Contents lists available at ScienceDirect

Digital Investigation

journal homepage: www.elsevier.com/locate/diin

Laying foundations for effective machine learning in law enforcement. Majura – A labelling schema for child exploitation materials

Janis Dalins [a, b, *], Yuriy Tyshetskiy [d], Campbell Wilson [a], Mark J. Carman [a], Douglas Boudry [c]

[a] Monash University, Caulfield, VIC, Australia
[b] Australian Federal Police, Melbourne, VIC, Australia
[c] Australian Federal Police, Barton, ACT, Australia
[d] Data61, CSIRO, Eveleigh, NSW, Australia

ARTICLE INFO

Article history:
Received 31 March 2018
Received in revised form
28 May 2018
Accepted 28 May 2018
Available online xxx

Keywords:
Neural networks
Digital forensics
Child exploitation
Forensic triage
Annotation schema

ABSTRACT

The health impacts of repeated exposure to distressing concepts such as child exploitation materials (CEM, aka ‘child pornography’) have become a major concern to law enforcement agencies and associated entities. Existing methods for ‘flagging’ materials largely rely upon prior knowledge, whilst predictive methods are unreliable, particularly when compared with equivalent tools used for detecting ‘lawful’ pornography. In this paper we detail the design and implementation of a deep-learning based CEM classifier, leveraging existing pornography detection methods to overcome infrastructure and corpora limitations in this field. Specifically, we further existing research through direct access to numerous contemporary, real-world, annotated cases taken from Australian Federal Police holdings, demonstrating the dangers of overfitting due to the influence of individual users' proclivities. We quantify the performance of skin tone analysis in CEM cases, showing it to be of limited use. We assess the performance of our classifier and show it to be sufficient for use in forensic triage and ‘early warning’ of CEM, but of limited efficacy for categorising against existing scales for measuring child abuse severity.

We identify limitations currently faced by researchers and practitioners in this field, whose restricted access to training material is exacerbated by inconsistent and unsuitable annotation schemas. Whilst adequate for their intended use, we show existing schemas to be unsuitable for training machine learning (ML) models, and introduce a new, flexible, objective, and tested annotation schema specifically designed for cross-jurisdictional collaborative use.

This work, combined with a world-first ‘illicit data airlock’ project currently under construction, has the potential to bring a ‘ground truth’ dataset and processing facilities to researchers worldwide without compromising quality, safety, ethics and legality.

Introduction

Reports of increasing workloads, employee ‘burn-out’ and psychological trauma are common across law enforcement and the judiciary, but the stresses and harms associated with exposure to psychologically harmful and offensive materials (typically child exploitation materials (CEM)[1] and violent imagery associated with online radicalisation) are now regarded as having been underestimated - particularly in instances of regular, lower level exposure. Law enforcement organisations such as the Australian Federal Police (AFP) traditionally employ a combination of regular psychological monitoring and mandatory staff rotations as a mitigating strategy, but these reduce skillsets within relevant teams (further exacerbating the problem), and tend to be reactive to persons already experiencing symptoms of harm.

In this paper we introduce the ‘Stonefish’ classifier - a machine learning (ML) tool demonstrating the feasibility of automated classifiers for CEM detection, both as triage tools and ‘early warning’ services for reviewers. This classifier uses supervised learning, an approach requiring high quality training and test data reflective of the ‘real world’ landscape. We assemble and utilise a collection of

* Corresponding author.
E-mail addresses: janis.dalins@monash.edu, janis.dalins@afp.gov.au (J. Dalins), yuriy.tyshetskiy@data61.csiro.au (Y. Tyshetskiy), campbell.wilson@monash.edu (C. Wilson), mark.carman@monash.edu (M.J. Carman), douglas.boudry@afp.gov.au (D. Boudry).

[1] aka ‘Child Pornography’, ‘Child Abuse Materials’, ‘Sexually Exploitative Imagery of Children (SEIC)’.

https://doi.org/10.1016/j.diin.2018.05.004
1742-2876/© 2018 Elsevier Ltd. All rights reserved.

AFP case data for training, and data from an unrelated case for testing. We detail challenges and safeguards implemented as part of the development process, specifically for practitioner welfare.

Furthermore, in response to practitioner complaints of incompatible tools and data, we introduce the Majura schema, a jurisdictionally independent labelling/annotation schema designed for use in developing ML techniques in the field.

Existing work

Existing work relevant to this paper can be broadly split into multiple categories - the impacts of exposure to CEM (and other offensive materials), the broader challenges in Digital Forensics affecting possible solutions, automated discovery of CEM (both in use and experimental), and the research limitations caused by a lack of relevant datasets.

Exposure to CEM

First-hand exposure to traumatic and offensive events is long documented as psychologically harmful. Surveys of police officers in provincial England and New York state (USA) by Brown et al. (1999) and Violanti and Aron (1995), respectively, indicated comparatively high levels of stress in exposure to traumatic events involving children. Both studies pre-date the mainstream emergence of online child sex abuse, but a key point of note appears to be stress associated with dealing with *victims* of crimes such as rape and child abuse being quite high, with police officers seen as potentially “becoming secondary victims” (Brown et al. (1999)) in such cases.

The absence of studies into the effects of exposure to child exploitation by forensic analysts and other persons involved in the investigation/prosecution process was observed by Edelmann (2010), who noted that employers such as the Metropolitan Police provide mandatory counselling to staff routinely exposed to such imagery.

More recently, Powell et al. (2015) conducted a survey of 32 law enforcement personnel across all Australian jurisdictions, specifically recording the reported impacts of exposure to child exploitation materials[2] within internet child exploitation investigations. Critically, the survey included not only sworn police, but also ‘computer analysts’ - a role arguably requiring even more regular and in-depth exposure to materials during the course of digital forensic analysis. Interestingly, some respondents indicated an experience akin to the previously mentioned ‘secondary victimhood’, though contrastingly, some perceived exposure to CEM as less harmful than direct ‘interaction with victims of assault’.[3]

Specific factors were listed by survey respondents as increasing a risk of long-term effects from exposure:

- Perceived resemblances between victims and children known to the reviewer (particularly the reviewer's own children);
- ‘Unexpected’ viewing of child exploitation materials;
- Repeated exposure to specific images or offenders;
- Viewing the progression of an offender from viewer to contact offender[4]; and
- Perhaps unexpectedly, some respondents also reported increased distress from text, as opposed to imagery & multimedia.

An anonymous survey of US law enforcement personnel by Seigfried-Spellar (2017) identified differences in psychological distress between investigators and forensic analysts, with persons conducting both duties in CEM related cases reporting higher levels of traumatic stress than those working single roles. The author hypothesizes this is due to their requirement to both review CEM and interact with victims and offenders, a theory consistent with the “secondary victimhood” identified by Brown et al. (1999). Furthermore, whilst respondents *generally* used healthy coping strategies, those working dual roles “may be more likely to use sedatives …as a coping mechanism.”

Powell et al. (2015) note that due to the large number of variables involved, individual investigators' reactions to CEM exposure are impossible to predict. Viewed together with the general reluctance by police to seek assistance, combined with a low (16%) level of mandatory counseling offered by the respondent's agencies, it appears quite feasible that the extent of exposure related stress and harm remains underreported across law enforcement.

As stated by Powell et al. (2015), “purchase of technological strategies for global reduction in exposure to images is therefore warranted”.

Challenges in digital forensics

In Powell et al. (2014), the aforementioned study's authors also questioned their respondents about the challenges they personally encounter in the field of Digital Forensics. Identified issues particularly of relevance to this article included:

- Access to “image scanning” software - most likely a reference to CETS (refer Table 1) or another cryptographic digest based content recognition system (refer Section Automated CEM Discovery);
- Inadequate staffing, including a lack of relevant digital forensics experience; and
- The need for “complete” examination - courts requiring every relevant item (image/video) to be reviewed and categorised, rather than accepting a representative sample. A respondent quotes a staff member “going through 500,000 images”.

More recently, Franqueira et al. (2017) conducted a targeted survey of Digital Forensic (DF) practitioners worldwide, seeking their comments on challenges in the field of online child exploitation. The survey returned similar results in regard to the stresses and impacts of exposure to such imagery, but the authors’ stronger focus on technical specialists[5] resulted in a differing set of reported challenges:

- Emerging technologies such as automatic age estimation are not ‘translating’ into workable tools for improving practices;
- Stressful working conditions associated with viewing CEM, with recommendations for improving automation to “minimize exposure in the first place”; and
- A need to standardise operations, procedures and legal frameworks globally, necessitating an *”internationally recognised scale of indecency levels and a taxonomy of terms to bridge language and cultural differences”*

The absence of standardisation as a challenge is glaring in Powell et al. (2014), most likely due to the paper's Australian focus. Nine

[2] Referred to as ‘internet child exploitation’ materials within the paper.

[3] It is unclear if this refers to *sexual* or physical assault, given the context).

[4] The *abuser*, as opposed to viewer of abuse.

[5] The authors use ‘DF’ in a broad sense, encompassing first responders, consultants and other roles regularly exposed to the crime type.