2023

Determining Child Sexual Abuse Posts based on Artificial Intelligence

Susan McKeever
Technological University Dublin, susan.mckeever@tudublin.ie

Christina Thorpe
Technological University Dublin, christina.thorpe@tudublin.ie

Vuong Ngo
Technological University Dublin, vuong.ngo@tudublin.ie

Recommended Citation

Susan Mckeever, Christina Thorpe and Vuong M. Ngo. 2023. Determining Child Sexual Abuse Posts based on Artificial Intelligence. In the 2023 International Society for the Prevention of Child Abuse & Neglect Congress (ISPCAN-2023), Edinburgh, Scotland, UK, September 24-27, 2023, DOI: 10.21427/S3GQ-3536

Funder: the Safe Online Initiative of End Violence and the Tech Coalition

Determining Child Sexual Abuse Posts based on Artificial Intelligence

Susan Mckeever*
School of Computer Science,
Technological University
Dublin, Ireland
Susan.Mckeever@tudublin.ie

Christina Thorpe
Informatics and Cybersecurity,
Technological University
Dublin, Ireland
Christina.Thorpe@tudublin.ie

Vuong M. Ngo
School of Computer Science,
Technological University
Dublin, Ireland
Vuong.Ngo@tudublin.ie

Keywords: *CSAM, CSEM, post content, machine learning, Dark Web*

1. Objectives

The volume of child sexual abuse materials (CSAM) created and shared daily both surface web platforms such as Twitter and dark web forums is very high ([1]). Based on volume, it is not viable for human experts to intercept or identify CSAM manually. However, automatically detecting and analysing child sexual abusive language in online text is challenging and time-intensive, mostly due to the variety of data formats and privacy constraints of hosting platforms. We propose a CSAM detection intelligence algorithm based on natural language processing and machine learning techniques ([2]). Our CSAM detection model is not only used to remove CSAM on online platforms, but can also help determine perpetrator behaviours, provide evidences, and extract new knowledge for hotlines, child agencies, education programs and policy makers.

2. Method

a. Labelled Dataset: Our first step is to create a labelled dataset that can be used for training or fine-tuning our classifier. The labelled dataset used for our study was collected and supplied by Web-IQ company[1], which provided us with over 352,000 posts from 8 dark web forums in 2022, of which approximately 221,000 were in English. Using a dictionary of 12,628 sexual abuse phrases extracted from THORN project[2] and Web-IQ dark web forums, we were able to detect about 177,000 English posts with no sexual abuse phrase and about 44,000 English posts with at least one sexual abuse phrase, which provides us with a high level grouping of posts, but with refinement required to allow for CSAM posts that does not contain any sexual abuse

[1] https://web-iq.com/
[2] https://www.thorn.org/

phrases, and vice versa. From the group of 177,000 posts, experts randomly selected 2,000 non-CSAM posts and 500 CSAM posts. From the group of 44,000 posts, experts randomly selected 2,000 CSAM posts and 100 non-CSAM posts. Ultimately, our manually labelled dataset contains 4,600 posts from the dark web, including 2,500 CSAM posts and 2,100 non-CSAM posts.

b. CSAM Classification Algorithm: The algorithm involves tokenizing the natural language post contents and transforming them into vector representation. Four of the most popular supervised learning methods in text classification were implemented to determine the most suitable method for classifying CSAM content. These methods are Naïve Bayes (NB), Support Vector Machine (SVM), Long Short-Term Memory (LSTM) and Bidirectional Encoder Representations from Transformers (BERT). The execution times (i.e., training time and prediction time) and the classification performance metrics (i.e., precision, recall and accuracy) were evaluated for each combination of methods using the 5-fold cross-validation methodology to avoid overfitting. This approach helps ensure that the resulting CSAM classification algorithm is both accurate and efficient.

3. Results

Average execution time and binary classification performance of the algorithms combining with NB, SVM, LSTM and BERT. The measures are derived from four categories in the confusion matrix: true positive (TP), true negative (TN), false positive (FP), and false negative (FN). "Positive" means that the posts were predicted as CSAM posts and "true" means that the posts were accurately predicted.

- NB: The training time and prediction time were 0.5 and 0.001 seconds. The TN was 280, FP was 140, FN was 55 and TP was 445. The precision was 76.1%, recall was 89% and accuracy was 78.8%.
- SVM: The training time and prediction time were 1.8 and 0.27 seconds. The TN was 385, FP was 35, FN was 79 and TP was 421. The precision was 92.3%, recall was 84.2% and accuracy was 87.6%.
- LSTM: The training time and prediction time were 32.5 and 1.01 seconds. The TN was 374, FP was 46, FN was 72 and TP was 428. The precision was 90.2%, recall was 85.5% and accuracy was 87.1%.

- BERT: The training time and prediction time 4,261 and 215.3 seconds. The TN was 352, FP was 68, FN was 85 and TP was 415. The precision was 86%, recall was 83% and accuracy was 83.4%.

4. Conclusions

We proposed and implemented a novel algorithm based on machine learning and natural language processing to automatically detect and classify CSAM user post content on the dark web. In the experimental evaluation on the dataset of 4,600 CSAM and non-CSAM posts with 5-fold cross-validation, the combination of SVM method performed the best in terms of classification performance and execution time. The second and third algorithms were those using LSTM and BERT methods, respectively. For the future work, we plan to implement functional APIs and develop a usable web application that allows partners and agencies to identify and analyse incidents of child sexual abuse.

5. Key Takeaways

- There is currently no existing CSAM text classifier. We have proposed, built and evaluated four CSAM text classification models using natural language processing and machine learning techniques.
- Our CSAM classifier may be useful for any user post text-based CSAM detection, either as a pre-trained model for fine tuning or for direct use, depending on the specific task.
- A manually labelled CSAM/non-CSAM dataset was created and available to use for training and/or testing CSAM detection algorithms.

6. Biography

Susan McKeever is a senior lecturer at TU Dublin School of Computer Science. She holds a Bachelor of Engineering (Telecommunications and Micro-Electronic) from Trinity College Dublin, an MSc in IT for Strategic Management (DIT). She received her PhD from UCD in 2011. Her main research areas are in the area of machine learning including deep learning, activity recognition, text mining, evidence theory and the general domain area of data analytics. Current projects include: Determining how to detect and automatically moderate and analyse content in social media – including abusive text and automated video analysis; Examining how best to detect and track the health and well-being of elderly people at home, through the use of activity monitoring. She is also a

collaborator in the CeADAR research group (Centre for Applied Data Analytics). CeADAR work with industry to bring data analytics in to address real world business and future business problems. Prior to joining DIT, Susan worked in IT sector, including Accenture as an IT Consultant, and as a contract project manager.

Christina Thorpe graduated with a B.Sc. in Computer Science from University College Dublin in 2005, and Ph.D. in Computer Science from University College Dublin in 2011. She is currently the Head of Cybersecurity in the Technological University Dublin.

Vuong M. Ngo received a B.E, M.E and Ph.D. degrees in Computer Science from HCMC University of Technology in 2004, 2007 and 2013 respectively. Currently, he is a Senior Researcher in Data Science at School of Computer Science, TU Dublin. Previously, he was a Researcher in Free University of Bozen/Bolzano, University College Dublin and Trinity College Dublin.

Acknowledgments

The paper is a part of the N-Light project which is funded by the Safe Online Initiative of End Violence and the Tech Coalition through the Tech Coalition Safe Online Research Fund (Grant number: 21-EVAC-0008-Technological University Dublin).

Reference

[1]. V.M. Ngo, C. Thorpe, C.N. Dang and S. Mckeever. 2022. Investigation, Detection and Prevention of Online Child Sexual Abuse Materials: A Comprehensive Survey. In *Proceedings of the* 2022 *RIVF International Conference on Computing and Communication Technologies (IEEE-RIVF),* Ho Chi Minh City, Vietnam, 2022, pp. 707-713, IEEE, doi: 10.1109/RIVF55975.2022.10013853

[2]. V.M. Ngo, T.V.T. Duong, T.B.T. Nguyen, P.T. Nguyen and O. Conlan. 2021. An Efficient Classification Algorithm for Traditional Textile Patterns from Different Cultures Based on Structures. In *Journal on Computing and Cultural Heritage* Volume 14, Issue 4, Article 53, pp. 1-22, doi: https://doi.org/10.1145/3465381