Guarding the Guardians: Automated Analysis of Online Child Sexual Abuse –Supplementary Material–

Juanita Puentes[1] Angela Castillo[1] Wilmar Osejo[2] Yuly Calderón[3]
Viviana Quintero[2] Lina Saldarriaga[2] Diana Agudelo[3] Pablo Arbeláez[1]
[1]Center for Research and Formation in Artificial Intelligence, Universidad de los Andes
[2]Aulas en Paz [3]Universidad de los Andes

Supplementary Material

1. Manual Cybercrime Analysis

According to the literature, cybercrime analysts are frequently exposed to traumatic experiences and narratives. Exposure to other people's traumatic experiences can lead to vicarious trauma, secondary traumatic stress, compassion fatigue, burnout, and even symptoms of Post Traumatic Stress Disorder [1]. A U.S. federal law enforcement agency study found that 36% of online CSAM researchers exhibited moderate and high levels of secondary traumatic stress [1].

Thus, a manual review of child sexual exploitation and abuse material can affect analysts' physical and mental health, job performance, interpersonal relationships, health, and overall well-being [2]. Over time, these levels of impairment can impact the accuracy of analysis and, therefore, the prioritization and referral of risk situations to authorities.

Online CSEA cases, on the other hand, have maintained a steady increase over the past few years. In 2020, NCMEC reported a 97.5% increase in "online enticement," a form of online exploitation that includes online grooming. Likewise, an analysis of dark web conversations conducted by CRISP found that the number of conversations between sex offenders to share online grooming strategies increased by 13% between 2019 and 2020. In the same period, NCMEC [3] reported a 63% increase in CSAM reports.

With a problem growing exponentially globally and taking into account the impact on the physical and mental health and well-being of analysts, it is necessary to explore new possibilities to reduce the burden of analysis and increase the processing capacity of reported cases.

2. Additional Experimental Validation

In this section, we present a comprehensive overview of the experimental details conducted to assess the performance of our language model. To achieve this, we conduct two sets of experiments: the first set focuses on evaluating the impact of hyperparameters without data augmentation. In contrast, the second set delves into the effects of data augmentation in conjunction with hyperparameters. Both sets of experiments employ fine-tuning.

In the first set of experiments, we thoroughly investigate the influence of critical hyperparameters on the performance of our fine-tuned language model without employing any data augmentation. The hyperparameters examined encompass the learning rate, batch size, dropout rate, and the number of training epochs. The best hyperparameter configurations for each dimension: Subject, Degree of Criminality, and Damage, are summarized in Table 1.

Hyperparameters	Subject	Degree of Criminality	Damage
Batch Size Train	41	167	54
Batch Size Test	68	39	171
Learning Rate	1.217 E-5	4.634 E-5	5.804 E-5
Epochs	144	116	10
Dropout	0.448	0.218	0.485
mAP	0.382 ± 0.0016	0.397 ± 0.040	0.429 ± 0.018
F-score	0.455 ± 0.001	0.593 ± 0.017	0.553 ± 0.004

Table 1: **Fine-Tuning experimentation details.** We present the hyperparameters used for Subject, Degree of Criminality and Damage dimensions.

In the second set of experiments, we present the impact of data augmentation on the performance of our fine-tuned language model. For this analysis, we carefully identify the optimal hyperparameters specific to the data augmentation configuration. Additionally, we introduced two additional augmentation-specific parameters: the Augmentation Factor (AF) and the Augmentation Deletion Rate (ADR). AF is the multiplier factor used to increase the training dataset size through data augmentation. ADR represents the likelihood of each word being deleted during augmentation. These identified optimal hyperparameter configurations, tailored for data augmentation for each dimension, are summarized in Table 2.

Hyperparameters	Subject	Degree of Criminality	Damage
Batch Size Train	75	221	200
Batch Size Test	212	89	169
Learning Rate	3.569 E-6	8.399 E-6	1.212 E-5
Epochs	140	13	91
Dropout	0.247	0.435	0.498
ADR	0.098	0.061	0.856
AF	4.354	8.77	1.532
mAP	0.386 ± 0.002	0.417 ± 0.032	0.458 ± 0.001
F-score	0.447 ± 0.014	0.598 ± 0.036	0.576 ± 0.019

Table 2: **Data Augmentation experimentation details.** We present the hyperparameters used for Subject, Degree of Criminality and Damage dimensions.

References

[1] Beth E. Molnar, Samantha A. Meeker, Katherine Manners, Lisa Tieszen, Karen Kalergis, Janet E. Fine, Sean Hallinan, Jessica D. Wolfe, and Muriel K. Wells. Vicarious traumatization among child welfare and child protection professionals: A systematic review. *Child Abuse & Neglect*, 110:104679, December 2020.

[2] Françoise Mathieu. *The Compassion Fatigue Workbook*. Routledge, May 2012.

[3] We Protect, CRISP, and PA Consulting. Global threat assessment 2021. working together to end the sexual abuse of children online, 2021.