Conversation Level Constraints on Pedophile Detection in Chat Rooms

Notebook for PAN at CLEF 2012

Claudia Peersman, Frederik Vaassen, Vincent Van Asch, Walter Daelemans

Computational Linguistics and Psycholinguistics Research Center, Antwerp University
{claudia.peersman, frederik.vaassen, vincent.vanasch, walter.daelemans}@ua.ac.be

Abstract. In this paper we present a new approach for detecting online pedophiles in chat rooms that combines the results of predictions on the level of the individual post, the level of the user and the level of the entire conversation, and describe the results of this three-stage system in the PAN 2012 competition. Also, we describe a resampling and a filtering strategy to circumvent issues regarding the unbalanced dataset. Finally, we describe the creation of a dictionary of words and expressions relating to predators' grooming stages, which we used to identify which posts in the predators' conversations were most distinctive for their grooming behavior.

1 Introduction

Between 2009 and 2011, the EU Kids Online project[1] organized a survey of nationally representative samples of children between the ages of 9 and 16 regarding their Internet use, in 25 different EU member states. This study not only showed that going online is very much embedded in children's lives — 9 to 16 year olds spend 88 minutes a day online on average, with 49% of these adolescents going online in their bedroom —, it also stated that 34% of children had added people they had never met face-to-face to their friends list, that 15% had sent strangers personal information, and that 14% had sent a picture or a video of themselves to a stranger. Moreover, the study showed that younger children usually do not possess the digital skills needed to manage their privacy settings in their user profiles or to block unwanted messages [5]. Unfortunately, it is impossible for social network moderators or law enforcement agencies to manually check the

[1] http://www2.lse.ac.uk/media@lse/research/EUKidsOnline/Home.aspx (last accessed on August 9th, 2012)

vast amount of communications online in order to tackle the risk of children being groomed by online sexual predators. We therefore turn to new, automated techniques for identifying Internet pedophiles to help create a safer Internet for children.

In this context, the author identification competition of the PAN 2012 lab created a sub-task in which the primary aim was to automatically identify Internet predators among other chatters. This paper describes a three-stage approach to tackling this problem. The approach plays on the strengths of two different text classifiers and models conversation-level constraints to improve the quality of the predictions.

One of the major challenges of this classification task lies in the class imbalance inherent to the problem: as in real life, the number of non-predator users in the dataset vastly outnumbers the number of predators. Section 2 briefly describes the data and the pre-processing steps we performed.

In Section 3, we explain how we used resampling and filtering techniques to circumvent this class imbalance. We will then describe in more detail how the three-stage approach combines the strengths of a high-recall post-level classifier with those of a high-precision user-based classifier, and how the combined predictions of these two classifiers are further improved by imposing conversation-level constraints, which boosts precision significantly.

The Sexual Predator Identification sub-task also had a secondary aim: to identify which posts in the predators' conversations were most distinctive for the predators' grooming behavior. Because there were no guidelines to which kind of posts were to be considered as grooming, we based our approach on Lanning's [4] analysis of the different stages of the grooming process, which include collecting information about the victim, lowering the victim's inhibitions, isolating the victim from adult supervision, initiating the abuse and (possibly) attempting to meet with the victim. In Section 4, we describe the creation of a dictionary of words and expressions that refer to these stages of grooming.

In Section 5, finally, we present the results of our system in the PAN 2012 competition and discuss some issues regarding the competition's evaluation measures.

2 Preprocessing of the Data

The PAN 2012 sexual predator identification training dataset consisted of 66,914 conversations involving 97,671 unique users of which 142 were labeled as a predator (0.15%). There were 2,015 conversations in which a predator was involved and these conversations constituted 4.52% of the total number of 900,631 posts. No conversation contained more than one predator. The majority of the conversations contained two users (68%), followed by single-user conversations (19%). The maximum number of users per conversation was 30. Most users (95%) were involved in only one conversation, while one user was involved in as many as 3,868 conversations. The most prolific predator produced posts in 182 conversations, while 20% of the predators were only represented by one conversation.

Our system was developed using two different splits of the PAN training data with each of the splits containing a training and a validation set. During the splitting, the conversations were clustered so that no user was present in two different clusters. Distributing clusters rather than conversations ensured that no user in training also appeared in validation, which prevented overfitting of user-specific features. In addition, no user in the validation set for split 1 is included in the validation set of Split 2. For example, for split 1, 13.2% of the conversations ended up in the validation set. There were 29 predators in this set, which constitutes 0.2% of the users. The complementary data ended up in the training set. The statistics for the two splits are given in Table 1.

Table 1. Statistics on the clustered splits of the PAN training set.

Partitions	**Training**		**Validation**	
	Conversations	*Predators*	*Conversations*	*Predators*
Split 1	86.83%	113 (0.14%)	13.17%	29 (0.18%)
Split 2	86.84%	110 (0.13%)	13.16%	32 (0.20%)

3 Task 1: Sexual Predator Identification

Due to both the criminal character of this topic and the privacy issues that are involved, so far there is only one dataset publicly available which displays chat conversations by sexual predators: the Perverted Justice[2] website (PJ) contains over 500 English chat conversations collected by adult volunteers pretending to be adolescents and as such were approached by an alleged pedophile. However, for machine learning algorithms to be effective in identifying online sexual predators, they would need to be trained with both illegal conversations between predators and their victims and sexually oriented conversations between consenting adults [8]. However, since such data are rarely made public, Pendar [8] only experimented with the PJ dataset. His kNN classification experiments based on word token n-grams achieved up to 93.4% F-score when identifying the predators from the pseudo-victims. To tackle the issue of limited dataset availability, Bogdanova et al. [1-2] as well as Rachid et al. [9] and Peersman et al. [7] investigated some other setups. Bogdanova et al. [1-2] experimented with new feature types such as emotional markers, emoticons and imperative sentences and computed sex-related lexical chains spanning over the PJ conversations. Using a corpus of cybersex chat logs[3] and the Naval Postgraduate School (NPS) chat corpus[4] as negative information for the predator class in the PJ dataset, their Naïve Bayes classifier yielded an accuracy of 92% for *PJ predators vs. NPS* and 94% for *PJ predators vs. cybersex* based on their high-level features. However, because these high-level features were partly derived from the PJ dataset itself, the experiment may have overestimated accuracy by detecting predators from the same PJ dataset. When dealing with the limited availability of data, both Rachid et al. [9] and Peersman et al. [7] proposed a system that in a first step classifies each user according to age group and gender, enabling the detection of predators using a fake adolescent user profile and distinguishing conversations between an adolescent and an adult from conversations between adults or adolescents only.

As the number of non-predator users in the PAN 2012 sexual predator dataset was far greater than the number of predator users and

[2] http://www.perverted-justice.com (last accessed on August 11th, 2012)
[3] www.oocities.org/urgrl21f/ (last accessed on August 11th, 2012)
[4] http://faculty.nps.edu/cmartell/NPSChat.htm (last accessed on August 11th, 2012)

no meta-information about the users' age and/or gender was available, in this paper we present a new approach that combines the results of predictions on the level of the individual post, the level of the user and of the entire conversation. Moreover, to circumvent problems resulting from the data imbalance in favor of the non-predator class, we describe both a resampling and a filtering strategy. More specifically, we trained a post-level classifier based on a balanced subset (described in Section 3.1), and a classifier on the user level based on a filtered subset of the training data (Section 3.2). We then combined the output of these two systems and imposed conversation-level constraints that significantly improved the quality of the output (Section 3.3). Figure 1 shows a simplified schematic representation of our three-stage approach to sexual predator identification.

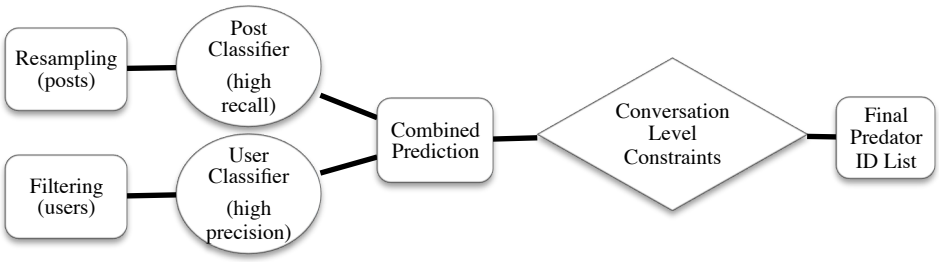

Figure 1. Overview of the system architecture

In all our experiments, we used LIBSVM [3] with the *c*-parameter set to 32.0, *g* to 0.0078125. We also set *b* to 1 in order to get the probability outputs for each class. These parameters were experimentally determined to provide good generalization during parameter optimization.

3.1 Classification on the Post Level

Our first classifier was designed to make a distinction between posts by predators and posts by other users. We first resampled the number of users to create an equal distribution of predator and non-predator posts in the training sets of both our splits. Because it is very unlikely that predators can be identified based on no more than a few lines in a conversation, in our first resampling process we only included data from users of whom at least ten posts were present in the training set.

This resulted in 40,722 posts for the predator class. We then attempted to match this number by randomly selecting data from other users which led to a set of 44,237 other users' posts, thus ensuring a balanced distribution of post instances over both classes. We tokenized each post and represented each token (words, punctuation and emoticons) using binary token unigram vectors. We chose this type of features, because during preliminary experiments these features performed better than more complex features types such as the approach words, communicative desensitization words and relationship words that were used by Bogdanova et al. [1-2].

Applying the clustered experimental setup we described in Section 2, using token unigram features only, the SVM classifier yielded a 64.6% accuracy and a 0.14 precision, a 0.77 recall and 0.23 F-score for the predator class for Split 1. For Split 2, the system achieved 76.9% accuracy, 0.23 precision, 0.67 recall and 0.34 F-score for the predator class. This means that our post classifier was able to identify 11,153 predator posts correctly out of 15,702 posts in both our validation sets combined, but also rendered 53,563 false positive predator posts.

After predicting which individual posts were possibly written by a predator, we aggregated the post-level predictions in order to identify which *users* in the data set were most likely to be predators. For this we used LiBSVM's probability outputs ($b = 1$) and took the average of the 10 most "suspicious" posts for each user —the 10 posts with the highest probability output for the predator class— as the final predator probability for each user. Then, to determine a good threshold for labeling a user as a predator based on this value, we performed a grid search on the validation set of the first split and then evaluated the threshold on the validation set of the second split. The best results were achieved when applying a threshold of 0.85, in which case the classifier identified 56 out of a total of 60 predators correctly (28 in each validation set). However, the classifier also returned 93 false positives (55 in the first and 38 in the second set). Table 2 (see Section 3.3) shows the precision, recall and F-scores for the predator class for both probability aggregation strategies.

As is clear from our results, our post classifier showed a high recall but a low precision score. Therefore, we decided to create a second system on the level of the user instead of the post level in order to produce a complementary system with a higher precision. In the next section we go into our user-based classification experiments and

describe the filter we used to resample the initial dataset.

3.2 Classification on the User Level

Because the PAN 2012 dataset also contains chat room conversations from a variety of domains other than the predator conversations (e.g. gaming, programming), we decided to compile a filter that would exclude all users that did not produce any "suspicious" posts. We therefore built a filter based on a list of words and expressions that could be linked to the typical stages in a predator's grooming process (e.g. [4], [6]), which are discussed in Section 4. It was our hypothesis that by only including data from users who had produced at least one post that was caught by our filter, we would be able to create a classifier that focuses on exactly those elements that distinguish pedophile chat from other chat that contained similar vocabulary, such as sexually oriented chat conversations between consenting adults and/or people making arrangements to meet up. This resulted in a resampled training set of 111 predators (98.2%) and 39,453 others (48.4%) in Split 1 and 107 predators (97.3%) and 39,500 others (46.7%) in Split 2. In our validation sets, using this filter, 24.6% (i.e. 7,891) of the users were automatically labeled as non-predator. Because our error analysis showed that only six predators —five in the training sets and one in the validation sets— who did not produce more than two posts in the entire PAN 2012 training dataset, were lost because of this filter, we decided not to adapt our filter.

Next, all posts from the same user were gathered into a single instance vector. This way, our second system would directly classify users as either a predator or a "non-predator" user, and no further aggregation steps were necessary. In the two splits, the users that were excluded by the filter were automatically labeled as non-predator in both the training and validation sets. As we expected, the "harder" classification experiment —because the data points in the filtered dataset lie closer together in the vector space— yielded a higher precision score than our aggregated post classifier, with only 7 false positives. The recall score, however, was much lower than for the aggregated post classifier, with 49 out of 60 predators identified correctly in the validation sets. The results of the user classification experiments are displayed in Table 2 (see Section 3.3).

Because the user classifier had a higher precision score than the post classifier, we decided to investigate which combination of the outputs achieved the best F-score for the predator class. It is to the results of these experiments we turn next.

3.3 Combining the Results with Post-processing on the Conversation Level

Starting from a post-based classifier that produced high recall and low precision scores for the predator class and a user-based classifier that achieved high precision and low recall scores, we decided to experiment with different ways of combining these outputs to create a system that played on the strengths of our high-precision user classifier, while still retaining some of the recall of the post-level classifier. One way to combine the outputs of several classifiers is to use ensemble methods, but these methods require an intricate experimental procedure (embedded cross-validation) to avoid overfitting on the test data. As we distributed clusters rather than conversations in both our partitions, we could not split these partitions into more sub-splits without users of the training sub-splits also appearing in the validation sub-splits and thus risking overfitting of user-specific features (see Section 2). Therefore, we decided to experiment with different ways of weighting the classifier outputs. Concretely, to determine the combined probability of a user being a predator, we used the following formula:

$$P(pred) = w_u * P_u(pred) + w_p * P_p(pred)$$

where w_u and w_p are the weights on the probability of the user being a predator according to the user classifier and the post classifier respectively.

After performing a grid search on the validation set of Split 1 and testing the weighted combinations on the Split 2 validation set, the best results were achieved when the user classifier was assigned a high weight ($w_u = 0.73$), while the post classifier was assigned a complementary weight of 0.27 (w_p). The resulting system managed to find 2 more predators than the user classifier alone (51 vs. 49), while still being relatively precise (10 false positives).

When performing an error analysis on the predicted predators, we discovered that in some cases both users in a conversation were labeled

as a predator. We suspect that the (pseudo-)victims, when replying to a predator during a conversation, mirrored some of the predator's vocabulary, which would explain why they were incorrectly labeled. To resolve this issue, we used the predator probabilities of the user classifier to determine which of both users in the conversation was in fact the predator. We again performed a grid search on the validation set of the first split to determine a good threshold so that as many false positives as possible could be excluded, without losing any of the true predators. By applying a threshold of 0.75, we were able to further reduce the number of false positives from 10 to 3: from 6 to 2 in the validation set of Split 1 and from 4 to 1 in the validation set of Split 2. Table 2 provides an overview of the best results of the single classifiers and their best performing weighted combinations.

Table 2. Overview of the combined results on the validation sets of the single and combined classifiers for the predator class.

Results	True Positives	False Negatives	False Positives	Predator Precision	Predator Recall	Predator F-score
Post Classifier	**56**	4	93	0.38	**0.93**	0.54
User Classifier	49	11	**7**	0.88	0.82	0.84
Post + User	51	9	10	0.84	0.85	0.84
After Post-processing	51	**9**	3	**0.94**	0.85	**0.90**

Based on the results of these experiments we finally retrained our models on the entire PAN 2012 training set, and performed the final classification experiment on the until then unknown PAN 2012 test set. This resulted in a list of 188 identified predators. We then applied the same post-processing strategy, which led to our final list of 170 predators. Of these predator IDs 152 were found to be correct, resulting in a 0.89 precision, a 0.60 recall and a 0.72 F-score for the predator class.

4 Task 2: Identifying the Grooming Posts

The second part of the Sexual Predator Identification sub-task consisted of detecting the specific posts in the predators' chat that were most distinctive for grooming behavior. Lanning [4] already argued that pedophiles groom their victims following five predictable stages: identifying a possible victim, collecting information about the intended victim, filling a need, lowering inhibitions and initiating the abuse. With regard to automatically detecting online grooming, McGhee et al. [6] were the first to incorporate the stage division into their research. Based on an expanded dictionary of terms they applied a rule-based approach, which categorized a post as belonging to the stage of *gaining personal information*, *grooming* (which included lowering inhibitions or *reframing* and sexual references*)*, *approach* or *none*. Their rule-based approach outperformed the machine learning algorithms they tested and reached up to 75.1% accuracy in determining whether a PJ post was predatory or not.

Because there were no gold standard labels available for this task that distinguished grooming from non-grooming posts in the predators' chat conversations, we decided to adopt a similar approach by creating a dictionary-based filter that only selects those posts that were linked to one of the following stages we adopted from Lanning [4] and McGhee et al. [6] in the predators' grooming processes: (1) the *sexual topic*, which includes discussing erogenous parts of the body, mentioning and performing sexual acts and sexually oriented adjectives and multi-word expressions; (2) *reframing*, which was already defined by McGhee et al. [6:8] as "the redefinition of sexual behaviors into non-sexual terms, such as connecting sexual acts to messing around, practicing or teaching" and (3) *approach*, which contains words and expressions that refer to meeting in person.

After manually analyzing part of the predator conversations in the PAN training set, we decided to add three extra stages that seemed typical of online grooming: (4) *requests for data*, i.e. pictures, videos or using the webcam; (5) *isolation from adult supervision* (e.g. "home alone?") and (6) *age*, which includes references to old(er) vs. young(er) and child-related vocabulary (e.g. "tummy", "tiny"). Although McGhee et al. [6] also mentioned the stage of building up a trusting or friendly relationship with the victim, we did not find this especially distinctive of pedophile grooming behavior and did not include this into our filter.

Next, we started compiling our grooming filter by adding the words from the dictionary by McGhee et al. [6] if they fit into one of our predefined stage categories. We then heavily expanded each category by manually selecting synonyms and related terms on the English Urban Dictionary website[5] and the English Synonyms website[6]. The complete dictionary can be obtained for research purposes by contacting the authors of this paper.

As we mentioned in Section 3.2, our filter alone could reduce the number of users by at least 24.6%, but it was definitely not enough to identify the highly limited number of predators. Therefore, we only applied this filter to perform a pre-selection for our user classifier and to select the grooming posts of the users that were already identified as a predator by our best scoring three-stage classification system (see Section 3.3). Using this strategy, our filter labeled 4,717 posts in the PAN 2012 test set as belonging to one of our six main grooming stages. Of these posts, 1,688 were found to be correct by the organizers, resulting in a 35.8 precision, a 26.1 recall and a 30.2 F-score.

5 Discussion

In this paper we proposed a new approach to detect Internet predators grooming their victims in chat rooms. Our experiments showed that a weighted combination of a high-recall post classifier and a high-precision user classifier achieved better results for the predator class than each system separately. Moreover, imposing conversation-level constraints boosted precision significantly, resulting in a final F-score of 0.90 for the predator class in cross-validation on the PAN 2012 training set. Interesting to see, was that the use of a dictionary-based filter could not only reduce the number of possibly suspicious users by 24.6 to 53.3%; it also enabled us to create a user classifier that focuses on exactly those elements that distinguish pedophile chat from other chat that contained similar vocabulary. This filter also proved to be very effective when detecting which of the posts in the identified predators' chat conversations were most distinctive for grooming behavior.

[5] http://www.urbandictionary.com/ (last accessed on June 22nd, 2012).
[6] http://www.synonyms.net/ (last accessed on June 22nd, 2012).

To calculate the final F-score the competition's organizers decided to set the β-factor to 0.5 to emphasize precision when detecting predators online to optimize the time needed for a police agent to check the output, while they set the β-factor to 3.0 to emphasize recall when detecting predators' grooming posts to collect as much evidence as possible. In our opinion, it would be more important to heavily reduce the number of possibly suspicious users that are to be checked manually, while still retaining all of the actual predators, in which case our post classifier produced the best results with a recall score of 0.93 and reducing the number of possibly suspicious users from over 32,000 to 149 in our validation sets while only losing 4 predators. Likewise, when manually checking a suspicious user's communications, a police officer or a moderator will need a swift access to the most striking posts to be able to quickly discard remaining false positives, which means precision is of high importance here. Therefore, in future research we will work on a system that combines a high-recall classifier with a grooming scoring system that will rank the remaining suspicious users according to the presence of the grooming stages in their conversations. This will also enable both law enforcement agencies and moderators to take action more quickly regarding imminent meetings and abuse, when these stages are attributed higher weights in the scoring system.

5 References

[1] Bogdanova, D., Rosso, P. and Solorio, T. Modelling Fixated Discourse in Chats with Cyberpedophiles. In: *Proceedings of the EACL 2012 Workshop on Computational Approaches to Deception Detection*. Avignon, France. p.86-90 (2012a)

[2] Bogdanova, D., Rosso, P. and Solorio, T. On the Impact of Sentiment and Emotion Based Features in Detecting Online Sexual Predators. In: *Proceedings of the 3rd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis*. Jeju, Republic of Korea. p.110-118 (2012b)

[3] Chang, C. and Lin, C. LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1--27:27 (2011)

[4] Lanning, K. Child Molesters: A Behavioral Analysis. For Professionals Investigating the Sexual Exploitation of Children.

National Center for Missing & Exploited Children, USA. http://www.missingkids.com/en_US/publications/NC70.pdf (2010)

[5] Livingstone, S., Haddon, L., Görzig, A., and Ólafsson, K. EU Kids Online final report (2011) http://www2.lse.ac.uk/media@lse/research/EUKidsOnline/EU%20Kids%20II%20(2009-11)/EUKidsOnlineIIReports/Final%20report.pdf

[6] McGhee, I., Bayzick, J., Kontostathis, A., Edwards, L., McBride, A. and Emma Jakubowski Learning to Identify Internet Sexual Predation. *International Journal of Electronic Commerce*, 15:3, pp. 103 (2011)

[7] Peersman, C., Daelemans, W. and Van Vaerenbergh, L. Predicting age and gender in online social networks. In: *Proceedings of the 3rd Workshop on Search and Mining User-Generated Contents*. Glasgow, UK. http://www.cpl.ua.ac.be/sites/default/files/smuc1504-peersman.pdf (2011)

[8] Pendar, N. Toward Spotting the Pedophile: Telling victim from predator in text chats. In: *The Proceedings of the First IEEE International Conference on Semantic Computing*. California, USA. p.235-241 (2007)

[9] Rashid, A., Rayson, P., Greenwood, P., Walkerdine, J., Duquenoy, P., Watson, P., Brennan, M. and Jones, M. Isis: Protecting Children in Online Social Networks. At *The International Conference on Advances in the Analysis of Online Paedophile Activity*. Paris, France. http://eprints.mdx.ac.uk/4738/1/Isis_overview_v3.pdf (2009)