Trigger Warnings: A Computational Approach to Understanding User-Tagged Trigger Warnings

Sarthak Tyagi
Vellore Institute of Technology, Chennai, India
sarthak.tyagi2019@vitstudent.ac.in

Adwita Arora, Krish Chopra and **Manan Suri**
Netaji Subhas University of Technology, New Delhi, India
{adwita.ug20, krish.ug20, manan.suri.ug20}@nsut.ac.in

Abstract

Content and trigger warnings give information about the content of material prior to receiving it and are used by social media users to tag their content when discussing sensitive topics. Trigger warnings are known to yield benefits in terms of an increased individual agency to make an informed decision about engaging with content Charles et al. (2022). At the same time, some studies contest the benefits of trigger warnings suggesting that they can induce anxiety and reinforce the traumatic experience of specific identities Bridgland et al. (2019). Our study involves the analysis of the nature and implications of the usage of trigger warnings by social media users using empirical methods and machine learning. Further, we aim to study the community interactions associated with trigger warnings in online communities, precisely the diversity and content of responses and inter-user interactions. The domains of trigger warnings covered will include self-harm, drug abuse, suicide, and depression. The analysis of the above domains will assist in a better understanding of online behaviour associated with them and help in developing domain-specific datasets for further research.

1 Introduction

Trigger warnings are "a statement at the start of a piece of writing, video, etc. alerting the reader or viewer to the fact that it contains potentially distressing material–often used to introduce a description of such content" Bellet et al. (2018). They are used frequently by users online when discussing sensitive issues which might trigger a detrimental response in certain users. Trigger warnings are used in multiple contexts with respect to sensitive content; common examples include narrating one's own experience, talking about someone else's experience, or discussing content that might be sensitive Bridgland et al. (2019); Bellet et al. (2018). Some domains that are commonly associated with the use of trigger warnings include Self-harm, Violence, Drug Abuse, Suicide, and Depression. Trigger warnings thus can be used as a tool for others to self-moderate the content that they are engaging with on the internet with their level of comfort. However, recent empirical studies have shown that trigger warnings may also lead to the centering of traumatic experiences in communities, thus having the exact opposite effect Jones et al. (2020).

While in the wrong circumstances, anything can act as a trigger for a person under distress, some content has a universally accepted nature of being distressing or troubling for a large group of people[1] Charles et al. (2022); Ballestrini (2022). An overwhelming consensus on the classification and typology of these content warnings does not exist, as some sources like providing a more general description of content and trigger warnings (as an example considering violence as the trigger warning). In contrast, others look at the sub-categories as a more appropriate way of describing the nature of the content (animal cruelty, and sexual violence all describe a sub-category of violence, but give the reader a better idea of the nature the content represents).

Nevertheless, these sources agree upon the fact that none of their lists represents an exhaustive account of content that can distress a reader. Through our research, we want to utilize publicly available data from social media to perform an analysis on the use of trigger warnings, the nature of discourse associated with the selected domains, and community modelling of online communities. The analysis would involve the linguistic study of the expressions demonstrated by users. It would also include

[1] University of Michigan, An Introduction to Content Warnings and Trigger Warnings.https://sites.lsa.umich.edu/inclusive-teaching/an-introduction-to-content-warnings-and-trigger-warni

Proceedings of the Student Research Workshop associated with RANLP-2023,
pages 44–54, Varna, Bulgaria, Sep 4–6, 2023.
https://doi.org/10.26615/issn.2603-2821.2023_005

topic modelling and keyword analysis to study the nature of posts. A comparison between different domains would help us understand the differences between the communities involved in the respective domain. An additional area of analysis includes the response received to posts tagged with trigger warnings. In our study, we build up a novel dataset of trigger warning posts from various subreddits on Reddit, with our target trigger warnings being self-harm, suicidal ideation, depression, and drug abuse. We perform a multitude of analyses on this extracted data, which includes sentiment analysis of the gathered posts, an analysis of how the frequency of posts in these different subreddits have evolved, extraction of keyphrases from these texts, a look into how common topic models perform in this analysis and finally we tackle a classification problem by assigning classes to the dataset based on what type of post (post asking questions, a post asking advice, rant posts, etc.). We plan to release our dataset after paper acceptance.

2 Related Work

While we are not aware of prior work on computational trigger warning analysis using social media data or specifically warning assignments, We have divided our related work into sections highlighting the advantages of BERTopic over traditional topic modelling techniques, Relevant tasks to our objective, and other relevant papers.

2.1 Employing BERTopic for analysis

In Ogunleye et al. (2023), the authors provide an evaluation of how different topic models measure up to the task of topic extraction from a corpus of tweets about Nigerian Banks. It evaluates how traditional topic modelling techniques like Latent Dirichlet Allocation (LDA), Hierarchical Dirichlet Process (HDP), and Latent Semantic Indexing (LSI) perform in comparison to BERTopic Grootendorst (2022) utilizing a Kernel PCA for dimensionality reduction and K-means for clustering, achieving a maximum coherence score of 0.8463, well above the other techniques studied. LDA with a more well-processed corpus performs well but not better than BERTopic, and it suffers from the different subtleties present in tweets that use Pidgin English. This study highlights the incredible performance of BERTopic as a way to model the key topics present in a corpus in a completely unsupervised way. de Groot et al. (2022) analyzes how the BERTopic performs on the multi-domain short text and tests its generalizability in this context in comparison to LDA in terms of topic coherence and diversity. It uses open-text documents of university students from various domains such as computer science to law. The data is short in its length, with a median of 14 to 20 words in each. BERTopic outperforms LDA on both short-form and long-form documents, however, the coherence declines similar to the decline in coherence for LDA models when we evaluate performance on short documents using BERTopic with HDBSCAN. The performance of BERTopic using k-means clustering shows that the model is the least susceptible to short documents, due to its ability to generate more interpretable topics and fewer outliers than HDBSCAN.

2.2 Identifying Trigger Warning from fictional text corpus

Wolska et al. (2022) presents a very similar task of looking at trigger warning assignments in a corpus built from fanfiction works present on Archive of Our Own (AO3), focusing on trigger warnings related to violence. They provide a binary classification system for assigning trigger warnings and assess their effectiveness on a similar unlabelled dataset of fanfiction documents. They evaluate an SVM classifier and a state-of-the-art BERT model and find out that the simpler SVM model outperforms the BERT classifier, which they reason is due to the limited context present in the BERT model compared to the SVM model which works over the entire set of tokens using a bag-of-words approach. Their study concludes that the task of trigger warning assignment is non-trivial and requires further refinement. We build up on this study to utilize the social media data highlighting the presence of trigger warnings and present an in-depth analysis of our dataset and how computational modelling would improve identifying and flagging such social media posts using keyword detection and intent classification.

2.3 Related Datasets

The establishment of a user-level database of users on Twitter who have self-disclosed their experiences with various mental health problems is described in Suhavi et al. (2022). More than 10,000 Twitter users who have tweeted about their experiences with mental health illnesses, such as anxiety, depression, bipolar disorder, and eating disorders,(all being a type of trigger warning) are included in the database known as Twitter-STMHD. The authors extracted user-level data, including demographic statistics and information about the particular disorder, from tweets relating to mental health issues using a combination of machine learning algorithms and manual annotation. It provides users with a large-scale and well-labelled dataset grouped into 8 disorder categories and may have practical applications for mental health professionals seeking to identify and reach out to individuals who may need support. Gautam et al. (2020) describes the creation of a dataset of Twitter messages related to the #MeToo movement. The authors used manual and automated methods to annotate the dataset on five linguistic aspects: relevance, sarcasm, hate speech, stance, and dialogue acts. It also contains geographical information in the form of the country of origin of the tweet. This dataset labelled #MeTooMA is of particular interest to study in both computational and social linguistics and model how different linguistic components like stance, hate, and sarcasm interact in a social media context. While these papers have made the effort to analyze social media text, They have not explored the presence of trigger warnings in text, Its analysis and classification of various intents present in such content. We use techniques like keyword and keyphrase extraction to understand frequently used words in with maximum similarity with such content and topic modelling to generalize the topic assignment ability of BERTopic over our corpora. We also propose machine learning models to identify the intent of the social media posts using the post content and other user metadata.

3 Methodology

In this section we will briefly explain our data collection and Pre-Processing techniques, Exploratory Data analysis to emphasize the increasing trend of trigger warning posts online, Keyword and KeyPhrase analysis of the collected data, and evaluate various topic generation metrics to generalize the topic modelling quality and benchmark classification performance to predict the intent of posts mentioning trigger warnings. The entire framework is described in Fig 1.

3.1 Dataset

The dataset created is extracted and compiled from Reddit by using the PRAW Tool [2]. for extracting posts on the topics of self-harm, suicide, depression, and drug abuse. For the given topics we extracted data from the following subreddits:

- Self-harm: r/selfharm, r/SelfHarmScars
- Suicide: r/SuicideWatch
- Depression: r/Depression, r/MentalHealth
- Drug Abuse: r/RedditorsInRecovery

We extracted the following fields for our analysis including, A unique identifier for the Reddit Post (id), The username of the author of the post (author), The Title of the post (title), The number of upvotes minus the number of downvotes given to the post give us a measure of how much other users sympathize or agree with the post (score), The subreddit that the data is extracted from (subreddit), The text content of the post (text), The time that the post went up (utc_time), The number of comments, which gives us an idea of the popularity of the posts (num_comments), The comments on the post (comments), Whether the post only contains text, or it has a link to another resource (image, video, etc.) (is_self), A link_flair_text that describes the type of post (a question, asking for advice, etc.), Whether the post is restricted for viewers under the age of 18 (over_18) and the url of the post.

For EDA, Keyphrase mining, Topic modelling, and Classification we remove HTML tags, URLs, emoticons (cases of using emojis where it semantically differs from text is prevalent on social media), and special characters, while punctuation is retained. We extract certain content warning keywords for each category and construct our search phrase to find posts that contain these terms either in the post's title or in its body. We create a search phrase using keywords from a pool of words that are most commonly present in posts from a particular topic from all the posts from the subreddit from 2013 to 2022. Here, *x = ["trigger warning", "tw", "TW", "Trigger Warning"]* is the list of mentions we search in a post, and *y = [keywords_list]* is a curated list of words used by users from different communities, The search phrase created is

$$z = x_i + y_i : x_i \in x \text{ and } y_i \in y$$

[2] PRAW https://praw.readthedocs.io/en/stable/

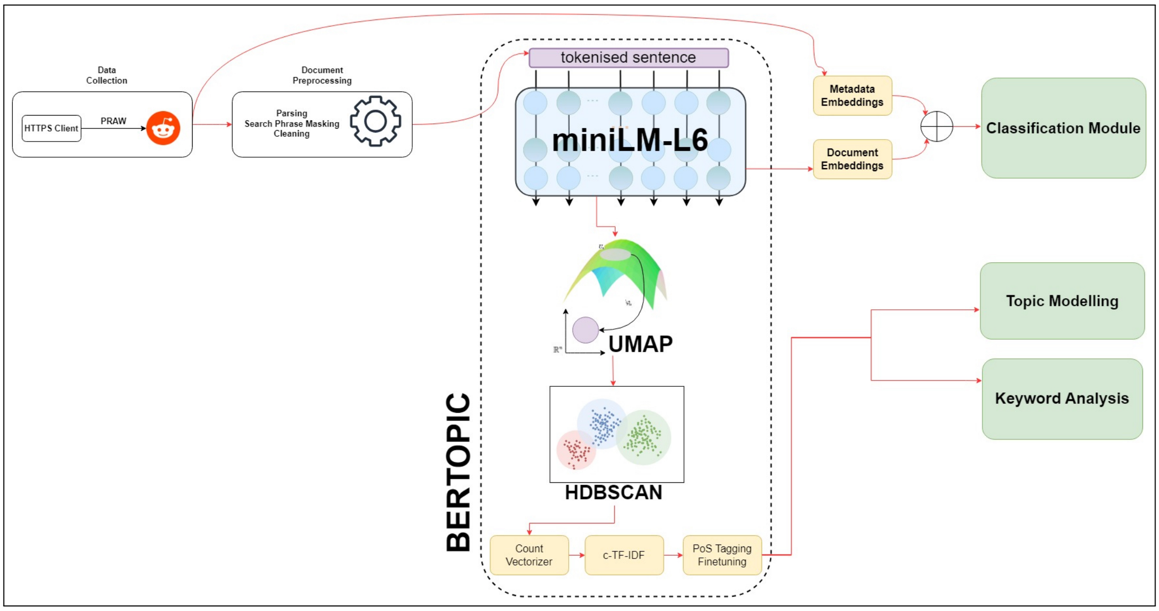

Figure 1: Overall Framework

	DATASET			
Measure	selfharm	suicide	depress	drugabuse
# Num of posts	4322	7090	11606	2437
Avg num of sentences	4.70	6.55	7.54	8.09
Mean Proportion of unique tokens	0.653	0.651	0.509	0.562

Table 1: Statistics of the dataset

3.2 Exploratory Data Analysis

3.2.1 Sentiment Analysis

In order to understand the sentiment, lexical and semantic depth of the user's post across different categories, we employ various techniques to extract such insights. We utilize a valence-aware dictionary and sentiment reasoner (VADER) tool. It is a lexicon and rule-based sentiment analyzer sensitive to the polarity and intensity of sentiments expressed in social media content due to its gold-standard lexicon attuned to a microblog-like context. Fig 2. depicts the Proportions of the documents with negative, neutral, or positive sentiments, These sentiments were calculated using the compound score, which is obtained by summing the valence score which is a value between -4 (negative) and 4 (positive) of each word in the lexicon/sentence which is added and is normalized between -1 and 1. Typical thresholds for a compound score for a positive sentiment are > 0.05, neutral sentiment for > -0.05 and < 0.05, and negative for < -0.05. A striking detail that we can notice is the prevalence of a high percentage of documents with positive sentiment from the drugabuse class. This highlights the point that the consumption of drugs and substance abuse is regarded as an activity that induces a feeling of euphoria and is not regarded as a negative belief within the community. The highest percentage of negative sentiment belongs to the depress class as people tend to convey their feelings and emotions on such social websites in an anonymous manner. The suicide class has the highest percentage of neutral sentiment as users often discuss narratives from films, tv shows, etc, or discuss laws pertaining to suicide in general.

3.2.2 Trend Analysis

We visualize the sudden surge in the frequency of posting on the platform with trigger warnings mentioned in subreddits as shown in Fig 3. corresponding to depress, drugabuse, selfharm, and suicide communities from the year 2019 when COVID-19 was declared and everyone was forced to be isolated and quarantined to ensure public safety. This explains the increase in user posting as people suffer from different mental issues during the surge of COVID-19. It is also evident that during this period

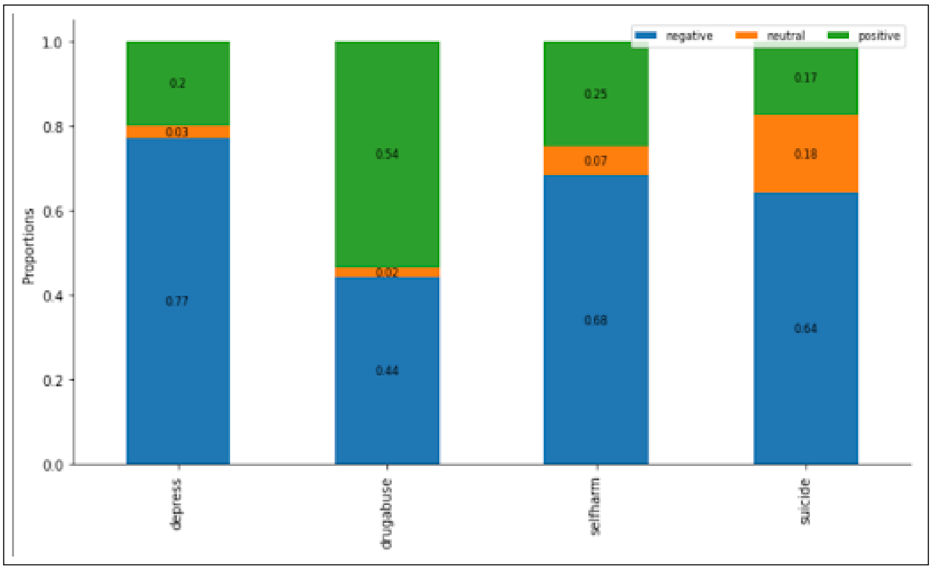

Figure 2: Percentage of posts in each sentiment category using the VADER tool

there was a rise in anxiety-related issues, domestic violence, and drug abuse as people had to stay at home against their choices with little to no physical interaction with the outer world. Many people who were dependent on alcohol and other substances experienced withdrawal symptoms, such as delirium and seizures, as a result of the abrupt closure of all liquor stores during the COVID-19 surge. Several alcohol "addicts," troubled by their urges, had turned to harmful drugs like hand sanitisers as replacements, These people also took to social media about their withdrawal symptoms and the distress they were going through [3].

3.2.3 Keyword/ Keyphrase Analysis

We present a thorough keyphrase analysis to inspect the different keywords and keyphrases used by the users from the collected corpora to understand the utilization of words when describing an incident, experience, or scene mentioning trigger warnings on social media websites. Keyword analysis provides us with a deep insight into the knowledge contained within the text and builds up an idea of the nature of the document. With the vast size of the text resources available on social media websites, manual extraction of keyphrases has become infeasible. Automatic keyword extraction [10] not only streamlines this process but also allows the reader to get an idea about the post's content in a very short period of time without going through the details and disregarding posts containing various sensitive/disturbing keywords or keyphrases. we utilize for KeyBert algorithm for our analysis. KeyBERT Grootendorst (2020) is a technique for extracting keywords and keyphrases from a document that utilizes BERT embeddings. It is simple and straightforward to use and generates keywords and key phrases that closely match the content of the document. The method employed by KeyBERT involves utilizing BERT embeddings and a basic cosine similarity technique to identify the sub-phrases within a document that is most closely related to the document as a whole. Initially, BERT is utilized to extract document embeddings, which results in a representation of the document at a high level. Following this, embeddings for N-gram words or phrases are obtained. Ultimately, cosine similarity is used to identify the words or phrases that are most similar to the document. We sample 5 keywords for each corpus with the highest similarity as it captures the semantics of the entire post's content.

The results from KeyBERT are as follows:

- Self Harm: 'fun': 0.8011, 'anxiety': 0.7559, 'normal': 0.7373, 'relapse': 0.731, 'blades': 0.69
- Depression: 'ptsd': 0.8209, 'desperate': 0.8132, 'hey': 0.8014, 'relapsed': 0.797, 'stories': 0.7636
- Drug abuse: 'relapse': 0.658, 'craving': 0.609, 'ptsd': 0.6087, 'caffeine': 0.6067, 'withdrawal': 0.5809

[3] https://encorerecovery.com/understanding-hand-sanitizer-addiction/

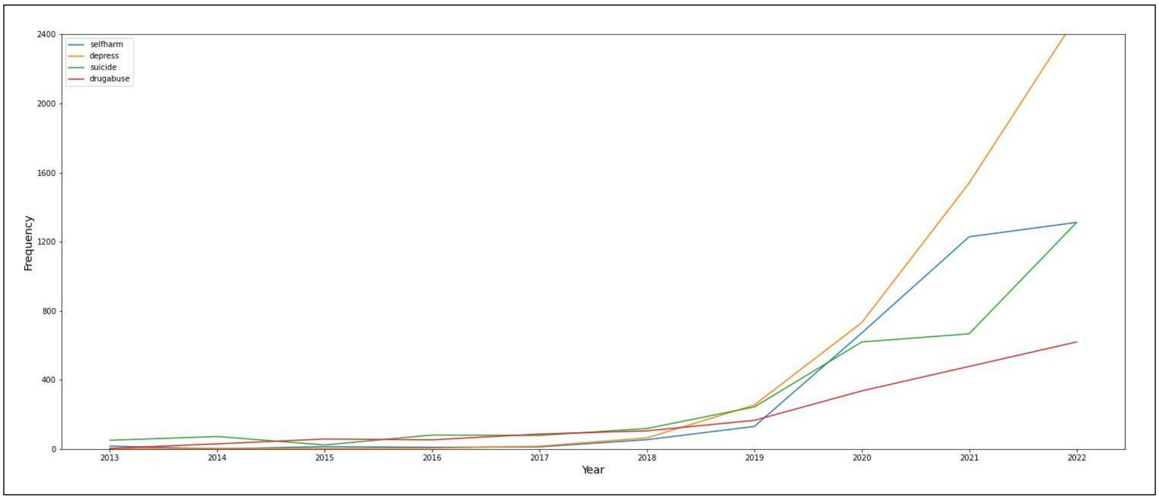

Figure 3: Trend Analysis of posting across different communities

- Suicide: ‘game’: 0.8591, ’hurts’: 0.8366, ’cut’: 0.8302, ’urgent’: 0.7709, ’relapse’: 0.7248, ’help’: 0.7217

Selfharm: The keyword "fun" suggests that there may be discussions or mentions of self-harm being perceived as pleasurable or enjoyable. It is important to approach this keyword with sensitivity, as it may indicate dark humour or the complexity of self-harm experiences. The presence of the keyword "anxiety" is likely to be mentioned when discussing anxiety triggers related to self-harm. Discussions involving trigger warnings revolve around understanding and normalizing self-harm experiences and that mention the keyword “normal” and direct a focus on reducing stigma and creating a safe space for individuals and promoting openness. Mentions of "relapse" indicate that individuals within the self-harm subreddit may share experiences or seek support related to relapses in self-harm behaviours. "blades" suggests that discussions or mentions of specific self-harm tools cause self-affliction.

Depression: "ptsd" suggests community discussion to address potential triggers related to post-traumatic stress disorder within the context of depression. This indicates that individuals may share experiences or seek support for co-occurring PTSD and substance abuse issues."desperate" indicates the use of the term in discussions involving desperate situations related to depression and individuals expressing feelings of desperation. The keyword "relapsed" and “stories” are likely to be used when discussing experiences, stories, or offering support, This signifies a likely presence of personal connections in the community.

Drugabuse: "relapse" indicates discussions involving experiences of relapse in drug abuse. It hints towards a discourse about preventing and recovering from relapse and its presence highlights the challenges and complexities individuals face in maintaining their recovery journey. "craving" might insinuate that discussions within the community revolve around the intense desire or urge to use drugs. "ptsd" suggests dialogue which addresses the intersection of drug abuse and post-traumatic stress disorder (PTSD) and draws the need to concurrently deal both at the same. “caffeine” can be used to compare the withdrawal symptoms from lack of caffeine to the withdrawal symptoms of drug addiction, “withdrawal” might imply the various challenges one would face during the recovery process.

Suicide: The presence of the word "game" suggests that discussions within the subreddit may involve references to video games or the gaming culture. This also highlights gaming as a source of comfort or a getaway, A user can also be simply sharing a gaming plot over a discussion thread. The use of the word ”hurts” in community discourse focuses on emotional pain, distress, or hopelessness. “cut” implies discussions revolving around self-harm behaviours especially cutting where a user might share their own story or reference other texts emphasizing the use of force to wound themselves. The use of words “urgent” and “help” and their interplay might be used to draw the community’s attention toward situations requiring immediate intervention and assistance.

4 Computational Modeling

4.1 Topic Modeling Evaluation

Topic models can automatically extract groups of related words from large collections of text without human guidance. By analyzing the words used in a particular document, these models can identify the key topics discussed within it. The need for coherence metrics Röder et al. (2015); Terragni et al. (2021b) has emerged in the field of text mining, where unsupervised learning methods such as topic models do not provide assurances about the interpretability of their results. We present a thorough evaluation of our trained topic models on both topic coherence and topic diversity to measure their interpretability of identifying appropriate topics from our trigger warnings corpora. In turn, This also helps us to understand the quality of topics discovered by our topic model and its importance in real-world applications to summarize documents with topics comprising potential trigger warnings to avert users from such content. Topic Coherence measures evaluate a particular topic by quantifying the level of semantic similarity between the most highly rated words within that topic Stevens et al. (2012). Such evaluations help distinguish between topics that can be semantically interpreted and those that are merely statistical artifacts of the inference process. We utilize the following topic coherence measures for evaluating the results produced by the BERTopic model: UMass Röder et al. (2015); Stevens et al. (2012) calculates how often two words, w_i and w_j appear together in the corpus and it’s defined as

$$C_{UMass} = \frac{2}{N \cdot (N-1)} \sum_{i=2}^{N} \sum_{j=1}^{i-1} \log \frac{P(w_i, w_j) + \epsilon}{P(w_j)} \tag{1}$$

where $D(w_i, w_j)$ indicate how many times words w_i and w_j appear together in documents, and $D(w_i)$ is how many time word w_i appeared alone. To determine the overall coherence of a topic, we take the mean coherence score for each pair of the top N words that best represent the topic. The UMass metric is unique in that it calculates these statistics based on the same corpus that was used to train the topic models, rather than an external corpus. This makes it an intrinsic metric that seeks to validate that the models have indeed learned the data present in all the corpora. C_V Röder et al. (2015)is a widely used method for measuring coherence which involves constructing content vectors based on the co-occurrences of words within a topic. This metric then uses normalized pointwise mutual information (NPMI) and cosine similarity to compute the coherence score. C_{UCI} Röder et al. (2015); Stevens et al. (2012) computes the coherence score by measuring the frequency of two words appearing in a document. Instead, We employ a sliding window method and calculate the pointwise mutual information between all pairs of the top N words by occurrence and examine their co-occurrence within the sliding window. If both words, w_i and w_j are present in a document but not within the same sliding window, we do not consider them as co-occurring.

$$C_{UCI} = \frac{2}{N \cdot (N-1)} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} PMI(w_i, w_j) \tag{2}$$

where

$$PMI(w_i, w_j) = \log \frac{P(w_i, w_j) + \epsilon}{P(w_i) \cdot P(w_j)} \tag{3}$$

The results are presented in Table 2. We see that the Drug abuse corpus achieves a considerably high coherence score of -0.693,0.701 and 0.701 followed by the suicide corpus with a score of -0.631, 0.642, 0.621 exhibiting that the topic model performs better on the aforementioned datasets in producing highly coherent topics, while the Selharm and Depression dataset shows an intermediate score suggesting some degree of semantic similarity with reasonably interpretable generated topics.

In addition to coherence, we also consider topic diversity as a measure. The greater the diversity among the resulting topics, the broader the coverage of various aspects of the analyzed corpus will be. It’s crucial to generate topics that differ from one another as it ensures the topic model produces non-redundant themes and learns a diverse topic distribution. Topic diversity Dieng et al. (2020) refers to the proportion of distinct words among the top n words in all topics. A diversity score of nearly 0 suggests that the topics are redundant, while a score close to 1 implies more diverse topics. Inverted Rank Biased Overlap (IRBO): It measures the overlap between two lists A and B over various depths by taking the number of elements intersecting between the two lists and then normalizing it by the depth (number of topics considered). Averaging over all the depth sizes gives us the RBO score for the two lists. The inversion of this RBO score gives us a measure of how distinct the two lists are Tan and Clarke (2014); Webber et al. (2010). The diversity results are presented in Table 3. We use the OCTIS framework Terragni et al. (2021a) to evaluate the topic diversity metrics. The results indicate that the BERTopic model gives us a diverse set of topics across all four datasets that are used. Suicide and Self Harm show comparatively higher diversity scores across both the used methods. Topic diversity scores for all the datasets are above 0.8 suggesting topics within each dataset are unique, non-repetitive, and cover a wide range of themes.

4.2 Classification

We use the extracted data from various different subreddits which are a part of self-harm, suicide & depression and drug abuse subreddits to determine user intent by using the link flair text as the exogenous variable. The posts with link flair texts were divided into training and test set respectively, Posts with no tags were manually labelled which included text with no trigger warning as well. Labelling was performed as follows:

- Created a lexicon-based dictionary of different link flair text categories.
- Assigned different tags with mentions of the same category into a singular category, Thus reducing the size of unique classes by binning data into semantically similar categories.

We evaluate the problem of user intent prediction as a multi-class classification problem to determine the nature of the post. This is crucial for the automatic intent detection of a user's post which includes several trigger warnings and helps users to view content of their interests by understanding the intention of the post. Our classification problem exhibits a significant imbalance in the distribution of the target classes: for instance, there are several times more posts asking questions, advice, etc. than actually mentioning any distressing event. We use stratified sampling for 10 folds to ensure that relative class frequencies are approximately preserved in each train and validation fold with a division of 70:30 for the training and validation set for each stratified sample comprising 1296, 2127, 3481, and 730 testing samples. We construct a strong baseline by first concatenating the BERTopic embedding of a subreddit's post text with other numerical features like num of comments, the num of likes, the score of the post, and the upvote ratio of the post represented using a singular prompt. This prompt is used to generate embeddings using MiniLM-L6, The same model used to train our Topic model and we then concatenate these embeddings to generate a singular feature vector. The MiniLM-L6 model Wang et al. (2020) is fine-tuned for topic classification and maps the text data to a 384-dimensional vector state, This model is trained on a set of 1B pairs of text using a contrastive learning objective. The model is compressed using a self-attention distillation process to reduce parameter size and make model serving easier. We train xgboost and lightgbm models as our baseline to predict the intent of the post due to their performance on highly sparse input data and make these models our first choice for training on combined embeddings of post content and user metadata. The evaluation metrics utilized are Accuracy, Precision, Recall, and F1 score. We report our scores in Table 4 where the best scores are highlighted in bold for each corpus and model. The xgboost and lightgbm perform well over classifying the posts for our classification task with both models performing quite evenly over the datasets, Only the drugabuse dataset-trained models report the lowest performance metrics due to the comparatively smaller dataset size with the lowest F1 score of 0.858 whereas for other datasets its 0.90 or above. Classification over such trigger warning datasets is the first one to be done which can benefit social media companies to label such posts and enhance user experience.

5 Conclusions

In this paper, we present an initial study of evaluating and inspecting data containing trigger warnings in social network groups dealing with different clinical disorders. We present a unique dataset that contains user-level mentions of trigger warnings present in their content, exploratory data analysis measuring the sentiment across posts using statistical and lexical techniques, trends of such posts

	DATASET			
Measure	**Depression**	**Drug Abuse**	**Self Harm**	**Suicide**
u_mass	-0.562	-0.693	-0.532	-0.631
c_v	0.563	0.701	0.567	0.642
c_uci	0.547	0.701	0.511	0.621

Table 2: Results of U_{Mass}, C_V, and C_{UCI} Topic Similarity measures on the datasets.

	DATASET			
Measure	**Depression**	**Drug Abuse**	**Self Harm**	**Suicide**
Topic Diversity	0.8	0.844	0.878	0.881
IRBO	0.821	0.854	0.882	0.884

Table 3: Results of Topic Diversity and Inverse Rank Based Overlap (IRBO) measures on the datasets.

Corpus	**Model**	**ACC**	**P**	**R**	**F1**
selharm	BERTopic+xgboost	0.895 ± 0.0052	0.946	0.835	0.887
	BERTopic+lightgbm	**0.899 ± 0.0051**	**0.948**	**0.857**	**0.900**
suicide	BERTopic+xgboost	**0.920 ± 0.0047**	0.944	**0.890**	**0.916**
	BERTopic+lightgbm	0.910 ± 0.0045	**0.948**	0.883	0.914
depress	BERTopic+xgboost	**0.930 ± 0.0475**	**0.947**	**0.887**	**0.912**
	BERTopic+lightgbm	0.920 ± 0.0048	0.945	0.879	0.910
drugabuse	BERTopic+xgboost	0.890 ± 0.0043	**0.930**	0.796	0.857
	BERTopic+lightgbm	**0.900±0.0042**	0.928	**0.798**	**0.858**

Table 4: Classification performance on the test set for all four datasets reported in accuracy (ACC), precision (P), recall (R), and F1 score.

online, and keyword/keyphrase analysis to discover words used in relation to events accentuating the presence of trigger warnings. We further evaluate topic modelling metrics to generalize the BERTopic model's topic generation ability over the different datasets and measure their quality. In the end, we evaluate the four labelled corpora in a text classification setting using the document's content and other metadata information and build classification models to detect trigger warnings at the document level.

6 Future Work

Our next goals are to study the same phenomenon on other social networking sites such as Twitter and the expansion of this initial study into a wider domain by covering multiple languages by employing multilingual transformer models. We plan to incorporate additional modalities including audio-visual data to be used because different modalities convey relevant psychological and social aspects of a social media user. We aim to include a demographic-based analysis using choropleth maps to represent a fine-grained analysis of trigger warning tags across the world and finally use manual labelling and other prompt-based techniques to create pseudo labels based on a list of seed words of posts and comments into a specific type of trigger warnings context and its stance.

7 Acknowledgements

We would like to convey our gratitude to Dr Divya Dabas Chaudhary, Dr Ian Gorton and Dr Bijender Kumar for their advice and feedback. We would also like to thank the anonymous reviewers for their constructive reflections on the paper.

References

Christine Ballestrini. 2022. University of Connecticut Office of the Provost | Trigger and Content Warning Guidance.

Benjamin W. Bellet, Payton J. Jones, and Richard J. McNally. 2018. Trigger warning: Empirical evidence ahead. *Journal of Behavior Therapy and Experimental Psychiatry*, 61:134–141.

Victoria M. E. Bridgland, Deanne M. Green, Jacinta M. Oulton, and Melanie K. T. Takarangi. 2019. Expecting the worst: Investigating the effects of trigger warnings on reactions to ambiguously themed photos. *Journal of Experimental Psychology: Applied*, 25(4):602–617.

Ashleigh Charles, Laurie Hare-Duke, Hannah Nudds, Donna Franklin, Joy Llewellyn-Beardsley, Stefan Rennick-Egglestone, Onni Gust, Fiona Ng, Elizabeth Evans, Emily Knox, Ellen Townsend, Caroline Yeo, and Mike Slade. 2022. Typology of content warnings and trigger warnings: Systematic review. *PLOS ONE*, 17(5):e0266722.

Adji B. Dieng, Francisco J. R. Ruiz, and David M. Blei. 2020. Topic modeling in embedding spaces. *Transactions of the Association for Computational Linguistics*, 8:439–453.

Akash Gautam, Puneet Mathur, Rakesh Gosangi, Debanjan Mahata, Ramit Sawhney, and Rajiv Ratn Shah. 2020. #MeTooMA: Multi-Aspect Annotations of Tweets Related to the MeToo Movement. *Proceedings of the International AAAI Conference on Web and Social Media*, 14:209–216.

Muriël de Groot, Mohammad Aliannejadi, and Marcel R. Haas. 2022. Experiments on Generalizability of BERTopic on Multi-Domain Short Text.

Maarten Grootendorst. 2020. Keybert: Minimal keyword extraction with bert.

Maarten Grootendorst. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. ArXiv:2203.05794 [cs].

Payton J. Jones, Benjamin W. Bellet, and Richard J. McNally. 2020. Helping or Harming? The Effect of Trigger Warnings on Individuals With Trauma Histories. *Clinical Psychological Science*, 8(5):905–917.

Bayode Ogunleye, Tonderai Maswera, Laurence Hirsch, Jotham Gaudoin, and Teresa Brunsdon. 2023. Comparison of Topic Modelling Approaches in the Banking Context. *Applied Sciences*, 13(2):797.

Michael Röder, Andreas Both, and Alexander Hinneburg. 2015. Exploring the Space of Topic Coherence Measures. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, pages 399–408, Shanghai China. ACM.

Keith Stevens, Philip Kegelmeyer, David Andrzejewski, and David Buttler. 2012. Exploring Topic Coherence over Many Models and Many Topics. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 952–961, Jeju Island, Korea. Association for Computational Linguistics.

Suhavi, Asmit Kumar Singh, Udit Arora, Somyadeep Shrivastava, Aryaveer Singh, Rajiv Ratn Shah, and Ponnurangam Kumaraguru. 2022. Twitter-STMHD: An Extensive User-Level Database of Multiple Mental Health Disorders. *Proceedings of the International AAAI Conference on Web and Social Media*, 16:1182–1191.

Luchen Tan and Clarke L. A. Clarke. 2014. A Family of Rank Similarity Measures based on Maximized Effectiveness Difference.

Silvia Terragni, Elisabetta Fersini, Bruno Giovanni Galuzzi, Pietro Tropeano, and Antonio Candelieri. 2021a. OCTIS: Comparing and Optimizing Topic models is Simple! In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 263–270, Online. Association for Computational Linguistics.

Silvia Terragni, Elisabetta Fersini, and Enza Messina. 2021b. Word embedding-based topic similarity measures. In *Natural Language Processing and Information Systems: 26th International Conference on Applications of Natural Language to Information Systems, NLDB 2021, Saarbrücken, Germany, June 23–25, 2021, Proceedings*, page 33–45, Berlin, Heidelberg. Springer-Verlag.

Wenhui Wang, Hangbo Bao, Shaohan Huang, Li Dong, and Furu Wei. 2020. MiniLMv2: Multi-Head Self-Attention Relation Distillation for Compressing Pretrained Transformers.

William Webber, Alistair Moffat, and Justin Zobel. 2010. A similarity measure for indefinite rankings. *ACM Transactions on Information Systems*, 28(4):1–38.

Magdalena Wolska, Christopher Schröder, Ole Borchardt, Benno Stein, and Martin Potthast. 2022. Trigger Warnings: Bootstrapping a Violence Detector for FanFiction.

8 Appendix

This section includes examples of the highest similarity from our collected datasets for the keyphrases detected which supports our speculations about the top 5 detected words for each keyword.

Selfharm: example of the keyword "fun" is "tw child abuse cutting I self-harm for two reasons the first is to help regulate emotions and to cope with things when I'm panicking very stressed and need to calm down quickly I often resort to self-harm I generally do a good job coping with this aspect of self-harm because I have years of healthy coping methods behind me resources like subreddits and friends who can help support me the only time I really slip up with it is if I'm restricted by time in some way the second reason enjoying the scarring it creates is a lot harder to deal with when I used to cut so far I've been able to resist going back to it mostly the scarring was something I liked about it I found the scars I had deeply important to me they were representative of an internal struggle".

Suicide: example of the keyword "game" is "trigger warning for mentions of suicide and sexual assault possible spoilers for cyberpunk number however im going to keep details vague im posting this here instead of on the cyberpunk subreddit because this has less to do with the game and more to do with my reaction to it this is really messy and cobbled together."

Depression: example of the keyword "ptsd" is "I am tired of fighting. My depression, anxiety, PTSD, and being jobless are destroying me."

Drugabuse: example of the keyword "relapse" is "I struggle deeply with anxiety and depression, and the need to relieve my pent-up anger and emotions led me to self-harm several times because I didn't want to relapse on smoking. But I can't decide what's better or worse for me at this point. Would it be so bad if I were to smoke again? I know I previously had issues of self-control, but I really don't know what else to do as I'm doing my very best to make all the positive changes in my life and I still feel horrid."