How technology can mitigate and counteract cyber-stalking and online grooming

Haider M. al-Khateeb, Gregory Epiphaniou
haider.alkhateeb@beds.ac.uk, gregory.epiphaniou@beds.ac.uk

Department of Computer Science and Technology
University of Bedfordshire, UK.

Abstract. With the virtual world becoming part of the social lives of adults and minors alike, new attack vectors emerged to increase the severity of human-related attacks to a level the community have not experience before. This article investigates and shares an outline on how technology could emerge further to counteract and mitigate the damage caused by online perpetrators. The review encourages approaching online harassment, stalking, bullying, grooming and their likes with an Incident Response methodology in mind. This includes a detection phase utilising automated methods to identify and classify such attacks, conduct digital forensic investigations to analyse the nature of the offence and reserve evidence, taking preventive measures as part of the reaction towards the problem such as filtering unwanted communications and finally looking at how we can rely on applicable computing to support and educate the victims.

1. The problem with online anti-social offences

The Internet has enabled people to socialise and exchange personal information online with a high degree of anonymity, together with lack of supervision, this has encouraged different forms of anti-social behaviour against vulnerable individuals including adolescents. Likewise, affected individuals could be victims of power abuse at the workplace. Whilst false accusations, defamation and threats are examples of different forms of harassment that could take place online, cyberbullying in this context is defined as the continuous harassment targeting minors and cyberstalking being the term when this behaviour is used against adults (some online references use these two terms interchangeably). Further, The National Centre for Cyberstalking Research (NCCR) recognised that the persistent nature of stalking behaviour is determined when harassment
events occur for ten or more times over a period of four weeks or more. In the UK, 38% of cyberstalking victims surveyed shared how they feared that the offensive behaviour could develop to a face-to-face confrontation whereas females feared physical injury as a consequence. [1]

These anti-social attacks come with a big price, research shows the impact is not limited to emotional and psychological suffering [2] as the severity of the consequences could lead to suicide or murder [1]. Additionally, significant financial loss was reported with some victims forced to move houses or change their work place.

Surveys also show that almost one of every two teens in the United States alone would experience cyberbullying at some point during their life. Likewise, online grooming has reportedly been a serious issue for minors with around 12% of those ages 11-16 years old in the UK received unwanted sexual messages as published by the National Society for Prevention of Cruelty to Children (NSPCC).

There are laws in place to criminalise these actions like the Protection from Harassment Act 1997 and Sexual Offences Act 2003 S15 in the UK, but law enforcements as well as victims require help in terms of support and advice in additional to other preventative measures that could only be implemented using technology. This should be a matter of common-sense since the offence itself is taking place online.

2. Why automated detection is needed

Despite the nature of the mitigation technique in place, detection is vital in the war against online predators. It is required to enable further actions to isolate the offence, alert victim, block communication or quarantine content to support potential digital forensic investigation. It could also trigger support for the victims by alerting a supervisor.

Automated detection of unwanted communication has very recently been covered within the area of machine learning, however there has already been satisfactory level of success despite the technical challenges; detecting anti-social behaviour can be very difficult and will presumably require human revision, one reason could be the fact that an intimidating message does not necessarily contain negativity or profanity, think of an anonymous love message not coming from your spouse at a time you are chased by a stalker at work. Likewise, without the help of artificial intelligence and machine learning to filter online content, detection could fail in many scenarios; put on the shoes of a website owner attempting to review thousands of user posts ‘manually’ to clean offensive text. As such, a balanced hybrid approach could very well be the answer to the problem.

3. Detection and classification methods

To detect a pattern of unwanted (e.g. harassing) text messages, data mining techniques are used in which algorithms are trained to recognise and classify text. Training datasets (corpus) must be real or directly collected from a relevant field such as YouTube comments, messenger chat-logs, or through academic studies. For instance, in a study by
Gupta et al., (2012) [3] examples of online grooming have been demonstrated and categorised. The next step for such experiments include pre-processing where stop-words and other unnecessary characters are removed from the sentences without affecting the meaning. Term-weight is the next phase were techniques like Term Frequency-Inverse Document Frequency (TF-IDF) is used, it is a statistical method to measure the significance of a term in a given document within the corpus, it takes into consideration frequency of occurrences. TF is first applied to determine frequency \((f)\) of term \((t)\) in a given document \((d)\), hence \(\text{tf}(t,d) = f(t,d)\), variations of TF include Boolean (binary) where the term weight is either occur \((=1)\) or not \((=0)\) and log normalisation where \(\text{tf}(t,d) = 1 + \log f(t,d)\), or zero if \(f(t, d)\) is zero. IDF is then applied to determine the value of the term cross all documents \((D)\), this is important to understand how much information \((\text{value})\) each term provides to the meaning as demonstrated in figure 1. Data will then be manually classified and annotated to give the algorithm sufficient information of how it should behave in the future. For instance, one study [4] argued that cyberbullying in high schools could be due to: physical appearance, sexuality, race & culture and intelligence so their study prepared these four classes for their algorithm. They have also assumed that harmful bullying can be described to include both negativity and profanity.

\[
\text{tfd}(t, d, D) = \text{tf}(t, d) \times \text{idf}(t, D)
\]

<table>
<thead>
<tr>
<th>\textbf{Term(s) 1}</th>
<th>\textbf{Doc 1}</th>
<th>\textbf{Doc 2}</th>
<th>...</th>
<th>\textbf{Doc n}</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>12</td>
<td>2</td>
<td>...</td>
<td>1</td>
</tr>
<tr>
<td>\textbf{Term(s) 2}</td>
<td>0</td>
<td>1</td>
<td>...</td>
<td>0</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>\textbf{Term(s) n}</td>
<td>0</td>
<td>6</td>
<td>...</td>
<td>3</td>
</tr>
</tbody>
</table>

\textbf{Figure 1: Term Frequency-Inverse Document Frequency (TF-IDF)}

Algorithms used for this purpose include famous generative and discriminative classifiers such as Naïve Bayes (NB), Support Vector Machines (SVM) and Decision Trees including J48 and C5.0. There are many other variations, and each works differently; some performed better when the dataset was limited while others can simply be more accurate giving a suitable length of text is available. After studying some of the recent studies, it is found that in Rule-based JRip gave more accurate results recognising cyberbullying compared to SVM, NB and J48 [4] while NB outperformed SVM and C4.5 in another study [5]. Key issue here is that the overall process matters e.g. stemming techniques used. There is still a lot of research required to tone the right measures for each type of anti-social behaviour but with the encouraging level of accuracy reported, it is possible to start applying the technology and provide a level of robust detection based on statistical methods. Assuming a hybrid approach as advised earlier, the end user could still manually report an incident whereas many others will be mitigated through the system.
These offences can be categorised into multiple stages, each should ideally be responded to with the appropriate Incident Response techniques. A good attempt is seen in a study [6] where classification models have been used to identify each stage of online grooming after studying the predators’ behaviour, these are; 1) deceptive trust development 2) grooming stage, and finally 3) to physically approach the minor.

Needless to say, each communication channel comes with various metadata and characteristics that is invaluable to enhance detection. For example, social graphs can be utilised as a white list to reduce false positives whereas personal profile details can be used to question the age, gender, interests and work of those attempting to communicate. It is therefore important that system designers take into account metadata, context together with statistical content analysis.

4. More on shielding unwanted communication

There is a requirement for a tool to shield communication; filter out and reject unwanted calls and messages. Most internet security solutions (Kaspersky, Norton etc) include parental control but they still lack a dedicated response to the social risks described in this article. That being said, in its simplest form, having a simple black-list based product in place is strongly recommended as compared to no protection installed in a device used by vulnerable individuals especially minors. Meanwhile, recent work demonstrates the advancement in machine learning where context-based algorithms are utilised to extract malicious messages and keep an online chat safe as discussed earlier. SafeChat [7] is an example of a proposed tool with an additional feature to create, authenticate and encrypt the communication channel. This proof-of-concept material from research gives clear clues to the type of products to look for.

Giving that some attackers are known to the victims, blocking IDs and mobile numbers is an obvious technique to consider. The second version of the Electronic Communication Harassment Observation (ECHO) survey by the NCCR shows that service providers including mobile operators tend to refuse to block numbers per victim’s request, instead, they would ask the customer to manually block numbers on their device which usually transfer the call into voice mail. Unfortunately this is not a good mitigation procedure, since customer will still get the unwanted communication anyway. The conclusion formed on this example and many others is that the response at the end-user side is limited, more must be done by service providers at their end to protect a service being sold to the community, let this be as part of their commitments, or from the viewpoint of their Corporate Social Responsibility (CSR).

5. Digital forensics

When the source of the attack is anonymous, digital investigations could help to trackback and recover the identity of the perpetrator. Forensic investigators should consider the unique nature of these attacks to use the right tools. However, in other cases, anonymity might not be the problem but rather the ability to preserve admissible digital evidence to a court of law. The National Stalking Helpline in the UK shared on their web pages that “67.8% of stalking reported [to them] between 2010 and 2011 were carried out by someone well known to the victim, such as a colleague, former partner or family
member”. It is clear then that the key challenge could be the technical expertise required that a traditional victim is unlikely to handle. As such, specialised software could be designed to record and control the scene when necessary with the user’s consent. We have already seen this approach proved effective when technical help is offered by companies to support their products. In this regard, such software (system) has to capture data at the local host of the victim to avoid the Man-In-The-Middle problem when a covert channel is utilised (encrypted data can not be monitored) [8]. Local data could then be fully monitored when a transparent network bridge is implemented to join and have full access to the communication with the internet. A design like this should also consider the ability to validate data integrity and keep it secure. Cryptographic hash functions are suitable for this purpose while the data itself can be secured using symmetric encryption with a suitable key size such as AES-256.

After the acquisition of digital evidence e.g. chat logs, content can be analysed to validate information related to the identity of the predator. Even when the anonymity of the attacker remains a challenge, authorship analysis could still be very useful to extract contextual clues to predict information such as the age of the person typing [9]. Some techniques attempt to establish a link between media files shared online and the identity of the owner, Sensor Pattern Noise (SPN) was tested for this purpose with satisfactory outcome to evidence its feasibility [10], SPN has the uniqueness to act as a fingerprint and could therefore be used to associate a given picture to a social media account where other similar images have been uploaded. When there is already a suspect, the investigation could be relatively much easier. SPN is unique due to the way digital photos are created using cameras as demonstrated in Figure 2, each of these steps contributes to the signature (and therefore uniqueness) of the photo. For example, de-mosaicking and post-processing interpolate and adjust colours to increase the overall quality (e.g. de-noising) and this can vary between devices.

![Image acquisition pipeline in typical camera devices. [10]](image)

Specialist investigation toolkits is another effective approach; an example is the Predator and Prey Alert (PIPA), an anti-cyberstalking implementation of a proactive software and hardware system to aid digital forensics investigations, we could assume its effectiveness to counteract similar anti-social crimes. PIPA’s key features include the ability to record and enforce data integrity; and evidence isolation and production to extend the Chain of Custody (CoC). While forensics tools are not easy to get or use, they are designed for professional incident responders rather than the normal user. However, the lack of such
resource could prevent law-enforcements in many countries from having the right tools to protect the community due to fund-related implications. In such cases, an open source approach is strongly recommended.

6. Support and training

Technology can simulate peer-support to empower victims. Not only will this make the service cost effective but it should provide means to share a unified evidence-based advices any time requested. For example, One proposal [11] focused on providing help to children to cope with the negative emotional impact of bullying. In this scenario, the victim describes their emotion state to an Embodied Conversational Agent (ECA). The agent will gather more information on what happened and attempt to refer to previous events when possible to build emotional links with the child. After the situation is realised, the agent shared an advice on what to do next. Another approach [12] includes building a database of previous experiences, this is then used to match the story of a new victim to a previous one for educational and support purposes, the key message here is, you are not alone and this is how you can handle the situation.

Using technology to provide better support and training may not be limited to victims, these same ideas can be used to train the trainer. This could include the police, the PIPA system discussed earlier was used as a training tool [13] not only to learn the technical side but to develop understanding for the behaviour of the perpetrators.

7. Conclusions

Detecting anti-social crimes can be extremely difficult using traditional tools, but there are currently satisfactory attempts from the area of machine learning where the context of the communication is included in the analysis to increase accuracy and therefore the overall reliability of any system. In view of that, prevention techniques can developed, published research include good proposals but these are still in the lab, more should be done for such methods to mature and provide sufficient help and support. This article has demonstrated that applicable technology can help investigators, social workers and victims, likewise, it is invaluable to provide emotional support and advice through agents as much as it could be useful to block risk.

For the near future, it is expected that many relevant research areas will consider the applicability of their technology to counteract anti-social behaviour. For instance, voice recognition methods can be used to verify the identity of the person calling, could this mean phone systems will one day be configured to recognise and automatically record incoming calls from a pre-identified predators as evidence? The answer is probably yes.

References


**Authors**

**Dr. Haider M. al-Khateeb**

Haider is a lecturer in the department of Computer Science and Technology, University of Bedfordshire and a Research Fellow in applicable computing and an active member of the National Centre for Cyberstalking Research (NCCR). He has a first-class BSc (Hons) in Computer Science, and PhD in Computer Security.

Haider has cross-disciplinary skills in computing but is mainly interested in Cyber Security and Digital Forensics. His current research is in areas including Memory Forensics, Authentication Methods, Internet-of-Things Security and Cyberstalking. Haider is also a Fellow of the Higher Education Academy (FHEA), UK.

Twitter: @H4ider

**Dr Gregory Epiphaniou**

Gregory is a Senior Lecturer in the Department of Computer Science and Technology at the University of Bedfordshire. He has driven and achieved quality delivery at BSc and MSc levels.

Dr Gregory has also worked as cyber security consultant and trainer for QA ltd with strong engagement with several industry partners in Information security domains. He delivered a wide range of industrial and bespoke certifications including but not limited to CISSP, CISMP, CEH and BCS. He has also been a leading trainer and developer for the GCHQ Summer School in Cyber Security.