

Using Machine Learning Methods to Study Technology-Facilitated Abuse: Evidence from the Analysis of UK Crimestoppers' Text Data

Felix Soldner, Leonie Maria Tanczer, Daniel Hammocks, Isabel Lopez-Neira, Shane D. Johnson

Abstract Quantitative evidence on technology-facilitated abuse (“tech abuse”) in intimate partner violence (IPV) contexts is lacking globally. This shortcoming creates barriers to the development of evidence-based interventions. This chapter draws on a data science-driven research project which aims to generate statistical evidence on the nature and extent of IPV tech abuse in the United Kingdom (UK). Using data from the independent UK charity Crimestoppers (2014-2019), we showcase an automated approach, facilitating Natural Language Processing and machine learning methods, to identify tech abuse cases within large amounts of unstructured text data. The chapter offers both useful insights into the types of tech abuse found within the data, as well as the challenges and benefits computational methodologies provide. The research team has released the code and trained machine learning algorithm along with the publication of this chapter. This hopefully allows other researchers to test, deploy, and further improve the automated approach and could facilitate the analysis of other text datasets to identify tech abuse.

Introduction

Technology-facilitated abuse, so-called “tech abuse”, in intimate partner violence (IPV) contexts describes the misuse of technical systems to harass, monitor, and control victims/survivors. This digitally-enabled mode of violence can take many forms (Dragiewicz et al., 2018; Harris & Woodlock, 2018; Woodlock, 2017). It may include abusive messages or calls, image-based abuse cases such as “revenge porn” (Citron & Franks, 2014; Henry, McGlynn, et al., 2020; McGlynn, Rackley, & Houghton, 2017), as well as means of being traced through phones, trackers, or other GPS- or Internet-enabled devices (Lopez-Neira, Patel, Parkin, Danezis, & Tanczer, 2019; Parkin, Patel, Lopez-Neira, & Tanczer, 2019). Due to the breadth of systems and means to harm victims/survivors through technology, previous publications have described the concept of tech abuse as “big bucket”, ranging from low-tech offences to more technically sophisticated crimes (Tanczer, forthcomingb).

Despite the rising uptake of digital technologies in our day-to-day lives, data on the scale, nature, and extent of this type of abuse is scarce internationally (Henry, Flynn, & Powell, 2020; Tanczer, Lopez-Neira, Parkin, Patel, & Danezis, 2018). Recorded data hint at the scope of the problem. In the UK, these include, for example, Refuge’s (2018) assessment of 920 tech abuse cases, and later statement that 72% of their service users experienced abuse through technology in 2019 (Refuge, 2020); Women’s Aid’s (2018) survey which indicated that 85% of

victims/survivors experience online and offline abuse in parallel; a study by Snook, Chayn, and SafeLives (2017) which found that nearly half of their 209 surveyed victims/survivors (47%) were monitored online by their partner; and most recently the Suzy Lamplugh Trust's evaluation of the UK's National Stalking Helpline figures which showcase that 100% of cases presenting to the Helpline now involve a cyber element. Akin dynamics having been identified in countries such as Australia (Henry, McGlynn, et al., 2020; Powell & Henry, 2019; Powell, Scott, Flynn, & Henry, 2020) and the United States (Messing, Bagwell-Gray, Brown, Kappas, & Durfee, 2020).

Whilst all these assessments are important and form a step in the right direction, the evaluations are opaque. None of these studies collect longitudinal evidence nor discuss the different forms of offences that fall under the broad category of tech abuse. Hence, present quantitative tech abuse studies do not account for the nuances of tech abuse that would enable to differentiate between distinct devices or platforms that are abused, nor their respective severity levels (Tanczer et al., 2018). Instead, they speak of tech abuse as an overarching category and are often primarily concerned with "conventional" cyber risks such as abuse patterns on social media and restrictions to devices such as laptops and phones (Burton, Tanczer, Vasudevan, & Carr, 2021; Slupska & Tanczer, forthcoming).

As the IPV tech abuse risk landscape is steadily transforming, our focus and attention must shift towards the different shapes and shades that tech abuse can adopt (Tanczer, forthcoming). A range of technical innovations including digital payment systems, Internet of Things (IoT) devices, blockchain technologies, or the ominous concept of Artificial Intelligence (AI) are here to stay (see Chapters X and X, this volume). The current chapter consequently outlines our recent attempt to add quantitative figures to the evidence-base on tech abuse. To facilitate this assessment, we analysed a dataset of 500,000 archival crime reports held by the UK charity Crimestoppers to look for cases of tech abuse. While such big datasets are of great value, it is unfeasible to manually inspect each entry to determine if it can be categorised as tech abuse. By utilising Machine Learning (ML), which is often placed under the umbrella term AI, we aimed to devise a system which can facilitate the automated detection of tech abuse within a large corpus of text data. Such a computerised identification mechanism would be beneficial not only for research, but for the practitioner community, including law enforcement and domestic abuse charities. Thus, in this chapter we outline how an automated system can look like, what potential problems can arise in deploying such tools, and how these can be mitigated. We also briefly touch upon the distinct forms of tech abuse we observed within the dataset.

Machine Learning (ML) & Natural Language Processing (NLP)

The term "AI" is most commonly used when referring to the concept of ML. The latter describes the automated learning from experience through iterative analysis processes on sample data (Murphy, 2012). ML can be seen as a part of data

science, which focuses on the ways we can extract information from data (Provost & Fawcett, 2013) and understand it within a given context (Dhar, 2013). Thus, data science is the integration of many disciplines, including computer science and mathematics, as well as knowledge from the (research) domain it is applied to (e.g., domestic abuse or mental health), which is crucial to help understanding the extracted information. Making use of ML systems is very helpful for a broad range of issues in several areas of science (Jordan & Mitchell, 2015). ML facilitates the addressing and solving of problems both quickly and automatically, without the need to give a machine precise instruction on how to do it. In the case of *supervised* ML methods, the idea is to present the system with a set of labelled datapoints, which it learns to identify automatically (Murphy, 2012). For example, we can present the system with a set of texts or reports, which are labelled as “tech abuse” and “other”. The system then learns to associate the text properties to the corresponding label with little or no direct human control. The goal is then to utilise the trained system to make predictions on a set of unlabelled text data (i.e., identify tech abuse reports based on the text properties it has studied previously). Hence, like a child that has come to know particular patterns, the system will apply the lessons it has been trained upon onto new information.

Since ML methods can only work with numerical data, it is not possible to present the system with plain text. To transform text data into numerical representation, we can use Natural Language Processing (NLP) methods. NLP works at the intersection of human-computer understanding and is a research domain of itself (Jurafsky & Martin, 2019; Nadkarni, Ohno-Machado, & Chapman, 2011). Using NLP methods to generate numerical text features can include simple frequencies or proportions of words (unigrams), or word-pairs (bigrams). Other NLP methods can include the generations of grammatical features of the text, in which each word is given its grammatical label (e.g., noun, verb, etc.), which are called part-of-speech tags (Jurafsky & Martin, 2019). Such part-of-speech tags can also be numerical represented in frequencies or proportions within a text. By adding feature sets together, text properties can be numerical represented, which can be understood by ML models. A supervised ML model can then learn to associate the patterns of these numerical text features with the corresponding labels (e.g., “tech abuse” and “other”). The task of differentiating between the texts labelled either “tech abuse” or “other” is also called a *classification task*. Hence, an algorithm (i.e., the ML method) is also referred to as a *classifier* (Murphy, 2012).

The Present Analysis

Given the significant opportunities that ML and NLP approaches provide, we set out to develop a tech abuse detection algorithm for unstructured text data. Such a method creates the means to identify and monitor tech abuse, generating data upon which evidence-based policy and interventions can be built. The underpinning research questions were: (a) What is the extent of technology-facilitated abuse evident in the Crimestoppers dataset?; and (b) What is the nature of technology-facilitated

abuse apparent in the Crimestoppers dataset? As part of our analysis, we encountered multiple hurdles that prevented us from explicitly answering these questions. However, our setbacks offer useful lessons on the limitations of ML and NLP tools. In this chapter, we outline the methodology, as well as our core findings, before discussing them through consideration of the challenges arising from the development of workable automated tech abuse identification systems. We end the chapter with recommendations and an overview of future research needs.

Method

For our data science-driven project, for which we received ethical approval¹, we worked with un-anonymised data collected as part of Crimestopper’s UK case management system. Crimestoppers is an independent charity that gives people the power to report crime anonymously either via a telephone number or an anonymous online form. The idea of Crimestoppers, which originated in the United States, is that reporting parties are not required to give their name or any personal information. This should lower the barrier for individuals to speak up and to report crime they observe within their surroundings, such as their community or neighbourhood. Thus, the reporting mechanisms encourage “bystanders”, meaning individuals who suspect particular offences to take place, as well as those directly involved in a crime.

Data

We received the Crimestoppers dataset as an Excel spreadsheet. The dataset contained in total 434,088 crime reports from across the UK (England, Wales, Scotland, and Northern Ireland), between 2014 and 2019. All reports were broadly categorised into different crime topics, such as fraud, domestic abuse, rape and sexual offences, drug trafficking, e-crime, murder & other killings. Across different columns, information about the date, type, and region/location of a reported incident are offered. Most important in the context of our study, the dataset included short text summaries of the nature of the report that was logged. These were content – suitably edited to preserve privacy – either provided via the anonymous online form or a written summary of the oral phone conversation noted down by one of Crimestopper’s call handlers. These text snippets are not verbatim transcripts. Instead, they are just a few lines long and are recorded in the call handlers’ words or those of the person reporting the incident. Examples found under the “domestic abuse”- labelled offences can be found below:

“[He] was physically abusing his wife and posting photos on FB of her injuries”; ‘He smashed her phone.’

“Her ex-partner has nude pictures of her and threats to post them around unless she has oral or full sex on a specific date.”

¹ University College London (UCL) Research Ethics Committee - Project ID Number: 10503/001.

“Raping her, drugging her and allowing other men to rape her [his girlfriend] and videoed rape.”

Data Storage

To receive this highly sensitive data, a data sharing agreement with Crimestoppers had been set in place, which included the arrangement that all information would be stored and analysed within the Jill Dando Institute Research Laboratory (JDRL) at University College London. The JDRL is a state-of-the-art, police-assured secure computer facility that allows researchers to store and analyse data classified up to OFFICIAL-SENSITIVE. The lab is an “air-gapped” facility, which means that the computers offer no connection to the outside world and users are not allowed to bring any electronic items into the lab. Data imports and exports are managed via a tightly controlled authorisation protocol and all users that wish to access the JDRL must undergo appropriate personal security checks.

Data Analysis

The purpose of our data analysis was to detect, examine, and compare reports of tech abuse across the Crimestoppers dataset through automated means. As no explicit offense category for tech abuse existed, qualitative descriptions from excerpts seen above were used to elucidate the nature and scope of tech abuse. To find an initial set of tech abuse cases, we conducted a keyword search (see Appendix A for used keywords) on all reports categorised as “domestic abuse”. We manually inspected the filtered reports ($n=700$) and annotated 133 of them as being tech abuse, while we marked 567 of them as non-tech abuse. From the annotated data, we compiled a balanced dataset (i.e., 50% tech abuse and 50% non-tech abuse reports) consisting of all labelled tech abuse cases and randomly matching them with non-tech abuse cases (but still related to domestic abuse). This resulted in 266 reports. Presenting *supervised* ML methods with a balanced dataset is common practice and ensures that the model appropriately learns to differentiate between the classes (Batista, Prati, & Monard, 2004).

We used these 266 reports to train and test the ML algorithm, using a k-fold cross-validation procedure (Kuhn & Johnson, 2013). Such a procedure splits the data into equally sized k sets, for which $k-1$ sets are used for training the model, and the remaining set is used of testing the model (in our case, $k=10$). Through an iterative process, in which each iteration is called a fold, the model is being trained and tested in each fold on a subset of the data. After completing all folds, all subsets were consequently used both for training and testing. To make this possible, each fold exchanges one of the nine subsets, included as part of the training data, with the one set, which is used as the testing set. That way, the model always trains on nine subsets, and is tested on one previously unseen data set. Since the data is split into k sets, the procedure is repeated k times, to complete a full round of training and testing (in our case ten times). Testing the model in each fold translates into predicting the data labels of each datapoint in the testing set. Because we know

what the true labels are for each datapoint, we can evaluate the models’ performances by looking at the correct and incorrect predictions and generate metric scores, such as accuracy for each fold. By averaging the performances scores across all folds, we obtain the average performance of the models on our data. A key advantage of a cross validation procedure is that the model is not evaluated on the same data it is trained on, to avoid “overfitting”. The danger of evaluating the model on the training data is, that the model relies more on specific sample characteristics rather than on the true relationships between the features and the labels. Furthermore, with cross validation we can utilize all labelled data for training and testing as well as observe performance variability in the models’ predictions across the folds. By averaging the performances across all folds, we can obtain a more robust estimation of the models’ prediction performances on unseen data.

In this study, we measured the models’ performances in accuracy, precision, recall, and f1 scores. Although accuracy is very important and most known, scores such as precision and recall can give us additional insights on how well a model is performing. Both scores give us further information on the predictions of the individual classes (i.e., tech-abuse, non-tech abuse). Precision assesses how well false positive cases are avoided, while recall (sensitivity) assesses how well the individual classes were predicted (e.g., how many reports were predicted to be tech abuse from all tech-abuse cases). The f1 score represents the harmonic mean from precision and recall.

To increase the performances of our model to detect tech facilitated abuse in a large corpus of text, we followed an iterative approach of (i) training the model, (ii) making predictions with it on the corpus (detecting tech abuse), and (iii) relabelling the predicted tech abuse cases. That way, we fine-tuned the model by including more correctly labelled reports, which were either correctly or incorrectly predicted by the algorithm. By correctly labelling falsely predicted tech abuse cases in the first iteration, the model used these cases in the next iteration of the training phase to improve its performance. We followed this procedure for two iterations, which lead us to compile a final balanced dataset of 294 reports² (i.e., 147 tech abuse and 147 non-tech abuse).

Since the cross-validation procedure trains and tests a separate model in each fold, to assess the methods general capability, none of the ten models is used for the final predictions on the large corpus. Instead, after determining, which methods works best, the classifier is trained on the whole labelled dataset, to make use of all datapoints. The final model is then used to make predictions on the corpus.

Data Cleaning and Feature Generations

To facilitate the generation of text features, the reports underwent several pre-processing steps. All texts were lowercased to make any further text analyses case

² The code and the trained model for making predictions can be found at: “https://osf.io/fea5j/?view_only=35786879fdee4d21bc1da71cba3661d1”.

insensitive. Next, all punctuations and English stopwords³ (e.g., “by”, “for”, “when”, etc.) were removed, to include more meaningful words, which carry content within sentences. Lastly, all remaining words were stemmed, converting each word to its stemm (e.g., “texting”, “texted” would be stemmed to “text”), to facilitate a more accurate count of the same word meanings. From the clean reports, we extracted Term Frequency-Inverse Document Frequency (TF-IDF) weighted unigrams, bigrams, and part of speech (POS) features⁴. TF-IDF weights are used to lower the importance of high frequency words in a document (e.g., “she”, “the”, “and”), which also occur more frequently across all documents. Thus, it assigns more weight to words which are important (higher frequency) for each individual document (Jurafsky & Martin, 2019). We used the python package “nltk” (Bird, Klein, & Loper, 2009) for text cleaning and feature generations.

Results

In this section we will first discuss the results of our automated approach to detect tech abuse, followed by a manual examination of 147 identified tech abuse reports describing the general observed content and properties of tech abuse cases. The results showcase that initial evaluation scores of our model are promising, but do not translate well into the practical detections of tech abuse reports. Specifically, most reports identified as tech abuse were false positives, showing that the current applied model is not workable to support a quantitative assessment of text data.

Automated Detection of Tech Abuse Reports

To decide which ML method would be most suited to detect tech abuse reports in the large corpus, we trained and tested ten different ML classifiers, utilising the 10-fold cross validation procedure as explained above. Since we used a two-step approach of training, testing and detection, we first trained our models on 226 and then on 294 Crimestoppers reports. Both datasets were always equally balanced between tech abuse and non-tech abuse. In both iterations, the “LinearSVC”⁵ seemed to perform best, for which the averaged performance scores across the ten folds of the last iteration are reported below. All analyses were completed in python using “scikit-learn” (Pedregosa et al., 2011), a programming package for using ML methods.

Model Performances

The accuracy of the “LinearSVC” classifier was 76.59% (SD = 4.18). Scores for each class and their averages are reported in Table 1.

³ For a full list of all stopwords, see Bird, Klein and Loper (2009).

⁴ We also extracted n-gram, and POS proportions, but they resulted in lower classification performances.

⁵ Model parameters were set to $l=1$ and $dual=False$, while the remaining parameters were unchanged (default settings).

	Precision	Recall	F1
Tech-abuse	80.71 (10.37)	73.01 (9.55)	75.63 (3.84)
Non-tech-abuse	75.48 (5.61)	80.06 (13.24)	76.89 (5.84)
Average	78.09 (5.15)	76.54 (4.31)	76.26 (4.15)

Table 1. Averaged performance scores for the “linearSVC” classifier. SD values in parentheses.

Predicting Tech Abuse

We used the “LinearSVC” classifier in both iterations of the training, testing, and detection of tech abuse reports. In both iterations, the evaluation metrics seemed promising (table 1) to warrant an application to detect tech abuse reports on the whole corpus of 434,088 unannotated reports. In the first iteration, the classifier identified 61,969 reports as tech abuse (14% of the whole corpus). As the case numbers seemed very high for such a specific abuse type within this dataset (as Crimestoppers commonly record many different forms of crimes), we manually inspected 700 entries and found 14 as being correctly attributed as tech abuse. After integrating the additional labelled reports to our data pool in our second iteration of training, testing and detection, the model identified 30,289 tech abuse reports across the whole corpus (7% of the whole 434,088 reports). In a second manual inspection, we examined a similar number of reports and found again only seven tech abuse cases.

Manual Analysis of the Nature of Tech Abuse

Having looked at the Crimestoppers data in more detail, tech abuse cases were primarily located in domestic abuse entries. The latter are often descriptive e.g., “commits DA”, “being abusive”, “is domestically violent”, “carries out DA”, making the extraction of information on distinct forms of abuse (e.g., financial, psychological, technical) rather difficult. The DA entries were also mostly reported for heterosexual couples, with incident logs regularly being accompanied by references to mental health issues such as perpetrators being “mentally unstable”.

Across the dataset, mainly “common” tech abuse offences were evident, echoing findings of previous research that attempted to cluster different forms of tech abuse (Brown, Reed, & Messing, 2018; Freed et al., 2017; Henry & Flynn, 2018; Southworth, Finn, Dawson, Fraser, & Tucker, 2007). These incidents included excessive, malicious, and/or unwanted “messages and emails”, some of which involved threats “to kill”. They also comprised of image-based sexual abuse cases, such as sending or threatening to send “nude pictures”, people having “videoed [a] rape”, or “posting photos on FB [Facebook] of her [the victims/survivors] injuries”. Mobile applications such as “Snapchat” and “WhatsApp” were further mentioned. Besides the importance of monitoring and controlling partners through technology was prevalent in the dataset. Products such as “Find my iPhone” were highlighted which allow perpetrators to track a victim/survivor— often without their knowledge.

The analysis of the data did not reveal the existence of tech abuse through “unconventional” systems such as smart Internet of Things devices nor drones or game consoles in IPV situations. However, a particular abuse category that has been discussed to a lesser extent within the tech abuse literature is the active withholding of access to technology (Henry, Vasil, Flynn, Kellard, & Mortreux, forthcoming). This was an element that was relatively prominent in the analysed tech abuse cases. It included perpetrators having “smashed her [the victims/survivors] phone” or them “confiscate[d] his [the victims/survivors] keys, hide his mobile phone to prevent him contacting someone to help him leave the address”.

Discussion

The present study aimed to develop an automated detection system, which could be used to find tech abuse cases within a large corpus of unstructured reports (e.g., charity or police records, administrative data). Since the manual inspections of big text corpora is unfeasible, an automated approach is needed to quantify and uncover tech abuse cases in such datasets. Once identified, these records may then be manually examined for a more detailed investigation. To realise this task and deliver a proof-of-concept, we utilised ML in combination with NLP methods. Previous research has shown that such tools can be effective in detecting specific text types within free text, including abuse types and victim/survivor injuries (Karystianis et al., 2019). However, in our study, the detection of tech abuse proved to be more difficult. Whilst our classifier showed promising performances it had difficulties to classify tech abuse accurately when deployed on a large dataset. Our findings consequently reveal the limitations around the generalisability of such automated methods that researchers, as well as practitioners, should closely consider.

Detecting Technology-Facilitated Abuse

As the presented evaluation scores show (Table 1), the expected performances of our trained classifier seemed to be acceptable in detecting tech abuse. Although the performance scores are not as high as in other works with similar goals (Karystianis et al., 2019), our model would have sufficed in serving as a filter to narrow down the possible reports we would have to inspect manually. Thus, the goal was not to develop a perfect detection system, but rather a sophisticated filtering method. However, after applying the trained classifier on the whole corpus, a large number (over 60k) were predicted as tech abuse and after closer inspection of these cases, only a small proportion (~2%) seemed to be of interest. The performance equated to a false positive rate of 98%, which is in a strong contrast to our expected performances.

To tackle this problem, we re-trained the classifier including newly labelled data from these false positive cases, which seemed to improve the predictions and lowering the predicted tech abuse case to around 30k. The idea for doing so was to train the algorithm on for the model difficult to classify cases, which in turn make it easier for the algorithm to find tech abuse in the future. Nevertheless, after a closer

inspection, the problem seemed to persist and only a small fraction of predicted tech abuse cases was in fact, tech abuse. This shows a strong discrepancy between the expected and the behaved performance of the ML method.

It seems that the evaluation scores did not translate well into assessing the classifiers practical applicability of detecting tech abuse reports. Hence, whenever we tried to generalise the model to look for these smaller values of tech abuse in the overarching sample (i.e., like finding a needle in a haystack), the method seems not to be as sensitive as needed. It is not immediately apparent, why the classifier is not working as anticipated, but some potential reasons are as follows. First, the training and evaluation sets did not represent the corpus well, effectually invalidating the evaluation scores. In our first iteration we only labelled reports originating from the domestic abuse category, while we applied the model to the whole corpus. Although, the domestic abuse category is not an adequate representation of the whole corpus, it includes reports, which are more difficult to differentiate (e.g., a report mentioning technology, which is not involved in the abuse vs. a report which mentions technology, which is involved in the abuse.). Thus, the initial training and testing set contained such difficult cases, which should be helpful for the model to learn about nuances in the data and classify them more accurately. Furthermore, in the second iteration of the model, the training set contained reports from other categories as well as reports which were initially falsely predicted, adding more difficult to differentiate cases, it can learn from. However, these measures did not lead to better detections, practically.

Second, the reports included in the data are very short (often ranging between five to ten short sentences). The brevity of the recorded call handler notes, limits the details and semantics which can be captured. Thus, the model can only be presented with very limited information. While this might be one of the problems within this study of detecting tech abuse, utilising short texts or “imperfect” recorded data will have to be addressed in the future.

Third, it is possible that the corpus does not contain a lot of tech abuse cases, which would make it very difficult to locate them. This problem is also referred to as a low base rate, which makes it even for highly accurate models very difficult to have low false positive rates (Axelsson, 2000).

Fourth, tech abuse as a concept and abuse form may not be well enough defined, highlighting the need to agree on exact features to allow for its possible application as a distinct offence category in future data collection processes. Currently, tech abuse is not an “official” concept, nor measurement category. For instance, Markwick et al. (2019) underlined that scholars have used different terminology to describe the perpetration of abuse and harassment via digital means. All used terms have further no agreed specifications but are frequently associated with a combination of “behaviours” (Dragiewicz et al., 2018), “areas” (Henry & Flynn, 2018) or “dimensions” (Powell & Henry, 2018). Additionally, different “forms” of tech abuse commonly intersect (Brown et al., 2018; Messing et al., 2020) and the subjective nature of tech abuse makes it further challenging to define, detect, and measure technology-enabled abuse incidents (Messing et al., 2020). Throughout

our iterative coding process, we encountered exactly this problem and frequently struggled to make a definite assessment whether a Crimestoppers report should or should not fall under our evaluation of tech abuse. It is possible, that this uncertainty of the labelling process is passed on to the model, making it more difficult to identify tech abuse.

Although all previously mentioned reasons could lead to reduced predictions performances, it is important to reiterate that the evaluation metrics did not transfer well into a practical prediction model. The model seems accurate but is not practical. This discrepancy raises questions on the practicality of exiting ML models with similar tasks and goals because the evaluation metrics might not always capture the actual prediction performances. It further illustrates that we need additional control mechanisms to ensure our ML models behave as anticipated, which we mention below in the “recommendations” section.

Insights from the Manual Inspections

With regards to manual analysis, our study sadly did not provide the anticipated in-depth details on the exact distribution and nature of the different nuances of tech abuse. However, we were able to observe elements that echo existing dynamics noted by scholars and practitioners alike, including image-based sexual abuse forms, malicious and/or unwanted messages, and stalking behaviours (Flynn & Henry, 2019; Henry & Flynn, 2019; McGlynn et al., 2017; McGlynn, Rackley, & Johnson, 2019; Messing et al., 2020; Women’s Aid, 2018; Yardley, 2020). Despite the uniqueness and qualitative difference of our dataset, there was nothing out of the ordinary that we were able to detect. Instead, the cases of tech abuse we reviewed replicated tech abuse dimensions and trends of which we already know or suspect to be happening.

However, one element worth pointing out is the huge proportion of cases that we classified as involving the active withholding of devices and technical systems such as phones. Instead of an “active” or a “kinetic” misuse of technology, its suppression may possibly be as serious as their deliberate manipulation. This observation also plays into our earlier posed question *what* tech abuse *is* and *what counts* as technology-mediated abuse. Indeed, the research team wondered, firstly, whether tech abuse did not flag up in the analysed datasets, because callers may themselves not consider aspects such as the confining of a device, as a crime. Secondly, they may also not perceive it severe “enough” to be worth reporting (McGlynn et al., forthcoming). The latter would explain why cases of physical, financial, or sexual abuse remain so prevalent. Thirdly, tech abuse may be an element of IPV that would be easier to observe in datasets where victims/survivors themselves (or people very close to them or the perpetrator) detail about wrongdoings (e.g., domestic abuse charity data). Unlike physical violence, where family, friends, and neighbours may hear or “see” something (e.g., screams, bruises), tech abuse may be much harder to detect and possibly also less likely to be reported by external, third parties.

Due to all these considerations, we are, therefore, attentive to accentuate that the absence of an observation (i.e., of tech abuse) does not imply the absence of the act itself. We consequently are hopeful that future research into tech abuse, will help to expand our understanding of the different subtleties that are part of a pattern of perpetrators not acknowledging victim's/survivor's boundaries and the difference shades and shapes that tech abuse may involve.

List of Recommendations and Future Research Needs

Across this chapter, we hoped to have revealed the practicalities of and challenges around doing experimental research on real world unstructured text data. We now want to end our article with a list of recommendations that derive from the lessons we learned throughout this study. These suggestions may profit researchers and other parties interested in further developing automated detection systems for tech abuse cases. Firstly, we could employ other more sophisticated methods to generate text features. Such methods could include using the Linguistic and Word Count Software (LIWC) (Pennebaker, Boyd, Jordan, & Blackburn, 2015) or word embeddings (Jurafsky & Martin, 2019). When training a classifier, it might be useful to tune the model's parameter to avoid false positive predictions, in exchange for more false negatives. Although this is not an optimal solution, as some positive cases will be missed, it would make the model somewhat more practical, as it would reduce the manual labour of checking all positive predicted cases. However, this approach is highly dependent on the classification task and the costs of increasing false negatives should be clearly considered. Lastly, there is a need for more iterative process when working with ML. As our study shows, humans must be in the loop and are required to audit the output as we did. Thus, we encourage others to always take a sample of your predicted cases to check if your predictions are practical, otherwise you might run the risk of falsely attributing prediction performances.

With that being said, we are also considerate that automated detection mechanisms may remain too shallow to understand the breadth of dynamics that come into play when studying the phenomena of tech abuse. We therefore encourage researchers and practitioners to not disregard the lived realities of victim/survivors, who remain unheard in such quantitative evaluations. Indeed, the research team believes that our study shows the necessity to continue qualitative research on tech abuse, especially whilst such other methodological tools remain ineffective. We should also not forget to pursue research into the root causes, consequences, escalation trajectories, and the causal pathways that precipitate tech abuse as well as the legal as well as technical instruments that could help improve the situation for victims/survivors. All these elements should never be replaced by the "sexiness" of latest instruments such as ML and NLP, who in many ways are only an add-on rather than a replacement to our existing research toolkit.

Author Biographies

Felix Soldner is a PhD student at University College London's Department of Security and Crime Science. He has a wide interest in many research areas, which led him to study Psychology, Biomimetics as well as Brain and Cognitive Sciences. Following this, he developed an interest in data science, machine learning, and natural language processing, which he now integrates in his work. His current research focuses on how data science methods can facilitate the detection and prevention of online fraud (e.g., counterfeit goods).

Leonie Maria Tanczer is a Lecturer in International Security and Emerging Technologies at University College London's Department of Science, Technology, Engineering and Public Policy (STeAPP) and affiliated with UCL's Academic Centre of Excellence in Cyber Security Research. Her research focuses on questions related to Internet security, and she is specifically interested in the intersection points of technology, security, and gender. Since 2018, she is the Principal Investigator of the "Gender and IoT" study, which examines the implications of the Internet of Things (IoT) and other digital technologies on victims/survivors of gender-based domestic violence and abuse.

Daniel Hammocks is a PhD student at the University College London's Department of Security and Crime Science and in the second of a four-year MRes+MPhil/PhD programme. He has a background in mathematics and data science, with his current research focussing on the detection and prediction of emerging crime trends, as well as future methods of perpetration. Across his research portfolio, he is interested in the application of data science, computer vision, and data visualisation to the Crime Science domain with a particular focus on policing.

Isabel Lopez-Neira is currently a **Sustainability Policy Officer at the European Consumer Organisation (BEUC) and the European Consumer Voice in Standardisation (ANEC)**. During her time working as Research Assistant at University College London's Department of Science, Technology, Engineering and Public Policy (STeAPP), she was an active member of the "Gender and IoT" research project, and since then maintains close links with the project. Isabel holds a UCL Master's degree in Science and Technology Studies. Her research interests focus on ethical and sociological questions surrounding research, technology, and innovation.

Shane D Johnson is the Director of the Dawes Centre for Future Crime, Professor of Future Crimes and Co-Director of the Centre for Doctoral Training in Cybersecurity at University College London. He has worked within the fields of criminology and forensic psychology for three decades. His research has explored how methods from other disciplines (e.g., complexity science) can inform understanding of crime and security issues, and the extent to which theories developed to explain common crimes can clarify more extreme events such as riots, maritime piracy, and cybercrime. He is currently interested in how technological and social change informs new opportunities for offending or approaches to crime prevention.

Acknowledgements

The authors are indebted to Crimestoppers for sharing their data, the Jill Dando Institute Research Laboratory (JDIRL) for hosting all records, and to Oli Hutt and Nigel Swift for providing us with technical support whilst using the JDIRL. We are also thankful to the book editors (Dr Asher Flynn, Dr Anastasia Powell, and Dr Lisa Sugiura) for their flexibility and support throughout our analysis and write-up process. Parts of the insights discussed in this publication stem from findings derived from UCL’s “Gender and IoT” research project. The latter has received funding from the UCL Social Science Plus+ scheme, UCL Public Policy, the PETRAS IoT Research Hub (EP/N02334X/1), the UK Home Office, and the NEXTLEAP Project (EU Horizon 2020 Framework Programme for Research and Innovation, H2020-ICT-2015, ICT-10-2015, grant agreement No. 688722). The work was also supported by the Dawes Centre for Future Crime at UCL.

References

Axelsson, S. (2000). The base-rate fallacy and the difficulty of intrusion detection.

ACM Transactions on Information and System Security, 3(3), 186–205.

<https://doi.org/10.1145/357830.357849>

Banerjee, M., Lee, J., & Choo, K.-K. R. (2018). A blockchain future for internet

of things security: A position paper. *Digital Communications and Net-*

works, 4(3), 149–160. <https://doi.org/10.1016/j.dcan.2017.10.006>

Batista, G. E. A. P. A., Prati, R. C., & Monard, M. C. (2004). A study of the be-

havior of several methods for balancing machine learning training data.

ACM SIGKDD Explorations Newsletter, 6(1), 20–29.

<https://doi.org/10.1145/1007730.1007735>

Bird, S., Klein, E., & Loper, E. (2009). *Natural Language Processing with Py-*

thon. Springfield: O’Reilly Media.

Blunden, M. (2018, August 28). Abusive partners use home technology to stalk

and abuse women, study shows. Retrieved from website:

<https://www.standard.co.uk/tech/abusive-partners-use-home-technology-to-stalk-and-abuse-women-study-shows-a3921386.html>

- Brown, M. L., Reed, L. A., & Messing, J. T. (2018). Technology-Based Abuse: Intimate Partner Violence and the Use of Information Communication Technologies. In J. R. Vickery & T. Everbach (Eds.), *Mediating Misogyny: Gender, Technology, and Harassment* (pp. 209–227). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-72917-6_11
- Burton, S., Tanczer, L. M., Vasudevan, S., & Carr, M. (2021). *The UK Code of Practice for Consumer IoT Cybersecurity: Where we are and what next* (pp. 1–74). London: Department of Digital Culture, Media and Sport; The PETRAS National Centre of Excellence for IoT Systems Cybersecurity. Retrieved from website: <https://discovery.ucl.ac.uk/id/eprint/10117734/>
- Citron, D., & Franks, M. A. (2014). Criminalizing revenge porn. *Wake Forest Law Review*, 49, 345–391.
- Dhar, V. (2013). Data science and prediction. *Communications of the ACM*, 56(12), 64–73. <https://doi.org/10.1145/2500499>
- Dragiewicz, M., Burgess, J., Matamoros-Fernández, A., Salter, M., Suzor, N. P., Woodlock, D., & Harris, B. (2018). Technology facilitated coercive control: Domestic violence and the competing roles of digital media platforms. *Feminist Media Studies*, 18(4), 609–625. <https://doi.org/10.1080/14680777.2018.1447341>

- Flynn, A., & Henry, N. (2019). Image-Based Sexual Abuse: An Australian Reflection. *Women & Criminal Justice*, 0(0), 1–14.
<https://doi.org/10.1080/08974454.2019.1646190>
- Freed, D., Palmer, J., Minchala, D. E., Levy, K., Ristenpart, T., & Dell, N. (2017). Digital Technologies and Intimate Partner Violence: A Qualitative Analysis with Multiple Stakeholders. *Proceedings of the ACM on Human-Computer Interaction*, 1(CSCW), 1–22. <https://doi.org/10.1145/3134681>
- Harris, B. A., & Woodlock, D. (2019). Digital Coercive Control: Insights from Two Landmark Domestic Violence Studies. *The British Journal of Criminology*, 59(3), 530–550. <https://doi.org/10.1093/bjc/azy052>
- Henry, N., & Flynn, A. (2019). Image-Based Sexual Abuse: Online Distribution Channels and Illicit Communities of Support. *Violence Against Women*, 25(16), 1932–1955. <https://doi.org/10.1177/1077801219863881>
- Henry, N., & Flynn, A. L. G. (2018). *Technology-Facilitated Abuse among Culturally and Linguistically Diverse Woman: A Qualitative Study* (p. 94). Canberra: Office of the eSafety Commissioner. Retrieved from website: <https://research.monash.edu/en/publications/technology-facilitated-abuse-among-culturally-and-linguistically->
- Henry, N., Flynn, A., & Powell, A. (2020). Technology-Facilitated Domestic and Sexual Violence: A Review. *Violence Against Women*, 26(15–16), 1828–1854. <https://doi.org/10.1177/1077801219875821>

- Henry, N., McGlynn, C., Flynn, A., Johnson, K., Powell, A., & Scott, A. J. (2020). *Image-based Sexual Abuse: A Study on the Causes and Consequences of Non-consensual Nude or Sexual Imagery*. London; New York: Routledge. <https://doi.org/10.4324/9781351135153>
- Henry, N., Vasil, S., Flynn, A., Kellard, K., & Mortreux, C. (forthcoming). Technology-Facilitated Domestic Violence Against Immigrant and Refugee Women: A Qualitative Study. *Journal of Interpersonal Violence*. <https://doi.org/10.1177/08862605211001465>
- Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, 349(6245), 255–260. <https://doi.org/10.1126/science.aaa8415>
- Jurafsky, D., & Martin, J. H. (2019). *Speech and Language Processing An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition—Third Edition draft*. Stanford. Retrieved from website: <https://web.stanford.edu/~jurafsky/slp3/>
- Karystianis, G., Adily, A., Schofield, P. W., Greenberg, D., Jorm, L., Nenadic, G., & Butler, T. (2019). Automated Analysis of Domestic Violence Police Reports to Explore Abuse Types and Victim Injuries: Text Mining Study. *Journal of Medical Internet Research*, 21(3), e13067. <https://doi.org/10.2196/13067>
- Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling*. New York: Springer.

- Lopez-Neira, I., Patel, T., Parkin, S., Danezis, G., & Tanczer, L. M. (2019). 'Internet of Things': How abuse is getting smarter. *Safe – The Domestic Abuse Quarterly*, (63), 22–26.
- Markwick, K., Bickerdike, A., Wilson-Evered, E., & Zeleznikow, J. (2019). Technology and Family Violence in the Context of Post-Separated Parenting. *Australian and New Zealand Journal of Family Therapy*, 40(1), 143–162. <https://doi.org/10.1002/anzf.1350>
- McGlynn, C., Johnson, K., Rackley, E., Henry, N., Gavey, N., Powell, A., & Flynn, A. (forthcoming). 'It's Torture for the Soul': The Harms of Image-Based Sexual Abuse: *Social & Legal Studies*. (Sage UK: London, England). <https://doi.org/10.1177/0964663920947791>
- McGlynn, C., Rackley, E., & Houghton, R. (2017). Beyond 'Revenge Porn': The Continuum of Image-Based Sexual Abuse. *Feminist Legal Studies*, 25(1), 25–46. <https://doi.org/10.1007/s10691-017-9343-2>
- McGlynn, C., Rackley, E., & Johnson, K. (2019). *Shattering lives and myths: A report on image-based sexual abuse*. Durham, Kent: Durham University; University of Kent.
- Messing, J., Bagwell-Gray, M., Brown, M. L., Kappas, A., & Durfee, A. (2020). Intersections of Stalking and Technology-Based Abuse: Emerging Definitions, Conceptualization, and Measurement. *Journal of Family Violence*, 35(7), 693–704. <https://doi.org/10.1007/s10896-019-00114-7>

- Murphy, K. P. (2012). *Machine learning: A probabilistic perspective*. Cambridge, MA: MIT Press.
- Nadkarni, P. M., Ohno-Machado, L., & Chapman, W. W. (2011). Natural language processing: An introduction. *Journal of the American Medical Informatics Association*, 18(5), 544–551. <https://doi.org/10.1136/amiajnl-2011-000464>
- Parkin, S., Patel, T., Lopez-Neira, I., & Tanczer, L. M. (2019). Usability analysis of shared device ecosystem security: Informing support for survivors of IoT-facilitated tech-abuse. *Proceedings of the New Security Paradigms Workshop*, 1–15. San Carlos, Costa Rica: Association for Computing Machinery. <https://doi.org/10.1145/3368860.3368861>
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. *The Journal of Machine Learning Research*, 12, 2825–2830.
- Pennebaker, J. W., Boyd, R. L., Jordan, K., & Blackburn, K. (2015). *The Development and Psychometric Properties of LIWC2015*. Austin, TX: University of Texas at Austin. <https://doi.org/10.15781/T29G6Z>
- Powell, A., & Henry, N. (2018). Policing technology-facilitated sexual violence against adult victims: Police and service sector perspectives. *Policing and Society*, 28(3), 291–307. <https://doi.org/10.1080/10439463.2016.1154964>
- Powell, A., & Henry, N. (2019). Technology-Facilitated Sexual Violence Victimization: Results from an Online Survey of Australian Adults: *Journal of*

- Interpersonal Violence*, 34(17), 3637–3665. (Sage CA: Los Angeles, CA).
<https://doi.org/10.1177/0886260516672055>
- Powell, A., Scott, A. J., Flynn, A. L. G., & Henry, N. (2020). *Image-based sexual abuse: An international study of victims and perpetrators – A Summary Report* (pp. 1–16). Melbourne: RMIT University. Retrieved from website: https://www.researchgate.net/publication/339488012_Image-based_sexual_abuse_An_international_study_of_victims_and_perpetrators
- Provost, F., & Fawcett, T. (2013). Data Science and its Relationship to Big Data and Data-Driven Decision Making. *Big Data*, 1(1), 51–59.
<https://doi.org/10.1089/big.2013.1508>
- Refuge. (2020, January 9). 72% of Refuge service users identify experiencing tech abuse. Retrieved from website: <https://www.refuge.org.uk/72-of-refuge-service-users-identify-experiencing-tech-abuse/>
- Slupska, J., & Tanczer, L. M. (forthcoming). Intimate Partner Violence (IPV) Threat Modeling: Tech abuse as cybersecurity challenge in the Internet of Things (IoT). In J. Bailey, A. Flynn, & N. Henry (Eds.), *Handbook on Technology-Facilitated Violence and Abuse: International Perspectives and Experiences*. London: Emerald Publishing.
- Snook, Chayn, & SafeLives. (2017). *Tech vs Abuse: Research Findings 2017* (pp. 1–56). London: Comic Relief.

- Southworth, C., Finn, J., Dawson, S., Fraser, C., & Tucker, S. (2007). Intimate Partner Violence, Technology, and Stalking. *Violence Against Women*, 13(8), 842–856. <https://doi.org/10.1177/1077801207302045>
- Tanczer, L. M. (forthcominga). Das Internet der Dinge: Die Auswirkung „smarte“ Geräte auf häusliche Gewalt. In N. Prasad & A. Hartmann (Eds.), *Digitalisierung geschlechtsspezifischer Gewalt*. Berlin: Transcript Verlag.
- Tanczer, L. M. (forthcomingb). Technology-facilitated abuse and the Internet of Things (IoT): The implication of the smart, Internet-connected devices on domestic violence and abuse. In B. Harris & D. Woodlock (Eds.), *Technology and Domestic Violence*. London: Routledge.
- Tanczer, L. M., Lopez-Neira, I., Parkin, S., Patel, T., & Danezis, G. (2018). *Gender and IoT (G-IoT) Research Report: The rise of the Internet of Things and implications for technology-facilitated abuse* (pp. 1–9). London: University College London. Retrieved from website: <https://www.ucl.ac.uk/steapp/sites/steapp/files/giot-report.pdf>
- Women's Aid. (2018). Online and digital abuse. Retrieved from website: <https://www.womensaid.org.uk/information-support/what-is-domestic-abuse/onlinesafety/>
- Woodlock, D. (2017). The Abuse of Technology in Domestic Violence and Stalking. *Violence Against Women*, 23(5), 584–602. <https://doi.org/10.1177/1077801216646277>

Yardley, E. (2020). Technology-Facilitated Domestic Abuse in Political Economy: A New Theoretical Framework. *Violence Against Women*.
<https://doi.org/10.1177/1077801220947172>

Appendix A: Used Keywords

Physical devices:

"smart", "device", "computer", "laptop", "alexa", "tablet", "keytracker", "tracker", "(smart)-heater", "light", "lock"

Online platform apps:

"online", "technology", "internet", "digital", "dating app", "facebook", "systems", "messages", "apps", "service", "account", "platform", "dating site", "instagram", "snapchat", "tinder", "app", "whatsapp", "spyware", "find my iPhone", "find my Friends", "gps", "youtube", "caller id", "profile", "sniffer", "Badoo", "messenger", "chat messenger", "fake account", "flirtfinder", "ipad", "snap chat", "what's app"

Verbs:

"dating", "stalking", "control", "victimisation", "report", "access", "texting", "calling", "sexting", "experience", "bullying", "rape", "video", "use", "abuse", "sexualise", "harass", "harm", "perpetrate", "experiment", "sharing", "threat", "intimate", "message", "phone", "post", "follow", "cyberbullying", "doxing", "tracking", "monitoring", "watching", "blackmailing", "humiliate", "restrict", "destroy", "punish", "force", "impersonate", "gaslight", "controlling", "distribute", "hacking", "attack", "expose", "film", "command", "spread", "s hout"