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 $\rightarrow$ Memories of the death of the readers dog.

triggers $\rightarrow$ 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]

@Wiegmann, 2023
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

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, English

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.

**Archive Warnings:**
- No Archive Warnings Apply
- Major Character Death
- Graphic Depictions Of Violence

**Category:** M/M
**Rating:** Mature

**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, wolffar
- 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]
@Wiegmann, 2023
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.

---

---

[MsKingBean89, 2018]

8 @Wiegmann, 2023
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, ...

- 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 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.
**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.

**Completeness:** ca. 80% of tags. **Effectiveness:** ca. 0.95 F1.
Data sampling: Sampling of a dense and reliable subset of data for the experimental evaluation.

Curation Criteria

<table>
<thead>
<tr>
<th>Language</th>
<th>English.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recency</td>
<td>Published after 2009.</td>
</tr>
<tr>
<td>Length</td>
<td>50–93,000 words.</td>
</tr>
<tr>
<td>Tag confidence</td>
<td>3–66 additional tags.</td>
</tr>
<tr>
<td>Popularity confidence</td>
<td>&gt;100 hits, &gt;5 kudos.</td>
</tr>
<tr>
<td>Remove near-duplicates.</td>
<td></td>
</tr>
</tbody>
</table>
Multilabel Trigger Warning Assignment

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

**Curation Criteria**

<table>
<thead>
<tr>
<th>Language</th>
<th>English.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recency</td>
<td>Published after 2009.</td>
</tr>
<tr>
<td>Length</td>
<td>50–93,000 words.</td>
</tr>
<tr>
<td>Tag confidence</td>
<td>3–66 additional tags.</td>
</tr>
<tr>
<td>Popularity</td>
<td>&gt;100 hits, &gt;5 kudos.</td>
</tr>
<tr>
<td>Remove</td>
<td>near-duplicates.</td>
</tr>
</tbody>
</table>

Distribution of warning labels in the dataset.

<table>
<thead>
<tr>
<th>Warning Category</th>
<th>Count (log)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pornography</td>
<td>616,971</td>
</tr>
<tr>
<td>Mental-health</td>
<td>155,541</td>
</tr>
<tr>
<td>Violence</td>
<td>153,248</td>
</tr>
<tr>
<td>Sexual-abuse</td>
<td>125,428</td>
</tr>
<tr>
<td>Abusive-language</td>
<td>121,563</td>
</tr>
<tr>
<td>Medical</td>
<td>80,140</td>
</tr>
<tr>
<td>Child-abuse</td>
<td>74,467</td>
</tr>
<tr>
<td>Suicide</td>
<td>69,020</td>
</tr>
<tr>
<td>Mental-illness</td>
<td>49,364</td>
</tr>
<tr>
<td>Addiction</td>
<td>41,138</td>
</tr>
<tr>
<td>Homophobia</td>
<td>34,089</td>
</tr>
<tr>
<td>Incest</td>
<td>30,950</td>
</tr>
<tr>
<td>Abusive-language</td>
<td>28,749</td>
</tr>
<tr>
<td>Aggression</td>
<td>28,272</td>
</tr>
<tr>
<td>Suicide</td>
<td>20,507</td>
</tr>
<tr>
<td>Procedural</td>
<td>19,092</td>
</tr>
<tr>
<td>Dysmorphia</td>
<td>18,002</td>
</tr>
<tr>
<td>Pregnancy</td>
<td>17,710</td>
</tr>
<tr>
<td>Abuse</td>
<td>13,586</td>
</tr>
<tr>
<td>Sexism</td>
<td>12,722</td>
</tr>
<tr>
<td>Animal-abuse</td>
<td>12,165</td>
</tr>
<tr>
<td>Racism</td>
<td>12,002</td>
</tr>
<tr>
<td>Graduation</td>
<td>11,894</td>
</tr>
<tr>
<td>Abolition</td>
<td>11,425</td>
</tr>
<tr>
<td>Religious</td>
<td>10,941</td>
</tr>
<tr>
<td>Classism</td>
<td>10,865</td>
</tr>
<tr>
<td>Body-shaming</td>
<td>1,265</td>
</tr>
</tbody>
</table>

@Wiegmann, 2023
Multilabel Trigger Warning Assignment

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

**Top Model (XGBoost)**

<table>
<thead>
<tr>
<th>Fine-grained (36 labels)</th>
<th>Macro</th>
<th>Precision</th>
<th>0.44</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Recall</td>
<td>0.25</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$F_1$</td>
<td>0.30</td>
<td></td>
</tr>
<tr>
<td>Micro</td>
<td>Precision</td>
<td>0.72</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Recall</td>
<td>0.40</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$F_1$</td>
<td>0.52</td>
<td></td>
</tr>
</tbody>
</table>
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.

<table>
<thead>
<tr>
<th>Top Model (XGBoost)</th>
<th>Fine-grained (36 labels)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Macro</td>
</tr>
<tr>
<td></td>
<td>Precision</td>
</tr>
<tr>
<td></td>
<td>Recall</td>
</tr>
<tr>
<td></td>
<td>$F_1$</td>
</tr>
<tr>
<td></td>
<td>Micro</td>
</tr>
<tr>
<td></td>
<td>Precision</td>
</tr>
<tr>
<td></td>
<td>Recall</td>
</tr>
<tr>
<td></td>
<td>$F_1$</td>
</tr>
</tbody>
</table>
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)

<table>
<thead>
<tr>
<th>Fine-grained (36 labels)</th>
<th>Macro</th>
<th>Precision</th>
<th>0.44</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Recall</td>
<td>0.25</td>
<td></td>
</tr>
<tr>
<td></td>
<td>F₁</td>
<td>0.30</td>
<td></td>
</tr>
<tr>
<td>Micro</td>
<td>Precision</td>
<td>0.72</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Recall</td>
<td>0.40</td>
<td></td>
</tr>
<tr>
<td></td>
<td>F₁</td>
<td>0.52</td>
<td></td>
</tr>
</tbody>
</table>
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.

<table>
<thead>
<tr>
<th>Fine-grained (36 labels)</th>
<th>Top Model (XGBoost)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Macro Precision</td>
<td>0.44</td>
</tr>
<tr>
<td>Recall</td>
<td>0.25</td>
</tr>
<tr>
<td>F₁</td>
<td>0.30</td>
</tr>
<tr>
<td>Micro Precision</td>
<td>0.72</td>
</tr>
<tr>
<td>Recall</td>
<td>0.40</td>
</tr>
<tr>
<td>F₁</td>
<td>0.52</td>
</tr>
</tbody>
</table>
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)

<table>
<thead>
<tr>
<th>Fine-grained (36 labels)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Macro Precision</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Micro Precision</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>
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