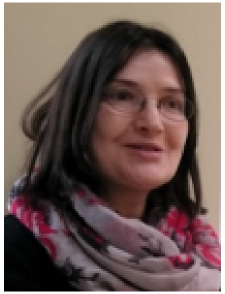
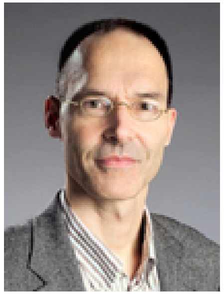
Overview of the Trigger Detection Task at PAN 2023



Matti Wiegmann | Magdalena Wolska | Martin Potthast | Benno Stein

Bauhaus-Universität Weimar | Leipzig University | ScaDS.AI

webis.de

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.”

$\xrightarrow{\text{evokes}}$ Memories of a past war.

$\xrightarrow{\text{triggers}}$ 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.”

$\xrightarrow{\text{evokes}}$ Memories of a past war.

$\xrightarrow{\text{triggers}}$ 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."

$\xrightarrow{\text{evokes}}$ Memories of a past war.

$\xrightarrow{\text{triggers}}$ Anxiety, feelings of loss or grief, . . .

Trigger Detection at PAN 2023:

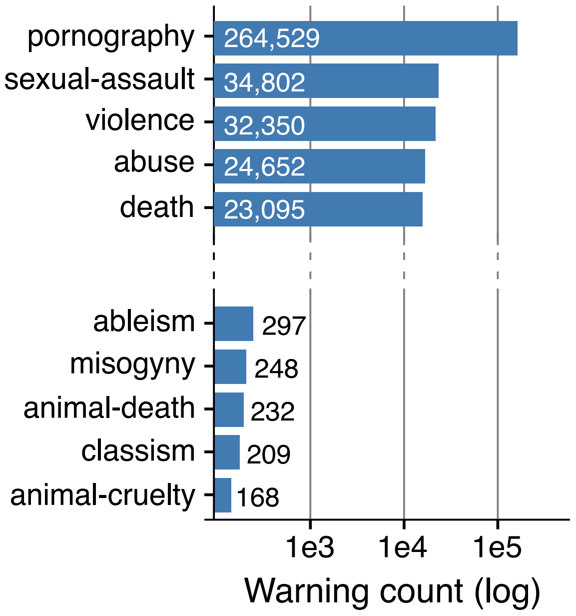
Given a fan fiction document, assign all appropriate trigger warnings from the given label set.

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.



Task Overview

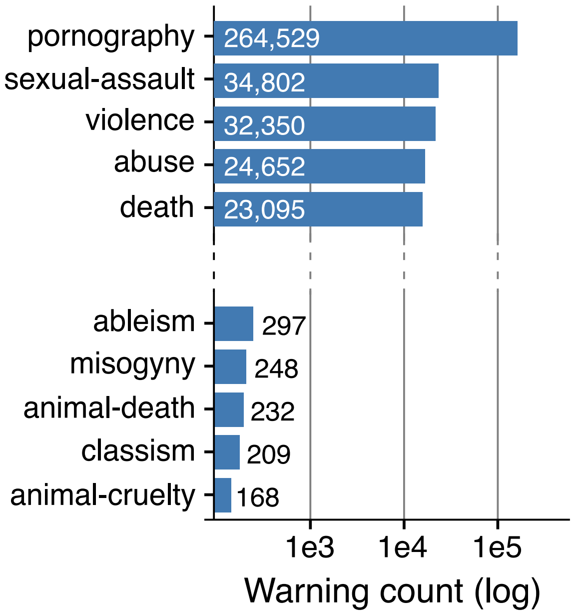
Dataset:

- 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.



Task Overview

Dataset:

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

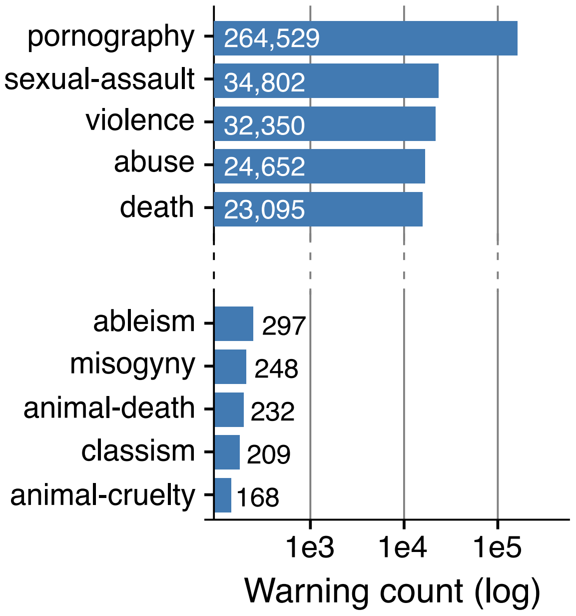
Evaluation:

- 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.



Dataset

Trigger Warning Taxonomy:

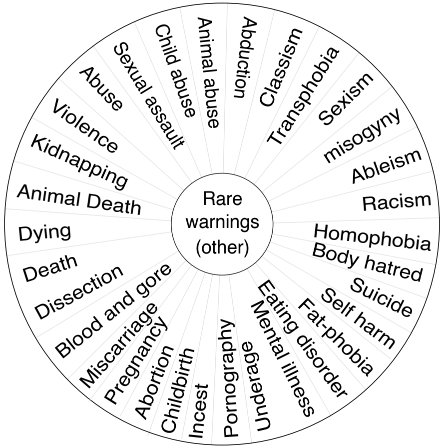
- We curated a trigger warning taxonomy based on university guidelines (Michigan and Reading).

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, ...`

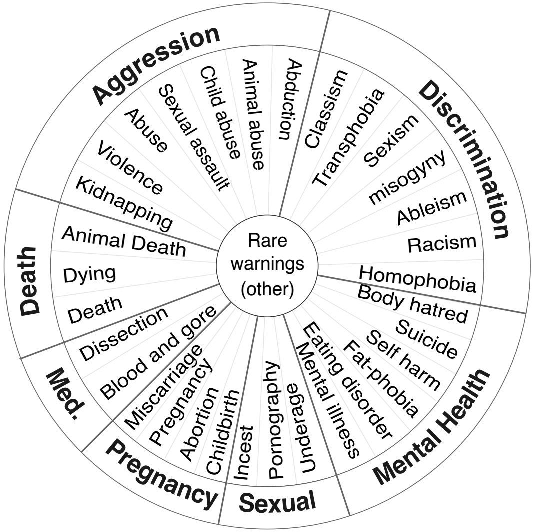


Dataset

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).

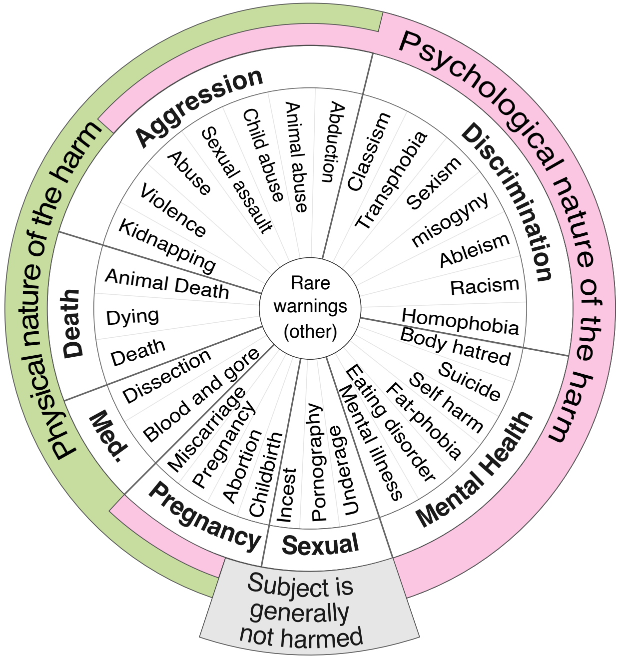


Dataset

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

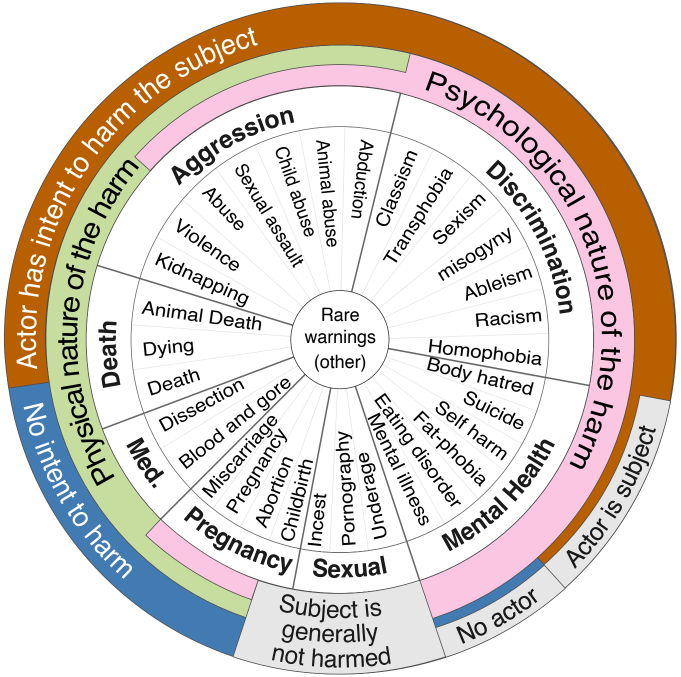
Dataset

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.



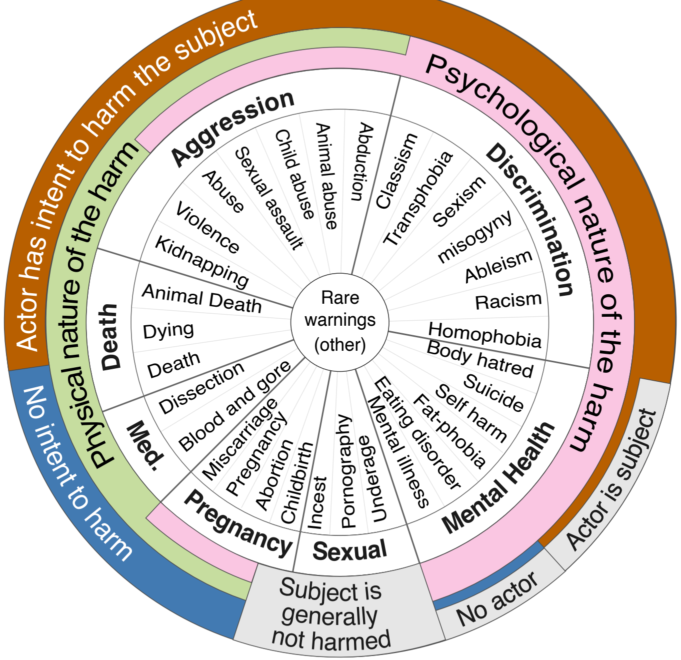
@Wiegmann, 2023

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.



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, 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]

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)

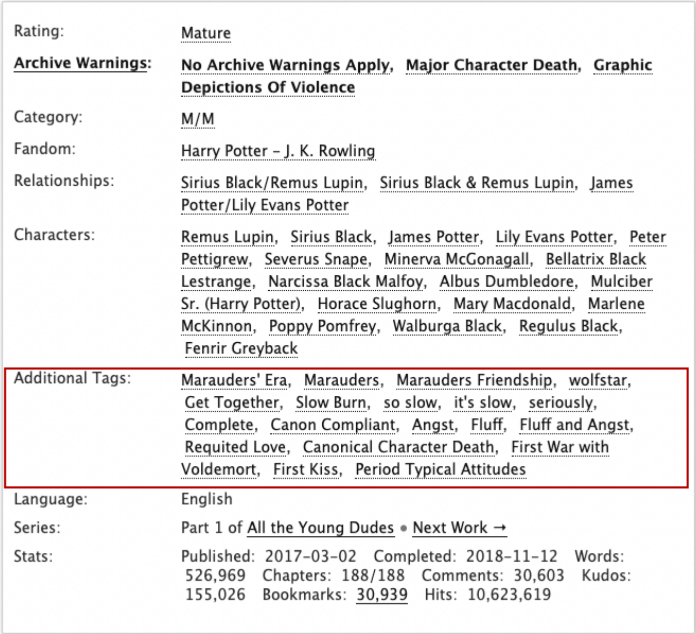
Rating:	Mature
Archive Warnings:	**No Archive Warnings Apply, Major Character Death, Graphic Depictions Of Violence**
Category:	M/M
Fandom:	Harry Potter - J. K. Rowling
Relationships:	Sirius Black/Remus Lupin, Sirius Black & Remus Lupin, James Potter/Lily Evans Potter
Characters:	Remus Lupin, Sirius 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]

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.



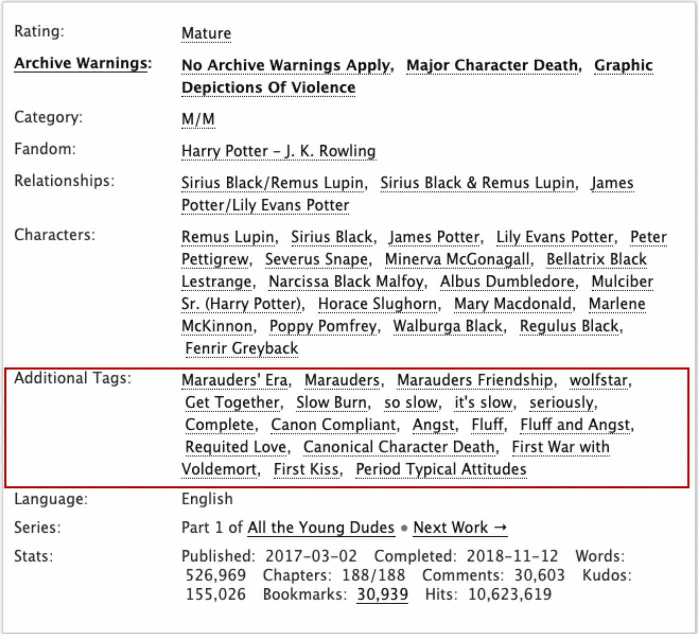
Rating:	Mature
Archive Warnings:	No Archive Warnings Apply, Major Character Death, Graphic Depictions Of Violence
Category:	M/M
Fandom:	Harry Potter - J. K. Rowling
Relationships:	Sirius Black/Remus Lupin, Sirius Black & Remus Lupin, James Potter/Lily Evans Potter
Characters:	Remus Lupin, Sirius Black, 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]

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.



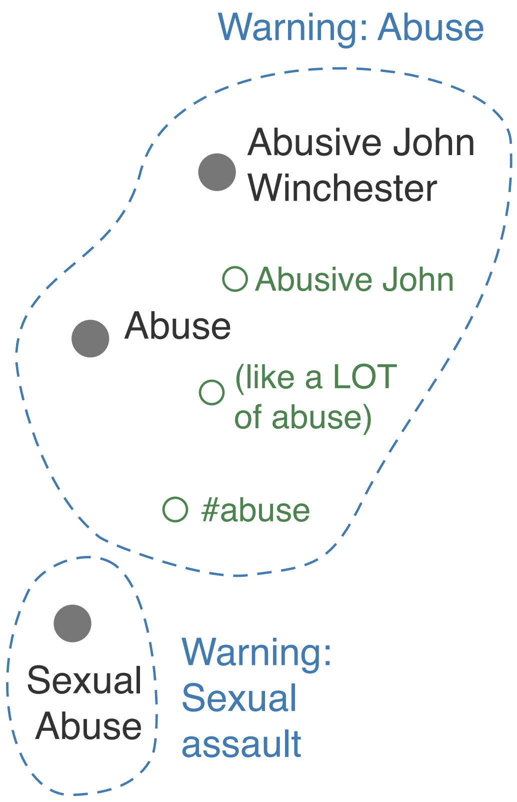
Rating:	Mature
Archive Warnings:	No Archive Warnings Apply, Major Character Death, Graphic Depictions Of Violence
Category:	M/M
Fandom:	Harry Potter - J. K. Rowling
Relationships:	Sirius Black/Remus Lupin, Sirius Black & Remus Lupin, James Potter/Lily Evans Potter
Characters:	Remus Lupin, Sirius Black, 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]

Dataset

Determining warning labels:

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

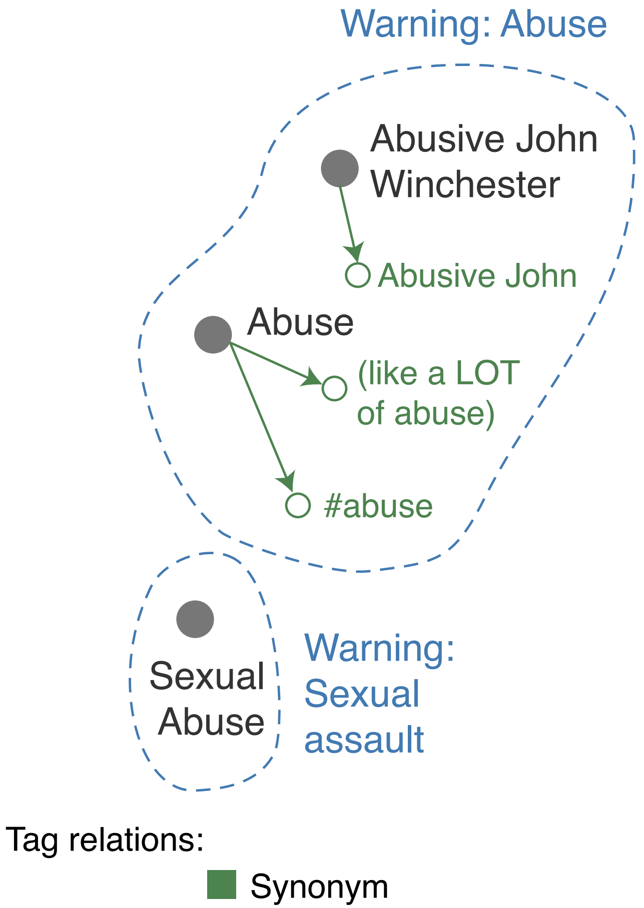


Tag relations:

Dataset

Determining warning labels:

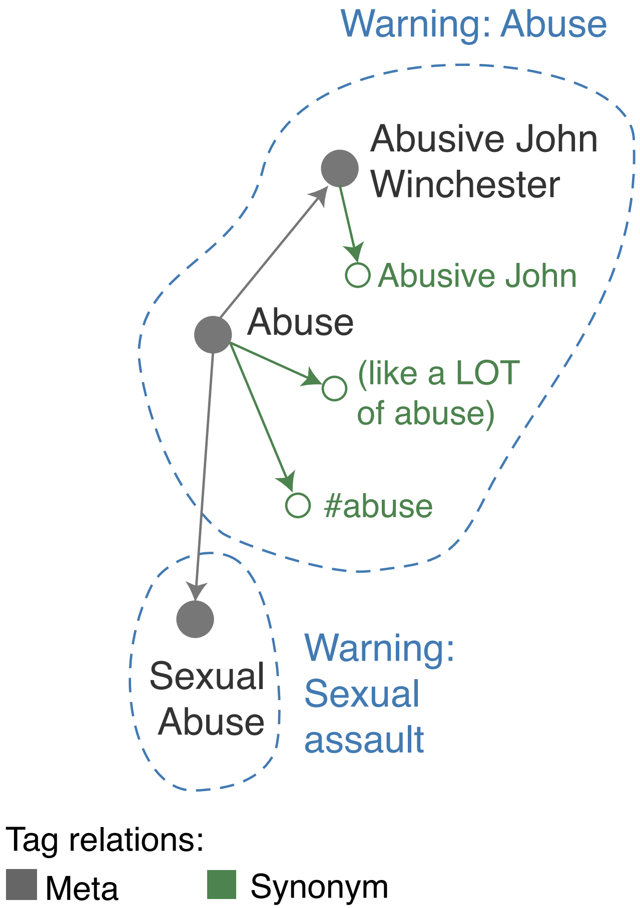
- Freeform tags are related through tag relations that were added by community experts
  ➜ Semi-automatic annotation.
- Synonymous tags are related.
  One synonym is marked as *canonical*.



Dataset

Determining warning labels:

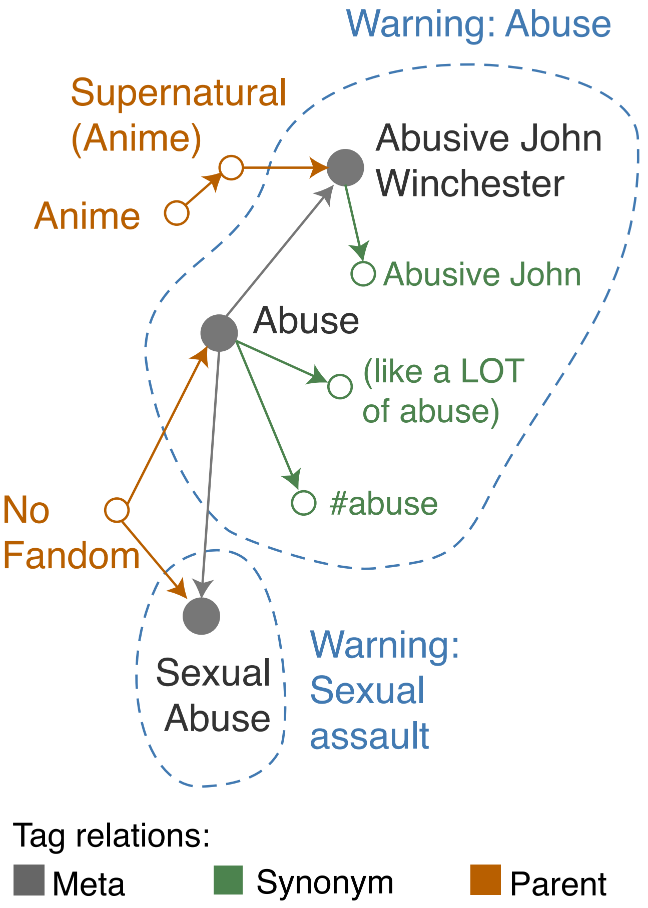
- Freeform tags are related through tag relations that were added by community experts
  ➜ Semi-automatic annotation.
- Synonymous tags are related.
  One synonym is marked as *canonical*.
- Canonical tags are in a meta-sub relation.
  Sources were annotated with a warning.



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.



Results

Submissions:

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

Participant	Macro		
	Prec	Rec	F_1
XGBoost	0.52	0.25	0.301

Participant	Micro		
	Prec	Rec	F_1
XGBoost	0.88	0.57	0.69

Results

Submissions:

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

Participant	Macro		
	Prec	Rec	F_1
Sahin	0.37	0.42	**0.352**
XGBoost	0.52	0.25	0.301

Participant	Micro		
	Prec	Rec	F_1
Sahin	0.73	0.74	0.74
XGBoost	0.88	0.57	0.69

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.

Participant	Macro		
	Prec	Rec	F_1
Sahin	0.37	0.42	**0.352**
Su	**0.54**	0.30	0.350
XGBoost	0.52	0.25	0.301

Participant	Micro		
	Prec	Rec	F_1
Su	0.80	0.71	**0.75**
Sahin	0.73	0.74	0.74
XGBoost	0.88	0.57	0.69

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.

Participant	**Macro**		
	Prec	Rec	F_1
Sahin	0.37	0.42	**0.352**
Su	**0.54**	0.30	0.350
XGBoost	0.52	0.25	0.301
Cao H.	0.24	0.29	0.228
Cao G.	0.28	0.22	0.225

Participant	**Micro**		
	Prec	Rec	F_1
Su	0.80	0.71	**0.75**
Sahin	0.73	0.74	0.74
XGBoost	0.88	0.57	0.69
Cao G.	0.58	0.66	0.62
Cao H.	0.43	0.79	0.56

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.

Participant	Macro		
	Prec	Rec	F_1
Sahin	0.37	0.42	**0.352**
Su	**0.54**	0.30	0.350
XGBoost	0.52	0.25	0.301
Cao H.	0.24	0.29	0.228
Cao G.	0.28	0.22	0.225
Felser	0.11	**0.63**	0.161

Participant	Micro		
	Prec	Rec	F_1
Su	0.80	0.71	**0.75**
Sahin	0.73	0.74	0.74
XGBoost	0.88	0.57	0.69
Cao G.	0.58	0.66	0.62
Cao H.	0.43	0.79	0.56
Felser	0.27	**0.82**	0.40

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.

Participant	**Macro**		
	Prec	Rec	F_1
Sahin	0.37	0.42	**0.352**
Su	**0.54**	0.30	0.350
XGBoost	0.52	0.25	0.301
Cao H.	0.24	0.29	0.228
Cao G.	0.28	0.22	0.225
Felser	0.11	**0.63**	0.161
Shashirekha	0.10	0.04	0.048

Participant	**Micro**		
	Prec	Rec	F_1
Su	0.80	0.71	**0.75**
Sahin	0.73	0.74	0.74
XGBoost	0.88	0.57	0.69
Shashirekha	0.82	0.50	0.63
Cao G.	0.58	0.66	0.62
Cao H.	0.43	0.79	0.56
Felser	0.27	**0.82**	0.40

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.

	Length		Popularity	
	short	long	low	high
Sahin	0.28	**0.34**	0.30	0.35
Su	**0.39**	0.27	0.22	0.35
XGBoost	0.24	0.29	0.16	0.30
Cao, H.	0.23	0.22	0.19	0.22

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).

	Length		**Popularity**	
	short	long	low	high
Sahin	0.28	**0.34**	0.30	0.35
Su	**0.39**	0.27	0.22	0.35
XGBoost	0.24	0.29	0.16	0.30
Cao, H.	0.23	0.22	0.19	0.22

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.

	Length		**Popularity**	
	short	long	low	high
Sahin	0.28	**0.34**	0.30	0.35
Su	**0.39**	0.27	0.22	0.35
XGBoost	0.24	0.29	0.16	0.30
Cao, H.	0.23	0.22	0.19	0.22

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).

	Length		**Popularity**	
	short	long	low	high
Sahin	0.28	**0.34**	0.30	0.35
Su	**0.39**	0.27	0.22	0.35
XGBoost	0.24	0.29	0.16	0.30
Cao, H.	0.23	0.22	0.19	0.22

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

	Length		**Popularity**	
	short	long	low	high
Sahin	0.28	**0.34**	0.30	0.35
Su	**0.39**	0.27	0.22	0.35
XGBoost	0.24	0.29	0.16	0.30
Cao, H.	0.23	0.22	0.19	0.22

Results

Observations from the Evaluation I:

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

	Porn.		**Common**		**Rare**	
	P	R	P	R	P	R
Sahin	0.95	0.96	**0.62**	0.48	0.12	0.51
Su	0.90	0.97	0.61	0.43	0.57	0.19
Cao H.	0.86	0.98	0.22	0.61	0.16	0.12

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.

	Porn.		**Common**		**Rare**	
	P	R	P	R	P	R
Sahin	0.95	0.96	**0.62**	0.48	0.12	0.51
Su	0.90	0.97	0.61	0.43	0.57	0.19
Cao H.	0.86	0.98	0.22	0.61	0.16	0.12

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).

	Porn.		Common		Rare	
	P	R	P	R	P	R
Sahin	0.95	0.96	**0.62**	0.48	0.12	0.51
Su	0.90	0.97	0.61	0.43	0.57	0.19
Cao H.	0.86	0.98	0.22	0.61	0.16	0.12

	Macro F_1	**Micro F_1**
Sahin	**0.35**	0.74
Su	0.35	**0.75**
XGBoost	0.30	0.69
Ensemble (Top 3)	**0.36**	**0.77**

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).

	Porn.		**Common**		**Rare**	
	P	R	P	R	P	R
Sahin	0.95	0.96	**0.62**	0.48	0.12	0.51
Su	0.90	0.97	0.61	0.43	0.57	0.19
Cao H.	0.86	0.98	0.22	0.61	0.16	0.12

	Macro F_1	**Micro F_1**
Sahin	**0.35**	0.74
Su	0.35	**0.75**
XGBoost	0.30	0.69
Ensemble (Top 3)	**0.36**	**0.77**

Contact `matti.wiegmann@uni-weimar.de`