Mental Disorder-Triggering Content Detection

Andrew Rippy, Klaijan Sinteppadon, Olivia Xu

{arippy, ksintepp, okx}@andrew.cmu.edu
Carnegie Mellon University
5000 Forbes Ave
Pittsburgh, Pennsylvania 15213

Abstract

Mental disorder triggers and trauma triggers often come without warning when users browse the Internet. For many of these users, reading triggering content may cause them to respond very severely and potentially worsen their disorder or trauma. As online content is an integral part of our lives, it is critical that we limit triggering content to lessen the harmful impact on those with mental disorder and trauma sensitivities. The areas of detecting triggering topics still have yet to be widely covered. In this study, we aim to create a mental disorder-triggering and trauma-triggering content detection model that covers a wide range of related topics, using data from Reddit, and we experiment with multiple models, including a SVM model and a BERT model. The application of the model is to deploy it as an extension of a web browser, specifically Chrome, to ensure safe content consumption for users, by their choice.

Introduction

Online content is constantly growing and as a result, it contains an abundance of mental disorder- and trauma-triggering content, especially without any warning. For example, this could range from content that could be triggering to those suffering from depression or post-traumatic stress disorder (PTSD) to those who are suicidal. Furthermore, content like this is even more apparent on social media platforms where any user can post or share content. It is estimated that around 5% of adults in the US suffer from PTSD yearly (US Department of Veterans Affairs 2023) and around 4% of adults in the US have suicidal thoughts yearly. Additionally, almost 17% of women alone have experienced sexual violence in the US (Schrading et al. 2015). These statistics only account for a few common mental disorders and traumas.

Furthermore, prior research and personal accounts have exemplified that such unexpected content can harm users and that adding warnings can be helpful for these users. For instance, a previous study interviewed medical students regarding their perspectives on trigger warnings, with one student commenting on their necessity: "When someone explained how suicide, because of their own personal experience, really upsets them without a warning in lectures, you can understand the need for [adding warnings]" (Nolan and Roberts 2022). The study also found that trauma-related content could cause adverse emotional reactions and that it could be minimized with warnings. Similarly, it was found that exposure to certain health-related online content could aggravate a user's anxiety (Ali et al. 2022). Many online users have talked about their experiences when reading or watching triggering content, with one user recalling how watching news broadcasts on various geopolitical events, including war and loss of life, triggered "a severe psychotic manic episode" resulting in hospitalization (Nelligan 2022). Another user explained how triggering content often brings them back to a place they "'do not want to go back [to right now]'" (Ponte 2022). All of these examples illustrate the potential harm that reading triggering content can bring to certain people. A key societal challenge is in ensuring that those suffering from these mental health issues and other similar issues are not put at risk or further harmed by reading such triggering content.

Even though the notion of a "trigger" differs between individual experiences, it is common for exposures to reminders similar to an individual's mental disorders or traumatic experiences to be a trigger. This is especially prominent for those suffering from PTSD and other related traumas (National Center for PTSD 2022). Having an appropriate warning from related content can help people employ effective management of their own mental health and allow them to engage with this content whenever they want to, at the amount they are comfortable with.

In response to this, we propose using AI to preemptively find and detect such mental disorder- and trauma-triggering content to ensure that those who suffer from these issues can be made aware of and avoid such content if desired. We propose building a baseline support vector machine (SVM) model and a model based on bidirectional encoder representations from transformers (BERT) to detect triggering content. Whereas many techniques have been developed to detect whether users suffer from these mental disorders and traumas through text using various machine learning (ML) techniques, little to no research has been done on detecting content that *triggers* those suffering from these mental disorders and traumas. Additionally, very few applications have actually integrated these techniques in practice. As a result, we also propose building a Chrome web extension that uses our model so that users can actively be made aware of potentially triggering content. We elaborate further on the work

that has been done in this field, as well as the current solutions that exist in the next section.

Related Work

In the space of detecting harmful online content, there has been a lot of work focused on detecting toxic content and triggering content in certain spaces. For instance, various ML algorithms ranging from logistic regression to decision trees to KNN classification have been used to accurately detect toxicity in text, with some results indicating the effectiveness of logistic regression with an accuracy around 90% (Kajla et al. 2020). Beyond these simple ML methods, there has also been research done on more advanced techniques that have resulted in better detection of cyberbullying, another form of harmful online content. These included using bidirectional neural networks, which gave an accuracy near 95% (Raj et al. 2021).

Recently, pre-trained large language models have also been studied in depth given their high performance in detecting harmful online content, most notably BERT. One prominent example of BERT was introduced with HateBERT, which is a re-trained BERT model for abusive language detection in English, and the model was trained on a large-scale dataset of Reddit comments in English from banned communities, with HateBERT managing to outperform other models (Caselli et al. 2020). Additionally, Dacon et al. (2022) found that BERT outperformed all of their other baseline models (i.e. logistic regression and SVMs) in detecting anti-LGBTQIA+ conversations online.

These recent works have shown the effectiveness of using BERT in the context of detecting content through text. Furthermore, BERT has been used in the context of detecting whether a person is suffering from depression through text and if a person is experiencing sexual harassment in messages, as examples (Senn et al. 2022; Yan and Luo 2021). While there has been a lot of work done on detecting content related to having or showing symptoms of these mental disorders and traumas, very little work has been done on detecting their *triggers*. Specifically, there has been little to no work done on detecting PTSD-, eating disorders-, suicide-, and abusive relationships-triggering content, and so we chose to focus on these areas for our project.

There has been work done in detecting whether a user has PTSD or shows signs of having PTSD in an online context (Ameer et al. 2022), but there does not seem to be any relevant research on using AI techniques to detect PTSD-*triggering* content, as far as the authors are aware. Similarly, there has been little to no research done on detecting eating disorders-*triggering* content. Existing research in this area has focused on classifying images as whether or not it shows a eating disorder (Counts, Manning, and Pless 2018) and early detection of anorexia using various ML models (Villegas, Cagnina, and Errecalde 2022), but they have not gone further to identify triggers. In another vein, there has been significant research done on detecting those who are at risk of committing suicide and / or are suicide ideators. For instance, Fodeh et al. (2019) proposed a ML framework, involving Latent Dirichlet Allocation and decision trees, to identify suicide risk factors on Twitter, while Aldhyani et al. (2022) also worked on using various ML models to classify whether a user is exhibiting suicidal ideation based on their posts. Yet, again, there does not seem to be extensive research done on detecting suicide-*triggering* content. There has been some minimal research done on studying how to predict whether text contains abusive relationships-*triggering* content, but only through more traditional ML techniques, such as SVMs, Naive Bayes, and decision trees (Roy, McClendon, and Hodges 2018; Amrit et al. 2016; Subramani, Vu, and Wang 2017). However, very little research has been done on detecting abusive relationships-triggering content with large language models, including BERT (Karystianis et al. 2021; Nayak and Baek 2022). We further build on these existing works by introducing and attempting to detect relevant mental disorder- and trauma-triggering topics, along with using large language models in this context.

We apply these aforementioned AI techniques to better detect PTSD-, eating disorders-, suicide-, and abusive relationships-*triggering* content, which is where the novelty of our work lies. This is important, especially given how common these mental health and health issues are and their far-reaching harmful effects.

Label	Number of Samples
SuicideWatch	2,977
ptsd	3,220
abusiverelationships	3,385
EatingDisorders	797
Others (Advice, Showerthoughts)	3,706

Table 1: Collected data

Methods

Data Collection

We focused on a subset of Internet content by using Reddit, given its abundance of content and how popular it is among online users. The data is scraped from Reddit, using the PMAW PushshiftAPI, under 6 Subreddits: r/ptsd, r/EatingDisorders, r/SuicideWatch, r/abusiverelationships, r/Advice, and r/Showerthoughts. The last two subreddits are scraped so as to illustrate non-mental disorder- / trauma-triggering topic-related texts and are combined into one category labelled as "Others".

Due to PushshiftAPI maintenance issues, we were not able to obtain Reddit data before November 2022, and there is no better way to access Reddit data in bulk other than through the API. The current data is only obtained from November 2022 until February 2023, at the maximum of 1,000 posts per month. In total, the data per each subreddit is a maximum of 3,594 posts and a minimum of 112 posts. Data obtained consisted of post information, such as `title`, `selftext`, and `subreddit`. More information on the amount of data we collected and used is in Table 1. Additionally, a small example of the data we collected is shown in Figure 1.

title	selftext	labels	utc_datetime_str	created_utc	text
possible weirdo knows where i live.	promised myself i'd never use reddit but here ...	Others	2023-02-09 02:11:16	1.675909e+09	possible weirdo knows where i live promised my...
what are your thoughts?	hi all,\n\n**i could really use your help and ...	ptsd	2022-12-02 12:39:54	1.669985e+09	what are your thoughts hi alli could really us...
just got a c on an exam, parents will be pissed.	probably gonna kill myself soon. fuck high exp...	SuicideWatch	2022-12-09 18:52:54	1.670612e+09	just got a c on an exam parents will be pissed...
roommate's sleep disorder symptoms are trigger...	title is pretty self explanatory, my roommate ...	ptsd	2023-01-20 06:28:04	1.674196e+09	roommates sleep disorder symptoms are triggeri...
straw that broke the camels back.	i have ptsd, some is accumulated from some chi...	ptsd	2023-02-23 00:39:32	1.677113e+09	straw that broke the camels back i have ptsd s...

Figure 1: Small example of data we collected

Learning rate	5e-6
Batch size	32
Number of epochs	15
Dropout	0.5
Optimizer	AdamW

Table 2: Hyparameter values and optimizer choice

Preprocessing

For modelling with a language model, we concatenated the `title` and `selftext` columns together to yield one long paragraph (refer to Figure 1 to see the raw data). The whole text is then converted to lowercase, and non-alphanumeric characters are removed.

Models

We created three models in order to try out various implementations to see which model would best predict triggering content from the categories we identified above (i.e. PTSD, eating disorders, suicide, and abusive relationships). Our models aim to detect their triggers and by training on Reddit data containing the experiences and accounts of those suffering from these mental disorders and traumas, we believe it captures their most common triggers, as discussed in the introduction. Here is a link to the code for our models: https://drive.google.com/drive/folders/1GWb2VDg7q6XBNhTLkAuwpjCmMx7LAmxK.

SVM We built a baseline SVM model in order to compare and evaluate it against the large-language models. The pipeline of this process started with the raw text input, which was then preprocessed and fed into the SVM model in order to generate a prediction. The text preprocessing consisted of tokenizing, removing stop words, and inputting it into a TF-IDF vectorizer.

BERT We selected BERT as a framework given its power in text-related areas, in that a large-language model would likely better understand the context of text better than a bag of tokenized words. The text input was first tokenized using the BERT Tokenizer from the pre-trained model `bert-base-uncased`, and then it was fed into the model. The model was built on top of a BERT transformer by huggingface, with a self-attention pooling layer and dense layers. The relevant hyperparameters and the optimizer we used are shown in Table 2. The model predicts the classification output as a subreddit classification.

BERT-BERT This is similar to our BERT model, with the main difference being that two BERT models are used with one learning from our dataset and the other learning from a sentiment dataset provided by huggingface. We leveraged this dual attention transfer network in order to account for the small size of the training data we had and the knowledge from other available and richer datasets. We built on the work proposed by Yu et al. (2018) on emotion classification with such a model. The hyperparameters and optimizer we used are the same as for the original BERT model.

	SVM	BERT	BERT-BERT
SuicideWatch	0.87	0.81	0.82
ptsd	0.75	0.76	0.77
abusiverelationships	0.85	0.67	0.73
EatingDisorders	0.73	0.83	0.83
Others	0.81	0.72	0.72

Table 3: F1-scores

Results and Evaluation

Each of the three models we used were evaluated. We primarily looked at the F1-score in order to balance specificity and sensitivity. Having too high specificity results in the risk of exposing online users to triggering content while having too high sensitivity results in hindering the Internet browsing experience. The F1-scores for each model are shown in Table 3. Additionally, the confusion matrix with the results of SVM is shown in Figure 2, and the confusion matrix with the results of BERT-BERT, which is the best model, is shown in Figure 4.

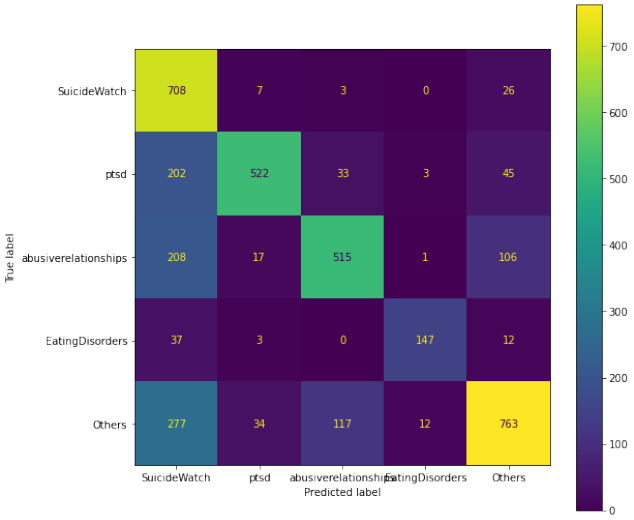

Figure 2: Confusion matrix of each category with SVM

Based on the F1-scores, there does not seem to be any single best model out of the SVM, BERT, and BERT-BERT models. They are all able to capture the text that belongs to each class quite well. For BERT-BERT, which utilizes a transfer network, it actually seems that the knowledge from the additional sentiment dataset does not seem to contribute to a significant improvement because most of our collected

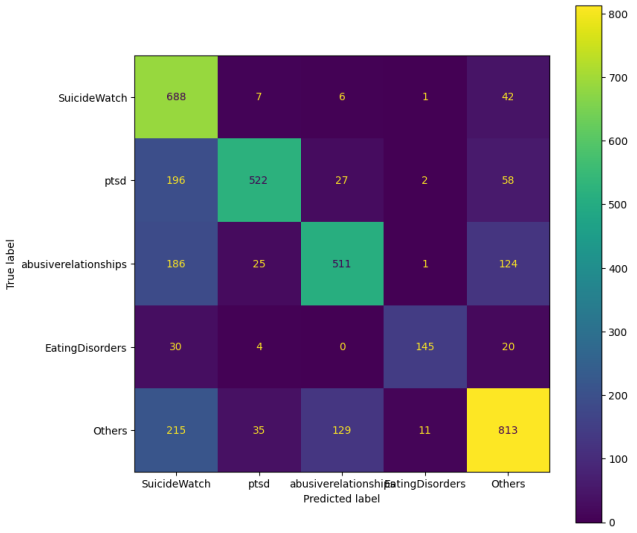

Figure 3: Confusion matrix of each category with BERT

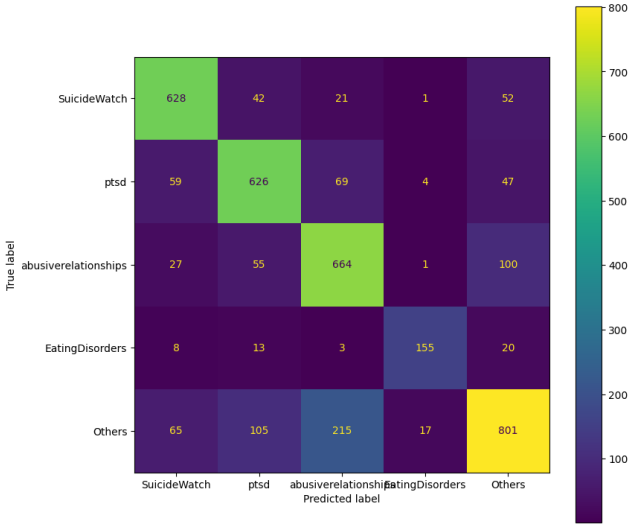

Figure 4: Confusion matrix of each category with BERT-BERT Dual Attention Transfer Network

data has an associated "negative" sentiment. The section below shows examples of how the models classified paragraphs of text and that the models are able to correctly classify in most relevant scenarios.

The following examples are the actual text scraped from Reddit, without any personally identifiable information mentioned here in our paper. It was preprocessed by removing non-alphanumeric characters and subsequently converting to lowercase.

This text contains content that is potentially triggering for those with PTSD as it contains descriptions about injuries, and the models correctly classified this as PTSD-triggering:

> when i was 4 i had a very severe ear infection ... i saw the doctor who didnt see any issues and sent me home ... later they found out the infection had gone through my skull ... i had surgery and had a tube implanted in my chest [for] 6 weeks ... i wonder if it is somehow subconsciously affecting me today ...
>
> **Classification**: ptsd

This text does not contain any PTSD-, eating disorders-, suicide-, or abusive relationships-triggering content, and the models correctly classified this as "Others":

> my boyfriend is coming over for the weekend and i want to sleep in the same bed as him ... my parents are very conservative and i often struggle for asking for things from them ... i dont want to push but i think [my mom] might understand if i phrase it right please help
>
> **Classification**: Others

This text may very likely be triggering for those with suicidal thoughts or suicidal ideations given that it mentions a drug that can be very dangerous if overdosed on (Mount Sinai 2023). The SVM model classified this text as "Others", whereas the large language models (BERT and BERT-BERT) classified this text as "SuicideWatch". This shows that in certain cases, the large language models are able to better distinguish content that may be triggering for people than the SVM model:

> helpadvice ive took a full pack of cocodamol 30mg500mg about 5 minutes ago whats going to happen to me
>
> **Large-language models' classification**: SuicideWatch
>
> **SVM classification**: Others

Even though the SVM model provides similar results as the large language models in most cases, given the scalability of the large language models, namely BERT-BERT (as it does better than the BERT model), they are preferred over the SVM model as they are able to better understand nuance and context within text (as seen especially in the last example above). With the use of BERT-BERT in a deployed setting, we would expect online users to be able to avoid triggering content, or at the very least get to decide if they want to be exposed to content that tends to be triggering.

Furthermore, for the BERT-BERT model we chose, we evaluated the negative examples predicted by the model. There are a total of 924 texts that were incorrectly predicted, of which the true and predicted labels can be viewed in the confusion matrix shown in Figure 4. We further investigated the nature of these wrongly predicted texts, and we discovered that even though some of them were predicted incorrectly without any evidence of why, some content within these samples are actually related to the class they were predicted as. One of the possible reasons for this is because

these mental disorder- and trauma-triggering topics are not completely separate categories from one another; each category covers a wide range of other mental health issues and related triggers. The actual label of these texts depends on the user and which subreddit they posted in, and this is sometimes ambiguous to qualify. The other explanation is that because our dataset is relatively small, it can pose difficulties to the model in always predicting correctly.

The following examples show some incorrectly-predicted texts.

This text was originally posted under the subreddit "r/Advice" (which corresponds to our label of "Others") but was predicted as "abusiverelationships". The text's contents mention an intense action from a significant other that made the poster feel betrayed. The text is about a relationship and unhappy feelings, which could be the reason why it was classified as "abusiverelationships":

```
i found out my significant other
did something pretty intense that
had a significant effect on my
trust for her she did not cheat
... i am wondering now how long
i will be stuck wrestling with
my feelings of betrayal and my
feeling of obligation to her and
her family in her time of crisis
```

Actual class: Others

Predicted class: AbusiveRelationships

This text was originally posted under the subreddit "EatingDisorders" but was predicted as "SuicideWatch". The text's contents mention depression and killing, which can be triggering for those with suicidal ideations, and it only mentions being "thin" once:

```
depression which was already bad
and chronic has gotten much worse
... tried to kill and get rid of
the idea that being thin would
finally give me the life i wanted
and make my mother love me
```

Actual class: EatingDisorders

Predicted class: SuicideWatch

This text was originally posted under the subreddit "ptsd" but was predicted as "EatingDisorders". The text's contents mention being unhealthy and nothing obvious about "ptsd". While this text in particular neither does not seem to exhibit PTSD- or eating disorder-triggering content, we believe that the mention of being unhealthy may be more related to being triggering for those with eating disorders rather than PTSD in many typical contexts.

```
anyone else smoke to help their
symptoms ive been smoking a lot
to cope lately ... i know its
not healthy as a long term thing
... now its daily and definitely
feels like im medicating ...
```

Actual class: EatingDisorders

Predicted class: ptsd

These examples illustrate how our BERT-BERT model does in both successful and unsuccessful instances. In comparison to the other models (the SVM and BERT model), the BERT-BERT model is able to achieve relatively high F1-socres and is also able to capture nuances within the text.

Deployment

In order to address the lack of deployed solutions in helping online users identify triggering content, we attempted to deploy our model in an online setting by developing a Chrome web extension to identify triggering text on a webpage and blocking out that text from the user (the user can choose to keep that information from being blocked out, if desired). The use of an extension gives the user the flexibility of choosing to block out triggering text or not. By using the extension, users are able to identify triggering content before actually reading it.

Implementation

Chrome Web Extension We deployed our SVM model for the purposes of creating our Chrome web extension. We were unable to get our BERT-BERT model incorporated into the Chrome web extension due to difficulties with importing the BERT-BERT model, but this is something we will work on in the future given that the BERT-BERT model was preferred in terms of performance. The implementation can be found in this GitHub repository: https://github.com/rippy1849/AIFSG_Trigger_Warning.

At a high level, the Chrome web extension interfaces with a WebPy backend local server to send page data from the user to be taken as input by the model. It then outputs a labelled embedding that the server subsequently sends as a `postMessage` to the original extension. The extension reads this in and finds the relevant post, and then based on the label, colors the text to indicate the type of content.

Workflow In order to understand this Chrome webextension in more detail, reference Figure 5 throughout the following explanation (begin at the observer and the webpage, which is in the upper left of Figure 5):

1. The User's current page is interfaced with via the "Trigger Warning" Chrome web extension `content.js`. It crawls the page and creates a consolidated array list of the text-related elements, specifically paragraph and header elements.
2. For every element e_i of the list, a consecutive id numbering scheme is applied, of the form $e_i.id = text$. We will refer to the labelled element as e_{l_i}. For each e_{l_i}, an `iframe` element is created, with a src of `webserverURL + ?id=i + ?text="textcontent"`.
3. `backend.py` accepts `id,text` as the input, and it converts these to Python-usable variables. `backend.py` passes the text to the imported trigger model function from `model.py`.
4. `model.py` loads the model weights via `svm.sav` and the text tokenizer via `tfidfvectorizer.pk`. It returns the label back to `backend.py` via labelText(inputText).

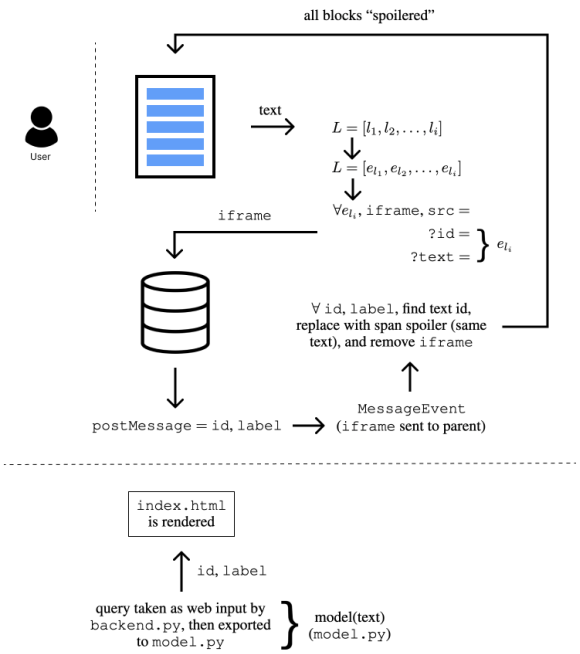

Figure 5: Workflow for the Chrome web extension

5. `backend.py` then renders `index.html` with the (id,label) pair.
6. Once rendered, `index.html` sends a `postMessage` event containing (id,label) out of the `iframe` into the original parent page.
7. The parent page receives this event, confirms the origin, and then parses the `event.data` into the unpacked (id,label) pair.
8. The text elements are then replaced with span classes that are defined in CSS style blocks injected into the head of the page via the `content.js` in the extension.
9. It is important to note that this parent-server exchange is not safe, as it *does not satisfy non-interference*. This is due to the circumvention of the same-origin policy, specifically the protocol (HTTPS → HTTP, HTTP → HTTPS). As a result, keep that loss of secure information in mind when querying.
10. The color of the hover element is dictated by the official awareness colors for each respective label:
    (a) Turquiose for ptsd
    (b) Lilac for EatingDisorders
    (c) Sky Blue for SuicideWatch
    (d) Purple for abusiverelationships
    (e) Black for Others

For each of the respective spoilers, there is a settings menu injected at the top of the page, along with our banner, as shown in Figure 6. The settings allow for toggling between blocking out triggering content or not for each category (i.e. PTSD, eating disorders, suicide, and abusive relationships). An example of the use of these settings is shown in Figure 7.

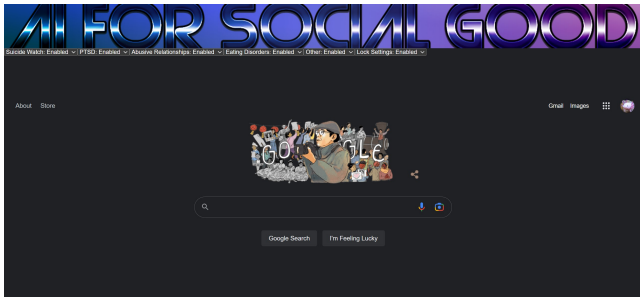

Figure 6: Banner and settings

Evaluation

We were able to test our Chrome web extension in the context of various Reddit webpages. For each of the respective Reddit webpages (e.g. https://www.reddit.com/r/ptsd/ for PTSD), relevant triggering content was shaded, in order to block out text that could be triggering. Screenshots showing our Chrome web extension in practice can be found in Figures 8, 9, 10 in the appendix. These examples show that we were able to deploy our model in a real-world setting. Due to this being a semester-long project, we have not had time to test out the use of this Chrome web extension with actual users yet. In order to evaluate the efficacy of this extension, we would ideally have users try it out and give feedback regarding how they used it and how helpful they found it, and this is also mentioned in the next section. When doing this, it will be important to consider the text used in order to ensure that little to no harm comes to the users testing the extension, especially in situations where the text may be especially triggering.

Limitations and Future Work

One of the main limitations of our work is that due to time constraints of only having one semester to work on this project, we were not able to evaluate other models to further evaluate and compare with the models we have currently implemented. Another limitation was that due to PushshiftAPI maintenance issues, we did not collect as much data as we would have liked. As a result, this likely affected the results of the BERT model just given the amount of data we had, and as a result, this work could be build on further by collecting more data.

Future work in this area would involve addressing the limitations mentioned above. Additionally, switching the sentiment dataset used in BERT-BERT to other datasets, such as a negative emotion dataset, could be incorporated into future work, as it was found that the used sentiment dataset in BERT-BERT did not aid much given that a majority of our data contains negative sentiments. In terms of future work for deployment, the Chrome web extension could use our BERT-BERT model and be tested with actual users in order to evaluate its efficacy and how to improve it. We could

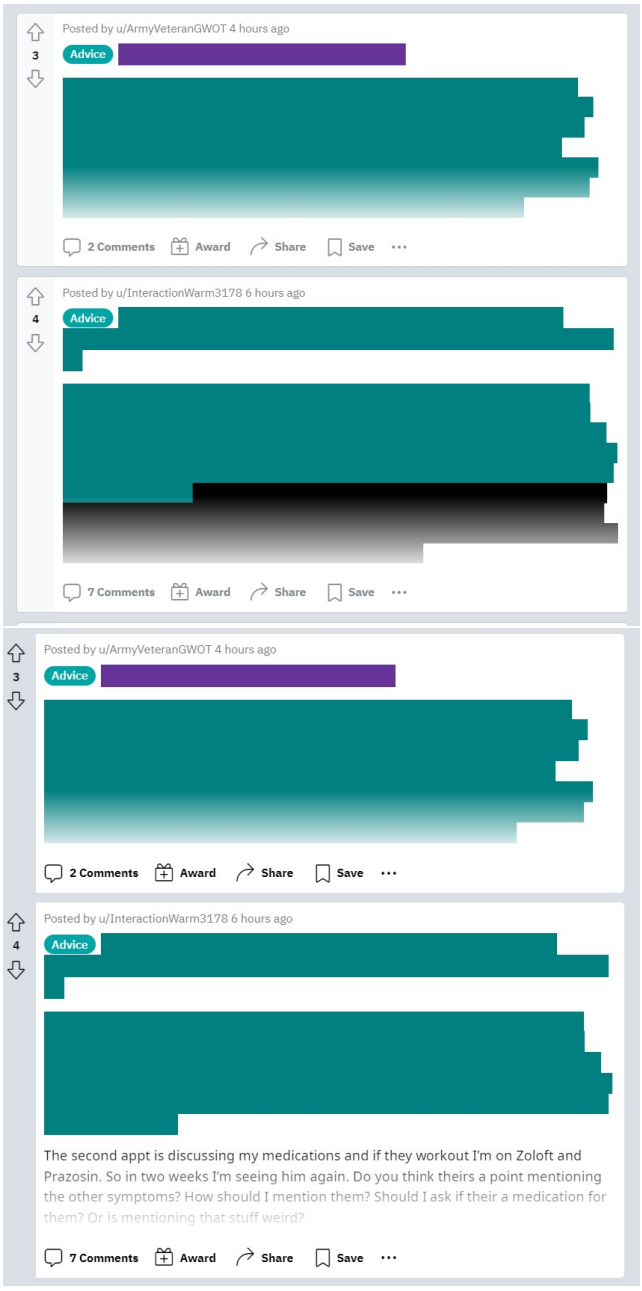

Figure 7: Example of toggling between blocking out certain content and not (top: "Others" enabled, bottom: "Others" disabled)

have various users try out our Chrome web extension and compare their experiences with a baseline of users who did not use our Chrome web extension. Using these results, we could refine our extension to further accomodate the needs of users, especially the needs of the users who would be using such an extension.

Lastly, we would like to discuss the ethics of our work. We sought to only work with public data to ensure that we were not compromising the private data of online users. We also do not make use of any personally identifiable data in implementing the extension; the text used for prediction on a given webpage is solely the text from the webpage and not other gathered data about the user. Any future work in this area should continue to consider the ethics behind design decisions and implementations, in order to respect the private information of all online users.

Conclusion

Browsing the Internet is often a challenge for those suffering from various mental disroders or past traumas, given that the Internet may contain text that is triggering for these people. For the most part, any potentially triggering content will likely show up without any prior warning, which could be harmful for these users. As a result, we proposed various models for identifying such text that could be triggering for those suffering from PTSD, eating disorders, suicidal thoughts and ideations, and trauma from abusive relationships, as these are all areas with little existing research on identifying their triggers within text, yet are still important to address given their wide prevalence among all populations. We scraped data from Reddit and built a SVM model, a BERT model, and BERT-BERT, and we evaluated each of these models. We found that while the F1-scores were relatively consistent between these models, the large language models were able to better capture nuances and scale, which is why they would be preferred in practice. Additionally, we implemented a preliminary Chrome web extension to apply our model in practice in the hopes of eventually reaching online users.

Acknowledgments

We are especially grateful to Professor Fei Fang and Melrose Roderick for their support and feedback on our project. We also acknowledge with thanks the feedback we got throughout the semester from our peers.

References

Aldhyani, T. H.; Alsubari, S. N.; Alshebami, A. S.; Alkahtani, H.; and Ahmed, Z. A. 2022. Detecting and analyzing suicidal ideation on social media using deep learning and machine learning models. *International journal of environmental research and public health*, 19(19): 12635.

Ali, M.; Goetzen, A.; Sapiezynski, P.; Redmiles, E.; and Mislove, A. 2022. All things unequal: Measuring disparity of potentially harmful ads on facebook. In *6th Workshop on Technology and Consumer Protection*.

Ameer, I.; Arif, M.; Sidorov, G.; Gòmez-Adorno, H.; and Gelbukh, A. 2022. Mental illness classification on social media texts using deep learning and transfer learning. *arXiv preprint arXiv:2207.01012*.

Amrit, C.; Paauw, T.; Aly, R.; and Lavric, M. 2016. Using text mining and machine learning for detection of child abuse. *arXiv preprint arXiv:1611.03660*.

Caselli, T.; Basile, V.; Mitrović, J.; and Granitzer, M. 2020. Hatebert: Retraining bert for abusive language detection in english. *arXiv preprint arXiv:2010.12472*.

Counts, S. N.; Manning, J.-L.; and Pless, R. 2018. Characterizing the Visual Social Media Environment of Eating Disorders. In *2018 IEEE Applied Imagery Pattern Recognition Workshop (AIPR)*, 1–6. IEEE.

Dacon, J.; Shomer, H.; Crum-Dacon, S.; and Tang, J. 2022. Detecting Harmful Online Conversational Content towards LGBTQIA+ Individuals. *arXiv preprint arXiv:2207.10032*.

Fodeh, S.; Li, T.; Menczynski, K.; Burgette, T.; Harris, A.; Ilita, G.; Rao, S.; Gemmell, J.; and Raicu, D. 2019. Using machine learning algorithms to detect suicide risk factors on twitter. In *2019 International Conference on Data Mining Workshops (ICDMW)*, 941–948. IEEE.

Kajla, H.; Hooda, J.; Saini, G.; et al. 2020. Classification of online toxic comments using machine learning algorithms. In *2020 4th international conference on intelligent computing and control systems (ICICCS)*, 1119–1123. IEEE.

Karystianis, G.; Cabral, R. C.; Han, S. C.; Poon, J.; and Butler, T. 2021. Utilizing text mining, data linkage and deep learning in police and health records to predict future offenses in family and domestic violence. *Frontiers in digital health*, 3: 602683.

Mount Sinai, N. 2023. Codeine Overdose.

National Center for PTSD. 2022. Trauma Reminders: Triggers. https://www.ptsd.va.gov/understand/what/trauma_triggers.asp. Accessed: 2023-05-05.

Nayak, R.; and Baek, H. S. 2022. Machine Learning for Identifying Abusive Content in Text Data. *Advances in Selected Artificial Intelligence Areas: World Outstanding Women in Artificial Intelligence*, 209–229.

Nelligan, M. 2022. My Case for Using and Respecting Trigger Warnings.

Nolan, H. A.; and Roberts, L. 2022. Medical students' views on the value of trigger warnings in education: A qualitative study. *Medical Education*, 56(8): 834–846.

Ponte, K. 2022. Understanding Mental Illness Triggers.

Raj, C.; Agarwal, A.; Bharathy, G.; Narayan, B.; and Prasad, M. 2021. Cyberbullying detection: hybrid models based on machine learning and natural language processing techniques. *Electronics*, 10(22): 2810.

Roy, T.; McClendon, J.; and Hodges, L. 2018. Analyzing abusive text messages to detect digital dating abuse. In *2018 IEEE International Conference on Healthcare Informatics (ICHI)*, 284–293. IEEE.

Schrading, N.; Alm, C. O.; Ptucha, R.; and Homan, C. 2015. An analysis of domestic abuse discourse on reddit. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, 2577–2583.

Senn, S.; Tlachac, M.; Flores, R.; and Rundensteiner, E. 2022. Ensembles of bert for depression classification. In *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, 4691–4694. IEEE.

Subramani, S.; Vu, H. Q.; and Wang, H. 2017. Intent classification using feature sets for domestic violence discourse on social media. In *2017 4th Asia-Pacific World Congress on Computer Science and Engineering (APWC on CSE)*, 129–136. IEEE.

US Department of Veterans Affairs, N. 2023. How Common is PTSD in Adults?

Villegas, M. P.; Cagnina, L. C.; and Errecalde, M. L. 2022. Anorexia Detection: A Comprehensive Review of Different Methods. In *Computer Science–CACIC 2021: 27th Argentine Congress, CACIC 2021, Salta, Argentina, October 4–8, 2021, Revised Selected Papers*, 170–182. Springer.

Yan, M.; and Luo, X. 2021. BERT-Based Detection of Sexual Harassment in Dialogues. In *2021 5th International Conference on Computer Science and Artificial Intelligence*, 359–364.

Yu, J.; Marujo, L.; Jiang, J.; Karuturi, P.; and Brendel, W. 2018. Improving multi-label emotion classification via sentiment classification with dual attention transfer network. ACL.

Appendix

Chrome Web Extension Examples

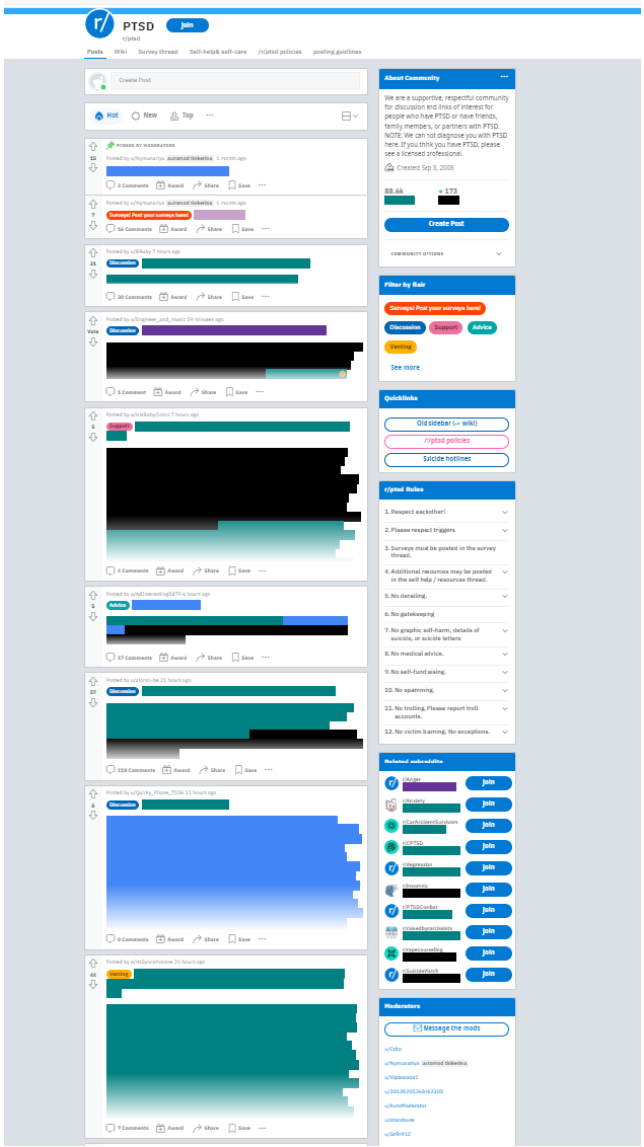

Figure 8: PTSD-triggers blocked out on r/ptsd

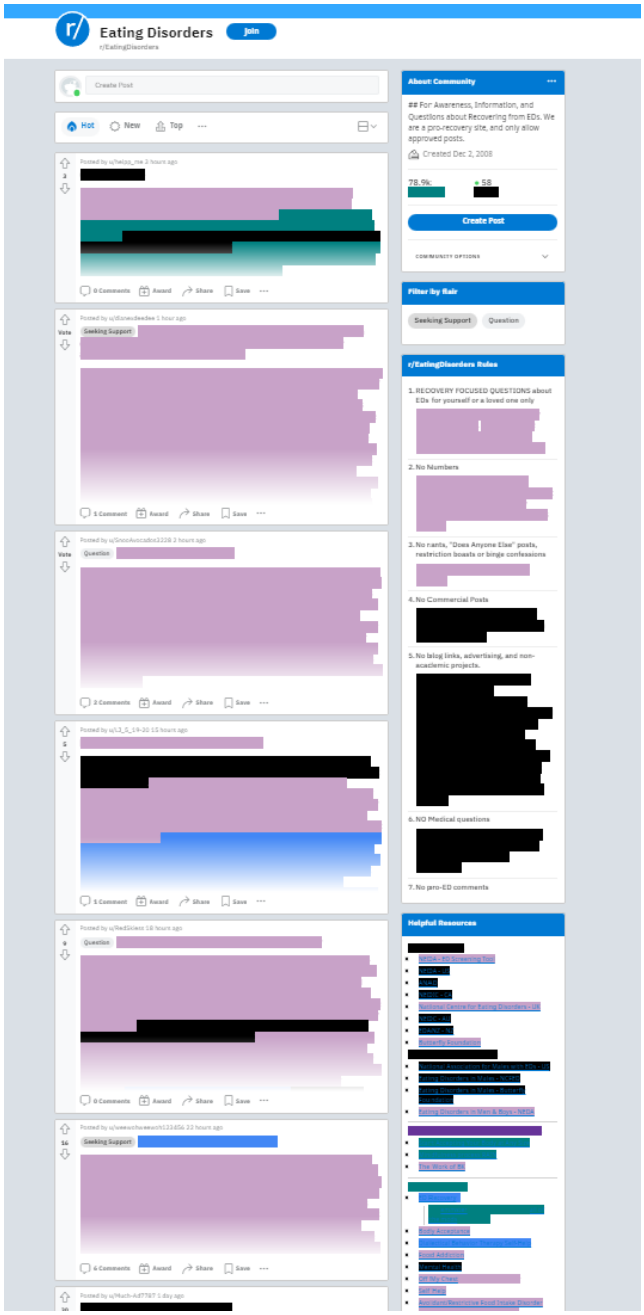

Figure 9: Eating disorders-triggers blocked out on r/EatingDisorders

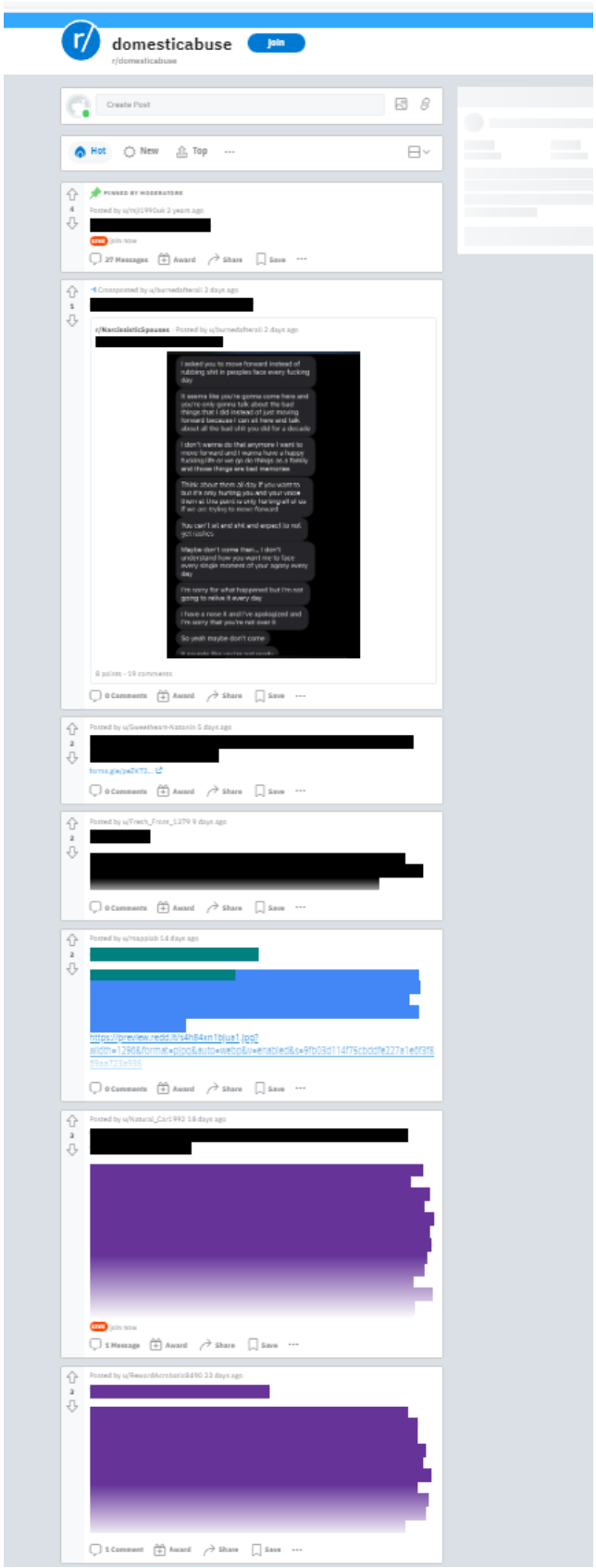

Figure 10: Abusive relationships-triggers blocked out on r/-domesticabuse