CLASSIFICATION OF ‘TRIGGERING’ CONTENT ON SOCIAL MEDIA

Keelin Sekerka-Bajbus B00739421

P-13

CSCI 4152 Natural Language Processing

What is a Trigger Warning?

- Trigger warnings are a cautionary label for sensitive digital content that may be distressing to viewers.
- Viewers can choose whether to engage with the content.
- Generally, employed to reduce distress for individuals with a history of trauma or Post Traumatic Stress Disorder (PTSD). [1].

The Problem

- People are increasingly discussing sensitive topics on social media in sub-communities, but also in more general spaces.
- The vast quantity of content online makes it difficult to avoid triggering content.
- Trigger warnings are elective and must be done manually by users.

The Project

- Automate the classification of social media posts from Reddit under a trigger warning label using natural language processing techniques.
- Employed 4 models using traditional classifiers and deep learning.
- There is **no existing published work** regarding this topic, so we establish baseline for dealing with content warnings.
- Related works in mental illness classification and detecting disturbing content helped shape our investigation, particularly work by Gkotsis et al. [2].

9 Classes

Abuse

Anxiety

Death

Depression

Domestic Violence

Dysmorphia

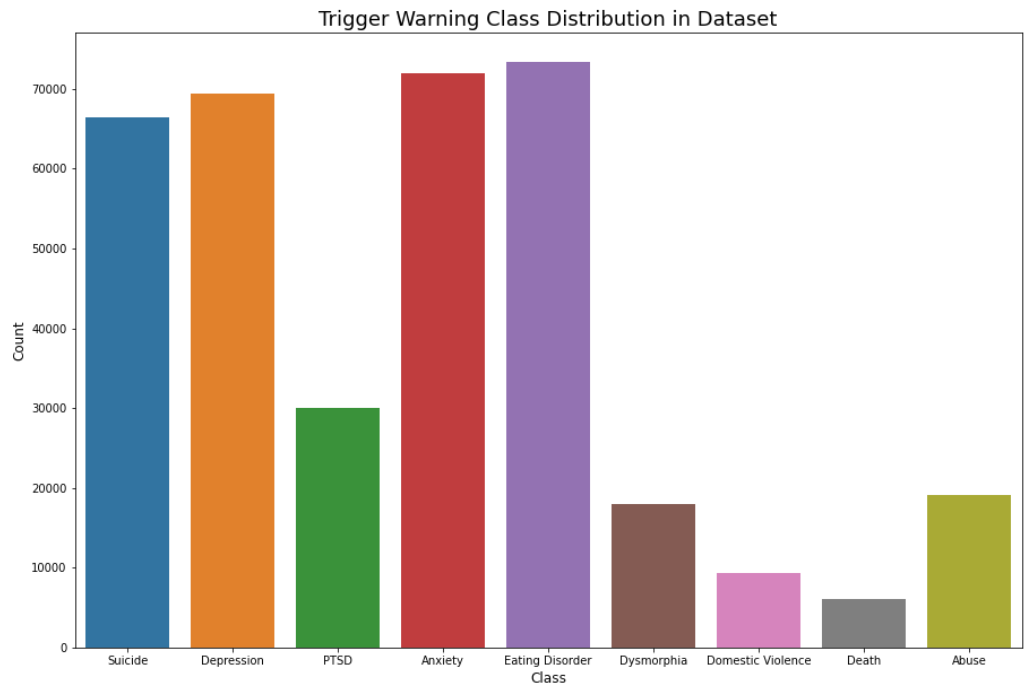
Eating Disorders

PTSD

Suicide

The Dataset

- Collected a dataset of **377,134** posts from Reddit communities (subreddits) related to classes.
- Stratified 80–20 split for training and testing



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**

Traditional Classifiers

Linear SVC

Multinomial Naïve Bayes

- Implemented using Sci-kit learn, including TF-IDF vectorizer to facilitate feature extraction [3]
- Converted words into document-weighted vectors with maximum of 5000 features.

Neural Networks

- Simple Feed Forward (FF) architecture designed by Gkotsis et al. [2].
- Convolutional (CNN) architecture proposed by Kim et al. [4].
- Word2Vec from *Gensim* [5] to create word embeddings with 300-dim vectors as input to both networks.

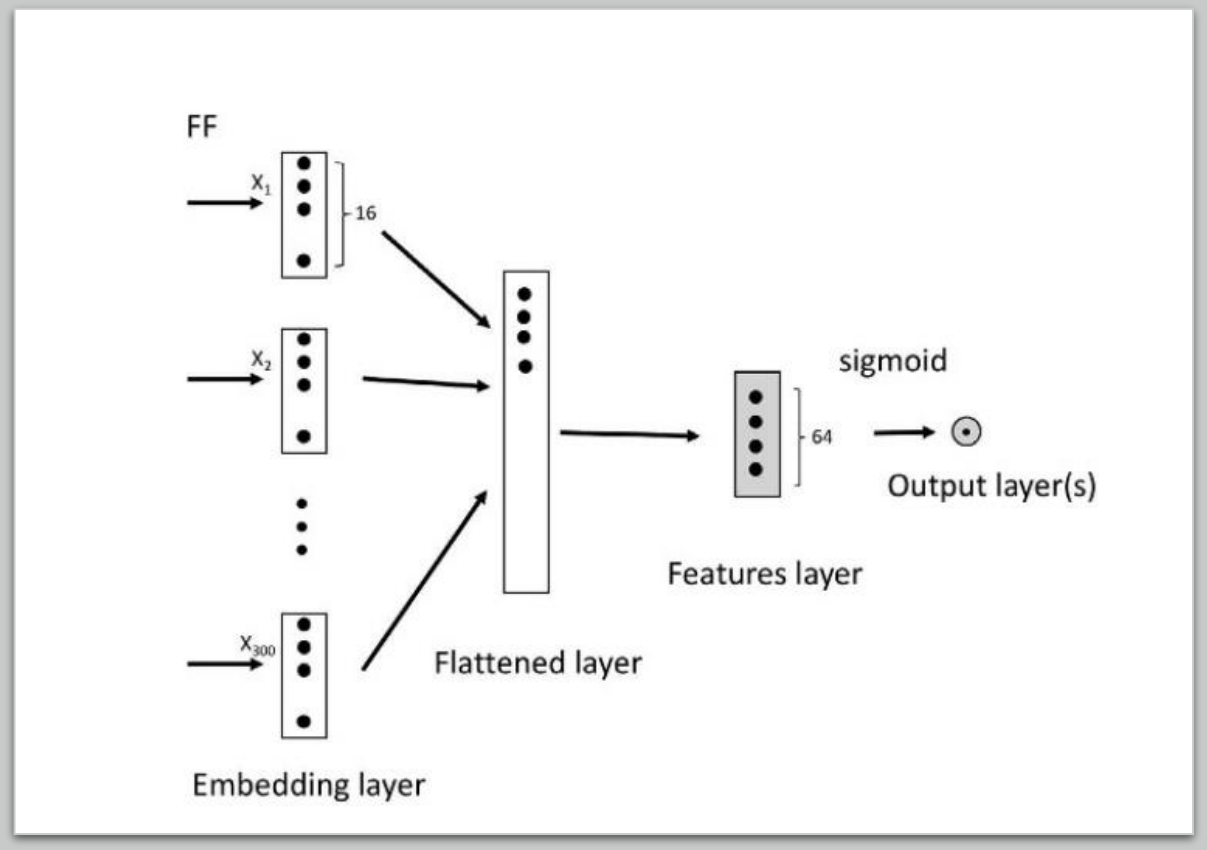


Figure 1. *Architecture for Feed Forward approach.*
Source: Adapted from: [2]

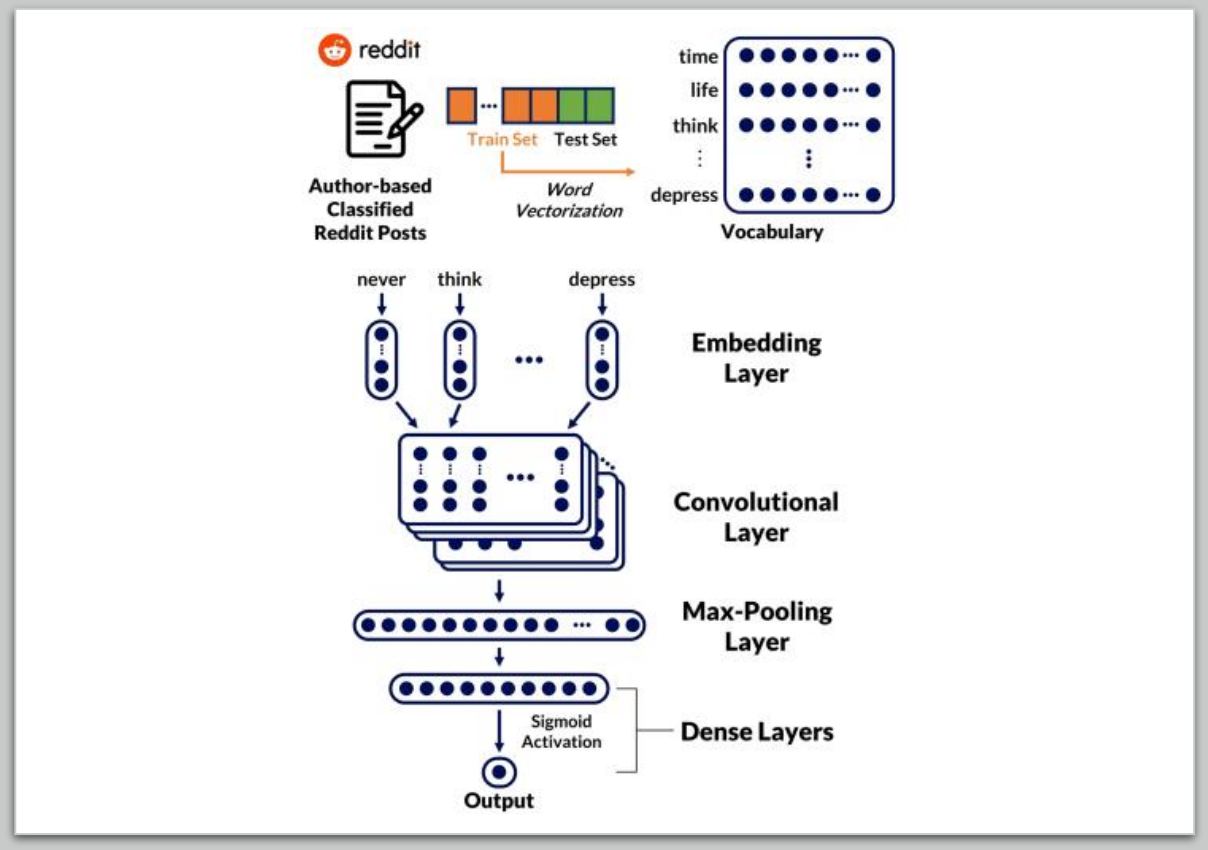


Figure 2. *Architecture for CNN approach.*
Source: Adapted from: [4]

Results

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

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

CNN

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

Key Findings

- All models struggled with distinguishing:
  - *Depression* and *Suicide* posts
  - *Depression* and *Anxiety* posts
  - *Abuse* and *Domestic Violence* posts
- Unbalanced classes fueling model bias.
- Very promising results for *Eating Disorder* and *Dysmorphia* classes, even with simplistic models.
- This problem is well captured by more complex models.

Recommendations for Future Work

- Increase size of dataset, achieve more class-balanced distribution to improve performance.
- Manual review of class labels to improve dataset.
- Explore advanced deep learning techniques.
  - Recurrent Neural Networks with Attention
  - BERT and RoBERTa language models
- Reconsider this problem as a multi-label classification task
  - Many classes are not mutually exclusive, high rate of theme overlap.

References

[1] 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.

[2] 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.

[3] “Scikit-Learn,” *scikit-learn*. [Online]. Available: https://scikit-learn.org/. [Accessed: 05-Dec-2021].

[4] 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.

[5] *Gensim: Topic modelling for humans*. (2021). [Online]. Available: https://radimrehurek.com/genism/.