CLASSIFICATION OF 'TRIGGERING' CONTENT OF SOCIAL MEDIA

by

Keelin Sekerka-Bajbus

B00739421

CSCI 4152 Natural Language Processing

Dalhousie University

Halifax, Nova Scotia

November 2021

1 Problem Statement

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 [1,2]. 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 [1]. While there has been discussion of trigger warnings in contemporary clinical psychology and sociology research [1], 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 [2]. 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., violence, death, sexualized 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. 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 [3,4], has been explored in recent research, no work appears to exist addressing broad trigger warnings. We will identify common content warning labels to define the broad classes for this complex multi-class classification problem.

2 Possible Approaches

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. [5] 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. [6] 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. [7] work provided a first attempt at automating the classification of mental health topics in Reddit posts using hierarchical RNN architectures with attention. They used the

document classification architecture proposed by Yang et al. [8] 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. [9]. 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 [4] 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. [3] 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. 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 Project Plan

This project will be completed in four phases: dataset building, data-preprocessing and model building, the experiment, and the final report. In terms of timelines, the first three phases will be completed by November 30th at the latest to provide time to complete the final report and prepare the presentation by December 7th. The dataset building phase will be completed within the first week of November, using Pushshift API [10] tools to collect text posts from the social media platform Reddit. These posts will be collected from sub-communities that pertain to the topics that will make up the classes for classification. The data will be cleaned such that no identifying information and unnecessary data is removed. The next phase will include data-preprocessing and data augmentation as necessary, for instance, selecting the text representation and using techniques for augmentation as necessary. We will also begin preparing for model selection at this stage and anticipate this stage to be complete within one week of Phase one's completion. We intend to use multiple classifiers to perform the experiment, including traditional models like SVM to benchmark performance against deep learning models (e.g. CNNs, BERT models). The experiment phase will follow, with its methodologies being revisited as necessary before completing the final report.

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] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

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

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