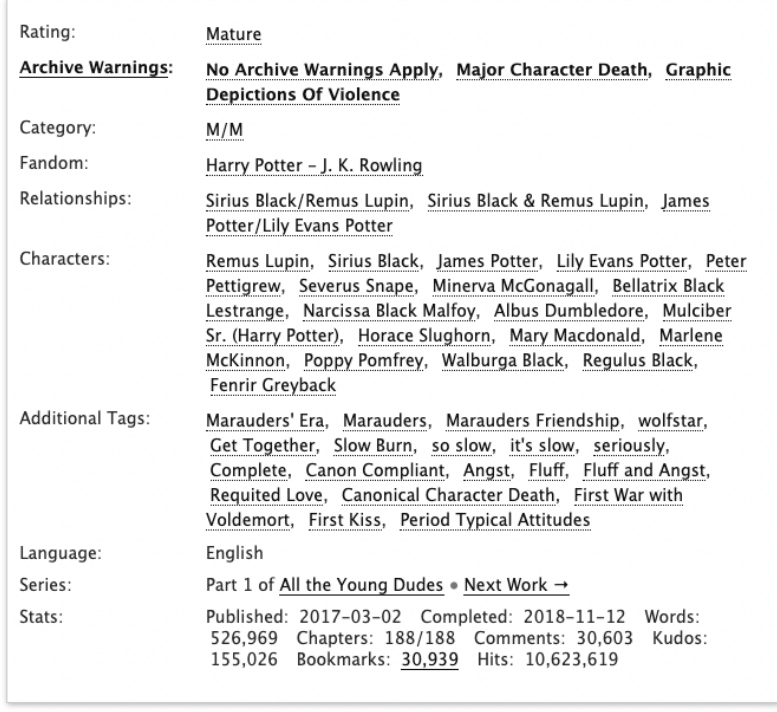
Words	58 billion (7.4K mean; 2,2K median)
Languages	91 (90.5% English)
Genre	Amateur narrative fiction

Metadata:

Fandom	Characters, Relationships, . . .
Stats	Hits, Kudos, Comments, . . .

Archive Warnings	3 coarse and specific warnings
	`Rape/Non-Con, Graphic Violence`
	`Character Death`

Additional Tags	9.7M unique, freeform content descriptions.



Rating:	Mature
Archive Warnings:	No Archive Warnings Apply, Major Character Death, Graphic Depictions Of Violence
Category:	M/M
Fandom:	Harry Potter - J. K. Rowling
Relationships:	Sirius Black/Remus Lupin, Sirius Black & Remus Lupin, James Potter/Lily Evans Potter
Characters:	Remus Lupin, Sirius Black, James Potter, Lily Evans Potter, Peter Pettigrew, Severus Snape, Minerva McGonagall, Bellatrix Black Lestrange, Narcissa Black Malfoy, Albus Dumbledore, Mulciber Sr. (Harry Potter), Horace Slughorn, Mary Macdonald, Marlene McKinnon, Poppy Pomfrey, Walburga Black, Regulus Black, Fenrir Greyback
Additional Tags:	Marauders' Era, Marauders, Marauders Friendship, wolfstar, Get Together, Slow Burn, so slow, it's slow, seriously, Complete, Canon Compliant, Angst, Fluff, Fluff and Angst, Requited Love, Canonical Character Death, First War with Voldemort, First Kiss, Period Typical Attitudes
Language:	English
Series:	Part 1 of All the Young Dudes • Next Work →
Stats:	Published: 2017-03-02 Completed: 2018-11-12 Words: 526,969 Chapters: 188/188 Comments: 30,603 Kudos: 155,026 Bookmarks: 30,939 Hits: 10,623,619

[MsKingBean89, 2018]

Trigger Warning Corpus

Contribution: A corpus of 7.9 million fan fiction documents with trigger warnings from Archive of Our Own (AO3).

Data:

Words	58 billion (7.4K mean; 2,2K median)
Languages	91 (90.5% English)
Genre	Amateur narrative fiction

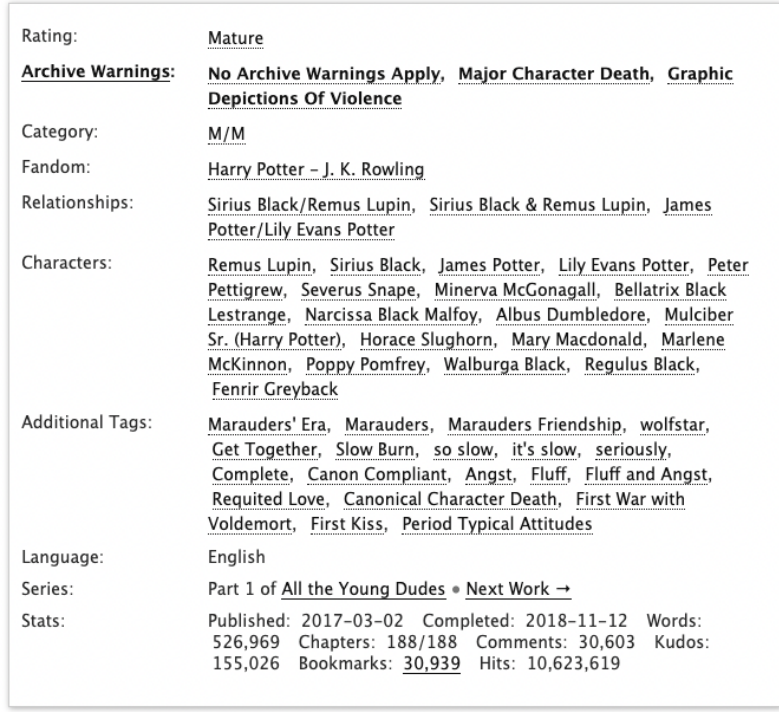
Metadata:

Fandom	Characters, Relationships, . . .
Stats	Hits, Kudos, Comments, . . .

Archive Warnings	3 coarse and specific warnings
	`Rape/Non-Con, Graphic Violence`
	`Character Death`

Additional Tags	9.7M unique, freeform content descriptions.

➜ We identified 240,000 unique Additional Tags as trigger warnings.



Rating:	Mature
Archive Warnings:	No Archive Warnings Apply, Major Character Death, Graphic Depictions Of Violence
Category:	M/M
Fandom:	Harry Potter - J. K. Rowling
Relationships:	Sirius Black/Remus Lupin, Sirius Black & Remus Lupin, James Potter/Lily Evans Potter
Characters:	Remus Lupin, Sirius Black, James Potter, Lily Evans Potter, Peter Pettigrew, Severus Snape, Minerva McGonagall, Bellatrix Black Lestrange, Narcissa Black Malfoy, Albus Dumbledore, Mulciber Sr. (Harry Potter), Horace Slughorn, Mary Macdonald, Marlene McKinnon, Poppy Pomfrey, Walburga Black, Regulus Black, Fenrir Greyback
Additional Tags:	Marauders' Era, Marauders, Marauders Friendship, wolfstar, Get Together, Slow Burn, so slow, it's slow, seriously, Complete, Canon Compliant, Angst, Fluff, Fluff and Angst, Requited Love, Canonical Character Death, First War with Voldemort, First Kiss, Period Typical Attitudes
Language:	English
Series:	Part 1 of All the Young Dudes • Next Work →
Stats:	Published: 2017-03-02 Completed: 2018-11-12 Words: 526,969 Chapters: 188/188 Comments: 30,603 Kudos: 155,026 Bookmarks: 30,939 Hits: 10,623,619

[MsKingBean89, 2018]

A Taxonomy of Trigger Warnings I

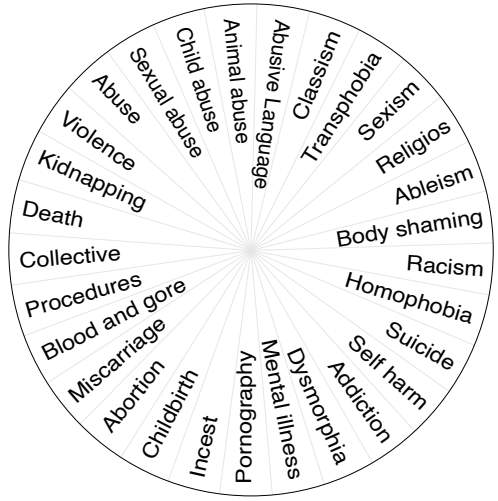
Contribution: A curated 36-label trigger warning taxonomy compiled from 8 university guides.

A Taxonomy of Trigger Warnings I

Contribution: A curated 36-label trigger warning taxonomy compiled from 8 university guides.

- 29 closed-set trigger labels.

  `Suicide, Eating disorders, Pornography, ...`



A Taxonomy of Trigger Warnings I

Contribution: A curated 36-label trigger warning taxonomy compiled from 8 university guides.

- 29 closed-set trigger labels.

  `Suicide, Eating disorders, Pornography, ...`

- 7 open-set trigger label groups.

  `Needles, Politics, ...`

- The long tail of the trigger warning labels is captured by the center, the rare warnings group, subordinate to each of the 7 coarse groups.

A Taxonomy of Trigger Warnings I

Contribution: A curated 36-label trigger warning taxonomy compiled from 8 university guides.

- 29 closed-set trigger labels.

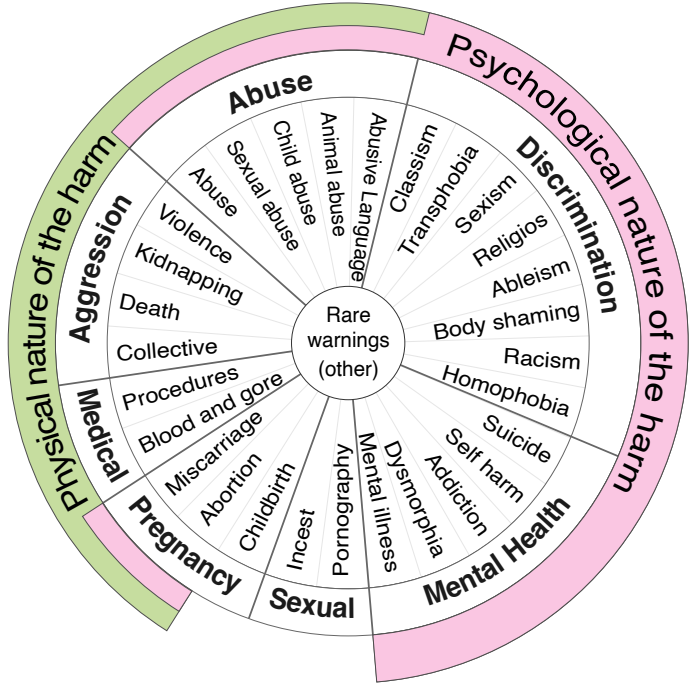
  `Suicide, Eating disorders, Pornography, ...`

- 7 open-set trigger label groups.

  `Needles, Politics, ...`

- The long tail of the trigger warning labels is captured by the center, the rare warnings group, subordinate to each of the 7 coarse groups.

- Characterization of the nature of the harm.



A Taxonomy of Trigger Warnings I

Contribution: A curated 36-label trigger warning taxonomy compiled from 8 university guides.

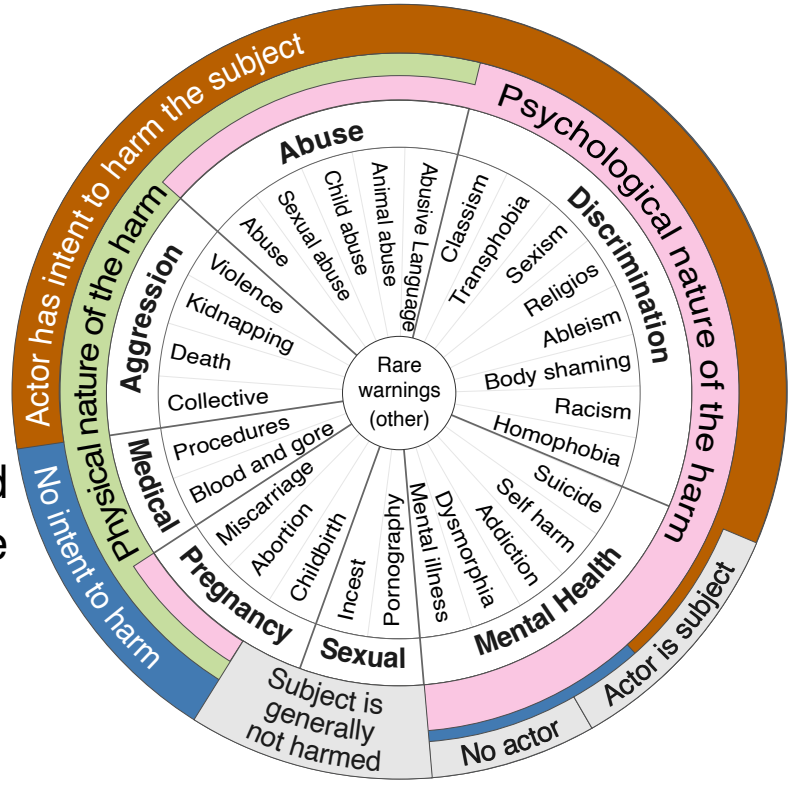
- 29 closed-set trigger labels.

  `Suicide, Eating disorders, Pornography, ...`

- 7 open-set trigger label groups.

  `Needles, Politics, ...`

- The long tail of the trigger warning labels is captured by the center, the rare warnings group, subordinate to each of the 7 coarse groups.
- Characterization of the nature of the harm.
- Characterization of the subject-actor-intent relation.



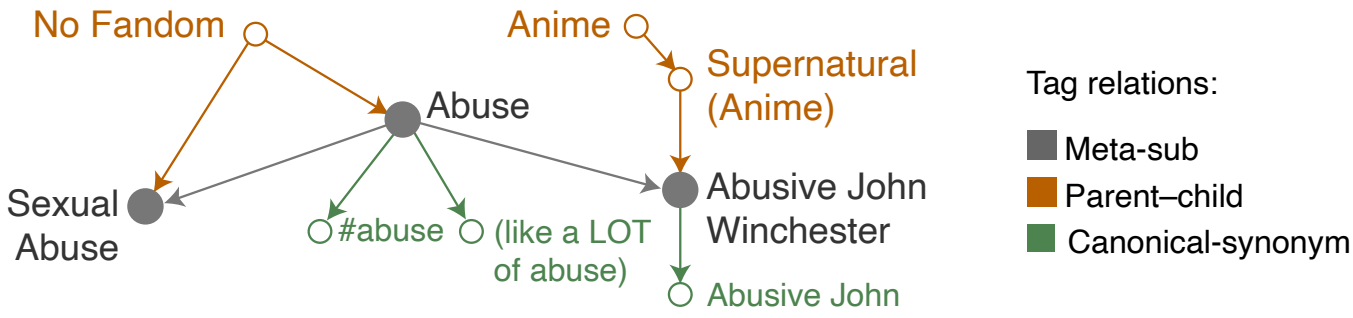
A Taxonomy of Trigger Warnings II

Contribution: A comprehensive analysis of 53 million user-defined additional tags and mapping of 41 million among them to our taxonomy to ensure a thorough grounding.

A Taxonomy of Trigger Warnings II

Contribution: A comprehensive analysis of 53 million user-defined additional tags and mapping of 41 million among them to our taxonomy to ensure a thorough grounding.

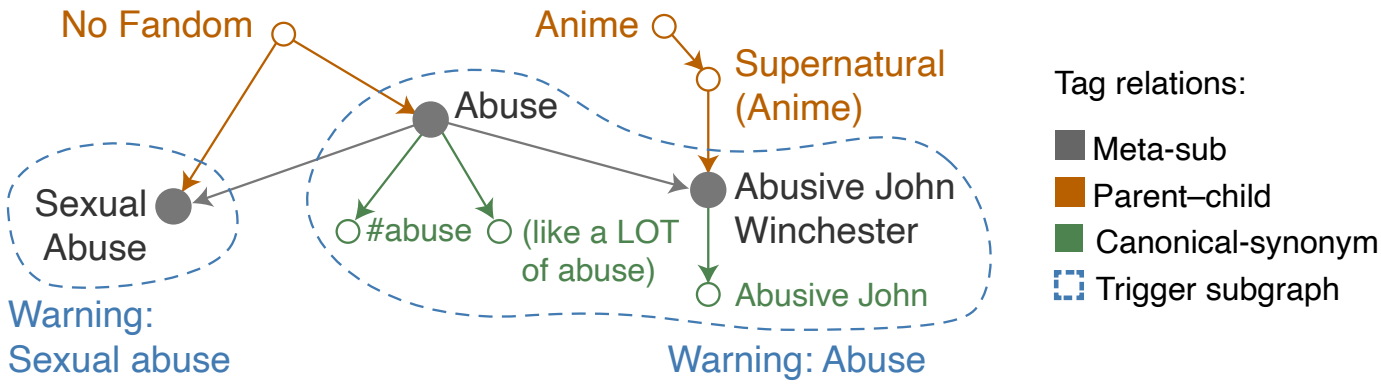
- Most tags are related to other tags through the acyclic tag (di)graph.
- Relations are added by community experts (*Wranglers*).



A Taxonomy of Trigger Warnings II

Contribution: A comprehensive analysis of 53 million user-defined additional tags and mapping of 41 million among them to our taxonomy to ensure a thorough grounding.

- Most tags are related to other tags through the acyclic tag (di)graph.
- Relations are added by community experts (*Wranglers*).
- Identify and classify central nodes, infer label for other nodes.



Multilabel Trigger Warning Assignment

Data sampling: Sampling of a dense and reliable subset of data for the experimental evaluation.

Curation Criteria

Language	English.	Tag confidence	3–66 additional tags.
Recency	Published after 2009.	Popularity confidence	>100 hits, >5 kudos.
Length	50–93,000 words.	Remove near-duplicates.	

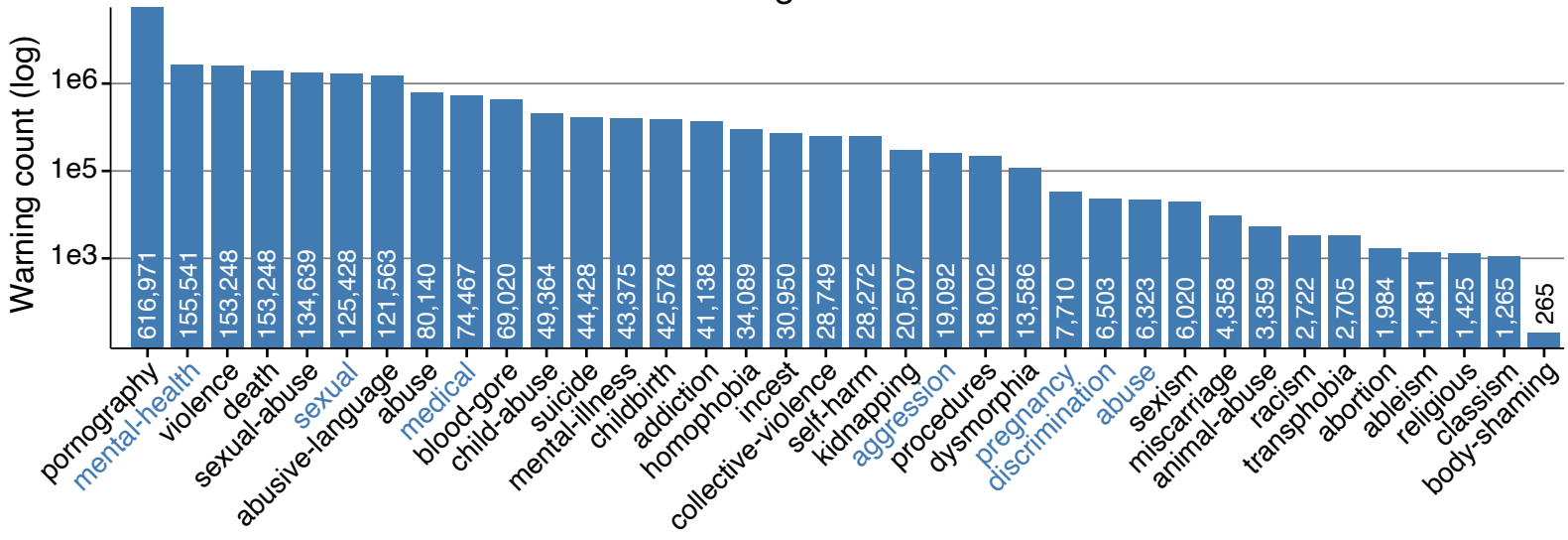
Multilabel Trigger Warning Assignment

Data sampling: Sampling of a dense and reliable subset of data for the experimental evaluation.

Curation Criteria

Language	English.	Tag confidence	3–66 additional tags.
Recency	Published after 2009.	Popularity confidence	>100 hits, >5 kudos.
Length	50–93,000 words.	Remove near-duplicates.	

Distribution of warning labels in the dataset.



Multilabel Trigger Warning Assignment

Experimental evaluation: Evaluated 4 models (SVM, XGBoost, BERT, Longformer) to assess classification difficulty and open problems.

Top Model (XGBoost)

Fine-grained (36 labels)		
Macro	Precision	0.44
	Recall	0.25
	F_1	0.30
Micro	Precision	0.72
	Recall	0.40
	F_1	0.52

Multilabel Trigger Warning Assignment

Experimental evaluation: Evaluated 4 models (SVM, XGBoost, BERT, Longformer) to assess classification difficulty and open problems.

Findings

1. Labels with open and closed-set semantics are equally difficult.

Top Model (XGBoost)

Fine-grained (36 labels)		
Macro	Precision	0.44
	Recall	0.25
	F_1	0.30
Micro	Precision	0.72
	Recall	0.40
	F_1	0.52

Multilabel Trigger Warning Assignment

Experimental evaluation: Evaluated 4 models (SVM, XGBoost, BERT, Longformer) to assess classification difficulty and open problems.

Findings

1. Labels with open and closed-set semantics are equally difficult.
2. Learning on full-text representations is essential. Models are less effective if input is truncated.

Top Model (XGBoost)

Fine-grained (36 labels)		
Macro	Precision	0.44
	Recall	0.25
	F_1	0.30
Micro	Precision	0.72
	Recall	0.40
	F_1	0.52

Multilabel Trigger Warning Assignment

Experimental evaluation: Evaluated 4 models (SVM, XGBoost, BERT, Longformer) to assess classification difficulty and open problems.

Findings

1. Labels with open and closed-set semantics are equally difficult.
2. Learning on full-text representations is essential. Models are less effective if input is truncated.
3. Recall is low, which is a key issue.

Top Model (XGBoost)

Fine-grained (36 labels)		
Macro	Precision	0.44
	Recall	0.25
	F_1	0.30
Micro	Precision	0.72
	Recall	0.40
	F_1	0.52

Multilabel Trigger Warning Assignment

Experimental evaluation: Evaluated 4 models (SVM, XGBoost, BERT, Longformer) to assess classification difficulty and open problems.

Findings

1. Labels with open and closed-set semantics are equally difficult.
2. Learning on full-text representations is essential. Models are less effective if input is truncated.
3. Recall is low, which is a key issue.
4. Poor effectiveness on rare labels.

Top Model (XGBoost)

Fine-grained (36 labels)		
Macro	Precision	0.44
	Recall	0.25
	F_1	0.30
Micro	Precision	0.72
	Recall	0.40
	F_1	0.52

Summary

1. A corpus of 7.9 million fan-fiction with content descriptors from `archiveofourown.org` (AO3).
2. A curated 36-label trigger warning taxonomy.
3. A distant supervision scheme to map additional tags to warnings.
4. A curated 1 million document dataset with dense labels.
5. Experimental evaluation of classification difficulty and open problems.

Data `https://doi.org/10.5281/zenodo.7976807`
Code `https://github.com/webis-de/ACL-23`
Contact `matti.wiegmann@uni-weimar.de`