Bogdanova, D.; Rosso, P.; Solorio, T. (2014). Exploring high-level features for detecting cyberpedophilia. Computer Speech and Language. 28(1):108-120. doi:10.1016/j.csl.2013.04.007.

Exploring High-Level Features for Detecting Cyberpedophilia

Dasha Bogdanova[a,1], Paolo Rosso[b], Thamar Solorio[c]

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

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 www.perverted-justice.com and two negative datasets of different nature: cybersex logs available online and the NPS chat corpus. The SVM classification results show that the NPS data and the pedophiles' conversations can be accurately discriminated with character n-grams, while in the more complicated case of cybersex logs high-level features 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

[1] Email: dasha.bogdanova@gmail.com

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 in the Internet usually involves very fast typing. The data usually contains a large amount of mistakes, misspellings, specific slang, character flooding etc. 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 dimesions 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 pedophiles' chats from non-pedophiles' 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 the 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; they are followed by discussion in Section 7. 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), who 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 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.

As it was already mentioned, pedophiles often create false profiles and pretend to be younger or of another gender. Moreover, they try to copy children's behavior. 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". Messages

were not longer than 150 messages and the percentage of predators was lower than 4%. The system that achieved the highest performance (Villatoro-Tello, 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 semantic dimensions, i.e. high-level features based on clues provided by psychology and sentiment analysis, helpful in distinguishing between pedophile and non-pedophile's chats.

3. Profiling the Pedophile

Pedophilia is a "disorder of adult personality and behavior" 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 behavior 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. On the one hand, pedophiles shift gradually to the sexual conversation, starting with more ordinary compliments:

  Offender: hey you are really cute
  Offender: u are pretty
  Offender: hi sexy

  On the other hand, the conversation then becomes 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 pedophiles 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
  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)

The most interesting from our research point of view is data of the type 1(a), but obtaining such data is not easy. However, the data of the type 1(b) is freely available at the web site www.perverted-justice.com. For our study, we have extracted chat logs from the perverted-justice (PJ) website. Since the victim is not real, and our goal is to learn the patterns of cyberpedophiles, we considered only the chat lines written by pedophiles.

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 [2]. From each dataset we randomly selected 20 files for testing.

The cybersex chat logs were downloaded from `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

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

Corpus name	**Positive or negative**	**Number of files in training data**	**Number of files in testing data**
Perverted-justice: subset 1	positive	40	20
Perverted-justice: subset 2	positive	40	20
Perverted-justice: subset 3	positive	40	20
Perverted-justice: subset 4	positive	40	20
Perverted-justice: subset 5	positive	40	20
Subset of NPS chat corpus	negative	45	20
Cybersex logs	negative	48	20

Table 1: Statistics on the experimental data.

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 (Forsyth and Martell, 2007), even though it is not of type 2, to explore how the 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 were much longer, with an average number of words and lines equal to 3618 and 526 respectively. While these numbers for the cybersex data were 1428 and 97. The NPS data had even shorter conversations. The average numbers of words and lines were 225 and 52. Balancing the data by reducing the conversations to the same size was not an option, because 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 pedophiles' chats from non-pedophiles' 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 caugth sometime

In this example the pseudo-victim is not answering:

Offender: hello
Offender: r u there
Offender:
Offender: thnx a lot
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 [3]. The frequencies of each type of markers are presented in Figure 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;

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

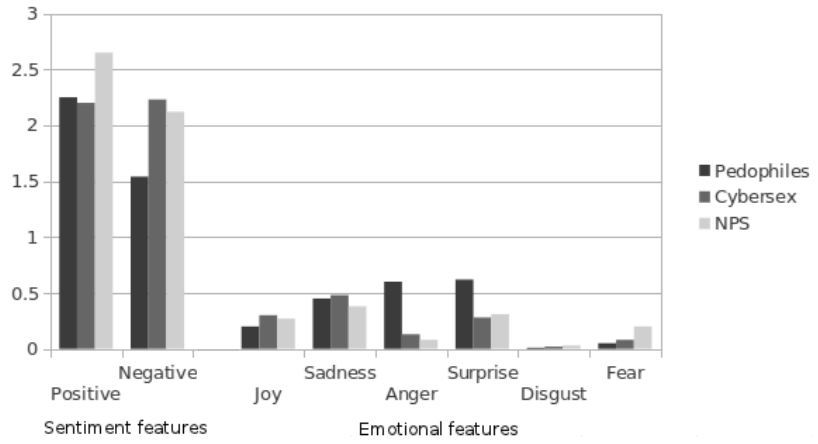

Figure 1: Average number of sentiment and emotional markers found for the three corpora.

- 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 such nouns 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*).
- 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.

Feature Class	**Feature**	**Example**	**Resource**
Sentiment and Emotional Markers	Positive Words	*cute, pretty*	SentiWordNet Baccianella et al. (2010)
	Negative Words	*dangerous, annoying*	
	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!*	

Table 2: Features used in the experiments.

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.

The semantic similarity measure that was proposed by Resnik (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

The experiments we performed included three major steps. First, experiments on finding an appropriate model for fixated discourse, one of the main characteristics of conversations with cyberpedophiles. Second, experiment on discriminating cyberpedophiles with classification techniques and high-level features described in above. Among the high-level features, there was a fixated discourse feature modelled according to the experimental results obtained in the first step. Third, we have performed feature analysis, where we have experimentally found which feature groups are more discriminative.

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)".

	Threshold			
	0.5		0.7	
	mean	**st.dev.**	**mean**	**st.dev.**
PJ	12.21	3.63	9.3	5.68
Cybersex	18.28	16.8	9.98	12.76
NPS	5.66	5.9	2.42	4.77

Table 3: Average length of the longest lexical chain (percentage in the total number of words) computed with Leacock and Chodorow's semantic similarity.

	Threshold			
	0.5		0.7	
	mean	**st.dev.**	**mean**	**st.dev.**
PJ	8.24	4.51	6.68	5.06
Cybersex	12.04	15.86	9.13	11.64
NPS	0.67	0.96	0.41	0.66

Table 4: Average length of the longest lexical chain (percentage in the total number of words) computed with Resnik's semantic similarity.

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 Table 3 and Table 4. As we expected, sex-related lexical chains in the NPS corpus are much shorter regardless 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.

We have used as a feature the length of the lexical chain constructed with the Resnik similarity measure (Resnik, 1995) with the threshold = 0.7, as it is seen from Table 4, it is good in discriminating all the datasets one from another.

6.2. Pedophiles Detection

To distinguish between pedophiles and not pedophiles we used an SVM classifier. To extract positive and negative words, we used SentiWordNet (Baccianella et al., 2010). The features borrowed from McGhee et al. (2011), were detected with the list of words authors made available for 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 contained 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 5. As it can be seen from the table, the NPS chats that are not sex-related in general, could be easily distinguished from cyberpedophiles' convesations with low-level features. SVMs based on character trigrams achieves 97% accuracy, while high-level features show only 81%. In case of 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 behavior 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 the following section we study the performance of different features groups.

6.3. Feature Analysis

The goal of this research is focused mostly in revealing semantic dimensions that could help in the detection of cyberpedophiles. In Section 5.1 we have suggested five groups of features. In this section we try to evaluate

	High-level features	Low-level features (baseline)				
		Bag of words	Term bigrams	Term 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

Table 5: Results of SVM classification.

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 6. From that table it is clear that emotional features (positive and negative words, and basic emotion markers) on their own distinguish cyberpedophiles' chats from cybersex chats with quite high accuracy that is only 6% lower than the accuracy while 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 still much worse than the accuracy achieved by low-level features (character trigrams) and worse than the accuracy of other feature groups. The fixated discourse group contains only one feature and thus not surprisingly, it is not enough for achieving a reliable performance on the PJ vs. Cybersex setting. In contrast, due to such big differences between lexical chains lengths in the NPS data and in the cyberpedophiles' logs, described in Section 6.1, this only feature distinguishes them with very high accuracy.

Table 7 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 contained imperatives and emoticons decreases the accuracy in both the NPS and the cybersex dataset cases. This could be due to the genre of the corpora used. All three datasets contain some type of online chat conversations, and the use of emoticons on this genre is widespread and not particularly related to soliciting sex from minors. Although he features used to detect the neuroticism levels by themselves achieve better results than other groups (See Table 6 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 case of the cybersex logs data, the

	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

Table 6: Classification using 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

Table 7: Classification when excluding one feature group.

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. This 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.

Taking into account the results presented above, we have performed classification with the most promising combinations of feature groups. Table 8 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 1% improvement.

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

	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

Table 8: Classification using feature groups combinations.

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, as opposed to one-on-one.

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. The maximal accuracy in this case is only 64%, but our proposed high-level features are able to boost accuracy to 94% on this dataset. In case of NPS data they do not outperform character n-grams and achieve only 81% accuracy.

The analysis of feature performance revealed that some features are more discriminative while others could be rather misleading for the algorithm. In particular, the features that did not appear to be discriminative are the frequency of emoticons and the features used in previous research on neuroticism level detection. 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 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 missclassified instances had one or more features with very unusual value, e.g. positive examples with very low level of words

expressing information sharing. In case of the NPS data, a number of missclassified examples were relatively short and containing many sex-related terms, so the lexical chain modelling the fixated discourse feature appeared to be as long as in case of cyberpedophiles chats. In this case if we changed this value to a lower one, the algorithm classified these instances correctly. With respect to cybersex logs, a few missclassified examples were different in 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 case of cybersex logs the opposite was much more common. Probably that was the reason that a few cybersex conversations were misclassified. Switching the values of positive and negative markers made the algorithm to make correct predictions for this instances.

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 real-world problem. The most reasonable way could be first, to find suspicious conversations with low-level features, and then apply high-level features to identify cyberpedophiles among them. In the future we plan to conduct experiments on bigger datasets.

The feature extraction we have implemented does not use any word sense disambiguation. This can as well cause mistakes. Consider for example the lemma "fit" that can be either a positive marker ("a fit candidate") or negative ("a fit of epilepsy"), depending on the context. Therefore, we also plan to employ word sense disambiguation techniques during the feature extraction phase.

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 IRSES project (grant no. 269180) within the FP 7 Marie Curie People, the MICINN research project TEXT-ENTERPRISE 2.0 TIN2009-13391-C04-03(Plan I+D+i), 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. Philadelphia, Xlibris.

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: 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. Proceedings of the EACL Workshop on Computational Approaches to Deception Detection.

Campbell, R. S., Pennebaker, J. W., Jan. 2003. The Secret Life of Pronouns. Psychological Science 14 (1).

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.

Forsythand, E. N., Martell, C. H., 2007. Lexical and discourse analysis of online chat dialog. International Conference on Semantic Computing ICSC 2007, 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. 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., Mar. 1991. Lexical cohesion computed by thesaural relations as an indicator of the structure of text. Comput. Linguist. 17 (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: 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, 70–74.

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., Villaseñor-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), 105–19.

World Health Organization, 1988. International statistical classification of diseases and related health problems: Icd-10 section f65.4: Paedophilia.