Project P-13:

CLASSIFICATION OF 'TRIGGERING' CONTENT ON SOCIAL MEDIA

by

Keelin Sekerka-Bajbus

B00739421

CSCI 4152 Natural Language Processing

Dalhousie University

Halifax, Nova Scotia

December 2021

Abstract

The growth of social media has changed the way people discuss sensitive topics, making it difficult for people with a history of trauma or Post Traumatic Stress Disorder (PTSD) to avoid content that may be triggering. As a result, the need to automatically warn users of sensitive content to suit user preferences has become increasingly important to promote safety and inclusion online. In this project, we collected text-based posts from communities about sensitive topics and mental health topics from the platform Reddit. We then analyzed these posts using traditional and deep learning classification methods to investigate whether we could accurately identify an appropriate trigger warning for the content. We employ four classifiers to establish a baseline for classifying social media posts under one of 9 identified trigger warning labels using our new dataset. Our experiments determine that this task is well captured by more complex models, with the *Linear SVC* and CNN models yielding well-rounded results. We discuss the implications of the models' performance and any difficulties in selecting the correct theme in this complex task. We believe that these results indicate that it would be possible to build a system to automate the labelling of triggering content with simple model architectures to support reduced computational requirements in some cases. This work appears to be the first attempt at automating trigger warnings using natural language processing techniques. We hope that this study brings attention to the topic for further research that will allow for further inclusivity and harm reduction in online spaces.

1 Introduction

With the growth of social media, the way people interact with one another and the variety of content available have changed drastically in recent years. Online platforms, like Reddit, have allowed people to join diverse communities to discuss or engage with content pertaining to sensitive topics readily. For instance, many social media communities exist where users can seek support and share their personal experiences with mental health issues, relationships, and more distressing issues such as domestic violence and suicide. Additionally, it is common for sensitive topics to be shared outside of the specific communities, making avoiding this content difficult for individuals with past trauma, mainly due to social media algorithms and the volume of such content. Currently, content warnings are used on an elective basis on social media and must be done manually by users [1]. As a result, the need to automatically warn users of sensitive content or filter social media feeds to suit user preferences has become increasingly important.

Trigger warnings are cautionary labels for sensitive digital content that may be distressing to viewers to allow users to choose whether to engage with the content and thus, allow them to prepare themselves adequately [2,3]. Such warning labels have largely been popularized by social media and have more recently become common in academia, with university professors increasingly providing trigger warnings for their course materials [2]. While there has been discussion of trigger warnings in contemporary clinical psychology and sociology research [2], there is limited research regarding their use in social media to address online safety or user preferences. Researching the use of trigger warnings in social media would prove helpful in platform design decisions and feature development [3]. For instance, the development of automated content warning systems for social media platforms would aid in reducing direct harms on users' mental health, whether by providing clear labels or improved content filtering, while also encouraging safety and inclusivity in online environments.

In this project, we will attempt to automate the classification of social media posts that may be considered triggering, explicitly pertaining to topics surrounding mental health and other disturbing content (e.g., death, domestic violence). Specifically, we will apply different types of classifiers to social media posts derived from Reddit to identify appropriate content warning labels using text classification techniques. We will identify common content warning labels to define the broad classes for this complex multi-class classification problem. While the text classification of social media posts pertaining to mental health and the classification of sensitive topics, such as hate speech and gender-based violence [4,5], has been explored in recent research, no work appears to exist addressing broad trigger warnings.

2 Related Work

Text-classification of social media posts using natural language processing techniques has been well-researched in recent years, with many works exploring public data from Reddit and Twitter to identify mental health issues through deep learning methods in particular. In 2017, Gkotsis et al. [6] proposed an approach to classify mental health-related content from Reddit using Convolutional Neural Networks (CNNs) for binary and multi-class classification tasks with strong results. They experimented with Feed Forward Neural Networks (FF), Support Vector Machines (SVM), and linear classifiers to determine if a text post was related to mental health and classified it as one of 11 mental health themes. More recently, Kim et al. [7] applied a CNN classification model with word-embeddings to Reddit posts using individual binary classifiers for each mental disorder considered to aid in performance.

Looking to Recurrent Neural Networks (RNNs), Ive et al. [8] work provided a first attempt at automating the classification of mental health topics in Reddit posts using hierarchical RNN architectures with attention, improving upon the work of Gkotsis et al.[6]. They used the document classification architecture proposed by Yang et al. [9] for this experiment, which builds the document representation at word and sentence levels, respectively. Additional work in detecting mental illness on social media has been done using a RoBERTa (Robustly Optimized BERT Pretraining Approach) language model by Murarka et al. [10]. Their results showed that the model could classify mental illness accurately and yielded robust results differentiating mental health-related posts from non-related content.

Deep learning and natural language processing techniques have also been applied to sensitive and disturbing topics unrelated to mental health. Soldevilla and Flores [5] applied BERT language models to successfully perform binary classification with text data from Reddit and Twitter to identify instances of gender-based violence messaging. Hate speech detection using traditional classifiers, deep learning models and BERT has been explored by Modha et al. [4] using social media data from Facebook and Twitter posts.

While the work outlined above has extensively covered the classification of mental illness and select disturbing topics separately, no work appears to be available regarding content warnings or dealing with mental health and disturbing content together in Natural Language Processing. Thus, to our knowledge, this project will be the first attempt at classifying mental health and non-mental health-related social media content as potentially distressing using traditional and deep learning methods.

3 Problem Definition and Methodology

In this project, we will classify social media posts under one of 9 trigger warning labels pertaining to a mixture of mental health and disturbing topics selected. Specifically, we collect social media posts from Reddit communities (subreddits) related to mental health and sensitive topics to investigate whether we can accurately identify an appropriate trigger warning for a specific post

based on its user-generated content automatically. We identify nine broad trigger warning classes [11]– *Abuse*, *Anxiety*, *Death*, *Depression*, *Domestic Violence*, *Dysmorphia*, *Eating Disorder*, *PTSD* (Post Traumatic Stress Disorder), and *Suicide*.

We will explore this classification problem by employing a combination of traditional classifiers and neural network architectures in our experiments to establish a baseline for dealing with content warnings for diverse social media content. In particular, a Linear Support Vector Machine (SVM) classifier, multinomial Naïve Bayes classifier, a simple Feed-Forward Neural Network (FF) and Convolutional Neural Network (CNN) architectures outlined by Gkotsis et al. [6] and Kim et al. [7] are employed using Feature Extraction and Word Embedding techniques**.** The traditional classifiers used Term Frequency Inverse Document Frequency (TF-IDF) to transform the words to document-weighted vectors, while the neural networks used embedded word vectors using a continuous bag-of-words (CBOW) model representation.

4 Experiment Design

We will investigate the problem of classifying social media posts as potentially triggering under an appropriate content warning label using traditional classifiers and neural network architectures.

4.1 Data Collection

We collected social media posts from the following subreddits, as each can be directly associated with a clear primary trigger warning: r/Anxiety, r/Depression, r/PTSD, r/EDAnonymous, r/BodyDysmorphia, r/DomesticViolence, r/Death, r/SuicideWatch, and r/AbusiveRelationships [12]. Based on the subreddit of origin, the posts were labelled under one of 9 trigger warning classes to facilitate training our classifiers. A summary of the collected data is shown in Table 4.1 below. Data was scraped from Reddit using the *Pushift* API [13], which allowed us to collect post titles, post text, and the subreddit of origin while filtering unnecessary fields and metadata. Post authors were not kept, anonymizing the dataset. Prior to data cleaning and data pre-processing, a total of 701,299 posts were collected for the dataset. Data was initially cleaned by dropping any duplicate posts, null posts, or posts that were removed or deleted by the original poster or subreddit, reducing the dataset to 377,134 posts.

4.2 Data Pre-Processing

The procedure for data pre-processing was conducted in several steps. After collecting data and dropping duplicate or deleted posts, any Unicode emoji characters were removed from post titles and body text. Next, the post titles and body text were concatenated to form a single text string for each post. We then made this string lowercase and removed any hyperlinks embedded in the posts' text. Using the *NLTK* API [14], the text was tokenized into words and English stop words were removed from the corpus since these words are frequently used and do not provide additional

meaningful features to the data. Finally, additional punctuation, whitespaces, and non-letter characters were removed from the tokenized text.

The distribution of dataset classes upon completion of dataset cleaning and pre-processing is shown in Table 4.2 below. We note that the dataset is unbalanced, with a large disparity between the smallest classes due to constraints in the available data from some subreddits. The fully processed dataset contains 363,641 observations.

Table 4.1. Summary of collected data from Reddit [12]

Subreddit	Trigger Class Label	Description
r/Anxiety	Anxiety	Discussion and support for sufferers and loved ones of any anxiety disorder.
r/Depression	Depression	Peer support for anyone struggling with a depressive disorder.
r/PTSD	PTSD	We are a supportive, respectful community for discussion and links of interest for people who have PTSD or have friends, family members, or partners with PTSD.
r/EDAnonymous	Eating Disorder	A public subreddit for discussing the struggles of having an eating disorder. Much like an Alcoholics Anonymous or Narcotics Anonymous group, we offer emotional support and harm reduction but no encouragement of furthering ED behaviors. This subreddit is not officially associated with the support group Eating Disorders Anonymous. We are not exclusive to or trying to "force" recovery on anyone.
r/BodyDysmorphia	Dysmorphia	Discussions and support on Body Dysmorphic Disorder, a type of obsessive-compulsive disorder that focuses on the body.
r/DomesticViolence	Domestic Violence	Information and support for victims, survivors, their friends and family.
r/Death	Death	Welcome to r/Death, where death and dying are open for discussion.
r/SuicideWatch	Suicide	Peer support for anyone struggling with suicidal thoughts.
r/AbusiveRelationships	Abuse	For everyone (male and female) who has ever been in an abusive relationship or is currently in one. This is a place for people to vent, share their stories and offer support to others in similar situations.

Table 4.2. Trigger Warnings Dataset Class Distribution

Class	Number of Posts
Eating Disorder	73,381
Anxiety	71,917
Depression	69,463
Suicide	66,509
PTSD	29,965
Abuse	19,144
Dysmorphia	17,904
Domestic Violence	9,329
Death	6,029

4.3 Model Implementation Details

In exploring this classification task, we trained four different classifiers using traditional methods and neural network architectures. The first two classifiers were implemented using the *Scikit-learn* library [15], namely using the *LinearSVC* model as our SVM-based classifier and the *MultinomialNB* model as our Naïve Bayes classifier. For these models, feature extraction using *TF-IDF Vectorizer* from *Scikit-learn* to convert words into n-dimensional, document-weighted vectors was used as input to the models. The *TF-IDF vectorizer* was set to keep a maximum of 5000 features by term frequency in the corpus. Training was limited to 1000 iterations.

The Neural network architectures were implemented using the *Keras* toolkit [16] following the designs outlined by Gkotsis et al. [6] for the Feed Forward (FF) model, and the CNN model proposed by Kim et al. [7] in their works regarding the classification of mental illnesses. For both networks, we used the word2vec API from the *Gensim* toolkit [17] to generate word vectors that were pre-trained with the corpus using a CBOW model, with the window size set to 5 as outlined by Kim et al. [7]. An embedding matrix was generated for the word vectors of dimension 300, generated by the word2vec model's vocabulary extracted from the corpus. This embedding matrix was used as the input weight to the Embedding layers of both neural networks.

The FF model was implemented to follow a similar architecture to that proposed by Gkotsis et al. [6]. The model is made up of an Embedding layer with word vector dimension set to 300 and using the word2vec generated embedding matrix as the weight. This input layer is followed by a Flatten layer and a Dropout layer, set to 0.25 to limit overfitting. Then a fully connected Dense layer with 64 nodes and ReLu activation function is used before the final output layer, consisting of 9 nodes (as needed for the classification) with Sigmoid activation.

The CNN model was implemented following the architecture to proposed by Kim et al. [7]. This architecture used the same Embedding layer as the FF network as input, followed by a single dimensional Convolution layer with 128 nodes and a filter size 5 and a single dimensional MaxPooling layer with 128 nodes. These layers are followed by a Flatten layer and a Dropout

layer set to 0.25 to account for overfitting issues. Finally, two fully connected Dense layers, with 128 nodes and ReLu activation and 9 nodes with Sigmoid activation respectively make up the output layers.

The dataset was divided into stratified training and testing sets using an 80-20 split, using the same randomized seed to ensure all models received the same input training and testing data before feature extractions and word embeddings procedures were performed. In training the neural networks, a maximum of 100 epochs was set with early stopping set for 5 epochs with no changes to the objective function (categorical cross entropy), as recommended by Gkotsis et al. [6]. Batch size was set to 64 and the Adam optimizer was used with a learning rate of 0.001 and an epsilon value of 1e-07.

4.4 Evaluation

To evaluate model performance, the following metrics were used as is standard in a classification task:

$$Precision = \frac{True\ Positives}{(True\ Positives\ +\ False\ Positives)}$$

$$Recall = \frac{True\ Positives}{(True\ Positives\ +\ False\ Negatives)}$$

$$F1 - Score = 2 \cdot \frac{Precision\ \cdot\ Recall}{Precision\ +\ Recall}$$

$$Accuracy = \frac{True\ Positives\ +\ True\ Negatives}{True\ Positives\ +\ False\ Negatives\ +\ True\ Negatives\ +\ False\ Positives}$$

We will consider the overall model performance and closely examine the class performances for each model.

5 Results and Discussion

As described above, we conducted experiments using four different classification models. Table 5.1 summarizes the overall performance of each model. From these results, we see that the best performing models were the Linear SVC and the CNN, with accuracies of 76.43% and 69.91% respectively. We see that both models are reasonably well-balanced and are able to effectively capture the complexity of the classification task, as shown from the F1-scores of 0.7630 and 0.6998 respectively. These results are satisfactory, and we note that the CNN architecture proposed by Kim et al. [7] performs similarly in our experiments to their published results in classifying mental

illnesses. We also present the class performances for each model in Table 5.2, Table 5.3, Table 5.4, Table 5.5 below. There are several notable trends that appear from these results.

Table 5.1. Evaluation results for overall model performance

Model	Accuracy	Precision	Recall	F1-Score
Linear SVC	0.7643	0.7631	0.7643	0.7630
Multinomial NB	0.7018	0.7104	0.7018	0.6994
FF NN	0.6399	0.6413	0.6399	0.6392
CNN	0.6991	0.7056	0.6991	0.6998

First, we see from the results for the Linear SVC model in Table 5.2, that the best performing classes are *Eating Disorder*, *Anxiety* and *Dysmorphia*. We see that the model can accurately classify *Dysmorphia* with high precision, despite being the third smallest class with a significant difference in size compared to *Eating Disorder* and *Anxiety* (the two largest classes). The worst class performances come from *Domestic Violence*, which can in part be attributed to the class-size disparity in the dataset. Upon examining the model's confusion matrix in Figure 5.1, the results indicate many *Domestic Violence* posts are misclassified as *Abuse*, while the converse is also true to a lesser degree. This result can be attributed to the fact that domestic violence is often considered a specific type of abuse, so the language used in these posts would have significant overlap between the classes. Additionally, it is possible there are posts labeled *Abuse* that explicitly discuss domestic violence that may be contributing to the classifier's confusion. Reviewing class labels manually to amend the dataset would prove helpful in model performance in this case. We also note from the confusion matrix, a high rate of misclassification between *Depression* and *Suicide*, which we can attribute to the fact that contemplation of suicide is a common symptom of depression. As such, many posts discussing suicidal urges may be contributing to this decrease in performance and may require class review. These trends are also evident in the other models, with some variability.

Table 5.2. Multiclass classification evaluation results using a Linear SVC.

Class	Precision	Recall	F1-Score
Abuse	0.69	0.69	0.69
Anxiety	0.83	0.86	0.84
Death	0.72	0.63	0.67
Depression	0.63	0.61	0.62
Domestic Violence	0.64	0.48	0.55
Dysmorphia	0.85	0.81	0.83
Eating Disorder	0.90	0.93	0.92
PTSD	0.82	0.76	0.79
Suicide	0.66	0.70	0.68
Weighted Average	**0.76**	**0.76**	**0.76**

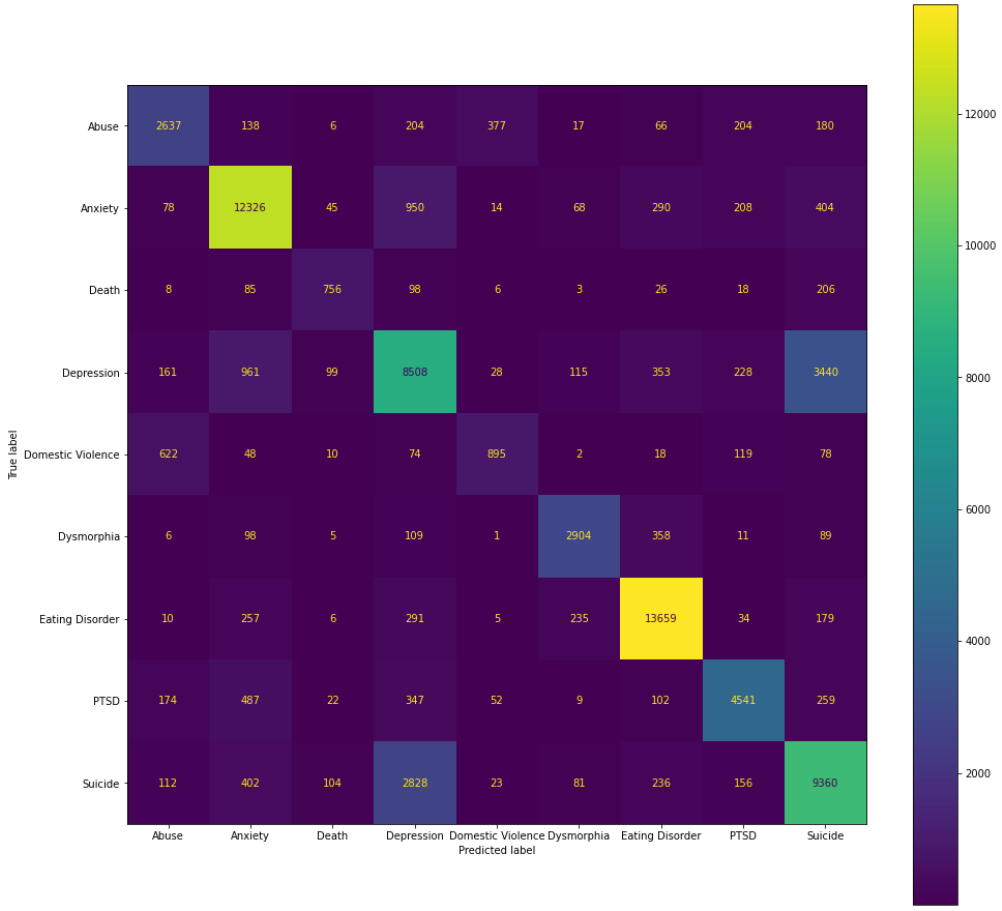

Figure 5.1. Linear SVC confusion matrix

Looking to the Multinomial Naïve Bayes model results in Table 5.3 below, we see a notable increase in the precision in classifying *Death* posts, however the recall and F1-score are the lowest metrics across all classes. From the confusion matrix of the model in Figure 5.2**,** we see the classifier also struggled with differentiating *Depression* posts from *Anxiety* and *Suicide*, with 4951 posts being misclassified under one of these two labels. The confusion between *Anxiety* and *Depression* posts is likely caused by similar language used in these posts, particularly since anxiety and depression symptoms are highly similar and many individuals may suffer from both conditions. Overall, in this model we see significant decreases across all metrics compared to the Linear SVC, which indicates that the classifier may not be as well suited to handle the complexity of this classification task.

Table 5.3. Multiclass classification evaluation results using Multinomial Naïve Bayes.

Class	Precision	Recall	F1-Score
Abuse	0.64	0.64	0.64
Anxiety	0.73	0.79	0.76
Death	0.89	0.24	0.38
Depression	0.53	0.59	0.56
Domestic Violence	0.64	0.36	0.46
Dysmorphia	0.83	0.71	0.76
Eating Disorder	0.89	0.90	0.90
PTSD	0.79	0.56	0.66
Suicide	0.62	0.67	0.65
Weighted Average	**0.71**	**0.70**	**0.70**

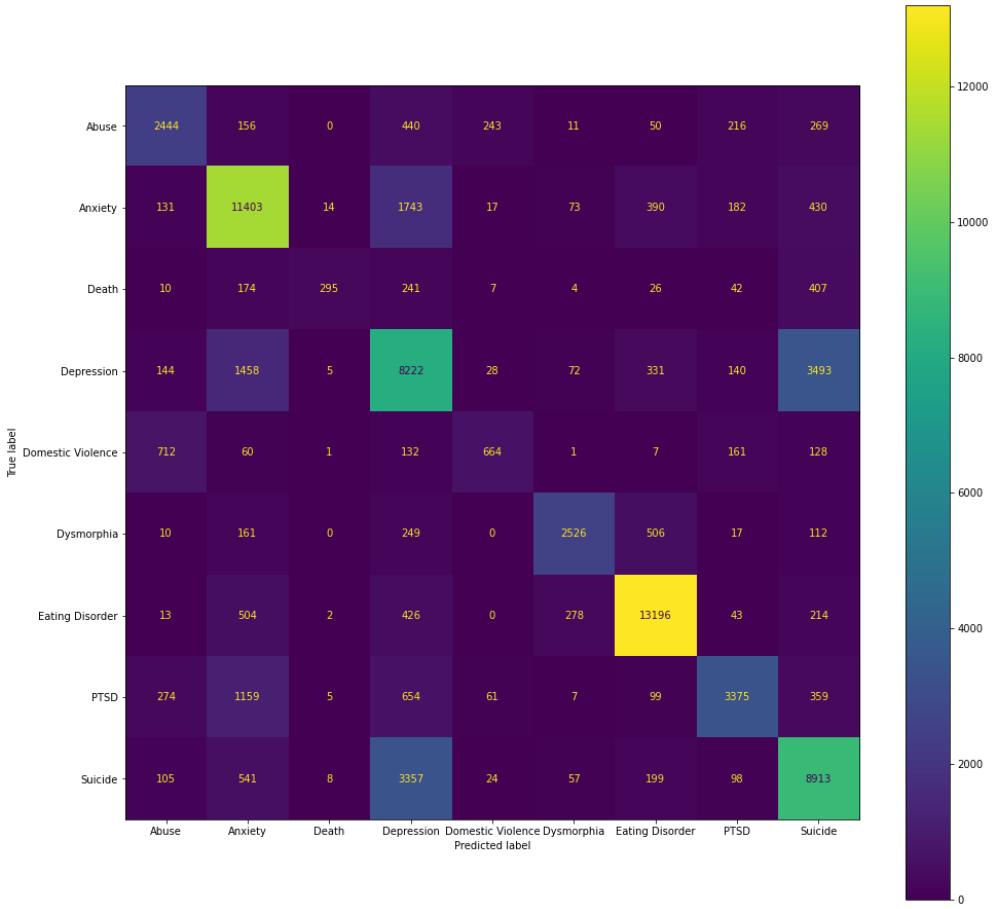

Figure 5.2. Multinomial Naïve Bayes confusion matrix

The worst performing model in our experiments followed the Feed Forward neural network architecture provided by Gkotsis et al [6]. This model was employed to serve as a benchmark for a simple neural network architecture in dealing with our high complexity task of classifying triggering content. From the class results shown in Table 5.4 and the confusion matrix in Figure 5.3, we see significant drops in performance in all classes except *Eating Disorder* and *Anxiety*.

This result indicates that a simple architecture is still able to achieve satisfactory results, and as such for applications where content warnings or filtering for these classes would benefit from the simplistic design and lower computation requirements. Additionally, we note again that *Eating Disorder* and *Anxiety* are the two largest classes in the dataset, so this may indicate that this architecture requires a larger amount of data from other classes to improve the overall performance.

Table 5.4. Multiclass classification evaluation results using a Feed Forward Neural Network.

Class	Precision	Recall	F1-Score
Abuse	0.58	0.47	0.52
Anxiety	0.73	0.74	0.74
Death	0.53	0.44	0.48
Depression	0.50	0.46	0.48
Domestic Violence	0.36	0.40	0.38
Dysmorphia	0.66	0.63	0.65
Eating Disorder	0.86	0.85	0.85
PTSD	0.60	0.54	0.57
Suicide	0.54	0.64	0.59
Weighted Average	**0.64**	**0.64**	**0.64**

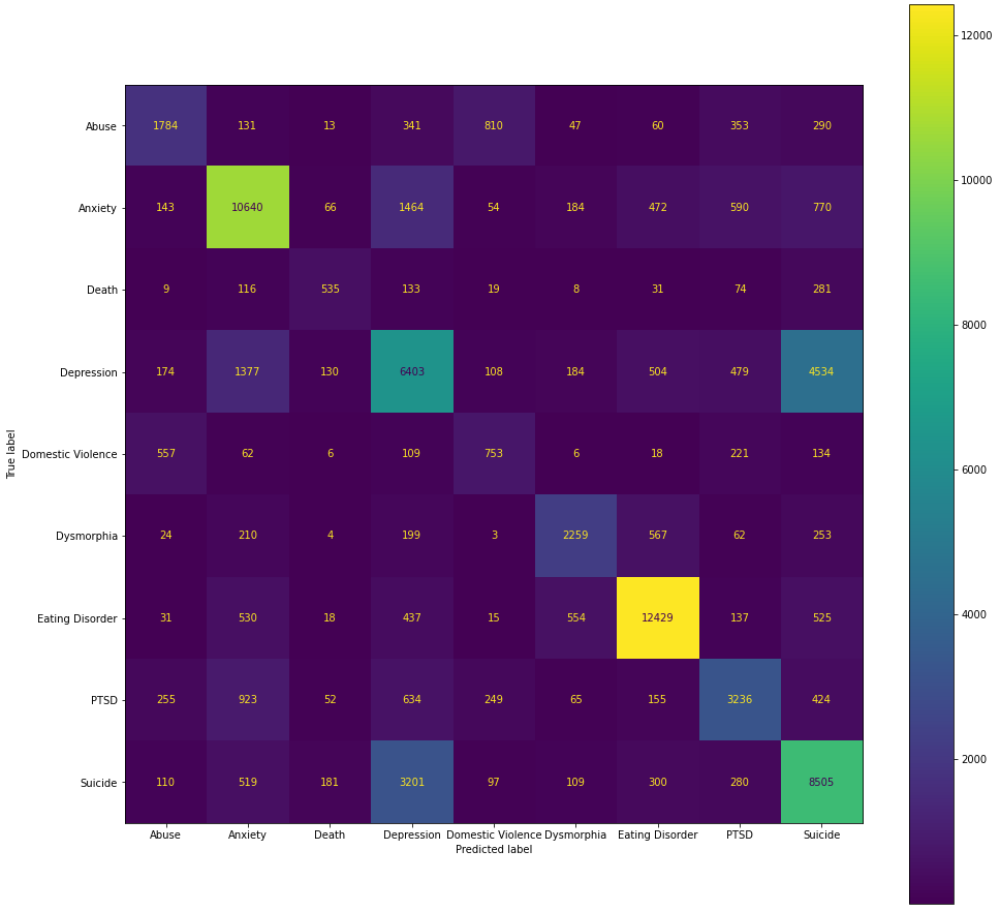

Figure 5.3. Feed Forward Neural Network confusion matrix

The Convolutional Neural Network performed significantly better than the Feed Forward architecture and managed to achieve a more balanced performance by class than the Naïve Bayes classifier as shown by the F1-scores in Table 5.5 below. We see again that many trends from the previous models also apply, as *Eating Disorder*, *Anxiety*, and *Dysmorphia* classes all achieve robust results. We note that *Death* and *PTSD* see significant improvements in their F1-scores compared to the Naïve Bayes and Feed Forward models, again emphasizing the relatively well-roundedness of the CNN model. This model appears to understand the complexity of the data relatively well, however performance among the less populous classes would likely be improved by increasing the size of the dataset, preferably to achieve a more balanced class distribution. This way, the model would be less inclined to favour the more populous classes for classification. For instance, we see in the confusion matrix (Figure 5.4) that many of the smaller classes are misclassified regularly as *Eating Disorder, Anxiety, Depression,* and *Suicide*. This tendency also highlights the fact that classes in this task are not mutually exclusive, and that many posts could be classified under multiple trigger warning labels due to interrelated themes and language.

Table 5.5. Multiclass classification evaluation results using a Convolutional Neural Network.

Class	Precision	Recall	F1-Score
Abuse	0.66	0.58	0.61
Anxiety	0.79	0.80	0.80
Death	0.67	0.53	0.59
Depression	0.58	0.52	0.55
Domestic Violence	0.53	0.48	0.51
Dysmorphia	0.79	0.70	0.74
Eating Disorder	0.87	0.84	0.86
PTSD	0.76	0.69	0.73
Suicide	0.55	0.70	0.61
Weighted Average	**0.71**	**0.70**	**0.70**

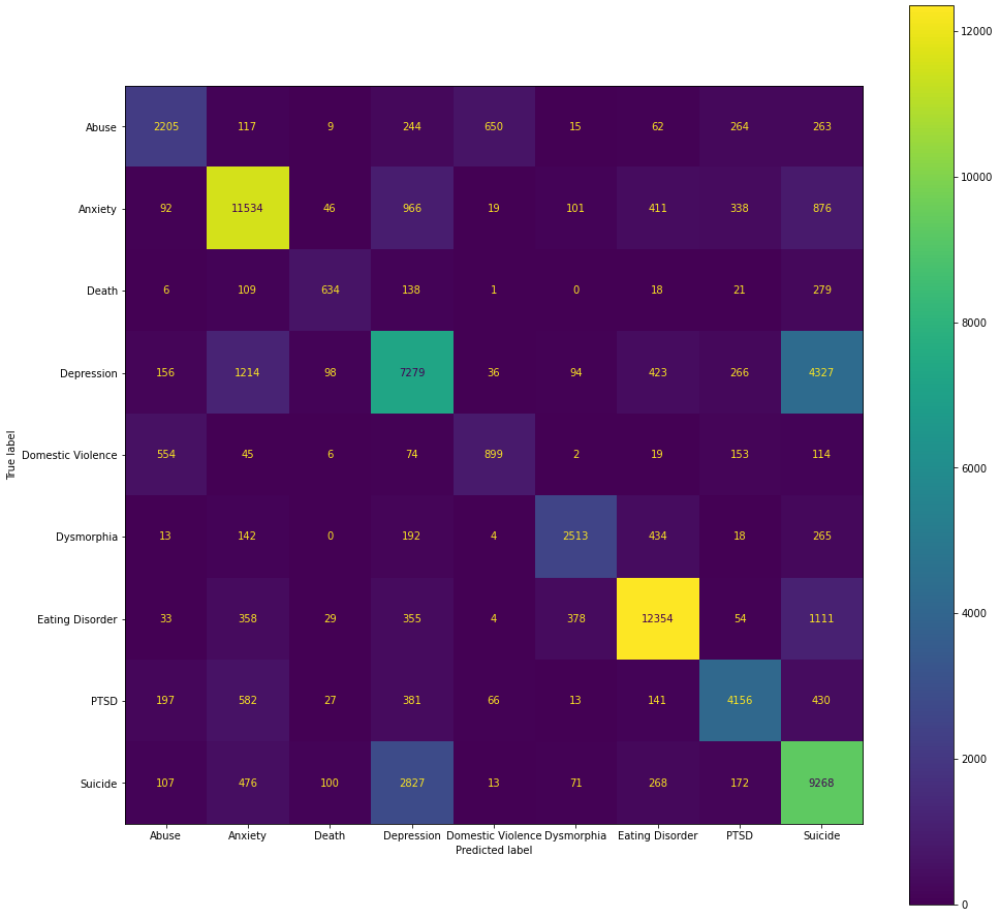

Figure 5.4. Convolutional Neural Network confusion matrix

6 Conclusion

In this project, we have applied four classifiers using traditional methods and deep learning approaches to classify social media posts that may contain sensitive content, requiring a trigger warning. This work, to our knowledge, appears to be the first of its kind regarding trigger warnings. Automating the labelling of social media posts under a trigger warning would improve inclusivity and reduce harm for individuals with a history of trauma in online spaces.

Our experiments have determined that this task is well captured by more complex models, with the *Linear SVC* and CNN models yielding well-rounded results. We also note that the *Eating Disorder* and *Dysmorphia* classes performed consistently well across models, even when using a simplistic model architecture. These results are auspicious and indicate that it would be possible to build a system to automate labelling of eating disorder and body dysmorphia-related content, at minimum, on social media with reduced computational requirements.

We hope to improve our dataset for future work by increasing the size and explicitly addressing the currently unbalanced class distribution to improve performance and reduce bias.

We also suggest a manual review of the dataset class labels to address any improperly labelled data from our initial collection to better suit the primary theme of the posts. Additionally, expanding the dataset to include further trigger warnings would benefit future works, despite requiring the benchmarking process to be repeated. Further investigations applying advanced deep learning techniques, including Recurrent Neural Network architectures with Attention, BERT and RoBERTA language models, would also prove helpful to explore the complexity of this task better and potentially improve upon our work. Finally, we suggest reconsidering this problem as a multi-label classification task, as many classes are not mutually exclusive and have a high rate of theme overlap. As a result, multiple labels (despite the increased complexity of this problem) would benefit system end-users as more than one trigger warning may be relevant to a given social media post. Thus, users would have better information available to them regarding the post's content, allowing them to make informed choices when engaging with content and reducing harm.

References

[1] A. Vingiano, "How The 'Trigger Warning' Took Over The Internet," *Buzzfeed News*, Buzzfeed, 05-May-2014.

[2] M. Sanson, D. Strange, and M. Garry, "Trigger warnings are trivially helpful at reducing negative affect, intrusive thoughts, and avoidance," *Clinical Psychological Science*, vol. 7, no. 4, pp. 778–793, 2019.

[3] O. L. Haimson, J. Buss, Z. Weinger, D. L. Starks, D. Gorrell, and B. S. Baron, "Trans time: Safety, Privacy, and Content Warnings on a Transgender-Specific Social Media Site," *Proceedings of the ACM on Human-Computer Interaction*, vol. 4, no. CSCW2, pp. 1–27, 2020.

[4] S. Modha, P. Majumder, T. Mandl, and C. Mandalia, "Detecting and visualizing hate speech in Social Media: A cyber watchdog for surveillance," *Expert Systems with Applications*, vol. 161, p. 113725, 2020.

[5] I. Soldevilla and N. Flores, "Natural language processing through Bert for identifying gender-based violence messages on social media," *2021 IEEE International Conference on Information Communication and Software Engineering (ICICSE)*, 2021.

[6] G. Gkotsis, A. Oellrich, S. Velupillai, M. Liakata, T. J. Hubbard, R. J. Dobson, and R. Dutta, "Characterisation of mental health conditions in social media using informed Deep Learning," *Scientific Reports*, vol. 7, no. 1, 2017.

[7] J. Kim, J. Lee, E. Park, and J. Han, “A deep learning model for detecting mental illness from user content on social media,” *Scientific Reports*, vol. 10, no. 1, 2020.

[8] J. Ive, G. Gkotsis, R. Dutta, R. Stewart, and S. Velupillai, “Hierarchical neural model with attention mechanisms for the classification of social media text related to mental health,” *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, 2018.

[9] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy, “Hierarchical Attention Networks for document classification,” *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Jun. 2016.

[10] Ankit Murarka, Balaji Radhakrishnan, and Sushma Ravichandran. 2020. Detection and Classification of mental illnesses on social media using RoBERTa. arXiv Prepr. arXiv2011.11226 (2020).

[11] “An Introduction to Content Warnings and Trigger Warnings,” *Univeristy of Michigan Inclusive Teaching LSA*. [Online]. Available: https://sites.lsa.umich.edu/inclusive-teaching-sandbox/wp-content/uploads/sites/853/2021/02/An-Introduction-to-Content-Warnings-and-Trigger-Warnings-Draft.pdf.

[12] “Reddit,” *reddit*. [Online]. Available: https://www.reddit.com/.

[13] *Pushshift.io.* (2019). Pushshift.io. Available: https://pushshift.io/

[14] *Natural Language Toolkit*. NLTK.Available: https://www.nltk.org/.

[15] *Scikit-Learn*. (2021). Scikit-Learn. Available: https://scikit-learn.org/.

[16] *Keras*. (2021). Tensorflow. Available: https://keras.io/.

[17] *Gensim: Topic modelling for humans*. (2021). Available: https://radimrehurek.com/gensim/.