A Machine Learning Framework to Detect And Document Text-based Cyberstalking

Zinnar Ghasem[1], Ingo Frommholz[1], and Carsten Maple[2]

[1] University of Bedfordshire,UK
[2] University of Warwick, UK
{zinnar.ghasem,ingo.frommholz}@beds.ac.uk
carsten.maple@warwick.ac.uk

Abstract. Cyberstalking is becoming a social and international problem, where cyberstalkers utilise the Internet to target individuals and disguise themselves without fear of any consequences. Several technologies, methods, and techniques are used by perpetrators to terrorise victims. While spam email filtering systems have been effective by applying various statistical and machine learning algorithms, utilising text categorization and filtering to detect text- and email-based cyberstalking is an interesting new application. There is also the need to gather evidence by the victim. To this end we discuss a framework to detect cyberstalking in messages; short message service, multimedia messaging service, chat, instance messaging and emails, and as well as to support documenting evidence. Our framework consists of five main modules: a detection module which detects cyberstalking using message categorisation; an attacker identification module based on cyberstalkers' previous messages history, personalisation module, aggregator module and messages and evidence collection module. We discuss our ongoing work and how different text categorization and machine learning approaches can be applied to identify cyberstalkers.

Keywords: Cyberstalking, digital forensics, email filtering, data mining, cyberharassment, machine learning, text categorisation

1 Introduction

With the proliferation of the use of the Internet, cyber security has become a major concern for users and businesses alike. While communication technologies have undoubtedly positively changed the way we communicate, it also provides cybercriminals with methods and techniques to be used for illegitimate purposes such as the distribution of offensive and threatening materials [25], spamming, phishing, cyberbullying, viruses, harassment and cyberstalking [18]. Cyberstalking is a complicated and pervasive problem, which affects and targets a huge

In: R. Bergmann, S. Görg, G. Müller (Eds.): Proceedings of the LWA 2015 Workshops: KDML, FGWM, IR, and FGDB. Trier, Germany, 7.-9. October 2015, published at http://ceur-ws.org

number of individuals [4], and unlike many other cybercrimes, cyberstalking does not occur on a single occasion [24], rather victims experience repeated, systematic and multiple attacks. Cyberstalking has been identified as a growing social problem [7], and a global issue [13], to an extent in which it is envisaged that almost twenty percent of people at one stages of their lives will become a victim of cyberstalking, where women will more likely become a victim than men [11]. In [12] Maple *et al.* have defined cyberstalking as a "course of actions that involves more than one incident perpetrated through or utilising electronic means that cause distress, fear or alarm". There is evidence that cyberstalking will increase in both frequency and intensity [16].

While cybercriminals such as cyberstalkers utilise an array of technologies, tools and techniques like chat rooms, bulletin boards, newsgroups, instant messaging (*IM*), short message service (*SMS*), multimedia messaging service (*MMS*), and trojans, email is one of the most commonly used methods of cyberstalking [21, 13, 20, 15]. A cyberstalker can send emails, SMS, IM, MMS, and chat to threaten, insult, harass, or disrupt e-mail communications by flooding a victim's e-mail inbox with unwanted mail [23, 22] anywhere at any time anonymously or pseudonymously, without fear of prosecution. This creates a new challenge for law enforcement and in digital forensic investigation. Anonymity in communication is one of the main issues exploited by cybercriminals [10]. Therefore, cyberstalkers could easily disguise themselves by spoofing email, and creating different pseudonym accounts mostly from free web mail providers. Similarly web based gateways are utilised to spoof *SMS* [5], and different anonymous chat IDs are easily created.

These techniques, coupled with the availability of remailers, unauthorised networks, public library's computers, internet cafs, and free anonymous communications through websites, in addition to free and unregistered mobile SIM cards, inexpensive and unregistered mobile handsets, give an upper hand to cyberstalkers in their attack and complicate the investigation of cyberstalking cases [20]. Cyberstalking prevention with text messages filtering might not be as effective as required, because it does not always hold cyberstalkers accountable for their misuse of emails, and other text-based messages communication. Therefore identifying the original sender of emails, SMS, MMS, and chat is an important factor in the prosecution of an attacker [25].

The discussion so far shows that it is imperative to deal with cyberstalking on the very earliest stage. We will therefore in the remainder of the paper discuss a framework to detect cyberstalking in text-based messages and to support the collection of evidence for law enforcement. Text categorization and analysis plays a crucial role in our framework.

2 The Need to Detect and Document Text-based Cyberstalking

Text-based cyberstalking includes sending abusive, hate, threatening, harassing and obscene emails, SMS, chats, MMS, IM, including video and photo; sending email or MMS with the intention to spread viruses to a victim's device, either with attachments containing viruses or directing victims to a malicious website through a hyperlink; taking over victim's email account; sending high volumes of junk emails, SMS, MMS, Chat and IM. To minimise the effect of text-based cyberstalking, we propose a system that monitors, detects, captures and documents evidence. Such a system requires that we provide means to analyse textual documents like messages and gather some information from this analysis, for instance to determine the true authorship of emails when we cannot trust any email header information. We therefore discuss how text categorization and processing can be utilized for several aspects of this task.

Our work is inspired by [2] where the authors propose a system that simply records data within a session, that is duration of victim's computer connection to and disconnection from Internet. However, their system has major limitations in handling text-based cyberstalking. We therefore propose a framework to detect and filter messages and to collect and analyse evidence.

A further aim of our system is to assist the victims in documenting evidence for the initial complain process, as well as to help law enforcement in early stages of their investigation. In order to persuade authorities to investigate or prosecute a cyberstalker, the responsibility is often on the victim to produce such evidence [21, 3], thus, it is imperative that the victims save and keep all copies of communication whether email or other communications and with all their headers available and readable to be given to law enforcement [8, 14]. Such documentation will clearly demonstrate the course of incident and provide valuable information for both the investigation and prosecution process [17].

Therefore an automated system will not only make the initial complaint process and investigation easier but will also speedup investigations with less effort. Furthermore it will encourage victims to come forward and complain to prosecute a cyberstalker, because "cyberstalking and stalking's victim reporting is an important consideration for the criminal justice system, not only to guarantee that offenders are held accountable for their actions, but also to ensure that crime victims receive the support and services needed"[19].

3 The ACTS Framework

Our proposed framework is called *Anti Cyberstalking Text-based System (ACTS).* To the best of our knowledge it is the first framework that specialises on the automatic detection and evidence documentation of text-based cyberstalking. A prototypical implementation of the framework is under development, and the data collection process is ongoing. ACTS will, e.g., run on a user's device to

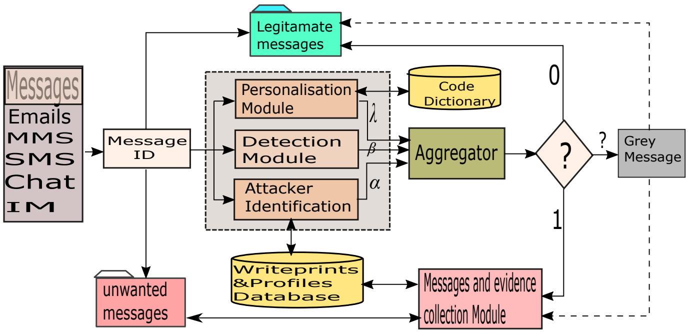

Fig. 1. ACTS Framework

detect text-based cyberstalking. The architecture of ACTS is depicted in Figure 1.

The proposed system utilises text mining, statistical analysis, text categorization and machine learning to combat cyberstalking. It consists of five main modules: detection, attacker identification, personalisation, Aggregator and messages and evidence collection.

ACTS first tries to detect cyberstalking based on a message ID list (the lists is optional for users and initially empty), which is automatically updated by the system. Messages whose IDs do not appear in the list are examined by the identification, personalisation, and detection modules; the results from three modules are passed to the aggregator for final decision.

Similar to some text categorization based email detection systems which can identify unwanted email, the *detection module* is employed to detect potential cyberstalking text based on their content. The received message is preprocessed by appying tokenisation, stop-word removal, stemming and presentation. Text mining techniques are utilised to extract required patterns from the message; a corresponding supervised algorithm like neural network/support vector machines is employed to detect and categorise emails to compute a value β based on three outputs, labelled as (00) not cyberstalking, (10) cyberstalking, and (01) grey email.

The *attacker identification module* is employed to identify whether received anonymous or pseudonymous messages are written by a cyberstalker or not, and to detect those messages from cyberstalkers where the message does not contain any known unwanted words. For this purpose, cyberstalker's writeprints including lexical, syntactic, structural and content-specific features [26] will be utilised. Unfortunately, due to character limitation of short messages like Twitter tweets, for e.g. SMS is limited to 160 $characters$ per message, writeprints might not always provide enough information to identify the author of the message. Nevertheless, because of their characters limitation, people tend to use unstandardised and informal language abbreviations and other symbols, which mostly depend on user's choice, subject of discussion and communities [9], where some of these abbreviations and symbols could provide valuable information in identify the sender. Thus to overcome this shortcoming and to enhance the identification process, we combines cyberstalker's writeprints with cyberstalker's profile including linguistic and behavioural profiles, utilising the existing cyberstalker's writeprints and profiles history in database.

Above considerations are based on the premise that, by definition, the victim must receive more than one attack to constitute cyberstalking. Thus there exist n messages where $n \in C_E\{m_1, ...m_n\}$ and $n \geqslant 2$ which will be used to check any new arriving message. Intuitively C_E will increase as the attack continues. A number of supervised algorithms have been used in authorship identification, where both authors and a set of their work are known prior to the identification process. Unfortunately, this is not the case in our approach. Identifying attackers is more challenging, firstly, because it will be implemented on user's device to detect messages, secondly, because we need to identify and detect the sender without prior knowledge. The system needs to decide whether a received message is written by a cyberstalker or not, thus supervised algorithms are not applicable. For this purpose, principal component analysis (PCA) could be employed to detect messages based on stylometrics and profiles; the new message's data is projected on PCA, and compared to the data matrix of all cyberstalking messages in C_E. The result is represented by the value α based on three outputs: not cyberstalking ($\alpha \geq r_2$), cyberstalking ($\alpha \leq r_1$) and grey ($r_1 < \alpha < r_2$). The α value is passed to the aggregator component. r_1 and r_2 are pre-defined threshold values in attacker identification (which have to be determined experimentally).

We also have to take into account that cyberstalking is often highly personalised; a bare word(s) or a phrase(s) of a message might have no inclination whatsoever towards bad feeling almost to anyone, but it might cause fear and distress to a cyberstalking victim. For instance, sending child birthday wishes may commonly be considered as positive, but not in case of somebody who lost their child or had undergone abortion. This complicates the process of developing a general tool to combat text-based cyberstalking. For this reason we define a *personalisation module* which is employed to enhance overall victim's control over incoming messages, where each victim can outline and define their own rule preferences. Therefore the personalisation module consists of rule based components and a code dictionary. The rule based component is optional, where rules are defined based on words, date and phrases provided by the user. A typical rule could be $if((date_A < today < date_B) \wedge (message\ contains\ "abc"))\ return\ true$. A code dictionary is created from sentiment and affect word(s) or phrases, which are commonly used in cyberstalking. Furthermore, the code dictionary could also be updated by the user. The received message would be preprocessed, for this purpose *k-shingling* [6] could be utilised. Where each k-length shingle is run against the dictionary, probabilistic disambiguation[1] is another possible method to be used; both the dictionary's returned result and rule-based result are represented by the value λ: either cyberstalking (1) or not cyberstalking (0) (when both returned results are negative).

The final decision whether a received message is cyberstalking or not is made in the *aggregator module*, utilising the outcome from the previous modules. α, β and λ are the final calculated result values for each individual received email by the attacker identifier and detection module, respectively. Messages are identified as either grey (?), cyberstalking (1) or not cyberstalking (0) based on $\psi(\beta,\ \alpha,\ \lambda)$ as follows

$$\psi(\beta,\ \alpha,\ \lambda) = \begin{cases} 0 & \text{if } (\beta = 00 \wedge (\ \alpha \geq\ r_2)\ \wedge (\ \lambda = 0)\ )\ , \\ 1 & \text{if } (\beta = 10 \vee \alpha\ \leq\ r_1\ \vee\ \lambda = 1)), \\ ? & \text{if } (\beta = 01 \wedge (r_1 < \alpha < r_2)\ \wedge\ \lambda = 0). \end{cases}$$

The final module is the *messages and evidence collection module*, which collects evidence from a newly arriving cyberstalking message, for instance, in the case of email the c source IP address or, if it is not available, the next server relay in the path, and the domain name (both addresses are automatically submitted to WHOIS and other IP geolocation website). The information with timestamp and email headers is saved for instance in the evidence database on the victims' device. The module also regularly updates and adds stylometrics, profiles and related information of the cyberstalking message to the database. Furthermore it will utilise statistical methods like multivariate Gaussian distribution and PCA to analyse the writeprint and profiles of cyberstalking, and text mining to extract similar features, attacker behavioural, greeting, farewell, etc, specifically between anonymous message and non anonymous message. The integrity and authenticity of a cyberstalking message, gathered evidence information whether saved on computer or through transmission are preserved using hash functions and asymmetric encryption keys.

4 Conclusions and Future Work

Combating cyberstalking is a challenging task, where technical solutions are a cornerstone in its prevention and mitigation. We therefore presented the ACTS framework that filters, detect and documents text-based cyberstalking. In the context of our framework we in particular discussed the potential use of textual machine learning approaches and discussed the difference to classical email categorization. The aim of our solution is not only to mitigate cyberstalking, but also to help victims in documenting evidence, which is required for law enforcement. Future work includes the implementation and evaluation of ACTS, in particular by measuring the effectiveness of the proposed text categorization methods in this emerging problem area.

References

1. A. Abbasi and H. Chen. Affect Intensity Analysis of Dark Web Forums. In *Intelligence and Security Informatics, 2007 IEEE*, pages 282–288. IEEE, 2007.

2. S. Aggarwal, M. Burmester, P. Henry, L. Kermes, and J. Mulholland. Anti-Cyberstalking: The Predator and Prey Alert (PAPA) System . In *Systematic Approaches to Digital Forensic Engineering, 2005. First International Workshop*, number iv, pages 195—-205. IEEE-CPS, 2005.
3. T. K. and et al Logan. Research on Partner Stalking: Putting the Pieces Together. *Lexington, KY: Department of Behavioral Science and Center on Drug and Alcohol Research, University of Kentucky*, pages 1–27, 2010.
4. M. Baer. Cyberstalking and the Internet Landscape We Have Constructed. *Virginia Journal of Law & Technology*, 15(154):153—-227, 2010.
5. A. Bose and K. G. Shin. On mobile viruses exploiting messaging and Bluetooth services. In *2006 Securecomm and Workshops*, pages 1–10. IEEE, 2006.
6. M. Chang and C. K. Poon. Using phrases as features in email classification. *The Journal of Systems & Software*, 82(6):1036–1045, 2009.
7. B. L. Ellison and Y. Akdeniz. Cyber-stalking: the Regulation of Harassment on the Internet. *Criminal Law Review*, 29:29–48, 2001.
8. J. Finn. A survey of online harassment at a university campus. *Journal of interpersonal violence*, 19:468–483, 2004.
9. J. M. Gómez Hidalgo, G. C. Bringas, E. P. Sánz, and F. C. García. Content based SMS spam filtering. In *Proceedings of the 2006 ACM symposium on Document engineering - DocEng '06*, pages 1–8. ACM, 2006.
10. R. Hadjidj, M. Debbabi, H. Lounis, F. Iqbal, A. Szporer, and D. Benredjem. Towards an integrated e-mail forensic analysis framework. *Digital Investigation*, 5(3-4):124–137, Mar. 2009.
11. D. A. Jurgens, P. D. Turney, and K. J. Holyoak. SemEval-2012 Task 2: Measuring Degrees of Relational Similarity. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1*, pages 356–364. Association for Computational Linguistics, 2012.
12. C. Maple, E. Short, A. Brwon, C. Bryden, and M. Salter. Cyberstalking in the UK: Analysis and Recommendations. *International Journal of Distributed Systems and Technologies*, 3(4):34–51, 2012.
13. A. Maxwell. Cyberstalking. Technical Report 7, Auckland University, July 2001.
14. R. Mccall. Online Harassment and Cyberstalking: Victim Access to Crisis , Referral and Support Services in Canada Concepts and Recommendations. *Canada: Victim Assistance Online Resources*, page 17, 2003.
15. E. Ogilvie. The internet and cyberstalking. (December), 2000.
16. N. Parsons-pollard and L. J. Moriarty. Cyberstalking: Utilizing What We do Know. *Victims and Offenders*, 4(4):435–441, 2009.
17. D. A. Pinals. *Stalking, Psychiatric Prespective and Practical Approach*. 2007.
18. K. Reynolds, A. Kontostathis, and L. Edwards. Using Machine Learning to Detect Cyberbullying. In *Proceedings ICMLA 2011*, pages 241–244. IEEE, Dec. 2011.
19. B. W. Reyns and C. M. Englebrecht. The stalking victim's decision to contact the police: A test of Gottfredson and Gottfredson's theory of criminal justice decision making. *Journal of Criminal Justice*, 38(5):998–1005, Sept. 2010.
20. D. Robert and J. Doyle. Study on Cyberstalking: Understanding Investigative Hurdles. *FBI Law Enforcement Bulletin*, 72(3):10–17, 2003.
21. L. Roberts. Jurisdictional and definitional concerns with computer-mediated interpersonal crimes: An Analysis on Cyber Stalking. *International Journal of Cyber Criminology*, 2(1):271–285, 2008.
22. L. L. Sheridan and T. D. Grant. Is cyberstalking different? *Psychology, Crime & Law*, 13(6):627–640, Dec. 2007.

23. C. Southworth, J. Finn, S. Dawson, C. Fraser, and S. Tucker. Intimate partner violence, technology, and stalking. *Violence against women*, 13(8):842–56, Aug. 2007.

24. J. L. Truman. *Examining intimate partner stalking and use of technology in stalking victimization.* PhD thesis, University of Central Florida Orlando, Florida, 2010.

25. O. D. Vel, A. Anderson, M. Corney, and G. Mohay. Mining E-mail Content for Author Identification Forensics. *ACM Sigmod Record*, 30(4):55–64, 2001.

26. R. Zheng, J. Li, H. Chen, and Z. Huang. A Framework for Authorship Identification of Online Messages: Writing-Style Features and Classification Techniques. *JASIST*, 57(3):378–393, 2006.