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,