Overview of the Trigger Detection Task at PAN 2023

Matti Wiegmann
Magdalena Wolska
Martin Potthast
Benno Stein

Bauhaus-Universität Weimar  Leipzig University  ScaDS.AI
Trigger Warnings

Trigger:

- A trigger in media content is a topic or situation that evokes images, memories, or emotions that cause discomfort or distress.

“Great infernos dotted the city here and there, charring and cremating the still bodies of those committed souls who now lay still forever.”

- evokes Memories of a past war.
- triggers Anxiety, feelings of loss or grief, …
Trigger Warnings

Trigger:

- A trigger in media content is a topic or situation that *evokes* images, memories, or emotions that *cause* discomfort or distress.

> “Great infernos dotted the city here and there, charring and cremating the still bodies of those committed souls who now lay still forever.”

- evokes → Memories of a past war.
- triggers → Anxiety, feelings of loss or 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 in media content is a topic or situation that evokes images, memories, or emotions that cause discomfort or distress.

“Great infernos dotted the city here and there, charring and cremating the still bodies of those committed souls who now lay still forever.”

- evokes Memories of a past war.
- triggers Anxiety, feelings of loss or grief, …

Trigger Detection at PAN 2023:

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

Dataset:
- Contains 341,246 English fan fiction documents.
- Documents are 50–6,000 words long.
- Annotated with 32 warning labels (multi-label).

Number of documents with the given warning label:

- Pornography: 264,529
- Sexual-assault: 34,802
- Violence: 32,350
- Abuse: 24,652
- Death: 23,095
- Ableism: 297
- Misogyny: 248
- Animal-death: 232
- Classism: 209
- Animal-cruelty: 168

Warning count (log)
Task Overview

Dataset:
- Contains 341,246 English fan fiction documents.
- Documents are 50–6,000 words long.
- Annotated with 32 warning labels (multi-label).

Evaluation:
- Precision, Recall, $F_1$, all micro and macro averaged.
- Best models: 0.35 macro $F_1$; 0.75 micro $F_1$. 

Number of documents with the given warning label.

<table>
<thead>
<tr>
<th>Warning Label</th>
<th>Number of Documents</th>
</tr>
</thead>
<tbody>
<tr>
<td>pornography</td>
<td>264,529</td>
</tr>
<tr>
<td>sexual-assault</td>
<td>34,802</td>
</tr>
<tr>
<td>violence</td>
<td>32,350</td>
</tr>
<tr>
<td>abuse</td>
<td>24,652</td>
</tr>
<tr>
<td>death</td>
<td>23,095</td>
</tr>
<tr>
<td>ableism</td>
<td>297</td>
</tr>
<tr>
<td>misogyny</td>
<td>248</td>
</tr>
<tr>
<td>animal-death</td>
<td>232</td>
</tr>
<tr>
<td>classism</td>
<td>209</td>
</tr>
<tr>
<td>animal-cruelty</td>
<td>168</td>
</tr>
</tbody>
</table>

Warning count (log):
- $10^3$
- $10^4$
- $10^5$
Task Overview

Dataset:
- Contains 341,246 English fan fiction documents.
- Documents are 50–6,000 words long.
- Annotated with 32 warning labels (multi-label).

Evaluation:
- Precision, Recall, $F_1$, all micro and macro averaged.
- Best models: 0.35 macro $F_1$; 0.75 micro $F_1$.

Submissions:
- 6 teams submitted.
- Different models, features, and strategies to deal with long documents and label imbalances.

Number of documents with the given warning label.

<table>
<thead>
<tr>
<th>Label</th>
<th>Warning count (log)</th>
</tr>
</thead>
<tbody>
<tr>
<td>pornography</td>
<td>264,529</td>
</tr>
<tr>
<td>sexual-assault</td>
<td>34,802</td>
</tr>
<tr>
<td>violence</td>
<td>32,350</td>
</tr>
<tr>
<td>abuse</td>
<td>24,652</td>
</tr>
<tr>
<td>death</td>
<td>23,095</td>
</tr>
<tr>
<td>ableism</td>
<td>297</td>
</tr>
<tr>
<td>misogyny</td>
<td>248</td>
</tr>
<tr>
<td>animal-death</td>
<td>232</td>
</tr>
<tr>
<td>classism</td>
<td>209</td>
</tr>
<tr>
<td>animal-cruelty</td>
<td>168</td>
</tr>
</tbody>
</table>
Dataset

Trigger Warning Taxonomy:

- We curated a trigger warning taxonomy based on university guidelines (Michigan and Reading).

Characterization of the nature of the harm.

Characterization of the subject-actor-intent relation.
Dataset

Trigger Warning Taxonomy:

- We curated a trigger warning taxonomy based on university guidelines (Michigan and Reading).
- It contains 32 closed-set warnings.

  Suicide, Eating disorders, Pornography, ...
Trigger Warning Taxonomy:

- We curated a trigger warning taxonomy based on university guidelines (Michigan and Reading).
- It contains 32 closed-set warnings. Suicide, Eating disorders, Pornography, ...
- The long tail of rare warnings is captured by 7 open-set warning groups (not used in the task).
Dataset

Trigger Warning Taxonomy:

- We curated a trigger warning taxonomy based on university guidelines (Michigan and Reading).
- It contains 32 closed-set warnings. Suicide, Eating disorders, Pornography, ...
- The long tail of rare warnings is captured by 7 open-set warning groups (not used in the task).
- Characterization of the nature of the harm.

@Wiegmann, 2023
Dataset

Trigger Warning Taxonomy:

- We curated a trigger warning taxonomy based on university guidelines (Michigan and Reading).
- It contains 32 closed-set warnings. Suicide, Eating disorders, Pornography, ...
- The long tail of rare warnings is captured by 7 open-set warning groups (not used in the task).
- Characterization of the nature of the harm.
- Characterization of the subject-actor-intent relation.
Dataset

Trigger Warning Taxonomy:

- We curated a trigger warning taxonomy based on university guidelines (Michigan and Reading).
- It contains 32 closed-set warnings. Suicide, Eating disorders, Pornography, ...
- The long tail of rare warnings is captured by 7 open-set warning groups (not used in the task).
- Characterization of the nature of the harm.
- Characterization of the subject-actor-intent relation.

Note: For an updated version of the taxonomy see [Wiegmann et al. (ACL 2023)]
Dataset

Documents:

- We scraped 7.9 million fan fiction documents with metadata 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, wolffman, 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 →

[MsKingBean89, 2018]
Dataset

Documents:

- We scraped 7.9 million fan fiction documents with metadata from Archive of Our Own (AO3).
- Select works based on recency (2009+), language (English), warning label confidence, length (50–6,000 words), popularity (1,000+ hits, 10+ kudos).

[MsKingBean89, 2018]
Dataset

Documents:

- We scraped 7.9 million fan fiction documents with metadata from Archive of Our Own (AO3).
- Select works based on recency (2009+), language (English), warning label confidence, length (50–6,000 words), popularity (1,000+ hits, 10+ kudos)
- Stratified sampling into training (307,102), validation (17,104), and test (17,040) documents.

[MsKingBean89, 2018]

[Wiegmann, 2023]
Dataset

Documents:

- We scraped 7.9 million fan fiction documents with metadata from Archive of Our Own (AO3).
- Select works based on recency (2009+), language (English), warning label confidence, length (50–6,000 words), popularity (1,000+ hits, 10+ kudos)
- Stratified sampling into training (307,102), validation (17,104), and test (17,040) documents.
- Determine warning based on the 10 million unique freeform tags.

[MsKingBean89, 2018]
Determined warning labels:

- Freeform tags are related through tag relations that were added by community experts → Semi-automatic annotation.
- Synonymous tags are related. One synonym is marked as canonical.
- Canonical tags are in a meta-sub relation.
- Sources were annotated with a warning.
- Parent relations indicate Genre/Fandom. Warnings are usually children of No Fandom.
- Annotate ca. 6,000 nodes, infer label for ca. 80% of tags used; 0.95 F1.
Determining warning labels:

- Freeform tags are related through tag relations that were added by community experts
  - Semi-automatic annotation.

- Synonymous tags are related.
  - One synonym is marked as *canonical*.
Determining warning labels:

- Freeform tags are related through tag relations that were added by community experts
  - Semi-automatic annotation.

- Synonymous tags are related. One synonym is marked as *canonical*.

- Canonical tags are in a meta-sub relation. Sources were annotated with a warning.
Dataset

Determining warning labels:

- Freeform tags are related through tag relations that were added by community experts → Semi-automatic annotation.
- Synonymous tags are related. One synonym is marked as *canonical*.
- Canonical tags are in a meta-sub relation. Sources were annotated with a warning.
- Parent relations indicate Genre/Fandom. Warnings are usually children of *No Fandom*.
- Annotate ca. 6,000 nodes, infer label for ca. 80% of tags used; 0.95 $F_1$. 

---

![Tag relations:](image)

- Warning: Abuse
- Supernatural (Anime)
- Abusive John Winchester
- Anime
- Abusive John
- (like a LOT of abuse)
- No Fandom
- #abuse
- Sexual Abuse
- Warning: Sexual assault
- Tag relations:
  - **Meta**
  - **Synonym**
  - **Parent**

@Wiegmann, 2023
Results

Submissions:

- **XGBoost baseline** based on TF-IDF document vectors.

<table>
<thead>
<tr>
<th>Participant</th>
<th>Macro</th>
<th>Micro</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Prec</td>
<td>Rec</td>
</tr>
<tr>
<td>XGBoost</td>
<td>0.52</td>
<td>0.25</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Participant</th>
<th>Micro</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Prec</td>
</tr>
<tr>
<td>XGBoost</td>
<td>0.88</td>
</tr>
</tbody>
</table>
Results

Submissions:

- **XGBoost baseline** based on TF-IDF document vectors.
- **Sahin et al.** Hierarchical classification with a RoBERTa-base and LSTM, use full documents.

<table>
<thead>
<tr>
<th>Participant</th>
<th>Macro</th>
<th>Micro</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Prec</td>
<td>Rec</td>
</tr>
<tr>
<td>Sahin</td>
<td>0.37</td>
<td>0.42</td>
</tr>
<tr>
<td>XGBoost</td>
<td>0.52</td>
<td>0.25</td>
</tr>
<tr>
<td>Felser</td>
<td>0.11</td>
<td>0.63</td>
</tr>
<tr>
<td>Shashirekha</td>
<td>0.10</td>
<td>0.04</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Participant</th>
<th>Micro</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Prec</td>
</tr>
<tr>
<td>Sahin</td>
<td>0.73</td>
</tr>
<tr>
<td>XGBoost</td>
<td>0.88</td>
</tr>
</tbody>
</table>
Results

Submissions:

- **XGBoost baseline** based on TF-IDF document vectors.
- **Sahin et al.** Hierarchical classification with a RoBERTa-base and LSTM, use full documents.
- **Su et al.** Hierarchical (siamese) classification with a RoBERTa-base and CNN, uses the first and last 500 words.

<table>
<thead>
<tr>
<th>Participant</th>
<th>Macro</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Prec</td>
<td>Rec</td>
</tr>
<tr>
<td>Sahin</td>
<td>0.37</td>
<td>0.42</td>
</tr>
<tr>
<td>Su</td>
<td><strong>0.54</strong></td>
<td>0.30</td>
</tr>
<tr>
<td>XGBoost</td>
<td>0.52</td>
<td>0.25</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Participant</th>
<th>Micro</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Prec</td>
<td>Rec</td>
</tr>
<tr>
<td>Su</td>
<td>0.80</td>
<td>0.71</td>
</tr>
<tr>
<td>Sahin</td>
<td>0.73</td>
<td>0.74</td>
</tr>
<tr>
<td>XGBoost</td>
<td>0.88</td>
<td>0.57</td>
</tr>
</tbody>
</table>
Results

Submissions:

- **XGBoost baseline** based on TF-IDF document vectors.
- **Sahin et al.** Hierarchical classification with a RoBERTa-base and LSTM, use full documents.
- **Su et al.** Hierarchical (siamese) classification with a RoBERTa-base and CNN, uses the first and last 500 words.
- **Haojie Cao et al.** and **Guiyuan Cao et al.** Classify chunks with RoBERTa-based voting ensemble.

<table>
<thead>
<tr>
<th>Participant</th>
<th>Macro</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Prec</td>
<td>Rec</td>
</tr>
<tr>
<td>Sahin</td>
<td>0.37</td>
<td>0.42</td>
</tr>
<tr>
<td>Su</td>
<td><strong>0.54</strong></td>
<td>0.30</td>
</tr>
<tr>
<td>XGBoost</td>
<td>0.52</td>
<td>0.25</td>
</tr>
<tr>
<td>Cao H.</td>
<td>0.24</td>
<td>0.29</td>
</tr>
<tr>
<td>Cao G.</td>
<td>0.28</td>
<td>0.22</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Participant</th>
<th>Micro</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Prec</td>
<td>Rec</td>
</tr>
<tr>
<td>Su</td>
<td><strong>0.80</strong></td>
<td>0.71</td>
</tr>
<tr>
<td>Sahin</td>
<td>0.73</td>
<td><strong>0.74</strong></td>
</tr>
<tr>
<td>XGBoost</td>
<td>0.88</td>
<td>0.57</td>
</tr>
<tr>
<td>Cao G.</td>
<td>0.58</td>
<td>0.66</td>
</tr>
<tr>
<td>Cao H.</td>
<td>0.43</td>
<td><strong>0.79</strong></td>
</tr>
</tbody>
</table>
Results

Submissions:

- **XGBoost baseline** based on TF-IDF document vectors.
- **Sahin et al.** Hierarchical classification with a RoBERTa-base and LSTM, use full documents.
- **Su et al.** Hierarchical (siamese) classification with a RoBERTa-base and CNN, uses the first and last 500 words.
- **Haojie Cao et al.** and **Guiyuan Cao et al.** Classify chunks with RoBERTa-based voting ensemble.
- **Felser et al.** MLP based on aggregate embeddings and topic model features.

<table>
<thead>
<tr>
<th>Participant</th>
<th>Macro</th>
<th>Micro</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sahin</td>
<td>0.37</td>
<td>0.42</td>
</tr>
<tr>
<td>Su</td>
<td><strong>0.54</strong></td>
<td>0.30</td>
</tr>
<tr>
<td>XGBoost</td>
<td>0.52</td>
<td>0.25</td>
</tr>
<tr>
<td>Cao H.</td>
<td>0.24</td>
<td>0.29</td>
</tr>
<tr>
<td>Cao G.</td>
<td>0.28</td>
<td>0.22</td>
</tr>
<tr>
<td>Felser</td>
<td>0.11</td>
<td><strong>0.63</strong></td>
</tr>
</tbody>
</table>

**Particpant**

<table>
<thead>
<tr>
<th>Participant</th>
<th>Prec</th>
<th>Rec</th>
<th>F₁</th>
</tr>
</thead>
<tbody>
<tr>
<td>Su</td>
<td>0.80</td>
<td>0.71</td>
<td><strong>0.75</strong></td>
</tr>
<tr>
<td>Sahin</td>
<td>0.73</td>
<td>0.74</td>
<td>0.74</td>
</tr>
<tr>
<td>XGBoost</td>
<td>0.88</td>
<td>0.57</td>
<td>0.69</td>
</tr>
<tr>
<td>Cao G.</td>
<td>0.58</td>
<td>0.66</td>
<td>0.62</td>
</tr>
<tr>
<td>Cao H.</td>
<td>0.43</td>
<td>0.79</td>
<td>0.56</td>
</tr>
<tr>
<td>Felser</td>
<td>0.27</td>
<td><strong>0.82</strong></td>
<td>0.40</td>
</tr>
</tbody>
</table>
Results

Submissions:

- **XGBoost baseline** based on TF-IDF document vectors.
- **Sahin et al.** Hierarchical classification with a RoBERTa-base and LSTM, use full documents.
- **Su et al.** Hierarchical (siamese) classification with a RoBERTa-base and CNN, uses the first and last 500 words.
- **Haojie Cao et al.** and **Guiyuan Cao et al.** Classify chunks with RoBERTa-based voting ensemble.
- **Felser et al.** MLP based on aggregate embeddings and topic model features.
- **Shashirekha et al.** LSTM based on GloVE embeddings.

<table>
<thead>
<tr>
<th>Participant</th>
<th>Macro</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Prec</td>
<td>Rec</td>
<td>F₁</td>
</tr>
<tr>
<td>Sahin</td>
<td>0.37</td>
<td>0.42</td>
<td>0.352</td>
</tr>
<tr>
<td>Su</td>
<td>0.54</td>
<td>0.30</td>
<td>0.350</td>
</tr>
<tr>
<td>XGBoost</td>
<td>0.52</td>
<td>0.25</td>
<td>0.301</td>
</tr>
<tr>
<td>Cao H.</td>
<td>0.24</td>
<td>0.29</td>
<td>0.228</td>
</tr>
<tr>
<td>Cao G.</td>
<td>0.28</td>
<td>0.22</td>
<td>0.225</td>
</tr>
<tr>
<td>Felser</td>
<td>0.11</td>
<td>0.63</td>
<td>0.161</td>
</tr>
<tr>
<td>Shashirekha</td>
<td>0.10</td>
<td>0.04</td>
<td>0.048</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Participant</th>
<th>Micro</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Prec</td>
<td>Rec</td>
<td>F₁</td>
</tr>
<tr>
<td>Su</td>
<td>0.80</td>
<td>0.71</td>
<td>0.75</td>
</tr>
<tr>
<td>Sahin</td>
<td>0.73</td>
<td>0.74</td>
<td>0.74</td>
</tr>
<tr>
<td>XGBoost</td>
<td>0.88</td>
<td>0.57</td>
<td>0.69</td>
</tr>
<tr>
<td>Shashirekha</td>
<td>0.82</td>
<td>0.50</td>
<td>0.63</td>
</tr>
<tr>
<td>Cao G.</td>
<td>0.58</td>
<td>0.66</td>
<td>0.62</td>
</tr>
<tr>
<td>Cao H.</td>
<td>0.43</td>
<td>0.79</td>
<td>0.56</td>
</tr>
<tr>
<td>Felser</td>
<td>0.27</td>
<td>0.82</td>
<td>0.40</td>
</tr>
</tbody>
</table>
Results

Observations from the Evaluation II:

1. Submissions with good representations of full documents are more effective (0.05–0.06) on long than on short documents.

<table>
<thead>
<tr>
<th></th>
<th>Length</th>
<th>Popularity</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>short</td>
<td>long</td>
</tr>
<tr>
<td>Sahin</td>
<td>0.28</td>
<td>0.34</td>
</tr>
<tr>
<td>Su</td>
<td>0.39</td>
<td>0.27</td>
</tr>
<tr>
<td>XGBoost</td>
<td>0.24</td>
<td>0.29</td>
</tr>
<tr>
<td>Cao, H.</td>
<td>0.23</td>
<td>0.22</td>
</tr>
</tbody>
</table>
Results

Observations from the Evaluation II:

1. Submissions with good representations of full documents are more effective (0.05–0.06) on long than on short documents.

2. Submissions with strong positional representation are more effective on short texts (< 500 words).

<table>
<thead>
<tr>
<th></th>
<th>Length</th>
<th>Popularity</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>short</td>
<td>long</td>
</tr>
<tr>
<td>Sahin</td>
<td>0.28</td>
<td>0.34</td>
</tr>
<tr>
<td>Su</td>
<td>0.39</td>
<td>0.27</td>
</tr>
<tr>
<td>XGBoost</td>
<td>0.24</td>
<td>0.29</td>
</tr>
<tr>
<td>Cao, H.</td>
<td>0.23</td>
<td>0.22</td>
</tr>
</tbody>
</table>
Results

Observations from the Evaluation II:

1. Submissions with good representations of full documents are more effective (0.05–0.06) on long than on short documents.

2. Submissions with strong positional representation are more effective on short texts (< 500 words).

3. Submissions are more effective on popular works.

<table>
<thead>
<tr>
<th></th>
<th>Length</th>
<th>Popularity</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>short</td>
<td>long</td>
</tr>
<tr>
<td>Sahin</td>
<td>0.28</td>
<td>0.34</td>
</tr>
<tr>
<td>Su</td>
<td>0.39</td>
<td>0.27</td>
</tr>
<tr>
<td>XGBoost</td>
<td>0.24</td>
<td>0.29</td>
</tr>
<tr>
<td>Cao, H.</td>
<td>0.23</td>
<td>0.22</td>
</tr>
</tbody>
</table>
Results

Observations from the Evaluation II:

1. Submissions with good representations of full documents are more effective (0.05–0.06) on long than on short documents.

2. Submissions with strong positional representation are more effective on short texts (< 500 words).

3. Submissions are more effective on popular works.

4. Submissions are less effective if documents have many freeform tags (0.06–0.12).

<table>
<thead>
<tr>
<th></th>
<th>Length</th>
<th>Popularity</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>short</td>
<td>long</td>
</tr>
<tr>
<td>Sahin</td>
<td>0.28</td>
<td>0.34</td>
</tr>
<tr>
<td>Su</td>
<td>0.39</td>
<td>0.27</td>
</tr>
<tr>
<td>XGBoost</td>
<td>0.24</td>
<td>0.29</td>
</tr>
<tr>
<td>Cao, H.</td>
<td>0.23</td>
<td>0.22</td>
</tr>
</tbody>
</table>
Observations from the Evaluation II:

1. Submissions with good representations of full documents are more effective (0.05–0.06) on long than on short documents.

2. Submissions with strong positional representation are more effective on short texts (< 500 words).

3. Submissions are more effective on popular works.

4. Submissions are less effective if documents have many freeform tags (0.06–0.12).

5. Submissions are less effective if documents have the "Choose Not To Use Archive Warnings" declaration (0.04–0.06).

<table>
<thead>
<tr>
<th></th>
<th>Length</th>
<th>Popularity</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>short</td>
<td>long</td>
</tr>
<tr>
<td>Sahin</td>
<td>0.28</td>
<td>0.34</td>
</tr>
<tr>
<td>Su</td>
<td>0.39</td>
<td>0.27</td>
</tr>
<tr>
<td>XGBoost</td>
<td>0.24</td>
<td>0.29</td>
</tr>
<tr>
<td>Cao, H.</td>
<td>0.23</td>
<td>0.22</td>
</tr>
</tbody>
</table>
Results

Observations from the Evaluation I:

6. Submissions are effective for common and less effective for rare warnings.

<table>
<thead>
<tr>
<th></th>
<th>Porn.</th>
<th></th>
<th>Common</th>
<th></th>
<th>Rare</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P</td>
<td>R</td>
<td>P</td>
<td>R</td>
<td>P</td>
</tr>
<tr>
<td>Sahin</td>
<td>0.95</td>
<td>0.96</td>
<td>0.62</td>
<td>0.48</td>
<td>0.12</td>
</tr>
<tr>
<td>Su</td>
<td>0.90</td>
<td>0.97</td>
<td>0.61</td>
<td>0.43</td>
<td>0.57</td>
</tr>
<tr>
<td>Cao H.</td>
<td>0.86</td>
<td>0.98</td>
<td>0.22</td>
<td>0.61</td>
<td>0.16</td>
</tr>
</tbody>
</table>
Results

Observations from the Evaluation I:

6. Submissions are effective for common and less effective for rare warnings.

7. Submissions favor either precision or recall, independently of overall effectiveness.

<table>
<thead>
<tr>
<th></th>
<th>Porn.</th>
<th>Common</th>
<th>Rare</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P</td>
<td>R</td>
<td>P</td>
</tr>
<tr>
<td>Sahin</td>
<td>0.95</td>
<td>0.96</td>
<td>0.62</td>
</tr>
<tr>
<td>Su</td>
<td>0.90</td>
<td>0.97</td>
<td>0.61</td>
</tr>
<tr>
<td>Cao H.</td>
<td>0.86</td>
<td>0.98</td>
<td>0.22</td>
</tr>
</tbody>
</table>
Results

Observations from the Evaluation I:

6. Submissions are effective for common and less effective for rare warnings.

7. Submissions favor either precision or recall, independently of overall effectiveness.

8. An ensemble of the (best) submissions improves $F_1$ marginally (0.01–0.03).

<table>
<thead>
<tr>
<th></th>
<th>Porn.</th>
<th>Common</th>
<th>Rare</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P</td>
<td>R</td>
<td>P</td>
</tr>
<tr>
<td>Sahin</td>
<td>0.95</td>
<td>0.96</td>
<td>0.62</td>
</tr>
<tr>
<td>Su</td>
<td>0.90</td>
<td>0.97</td>
<td>0.61</td>
</tr>
<tr>
<td>Cao H.</td>
<td>0.86</td>
<td>0.98</td>
<td>0.22</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Macro $F_1$</th>
<th>Micro $F_1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sahin</td>
<td>0.35</td>
<td>0.74</td>
</tr>
<tr>
<td>Su</td>
<td>0.35</td>
<td>0.75</td>
</tr>
<tr>
<td>XGBoost</td>
<td>0.30</td>
<td>0.69</td>
</tr>
<tr>
<td>Ensemble (Top 3)</td>
<td><strong>0.36</strong></td>
<td><strong>0.77</strong></td>
</tr>
</tbody>
</table>
Results

Observations from the Evaluation I:

6. Submissions are effective for common and less effective for rare warnings.

7. Submissions favor either precision or recall, independently of overall effectiveness.

8. An ensemble of the (best) submissions improves $F_1$ marginally (0.01–0.03).

<table>
<thead>
<tr>
<th></th>
<th>Porn.</th>
<th>Common</th>
<th>Rare</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P</td>
<td>R</td>
<td>P</td>
</tr>
<tr>
<td>Sahin</td>
<td>0.95</td>
<td>0.96</td>
<td>0.62</td>
</tr>
<tr>
<td>Su</td>
<td>0.90</td>
<td>0.97</td>
<td>0.61</td>
</tr>
<tr>
<td>Cao H.</td>
<td>0.86</td>
<td>0.98</td>
<td>0.22</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Macro $F_1$</th>
<th>Micro $F_1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sahin</td>
<td>0.35</td>
<td>0.74</td>
</tr>
<tr>
<td>Su</td>
<td>0.35</td>
<td>0.75</td>
</tr>
<tr>
<td>XGBoost</td>
<td>0.30</td>
<td>0.69</td>
</tr>
<tr>
<td>Ensemble (Top 3)</td>
<td><strong>0.36</strong></td>
<td><strong>0.77</strong></td>
</tr>
</tbody>
</table>

Contact matti.wiegmann@uni-weimar.de