Security Informatics
a SpringerOpen Journal

RESEARCH Open Access

"Our Little Secret": pinpointing potential predators

Anna Vartapetiance* and Lee Gillam

Abstract

The word "Paedophilia" has come a long way from its Greek origin of child-companionship to a Mental Disorder, Social Taboo and Criminal Offence. Various laws are in place to help control such behaviour, protect the vulnerable and restrain related criminal offences. However, enforcement of such laws has become a significant challenge with the advent of social media creating a new platform for this old crime. This move necessitates consideration of approaches that are suited to this new platform and the way in which it affects the Cycle of Entrapment. This paper reviews definitions of, and features of, paedophilia and other related –philias, and sexual offences against children, and seeks through the understanding of these to determine where specific detection approaches are effective. To this end, we present our own detection approach which is geared towards predatory behaviours, which can be a precursor to sexual offences against children, and which directly references this Cycle of Entrapment. Our approach has shown early promise with an F1 score of 0.66 for training data but only achieving 0.48 for testing data on a collection of chat logs of sexual predators. The results were later improved to achieve an F1 score of 0.77 for train and 0.54 for test data based on the approach.

Keywords: Paedophile; Hebephile; Sexual offender; Predator; Detection

Introduction

Paedophilia, whilst perhaps historically of little consequence – perhaps even a socially acceptable form of entertainment [1] is now variously considered a Mental Disorder, a Social Taboo, and a Criminal Offence [2,3]. Various laws exist in various countries that aim to control or prevent such behaviours and protect the vulnerable. These laws rely on the ability to detect the occurrence of such behaviours, and in recent years this ability has been challenged by the emergence of social media. Social media has created a new platform for an old crime, challenging authorities in both applicability of laws and in possibilities of crime detection.

If we go back only as far as the early 1990s, predators who were unknown to their victims would have to approach them in real world settings, with concomitant risks of being identified, and prevented, by eyewitnesses. A mere 20 years on, and social media can facilitate much more ready access with rather lesser risk of eyewitnesses. The principal difference is one of familiarity: predators can get to know their prey in advance of the physical approach that was needed previously, making the approach much easier, and offering the predator the opportunity to control the surroundings of this physical encounter - either reducing the likelihood of eyewitnesses being present, or making the situation appears entirely normal. Further, it can be unclear whether the predator convinces the prey to take actions based on false beliefs or ill-perceived risks, or whether the predator is entirely open about their intentions and the prey is merely lulled into a false sense of security. In being able to control such situations, paedophiles, hebephiles, and others intent on commission of sexual offences against children seem to have a dangerously lessened risk of detection.

The principal aim of this paper is to present our understanding of paedophilia and related issues of hebephilia and sexual offences against children, and through these to appreciate what would be detectable in the predatory activities as might precede these. In Background section, we discuss the clinical and legal perspectives on these matters, and note how variation in age is a feature and that mainstream use of such labels can be inconsistent with such definitions. We note that predatory activities tend to involve a degree of effort on the part of the predator, and that the Cycle of Entrapment is where certain efforts

* Correspondence: a.vartapetiance@surrey.ac.uk
Department of Computing, Faculty of Engineering & Physical Sciences, University of Surrey, Guildford, UK

may be focussed. In Section Cyber-predators, we extend our discussion to the online world, noting the implications this brings for the Cycle of Entrapment and the difficulties of applying technological controls and also of how the blurring of geographies can create issues, and ways in which predators can use the online world. In Section Empowering investigators, we discuss technologies that can be deployed against such predators, as well as our own approach to detection around identifying requests for information that relate to the Cycle of Entrapment, and briefly conclude the paper in Section Investigating the possible gain from machine learning.

Background

In this section we offer a brief discussion of differences in definitions as relate to clinical (paedophilia, hebephilia, World Health Organisation and Diagnostic and Statistical Manual of Mental Disorders) and legal (laws on sexual offences) interpretation of sexual relationships with children. This leads on to an exploration of the kinds of predatory behaviours involved, and finally to the features which make the online world such an appealing place for those wishing to behave in such a way.

Paedophile, hebephile, or child sexual offender?

The word Paedophilia derives from the Greek words "child" (παιδί/paidí) and "Friendship/ Companionship" (φιλία/philía). But this historical derivation seems somehow inconsistent with it being a disorder of sexual preference according to the World Health Organisation (WHO) in International Classification of Diseases (ICD-10 [4]):

A. The general criteria for F65 Disorders of sexual preference must be met [which are outlined here]

G1. Recurrent and intense sexual urges and fantasies involving unusual objects or activities.
G2. Acts on the urges or is markedly distressed by them.
G3. The preference has been present for at least six months

B. A persistent or a predominant preference for sexual activity with a *prepubescenta*[a] child or children.
C. The person is at least 16 years old and at least five years older than the child or children in B.

The Diagnostic and Statistical Manual of Mental Disorders (DSM-IV_TR [5]) relates the persistent or predominant preference as *Exclusive* and *Non-exclusive*, depending on whether it involves sexual relations only with children or with adults as well [2,6]. Seto [2], p.14 further extends these to incorporate a more vivid description of the kinds of acts involved (underlined for emphasis):

> *"...sexual preference for* ***prepubescent***[b] *children when sexually mature partners are potentially available, whether it is reflected in recurrent self-reported thoughts, fantasies, or urges about sexual contact with children; exhibited in greater sexual arousal to stimuli depicting prepubescent children relative to stimuli depicting adults; or manifested in a pattern of sexual behaviour involving children."*

This, then, seems to involve either a desire to do something with prepubescent children, or the actual doing of those desired things – with the former potentially difficult to identify. This notion of prepubescent seems important. However, the label *"Paedophile"* seems to be used quite broadly by mainstream media, investigative organizations, and other agencies, to apply to anyone who has an interest in children (under the age of 16) or commits an offence against them. Interpreting such definitions would seem to require clarification of prepubescent. The age of puberty can vary from person to person, but a recent article in a paediatric journal suggested "patients ≦ 8years old (prepubescent), 8–13 (pubescent), 13–18 (post-pubescent)" [7]. It is possible that these age ranges suit the country and the study, but they appear unduly low and it may be more typical to distinguish amongst those sexually attracted to infants from 0-5 years, (called Nepiophilia or infantophilia), children younger than 11 (Paedophilia, which may incorporate Nepiophilia), ages 11–14 (Hebephilia) and 15-19 (Ephebophilia) e.g. [2,8].

It has been reported that the largest proportion of sexual offenses in the US occurs against 14 year old children (Figure 1) [9], which means these offenders are interested in early ***pubescent*** children (recall that Hebephilia is 11- 14), and this may not even be considered as a crime or disorder in some countries [8].

These labels on child-adult sexual relations at specific ages also do not necessarily cohere with national laws. For example, in the United Kingdom, sexual relations between an adult and a 15 year old girl are crimes according to the Sexual Offences Act 2003, which makes the key distinctions for ages at 13 and 16. An adult that falls foul of such a law with a 15 year old girl would, by the previous definitions, be considered neither a paedophile nor a hebephile but an ephebophile. However, the paedophile label will still be used by mainstream media[c]. For such purposes, Lanning's [10], p.18 definition of Child Sexual Offender/Child Molester seems more apt:

> *"... as a significantly older individual who engages in any type of sexual activity with individuals legally defined as children."*

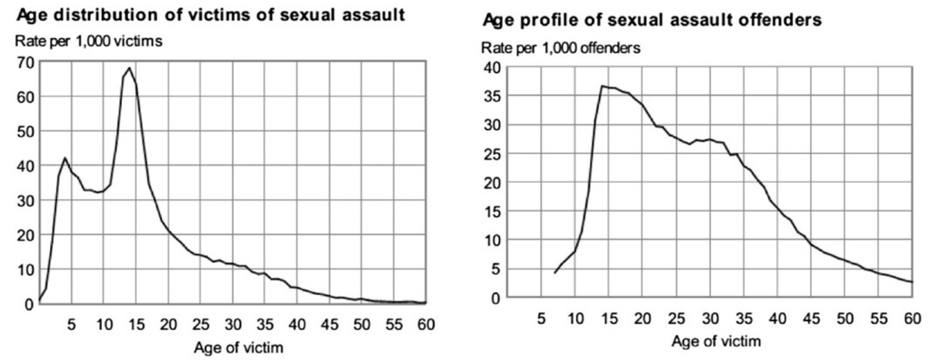


Figure 1 From Snyder [9] (Figures one and six): relation between sexual offences, victims and offenders.

Two notions are important here:

1. *"Significantly Older"*: Based on both DSM-IV-TR and ICD-10, an age difference of five years or more although this might vary in different jurisdictions.
2. *"Defined as child"*: "Age of Consent" defines who may be considered a child *in a given jurisdiction*.

But when sexually mature partners are potentially available, what compels adults to become sexually interested in and involved with children? Finkelhor & Araji [6] suggest four reasons for paedophilic behaviour.

1. Emotional congruence - the adult has an emotional need to relate to a child
2. Sexual arousal - the adult could only become sexually aroused by a child;
3. Blockage - alternative sources of sexual and emotional gratification are not actually available;
4. Disinhibition - the adult is not deterred from such an interest by normal prohibitions.

The nature of the need, then, is likely to influence the ways in which they seek to satisfy that need. Offenders can be further divided based on their approach [10,11]:

- Situational or Short Term Strategic Placement: the offender has weak motives (intensity) and might not have acted upon it before. There are chances for it to be accidental and opportunistic approaches (not planned).
- Preferential or Long Term Strategic Placement: the offender has strong intensity and persistence. The offender would usually try to place him/herself in a position to assure his/her access to children

Table 1 relates the above distinctions more clearly. However, it is important to note that such differences in predatory activity need not be clearly fit to one column or the other.

Relationship with victims (Children)

The relationship of a predator with a child can be either as a stranger, an acquaintance, or a familiar, described as follows:

Stranger: Sexual abuse of children where the person is unknown/ not well known to the child. Most such offenders do not have or want long term access to children nor previously build relationships with them. Therefore, in order to lure the children, they are more likely to use threats and physical forces [10].

Acquaintance: Sexual abuse of children where the person is known/ thought to be known by the child. These offenders usually build access to children, however they do not use violence to lure the children. They would spend time to create the relationship both to give them access to the child and decrease the likelihood of disclosure. Every acquaintance offender starts from being a "stranger" and then builds the abusing relation. Depending on the child's (victim's) age, the offender

Table 1 Relation between different characteristics defined for sexual offenders

Predatory activity		Reference
Situational predators	Preferential Predator	[10]
Short term strategic placement	Long Term Strategic Placement	[11]
Less exclusive	More exclusive	[2,5,6]
Less intensity	More Intensity	[2,4,5]
Less persistence	More Persistence	[2,4,5]

might need to build a relation with the parents as well to gain their trust and consequently access to the child. Lanning [10] identifies that seduction techniques for young children revolve around fun, games and plays while for older children it revolves around sexual arousal, curiosity, rebelliousness and inexperience.

Familiar: Sexual abuse of children where the person is within close family circle; e.g. step/father, step/brother, uncle, grandfather, female family members. These offenders usually have long term access to the children and they use their authority and status to control the children. As they are part of family, some of these relations may never be disclosed by the child.

Snyder ([9], Figure six) present the results of studies on victim-offender relationships in sexual assaults (no age limit) and shows that except for victims under age 6, most sexual assault offenders were acquaintances, with 60% in general across all ages.

Table 2 summarises our understanding across different classifications introduced in various research in regard to the three types of relationships. Note how these incorporate Situational and Preferential as discussed above and shown in Table 1.

Luring the prey

Olson et al. [11] introduces the theory of luring communication (LCT) based on grounded theory [13] to address the process of entrapment used by child sexual predators to lure their victims into an on-going sexual relationship (Figure 2); this is believed to be very similar to strategies used by rapist and stalkers [14].

They define the following phases which are of interest here:

- Gaining Access: This is possible through *strategic placement* of the predator, where they will have the chance to have access to children;
- Cycle of Entrapment: the core of the cycle is defined as deceptive trust development. The success of a predator in luring a child depends on its ability to build the trust. Oslon et al. [11] suggest that most predators create the trust by strategically placing themselves in authoritative position such as teachers, priests or coaches and they will engaging in relationship-building activities such as dating, buying them gifts and showing them attention and affection
  - Grooming: predator engages in sexually explicit conversation to desensitise them in order to secure the cooperation of the victim and reduce risk of discovery or exposure[d] [15].
  - Isolation: predator tries to isolate their victim, mentally and physically, from support networks; e.g. friends, family/parents, and guardians
  - Approach: the initial physical contact or verbal lead-ins that occur prior to the actual sex act

This process is followed with Communicative Responses to Sexual Acts that may, but may not, result in sexual abuse.

Discussion

In this section, we briefly discussed definitions as relate a clinical (paedophilia, hebephilia, ICD-10 and DSM-IV-TR) and a legal (sexual offenders) interpretation of sexual relationships with children. It is apparent that the label ‘paedophile’ is widely used even if inconsistently with related definitions. Further, that variations in national laws also leave room for interpretation. This is certainly not an argument *for* such activities, but careless use of such terms is not necessarily helpful in interpreting such matters and in identifying the kinds of approaches as are relevant to detection and prevention. In considering the kinds of predatory behaviours involved, we see that it takes *intensity* and *persistence* for a predator to *gain access* and so to move from *stranger* to

Table 2 Relationship between different types of predatory characteristics

Class	Characteristics			Discussed in
Relation	**Stranger**	**Acquaintance**	**Family**	[10,11]
Label	Grabbers	Groomers	Granters	[10,12]
Preferences	***Situational***	***Preferential***	***Both***	[10]
Strategic placement	Short term	Long term	Long term	[11]
Exclusivity	More Non-exclusive	Both	Both	[2,5,6]
Intensity	Less Intense	Intense	Intense	[2,4,5]
Persistence	Less Persistence	Persistence	Persistence	[2,4,5]
Dangerous to other children	Yes	Yes	Usually No	[10]
Children they approach	1 to many	1 to many	Usually one	[10]
Individuals involved	One-to-one, One-to-many	One-to-one, One-to-many	Usually one-to-one	[10]

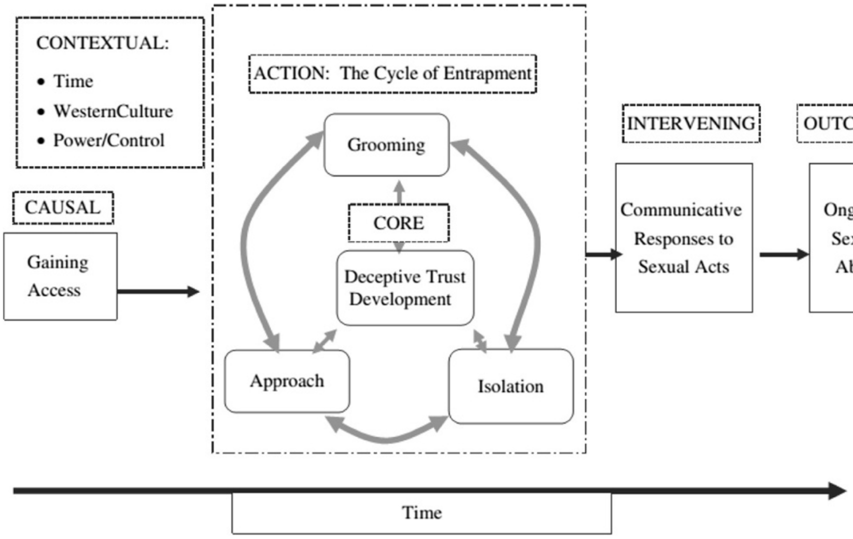


Figure 2 **A model of luring communication theory.**

acquaintance. Having gained access, the focus moves to the four important aspects of the Cycle of Entrapment. In the next section, we will discuss how the efforts previously needed to *gain access* are less in the online world, and hint towards effective detection being best undertake in the Cycle of Entrapment.

Cyber-predators

Computer Mediated Communications (CMCs) and Cyberspace undermine the traditional relationship between physical context and social situation. What is introduced to provide users with new, easy and cheap means to do things they used to do before also offers new ways to commit old crimes - e.g. money laundering, identity theft and child sexual abuse. In most cases, moves online are accompanied by redefinition of extant words such that a Facebook "friend" is somewhat different to a real friend, and "sharing", "stalking", and "grooming", can have different interpretations also [16,17].

Internet communication brings much autonomy: people can be whoever, whatever and wherever they wish. The downside is that such autonomy can be a ready cover for vice. In the present context, it can be very easy for a predator to *gain access* to children in chat rooms to have a "private chat" and become "friends" while, initially, hiding their real intentions. Moreover, social networks make children even more accessible, with their ease of posting of personal information, including their location, also making it potentially easier for predators to *approach* them [16]. If they include a *stranger* into their network, the transition to an *acquaintance* is already apparent – and there is also the possibility based on this to *gain access* to other *acquaintances*.

A YISS-2[e] survey, conducted on 1,500 youth Internet uses (age 10 to 17) by the American National Center for Missing & Exploited Children (NCMEC) and the Office of Juvenile Justice and Delinquency Prevention (OJJDP) [18] on Online Victimization of Youth revealed that [19]:

- 13% (1 in 7) of youth has been exposed to unwanted sexual solicitation online
  - 4% asked them for nude or sexually explicit photographs of themselves
  - 14% of such solicitors were acquaintances
  - 31% of such solicitors were aggressive - where solicitors made or attempted to make offline contact with youth
    - 75% of them asked to meet the youth in person
    - 34% called youth on the telephone
    - 18% came to youth's home
    - 12% gave youth money, gifts, or other items
    - 3% bought travel tickets for youth
- 34% (1-3) were exposed to unwanted sexual material; an increase of 9% over YISS-1[f] despite increased use of filtering, blocking, and monitoring software in households of youth Internet users (from 33% in YISS-1 to 55% YISS-2).
- 9% (1 in 11) were harassed, with threatening or other offensive behaviour
- 34% communicated online with people they did not know in person.

- 11% formed close online relationships with people they met online.
- Only 5% of solicitations and 9% of unwanted exposures were reported to law enforcement, Internet service providers, or other authorities.

Most of those affected do not have the digital skills and knowledge to protect themselves[g] and it appears that law enforcement agencies also lack the resources to address it. Organizations such as the Virtual Global Taskforce (VGT) [20], the Child Exploitation and Online Protection centre (CEOP) [21] and Internet Watch Foundation (IWF) [22] in the UK, and many more, have been created to address such problems, and even more laws have been crafted such as EU Directive 2011/92/EU[h] and UK acts such as Sexual Offence Act 2003 and Children Act 2004. However, such laws assume predator detection, without which enforcement will be difficult. Such detection faces major challenges: (1) cyberspace empowers the predators by ease of access; (2) access can occur in private online settings; (3) law enforcement agencies may not have access to these private online settings; (4) the volume of communication in general is large, and monitoring all of them for a relative few such problems is non-trivial. Here, we address only (1) in how cyberspace can empower the predators. Addressing (2) and (3) would require technological considerations, and (4) is a challenge of scale; all of these are beyond the scope of this paper.

Empowering predators?

In the recent past, children would spend a lot of time outside. Predators would need to approach these children physically. Some might suggest that fears of such predators have led to children now spending most of their time indoors. Whilst indoors, these children have ready access to social media. Ironically, this can increase access to children, because:

- They no longer need to be physically present in a certain location in order to contact children
- They can target more than one child at a time, especially in chat rooms
- Cyberspace bypasses parental supervision
- There is the possibility to avoid eyewitnesses
- Predators can assume any persona - e.g. age, gender and image - and craft elaborate and apparently exciting stories as they wish

The real world places limits on time, space and communications, where a person/predator can only be in one place at one period of time having conversations with people who share the same environment. Cyberspace compresses all of these, providing predators with the ability to have parallel conversations with children/other predators in different places at the same time; giving possibilities for one-to-many solicitations. A further complication in terms of space is that laws are still supposed to respect geographies. Can an adult from Spain be arrested for solicitation with a 15 year old girl from Portugal - the age of consent at the time of writing was 13 in Spain and 16 in Portugal – and what if they meet in Spain? How about an adult from Nevada and a 17 year old girl in any of the neighbouring states – age of consent in Nevada is 16 while in neighbouring states is 18. Again, does it depend where they meet?

The predators are able to use the Internet in at least four ways [23]: (1) to locate children; (2) to engage in sexual communication with children (3) to exchange materials; e.g. stories and child pornography (4) to communicate with others predators; e.g. for self-validation or additional information. Related to these, Hall & Hall [24] label predators as:

1. Stalkers[i]: approaching children in chat rooms in order to get physical access to them
2. Cruisers: interested in inappropriate sexual communication and file exchange with children but not with an intention to meet them offline
3. Masturbators: watchers of child pornography
4. Swappers: trading information, stories and pornography

Although all of the classes and factors mentioned above may result in sexual victimisation of a child or a criminal offence, it does not automatically follow that it will. However, it would be undesirable to test such a notion. In fact it would be desirable to detect the signals irrespective of harm. It is apparent that predators may be empowered in cyberspace, and so in the next section we look at ways in which prevention and investigation can be empowered.

Empowering investigators

Detection of predators, then, appears to become split across (1) detection of child pornography and identification of the people involved; creators, distributors, websites, etc. and (2) detection of predators who are attempting to meet children offline for sexual purposes (Penna et al. [25]). The first may entail and require investigation of prior offences, but the second also has a goal of prevention. It is important also to highlight the difference in (2) between merely detecting predatory communications and identifying the perpetrator. Perhaps the most effective way to affect an arrest of such predators is the so-called “sting” in which law enforcement officers, and even volunteers, are trained to pose as children – usually in chat rooms. One such voluntary organisation is the Perverted Justice Foundation [26], a non-profit organization where the volunteers, imitating children, attract predators. They claim to have helped to convict 550 predators since 2004.

State-of-the-art

Since manual tracking is a labour intensive activity, software that implements parental controls may help. But this needs to be installed and appropriately configured on all devices. Some software even claims to flag potential predators alongside cyber-bullies. For example, Net Nanny [27] claims to have an "anti-predator phrase list", perhaps similar to that of ContentBarrier [28] which apparently flags on phrases such as: "are you alone"; "believe me"; "can i see you"; "can we meet"; "come alone". Of course, such a detection approach will only work if predators use these phrases specifically. Also, not every parent is sufficiently technologically knowledgeable to use the right software to prevent predators from gaining access to their child[j] and one would have to wonder about responsibility and legal liability in the event that a predator is not detected despite such software being deployed.

Each software installation would need to update such lists regularly, leading to inconsistencies in what was detectable each time. And this leads towards such detection being undertaken at the network layer by internet service providers using similar natural language processing tools and techniques for predatory language/keyword profiling [29].

There have been various approaches taken to cyber-predator detection, many using the dataset provided by Perverted Justice. These approaches can be divided into:

- Group A: Distinguishing between predators and victims/ children
- Group B: Identifying inappropriate chat-conversations with victims/ children
- Group C: Identifying the grooming/ predator

Pendar [30] addresses Group A by removing stop-words, generating word unigrams, bigrams and trigrams and using a Support Vector Machine (SVM) and k-NN to achieve an f-of 0.943. RahmanMiah et al. [31] address Group B through three classes of chat: (i) **Child Exploitation:** adult-child sexual conversation; (ii) **Sexual Fantasies:** adult-adult sexually explicit conversation; (iii) **General:** conversations with no sexual content. They combine text categorisation, category information provided by LIWC (Linguistic Inquiry and Word Count) and a Naïve Bayes classifier [32,33]. Unlike Pendar [30]), they do not use pre-processing or spellchecking. McGhee et al. [34] use a rule based approach and k-NN - achieving 83% accuracy - for Group C, labelling conversations – partially similar to Olson et al. [11] as (i) Exchange of personal information; (ii) Grooming; (iii) Approach; (iv) None of the above. Similarly for Group C, Michalopoulos and Mavridis [35] used decision-making methods and Naïve Bayes after removing stop-words and applying a spelling correction strategy, to achieve 96% accuracy, based on: (i) **Gaining Access**: predators intention to gain access to the victim/child; (ii) **Deceptive Relationship**: the deceptive relationship that the predator tries to establish with the minor, as a preliminary to a sexual attack (as mentioned in [10,11]) and (iii) **Sexual Affair**: indicates the predator's intention for a sexual affair with the victim/ child.

Other related research tends towards variations on these approaches; Bogdanova et al. [36] addresses both predator and the conversation using psychological cues and sentiment analysis approaches with Naïve Bayes; Strapparava & Mihalcea [37], McGhee et al. [34], Argamon et al. [38], etc. Peersman et al. [39] use the combination of features from Lanning [10]) and McGhee et al. [34] to create predator dictionaries and apply:

> *"… both a resampling and a filtering strategy. More specifically, we trained a post-level classifier based on a balanced subset and a classifier on the user level based on a filtered subset of the training data .We then combined the output of these two systems and imposed conversation-level constraints that significantly improved the quality of the output."*

Research on topics related to detection of predators in chat-logs includes:

- Detecting child pornography in peer-to-peer networks: e.g. Hughes et al. [16]
- Detecting conversation topic: e.g. Adams and Martell [40]
- Detecting harassment: Yin et al. [41]
- Authorship Attribution: e.g. Juola [42]
- Authorship Profiling (detecting age and gender): e.g. Peersman, Vaassen, Asch and Daelemans [39])

PAN 2012

In 2012, a workshop on "Uncovering Plagiarism, Authorship, and Social Software Misuse" (PAN) introduced the challenge of Sexual Predator Identification. A set of chat logs were provided for participating research teams against which to evaluate systems. Two different tasks were involved:

- Task 1: Identify the predators among all users in the different conversations.
- Task 2: Identify the part (the lines) of the conversations, which are the most distinctive of the predator behaviour.

The dataset comprises chat logs and includes real world predatory scenarios amongst other (non-predatory) exchanges[k]. This is a large collection (~358K lines) with following properties:

- Few True Positives (real conversations with a potential "predator") - 4% only: selected from chat-logs provided on the Perverted Justice Foundation website (PJ) – which publishes logs of online conversations between convicted predators and volunteers posing as underage teenagers.
- Large number of potential False Positives (people talking about sex or shared topic with the "predator"). The organisers used Omegle [43] repository - which presents a random sample of more than 1 million anonymous conversations of strangers – as it contains "abusive language and general silliness online" and sometimes users "engage in cybersex" [44].
- Large number of Negatives (general conversations). Regular conversations collected from IRClog [45] and Krinj [46].

Data are divided into Training and Testing sets, with 30% of the collection for Training comprising 66,927 conversations with no more than 150 lines each. 291 of these 66,927 involved 142 unique predators. The Test Corpus comprises 155,128 conversations with more than 150 lines each, 440 of which involve 254 unique predators (task 1). Table 3 below presents the details of the created dataset.

The data are presented in XML as shown in Figure 3.

For Training, predator IDs were known (task 1) but the predatory elements of the chat were not identified (task 2). For the latter, the organisers assessed 113,888 lines submitted by participants and identified 6478 that *they* considered demonstrated predatory behaviour.

Surrey detecting sexual predators at PAN12

We participated in this PAN challenge [47] to explore the patterns of offender behaviour. Having never attempted the analysis of such a corpus previously, our approach – described below - was largely built around heuristics relating to commonality of requests for key personal information, with similarities in the type of request but variation in the wording of the request.

Table 3 Properties of the collection by data provider

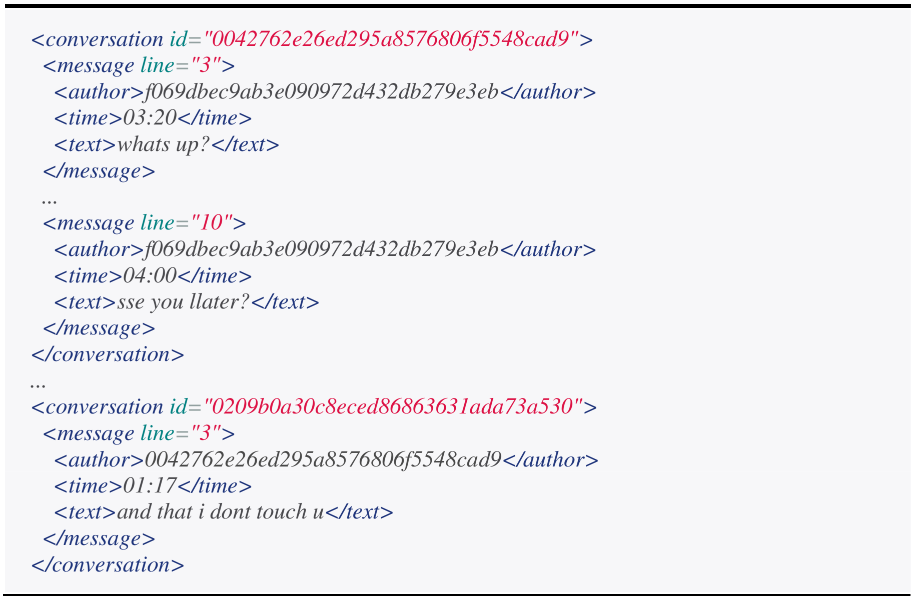
	PJ	Omegle	IRClog	Krjin
#conversations	11350	267261	28501	50510
conv.length ≤ 150	9076	265747	21896	48569
Training Set				
conv. length ≤ 150	2723	43064	6569	14571
Unique user (perverted)	291(142)	84131	10613	2660
Test Set				
#conv.length ≤ 150	5321	100482	15327	33998
Unique user	440 (254)	196130	17788	4358

Samples selected from the training corpus showed that the following four classes of information request appeared to be common, and led us to produce sets of indicators for accepting or rejecting passages (Table 4); this offers the possibility to filter the chat logs to identify potential predators.

- Address (*Approach*): Most ask for the address of the house or somewhere nearby to travel to.
- Parents (*Isolation*): Questions about parents are usually because of:
  - Secrecy
    - Making sure children are unsupervised while chatting
    - Making sure the chat history will be deleted later
    - Saying nothing to their parents
  - Seclusion
    - How isolated the child usually is? Relation with family members
    - To determine whether parents are around
    - To ascertain how long they would be gone for
- Age (*Deceptive Trust Development*): Some predators might lie about their age. Interestingly, most of them can be quite open about their age. They would usually highlight the fact that they are older, wishing the child were older, the fact that they might end up in a jail or trouble because of chatting with an underage children and so on, and so an expectation on the child to keep "Our Little Secret".
- Activities (*Grooming / Approach*): References to sexual activities. They usually focus more on the concept of meeting and having fun, watching TV and listening to music. But these conversations can be shifted depending on the child's age as in the chat with adolescents the conversations are more explicit.

These classes correspond with the Cycle of Entrapment, discussed earlier [11] which includes Deceptive Trust Development, Grooming, Isolation, and Approach. We note again, as before, that Deceptive Trust Development is usually considered by reference to these other phases.

In the training corpus, there were phrases solely used to emphasise this, e.g. "u know I would go to jail if some on figures out", "this is out little secret", "I can get in trouble talking to u". In all of these, there is an attempt to create a fake trust. We denote all phrases as belonging to the "Age" class as most refer to the age difference. We also consider some Activities to cover both



```
<conversation id="0042762e26ed295a8576806f5548cad9">
  <message line="3">
    <author>f069dbec9ab3e090972d432db279e3eb</author>
    <time>03:20</time>
    <text>whats up?</text>
  </message>
  ...
  <message line="10">
    <author>f069dbec9ab3e090972d432db279e3eb</author>
    <time>04:00</time>
    <text>sse you llater?</text>
  </message>
</conversation>
...
<conversation id="0209b0a30c8eced86863631ada73a530">
  <message line="3">
    <author>0042762e26ed295a8576806f5548cad9</author>
    <time>01:17</time>
    <text>and that i dont touch u</text>
  </message>
</conversation>
```

Figure 3 Structure of the corpus data.

Grooming and Approach. As mentioned, Lanning [10] identifies that seduction/grooming techniques for young children revolve around fun, games and plays while for older children it revolves around sexual arousal, curiosity, rebelliousness and inexperience. On the other hand, Olson et al. [11] identify Approach as "the initial physical contact or verbal lead-ins that occur prior to the actual sex act." Hence, we labelled the Activity class as "Grooming/Approach" to fit both definitions, but retained Address information separately. Examples of phrases involved (accept) in each information request, and exceptions (reject) are shown in Table 4.

System design

The approach is relatively straightforward, was implemented using a variety of (Linux) shell scripts, and appeared to offer good performance on the Training Corpus (up to f1 of 0.66 for training). We removed the XML markup and structured the data by Author ID and Conversation ID as shown in Table 5. We did not use any pre-processing of date information, and used the chats as they were presented; including those where volunteers posed as children (pseudo-victims).

The system algorithm is presented below by Table 6 and Figure 4.

Our process of Sexual Predator Identification can be explained as:

1. For all the $q \in Q$, Store the Lines of the Corpus L where the Text column T_W includes words from Accept List of one of the categories C_A^X and not C_R^X in file *Data*

Table 4 Contents of accept and reject classes

Categories	Accept/reject size		Samples of accept/reject content
Address	Accept	13	Different spelling combination of following words: "your addres", "ur addres", "the addres"
	Reject	78	IT and social networking related topics such as URL address, Facebook address, email address, IP address
Parents	Accept	11	Different spelling combination of following words: "your mom", "your dad", "your Parent"
	Reject	26	Reference to parents' objects or characteristics such as "Ur dads car", "Your mom is nice, young"
			Reference to technical terms such as "parent class"
Age	Accept	11	Different spelling combination of following words: "you are young", "get in trouble", "underage", "to jail", "wish you were"
	Reject	33	Self-reference such as "I'm underage"
			Reference to the others such as sister, brother, friend
			Excluding, "wish you were here /with me"
Activities	Accept	6	Different spelling combination of following words: "go down on you", "make you come"

Table 5 Data structure after removing XML tags

Lines	Conversation ID	Message line	Author ID	Time	Text
1	0042762e26ed2	3	f069dbec9ab3	03:20	whats up?
2	0042762e26ed2	10	f069dbec9ab3	04:00	sse you llater
3	0209b0a30c8ec	3	0042762e26ed	01:17	and that i dont touch u

2. For Task 1: Select the *Unique* (*Author$_{ID}$*) of the authors whose count of occurrence was more than or equal the defined *Confidence Threshold*
3. For Task 2: Select the Lines of the Corpus *Data* where the count of *Unique (Author$_{ID}$)* is more than the *Threshold*

Training approach

We tested all information request types (Categories) mentioned in Table 4 individually, varying the confidence threshold (CT) required for a detection, and also in various combinations. The results are presented in Table 7 that shows the number of predatory lines “Flagged” per experiment (task 2), the number of “Unique” predator IDs (task 1), the correct detections (True Positives, TP), incorrect one (False Positives, FP) and False Negatives (FN), followed by values for Precision, Recall and F1.

We could detect 113 out of 142 predators using two or more occurrences of all four indicators (**). Moreover, we tested each indicator individually - varying the number of occurrences and in various combinations - to analyse the importance of each. Although all four classes play an important role, the combination of two or more occurrences of Parents and Address classes together correctly flagged at least 105 predators; showing the information need in these two.

Testing - competition results

For the Test Corpus, we used the combination of all four categories that occurred twice or more, as this offered the optimal f1 score on the Training Corpus (precision = 0.7, recall = 0.62 and F1 = 0.66). However, for test data this did not perform as well (precision = 0.62, recall = 0.39 and F1 = 0.48) and it could be argued that for real detection the false negatives would be of particular concern making focus on recall rather more important than precision (Table 7, *).

Table 6 Table of notations

Symbol	*Meaning*
Q	Set of Queries
q	A single query where $q \in Q$
L	The single Line from the Corpus ∈{*Conversattion$_{ID}$, Message #, Author$_{ID}$, Time, T_W*}
Author$_{ID}$	The unique ID for each Author presented in Author ID column
T_W	The content presented in Text Column
C	Set of Categories where C ∈ {*Address, Parent, Age, Intention*}
C^X	A single Category where $X \in C$
C^X_A, C^X_R	C^X for Accept (A) and Reject files (R)
Threshold	The Value defined as the confidence threshold for flagging someone as predator

Results suggest, then, that predators do use patterns that would fit with the Cycle of Entrapment, so it may be possible that such a system can generalise. However, it also appears that accumulation of cues during conversations is important. This finding, and others like it, are perhaps not readily tested in the wild.

Post competition evaluation

Our participation in PAN12 demonstrated that the chat logs did indeed bear evidenced of Cycle of Entrapment related communications, and our own post-competition analysis highlighted the need for a more comprehensive coverage of each aspect of the Cycle, not least of which involves coverage of predatory chats that are more sexually explicit in Activities (Grooming / Approach).

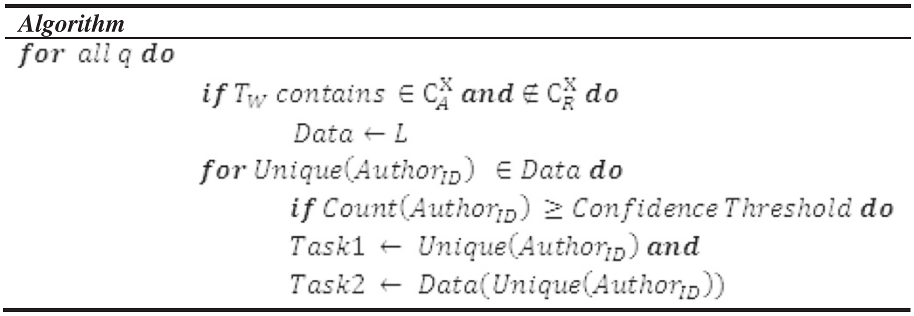
Initially, we improved our approach through contrastive frequency analysis, which increased the number of attributes in each category to those shown in Table 8. This improved the number of True Positives, and so the Recall.

Using a confidence threshold, we can increase the F1 score from 0.66 to 0.74 for Train, which would have achieved an increase for F1 from 0.48 to just 0.52 for Test. The Confidence threshold is, simply, a requirement for 3 occurrences or more of predatory behaviour (Table 9, row 6, 7 and 11, 12[l]). The best recall, i.e. retrieving the majority of predatory conversations, could be achieved absent such a threshold, with Recall increased from 0.8 to 0.94 but with 683 False Positives which might not be ideal for investigatory purposes (Table 9, row 2); F1 gives us the trade-off.

To reduce False Positives, we looked to filtering based on:

1. Removing all conversations centred on computer-related conversations. (135 additional attributes)
2. Removing sex chats which do not contain other Cycle of Entrapment indications (47 additional attributes)

These two are significant distractors in the PAN2012 corpus (based on Sections PAN 2012). Filter 1 leads to an increase in Precision, Recall and F1 (Table 9, row 7). Filter



```
Algorithm
for all q do
        if $T_W$ contains $\in C_A^X$ and $\notin C_R^X$ do
                Data ← L
        for Unique($Author_{ID}$) ∈ Data do
                if Count($Author_{ID}$) ≥ Confidence Threshold do
                Task1 ← Unique($Author_{ID}$) and
                Task2 ← Data(Unique($Author_{ID}$))
```

Figure 4 **Algorithm of our system for PAN2013.**

Table 7 Results of experiments

# of occurrence	*Flagged*	*Unique*	*TP*	*FP*	*FN*	*Precision*	*Recall*	*F1*
Address cues category								
Once or more	159	117	58	59	84	0.5	0.41	0.45
Twice or more	74	33	28	5	114	0.85	0.20	0.32
Three times or more	18	9	8	1	134	0.89	0.06	0.11
Parents cues category								
Once or more	440	255	84	172	58	0.33	0.59	0.42
Twice or more	257	72	49	24	93	0.68	0.35	0.46
Three times or more	151	38	32	6	110	0.84	0.23	0.36
Age cues category								
Once or more	124	88	33	55	109	0.38	0.23	0.29
Twice or more	62	25	17	8	125	0.68	0.12	0.20
Three times or more	21	10	9	1	133	0.90	0.06	0.12
Intentions cues category								
Once or more	39	35	14	21	128	0.40	0.10	0.16
Twice or more	8	5	4	1	138	0.80	0.03	0.05
Three times or more	0	0	0	0	0	0	0	0
Combining two cue categories of address and parents								
Once or more	598	333	105	228	37	0.32	0.74	0.44
Twice or more	366	101	74	27	68	0.73	0.52	0.61
Three times or more	217	53	46	7	96	0.87	0.32	0.47
Combining three cue categories of address, parents and age								
Once or more	722	388	112	276	37	0.29	0.79	0.42
Twice or more	458	124	85	39	57	0.69	0.60	0.64
Three times or more	280	69	58	11	84	0.84	0.41	0.55
Combining all four								
Once or more	761	410	113	297	29	0.28	0.80	0.41
**** Twice or more**	**478**	**126**	**88**	**38**	**54**	**0.70**	**0.62**	**0.66**
Three times or more	298	72	62	10	80	0.86	0.44	0.58
Test Corpus								
***** Twice or more	630	160	99	61	155	0.62	0.39	**0.48**

Table 8 Comparing original and improved contents of accept and reject classes

Category	*Accept/reject attributes*	*Class*	*No.*	*TP*	*FP*	*FN*	*Precision*	*Recall*	*F1*
Address	Original	Accept	13	58	59	84	0.5	0.41	0.45
		Reject	78						
	Improved	Accept	27	92	202	50	0.31	0.65	0.42
		Reject	82						
Parent	Original	Accept	11	84	172	58	0.33	0.59	0.42
		Reject	26						
	Improved	Accept	20	117	399	25	0.23	0.82	0.36
		Reject	22						
Age	Original	Accept	11	88	33	55	0.38	0.23	0.29
		Reject	33						
	Improved	Accept	37	84	186	58	0.31	0.59	0.41
		Reject	34						
Activities	Original	Accept	6	35	14	21	0.40	0.10	0.16
		Reject	0						
	Improved	Accept	51	61	132	81	0.32	0.43	0.36
		Reject	20						
Overall	Original	Accept	41	113	297	29	0.28	0.80	0.41
		Reject	137						
	Improved	Accept	135	134	683	8	0.16	0.94	0.28
		Reject	296						

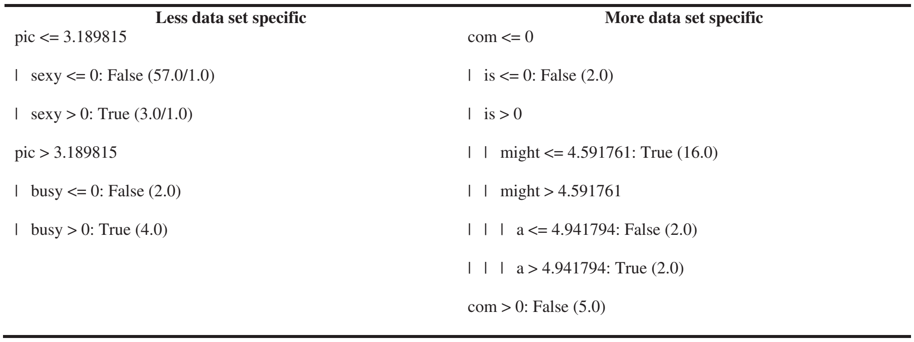
2, and a combination of Filter 1 and Filter 2 did tend to filter some of the sexually explicit predatory chats as well (Table 9, row 8, 9), so require further adaptation. The computer-related conversations filter offered greatest promise.

Investigating the possible gain from machine learning

Such an approach, whilst supported by theory, has been criticised as "Too Simplistic" for not using machine learning approaches for detection. However, we argue

Table 9 Results for improvements on our PAN2012 approach

	Filter	*Categories*	*CT*	*TP*	*FP*	*FN*	*Precision*	*Recall*	*F1*
Results from Train Corpus based on highest recall									
1.	Original	All 4	1	113	294	29	0.28	0.8	0.41
2.	Improved -no Filter	All 4	1	134	683	8	0.16	0.94	0.28
3.	Computer	All 4	1	131	529	11	0.2	0.92	0.33
4.	Sex	All 4	1	130	540	12	0.19	0.92	0.32
5.	Combined	All 4	1	127	399	15	0.24	0.89	0.38
Results from Train Corpus based on highest F1									
6.	Original	All 4	2	88	38	54	0.7	0.62	0.66
7.	Improved - no Filter	All 4	3	114	53	28	0.68	0.8	0.74
8.	Computer	All 4	3	111	34	31	0.77	0.78	**0.77**
9.	Sex	All 4	3	99	47	43	0.68	0.7	0.69
10.	Combo	All 4	3	98	30	44	0.77	0.69	0.73
Results from Test Corpus based on highest F1									
11.	Original	All 4	2	99	61	155	0.62	0.39	0.48
12.	Computer	All 4	3	115	57	139	0.67	0.45	**0.54**



Less data set specific	More data set specific
pic <= 3.189815	com <= 0
| sexy <= 0: False (57.0/1.0)	| is <= 0: False (2.0)
| sexy > 0: True (3.0/1.0)	| is > 0
pic > 3.189815	| | might <= 4.591761: True (16.0)
| busy <= 0: False (2.0)	| | might > 4.591761
| busy > 0: True (4.0)	| | | a <= 4.941794: False (2.0)
	| | | a > 4.941794: True (2.0)
	com > 0: False (5.0)

Figure 5 Sample of attributes selected by Weka and decision tree made using them.

that machine learning approach can only be successful if there is a strong theory underlying it or if it leads to theory generation. Otherwise, as is too often seen, results lack explanation[m]. In this section, we address the potential for machine learning to:

1. Improve detection based on our approach
2. Produce results independent of our approach, which might subsequently be interpreted with respect to theory.

We use the popular Java-based Weka software, from the University of Waikato, with default settings for each approach -if not otherwise mentioned, to evaluate this potential.

A decision tree for attributes: J48 classifier

As a first step, we let the Weka select attributes (using StringToWordVector, using 500 word fields and TF-IDF), and use the J48 classifier to identify suitable decision (TF-IDF) values for chat conversations per author (both with and without any pre-processing or filtering). We observe for this approach that (1) there is no improvement over our approaches (Table 9, row 8). (2) Branching values are very specific and may cause difficulties for generalisation (Figure 5).

A decision tree for categories: J48 classifier

The J48 classifier was next tested against classes extracted following computer conversation filtering, using three approaches:

1. Categories only: A profile of author based on the categories and related class (predator, or not) was created. For Train, this resulted in a tree with 11 leaves and size of 21, which achieved an F1 of 0.67 (Table 10, row 2).
2. Occurrences only: Since predatory chats might extend across many conversation lines, and not all of the categories might be satisfied, there might be enough occurrences still to classify it as predatory chat. Therefore, we consider the total number of occurrences of attributes in each categories (Sum)

Table 10 J48 classifier's results

	Experiment	*Leave*	*Size*	*TP*	*FP*	*FN*	*Precision*	*Recall*	*F1*
Results from Train Model									
1.	Raw data per Author	40	79	73	56	69	0.57	0.51	0.54
2.	Categories	11	21	86	28	56	0.75	0.60	0.67
3.	Occurrences	3	5	103	32	39	0.76	0.73	0.74
4.	Categories & Occurrences	10	19	104	25	38	0.81	0.73	**0.77**
Results from applying Train Model on test									
5.	Categories	11	21	101	33	153	0.75	0.40	0.52
6.	Occurrences	3	5	115	57	139	0.67	0.45	**0.54**
7.	Categories & Occurrences	10	19	102	33	152	0.76	0.40	0.52
Results from Test Model									
8.	Categories	9	17	92	20	162	0.82	0.36	0.50
9.	Occurrences	3	5	115	57	139	0.67	0.45	**0.54**
10.	Categories & Occurrences	4	7	93	22	161	0.81	0.37	0.50

```
                Train (leaves = 10 , size = 19)                                Test (leaves = 4 , size = 7)
Sum <= 1: No (97476.0/23.0)                                  Parents <= 0: No (218163.0/130.0)
Sum > 1                                                      Parents > 0
|  Sum <= 2: No (63.0/8.0)                                   |  Age <= 0
|  Sum > 2                                                   |  |  Address <= 0: No (430.0/34.0)
|  |  Sum <= 6                                               |  |  Address > 0: Yes (29.0/6.0)
|  |  |  Address <= 0                                        |  Age > 0: Yes (75.0/8.0)
|  |  |  |  Sum <= 4
|  |  |  |  |  Age <= 0: No (6.0)
|  |  |  |  |  Age > 0
|  |  |  |  |  |  Parents <= 0
|  |  |  |  |  |  |  Age <= 2: Yes (3.0/1.0)
|  |  |  |  |  |  |  Age > 2: No (6.0)
|  |  |  |  |  |  Parents > 0: Yes (7.0/2.0)
|  |  |  |  Sum > 4: Yes (12.0/2.0)
|  |  |  Address > 0
|  |  |  |  Address <= 4: Yes (56.0/11.0)
|  |  |  |  Address > 4: No (3.0)
|  |  Sum > 6: Yes (52.0/3.0)
```

Figure 6 Different decision trees for categories & occurrences experiment, train vs test.

which would indicate a confidence threshold. As shown in (Table 7, row 3), have number of occurrences only improves the F1 value to 0.74

3. Categories and Occurrences: A combination of 1 and 2. As shown in (Table 10, row 4,), results equal those already obtained by our improved approach (Table 9, row 8)

We used the classification models developed from the Train Corpus on the Test Corpus, resulting in an F1of 0.54 for "Occurrences only". To quickly evaluate the ability to generalise, we used J48 on Test, and assessed tree similarity (Figure 6); however, this did not indicate how to improve F1 for Test using Train and did not improve Test results (Table 10, row 6,9 vs Table 9, row 12).

Table 11 Naïve and Multinomial Naïve Bayes classifiers' results[t]

	Filter	*Stopwords filter*	*Attributes selection*	*TP*	*FP*	*FN*	*Pre.*	*Recall*	*F1*
Naïve Bayes, Unigram									
1.	No Filter	No	No – 703	137	4909	5	0.3	0.96	0.05
2.	No Filter	No	Yes – 569	137	5773	5	0.2	0.97	0.05
3.	No Filter	Yes	No – 764	136	5218	6	0.03	0.96	0.05
4.	No Filter	Yes	Yes – 572	136	6270	6	0.02	0.96	0.04
5.	Computer	No	No – 654	137	5992	5	0.02	0.96	0.04
6.	Computer	No	Yes – 576	137	6127	5	0.02	0.97	0.04
7.	Computer	Yes	No – 654	136	6537	6	0.02	0.96	0.04
8.	Computer	Yes	Yes – 586	136	6710	6	0.02	0.96	0.04
Multinomial Naïve Bayes, Unigram									
9.	No Filter	No	No – 703	138	1465	4	0.09	0.97	0.16
10.	No Filter	No	Yes – 569	137	911	5	0.13	0.96	0.23
11.	No Filter	Yes	No – 764	139	2161	3	0.06	0.98	0.11
12.	No Filter	Yes	Yes – 572	138	1166	4	0.11	0.97	0.19
13.	Computer	No	No – 654	137	713	5	0.16	0.96	0.28
14.	Computer	No	Yes – 576	137	623	5	0.18	0.96	0.30
15.	Computer	Yes	No – 654	138	952	4	0.13	0.97	0.22
16.	Computer	Yes	Yes – 586	137	831	5	0.14	0.96	0.25
Naïve Bayes Applied to Accept file (improved) Approach									
17.	Computer	NA	No – 136	91	76	51	0.54	0.64	**0.59**
18.	Computer	NA	Yes – 103	91	76	51	0.54	0.64	**0.59**
Multinomial Naïve Bayes Applied to Accept file (improved) Approach									
19.	Computer	NA	No – 136	13	3	129	0.81	0.09	0.16
20.	Computer	NA	Yes – 103	11	4	131	0.73	0.08	0.14

Table 12 Sample of attribute probabilities defined by best scored Multinomial Naïve Bayes classifier

Attributes	*False*	*True*	*Diff.*	*Attributes*	*False*	*True*	*Diff.*
ok	0.0051	0.0065	0.0014	**i**	0.0162	0.0051	-0.0111
to	0.0111	0.0057	-0.0054	**you**	0.0143	0.0037	-0.0106
get	0.0034	0.0056	0.0022	**u**	0.0120	0.0043	-0.0077
be	0.0044	0.0055	0.0011	**a**	0.0119	0.0040	-0.0080
me	0.0081	0.0055	-0.0025	**m**	0.0113	0.0008	-0.0105
want	0.0041	0.0053	0.0012	**to**	0.0111	0.0057	-0.0054
call	0.0011	0.0053	0.0042	hi	0.0110	0.0007	-0.0103
if	0.0044	0.0052	0.0008	asl	0.0099	0.0000	-0.0099
when	0.0023	0.0052	0.0029	hey	0.0096	0.0016	-0.0080
so	0.0069	0.0051	-0.0018	**the**	0.0095	0.0046	-0.0050
i	0.0162	0.0051	-0.0111	**and**	0.0095	0.0045	-0.0050
will	0.0027	0.0051	0.0023	**it**	0.0091	0.0051	-0.0040
it	0.0091	0.0051	-0.0040	**f**	0.0084	0.0002	-0.0082
can	0.0052	0.0050	-0.0003	**are**	0.0084	0.0031	-0.0053
on	0.0059	0.0049	-0.0010	**is**	0.0083	0.0038	-0.0045
do	0.0079	0.0048	-0.0031	**what**	0.0081	0.0047	-0.0035
see	0.0033	0.0048	0.0015	**me**	0.0081	0.0055	-0.0025
would	0.0026	0.0048	0.0022	**do**	0.0079	0.0048	-0.0031
just	0.0053	0.0047	-0.0006	**my**	0.0076	0.0041	-0.0035
lol	0.0054	0.0047	-0.0007	**have**	0.0071	0.0047	-0.0024

This experiment did, however, suggest: (i) it is not necessary to expect a specific category in detection; (ii) dataset difference impacts performance; (iii) a system relying on such an approach may obtain worse results.

Naïve and Multinomial Naïve Bayes classifiers

Classifiers such as Naïve Bayes and Multinomial Naïve Bayes were suggested for such a machine learning evaluation. Naive Bayes calculates the probability of having a feature (attribute) given a class; Multinomial Naïve Bayes considers these features as independent variables. For Predator Identification, we are interested in co-occurrences of attributes and independence is unclear. Weka helpfully provides Java classes to simplify the transformation of data to a suitable form for analysis with same parameters as J48. We evaluate the effects of (shown in Table 11):

1. Computer based conversation filter only (column 2, Filter)
2. Including and excluding stopwords[n] (column 3, Stopwords Filter)
3. Attribute Selection using Information Gain based on the rank values of zero and more (column 4, Attributes Selection)

Experiments involved both unigrams and trigrams[o] for attribute selection. Trigrams were selected to seek a balance between attributes chosen by Weka and those we had manually selected. For Naïve Bayes, results for trigrams were marginally better than for unigrams – 0.01 – but selected attributes show preference for unigrams - 51 trigrams, 256 bigrams and 413 unigrams).

Although these results are not particularly illuminating or performative –they bear similarity with those seen for at least one other PAN12 participant[p] - it is worth looking to the attributes and probabilities as relate to the classification. Table 12 shows attributes, probability of

Table 13 Sample of attribute probabilities defined by Naïve Bayes classifier applied on accept file of our approach

Attributes	*False*	*True*	*Difference*	*Attributes*	*False*	*True*	*Difference*
ur mom	0.0021	0.9968	**0.9947**	ur mum	0.0006	0	–0.0006
sweetie	0	0.8369	0.8369	your mum	0.0006	0	–0.0006
your mom	0.0013	0.7356	0.7343	you are so young	0.0001	0	–0.0001
in trouble	0.0007	0.417	0.4163	much young	0.0001	0	–0.0001
ur dad	0.0007	0.4161	0.4154	in troble	0	0	0

Table 14 SMO classifiers' results

	Filter	*Stopwords filter*	*Attributes Selection*	*TP*	*FP*	*FN*	*Pre.*	*Recall*	*F1*
SVM, Unigram									
1.	No Filter	No	No – 703	110	16	32	0.87	0.77	0.82
2.	No Filter	No	Yes – 569	106	16	36	0.87	0.75	0.80
3.	No Filter	Yes	No – 764	99	15	43	0.86	0.70	0.77
4.	No Filter	Yes	Yes – 572	98	17	44	0.85	0.70	0.76
5.	Computer	No	No – 654	113	11	29	0.91	0.80	0.85
6.	Computer	No	Yes – 576	114	10	28	0.92	0.80	**0.86**
7.	Computer	Yes	No – 654	106	12	36	0.90	0.75	0.82
8.	Computer	Yes	Yes – 586	106	12	36	0.90	0.75	0.82
SVM, Trigram									
9.	No Filter	No	No – 743	104	17	38	0.86	0.73	0.79
10.	No Filter	No	Yes – 639	104	16	38	0.87	0.73	0.79
11.	No Filter	Yes	No – 783	106	13	36	0.89	0.75	0.81
12.	No Filter	Yes	Yes – 648	105	13	37	0.89	0.74	0.81
13.	Computer	No	No – 720	107	12	35	0.90	0.75	0.82
14.	Computer	No	Yes – 602	104	13	38	0.89	0.73	0.80
15.	Computer	Yes	No – 758	102	11	40	0.90	0.72	0.80
16.	Computer	Yes	Yes – 648	105	13	38	0.89	0.74	0.81
SVM Applied on Accept file of our Approach									
17.	Computer	NA	No – 136	61	10	81	0.86	0.43	0.57
18.	Computer	NA	Yes – 103	61	10	81	0.86	0.43	0.57
Results from Applying SVM Train Model on Test									
19.	Computer	NA	No – 654	162	8	92	0.95	0.64	**0.76**
20.	Computer	NA	Yes – 576	158	9	96	0.95	0.62	0.75

their occurrence for a classification of False (Non-Predator), probability for True (Predator), and the difference (positive values imply True) in two different columns sorted by True and False values respectively. Of the 40 attributes in the table, 34 (excluding 'ok', 'call' and 'lol' on the left and 'hi', 'asl' and 'hey' on the right) are members of a well-known stoplist generated for Information Retrieval from the Brown corpus[q].

Of interest for Deception Detection theorists would be the relationship between first person singular pronouns (underlined in the Table 12) and predatory conversations. Deception Detection theory suggests a reduction in use of first person singular pronoun by those attempting to deceive; as a means of distancing self from deception[r].

Although results appear disappointing, applying the same approach to our 136 accept attributes enables us to verify their value. For computer-filtered conversations (no selection), we find 4 attributes have small negative values and 28 have a zero probability (Table 11, row 17, 18). These cover possible typos/misspellings that we have included to offer coverage for a Test Corpus knowing that Train is only a possible sample. Deleting all attributes with non-positive values does not change the outcome, so they are at least not harmful to the approach. Table 13 shows the top 5 True and top 5 False attributes.

As Table 13 indicates, there is a big difference between those attributes indicating predators and those offering negative evidence. However, these sets do offer up variations of potential interest between 'mom' and 'mum', 'trouble' and 'troble'.

Support vector machine: SMO classifier

SVMs have been shown to "consistently achieve good performance on text categorization tasks", and have been demonstrated to outperform various other approaches[s]. To understand what is possible if using SVM, we

Table 15 Sample of attribute weights assessed by SMO

Predatory words		*Non-predatory words*	
Weight	*Attribute*	*Weight*	*Attribute*
0.32	yes	–0.08	yep
0.30	mmm	–0.09	umm
0.20	hello	–0.06	hey
0.08	ya	–0.06	yo
0.08	aww	–0.20	awww

made use of Weka's SMO with the same data structures as for Naïve Bayes and Multinomial Naïve Bayes.

In most cases, results were either equal to or better than those from *any* of the previous experiments. Best results were achieved for attributes selected from computer conversation filtered dataset with stopwords not removed (Table 14, row 6).

Table 15, shows a set of words identified by SMO (stopwords included) and their related effect.

Use of SVM improved results dramatically. However, Table 16 shows some of the attributes which result in high scores being produced, where "aww" is considered to be predatory but "awww" is not. It would seem peculiar to suggest that predators rely on a shared style guide, and yet this would be one simple theory that could be derived from such an analysis.

The preponderance of stopwords in Table 12 brings about one final question regarding the extent of their influence. We tested for this by removing all but stopwords from all computer filtered chats and running SVM over these data. Results (F1, 0.58) seem to indicate that stopwords alone could be used to some extent for predator detection, which would outperform a number of other tested approaches in PAN2012. How meaningful such a result is, and what it can tell us about the conversations in general and the machine learning approaches in particular, remains to be understood.

Of the machine learning approaches attempted, SVM leads to the best result, which outperforms our approach and based on results here could indicatively have offered 4th place in the PAN2012 competition compared to the 9th achievable through post-competition improvements. This is unsurprising: competition participants who featured in 1st, 3rd (2nd unreported) and 4th all used SVM. However, it is important also to evaluate the basis for performance. If we think of the features derived by SVM as correlations from attributes to a binary classification, we can then judge whether attributes would generally make sense in such a correlation. Further, we can consider how such attributes would relate to a theory such as that of the Cycle of Entrapment and whether we would expect correlation values to remain stable. Moreover, in this particular scenario, we could consider whether such a correlation might offer sound evidence for judicial purposes. Whilst use of the word 'address' fits to our theory (in the right context), it is more difficult to suggest a theoretical basis upon which the words 'were', 'call' and 'there' would be predatory indicators.

The question still remains whether or not we can easily explain the reasons that would underlie a detection, not least so that it is readily possible to reason over any false detection, which in reality might have a devastating effect on an innocent person.

Discussion and conclusion

In this paper, we have presented our understanding of paedophilia and related issues of hebephilia and sexual offences against children, and through these to understanding what would be detectable in the predatory activities as might precede these. In discussing the clinical and legal perspectives on these matters, we noted how variation in age is a feature and that mainstream use of such labels can be inconsistent with such definitions. We noted that predatory activities tend to involve a degree of effort on the part of the predator, and that the Cycle of Entrapment is where such effort may be focussed. In extending our discussion to the online world, we noted implications for the Cycle of Entrapment and the difficulties of applying technological controls and also of how the blurring of geographies can create issues, and ways in which predators can use the online world. We also discussed technologies that can be deployed against such predators, as well as our own approach to detection which accounts for the identification of requests for information from children that would relate to the Cycle of Entrapment, and approaches which were

Table 16 Top 10 attribute probabilities from each class defined by SMO classifier

Predatory words				*Non-predatory words*			
Weight	*Attribute*	*Weight*	*Attribute*	*Weight*	*Attribute*	*Weight*	*Attribute*
0.89	ok	0.52	were	−0.65	f	−0.45	thought
0.77	call	0.52	there	−0.53	they	−0.44	stuff
0.68	hun	0.48	address	−0.52	kewl	−0.43	long
0.65	leave	0.48	u	−0.52	nice	−0.42	tired
0.65	soon	0.47	mind	−0.50	email	−0.42	my
0.59	sweetie	0.47	older	−0.49	which	−0.42	means
0.57	pm	0.46	very	−0.48	omg	−0.41	idk
0.54	hour	0.43	tight	−0.46	cuz	−0.41	bed
0.52	around	0.43	going	−0.46	a	−0.38	ve
0.52	right	0.42	sure	−0.46	like	−0.38	waiting

geared to improving our own system and evaluating it against other approaches.

Children are often more skilled in their use of social media and general ICT than parents, but likely less adept at identifying suspicious behaviour or appropriately assessing the risks from their interactions. Determined predators are likely to use all possible devices at their disposal to satisfy their needs, whilst defences against these are at present likely to be minimal. This asymmetry makes social media and related online activities into a rich hunting ground for these predators, with lawmakers and enforcers always trying to keep up. Detection of any kind, appropriately deployed, as can make the tasks of such predators rather more difficult again will begin to address this asymmetry and help make the online world a safer place for children of any age.

Endnotes

[a] We have emphasised the term prepubescent, and will refer back to this later.

[b] As above.

[c] E.g. [48].

[d] National Criminal Intelligence Service, *UK Threat Assessment of Serious and Organised Crime, Paedophile Crime Including Online Child Abuse*. 2002, National Criminal Intelligence Service.

[e] Wolak et al. [19].

[f] Finkelhor et al., [49].

[g] Based on EU Kid report only 56% can change privacy settings on a social networking profile, and 51% can block junk mails and spam [50].

[h] Directive 2011/92/EU of the European Parliament and of the Council of 13 December 2011 on combating the sexual abuse and sexual exploitation of children and child pornography, and replacing Council Framework Decision 2004/68/JHA.

[i] Cyber-stalking is recognised as an offence that can be prosecuted under a range of existing legislation (such as the UK Protection from Harassment Act, 1997).

[j] For more information and statistics please check [19,49,50].

[k] See Inches & Crestani [51] for further detail related to creation of the Dataset.

[l] Values differ to those presented on the PAN2012 website due to organisers uses F0.5 which puts emphasis on Precision over Recall. We would consider a better system as one able to detect more suspects, hence our discussion tends to draw attention to Recall values. As a compromise, we only present F1 values.

[m] Spinning the Election (Skillicorn and Little) "We do not yet completely understand these models of word use, so the results should be taken with a grain of salt [52]."

[n] Stopwords list extracted from [53].

[o] In Weka, n-grams include lower values for n, so trigrams covers bigrams and unigrams also.

[p] See participant run for gomezhidalgo12-2012-06-15-1 in Inches & Crestani [51].

[q] A list of 421 words produced from the Brown Corpus as described by Fox [54].

[r] We have addressed this in greater detail in [55], section Relationship with victims (Children).

[s] Thorsten [56].

[t] Tables from the trigram experiments have not been presented as they did not produce an effect on the results.

Received: 7 August 2013 Accepted: 29 April 2014
Published online: 18 September 2014

References

1. J Gentry, Ancient Pedophilia, The Ohio State University, 2009 [https://kb.osu.edu/dspace/handle/1811/37221]
2. MC Seto, *Pedophilia and Sexual Offending Against Children: Theory, Assessment, and Intervention* (American Psychological Association, Washington, DC, US, 2008)
3. The Sexual Offence Act 2003 (c.42), [http://www.legislation.gov.uk/ukpga/2003/42/contents]
4. ICD-10, *Classification of Mental and Behavioural Disorders Diagnostic Criteria for Research* (World Health Organization, Geneva, 2010). http://apps.who.int/classifications/icd10/browse/2010/en
5. American Psychiatric Association, *Diagnostic and Statistical Manual of Mental Disorders (DSM-IV-TR)*, 2000
6. D Finkelhor, S Araji, Explanations of pedophilia: a four factor model. J. Sex Res. **22**, 145–161 (1986)
7. L Chiang, Y Lin, H Chan, B Chiang, Differential manifestations of prepubescent, pubescent and postpubescent pediatric patients with systemic lupus erythematosus: a retrospective study of 96 Chinese children and adolescents. Pediatr. Rheumtol. **10**(12), 1–9 (2012)
8. R Blanchard, AD Lykins, D Wherrett, ME Kuban, JM Cantor, T Blak, R Dickey, PE Klassen, Pedophilia, hebephilia, and the DSM-V. Arch. Sex Behav. **38**, 335–350 (2009)
9. HN Snyder, *Sexual Assault of Young Children as Reported to Law Enforcement: Victim, Incident, and Offender Characteristics (Report No. NCJ 18399)* (US: Department of Justice, Washington, DC, 2000)
10. K Lanning, *Child Molesters: A Behavioral Analysis. For Professionals Investigating the Sexual Exploitation of Children* (National Center for Missing & Exploited Children, USA, 2010)
11. LN Olson, JL Daggs, BL Ellevold, TKK Rogers, Entrapping the innocent: toward a theory of child sexual predators' luring communication. Comm. Theor. **17**(3), 231–251 (2007)
12. C Van Dam, *The Socially Skilled Child Molester: Differentiating the Guilty from the Falsely Accused* (The Haworth Press, Binghamton, NY, 2006)
13. A Strauss, J Corbin, *Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory*, 2nd edn. (Sage, Thousand Oaks, CA, 1998)
14. BH Spitzberg, L Marshall, WR Cupach, Obsessive relational intrusion, coping, and sexual coercion victimization. Comm. Rep. **14**, 19–30 (2001)
15. C Harms, Grooming: an operational definition and coding scheme. Sex Offender Law Rep. **8**(1), 1–6 (2007)
16. D Hughes, P Rayson, J Walkerdine, K Lee, P Greenwood, A Rashid, C May-Chahal, M Brennan, Supporting law enforcement in digital communities through natural language analysis, in *Proceedings of the Second International Workshop on Computational Forensics (IWCF '08)* (Springer, Washington, DC, 2008), pp. 122–134
17. L Ellison, Cyberstalking: Tackling Harassment on the Internet, in *14th BILETA Conference: CYBERSPACE 1999* (Crime, Criminal Justice and the Internet, 1999)
18. Office of Juvenile Justice and Delinquency Prevention (OJJDP), http://www.ojjdp.gov/
19. J Wolak, K Mitchell, D Finkelhor, *Online Victimization of Youth: Five Years Later. National Center for Missing & Exploited Children* (The Crimes Against Children Research Center, USA, 2006)
20. Virtual Global Taskforce, http://www.virtualglobaltaskforce.com/

21. Child Exploitation & Online Protection Centre: Internet Safety (CEOP), http://ceop.police.uk/
22. Internet Watch Foundation, http://www.iwf.org.uk/
23. K Durkin, Misuse of the Internet by Pedophiles: implication for law enforcement and probation practice. Federal Probation **61**(3), 14–18 (1997)
24. RCW Hall, RCW Hall, A profile of pedophilia: definition, characteristics of offenders, recidivism, treatment outcomes, and forensic issues. Mayo Clin Proc **82**(4), 457–471 (2007)
25. L Penna, A Clark, G Mohay, *Challenges of Automating the Detection of Paedophile Activity on the Internet. First International Workshop on Systematic Approaches to Digital Forensic Engineering (SADFE'05)* (IEEE Computer Society, 7-9 November 2005, Taipei, Taiwan, 2005), pp. 206–222
26. Perverted Justice, http://www.perverted-justice.com/
27. Net Nanny, https://www.netnanny.com/products/net-nanny-social/
28. ContentBarrier, http://www.intego.com/manuals/en/cb/1-welcome-to-contentbarrier.html
29. M Scott, Focusing on the text and its key words, in *Rethinking Language Pedagogy from a Corpus Perspective*, ed. by L Burnard, T McEnery (Peter Lang, Frankfurt, 2000), pp. 104–121
30. N Pendar, Toward spotting the pedophile: telling victim from predator in text chats, in *First IEEE International Conference on Semantic Computing (ICSC 2007)* (IEEE Comput Society, 17-19 Sep 2007, Irvine, USA, 2007), pp. 235–241
31. MW RahmanMiah, J Yearwood, S Kulkarni, Detection of Child Exploiting Chats from a Mixed Chat Dataset as a Text Classification Task, in *Australasian Language Technology Association Workshop 2011 (ALTA 2011)*, ed. by D Molla (Association for Computational Linguistics (ACL), 1-2 December 2011, Canberra, Australia, 2011), pp. 157–165
32. R O'Connell, *A Typology of Child Cyber- sexploitation and Online Grooming Practices. In Cyberspace Research Unit* (University of Central Lancashire, 2003)
33. JW Pennebaker, ME Francis, RJ Booth, *Linguistic Inquiry and Word Count (LIWC)* (Erlbaum Publishers, 2001)
34. I McGhee, J Bayzick, A Kontostathis, L Edwards, A McBride, E Jakubowski, Learning to identify internet sexual predation. Int J Electron Commerce **15**(3), 103–122 (2011)
35. D Michalopoulos, I Mavridis, Utilizing document classification for grooming attack recognition, in *IEEE Symposium on Computers and Communications (ISCC)* (IEEE, 28 June - 1July, 2011, Kerkira, Greece, 2011), pp. 864–869
36. D Bogdanova, P Rosso, T Solorio, On the impact of sentiment and emotion based features in detecting online sexual predators, ed. by P Forner, R Navigli, D Tufis (Working Notes Papers of the CLEF 2012 Evaluation Labs, Rome, Italy, 2012), pp. 110–118
37. C Strapparava, R Mihalcea, SemEval-2007 Task 14: Affective Text. The 4th International Workshop on Semantic Evaluations (SemEval '07), (Association for Computational Linguistics (ACL), 23-24 June, Prague, Czech Republic, 2007), pp. 70–74
38. S Argamon, M Koppel, J Pennebaker, J Schler, Automatically profiling the author of an anonymous text. Comm ACM **52**(2), 119–123 (2009)
39. C Peersman, F Vaassen, V Van Asch, W Daelemans, Conversation Level Constraints on Pedophile Detection in Chat Rooms - Notebook for PAN at CLEF 2012, ed. by P Forner, R Navigli, D Tufis (Working Notes Papers of the CLEF 2012 Evaluation Labs., 17-20 September 2012, Rome, Italy, 2012)
40. PH Adams, CH Martell, Topic detection and extraction in chat, in *Proceedings of the 2008 IEEE International Conference on Semantic Computing* (IEEE Computer Society Press, Los Alamitos, 2008), pp. 581–588
41. D Yin, Z Xue, L Hong, BD Davison, A Kontostathis, L Edwards, *Detection of Harassment on Web 2.0* (2009). Paper presented at the first Content Analysis in Web 2.0 Workshop, Madrid, April 15
42. P Juola, An overview of the traditional authorship attribution subtask, in *CLEF 2012 Evaluation Labs and Workshop - Working Notes Papers*, ed. by P Forner, J Karlgren, C Womser-Hacker (Rome, Italy, 2012)
43. Omegle: Talk to strangers! http://omegle.inportb.com
44. The Omeglean Society, http://inportb.com/2010/02/21/the-omeglean-society/
45. IRC logs, http://irclog.org/
46. IRC Logs Archive, http://krijnhoetmer.nl/irc-logs/
47. A Vartapetiance, L Gillam, Quite Simple Approaches for Authorship Attribution, Intrinsic Plagiarism Detection and Sexual Predator Identification - Notebook for PAN at CLEF 2012, in *CLEF 2012 Evaluation Labs and Workshop - Working Notes Papers*, ed. by P Forner, R Navigli, D Tufis (Working Notes Papers of the CLEF 2012 Evaluation Labs., 17-20 September 2012, Rome, Italy, 2012)
48. G Sheldrick, D Churchill, Abducted schoolgirl is banned from visiting paedophile teacher Jeremy Forrest. (2013). 1 July. http://www.express.co.uk/news/uk/411477/Abducted-schoolgirl-is-banned-from-visiting-paedophile-teacher-Jeremy-Forrest
49. D Finkelhor, K Mitchell, J Wolak, Online Victimization, *A Report on the Nation's Youth. National Center for Missing & Exploited Children* (The Crimes Against Children Research Center, USA, 2000)
50. S Livingstone, L Haddon, A Görzig, K Olafsson, EU kids online final report. (2011). http://www2.lse.ac.uk/media@lse/research/EUKidsOnline/EU%20Kids%20II%20(2009-11)/EUKidsOnlineIIReports/Final%20report.pdf
51. G Inches, F Crestani, Overview of the international sexual predator identification competition at pan-2012, in *CLEF 2012 Evaluation Labs and Workshop - Working Notes Papers*, ed. by P Forner, J Karlgren, C Womser-Hacker (Rome, Italy, 2012)
52. D Skillicorn, A Little, Spinning the Election, http://research.cs.queensu.ca/home/skill/election/election.html
53. Stopwords: Default English Stopwords List, http://www.ranks.nl/resources/stopwords.html
54. C Fox, A stop list for general text. ACM Sigir **24**(2), 19–21 (1989)
55. A Vartapetiance, L Gillam, Deception detection: dependable or defective? Springer Journal of Social Network Analysis and Mining, Special Issue on Uncovering Deception in Social Media **4**(166), 1–14 (2013)
56. J Thorsten, *Text Categorization with Support Vector Machines: Learning with Many Relevant Features, Technical Report LS VIIIReport, Universit'at Dortmund, Dortmund, Germany*, (1997). http://www.cs.cornell.edu/people/tj/publications/joachims_98a.pdf

doi:10.1186/s13388-014-0003-7
Cite this article as: Vartapetiance and Gillam: **"Our Little Secret": pinpointing potential predators.** *Security Informatics* 2014 **3**:3.