Identifying Predators Using ChatCoder 2.0
Notebook for PAN at CLEF 2012

April Kontostathis+, Will West*, Andy Garron~, Kelly Reynolds*, Lynne Edwards+

+ Ursinus College, * Lehigh University, ~ The University of Maryland
akontostathis@ursinus.edu

Abstract. This article describes our experiments for the Sexual Predator Identification tasks at PAN2012. We have previously developed a software application, ChatCoder, for the identification of predatory posts in an online conversation. This paper extends this research to the detection of authors in addition to individual lines of text. We show that we were able to detect up to 98% of the predatory authors in the training data and 87% of the authors in the test set, using our fully automated system. Our recall is high, but it comes at a cost, as we return many false positives. This article describes our experimental method and results, and suggests improvements to our system that will improve precision without hurting recall.

1 Introduction

The PAN2012 Sexual Predator Identification competition is comprised of two subtasks: the first is identification of userids (anonymized) that are "owned" by Internet sexual predators; the second is identification of sexually explicit or predatory posts. We participated in both subtasks.

Our primary interest was to determine if our existing application software, ChatCoder 2.0 [2] could be adapted for use to identify predatory authors. ChatCoder 2.0 was previously limited to classification of individual posts, thus the contribution of this paper is the expansion from detection of individual lines to detection of authors based on a body of text. In Section 2 of this article we give an overview of the ChatCoder project in general and ChatCoder 2.0 specifically. In Section 3 we describe our experimental method, including our efforts to apply machine learning techniques for the detection of Internet sexual predators, using features that were extracted from the training data using ChatCoder 2.0. Sections 4 and 5 present our results and analysis. Conclusions follow in Section 6.

2 Background

Crimes against children receive a lot of attention in the popular press. Increasingly these crimes are facilitated or perpetrated using the Internet [3]. This threat is of

particular interest to researchers, law enforcement, and youth advocates because of the potential for it to get worse as membership in online communities continues to grow and as new social networking technologies emerge [5]. Despite safeguards put in place by some social networking sites, underage children still use these sites, exchanging personal information and photos with friends and strangers [6].

To address these ongoing concerns, we embarked on an interdisciplinary approach for studying cyber aggression, in particular the communicative patterns of cyber predators and their victims. This project, ChatCoder, led to the development of labeled collections for studying this problem, and formulated and operationalized communication theories. These resources and theories have been used to develop computer algorithms for detecting cyber crimes. The project home page is http://www.chatcoder.com.

In previous work, we used machine learning approaches, combined with an in-depth study of communicative patterns, to identify posts that fall into one of three categories that are often used by cyber predators when they communicate with their victims [2]. The three categories of interest are: Personal Information Exchange, Grooming, and Approach. In [2] we are able to correctly assign class labels to individual lines of a conversation approximately 63% of the time using both a custom made rule-based learner (ChatCoder 2.0) and various machine learning algorithms. Both ChatCoder 2.0 and the machine learning algorithms are dependent on dictionary and a set of 15 attributes which were collected for each post. The attributes are:

- Total number of words in a line (words are defined as strings of characters separated by white space)
- Number of first person pronouns in a line (e.g. I, me)
- Number of second person pronouns in a line (e.g. you, your)
- Number of third person pronouns in a line (e.g. he, them)
- Number of personal information nouns (e.g. age, pic)
- Number of relationship nouns (e.g. boyfriend, date)
- Number of activities nouns (e.g. movie, favorite)
- Number of family nouns (e.g. mom, sibling)
- Number of communicative desensitization verbs (e.g. kiss, suck)
- Number of communicative desensitization nouns (e.g. bra, orgasm)
- Number of communicative desensitization adjectives (e.g. horny, naked)
- Number of communicative desensitization words (e.g. sex, penis)
- Number of reframing verbs (e.g. teach, practice)
- Number of approach verbs (e.g. meet, see)
- Number of approach nouns (e.g. hotel, car)

See [2] for a detailed description of how these attributes are used to determine if a post is labeled as Personal Information Exchange, Grooming, or Approach

3 Experimental Design

In this section we describe, in detail, the process we used to develop our submission to the PAN2012 competition.

3.1 Preprocessing

We first separated the training data by author, creating one file per author (for convenience). This allowed us to quickly and easily collect statistics at the author level. The provided truth set was incorporated into the output, so we knew which authors were predators. We used the ChatCoder 2.0 rules to extract two sets of attributes for each author: one that used the summary labeling of lines (ex. number of personal information lines, number of grooming lines, number of approach lines), and one that drilled down a level, using the 15 attributes.

For the first set of experiments (ChatCoder 2.0), we generated a count of the number of lines for each author that were labeled as containing personal information, grooming, and approach (we refer to these as 200, 600, and 900 lines). We also counted the total number of lines for a given author, and used this to calculate the percentage of 200, percentage of 600, percentage of 900, percentage that were either 600 or 900, and percentage of lines coded. The percentages and counts were both used as features in our machine learning experiments.

In the second set of experiments (ChatCoder 2.0DD), we drilled down a level, and extracted a count at the author level for each of the 15 attributes that were used in our ChatCoder 2.0 rules. For example, we counted the number of first person pronouns, the number of communicative desensitization words, etc. used by each author.

The extracted datasets were used as input to our machine learning tool.

3.2 Machine Learning

We used the Weka data mining tool kit [7] to identify the predatory authors. We were interested in deterministic algorithms that could easily be understood by a human, and could be implemented in code, so we used only the C4.5 decision tree learner (implemented as J48 in Weka)[4], and the RIPPER rule-learning algorithm using N minimum occurrences of a given rule (implemented as JRip in Weka)[1], for our learning experiments.

The dataset size used in the competition was too large to be handled by Weka on the systems we used, and the number of true positives was very sparse (only 142 positive instances out of 97,689 authors in the training files were marked as predators). Thus, we needed to weight our positive instances more heavily and sample our negative instances. We produced a dataset containing approximately 25,000 records, which were approximately evenly split between positive and negative instances. This sampling was done twice, with different seeds, in order to ensure that we did not inadvertently end up with an unusual sampling, which is possible when random processes are used. We also adjusted our learning algorithm parameters, eventually choosing settings that required a minimum number of items at each leaf or captured in each rule.

We identified a compact C4.5 decision tree with only a few leaves (Figure 1) for the ChatCoder 2.0 experiment. We set the parameters so that a minimum of 100 instances had to be covered by each leaf. We also discovered a set of 11 rules from JRIP (Figure 2) as well as a C4.5 tree (Figure 3) for the ChatCoder 2.0DD data, both

requiring a minimum of 50 items at each leaf or rule. These three classification systems were implemented in code.

```
If ApproachCount <= 1
        If ApproachPercent <= 3.78:  Not Predatory
        Else If GroomingOrApproachPercent <= 4:  Predatory
            Else:  Not Predatory
Else    If ApproachCount <= 4
            If GroomingCount <=1
                    If ApproachPercent <= 26.67:  Not Predatory
                    Else:  Predatory
                Else:  Predatory
            Else:  Predatory
```

Figure 1: Decision Tree for ChatCoder 2.0 experiments

Predatory if:
1: (approachverb >= 4) and (secondpronoun >= 24) and (familynoun <= 6) and (approachverb >= 13)
2: (approachverb >= 4) and (secondpronoun >= 23) and (commadj >= 3) and (commverb >= 35)
3: (approachverb >= 4) and (secondpronoun >= 20) and (familynoun >= 1) and (firstpronoun <= 61) and (secondpronoun >= 36)
4: (approachverb >= 4) and (approachverb >= 7) and (secondpronoun >= 15) and (thirdpronoun <= 28) and (commverb <= 8)
5: (approachverb >= 4) and (secondpronoun >= 23) and (firstpronoun <= 28) and (commverb >= 9)
6: (approachverb >= 4) and (secondpronoun >= 94) and (secondpronoun <= 159)
7: (firstpronoun >= 3) and (approachverb >= 4) and (commnoun <= 0) and (infonoun <= 1) and (thirdpronoun <= 9)
8: (linkingverb >= 9) and (relationshipnoun <= 0) and (commnoun >= 7) and (commdesens >= 4)
9: (linkingverb >= 10) and (thirdpronoun <= 9) and (approachnoun >= 1)
10: (thirdpronoun <= 0) and (actionverb >= 3) and (secondpronoun <= 1) and (firstpronoun >= 3)
Otherwise **Not Predatory**

Figure 2: Rule set for ChatCoder 2.0DD (JRIP) experiments

```
If approachverb <= 3
        If linkingVerb <= 9
            If actionVerb <= 2:  Not Predatory
            Else if thirdPronoun<=0:  Predatory
                Else:  Not Predatory
        Else:  Predatory
Else:  Predatory
```

Figure 3: Decision Tree for ChatCoder 2.0DD (J48) experiments

An analysis of the trees and rule sets shows that the approach and grooming (communicative desensitization) categories are the most important attributes (Figures 1, 2, 3). Additionally, in the J48 tree (Figure 3) and JRIP rule set (Figure 2) that were created for the ChatCoder 2.0DD experiments, we note that the number of second person pronouns appears to indicate a higher chance of predation. The fine grained breakdown of the three original categories gives us a better understanding of why each of the attributes from ChatCoder 2.0 tree (Figure 1) are the most influential.

We also identified all of the authors that participated in conversations either by themselves (no one else in the chatroom) or with multiple other participants (3 or more chatters). These authors were marked as non-predators in all experiments

except for the original ChatCoder 2.0 run. The rationale for this decision is that we believe only one-on-one conversations could be predatory in nature, as the presence of additional parties would deter predators. A predatory conversation with only one participant is trivially impossible. Furthermore, all of the positive examples in the training data were authors that participated in conversations with exactly two parties.

The second phase of the task was to identify the lines that were indicative of predation. For all runs, we extracted the 600 (Grooming) and 900 (Approach) labeled lines for submission to this subtask. We extracted only the 600 and 900 lines for the authors that we labeled as predators. We discuss the results from the second subtask in Section 5.

3.3 Determining Submission Runs

Before identifying the runs to submit, we calculated accuracy statistics using the instances in the full training set. The results of these experiments are shown in Table 1 (lines 1-3). *Precision* is the number of true positives divided by the number labeled as predatory by our learning algorithms. *Recall* is the number of true positives divided by actual the number of predators (142). The run with only the authors who participated in one-on-one conversations is shown in line 4 of Table 1.

Line	Run	Num Marked as Pred	NumTP	NumFP	Data Set Size	Precision	Recall
1	V1 - ChatCoder 2.0	2655	137	2518	97689	0.05	0.96
2	ChatCoder 2.0DD (J48)	1676	139	1537	97689	0.08	0.98
3	ChatCoder 2.0DD (JRIP)	382	139	243	97689	0.36	0.98
4	V2 - ChatCoder 2.0 (only 1-1 conv.)	1786	137	1649	97689	0.08	0.96
5	**JRIP && V2**	298	137	161	97689	0.46	0.96
6	J48 || V2	2657	139	2518	97689	0.05	0.98

Table 1: Statistics collected using the training data for our runs

We had some extra time, and decided that it would be interesting to see if combinations of these algorithms would be useful. As a result we took the intersection and union of each combination of runs. Lines 5 and 6 show the outliers from these experiments (the run with the fewest returned authors, and the run with the most returned authors). The combination of Chatcoder 2.0DD (JRIP) and Chatcoder 2.0 (only 1-1 conv.) had the best performance overall on the training data, and it was chosen as our competition run. All runs in Table 1, except for ChatCoder 2.0DD (J48) were sent for evaluation.

4 Results and Discussion

The results for the first subtask, identification of the predatory authors, are shown in Table 2. As expected, our competition run (J48 && V2), shown in line 5, was our best run, using the competition metric of F1, a metric which provides a balanced trade-off between precision and recall. It also achieved the best precision. The best

recall was given by our J48 || V2 run. Unsurprisingly, it achieves this recall level by returning more authors overall, over 11 times more than our top run. Also clear is that an author must participate in a one-on-one conversation to be labeled as a predator. The number of true positives returned by V1 and V2 are the same, returning fewer authors increases precision, without hurting recall at all.

Line	Run	Num Marked as Pred	NumTP	Precision	Recall	F1	Fbeta0.5
1	V1 - ChatCoder 2.0	5225	206	0.04	0.81	0.08	0.05
2	J48 || V2	5625	221	0.04	0.87	0.08	0.05
3	V2 - ChatCoder 2.0 (only 1-1 conv.)	3696	206	0.06	0.81	0.10	0.07
4	ChatCoder 2.0DD (JRIP)	688	172	0.25	0.68	0.37	0.29
5	**JRIP && V2**	475	170	0.36	0.67	0.47	0.39

Table 2: Statistics returned by the organizers using the test data for our runs

Note that our experiments and submissions were fully automated. We did not manually review the submitted results until after we obtained the list of true predators in the test set from the conference organizers. We were able to then do some manual review and analysis. We focused on a comparison between the true positives (the predators we correctly identified), the false positives (the authors we identified, who were not predators), and the false negatives (the authors we did not identify, who were predators) for our competition run (JRIP && V2).

Table 3 provides a brief look at some descriptive statistics, comparing the true positives to the false positives and false negatives. Clearly our current processing is favoring longer bodies of text, in general. Both the true and false positives have significantly higher counts, when compared to the false negatives. However, the minimum values show that some of the authors with only a few lines of conversation were returned by our system.

A manual review of all of the predator posts in the set of false negatives we had for the JRIP && V2 run detected some interesting patterns. In some cases, we do not see anything predatory in these transcripts. For example, one of the userIDs in this set (5904488cf6bfcd01beaf225ac00efd99) had three lines of text, each containing the word "sup." We believe that over 25% of this false negative set contains cases that are not predatory (based solely on the information provided), or are borderline cases.

However, two significant categories were also identified during this manual review, and we believe that capturing these categories can be used to dramatically improve our system. First of all, we noticed that there were certain words or phrases that indicated an age difference between the participants in a conversation. Comments such as "too old for you," "you are very young," "go to school," "get someone to drive you," etc. are indicative of age, and we are not capturing these in a separate category. Age information is included as part of our 200 category (Personal Information Exchange), but we are now inclined to believe the age comments, particularly comments pertaining to evidence of youthfulness and/or age differences between participants, should stand alone as a separate category.

The second significant category we discovered is something we refer to as "awareness of guilt." Much of the evidence for predation that was found in the false negatives revolves around comments like "are you a cop," "we can get into a lot of trouble for this," "I wish you were older," etc. Again, we capture this language in our

approach category (as evidence of isolation – trying to distance a youngster from his or her support network), but it appears to be particularly significant for detection of predation and should stand alone.

Summary Statistics for True Positives							
	200s	600s	900s	Percent 200s	Percent 600s	Percent 900s	TotalLines
Average	10	38	24	0.03	0.10	0.08	358
Min	0	0	2	0.00	0.00	0.01	10
Max	108	495	181	0.09	0.27	0.30	3189
Summary Statistics for False Negatives							
	200s	600s	900s	Percent 200s	Percent 600s	Percent 900s	TotalLines
Average	2	5	3	0.04	0.07	0.06	61
Min	0	0	0	0.00	0.00	0.00	1
Max	13	40	33	0.50	0.36	1.00	546
Summary Statistics for False Positives							
	200s	600s	900s	Percent 200s	Percent 600s	Percent 900s	TotalLines
Average	29	73	66	0.03	0.13	0.09	1038
Min	0	0	2	0.00	0.00	0.01	3
Max	1079	2291	2319	0.28	0.61	1.00	36689

Table 3: Summary Statistics for Comparison

In our analysis of the false positives, we were not able to review all of the communication for each of the authors we identified, but a quick review of a sampling from this set found some interesting trends. Some of our false positives were clearly just sexual talk, presumably between adults because there was no discussion about age. Other false positives were the victims in a predatory conversation (which we should have been able to easily identify, but clearly just overlooked). We found one or two bullying conversations, and, to our great amusement, several instances of highly technical discussions. Apparently geeks who are frustrated with technology use a lot of terms that can be misconstrued as sexual in nature by our system (ex "BUTTon", and "this browser SUCKS"). We also found at least one transcript that appears to be predatory, based on the text available. In conversations involving author 0a58a246104192e7e7568e5edb90e60c, the author receives confirmation that the girl he is chatting with is only 14, but continues with highly sexual conversation anyway.

As a final note, we disagree with the use of the Fbeta0.5 statistic as proposed by the conference organizers. This version of the F statistic gives higher weight to precision and reduces the contribution from recall. We feel that the justification for this metric is particularly troubling. Reducing the number of false positives will, indeed, require the use of fewer resources for police investigations, but these crimes are so onerous and it is well-known that the perpetrators have abused multiple victims. We would rather expend investigative resources to exclude someone as a criminal, instead of overlooking a single true positive. In fact, we would argue strenuously for the use of recall as the primary statistic, instead of F1.

Another possibility is to change the competition slightly and require teams to provide ranked results rather than binary results. This enhancement would make a large number of cases manageable for real-world criminal investigators. Those cases

which have a high probability of being predatory would appear higher on the list, and would thus receive higher priority from law enforcement.

5 Identification of the Predatory Lines

As mentioned above, we submitted all of the lines which we identified as containing either Grooming (600) or Approach (900) language as evidence of predatory behavior. We submitted these lines for each author we identified as a predator. This section describes our results for the Identification of Predatory Lines subtask.

Our ChatCoder system was designed specifically to detect predatory lines, rather than authors, and we were pleased with the results we achieved on this subtask. We submitted 19535 lines as evidence of predatory behavior (again automatically generated by our system). Of these 3249 were identified as predatory by the subject matter expert who did the evaluation for this subtask. Our recall was fairly low, at .17, but recall was significantly better at .50. This is particularly interesting because we only found 170 of the 254 predatory authors on this run. Thus, even though we are missing almost a third of the predatory authors, we are still finding approximately half of the predatory lines.

After the truth set was released for this subtask, we reran our process, using the 254 authors that were identified as predators. The results are quite different, as shown in Table 4. The run (still fully automated) with all 254 predatory authors tells a interesting story. Recall remains the almost the same (increasing by just .03). Precision, however, almost doubles from .17 to .31. Using the competition metrics, the F1 for this run was .39 and the F3 was .45. The small difference in recall confirms our system's bias against short conversations.

Comparison of Submitted Predatory Authors to All Predatory Authors				
	Num Lines Identified	Num True Positives	Precision	Recall
Submitted Run	19535	3249	0.17	0.50
Run with 254 Predatory Authors	11104	3428	0.31	0.53

Table 4: Precision and Recall, submitted vs. maximum

We briefly looked at the lines that were identified as true positives, false positives and false negatives for our submitted run. The 3249 lines that we correctly identified as evidence of predatory behavior are almost all very explicit sexual remarks. There are also few remarks that are more romantic in nature ("I love you"), and a few that are associated with isolation ("when does mom or dad get home").

The 3229 false negatives are more interesting. These are lines that our system did not identify as predatory, but which were found to be predatory by the subject matter expert. Only 179 of these lines would have moved into the "true positive" column if we had correctly identified all of the predatory authors. These lines seem to fall into a few different categories. Many are subtly rather than explicitly sexual ("what are you wearing to bed"). Some are discussions of age of the participants ("you are 15 and I'm 44") and of concerns about the illegality of this interaction ("don't get caught together"). These remarks lend more weight to our belief that our system can be

improved dramatically if we develop new categories for age discussion and awareness of guilt. These two additional categories would allow us to identify specific lines as well as authors more accurately. We also noticed that 78 of the false negatives are "EDIT PORN LINK" that our system would most likely never detect because it is meant to handle real time communication, which would not be edited.

The 16286 false positives are harder to analyze and categorize. There are many technical remarks which are being categorized as grooming or approach by our system. There are also comments from younger chatters ("get busted by my parents"), which may or may not be from victims in a predatory conversation. The comments that are most interesting are those that are part of predatory conversations, but which pertain to approach. We have quite a few rules for approach in our system, everything from asking for a phone number, or for a call from the victim, to making specific arraignments to meet. From a legal perspective, attempts to meet are particularly disturbing because the predatory is attempting to move out of the virtual world, into physical victimization. Thus, we believe comments such as ("do you want me to come and get you") are particularly inflammatory, and should be considered as evidence of predation when it is clear from context that an adult author expects the meeting to result in a sexual interaction with a teen or tween.

6 Conclusions

In this article we present our results from the PAN2012 competition. Our system did not do as well as we would like, and we have determined from a brief analysis of the false positives and false negatives that enhancements are needed. In particular, we need to develop a category pertaining to age, specifically to discussion of age differences or youthfulness in an online conversation that has sexual content. Also, we need to provide markers to identify when an author is aware of his criminality, based solely on elements in the conversation itself.

Some additional enhancements to reduce the number of false positives are also needed. Specifically we need to remove any conversations that are clearly not sexual at all, such as technical discussions, and also remove authors who are participating in predatory conversations because they are the victims, not the predators.

Although we were more successful when attempting to identify predatory lines, we still find that there is a lot of room for improvement. Our preliminary analysis confirms the need for the two additional categories mentioned above, and also a need within the research community for a discussion of what specifically do we mean by predatory conversation and what specific types of communication are of particular interest to law enforcement agencies.

The PAN2012 Sexual Predator Identification competition is an interesting and important research opportunity. It was particularly rewarding to see the large number of participants. We look forward to additional opportunities for research groups to work collaboratively to develop solutions for the automatic detection of crimes against children. Because criminals are increasingly using Internet resources to find new victims, we need to develop technology to ensure that we can protect children to the greatest extent possible.

Acknowledgements
We would like to thank the organizers of the PAN labs for their work on conceiving and coordinating this competition. This material is based upon work supported by the National Science Foundation under Grant No. 0916152. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Bibliography
[1] Cohen, W.W. (1995). Fast Effective Rule Induction. In: *Proceedings of the Twelfth International Conference on Machine Learning*, 115-123, 1995.
[2] McGhee, I., J. Bayzick, A. Kontostathis, L. Edwards, A. McBride, and E. Jakubowski. (2011). Learning to Identify Internet Sexual Predation. *International Journal on Electronic Commerce*. Volume 15, Number 3. Spring 2011.
[3] National Center for Missing and Exploited Children. Reviewed Dec 2011. http://www.missingkids.com/en_US/documents/CyberTiplineFactSheet.pdf.
[4] Quinlan, R. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers, San Mateo, CA.
[5] Smith, A. 2011. Why Americans use social media. *Pew Internet and American Life Project*. November 15, 2011. http://pewinternet.org/Reports/2011/why-Americans-use-social-media.aspx.
[6] Weichselbaum, S. and E. Durkin. (2011). Facebook lures youngsters with parents' OK. *NYDailyNews.com*. Posted December 11, 2011.
[7] Witten, E. and I. Frank. (2005). *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann Publishers.