Available online at www.sciencedirect.com

SciVerse ScienceDirect

Computer Speech and Language 28 (2014) 108–120

COMPUTER SPEECH AND LANGUAGE

www.elsevier.com/locate/csl

Exploring high-level features for detecting cyberpedophilia

Dasha Bogdanova [a,*], Paolo Rosso [b], Thamar Solorio [c]

[a] *University of Saint Petersburg, Russian Federation*
[b] *NLE Lab, ELiRF, Universitat Politècnica de València, Spain*
[c] *CoRAL Lab, University of Alabama at Birmingham, USA*

Received 1 August 2012; received in revised form 10 April 2013; accepted 24 April 2013
Available online 3 May 2013

Abstract

In this paper, we suggest a list of high-level features and study their applicability in detection of cyberpedophiles. We used a corpus of chats downloaded from http://www.perverted-justice.com and two negative datasets of different nature: cybersex logs available online, and the NPS chat corpus. The classification results show that the NPS data and the pedophiles' conversations can be accurately discriminated from each other with character n-grams, while in the more complicated case of cybersex logs there is need for high-level features to reach good accuracy levels. In this latter setting our results show that features that model behaviour and emotion significantly outperform the low-level ones, and achieve a 97% accuracy.

Keywords: Cyberpedophilia; Sentiment analysis; Emotion detection

1. Introduction

Child sexual abuse and pedophilia are both problems of great social concern. On the one hand, law enforcement is working on prosecuting and preventing child sexual abuse. On the other hand, psychologists and mental specialists are investigating the phenomenon of pedophilia. Even though pedophilia has been studied from different research points, it remains to be a very important problem that requires further research, especially from the automatic detection point of view.

Previous studies report that in the majority of cases of sexual assaults the victims are underaged (Snyder, 2000). On the Internet, attempts to solicit children have become common as well. Wolak et al. (2003) found out that 19% of children have been sexually approached online. However, manual monitoring of each conversation is impossible, due to the massive amount of data and privacy issues. A good and practical alternative is the development of reliable tools for detecting pedophilia in online social media.

In this paper, we address the problem of distinguishing pedophiles in chat logs with natural language processing (NLP) techniques. This problem becomes even more challenging because of the chat data specificity. Chat conversations are very different not only from the written text but also from other types of social media interactions, such as blogs and forums, since chatting on the Internet usually involves very fast typing. The data usually contains a large amount

* Corresponding author. Tel.: +7 9516547238.
E-mail address: dasha.bogdanova@gmail.com (D. Bogdanova).

0885-2308/$ – see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.csl.2013.04.007

of mistakes, misspellings, specific slang, and character flooding. Therefore, accurate processing of this data with automated analyzers is quite challenging and can result in very noisy output.

Previous research on pedophilia reports that the expression of certain emotions in text could be helpful to detect pedophiles in social media (Egan et al., 2011). Following these insights we suggest a list of features, including sentiments as well as other content-based features that could unveil semantic dimensions important in detecting cyberpedophilia. We propose a model of fixated discourse, one of the characteristics of cyberpedophile conversations described in previous research. The model we propose is based on lexical chains. We include this feature in further experiments as well as other high-level features. We investigate the impact of the proposed features on the problem of distinguishing pedophile chats from non-pedophile chats. Our experimental results show that binary classification based on such features discriminates pedophiles from non-pedophiles with high accuracy.

The remainder of the paper is structured as follows: Section 2 overviews related work on the topic. Section 3 outlines the profile of a pedophile based on previous research. Our approach to the problem is presented in Section 5. Experimental data is described in Section 4. We show the results of the conducted experiments in Section 6. In Section 7 we discuss in more detail the findings from our research. We finally draw some conclusions and share plans for future research in Section 8.

2. Related research

The problem of automatic detection of pedophiles in social media has been rarely addressed so far. In part, this is due to the difficulties involved in having access to useful data. There is an American foundation called Perverted Justice (PJ), that investigates cases of online child sexual abuse: adult volunteers enter chat rooms as juveniles (usually 12–15 year old) and if they are sexually solicited by adults, they work with the police to prosecute the offenders. Some chat conversations with cyberpedophiles are available at http://www.perverted-justice.com and they have been the subject of analysis of recent research on this topic.

Pendar (2007) experimented with PJ data. He separated the lines written by pedophiles from those written by pseudo-victims and used a kNN classifier based on word n-grams to distinguish between them.

Another related research has been carried out by McGhee et al. (2011). The chat lines from PJ were manually classified into the following categories:

1. Exchange of personal information.
2. Grooming.
3. Approach.
4. None of the classes listed above.

Their experiments have shown that kNN classification achieves up to 83% accuracy and outperforms a rule-based approach.

It is well known that pedophiles often create false profiles and pretend to be younger or of the opposite sex. Moreover, they try to copy children's behaviour. Automatically detecting age and gender in chat conversations could then be the first step in detecting cyberpedophilia. Peersman et al. (2011) have analyzed chats from the Belgium Netlog social network. Discrimination between those who are older than 16 from those who are younger based on a Support Vector Machine classification yields 71.3% accuracy. The accuracy is even higher when the age gap is increased (e.g. the accuracy of classifying those who are less than 16 from those who are older than 25 is 88.2%). They have also investigated the issues of the minimum amount of training data needed. Their experiments have shown that with 50% of the original dataset the accuracy remains almost the same, and with only 10% it is still much better than the random baseline performance.

NLP techniques were as well applied to capture child sexual abuse data in P2P networks (Panchenko et al., 2012). The proposed text classification system is able to predict with high accuracy if a file contains child pornography by analyzing its name and textual description.

A shared task on a similar problem was organized at PAN 2012 (http://pan.webis.de/). Given many short conversations, the task was to identify which user was convincing others"to provide some sexual favour". Conversations were not longer than 150 messages and the percentage of predators was lower than 4%. The system that achieved the highest performance (Villatoro-Tello et al., 2012) was based on lexical features, prefiltering and a two-step classification. First,

conversations were prefiltered, e.g. by removing those containing only one user. Then, suspicious conversations were identified and lastly, "predators" were detected among the suspicious conversations. In contrast, this research is not about identifying users convincing others to provide some sexual favour. It neither aims at classification of chat lines into categories, as it was done by McGhee et al. (2011), nor at discriminating between victim and pedophile as it was done by Pendar (2007). Our goal is to reveal behaviour and emotion dimensions, i.e. high-level features based on clues provided by psychology and sentiment analysis, that can help to distinguish chats that belong to a pedophile from those of non-pedophiles.

3. Profiling the pedophile

Pedophilia is a "disorder of adult personality and behaviour" which is characterized by sexual interest in prepubescent children (World Health Organization, 1988). Even though solicitation of children is not a medical diagnosis, Abel and Harlow (2001) reported that 88% of child sexual abuse cases are committed by pedophiles. Therefore, we believe that understanding behaviour of pedophiles could help to detect and prevent children sexual abuse in social media. While an online sexual offender is not always a pedophile, in this paper we use these terms as synonyms.

Previous research reports that about 94% of sexual offenders are males. With respect to female sexual molesters, it is reported, that they tend to be young and, in these cases, men are often involved as well (Vandiver and Kercher, 2004). Sexual assault offenders are more often adults (77%), though in 23% of cases children are solicited by other juveniles.

Analysis of pedophiles' personality characterizes them with feelings of inferiority, isolation, loneliness, low self-esteem and emotional immaturity. Moreover, 60–80% of them suffer from other psychiatric illnesses (Hall and Hall, 2007). In general, pedophiles are less emotionally stable than mentally healthy people.

Hall and Hall (2007) noticed that five main types of computer-based sexual offenders can be distinguished: (1) the stalkers, who approach children in chat rooms in order to get physical access to them; (2) the cruisers, who are interested in online sexual molestation and not willing to meet children offline; (3) the masturbators, who watch child pornography; (4) the networkers or swappers, who trade information, pornography, and children; and (5) a combination of the four types. In this study we are interested in detecting stalkers (type (1)) and cruisers (type (2)).

The language sexual offenders use was analyzed by Egan et al. (2011). The authors considered the chats available from PJ. The analysis of the chats revealed several characteristics of pedophiles' language:

- Implicit/explicit content. Typically, pedophiles shift gradually to the sexual conversation, starting with ordinary compliments:
  Offender: hey you are really cute
  Offender: u are pretty
  Offender: hi sexy
  Then they shift the conversation to make it overtly related to sex. They do not hide their intentions:
  Offender: can we have sex?
  Offender: you ok with sex with me and drinking?
- Fixated discourse. Pedophiles are not willing to step aside from the sexual conversation. For example, in this conversation the pedophile almost ignores the question of pseudo-victim and comes back to the sex-related conversation:
  Offender: licking dont hurt
  Offender: its like u lick ice cream
  Pseudo-victim: do u care that im 13 in march and not yet? i lied a little bit b4
  Offender: its all cool
  Offender: i can lick hard
- Offenders often understand that what they are doing is not moral:
  Offender: i would help but its not moral
- They transfer responsibility to the victim:
  Pseudo-victim: what ya wanta do when u come over
  Offender: whatever – movies, games, drink, play around – it's up to you – what would you like to do?
  Pseudo-victim: that all sounds good
  Pseudo-victim: lol

Table 1
Statistics on the experimental data. The type information refers to the categorization provided by Pendar (2007).

Corpus	Type	Training set	Test set
Perverted-justice: subset 1	1(b)	40	20
Perverted-justice: subset 2	1(b)	40	20
Perverted-justice: subset 3	1(b)	40	20
Perverted-justice: subset 4	1(b)	40	20
Perverted-justice: subset 5	1(b)	40	20
Subset of NPS chat corpus	Other	45	20
cybersex logs	2	48	20

Offender: maybe get some sexy pics of you:-P
Offender: would you let me take pictures of you? of you naked? of me and you playing?:-D

- Offenders often behave as children, copying their linguistic style. Colloquialisms appear often in their messages:
  Offender: howwwww dy
  . . .
  Offender: i know PITY MEEEE
- They try to minimize the risk of being prosecuted: they ask to delete chat logs and warn victims not to tell anyone about the talk:
  Offender: don't tell anyone we have been talking
  Pseudo-victim: k
  Pseudo-victim: lol who would i tell? no one's here.
  Offender: well I want it to be our secret
- Though they finally stop being cautious and insist on meeting offline:
  Offender: well let me come see you
  Pseudo-victim: why u want 2 come over so bad?
  Offender: i wanna see you

In general, Egan et al. (2011) have found online solicitation to be more direct, while in real life children seduction is more deceitful.

4. Datasets

Pendar (2007) has summarized the possible types of chat interactions with sexually explicit content:

1. Offender/other
   (a) Offender/victim (victim is underage).
   (b) Offender/volunteer posing as a child.
   (c) Offender/law enforcement officer posing as a child.
2. Adult/adult (consensual relationship)

For our current study, the most interesting data is that of the type 1(a). However, obtaining data from actual cases of offender/other is not easy. In contrast, data of the type 1(b) is freely available at the web site http://www.perverted-justice.com. Therefore, have extracted chat logs from the perverted-justice (PJ) website, where pedophiles have been identified and prosecuted by law enforcement agencies. Since the victim is not real, and our goal is to learn the patterns of cyberpedophiles, we considered only the chat prompts written by the pedophiles.

Table 1.

For our task of distinguishing sex-related chat conversations where one of the parties involved is a pedophile, the ideal negative dataset would be chat conversations of type 2 (consensual relations among adults). However the PJ data will not meet this condition for the negative instances. We need additional chat logs to build the negative dataset. We

used two negative datasets in our experiments: cybersex chat logs and the NPS chat corpus.[1] From each dataset we randomly selected 20 files for testing.

The cybersex chat logs were downloaded from http://oocities.org/urgrl21f/. This dataset belongs to type 2. We assume that the users on these chats are adults, although no explicit attempt was done to verify this. The archive contains 34 one-on-one cybersex logs. We have separated lines of different authors, thereby obtaining 68 files.

We decided to use a subset of the NPS chat corpus (Forsythand and Martell, 2007), even though it is not of type 2, to explore how the high-level features work on the data when distinguishing cyberpedophiles from ordinary conversations. We have extracted chat lines only for those adult authors who had more than 30 lines written. Finally the dataset consisted of 65 authors.

The datasets differ in length. The PJ conversations are much longer, with an average number of words and lines equal to 3618 and 526 respectively. For the cybersex data the averages are 1428 (words) and 97 (lines). The NPS data has even shorter conversations with, an average of 225 words and 52 lines. Balancing the data by trimming the conversations to the same size was not an option because our high level features attempt to model behaviour, thus some of the features we use span over the whole conversation. Moreover, as it is reported by previous research (Egan et al., 2011), cyberpedophile's behaviour changes during the conversation. So, instead of trimming the conversations, we normalize all the features we use by the document length.

5. Our approach

As already mentioned, while previous studies were focused on classifying chat lines into different categories (McGhee et al., 2011) or distinguishing between offender and victim (Pendar, 2007), in this work we address the problem of revealing which high-level features are discriminative when distinguishing pedophile chats from non-pedophile ones.

We formulate the problem of detecting pedophiles in social media as the task of binary text categorization: given a text (a set of chat lines), the aim is to predict whether it is a case of cyberpedophilia or not. We describe our proposed features in the following sections.

5.1. Features

On the basis of previous analysis of pedophiles' personality (described in the previous section), we consider as features those emotional markers that could unveil a certain degree of emotional instability, such as feelings of inferiority, isolation, loneliness, low self-esteem and emotional immaturity.

It has been observed that pedophiles try to be nice with a victim and make compliments, at least in the beginning of a conversation. Therefore, the use of positively charged words is expected. However, pedophiles tend to be emotionally unstable and prone to loose temper easily. Hence words expressing anger and negative lexicon are an expected pattern in their chat logs. Other emotions can be as well a clue to detect pedophiles. For example, offenders often demonstrate fear, especially with respect to being prosecuted, and they express anger and emotions reflecting frustration:

Pseudo-victim: u sad didnt car if im 13. now u car.
Offender: well, *I am just scared* about being in trouble or going to jail
Pseudo-victim: u sad run away now u say no. i gues i dont no what u doin
Offender: *I got scared*
Offender: we would get caught sometime

In this example the pseudo-victim is not answering:

Offender: hello
Offender: r u there
Offender:
Offender: thnx a lot

[1] http://faculty.nps.edu/cmartell/NPSChat.htm.

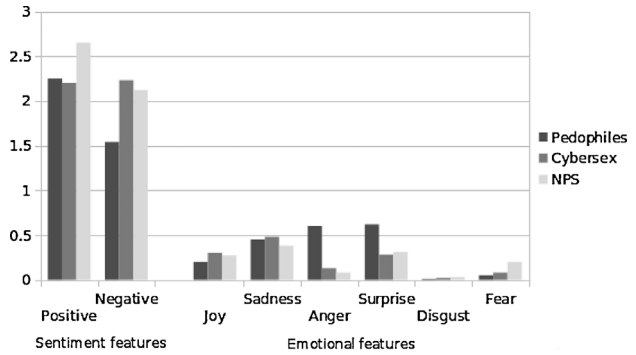

Fig. 1. Average number of sentiment and emotional markers found for the three corpora.

Offender: thanx a lot
Offender:
Offender: *u just wast my time*
Offender: drive down there
Offender: can u not im any more

Here the offender is angry because the pseudo-victim did not call him:

Offender: u didnt call
Offender: *i m angry with u*

Therefore, we have decided to use markers of basic emotions as features. At the SemEval 2007 task on "Affective Text" (Strapparava and Mihalcea, 2007) the problem of fine-grained emotion annotation is defined as: given a set of news titles, the system is to label each title with the appropriate emotion out of the following list: ANGER, DISGUST, FEAR, JOY, SADNESS, SURPRISE. In this research work we only use the percentages of the markers of each emotion.[2] The frequencies of each type of markers are presented in Fig. 1.

Finally, we suggest the following sentiment and emotional markers as the features:

- percentage of positive words;
- percentage of negative words;
- percentage of JOY markers;
- percentage of SADNESS markers;
- percentage of ANGER markers;
- percentage of SURPRISE markers;
- percentage of DISGUST markers;
- percentage of FEAR markers;

We have also borrowed several features from (McGhee et al., 2011):

- Percentage of *approach words*. Approach words include verbs such as *come* and *meet* and nouns such as *car* and *hotel*.
- Percentage of *relationship words*. These words refer to dating (e.g. *boyfriend*, *date*).
- Percentage of *family words*. These words are the names of family members (e.g. *mum*, *dad*, *brother*).
- Percentage of *communicative desensitization words*. These are explicit sexual terms offenders use in order to desensitize the victim (e.g. *penis*, *sex*).

[2] Obtained with WordNet-Affect: http://wndomains.fbk.eu/wnaffect.html.

Table 2
Features used in the experiments.

Feature group	Feature	Example	Reference
Sentiment and emotional markers	Positive words	*Cute, pretty*	SentiWordNet
	Negative words	*Dangerous, annoying*	Baccianella et al. (2010)
	JOY words	*Happy, cheer*	WordNet-Affect (Strapparava and Valitutti, 2004)
	SADNESS words	*Bored, sad*	
	ANGER words	*Annoying, furious*	
	SURPRISE words	*Astonished, wonder*	
	DISGUST words	*Yucky, nausea*	
	FEAR words	*Scared, panic*	
Features borrowed from McGhee et al. (2011)	Approach words	*Meet, car*	McGhee et al. (2011)
	Relationship nouns	*Boyfriend, date*	
	Family words	*mum, dad*	
	Communicative desensitization words	*Sex. penis*	
	Information words	*Asl, home*	
Features helpful to detect neuroticism level	Personal pronouns	*I, you*	Argamon et al. (2009)
	Reflexive pronouns	*Myself, yourself*	
	Obligation verbs	*Must, have to*	
Features derived from pedophile's psychological profile	Fixated Discourse	see in Section 5.2	Bogdanova et al. (2012)
Other	Emoticons	*8),*	
	Imperative sentences	Do it!	

- Percentage of *words expressing sharing information*. This implies sharing basic information, such as age, gender and location, and sending photos. The words include *asl, pic*.

Since pedophiles are known to be emotionally unstable and suffer from psychological problems, we consider features reported to be helpful to detect neuroticism levels by Argamon et al. (2009). In particular, the features include *percentages of personal* and *reflexive pronouns* and *modal obligation verbs* (have to, has to, had to, must, should, mustn't, and shouldn't).

We consider the use of imperative sentences and emoticons to capture the pedophiles' tendencies to be dominant and copy childrens' behaviour respectively. The full list of features is presented in Table 2.

5.2. Modelling fixated discourse

As it was mentioned above, the study of Egan et al. (2011) has revealed several recurrent themes that appear in PJ chats. Among them, *fixated discourse:* the unwillingness of the cyberpedophile to change the topic. We believe that lexical chains are appropriate to model the fixated discourse of the pedophiles chats. We follow the definition of lexical chains discussed below and include these as higher-level features in our approach.

Lexical chains have applications in many tasks including Word Sense Disambiguation (WSD) (Galley and McKeown, 2003) and Text Summarization (Barzilay and Elhadad, 1997). A lexical chain is a sequence of semantically related terms (Morris and Hirst, 1991). In order to find semantically related terms, we used metrics of semantic similarity. In particular, the similarity of Leacock and Chodorow (Leacock and Chodorow, 1998), and the Resnik similarity (Resnik, 1995). Leacock and Chodorow's semantic similarity measure is defined as:

$$Sim_{L\&Ch}(c_1, c_2) = -\log \frac{length(c_1, c_2)}{2 * depth}$$

where $length(c_1, c_2)$ is the length of the shortest path between the concepts c_1 and c_2 and $depth$ is depth of the taxonomy.

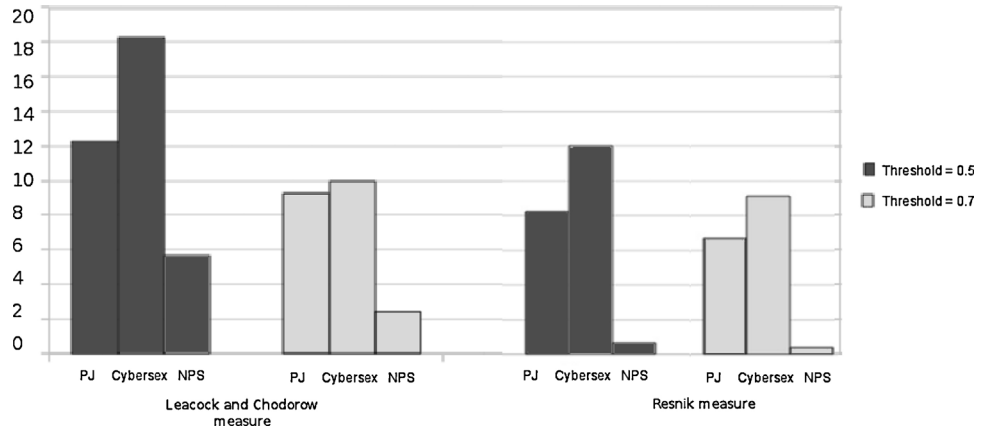

Fig. 2. Average length of lexical chains calculated with Leacock and Chodorow semantic similarity and Resnik semantic similarity.

The semantic similarity measure that was proposed by Resnik (1995) relies on the Information Content concept:

$$IC(c) = -\log P(c)$$

where $P(c)$ is the probability of encountering the concept c in a large corpus. Thus, Resnik's similarity measure is defined as follows:

$$Sim_{Resnik}(c_1, c_2) = IC(lcs(c_1, c_2))$$

where $lcs(c_1, c_2)$ is the least common subsumer of c_1 and c_2.

6. Experiments

In this study we report on three sets of experiments. The first set of experiments focuses on finding an appropriate model for fixated discourse, since we assume this will be one of the main characteristics of conversations with cyberpedophiles. The second set of experiments evaluates the performance on discriminating cyberpedophiles with a machine learning algorithm trained with the high-level features described above. In the last set of experiments we try to assess the contributions from the different features by ablation testing.

6.1. Modelling fixated discourse

We carried out experiments on estimating the length of lexical chains with sexually related content in PJ chats. We have constructed lexical chains (Morris and Hirst, 1991) starting with the anchor word "sex" in the first WordNet meaning: "sexual activity, sexual practice, sex, sex activity (activities associated with sexual intercourse)".

We used Java WordNet Similarity library (Hope, 2008), which is a Java implementation of Perl Wordnet::Similarity (Pedersen et al., 2008). The average length of the longest lexical chains (with respect to the total number of words in a document) found for the different corpora are presented in Fig. 2. As we expected, sex-related lexical chains in the NPS corpus are considerably shorter than in the other two corpora irrespective of the similarity metric used. The lexical chains in the cybersex corpus are even longer than in PJ corpus. This is probably due to the fact that whilst both corpora contain conversations about sex, cyberpedophiles are switching to this topic gradually, whereas cybersex logs are entirely sex-related.

In the next round of experiments we include as a high-level feature the length of the lexical chains constructed with the Resnik similarity measure (Resnik, 1995) (threshold = 0.7). We chose this metric and threshold because the results of Fig. 2 show larger differences among the three corpora with these parameters.

Table 3
Prediction accuracy for the task of detecting pedophiles using high-level features and comparison with different n-gram based lexical features. All results use an SVM classifier. The set of high-level features is that described in Table 2.

	High-level features	Low-level features (baseline)				
		Bag of words	Word bigrams	Word trigrams	Character bigrams	Character trigrams
PJ vs. NPS	0.81	0.60	0.83	0.57	0.95	**0.97**
PJ vs. cybersex	**0.94**	0.50	0.57	0.50	0.52	0.64

The maximum achieved accuracy is in bold.

6.2. Detection of pedophiles

We perform the discrimination between pedophiles and non-pedophiles with a Support Vector Machine (SVM) classifier (Cortes and Vapnik, 1995) (we use LIBLINEAR[3] library). We want to assess the value of the high-level features shown in Table 2.

To extract positive and negative words, we used SentiWordNet (Baccianella et al., 2010). The features borrowed from McGhee et al. (2011), were detected from the lexicon the authors made available to us. Imperative sentences were detected as affirmative sentences starting with verbs. Emoticons were captured with simple regular expressions.

Our dataset is imbalanced, the majority of the chat logs are from PJ. To make the experimental data more balanced, we have created 5 subsets of PJ corpus, each of which contains chat lines from 60 randomly selected pedophiles.

For the cybersex logs, half of the chat sessions belong to the same author. We used this author for training, and the rest for testing, in order to prevent the classification algorithm from learning to distinguish this author from the other pedophiles.

For comparison purposes, we experimented with several baseline systems using low-level features based on n-grams at the word and character level, which were reported as useful features by related research (Peersman et al., 2011). We trained SVMs using word level unigrams, bigrams and trigrams. We also trained SVM classifiers using character level bigrams and trigrams.

The classification results averaged over the five runs are presented in Table 3. As it can be seen from the table, the NPS chats that are not sex-related in general, could be easily distinguished from cyberpedophiles' conversations with low-level features. SVMs based on character trigrams achieves 97% accuracy, while high-level features show only 81%. In case of the cybersex chat logs, which are supposed to have similar vocabulary as the PJ chats, low-level features achieve only up to 64% accuracy, but high-level features provide an accuracy of 94%. These results support the need to extract features that model behaviour and emotion. In particular since it is more likely than in a real world scenario the test data will be more similar to the PJ vs. cybersex than that of PJ vs. NPS. In other words, there is less interest in detecting pedophiles on chat interactions that are not related to sex.

6.3. Feature analysis

In Section 5.1 we have suggested five groups of features. In this section we try to evaluate empirically the individual contributions of each group in the task with two sets of experiments. In the first set, we train the classifier using only one group at a time. In the second set we train the classifier using all but one feature group. The results of the first set are shown in Table 4. From that table it is clear that emotional features (positive and negative words, and basic emotion markers) on their own distinguish the cyberpedophile chats from the cybersex chats with high accuracy that is only 6% lower than the accuracy from employing all high-level features. A similarly high accuracy is achieved by using features from McGhee et al. (2011). However, in the case of the NPS chats, these features are not as reliable and achieve only up to 69% accuracy, which is lower than the accuracy achieved by some of the low-level features (character trigrams) and worse than the accuracy of other feature groups. In the case of the Fixated discourse feature, it was not surprising to see better results for the setting of PJ vs. NPS than PJ vs. cybersex, since this feature represents the length of the lexical

[3] http://www.csie.ntu.edu.tw/ cjlin/liblinear/.

Table 4
Classification using one feature group.

	Keep one feature group				
	Emotional features	McGhee features	Neuroticism features	Fixated discourse	Other
PJ vs. NPS	0.69	0.64	0.71	**0.80**	0.70
PJ vs. cybersex	**0.88**	0.87	0.64	0.62	0.52

The maximum achieved accuracy is in bold.

Table 5
Classification when excluding one feature group.

	Remove one feature group				
	Emotional features	McGhee features	Neuroticism features	Fixated discourse	Other
PJ vs. NPS	0.86	0.90	0.89	0.78	**0.92**
PJ vs. cybersex	0.94	0.89	**0.98**	0.94	0.97

The maximum achieved accuracy is in bold.

chains related to sex. Thus, we assume the sex related lexical chains can discriminate with high accuracy in the former setting where the NPS data has presumably less sex related content, than the other two datasets, while in the latter setting they loose this discrimination power as both corpora, PJ and cybersex, have a high content of sex related discussions.

Table 5 presents the results of classification excluding one feature group at a time. Surprisingly, excluding some feature groups increased the overall accuracy of the approach. In particular, including the last group (Other) that contains imperatives and emoticons, decreases the accuracy in both test scenarios. This could be due to the genre of the corpora used. All three datasets contain online chat conversations, and the use of emoticons in this genre is widespread and not specifically related to soliciting sex from minors. We also noticed that while the features used to detect the neuroticism levels by themselves achieve better results than other groups (see Table 4 columns 4 and 6), they are not helpful when the other groups are present. Excluding either one of the first two groups, emotional features or features from the study of McGhee et al. (2011), slightly improves the results on the PJ vs. NPS setting. But in the case of the cybersex logs data, the accuracy decreases, with a more noticeable decrease when removing McGhee features. Because the features in the work by McGhee et al. (2011) were selected with the PJ dataset in mind, it is not surprising to see this drop in accuracy when removing these features. These differentiated patterns for the two negative datasets used highlight the need for domain specific feature selection, as the nature of the datasets used will affect the effectiveness of the features. But more important, for our problem of automatic detection of pedophiles in chats, it shows that when there is a significant presence of sex-related discussions in the data, it becomes more critical to rely on behavioural and emotion based features inspired by research on pedophilia.

Taking into account the results presented above, we have performed classification with the most promising combinations of feature groups. Table 6 presents the results of classification based on subsets of features. To compare the results, we also show the accuracy of all high-level features and character trigrams. The emotional features and the McGhee features outperform all the features in the cybersex logs data. However in the NPS data they do not provide acceptable performance. Adding the fixated discourse feature significantly improves the accuracy on the NPS chats, and in the case of the other dataset brings only 1% improvement.

Table 6
Classification using feature groups combinations.

	Emotional +McGhee	Emotional +McGhee +Fix. discourse	All high-level	Character trigrams
PJ vs. NPS	0.67	0.87	0.81	**0.97**
PJ vs. cybersex	0.96	**0.97**	0.94	0.64

The maximum achieved accuracy is in bold.

7. Discussion and error analysis

We have conducted experiments on detecting pedophiles in chats with a binary classification algorithm. In the experiments, we used two negative datasets of different nature. The first one, cybersex logs, is more appropriate for our task. It contains one-on-one cybersex conversations. The second dataset was extracted from the NPS chat corpus and contains logs from chat rooms, and, therefore, is less appropriate since the conversations are not necessarily sex related and include multi-party interactions.

It is reasonable to expect that in the case of the negative data consisting of cybersex logs, distinguishing cyberpedophiles is a harder task, than in the case of the NPS data. The results obtained with the baseline systems support this assumption: we obtain very high accuracy for the NPS chats using only character bigrams and trigrams, while the cybersex logs and the pedophiles' conversations are not distinguishable with these low-level features. This is probably because in the NPS vs. PJ setting, the fact that there is a sex-related topic in the discussion is by itself a high indicator of a pedophile. But in the other setting, PJ vs. cybersex logs, the topic correlation is no longer informative as a lot of the content is related to sex, and therefore features that model behaviour and emotional states are more promising, while low level features show a mediocre performance, 64% accuracy. The best accuracy of 97% on the cybersex data is achieved by combining emotional, fixated discourse features and those from McGhee et al. (2011).

The feature ablation experiments showed some interesting findings. First, we learned that the use of emoticons is not useful, so better performance in both settings is achieved when we remove this feature. Although we first thought this to be a promising group, since it has been successfully used in many tasks such as sentiment analysis in social media, it is likely that their pervasive use is so widespread that their presence or absence is not a strong indicator of the behaviour of pedophiles. Perhaps the appearance of a specific set of emoticons (those signalling anger, for instance) could be more informative.

The analysis of the misclassified data revealed that there are several common cases when the algorithm fails to perform correct classification.

- Positive examples with very low level of information words.
- NPS conversations with long sex-related chains
- cybersex conversations with very low level of negative words and high level of positive words

In most of the cases the misclassified instances had one or more features with very unusual values, e.g. positive examples with a very low level of words expressing information sharing.

In the case of the NPS data, a number of misclassified examples were relatively short and contained many sex-related terms, which means the feature modelling lexical chains had similar values as that for PJ. The classifier was giving too much weight to this feature and therefore these instances were mislabelled. With respect to the cybersex logs, a few misclassified examples were different in the sentiment frequencies from other cybersex logs. As it was described earlier, conversations with cyberpedophiles in general contained more positive and less negative words, while in the case of cybersex logs the opposite was much more common. Probably that was the reason that a few cybersex conversations were misclassified.

8. Conclusions and future work

This paper presents the results of a research project on the detection of cyberpedophilia. Following the clues given by research in psychology, we have suggested a list of high-level features that aim to model the level of emotional instability of pedophiles, as well as their feelings of inferiority, isolation, loneliness, and low self-esteem. We have considered as well such low-level features as character bigrams and trigrams and word unigrams, bigrams and trigrams. The SVM classification based on combinations of high-level features achieves 97% accuracy in distinguishing conversations with cyberpedophiles from cybersex chat logs, whereas low-level features achieve only 50–64% on the same data. In case of the common chat conversations (the NPS data), the low-level features, character trigrams in particular, are the most discriminative.

Here we have presented experiments on two toy datasets, but the obtained results give some clues for solving the real-world problem. The most reasonable way could be first, to find suspicious conversations (sex related topics) with low-level features, and then apply high-level features to identify cyberpedophiles among them.

For future work we want to gather a much larger data set and if possible one with actual victims involved. That will give us the opportunity to try to model the mental state of the victim with a similar approach to what we have used here. Being able to predict the vulnerability of children and young adults in social media interactions could also have large impact on society as it can be used to trigger intervention strategies to prevent cyberpedophilia. In addition, having the ability to model both, the potential victim and the pedophile can increase the detection rate of pedophiles.

Acknowledgements

The work of Dasha Bogdanova was partially carried out during the internship at the Universitat Politècnica de València (scholarship of the University of St. Petersburg). Her research was partially supported by Google Research Award. The collaboration with Thamar Solorio was possible thanks to her one-month research visit at the Universitat Politècnica de València (program PAID-PAID-02-11 award n. 1932). The research work of Paolo Rosso was done in the framework of the European Commission WIQ-EI Web Information Quality Evaluation Initiative (IRSES Grant No. 269180) project within the FP 7 Marie Curie People, the DIANA-APPLICATIONS – Finding Hidden Knowledge in Texts: Applications (TIN2012-38603-C02-01) project, and the VLC/CAMPUS Microcluster on Multimodal Interaction in Intelligent Systems.

References

Abel, G.G., Harlow, N., 2001. The Abel and Harlow Child Molestation Prevention Study. Xlibris, Philadelphia.

Argamon, S., Koppel, M., Pennebaker, J., Schler, J., 2009. Automatically profiling the author of an anonymous text. Communications of the ACM 52 (2), 119–123.

Baccianella, S., Esuli, A., Sebastiani, F., 2010. Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. The Seventh Conference on International Language Resources and Evaluation.

Barzilay, R., Elhadad, M., 1997. Using lexical chains for text summarization. In: Proceedings of the ACL Workshop on Intelligent Scalable Text Summarization, pp. 10–17.

Bogdanova, D., Rosso, P., Solorio, T., 2012. Modelling fixated discourse in chats with cyberpedophiles. In: Proceedings of the EACL Workshop on Computational Approaches to Deception Detection.

Cortes, C., Vapnik, V., 1995. Support-vector networks. Machine Learning 20 (3), 273–297.

Egan, V., Hoskinson, J., Shewan, D., 2011. Perverted justice: a content analysis of the language used by offenders detected attempting to solicit children for sex. Antisocial Behavior: Causes, Correlations and Treatments 20 (3), 273–297.

Forsythand, E.N., Martell, C.H., 2007. Lexical and discourse analysis of online chat dialog. In: International Conference on Semantic Computing ICSC 2007, pp. 19–26.

Galley, M., McKeown, K.,2003. Improving word sense disambiguation in lexical chaining. In: Proceedings of the 18th International Joint Conference on Artificial Intelligence. IJCAI'03. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, pp. 1486–1488.

Hall, R.C.W., Hall, R.C.W., 2007. A profile of pedophilia: Definition, characteristics of offenders, recidivism, treatment outcomes, and forensic issues. In: Mayo Clinic Proceedings.

Hope, D., 2008. Java Wordnet Similarity Library. http://www.cogs.susx.ac.uk/users/drh21

Leacock, C., Chodorow, M., 1998. Combining local context with WordNet similarity for word sense identification. In: WordNet: A Lexical Reference System and Its Application.

McGhee, I., Bayzick, J., Kontostathis, A., Edwards, L., McBride, A., Jakubowski, E., 2011. Learning to identify internet sexual predation. International Journal on Electronic Commerce 15 (3).

Morris, J., Hirst, G., 1991. Lexical cohesion computed by thesaural relations as an indicator of the structure of text. Computational Linguistics 17 (March (1)), 21–48.

Panchenko, A., Beaufort, R., Fairon, C., 2012. Detection of child sexual abuse media on p2p networks: normalization and classification of associated filenames. Language Resources for Public Security Applications, 27.

Pedersen, T., Patwardhan, S., Michelizzi, J., Banerjee, S., 2008. Wordnet:similarity. http://wn-similarity.sourceforge.net/

Peersman, C., Daelemans, W., Van Vaerenbergh, L., 2011. Predicting age and gender in online social networks. In: Proceedings of the 3rd International Workshop on Search and Mining User-generated Contents. SMUC'11, ACM, New York, NY, USA, pp. 37–44.

Pendar, N., 2007. Toward Spotting the Pedophile: Telling Victim From Predator in Text Chats. Irvine, California, pp. 235–241.

Resnik, P., 1995. Using information content to evaluate semantic similarity in a taxonomy. In: Proceedings of the 14th International Joint Conference on Artificial Intelligence, pp. 448–453.

Snyder, H., 2000. Sexual assault of young children as reported to law enforcement: Victim, incident, and offender characteristics. A NIBRS statistical report. Bureau of Justice Statistics Clearinghouse.

Strapparava, C., Mihalcea, R., 2007. Semeval-2007 task 14: affective text. In: Proceedings of the 4th International Workshop on Semantic Evaluations, SemEval'07, pp. 70–74.

Strapparava, C., Valitutti, A., 2004. WordNet-Affect: an Affective Extension of WordNet. In: Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004), Lisbon, pp. 1083–1086.

Vandiver, D.M., Kercher, G., 2004. Offender and victim characteristics of registered female sexual offenders in Texas: a proposed typology of female sexual offenders. Sex Abuse 16, 121–137.

Villatoro-Tello, E., Juárez-González, A., Escalante, H.J., Montes-y-Gómez, M., Villase nor-Pineda, L., 2012. A two-step approach for effective detection of misbehaving users in chats. In: Proceedings of CLEF 2012.

Wolak, J., Mitchell, K., Finkelhor, D., 2003. Escaping or connecting? Characteristics of youth who form close online relationships. Journal of Adolescence 26 (1), 105–119.

World Health Organization, 1988. International Statistical Classification of Diseases and Related Health Problems: Icd-10 Section f65.4: Paedophilia.