Proceedings of the Nineteenth International AAAI Conference on Web and Social Media (ICWSM 2025)

Examining the Makeup of Media Trigger Warnings Online

Peixian Zhang[1], **Yupeng He**[1], **Ehsan-Ul Haq**[1], **Gareth Tyson**[12]

[1] Hong Kong University of Science and Technology (Guangzhou)
[2] Queen Mary University of London
{pzhang041, yhe382}@connect.hkust-gz.edu.cn, euhaq@hkust-gz.edu.cn, gtyson@ust.hk

Abstract

In today's digital landscape, the prevalence of sensitive online content has made *trigger warnings* essential. These warnings inform viewers that the content they are about to see contains sensitive artifacts (*e.g.* violence). This paper studies the use of trigger warnings, exploiting data from two major platforms: Does the Dog Die, a crowdsourcing trigger warnings platform, and IMDb, a media database. We first study how different media types (*e.g.* films, video games, and TV shows) are labeled with varying trigger warnings and the different co-occurrence patterns among different trigger warnings. We also discover controversy surrounding certain trigger warnings, with inconsistent opinions stated by different people. We further show that different jurisdictions (*e.g.* USA vs. UK) assign different content ratings (*e.g.* R-18) for the same media, even when the same trigger warnings are present. Finally, we develop automatic detectors to identify trigger warnings from IMDb text. We achieve F1 scores exceeding 0.7 for all 10 selected trigger warnings.

Introduction

Trigger warnings have become increasingly common in various mediums, including books, movies, and online media media (Stratta, Park, and deNicola 2020; Wiegmann et al. 2023; Blackwell et al. 2019; Ling, Gummadi, and Zannettou 2023). A trigger warning[1] is a statement warning people that they may find certain parts of the content (*e.g.* blog post or movie) upsetting (Charles et al. 2022). It can also be indicated by ratings, often using age suitability for the content, such as over-13 (Barranco, Rader, and Smith 2017). Prior research shows that people value such warnings for sensitive topics such as sexual assault and suicide (Boysen et al. 2021; Bryce et al. 2023; Lockhart 2016). Given that online content includes material that may be disturbing to some individuals, it has also become increasingly noticeable in online communities (Bridgland, Bellet, and Takarangi 2023; Scott et al. 2023). Consequently, there have been attempts to standardize trigger warnings for online environments (Bryce et al. 2023; Hirsch 2020; Charles et al. 2022).

There are significant challenges in establishing standardized and systematic trigger warnings for online content due to the variance in content types and the demographics of their targeted audience. We identify two specific challenges: **(i) Complexity of content:** The multimodality of online platforms enables users to consume all types of content, including books, movies and video games. Earlier research has highlighted that movies and video games may be associated with violence, and risky behavior-related content, respectively (Charles et al. 2022). We currently lack a comprehensive analysis of content-specific trigger warnings. This could help uncover multiple viewing risks associated with each type. **(ii) Diverse sensitivities of users:** The use of trigger warning varies with demographics (Charles et al. 2022), and the effectiveness of trigger warnings is closely linked with individual experiences (Blackwell et al. 2019). There is currently a lack of knowledge on how such trigger warnings are interpreted across different countries and platforms.

In this paper, we characterize trigger warnings based on the type of content and the target users (*e.g.* based on geography). We focus on two platforms: the crowd-sourcing trigger warning platform, *Does the Dog Die* (referred to as DTDD), and a media content information database, *IMDb*. DTDD enables users to create trigger warnings by posing specific questions about potential triggers in media content (*e.g.* a movie or a video game). These questions are then answered by other users. This question-and-answer format offers users the flexibility to identify specific trigger warnings they care about. We use the warnings and rating systems to explore the cultural and social sensitivity across different countries (Barranco, Rader, and Smith 2017). This allows us to gain a comprehensive understanding of content sensitivity, and the interplay between specific trigger warnings and rating systems (*e.g.* R-18, 13+). We answer the following research questions:

RQ1: What are the differences between the trigger warnings associated with different types of content (movie, TV series, video games)?

RQ2: To what extent are individual jurisdictions (*e.g.* USA, Europe, Asia) sensitive to different types of trigger warnings? How do these differences manifest in the rating systems (*e.g.* R-18, 15+) of each country?

RQ3: Can we build a tool to automatically identify trigger warnings, using free-text information (*e.g.* film descriptions)?

Copyright © 2025, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

[1] Sometimes referred to as a content warning

By answering these questions, we make several key findings:

- The distribution of trigger warnings vary across different content types (*i.e.* movies, video games, and TV series). For example, TV series have a higher percentage of trigger warnings related to mental and physical health as compared to movies.
- Users demonstrate diverse sensitivities to different trigger warnings, as demonstrated by their level of agreement/disagreement (captured via votes). Trigger warnings related to `Diversity and Inclusion` create the greatest disagreement among users.
- There are differences between the official content ratings and the user-promoted trigger warnings. For instance, we find G-rated movies (deemed fit for all ages) tagged with `Social Concerns` and `Fear and Aversions` trigger warnings.
- The USA has stricter ratings compared to other jurisdictions. This shows that the use of a non-USA rating in the USA may result in an unexpected viewing experiences. For instance, Mexico has lower age rating for alcohol and drug-related content as compared to the USA.

Building on these findings, we develop automated detectors to identify trigger warnings using free-text from items on IMDb. We achieve F1 scores exceeding 0.7 for identifying 10 popular trigger warnings. Notably, we obtain the highest F1 score (81.13%) for the trigger warning "Is there body horror?", and 76.73% for "Are there jump scares?". These results underscore the potential of our detectors in automating the detection of trigger warnings.

Related Work

Online Content Trigger Warnings

Trigger warnings originated from trauma therapy, and then spread to the internet in an attempt to help viewers identify problematic content (Knox 2017). Previous research indicates that online warnings enhance the experiences of communities (Dillahunt et al. 2017; Scott et al. 2023). A recent study (Blackwell et al. 2019) examines the limitations of trigger content within social virtual reality environments. However, this research is primarily focused on a specific context and lacks broader investigation into the application of trigger warnings in more generalized settings. (Charles et al. 2022) proposed the Narrative Experiences Online (NEON) taxonomy, a framework for multimedia warnings that encompasses a diverse array of data sources, including online content. This taxonomy serves as a foundational framework for future research in trigger warnings. This work also highlights the varying demands for diversity across different types of media content. (Wiegmann et al. 2023) provides a more precise definition grounded in a semantically motivated classification system, utilizing a dataset derived from the fan fiction community. Recently, (Horne 2024) explored the feasibility of providing warning labels using artificial intelligence. In contrast to our work, this prior work focuses on warning labels related to false information on social media platforms, whereas we focus on trigger warnings for content.

Automatic Trigger Warning Generation

(Stratta, Park, and deNicola 2020) was the first work on automatic trigger warnings, building a browser plugin. The authors conclude that client-side warnings are feasible and that users respond positively. However, this study is limited in scope, as it only addresses sexual assault warnings using a simple dictionary-based approach. Subsequently, (Wolska et al. 2022) examined binary document classification of fan fiction documents. Another study (Zhang et al. 2021) analyzes trigger warning classification based on the IMDb Parents Guide information, whereas (Khan et al. 2019) focuses on predicting instances of violence. Later, (Wiegmann et al. 2023) extended the previous work and trained a multi-label model. (Wiegmann et al. 2024) proposes further work in the dictionary-based approach to detect passage-level trigger warnings based on the same fan fiction dataset. The limitations of these works include the reliance on datasets from a single fan fiction platform, which may not provide comprehensive insights applicable to other contexts.

Our work is very different to the above. We investigate trigger warnings using a unique user-generated ask-and-answer dataset from DTDD. Our analysis explores the differences in trigger warnings across various types of media. Additionally, we evaluate overall social content sensitivity by analyzing trigger warnings in conjunction with content rating systems. By integrating datasets from DTDD and IMDb, we further develop models to automatically detect user-generated trigger warnings using both machine learning models and large language model (LLM) methods.

Dataset

In this section, we introduce our data processing pipeline, including the data collection from *Does the Dog Die* (DTDD) and *IMDb*.

Data Collection

To answer our RQs, we need both crowdsourcing trigger warnings and rating systems from different countries. Thus, we collect data from DTDD and *IMDb*. DTDD[2] is a crowdsourcing trigger warnings platform, and *IMDb* is an online database containing movie-related information. Because both platforms include more than one type of media (movies, TV series, and video games), we use the term *item* in the following sections.

Does the Dog Die (DTDD) DTDD is an online crowdsourced trigger warning community site, which covers films, TV, books, and video games. On this site, users pose questions, *e.g.* asking "does the dog die"? Then, the questions are answered by other users by voting yes or no, alongside optional text to explain their stance. If the number of yes votes is larger than no, it is labeled as a valid trigger warning for the item. There are 808 media items mentioned on DTDD. Each media item includes basic information such as its type

[2] https://www.doesthedogdie.com/

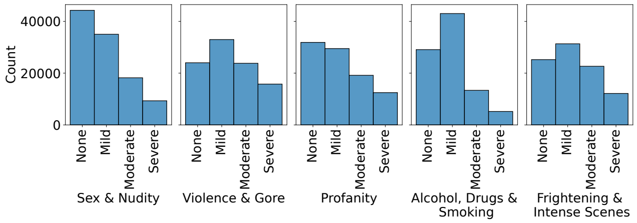


Figure 1: Level distribution for each category of trigger warnings on IMDb. The y-axis is the number of films in each level.

(*e.g.* movie or TV series), the year it was published, and the corresponding ID number on IMDb. The latter allows us to easily link DDTD entries to metadata on IMBd. Each media item can be associated with trigger warnings, taken from a standardized list of 197. If a new trigger warning question is proposed, it must receive more than 1,000 votes from the supporters (paying users) before being added to this standard list. Once approved, the new question is automatically added to the list for all media items. These trigger warnings are classified into 34 groups by the platform, such as Race, Relationships, Mental Health, etc. Overall, there are 19,889 unique users with an average of 4.5 comments and 10.2 votes. The full list of trigger warnings is in Appendix .

IMDb IMDb is an online database of information related to films, television series, etc. It provides basic information including directors, characters, ratings, and so on. We compile a large IMDb dataset using the IMDBpy python library. This covers all films on DTDD, including the Parents Guide information provided by IMDb. The Parents Guide provides official age-sensitive rating information (from several countries), and a voting function that allows people to vote if the film contains certain types of sensitive content. The vote function covers five categories of trigger warnings: (*i*) Sex&Nudity, (*ii*) Violence&Gore, (*iii*) Profanity&Alcohol, (*iv*) Drugs&Smoking, and (*v*) Frightening&Intense Scenes. Each of the five categories offers four levels of trigger warnings: None, Mild, Moderate, and Severe. Users can therefore vote for which level each category has, *e.g.* mild Profanity&Alcohol. Based on the votes, IMDb assigns a final trigger warning level for each category. Note, not all the films have the Parents Guide information. In total, there are 114,283 items with 478,462 trigger warning levels. For context, Figure 1 plots the distribution of levels, for each individual category of trigger warning.

We then link each DTDD entry to its equivalent IMDB page (recall, DDTD contains a link to the equivalent IMDb page for each entry). Our dataset contains 676 items (434 movies, 208 TV series, and 34 video games) with 37,697 trigger warnings on DTDD. We collect the description and storyline of each item from IMDb. Note, 16 TV series are tagged as "TV shows" on DTDD. After manual confirmation, we switch these labels to "TV series". The complete list of item IMDb ids are provided in the supplementary materials.[3]

Data Pre-processing

Before analysis, we perform a number of pre-processing steps.

Categorizing Trigger Warnings Recall, DTDD has 34 warning groups for 197 categories, with the number of warnings per group varying from 1 to 34. A manual review of these groups shows that there are several similar warnings. For example, the Children group includes a trigger warning "Does a kid die", while the Death group contains a trigger warning "Does someone die", and the Family group includes "Does a family member die". Thus, we merge similar groups to overcome the imbalance in between groups and make the categories more coherent.

For this, we first filter any two trigger warnings with similarity ($>= 0.8$) by sequence comparison using `difflib`. 23 trigger warning pairs are classed as similar using this approach. Following this, the first created is saved. We then manually classify the remaining groups according to the taxonomy from multimedia triggers of Narrative Experience Online (NEON) (Charles et al. 2022). This is a systematic typology of content warnings covering five factors (Charles et al. 2022). Table 1 shows our distribution across the five categories. To provide brief insight into the categories, we generate the top words in each category by TF-IDF in Figure 8 of the Appendix.

The description of each category is below:

- **Diversity and Inclusion** refers to the groups that are often stereotyped or overall negatively portrayed in films (Thomson 2021), such as those related to ethnicity and sexual orientation.
- **Fear and Aversions** covers trigger warnings relevant to personal fears and traumatic experiences such as PTSD. This type of content can aggravate a viewer's condition (Bellet, Jones, and McNally 2018).
- **Mental and Physical Health** These warnings are concerned with well-being, with a special focus on mental health, *e.g.* content related to suicide and disability.
- **Social Concerns** This category refers to warnings related to social aspects such as family, child safety, and law enforcement.
- **Violence and Sex** This refers to violent or sexual content such as gore, violence, and obscenity.

Characterizing Trigger Warnings in Different Content

We start by characterizing the trigger warnings that we observe across all content types (**RQ1**).

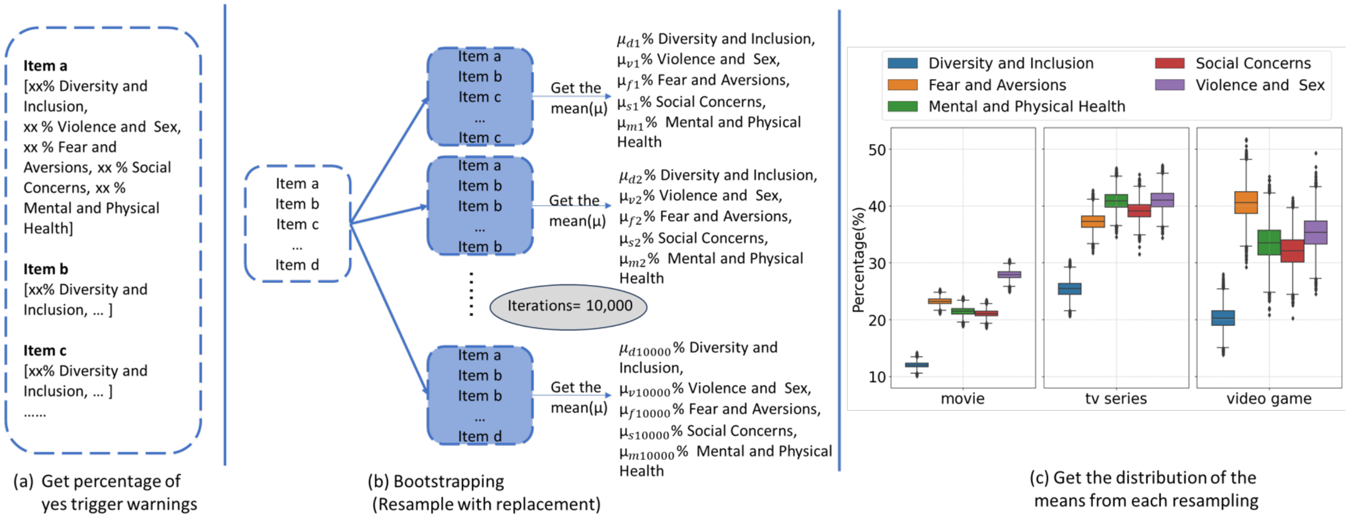
Distribution of Trigger Warnings

We first analyze whether the distribution of warnings is the same among the three content types (movies, TV, video games).

[3] https://github.com/PeixianZhang/Trigger-Warning-Dataset

Category(n)	Definition	Sub-categories(groups)	Examples
Diversity and Inclusion (n=20)	Content depicts negative stereotypes about or attitudes towards a specific group, such as racism or sexism.	LGBTQ+, Prejudice, Race, Religious, Sexism	*is there bisexual cheating*
Fear and Aversions (n=39)	Content contains imagery, sounds, or effects that may frighten, disgust, or scare	Animal, Creepy Crawly, Fear, Gross, Loss, Noxious, Paranoia	*does someone wet/soil themselves*
Mental and Physical Health (n=32)	Content relates to mental health issues.	Addiction, Disability, Drugs/Alcohol, Medical, Mental Health, Sickness	*does someone attempt suicide*
Social Concerns (n=32)	Content includes social or political issues.	Abandonment, Abuse, Children, Family, Law Enforcement, Pregnancy, Social	*is a child abandoned by a parent*
Violence and Sex (n=51)	Content contains both violence and sexual themes. The sexual themes include nudity, sexual content, and relationships.	Assault, Bodily Harm, Large-scale Violence, Sex, Relationships, Vehicular, Death, Violence	*is there gun violence*

Table 1: The category of the trigger warnings with definition and examples. **n** is the number of trigger warnings in the category.


Item a
[xx% Diversity and Inclusion, xx % Violence and Sex, xx % Fear and Aversions, xx % Social Concerns, xx % Mental and Physical Health]
Item b
[xx% Diversity and Inclusion, …]
Item c
[xx% Diversity and Inclusion, …]
……
Item a
Item b
Item c
…
Item d
Item a
Item b
Item c
…
Item c
Get the mean(μ)
μ_{d1}% Diversity and Inclusion,
μ_{v1}% Violence and Sex,
μ_{f1}% Fear and Aversions,
μ_{s1}% Social Concerns,
μ_{m1}% Mental and Physical Health
Item a
Item b
Item b
…
Item b
Get the mean(μ)
μ_{d2}% Diversity and Inclusion,
μ_{v2}% Violence and Sex,
μ_{f2}% Fear and Aversions,
μ_{s2}% Social Concerns,
μ_{m2}% Mental and Physical Health
Iterations= 10,000
Item a
Item b
Item b
…
Item d
Get the mean(μ)
μ_{d10000}% Diversity and Inclusion,
μ_{v10000}% Violence and Sex,
μ_{f10000}% Fear and Aversions,
μ_{s10000}% Social Concerns,
μ_{m10000}% Mental and Physical Health
Diversity and Inclusion
Fear and Aversions
Mental and Physical Health
Social Concerns
Violence and Sex
Percentage(%)
50
40
30
20
10
movie
tv series
video game

(a) Get percentage of yes trigger warnings

(b) Bootstrapping (Resample with replacement)

(c) Get the distribution of the means from each resampling

Figure 2: The figure explains the bootstrapping process on the dataset. (a) for the first step to get the yes trigger warning in each category, normalize by percentage. (b) for the resampling process (c) The percentage of included trigger warnings within the categories in different types of content.

Frequency of Trigger Warnings. First, we measure the distribution of trigger warnings across all items. However, the limited number of samples in the dataset may lead to skewed results. To alleviate this issue, we employ bootstrapping (Mooney, Duval, and Duvall 1993). Bootstrapping is commonly used for reliable estimation of statistical properties by resampling the available data with replacement (Lukin and Walker 2017). The detailed steps are as follows.

The first step is shown in Figure 2(a). For each item, we calculate the percentage of assigned warning labels for each of the five warning categories. For example, if a movie is labeled with 4 trigger warnings out of a possible 20 in the `Diversity and Inclusion` category, the percentage usage of this category for the movie is 20%. We then calculate the percentages of five warning categories for all items across the three content types (movies, TV, video games). Therefore, each item obtains 5 percentage values for the corresponding 5 warning categories.

The next step is shown in Figure 2(b), where we process the resampling. We divide all items into three parts (movies, TV, video games) based on their content types. Each time we perform resampling with replacement, we generate a dataset that is the same size as the original data. Then, we count the mean value of the percentages of each category in the resampled dataset. Finally, for each content item, we have 10,000 mean percentages of each trigger warning category.

Figure 2(c) displays the distribution of the 10,000 mean percentages from each iteration of bootstrapping across the various contents. Different types of content exhibit notable distinctions in their percentage of trigger warnings. Among all types of content, movies have the lowest mean percentage (21.10%) of trigger warnings among categories. This can be compared to TV series (36.72%), and video games (32.32%). One explanation is that movies typically have a shorter runtime than TV series and video games. This finding also hints that greater attention should be given to assigning trigger warnings in TV series and video games.

For all types of content after bootstrapping, the top two frequent trigger warnings are `Violence and Sex` (34.71%) and `Fear and Aversions` (33.65%). Comparing among content, TV series show higher Mental and Physical Health content than the others. For video games, the highest mean percentage of Fear and Aversions is 40.55%.

Co-occurrence Among Trigger Warnings To better understand warning differences across content types, we next investigate the co-occurrence of trigger warnings for each content. We again perform bootstrapping with 10,000 iterations. At each iteration, we take a random sample of the data for each content type. For the given sample, we create a 5x5 matrix (5 being the number of warning categories). Thus, each cell in this matrix will represent the number of items that have two warning categories assigned to it that correspond to the column and row of the cell. We then use this matrix calculate the phi-correlation (ϕ) (Matthews 1975) We compare category pairs individually and perform 10,000 bootstrapping iterations for each comparison. After each iteration, we calculate the phi-correlations. Subsequently, we compute the mean value of all the phi-correlation results across 10,000 iterations. We repeat this process for all the content types. The bootstrapping helps decrease the effect of the outliers.

Figure 3 presents the phi-correlation among trigger warning categories. We interpret the phi-correlation (ϕ) as follows: none or negligible (0-0.19), weak positive relationship (0.20-0.29), moderate positive relationship (0.30-0.39), strong positive relationship (0.40-0.69), and very strong positive relationship (higher than 0.7). Different types of content exhibit varying correlation patterns. The diversity of co-occurrence in trigger warnings highlights the unique content dynamics across various content types.

For movies, the Diversity and Inclusion category does not correlate with the other categories. There is a weak positive relationship between Fear and Aversions and Social Concerns. A moderate positive relation exists among Social Concerns, Mental and Physical Health, and Violence and Sex. For TV series, Diversity and Inclusion has stronger relationships with other categories than in movies, especially Mental and Physical Health. One possible explanation is that TV series often contain more offensive joke language compared to movies. This indicates the need for a customized trigger warning strategy for different content types. Interestingly, these correlations are extremely high in video games. However, Fear and Aversions have no correlation with other categories, for video games. Fear and Aversions include several factors, such as animals, ghosts, or even razors. They are relevant to the audience's personal experience (Blackwell et al. 2019). We posit that these correlations may be useful for helping automate the identification of trigger warnings.

Users' Vote and Comment Activities

Next, we examine users' voting and commenting patterns. Disagreement between users' voting and comment patterns may reveal differing opinions on which trigger warnings are valid.

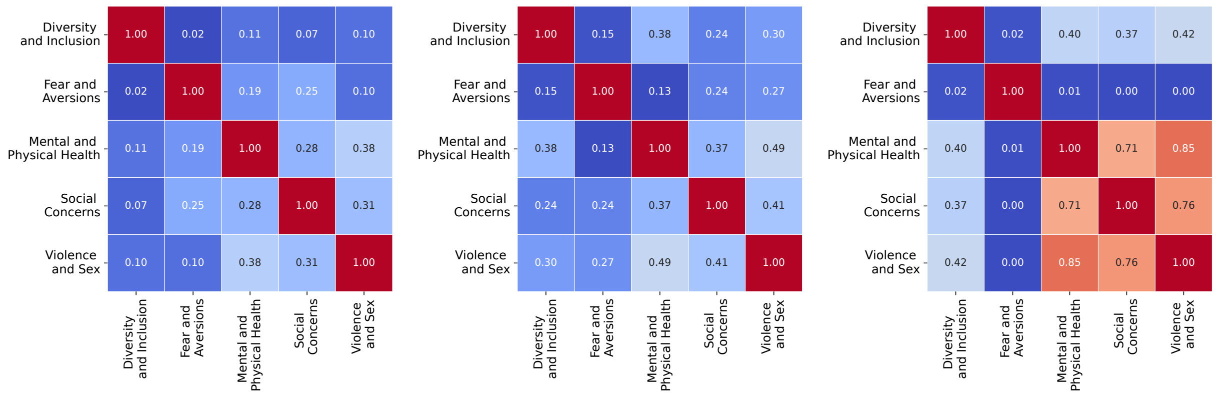
Voting Patterns Since trigger warnings on DTDD are based on user votes, we begin by analyzing the patterns of confirmed and rejected trigger warnings. Confirmed votes for an item indicate a greater number of "yes" votes than the "no" votes for a given trigger warning. Conversely, rejected trigger warnings refer to those that have a greater number of "no" votes than the "yes" votes.

To evaluate the diversity of the votes, we use Shannon Entropy (Shannon 1948). Given that there are two possible vote options (yes and no), the entropy is constrained within the range of 0 to 1. An entropy value approaching 1 signifies a higher level of controversy associated with the trigger warning, whereas an entropy value near 0 indicates a lower degree of controversy.

Figure 4 shows the distribution of the entropy for different content items for (a) confirmed and (b) rejected trigger warnings, separately. We observe that the entropy is lower for the rejected trigger warnings compared to confirmed trigger warnings. We conjecture that this difference is attributed to the varying sensitivities on specific trigger warnings. More specifically, users tend to foster consensus when rejecting trigger warnings but generate varied opinions when confirming trigger warnings. Overall,Diversity and Inclusion is the most controversial category across all the content, with the most controversial trigger warning is "Is there ableist language or behavior?"

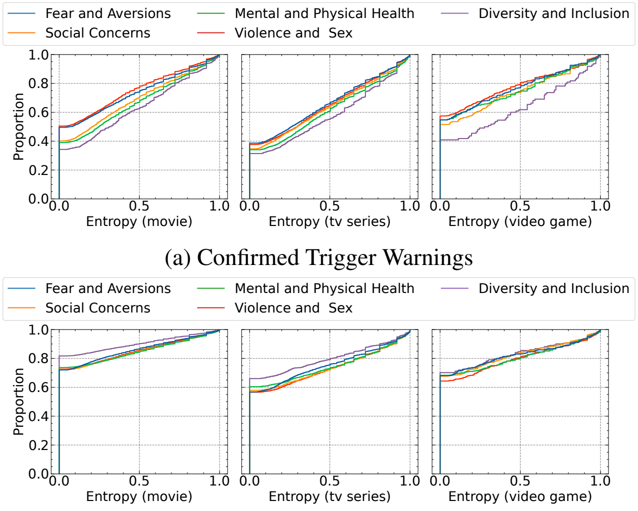
Votes *vs.* Comments Next, we take a closer look at the ratio of user comments and user vote activity. We count each trigger warning's comment rate by the ratio of the number of comments vs. the number of votes. For each trigger warning question, we suppose that a higher comment frequency indicates a greater level of concern. To obtain an overall estimate, reducing the bias from small samples, we conduct 10,000 bootstrapping iterations again (Mooney, Duval, and Duvall 1993). We separately process interactions of confirmed and rejected trigger warnings and calculate the mean comment ratio for each interaction. Figure 5 shows the distribution of the mean comment rates from 10,000 bootstrappings of different categories. Across all content, Diversity and Inclusion exhibits the highest comment rate. To explain this phenomenon, we manually check the comments' content. Typically, these comments articulate the associated plots of the corresponding trigger warning, alongside any justification for the chosen vote. This observation suggests that, in the majority of categories, a higher percentage of users demonstrate higher confidence in the rejected trigger warnings than accepting them. Conversely, within the category of Fear and Aversions, a higher percentage of users exhibit assurance in recognizing and affirming the relevant trigger warnings.

Our analysis reveals that TV series and video games possess a greater number of potential trigger warnings compared to movies. While movies and TV series include more Violence and Sex trigger warnings, video games include Fear and Aversions. Moreover, trigger warnings related to Diversity and Inclusion demonstrate the highest disagreement (among votes) and comments ratio, suggesting a wide range of sensitivities regard-



(a) Movie (b) TV Series (c) Video Game

Figure 3: Phi-correlation of 5 trigger warning categories in different content.



(a) Confirmed Trigger Warnings

(b) Rejected Trigger Warnings

Figure 4: The distributions of the entropy in different content. (a) for confirmed trigger warnings and (b) for rejected trigger warnings.

ing this topic.

Exploring Rating Systems Among Countries

We next inspect the individual rating systems of each country. Here, we seek to explore how each country assigns (different) ratings to each media item (**RQ2**).

Aligning Rating System

To compare the different rating systems across the countries, we first align the rating scores across the content types and countries. For example, the MPA system is used for movies, whereas the PEGI rating system is used for video games. These different rating systems make it difficult to compare different types of content directly.

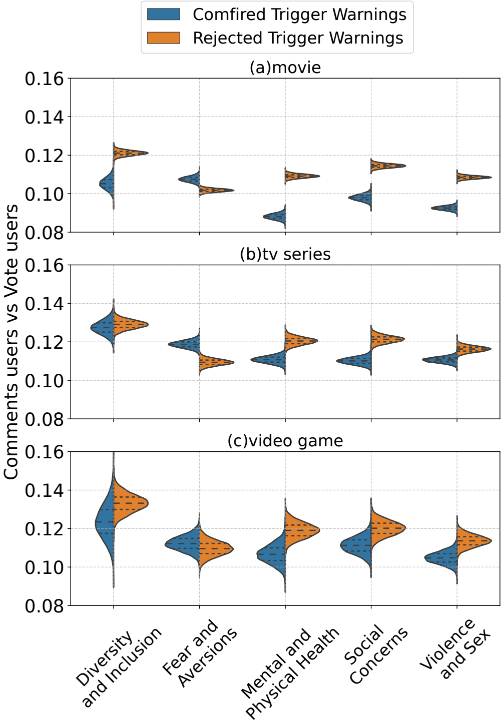


Figure 5: The distribution of comments rates within the categories in different types of content.

To address this, we note that most rating systems rely on age ranges. Thus, we use age as the common ground to align differences in rating systems. Moreover, age can be used as a standard comparison scale to see the difference in country ratings. We compare each country's ratings to the USA as a common baseline to simplify later comparisons. Figure 6 shows the labels across various countries' rating systems, mapped to the age range.

We identify 3 ratings, which all country's ratings are then mapped to: *G*, *PG*, and *R*. *G* indicates that the content is suit-

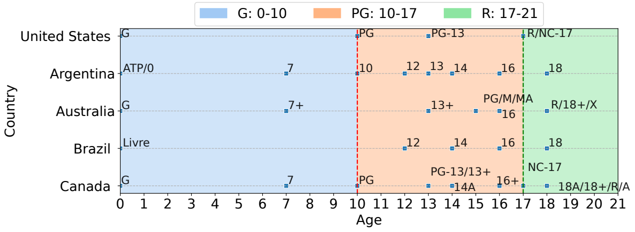


Figure 6: Manual alignment of the age-based ratings among four other countries with the U.S.

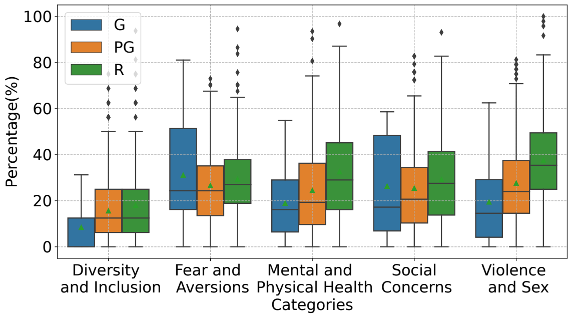


Figure 7: The percentage distribution of trigger warnings for individual items across various categories, segmented by different ratings (G, PG, R). The green triangles represent the mean average values.

Categories	KW H(2)	Mean diff (post-hoc)	p-value (post-hoc)
Diversity and Inclusion	16.90(***)	G<PG G<R PG<R	ns * ns
Fear and Aversions	3.40(ns)	G>PG G>R PG<R	- - -
Mental and Physical Health	32.53(***)	G<PG G<R PG<R	ns ** ***
Social Concerns	4.73(ns)	G<PG G<R PG<R	- - -
Violence and Sex	62.57(***)	G<PG G<R PG<R	ns *** ***
All	67.54(***)	G<PG G<R PG<R	ns ** ***

Table 2: Pair-wise comparison of groups by the Kruskal Wallis test with Dunn's post-hoc test. There are four types: *** ($p < 0.001$), **($p < 0.01$), *($p < 0.05$) and ns ($p > 0.05$).

able for a general audience, *PG* signifies that parental guidance is advised, and *R* denotes that the content is restricted to adult audiences.

Correlation Between Warnings and Ratings

Recall, we group all the crowdsourcing trigger warnings into five categories. Because there is more than one content type, we use the term "item", and each item has an aligned rating from G, P, and R. We begin by analyzing ratings from the USA to assess their relationship with trigger warnings. Because higher ratings may suggest that the item is suitable for a more limited group, we hypothesize that items with higher ratings may have more trigger warnings. We, therefore, calculate the percentage of trigger warnings associated with each item within each category.

Figure 7 presents the percentage of media items that have each category of trigger warning. We divide films based on their rating (G, PG, R). Intuitively, items rated R exhibit a higher mean percentage of trigger warnings compared to those rated PG and G. We notice that G-rated items show the highest average in `Fear and Aversions`. Moreover, all three ratings have a relatively similar average in `Social Concerns`. To confirm the differences in the distribution between the three ratings, we perform a non-parameter test (the Kruskal-Wallis test) on the number of trigger warnings in different categories with different ratings. Table 2 shows the Kruskal-Wallis result followed by post-hoc Dunn's test. The results indicate that the categories `Violence and Sex` and `Mental and Physical Health` exhibit statistically significant differences between the G and R ratings, as well as between the PG and R ratings. There are also fewer statistically significant differences in `Diversity and Inclusion`. The reasoning is intuitive, however, we also find that the different ratings do not demonstrate statistically significant differences for `Fear and Aversions` and `Social Concerns` trigger warnings. This suggests that rating systems may not effectively convey information about potential trigger warnings related to `Fear and Aversions` and `Social Concerns`.

Rating Agreement Among Countries

Next, we explore how the ratings assigned in each region differ (for the same media item). This helps us understand the cultural and social sensitivity among countries.

Rating Agreement Among Countries We notice that different countries sometimes give different content ratings for the same items. To evaluate the difference, we take the USA as a baseline and compare the 18 other countries that have at least 100 rated items in common with the USA. The number of common items shared by each country with the USA is presented in Appendix Table 6.

We first compare the agreement between the USA and the other countries using Cohen's Kappa score. Interestingly, the average Cohen's Kappa among 18 countries is 0.119, indicating only a slight agreement. There are six countries with a Cohen's Kappa less than 0. Only Spain and Canada have a moderate agreement (ranging from 0.4 to 0.6). This reflects the diversity in ratings assigned by each country.

To further understand the disagreement among countries, we calculate the risk ratio (RR) metric. For each country c and for each rating r, we calculate the following:

$$RR(r,c) = \frac{\frac{N(r,c)}{M}}{\frac{N(r,USA)}{M}} \tag{1}$$

Countries (Weighted F1-score)	Brazil (0.88)	Japan (0.80)	Mexico (0.70)	Singapore (0.71)	South Korea (0.71)	United Kindom (0.80)
1	Profanity 🔵 🟠 🟢	Sex&Nudity 🔵 🟠 🟢	Profanity 🔵 🟠 🟢	Sex&Nudity 🔵 🟠 🟢	Sex&Nudity 🔵 🟠 🟢	Sex&Nudity 🔵 🟠 🟢
2	Alcohol, Drug&Smoking 🔵 🟠 🟢	Violence&Gore 🔵 🟠 🟢	Fighting&Instense 🔵 🟠 🟢	Fighting&Instense 🔵 🟠 🟢	Profanity 🔵 🟠 🟢	Profanity 🔵 🟠 🟢
3	Fighting&Instense 🔵 🟠	Fighting&Instense 🔵 🟠 🟢	Sex&Nudity 🔵 🟠 🟢	Alcohol, Drug&Smoking 🔵 🟠 🟢	Alcohol, Drug&Smoking 🔵 🟠 🟢	Alcohol, Drug&Smoking 🔵 🟠 🟢
4	Violence&Gore 🔵 🟠	Alcohol, Drug&Smoking 🔵 🟠 🟢	Violence&Gore 🔵 🟠 🟢	Violence&Gore 🔵 🟠 🟢	Fighting&Instense 🔵 🟠 🟢	Violence&Gore 🔵 🟠 🟢
5	Sex&Nudity 🔵 🟠 🟢	Profanity 🔵 🟠 🟢	Alcohol, Drug&Smoking	Profanity 🔵 🟠 🟢	Violence&Gore 🔵 🟠 🟢	Fighting&Instense 🔵 🟠 🟢

Table 3: The SHAP results for the countries with a weighted F1 score higher than 0.7. The first column is the ranking of feature importance. 🔵 represents the feature influence on the G class. Similarly, 🟠 stands for PG and 🟢 stands for R.

where N (.) is the number of content items for rating r and country c. M is the total number of common items across the country c and the USA. The result from RR is a positive value. For the same item, a $RR < 1$ indicates that a country has fewer items of the given rating compared to the United States. A $RR = 1$ indicates that a country has the same number of items for the given rating as the United States. A $RR > 1$ indicates that a country has more items of the given rating compared to the United States. The full risk ratio results are given in Appendix Figure 9. Overall, the mean risk ratio for R rated media items among all the countries is 0.15. However, the mean risk ratio for PG is 2.64 and G is 5.98. This means the other countries give a lower rating than the USA, which aligns with the risk of unsuitable content being exposed to younger audiences. To gain deeper insights, we examine the genre tags associated with the items. 87.41% of items in the *Crime* genre get lower ratings in other countries than the USA. 86.37% of items in the the *Adult* genre receive consistent ratings. One exception is Sweden, which prefers to give lower ratings to adult items compared to the USA.

Impact of Trigger Warnings on Ratings Per-Country The above analysis reveals discrepancies among the ratings assigned by different countries. We next explore the specific trigger warnings that contribute to these discrepancies.

To investigate how the ratings assigned to each item differ on a per-country basis, we train separate decision trees to try and predict the ratings for each country, based on the trigger warnings listed on the IMDb parents guide. If the model performs well, it will confirm a strong link between specific trigger warnings and the rating assigned by the country. As an input to the model, we use the IMDb parents guide for each of the five categories of trigger warnings mentioned in Dataset Section ((*i*) Sex&Nudity, (*ii*) Violence&Gore, (*iii*) Profanity, Alcohol, (*iv*) Drugs&Smoking, and (*v*) Frightening&Intense Scenes). This produces a 5-item vector, with each item corresponding to one of the five specified trigger warnings. Categorical data is numerically encoded as follows: 1 for None, 2 for Mild, 3 for Moderate, and 4 for Severe, with a value of 0 representing the absence of data in the category. The model is then trained to use these features to predict the content rating for the media term (G, PG, or R). Note, we train a separate model for predicting each country's assigned rating. We employ a grid search to optimize the hyper-parameters of each model. 6 out of 18 countries get a weighted F1 score higher than 0.7. The hyperparameters of the models are in Appendix Table 7.

To measure which trigger warnings influence the ratings, we utilize SHAP (SHapley Additive exPlanations) (Lundberg 2017) to interpret feature influence. SHAP assigns each feature a Shapley value, which represents its average marginal contribution to the prediction across all possible combinations of features.

Table 3 presents the rank of features and whether it influences a specific class based on the Shapley value. The first column denotes the rank order of feature importance based on their SHAP value. The table highlights the diverse impact of trigger warnings across different countries. Unsurprisingly, `Sex &Nudity` is the most important feature in predicting the rating, often resulting in R. It ranks as the second most important feature on an average of six countries. However, the lowest ranked is `Violence &Gore` with an average rank of 3.8. The diverse rank of different trigger warnings among countries offers insights into the interpretation of trigger warnings across cultures. For example, based on ratings, audiences are more likely to be shielded from sex-related content, while being more exposed to unexpected violent content. These analyses could enhance the effectiveness of practical trigger warnings in different countries. For instance, trigger warnings related to alcohol, drugs, and smoking may be particularly relevant for users in Mexico, as this factor is not considered so important in the official ratings.

Automating Trigger Warnings Detection

As a crowdsourcing platform, DTDD offers detailed trigger warning. Specifically for each item, there are 197 potential trigger warnings (see the full list in Appendix), which are absent from traditional platforms like IMDb. These more fine-grained trigger warnings assist audiences in making more informed choices, and avoid potentially sensitive content. However, replicating this crowdsourcing approach for all items on IMDb would be both labor-intensive and time-consuming. To address this challenge (**RQ3**), we develop detectors to identify trigger warnings in items on IMDb.

Trigger warnings	Num of samples	
	Positive	Negative
Religion is discussed (*Diversity and Inclusion*)	217	214
Someone speaks hate speech (*Diversity and Inclusion*)	299	312
There's audio gore (*Fear and Aversions*)	263	337
An animal is sad (*Fear and Aversions*)	307	190
Someone uses drugs (*Mental and Physical Health*)	313	299
There are jump scares (*Mental and Physical Health*)	290	322
Someone is kidnapped (*Social Concerns*)	273	331
There's incarceration (*Social Concerns*)	240	194
There's body horror (*Violence and Sex*)	287	276
There's abusive parents (*Violence and Sex*)	248	285

Table 4: The selected 10 trigger warnings, and the positive/negative sample sizes for each.

Experimental Design

Our objective is to develop a tool that can automatically identify trigger warnings. We focus on the 646 items that appear on both DTDD and IMDb. We utilize data from DTDD as the ground truth for each item. For each trigger warning listed on DTDD, an item is marked as 1 (positive) for the corresponding trigger warning if it receives more "Yes" votes than "No" votes. Conversely, if an item receives more "No" votes than "Yes" votes, it is marked as 0 (negative) for that trigger warning. Based on these votes, we establish the ground truth labels of all 646 items for each of the 197 trigger warnings.

Considering the cost of running the experiments, we focus on 10 trigger warnings. Specifically, we select 2 trigger warnings with the most balanced distribution of positive and negative samples from each of the 5 categories outlined in Dataset Section. Table 4 displays the ten selected trigger warnings, along with the quantities of positive and negative samples associated with each trigger. The first column lists the trigger warnings and their respective categories. The subsequent two columns indicate the number of positive and negative samples.

Based on the ground truth labels mentioned above, we then train detectors (one per trigger warning) to predict whether an item should be tagged with the given trigger warning (1) or not (0). As input features, we utilize the item's text-based description and storyline on IMDb.

Trigger Warning Detector Design

In designing our trigger warning detector tool, we employ two types of models: large language models (LLMs) and traditional machine learning models. We evaluate these models in the context of the classification task, comparing their performance to identify the most effective model for trigger warning detection.

LLMs. We select GPT-4o as the representative of LLMs. We utilize the official API provided by OpenAI and design a tailored prompt specifically for detecting trigger warnings. An example of our prompt, used to identify a particular trigger warning, is shown below. The bold "Is religion discussed" serves as the specific trigger warning in this example.

*"role": "user", "content": "'Prompt': You're an expert in item and content moderation. Your task is identify **Is religion discussed** in this item based on its description and storyline. 'Labels': ['Yes', 'No'], 'item': item description + item story line."*

Machine Learning Models. We experiment with 7 machine learning (ML) algorithms: Random Forest Classifier (RF), Logistic Regression (LR), Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Gradient Boosting Classifier (GB), Decision Tree Classifier (DT) and Gaussian Naive Bayes (GNB).

For the seven traditional machine learning models, we employ vector embeddings of the IMDb item descriptions and storylines as features. We utilize *all-MiniLM-L6-v2*, a widely recognized model from the SentenceTransformer series (Reimers and Gurevych 2019), to generate vector representations of these texts. The SentenceTransformer series comprises pre-trained models that excel at transforming sentences into dense vector representations, effectively capturing the semantic meaning of the text. The *all-MiniLM-L6-v2* model is extensively used in studies for generating vector embeddings of text sentences (He et al. 2024, 2023).

We then train binary classifiers for predicting if a media item should contain a particular trigger warning. Note, we train a separate classifier for each trigger warning. For each trigger warning, the item samples are divided into a training set and a test set (80:20 ratio). During the training phase, we employ 5-fold cross-validation and use grid search to optimize the parameters of each classifier. The specific parameters for each classifier are detailed below:

- **Random Forest**: 'n_estimators': [5, 50, 100, 250]; 'max_depth': [2, 4, 8, 16, 32, None].
- **Logistic Regression**: 'penalty': ['l1','l2']; 'C': [0.001, 0.01, 0.1, 1, 10, 100, 1000].
- **Support Vector Machine**: 'C': [0.001, 0.01, 0.1, 1, 10, 100, 1000]; 'kernel': ['rbf', 'linear', 'sigmoid'].
- **K-Nearest Neighbors**: 'n_neighbors': [1, 3, 5, 7, 9]; 'algorithm': ['auto', 'ball_tree', 'kd_tree', 'brute'].
- **Gradient Boosting**: 'n_estimators': [100, 300, 500, 800]; 'learning_rate': [0.01, 0.1, 1, 10].
- **Decision Tree**: 'max_depth': [2, 4, 8, 16, 32, None].
- **Gaussian Naive Bayes**: 'var_smoothing': [1e-8, 1e-9, 1e-10].

Evaluation

Table 5 presents the performance metrics for all models tasked with identifying the selected 10 trigger warnings. We achieve F1 scores exceeding 70% across all these trigger warnings. The highest F1 score is 81.13%. This is attained for the trigger warning "Is there body horror" in the `Diversity and Inclusion` category, while a score of 76.73% is recorded for the trigger "Are there jump scares" in the `Mental and Physical Health` category.

GPT-4o demonstrates surprisingly poor performance in detecting all trigger warnings, as evidenced by the last column of Table 5. Examination of GPT-4o's confusion matrix reveals a high incidence of false negatives, with most items being incorrectly classified as negative (indicating the absence of the corresponding trigger warning). This trend is

Trigger warnings	Performance of models (measured by F1-score (%))							
	RF	LR	SVM	KNN	GB	DT	GNB	GPT-4o
Religion is discussed	**70.71**	67.99	69.51	69.51	65.83	62.19	70.45	31.28
Someone speaks hate speech	70.17	**71.90**	70.78	60.52	65.77	65.47	70.18	16.56
There's audio gore	**74.11**	72.46	73.28	67.11	63.81	63.35	70.48	20.74
An animal is sad	63.21	68.26	**73.01**	67.16	56.65	59.03	71.59	13.16
Someone uses drugs	**71.23**	68.69	69.57	64.34	57.34	61.74	68.64	35.71
There are jump scares	**76.73**	74.32	74.81	68.07	68.89	63.11	75.25	48.33
Someone is kidnapped	**72.68**	65.71	71.27	65.77	59.14	60.18	67.40	21.05
There's incarceration	69.63	71.44	70.29	71.61	**71.75**	65.83	67.04	17.04
There's body horror	**81.13**	71.44	71.61	68.73	74.14	65.02	73.55	53.30
There's abusive parents	**73.68**	69.80	69.29	65.93	63.35	69.26	68.52	21.18

Table 5: Classification results of models for all trigger warnings. For convenience, we embolden the best result for every trigger warning.

consistent across almost all trigger warnings evaluated by GPT-4o. In contrast, the classification outcomes from the seven other machine learning models do not exhibit such a skew, maintaining a balanced distribution of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN).

To explore this, we manually review the text output from ChatGPT. To facilitate this, our prompts asks ChatGPT to provide an explanation of all labels that it predicts (see Apendix). Two independent authors then manually review these explanations on all ten trigger warnings' output results. We find that ChatGPT often has a different explanation of trigger warnings compared to crowdsourced evaluations. For example, in the best-performing trigger warning, "Is there body horror", ChatGPT's explanation of its prediction highlights that body horror refers to unnatural or fantastical transformations of the body. As a result, it states that killing sprees or physical harm are not considered forms of body horror. In contrast, such elements *are* considered body horror in the crowdsourced DDTD explanations, resulting in the discrepancy. Machine learning models can leverage these patterns during the training process (which a zero-shot GPT approach cannot) to enhance performance in trigger warning detection. Although the limited number of samples pose a challenge to further improving the models' performance, our detector achieves an F1 score of up to 81.13%, demonstrating its potential for effective trigger warning detection. This labor-saving automatic annotation tool could enable media websites, such as IMDb, to automatically apply more fine-grained trigger warnings to their content. Additionally, it could facilitate the development of various features designed to enhance user experience, including advanced search systems.

Discussion and Conclusion

We have examined trigger warnings across different media types, and analyzed their relation to the rating systems of various countries. Our work highlights the need for a more comprehensive approach to evaluating and flagging trigger warnings in media.

Based on our analysis, we have also developed detectors to identify trigger warnings automatically using free-text from IMDb. We achieve F1 scores exceeding 0.7 for all of 10 selected trigger warnings. The highest F1 score of 81.13% underscores the potential of our detectors in automating the detection of trigger warnings. This could be applied to platforms like IMDb or integrated into streaming platforms like Netflix to automatically assign context-aware trigger warnings. Our results highlight that such integration should be tailored to each country, and offer users the option to choose which trigger warnings they prefer. Considering the poorer performance of LLMs, to enhance real-world applicability, further fine-tuning could involve combining user survey results in the future.

Our study acknowledges key limitations that may affect our findings. We relied solely on IMDb and DTDD data, which, while a rich source for mainstream film and television data, lacks diversity of media types and user-generated content found on platforms like TikTok and YouTube. Arguably, this restricts our ability to capture the variety of media utilizing trigger warnings. Additionally, our focus on English-language content limits the applicability of our findings to non-English contexts. In the future, we plan to incorporate data from additional platforms and languages to provide a broader perspective. Despite these constraints, our study is one of the first to quantitatively analyze trigger warnings across three media types, laying the groundwork for future research to expand on these findings with a more global and diverse dataset.

We have incorporated data from the user-generated platform DDTD, which offers a unique perspective on trigger warnings across various media. However, relying on user-generated warnings potentially introduces sampling biases that must be considered. The users contributing to DTDD are a small self-selecting subset of the population. Further, users on such platforms may focus on popular films or games, potentially skewing the representation of trigger warnings toward more mainstream content. This bias could therefore limit the diversity of media types and the breadth of trigger warning usage captured in our study. To mitigate risks, we use bootstrapping to resample with replacement. However, exploring and addressing this bias constitutes a key line of future work. We also note that the absence of users' demographic information limits our ability to analyze how demographic factors influence users' trigger warning votes. We conjecture that demographics may have a major impact on the specific trigger warnings flagged, and would therefore be interesting to investigate.

Acknowledgements

This work was supported in part by the Guangzhou Science and Technology Bureau (2024A03J0684), Guangdong provincial project 2023QN10X048, the Guangzhou Municipal Key Laboratory on Future Networked Systems (2024A03J0623), the Guangdong Provincial Key Lab of Integrated Communication, Sensing and Computation for Ubiquitous Internet of Things (No.2023B1212010007), the Guangzhou Municipal Science and Technology Project (2023A03J0011), Guangdong provincial project (2023ZT10X009), and the 111 Center (No. D25008).

References

Barranco, R. E.; Rader, N. E.; and Smith, A. 2017. Violence at the box office: Considering ratings, ticket sales, and content of movies. *Communication Research*, 44(1): 77–95.

Bellet, B. W.; Jones, P. J.; and McNally, R. J. 2018. Trigger warning: Empirical evidence ahead. *Journal of Behavior Therapy and Experimental Psychiatry*, 61: 134–141.

Blackwell, L.; Ellison, N.; Elliott-Deflo, N.; and Schwartz, R. 2019. Harassment in social virtual reality: Challenges for platform governance. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW): 1–25.

Boysen, G. A.; Isaacs, R. A.; Tretter, L.; and Markowski, S. 2021. Trigger warning efficacy: The impact of warnings on affect, attitudes, and learning. *Scholarship of Teaching and Learning in Psychology*, 7(1): 39.

Bridgland, V. M.; Bellet, B. W.; and Takarangi, M. K. 2023. Curiosity disturbed the cat: Instagram's sensitive-content screens do not deter vulnerable users from viewing distressing content. *Clinical psychological science*, 11(2): 290–307.

Bryce, I.; Horwood, N.; Cantrell, K.; and Gildersleeve, J. 2023. Pulling the trigger: a systematic literature review of trigger warnings as a strategy for reducing traumatization in higher education. *Trauma, Violence, & Abuse*, 24(4): 2882–2894.

Charles, A.; Hare-Duke, L.; Nudds, H.; Franklin, D.; Llewellyn-Beardsley, J.; Rennick-Egglestone, S.; Gust, O.; Ng, F.; Evans, E.; Knox, E.; et al. 2022. Typology of content warnings and trigger warnings: Systematic review. *PloS one*, 17(5): e0266722.

Dillahunt, T. R.; Erete, S.; Galusca, R.; Israni, A.; Nacu, D.; and Sengers, P. 2017. Reflections on design methods for underserved communities. In *Companion of the 2017 ACM conference on computer supported cooperative work and social computing*, 409–413.

He, J.; Zia, H. B.; Castro, I.; Raman, A.; Sastry, N.; and Tyson, G. 2023. Flocking to mastodon: Tracking the great twitter migration. In *Proceedings of the 2023 ACM on IMC*, 111–123.

He, Y.; Gu, Y.; Shekhar, R.; Castro, I.; and Tyson, G. 2024. Making the Pick: Understanding Professional Editor Comment Curation in Online News. In *Proceedings of ICWSM*, volume 18, 557–568.

Hirsch, T. 2020. Practicing without a license: Design research as psychotherapy. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–11.

Horne, B. D. 2024. Does the Source of a Warning Matter? Examining the Effectiveness of Veracity Warning Labels Across Warners. *arXiv preprint arXiv:2407.21592*.

Khan, S. U.; Haq, I. U.; Rho, S.; Baik, S. W.; and Lee, M. Y. 2019. Cover the violence: A novel Deep-Learning-Based approach towards violence-detection in movies. *Applied Sciences*, 9(22): 4963.

Knox, E. J. 2017. *Trigger warnings: History, theory, context*. Rowman & Littlefield.

Ling, C.; Gummadi, K. P.; and Zannettou, S. 2023. " Learn the Facts About COVID-19": Analyzing the Use of Warning Labels on TikTok Videos. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 17, 554–565.

Lockhart, E. A. 2016. Why trigger warnings are beneficial, perhaps even necessary. *First Amendment Studies*, 50(2): 59–69.

Lukin, S.; and Walker, M. 2017. Really? well. apparently bootstrapping improves the performance of sarcasm and nastiness classifiers for online dialogue. *arXiv preprint arXiv:1708.08572*.

Lundberg, S. 2017. A unified approach to interpreting model predictions. *arXiv preprint arXiv:1705.07874*.

Matthews, B. W. 1975. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta (BBA)-Protein Structure*, 405(2): 442–451.

Mooney, C. Z.; Duval, R. D.; and Duvall, R. 1993. *Bootstrapping: A nonparametric approach to statistical inference*. 95. sage.

Ramos, J.; et al. 2003. Using tf-idf to determine word relevance in document queries. In *Proceedings of the first instructional conference on machine learning*, volume 242, 29–48. Citeseer.

Reimers, N.; and Gurevych, I. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Scott, C. F.; Marcu, G.; Anderson, R. E.; Newman, M. W.; and Schoenebeck, S. 2023. Trauma-informed social media: Towards solutions for reducing and healing online harm. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–20.

Shannon, C. E. 1948. A mathematical theory of communication. *The Bell system technical journal*, 27(3): 379–423.

Stratta, M.; Park, J.; and deNicola, C. 2020. Automated Content Warnings for Sensitive Posts. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing*

Systems, CHI EA '20, 1–8. New York, NY, USA: Association for Computing Machinery. ISBN 9781450368193.

Thomson, K. 2021. An analysis of LGBTQ+ representation in television and film. *Bridges: An Undergraduate Journal of Contemporary Connections*, 5(1): 7.

Wiegmann, M.; Rakete, J.; Wolska, M.; Stein, B.; and Potthast, M. 2024. If there's a Trigger Warning, then where's the Trigger? Investigating Trigger Warnings at the Passage Level. *arXiv preprint arXiv:2404.09615*.

Wiegmann, M.; Wolska, M.; Schröder, C.; Borchardt, O.; Stein, B.; and Potthast, M. 2023. Trigger warning assignment as a multi-label document classification problem. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 12113–12134.

Wolska, M.; Schröder, C.; Borchardt, O.; Stein, B.; and Potthast, M. 2022. Trigger warnings: Bootstrapping a violence detector for fanfiction. *arXiv preprint arXiv:2209.04409*.

Zhang, Y.; Shafaei, M.; Gonzalez, F.; and Solorio, T. 2021. From none to severe: Predicting severity in movie scripts. *arXiv preprint arXiv:2109.09276*.

Paper Checklist

1. For most authors...
   (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes, see Section - Ethics and Broader Impacts.
   (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes, see Abstract
   (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes
   (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? Yes
   (e) Did you describe the limitations of your work? Yes
   (f) Did you discuss any potential negative societal impacts of your work? Yes, see Section - Ethics and Broader Impacts.
   (g) Did you discuss any potential misuse of your work? Yes, see Section - Ethics and Broader Impacts.
   (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? Yes
   (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes
2. Additionally, if your study involves hypotheses testing...
   (a) Did you clearly state the assumptions underlying all theoretical results? Yes
   (b) Have you provided justifications for all theoretical results? Yes
   (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? Yes
   (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? Yes
   (e) Did you address potential biases or limitations in your theoretical framework? Yes
   (f) Have you related your theoretical results to the existing literature in social science? Yes
   (g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? Yes
3. Additionally, if you are including theoretical proofs...
   (a) Did you state the full set of assumptions of all theoretical results? Yes
   (b) Did you include complete proofs of all theoretical results? Yes
4. Additionally, if you ran machine learning experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? No, we cannot release the Does the Dog Die and IMDb datasets considering the copyright. That said, it is possible for other researchers to collect an equivalent dataset, and we will make our code available to facilitate this.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? Yes
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? No, because random seed is fixed in our experiments.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? No, because the resources required for our experiments are only CPUs on our own devices, not huge.
   (e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? Yes
   (f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? NA
5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity**...
   (a) If your work uses existing assets, did you cite the creators? NA
   (b) Did you mention the license of the assets? NA
   (c) Did you include any new assets in the supplemental material or as a URL? NA
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? NA
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? NA

(f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR? NA

(g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset? NA

6. Additionally, if you used crowdsourcing or conducted research with human subjects, **without compromising anonymity**...

(a) Did you include the full text of instructions given to participants and screenshots? NA

(b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? NA

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? NA

(d) Did you discuss how data is stored, shared, and deidentified? NA

Ethics and Broader Impacts

We rely on public data, i.e., Does the Dog Die users have shared it knowingly that it is public on the Internet. We obtain our dataset using the official AIP for every user of Does the Dog Die. We do not interact with or solicit any data from any human subject, directly. That said, it is possible for other researchers to collect an equivalent dataset using the same methodology. To assist in this, we would be happy to offer code and support upon request.

We also note there are important positive outcomes of our work. We build a tool to help identify subtle trigger warnings for items on IMDb. This can assist audiences in making more informed choices and avoiding potentially sensitive content. Since the tooling built based on trigger warnings from DTDD, we do not expect that the approach will suppress voices of other audiences out of DTDD. That said, our models could certainly identify potential trigger warnings. Thus, we adopt a human-in-the-loop model, whereby we only shortlist potential trigger warnings — voices of other audiences and moderators must be considered to make ultimate decisions.

Appendix

All Trigger Warnings

The all 197 trigger warnings from Does the Dog Die as following:

- **Animal**
  - Does the dog die
  - Are animals abused
  - Does an animal die
  - Does a cat die
  - Does a pet die
  - Is there a dead animal
  - Is there dog fighting
  - Were animals harmed in the making
  - Does a horse die
  - Is an animal sad
  - Are rabbits harmed
  - Are there spiders
  - Are there bugs
  - Does a dragon die
  - Are there snakes
  - Are there sharks
  - Are there alligators/crocodiles?
- Abandonment
  - Is an animal abandoned
  - Is a child abandoned by a parent
  - Does someone leave without saying goodbye
- Abuse
  - Is a child abused
  - Is there domestic violence
  - Is a woman brutalized for spectacle
  - Are there abusive parents
  - Does an abused person forgive their abuser
  - Is someone gaslighted
  - Does the abused become the abuser
  - Is someone stalked
  - Is someone abused with a belt
- Addiction
  - Is there addiction
  - Does someone abuse alcohol
  - Does someone use drugs
- Assault
  - Is someone sexually assaulted
  - Is someone raped onscreen
  - Is there pedophilia
  - Is rape mentioned
  - Are there jokes about sexual assault on men
  - Is someone drugged
  - Is someone restrained
  - Is someone beaten up by a bully
  - Is someone held under water
  - Is someone's mouth covered
  - Does a woman get slapped
- Bodily Harm
  - Is someone tortured
  - Is there eye mutilation
  - Is there excessive gore
  - Is there genital trauma/mutilation
  - Does a head get squashed
  - Is there body horror
  - Is there finger/toe mutilation
  - Is there cannibalism
  - Is there a hanging

  - Is someone burned alive
  - Are any teeth damaged
  - Is there shaving/cutting
  - Does someone asphyxiate
  - Is someone crushed to death
  - Does someone struggle to breathe
  - Is there decapitation
  - Does someone break a bone
  - Are any hands damaged
  - Is there Achilles Tendon injury
  - Is someone buried alive
  - Is someone choked
  - Is there amputation
  - Is there throat mutilation
  - Does someone have a seizure
  - Does someone fall to their death
  - Is someone stabbed
  - Does someone become unconscious
  - Are there dislocations
  - Does someone fall down stairs
- Children
  - Is a minor sexualized
  - Does a kid die
  - Is an infant abducted
- Creepy Crawly
  - Are there bedbugs
- Death
  - Does a major character die
  - Does someone die
  - Does a non-human character die
  - Does someone sacrifice themselves
- Disability
  - Is the r-slur used
  - Is someone disabled played by able-bodied
- Drug/Alcohol
  - Does someone overdose
- Family
  - Is a child's toy destroyed
  - Does a family member die
  - Does someone cheat
  - Does a parent die
  - Is someone kidnapped
- Fear
  - Are there jumpscares
  - Is trypophobic content shown
  - Is someone possessed
  - Are there clowns
  - Are there ghosts
  - Are there razors
  - Are there mannequins
  - Is there a shower scene
  - Are there natural bodies of water
- Gross
  - Does someone vomit
  - Is there audio gore
  - Is someone eaten
  - Is there on-screen pooping
  - Does someone wet/soil themselves
  - Does someone spit
  - Is there farting
- Large-scale Violence
  - Are there 9/11 depictions
- Law Enforcement
  - Is there copaganda
  - Is there incarceration
- LGBTQ+
  - Is a trans person depicted predatorily
  - Are there transphobic slurs
  - Is there deadnaming or birthnaming
  - Is there bisexual cheating
  - Is an LGBT+ person outed
- Loss
  - Is a priceless artifact destroyed
- Medical
  - Are needles/syringes used
  - Is electro-therapy used
  - Is there a mental institution scene
  - Does someone have cancer
  - Is there a hospital scene
  - Is there menstruation
- Mental Health
  - Does someone die by suicide
  - Does someone attempt suicide
  - Does someone self harm
  - Does someone suffer from PTSD
  - Is there autism specific abuse
  - Is there misophonia
  - Does someone have an eating disorder
  - Are there anxiety attacks
  - Is autism misrepresented
  - Is a mentally ill person violent
  - Does someone say "I'll kill myself"
  - Is there body dysmorphia
  - Does someone have a mental illness
  - Is there a claustrophobic scene
  - Is there ABA therapy

  - Is there body dysphoria
  - Is reality unstable or unhinged
  - Is there D.I.D. misrepresentation
  - Does someone have a meltdown
  - Is there dissociation, depersonalization, or derealization
- Noxious
  - Is there shakey cam
  - Are there flashing lights or images
  - Are there sudden loud noises
  - Does a baby cry
  - Are there underwater scenes
  - Is there screaming
  - Is there obscene language/gestures
- Paranoia
  - Is someone watched without knowing
  - Is the fourth wall broken
- Pregnancy
  - Does someone miscarry
  - Is there childbirth
  - Is a baby stillborn
  - Does a pregnant person die
  - Are there abortions
  - Are there babies or unborn children
- Prejudice
  - Are there homophobic slurs
  - Does an LGBT person die
  - Is there ableist language or behavior
  - Are there fat jokes
  - Is someone misgendered
  - Are there ”Man in a dress” jokes
  - Is there hate speech
  - Are there n-words
  - Is there antisemitism
  - Is there aphobia
  - Does the black guy die first
  - Is a minority is misrepresented
- Race
  - Is there blackface
- Relationships
  - Is there a large age gap
- Religious
  - Are there demons or Hell
  - Is religion discussed
- Sex
  - Are there incestuous relationships
  - Is there bestiality
  - Is there sexual content
  - Is someone sexually objectified
  - Are there nude scenes
  - Is there BDSM
  - Does someone lose their virginity
- Sexism
  - Is a male character ridiculed for crying
- Sickness
  - Is there dementia/Alzheimer’s
  - Is someone terminally ill
  - Does someone have a chronic illness
  - Does someone have a stroke
- Social
  - Are there anti-abortion themes
  - Are there fat suits
  - Is existentialism debated
  - Is someone homeless
- Spoiler
  - Does it have a sad ending
  - Are there end credit scenes
  - Is Santa (et al) spoiled
- Vehicular
  - Is someone hit by a car
  - Does a car crash
  - Does a plane crash
  - Does a car honk or tires screech
- Violence
  - Is there blood/gore
  - Does someone drown
  - Is there gun violence
  - Is there a nuclear explosion

To have an overall understanding of the trigger warnings of each category, we present the distinctive words in each category. we perform lemmatization and remove the stopwords on all the trigger warnings. Figure 8 presents the top 20 words in the description of each category in lowercase generated by TF-IDF (Ramos et al. 2003).

Risk Ratio

To select more countries for comparison, we set the threshold as 100 and find 18 countries. Table 6 shows the details of the selected countries and the common item with the US. Figure 9 illustrates the risk ratio results, comparing the United States with other countries. The red line is equal to RR=1.

Hyperparameters of Models

To better understand cultural diversity across countries, we take SHAP to explain the machine learning models. We obtain the hyperparameters by using grid search and get 6 countries with a weighted F1 score greater than 0.7. Table 7 presents the specific hyperparameters for the decision tree classification models of these countries.

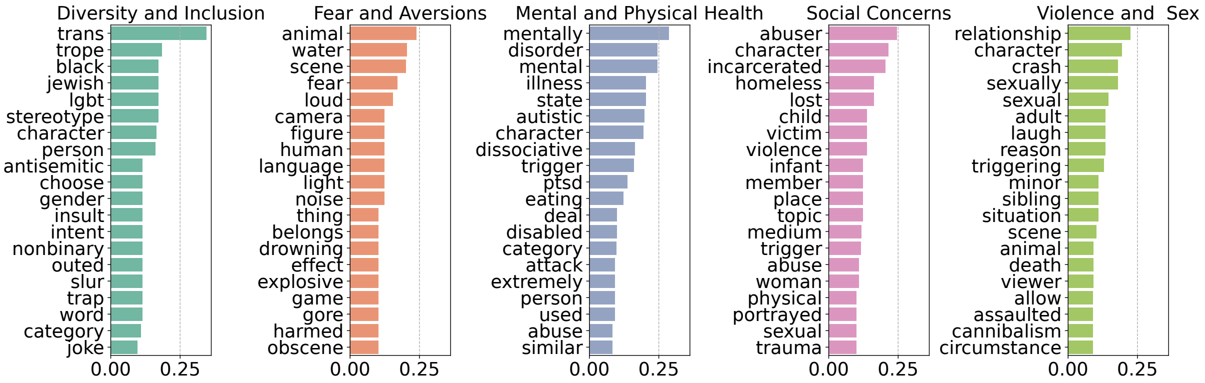


Figure 8: The Top-20 TF-IDF words from the description of each group. The x-axis shows the TF-IDF score.

Countries	No. of Items
Sweden	10181
United Kingdom	9385
Australia	8759
Netherlands	4599
Japan	2123
Brazil	1959
Argentina	1561
Germany	1474
Finland	903
Spain	711
France	662
Canada	659
Singapore	600
Norway	478
South Korea	299
Mexico	286
Greece	191
Philippines	121

Table 6: The number of common items with the US.

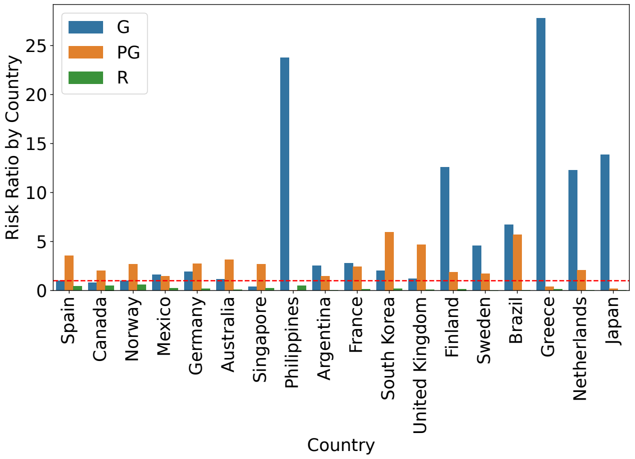


Figure 9: The risk ratio between the county and the United States. The red line is equal to RR=1.

Countries	Hyperparameters			
	criterion	max_depth	min_samples_leaf	min_samples_split
Brazil	entropy	None	2	30
Japan	entropy	None	1	2
Mexico	gini	None	1	2
Singapore	gini	None	1	2
South Korea	gini	None	1	2
United Kingdom	entropy	None	4	10

Table 7: The hyperparameter of the decision trees.