Overview of the International Sexual Predator Identification Competition at PAN-2012

Giacomo Inches and Fabio Crestani

Faculty of Informatics
University of Lugano (USI), Switzerland
{giacomo.inches, fabio.crestani}@usi.ch

Abstract This contribution presents the evaluation methodology for the identification of potential "sexual predators" in online conversations as part of PAN 2012. We provide details of the realized collection and analyse the submissions of the participants, who had to solve two problems: identify the predators among all the users in the different conversations and identify the part (the lines) of the predator conversations which are the most distinctive of the predator bad behaviour. The methods proposed by the 16 teams participating in the contest made possible the recognition of common pattern for predator identification (e.g. no preprocessing of the conversations, lexical and behavioral analysis, blacklisting of predator terms) as well as possible extension to existing systems (e.g. victim-predator distinction, pre-filtering of not relevant conversations).

1 Introduction and Motivations

"Chat messages" or "online/IRC conversations" are part of almost everybody's everyday life with services like Skype, Yahoo Messanger, MSN Messanger, ICQ but also IRC networks like Freenode or Quakenet. Although these services facilitate the establishment of new connections between persons or reinforce existing ones, they also allow for misbehaviours or cybercriminal acts. The *Sexual Predator Identification* competition ran for the first time in 2012 within PAN[1] and aimed at providing researches with a common framework to test methods for identifying such misbehaviours or cybercryminal activities. For simplicity, in the competition we only concentrate on the identification of "sexual predator" inside a chat, not dealing with other kind of misbehaviour or media. A "sexual predator" is defined in the New Oxford American Dictionary as *"a person or group that ruthlessly exploits others"* while Wikipedia noticed how the definition *"is used pejoratively to describe a person seen as obtaining or trying to obtain sexual contact with another person in a metaphorically "predatory" manner"*. We refer to these interpretations of the term "sexual predator' for the competition.

In defining the tasks for the competition, we were also inspired by some previous works [12,9,16] that addressed similar problem, even if none of them aimed at being an evaluation laboratory or containing a challenging collection to be used as a reference. In fact, we were the firsts to propose the following two kind of problems: given a collection containing chat logs involving two (or more) persons the participants had to:

[1] A benchmarking activity on uncovering plagiarism, authorship and social software misuse http://pan.webis.de

1. identify the predators among all users in the different conversations (problem 1)
2. identify the part (the lines) of the conversations which are the most distinctive of the predator behaviour (problem 2).

We are presenting in Section 2 the details of the collection used in the competition and in Section 3 the analysis of the methods employed by the participants to the task. Finally, in Section 4 we are presenting the results of the competition, concluding with Section 5.

2 Description of the Evaluation Framework

In this section we present the corpus realized specifically for this competition and highlight its properties and novelty with respect to existing collections. We also describe the measures of performance used for evaluating the submission of the participants the two problems of the task.

2.1 Corpus

In creating our collection we were animated by the same spirit of TREC and, more recently, CLEF: we wanted to build a *large* collection that could serve as *common reference point* for researchers of different fields (from Information Retrieval to Natural Language Processing, from Text Mining to Machine Learning) and where they could *compare the performances of their different approaches*. The realistic (large) size of the collection is very important and is one of the central aspect of TREC tracks [20] and PAN laboratories [3]. It serves to fill the gap between the research and the industrial application of the technologies developed in lab. For this reason we created a large collection (hundred of thousands of conversations) with realistic properties: few number of true positives (conversations with a potential "sexual predator"), large number of false positives (people talking about sex or shared topic with the "sexual predator") and large number of false negatives (general conversations between users on different topics). We believe that in a realistic scenario the percentage of "predator" conversations with respect to the "regular" ones should be very low. In a different field (paedophile queries in peer-to-peer system) the number of "predator" queries was found to be 0.25% of the total [11]. In our collection we therefore tried to respect that number but, in order to make the identification of the predator a doable investigation, we increased the percentage of one order of magnitude and set this to less than 4%.

When looking for the "predator" collections, we found a common source for the different datasets that were already used in the literature [12,9,16]: the `http://www.perverted-justice.com/` (PJ) website. This is a website where logs of online conversations between convicted sexual predators and volunteers posing as underage teenagers are published. The controversial creation and preliminary usage of these data has been already discussed in [9], where the authors also give a detailed overview of other collections [16,21], tools and approaches to cybercrime and online deception detection. We therefore started with the PJ data for building our collection and kept in mind the observations present in [16], where two kinds (and different subkinds) of suspicious interactions were identified: I) Predator/Other interaction, subdivided into: (Ia)

Predator/Victim (victim is underage); (Ib) Predator/Pseudo-Victim (volunteer posing as child); (Ic) Predator/Pseudo-Victim (law enforcement officer posing as child) and II) Adult/Adult (consensual relationship). Data of type (Ia) and (Ic) are difficult to obtain, since it involves the police or law enforcement agencies in the process of data acquisition. To our personal experience, police and law enforcement agencies are reluctant and not very enthusiastic in collaborating on this sensible topic, therefore we ignored this approach to data acquisition and focused on (Ib), which corresponds to the PJ data. PJ data constitutes therefore our true positive set.

Regarding interaction of type II) we found at first several online sources[2] that could had been of help but we later discarded them, because they were based on a single person experience or were not of sufficiently large size (only some hundreds conversations) to be successful employed in our collection. The documents present in the Omegle repository[3], to the contrary, served exactly our purpose. The original service Omegle (where the documents come from) is a website that allows two strangers, connected at the same time to the website, to have an anonymous online conversation. The repository presents a random sample of more than 1 million original Omegle conversations and by admission of the provider contains *"abusive language and general silliness online"* and sometimes users *"engage in cybersex"*[4]. The quantity of conversations as well their nature and characteristics made this repository perfect to augment the level of false positives in our collection, thus to make it more challenging and somehow real.

The major difficulty that we encountered was in crawling "regular" online conversations to complete the false negative set of documents and add a variety of topic of discussion, to possibly hide an eventual general topicality of our true positive conversations. We already mitigated the fact that the "predator" conversations are between just two users by introducing the conversations extracted from Omegle, so now we just needed to focus on topics about general discussions. To our surprise, the Internet lacks of this kind of conversations: few people share their (private) conversations online and the massive crawling of the public channels of the major IRC networks[5] is neither trivial nor encouraged[6]. We decided then to rely on those IRC logs that included thousand of conversations and that were already made available on the website of the IRC channel managers, namely `http://www.irclog.org/` and `http://krijnhoetmer.nl/irc-logs/`. Having a large volume of conversations allowed us to increase the probability of having general discussions, interactions between just few users and a variety of messages in length and duration, despite the topical similarity between these conversations.

A few other issues had to be solved in merging so different collections together. A first problem occurred when deciding about the semantic definition of *conversation*. In fact, we downloaded files from different sources of different formats, containing from continuous logs of conversations on a daily basis to transcripts of unique conversations

[2] See for example: `http://www.oocities.org/urgrl21f/`,`http://www.fugly.com/victims/` or `http://chatdump.com/`

[3] `http://omegle.inportb.com/`

[4] See: `http://inportb.com/2010/02/21/the-omeglean-society/`

[5] See: `http://irc.netsplit.de/`

[6] See: `http://wiki.vorratsdatenspeicherung.de/IRSeeK-en`

of few lines, and we needed to combine them together in a single collection. To make the conversations contained in the different files comparable, we decided to segment all the messages exchange between the users in threads, where the cut was a break in the messages exchange of more than 25 minutes. We empirically observed that this was a reasonable threshold for a topic change in the conversation or the starting of a total new one. After this step we obtained a consistent collection of hundred of thousand conversations. We then noticed, by studying the length of the conversations, that the wast majority (from 77% to 99% depending on the source) were below 150 messages exchange. We therefore decided to include in the collection all the conversations that were less or equal to 150 message exchange. Finally, we decided to generate an arbitrary unique id for each conversation and also for each user and to replace nicknames within each message with the corresponding user ids. Where possible we also substituted real email addresses with arbitrary tags, in order to avoid the identification of real users.

To the purpose of the competition we divided the collection into two parts, a training one and a testing one. Given the fact that the training part is intended as "practicing" rather than "training" as in Machine Learning, we decided to release 30% of the collection as training set. In Table 1 we report the main properties of the whole collection.

Table 1. Properties of the collection

	PJ perverted-justice.com	krjin krijnhoetmer.nl/irc-logs	irclog irclog.org	omegle omegle.inportb.com
#conversations	11350	50510	28501	267261
#conv. length $\leq$150	9076	48569	21896	265747
(% all)	(80%)	(96%)	(77%)	(99%)
Training set				
#conv. length$\leq$150	2723	14571	6569	43064
" and exactly 2 user	984	2420	1146	41067
(% training)	(36%)	(17%)	(17%)	(95%)
unique (perverted) users	291 (142)	2660	10613	84131
Testing set				
#conv. length$\leq$150	5321	33998	15327	100482
" and exactly 2 user	1887	5648	2673	95648
(% testing)	(35%)	(17%)	(17%)	(95%)
unique (perverted) users	440 (254)	4358	17788	196130

2.2 Performance Measures

For the evaluation of the performance of the participants of the two problems, we referred to the standard Information Retrieval measure of Precision (P), Recall (R) and F (weighted harmonic mean between Precision and Recall):

$$\text{Precision (P)} = \frac{\#(\text{relevant items retrieved})}{\#(\text{retrieved items})} \tag{1}$$

$$\text{Recall (R)} = \frac{\#(\text{relevant items retrieved})}{\#(\text{relevant items})} \tag{2}$$

$$\text{F} = \frac{1}{\alpha\frac{1}{\text{P}} + (1-\alpha)\frac{1}{\text{R}}} = \frac{(\beta^2+1)\text{PR}}{\beta^2\text{P} + \text{R}} \text{ where } \beta^2 = \frac{1-\alpha}{\alpha} \tag{3}$$

The "items" retrieved are in one case (problem 1, identify the predators) the ids of the authors considered perverted and in the second one (problem 2, identify the predators' lines) the line numbers considered indicative of a bad behaviour within a conversation. We also noticed that, while the standard F measure equally weighted P and R with β equal to 1, this is not always desired. In our case, in fact, for the first problem, despite we observed that retrieving lot of relevant authors is important (Recall), to facilitate the work of a police agent who would like to receive the largest number of suspect, what is more important is the fact that the retrieved authors are relevant (Precision). This to optimize the time of the police agent towards the "right" suspect rather than "all" the possible suspects. For this reason we used a measure of F with the β factor equal to 0.5, in order to emphasize Precision [2]. For the second problem, instead, we observed that retrieving lot of relevant lines (Recall) is more important than finding only the relevant ones (Precision). Having lot of relevant lines, in fact, augments the possibility of finding good evidences towards a suspect and for this reason we used a measure of F with the β factor equal to 3, for emphasizing Recall [2].

It is also to be said that, while for the first problem the evaluation was quite straightforward, having an a-priori indication of convicted perverted from the PJ website, for the second it was harder (and more discussed) to define the ground truth. We decided to adopt a TREC-like methodology for the evaluation and manually evaluated all submitted lines by at least one participants (this accounted for 91% of all the predators' lines). Given the particular nature of the task, that requires a particular training for the evaluator in order to be able to distinguish between a predator chat and a regular chat, this could not be done in a distributed way (e.g. mechanical turk). Moreover given the limited time for the evaluation, we could not train other experts than us, thus relying on the evaluation of a single expert in our group. For this reason, evaluations contain a certain grade of subjectivity that we could not avoid. This is certainly a weak point in this year competition that we will try to address better next year.

3 Overview of the Participants' Approaches

We received 16 submissions for the first problem (identifying the predators) and 14 for the second problem (identifying the distinctive chat lines of the predator behaviour) of the Sexual Predator Identification competition. Few users decide not to submit a notebook paper to explain their used methods, therefore we are presenting an analysis based on the 12 notebook paper received.

3.1 Problem 1: identify predators

Pre-filtering For the first problem, where the participants had to return a list of potential predator, different pre-filtering techniques as well as classification methods have

been applied. The collection given to the participants was by design very unbalanced (as most of them noticed) having few true positive authors (1% or less) in both training and testing dataset and containing lot of false negatives that needed to be filtered out. A common approach to overcome this problem was the use of a two stage classifier, where in the first stage the classifier had to distinguish between conversation involving a predator (true positive) and conversation without a predator (false negatives) [19,13,15,6]. In addition to this, one of the most successful approaches [19] decides for the pre-filtering of all the conversations that manifested some particular patterns: presence of 1 participants only, those with less then 6 interventions per user or those that contained 3 long sequences of unrecognised characters. Similar attempts were done by other participants but with a rule-based approach and on different features for different approaches [14].

Features Apart from one case [17], where participants used machine learning approaches that work at character level (kernel with character 5-gram presence bit), in all the others submissions we can divide the used features into two main categories: "lexical" features and "behavioural" features. Lexical features are those that can be derived from the raw text of the conversation: example of these features are unigram or bigram [19,13,4,14], their weighting using TF-IDF or the cosine similarity and emoticons counting. Other examples are the name recognition of the participants in the conversation (self, other, group) [4] but also features obtained by the LIWC tool[7] that calculates the degree to which people use different categories of words across a wide array of texts [14,18]. It is to be noted that, in general, lexical features have been used without any stemming or stopword removal, to preserve each author own style, including misspelling and grammatical errors.

Behavioural are all those features that captures the "actions" of a user within a conversation [6,18]: the number of times a user starts a dialogue, the response time after a message of the partner in the conversation, the number of questions asked, the frequency of turn-taking, intention (grooming, hooking, ...), etc. One of the most common approach was the creation of a single set of features for each author, to be able to profile him and exploit his predator potential. Some participants decided to build up not just the Language Model (LM) of a single author, but also a LM as a combination of the LMs of the two participants in the chat [4]. Some other approaches were working, instead, at a conversation or at a line level, therefore participants that used this strategies had to aggregate the partial scores relative to all the lines or conversations of an author to obtain a unique set of features for each author [1,10,6,15,8].

Classification approaches In the classification step we could observe different proposed method, but Support Vector Machines (SVM) were the most used [13,14,15,19]. In general, they were used in most cases for the first (predator-vs-all), then also for the second step of the classification (predator-vs-victim). Sometime participants found out that other solutions worked better than SVM, for example when they used a Neural Network classifier [19]. Other classifier applied were based on Maximum-Entropy [4,8], decision trees[10], k-NN [7,17] and/or random forest [17] as well as Naïve Bayes [6,1]. In combination with the classifier sometimes we observed a filtering approach based on a self-compiled dictionary of predatory terms.

[7] `http://www.liwc.net/`

To conclude, we should noticed that for this first problem we release a training set, that allowed for supervised algorithms to be easily used. The situation was different for the second problem, were no training data was available.

3.2 Problem 2: identify predators' lines

For this second problem, no training data was available for the participants. This was intentionally done, mostly to test how participants approached the problem without a-priori relevance.

The difficulty of the problem reduced the number of submissions (from 16 to 14) and obliged the participants to use different approaches, compared with the supervised ones of problem 1. The straightforward solution was to return as relevant all the conversations lines of all the identified predators from the first problem [17]. One of the most used method was a filtering of all the predator conversations through a dictionary of "perverted" terms or with a particular score (e.g. TF-IDF weighting) [15,13,14,4]. Similar to this approach, another first computed the LMs of the part of the conversation considered predatory and then computed the differences between the actual conversation and the LMs [19]. To conclude, the last approach was simply to return those lines already labelled as predatory in the proposed algorithm by the default method for problem 1 (working at line level) [10,6,8].

4 Evaluation Results of the Participants' Approaches and Discussion

As reported in Table 1, participants received a training and a testing set, the first containing 142 users labelled as predator, the second containing 254 predators to be discovered. This was useful for the first problem (identify the predators), while for the second problem (identify the lines manifesting the predators bad behaviour) we did not release any training set. We wanted, in fact, to test how such a problem could be addressed without any evidence. We later evaluated manually all the 113888 lines submitted by the participants and identified 6478 that we considered expression of a predator bad behaviour. In Table 2 and Table 3 we present the results for the first and second problem, with the measures of evaluation explained in Section 2.2.

4.1 Problem 1: identify predators

If we analyse in details the results for the first problem, in particular the ranking in the case of the two different metrics F with $\beta = 1$ and F with $\beta = 0.5$, we can notice that only two positions swaps (1^{st} and 5^{th}) in case we consider one or the other measure of F. This is due to the fact that we emphasised Precision with the F with $\beta = 0.5$. This choice did not encountered the favour of all the participants, in fact some manifested their disagreement and suggested giving more weight to Recall (thus, having a F measure with $\beta \geq 2$). In a real scenario, the proposed idea is to let the police agent decide who is a predator and "manually" filter the results automatically obtained. Another

suggestion into this directions is the creation of a ranked list of suspects, that could serve to prioritize the investigations.

Besides this issues, from an operational point of view, it is interesting to notice how important was the pre-filtering of unrelated conversations (at the cost of few true positive) [19] and the similar use of lexical features in all the first ranked approaches: bag-of-words with boolean weighting scheme [13,19], unigrams with TF-IDF weighting scheme [14], unigram and bigram [4]. Participants also created a unique profile for each author, by computing the features on an author-based file that collects all the posts/messages of that author [4,14,13]. Behavioural/conversational features were, on the other hand, used by all [4,14,13] except one [19] of the top-5 participants. These last one [19] also choose to use a Neural Network classifier instead of SVM (in both cases, two step classifiers) that were instead used by two others [14,13], while others employed a Maximum-Entropy Classifier.

Despite the similar features used and the relatively closeness of the performance measures, the different classification strategies are a signal of still possible improvement possibilities in the problem.

4.2 Problem 2: identify predators' lines

As mentioned before, problem 2 was more difficult than problem 1 and presented more open-issues than problem 1 too. Despite the suggestion of giving more weight to Precision than to Recall, we should mention at least two issues that touched this part of the competition. The first one is a certain dependency from the first problem: identifying lines of the predator conversation requires at the beginning the correct identification of a good number of predators. This might disadvantage participants that performed poorly in the first part of the task. A solution to this problem might be having two stages for the competition that corresponds to the two problems. The best result set of the first problem could be used as a starting point for the second task. It has to be noticed, however, that in the best-performers list (first-half of the ranking) we find also participants that were not in the top-5 of the first problem. A preliminary explanation for this is that few conversations of relatively few predators contributes to generate the ground truth for the predators' lines, therefore it is enough to identify such predators to obtain a good score for problem 2. This fact leads to a second issue for problem 2, the creation of the ground truth for the predators' lines. At the beginning of the competition, there was no ground truth for this second problem and we generated it on the basis of the received submissions. We could have generated the ground truth by analysing all the predators' conversation but by labelling only the submitted lines we spared 10% of all the conversations and approximatively 1 week of work time. The real issue was determined by the fact that one expert only labelled the lines of the conversation, leading to exclusion of possibly relevant lines or the over-consideration of some others. We would have liked to have more experts (at least 2 or more) for labelling the relevant lines in all the predator conversations, but due to time and resource constraints that was not possible this year. For a future edition of the Sexual Predator Identification task, we should plan more time and resources for generating the ground truth and maybe we should consider the release of a training set for this part of the problem as well.

Table 2. Results for problem 1): identify predators. The table reports the evaluation of all the runs submitted ordered by value of F score with $\beta = 0.5$. Runs with ranking number are the ones used for official evaluation. RET. = Retrieved documents, REL. = Relevant document retrieved. P = Precision. R = Recall

Participant run	RETR.	REL.	P	R	$F_{\beta=1}$	$F_{\beta=0.5}$	Official run rank
villatorotello-run-2012-06-15-2157g	204	200	0.9804	0.7874	0.8734	0.9346	1
snider12-run-2012-06-16-0032	186	183	0.9839	0.7205	0.8318	0.9168	2
villatorotello-run-2012-06-15-2157c	*211*	*200*	*0.9479*	*0.7874*	*0.8602*	*0.9107*	
parapar12-run-2012-06-15-0959j	181	170	0.9392	0.6693	0.7816	0.8691	3
morris12-run-2012-06-16-0752-main	159	154	0.9686	0.6063	0.7458	0.8652	4
eriksson12-run-2012-06-15-1949	265	227	0.8566	0.8937	0.8748	0.8638	5
parapar12-run-2012-06-15-0959g	*171*	*162*	*0.9474*	*0.6378*	*0.7624*	*0.8635*	
morris12-run-2012-06-17-0126	*152*	*147*	*0.9671*	*0.5787*	*0.7241*	*0.8527*	
parapar12-run-2012-06-15-0959i	*173*	*161*	*0.9306*	*0.6339*	*0.7541*	*0.8510*	
parapar12-run-2012-06-15-0959e	*182*	*164*	*0.9011*	*0.6457*	*0.7523*	*0.8350*	
peersman12-run-2012-06-15-1559	170	152	0.8941	0.5984	0.7170	0.8137	6
parapar12-run-2012-06-15-0959d	*175*	*151*	*0.8629*	*0.5945*	*0.7040*	*0.7914*	
parapar12-run-2012-06-15-0959c	*169*	*145*	*0.8580*	*0.5709*	*0.6856*	*0.7796*	
villatorotello-run-2012-06-15-2157a	*108*	*103*	*0.9537*	*0.4055*	*0.5691*	*0.7507*	
parapar12-run-2012-06-15-0959b	*205*	*160*	*0.7805*	*0.6299*	*0.6972*	*0.7449*	
grozea12-run-2012-06-14-1706b	215	163	0.7581	0.6417	0.6951	0.7316	7
parapar12-run-2012-06-15-0959f	*202*	*154*	*0.7624*	*0.6063*	*0.6754*	*0.7250*	
sitarz12-run-2012-0615-1515	218	159	0.7294	0.6260	0.6737	0.7060	8
parapar12-run-2012-06-15-0959h	*223*	*161*	*0.7220*	*0.6339*	*0.6751*	*0.7024*	
parapar12-run-2012-06-15-0959a	*200*	*128*	*0.6400*	*0.5039*	*0.5639*	*0.6072*	
vartapetiance12-run-2012-06-15-1411	160	99	0.6188	0.3898	0.4783	0.5537	9
villatorotello-run-2012-06-15-2157f	*269*	*143*	*0.5316*	*0.5630*	*0.5468*	*0.5376*	
grozea12-run-2012-06-14-1706a	*322*	*142*	*0.4410*	*0.5591*	*0.4931*	*0.4604*	
kontostathis-run-2012-06-16-0317e	475	170	0.3579	0.6693	0.4664	0.3946	10
kontostathis-run-2012-06-16-0317d	*688*	*172*	*0.2500*	*0.6772*	*0.3652*	*0.2861*	
kang12-run-2012-06-15-0904b	930	203	0.2183	0.7992	0.3429	0.2554	11
kang12-run-2012-06-15-0904a	*1049*	*202*	*0.1926*	*0.7953*	*0.3101*	*0.2270*	
kern12-run-2012-06-18-1827b	1172	177	0.1510	0.6969	0.2482	0.1791	12
kern12-run-2012-06-18-1827a	*1172*	*177*	*0.1510*	*0.6969*	*0.2482*	*0.1791*	
villatorotello-run-2012-06-15-2157d	*240*	*36*	*0.1500*	*0.1417*	*0.1457*	*0.1483*	
kontostathis-run-2012-06-16-0317c	*3696*	*206*	*0.0557*	*0.8110*	*0.1043*	*0.0685*	
villatorotello-run-2012-06-15-2157b	*204*	*12*	*0.0588*	*0.0472*	*0.0524*	*0.0561*	
kontostathis-run-2012-06-16-0317a	*5225*	*206*	*0.0394*	*0.8110*	*0.0752*	*0.0487*	
kontostathis-run-2012-06-16-0317b	*5625*	*221*	*0.0393*	*0.8701*	*0.0752*	*0.0486*	
vilarino12-run-2012-06-14-2121a	*9071*	*236*	*0.0260*	*0.9291*	*0.0506*	*0.0323*	
bogdanova12-run-2012-06-14-1117	2109	55	0.0261	0.2165	0.0466	0.0316	13
prasath12-run-2012-06-15-2122	10289	207	0.0201	0.8150	0.0393	0.0250	14
vilarino12-run-2012-06-14-2121b	5225	98	0.0188	0.3858	0.0358	0.0232	15
villatorotello-run-2012-06-15-2157e	*305*	*6*	*0.0197*	*0.0236*	*0.0215*	*0.0204*	
gomezhidalgo12-2012-06-15-1900	150	1	0.0067	0.0039	0.0050	0.0059	16

Table 3. Results for problem 2): identify predators' lines. The table reports the evaluation of all the runs submitted ordered by value of F score with $\beta = 3$. RET. = Retrieved documents, REL. = Relevant document retrieved. P = Precision. R = Recall

Participant run	RETR.	REL.	P	R	$F_{\beta=1}$	$F_{\beta=3}$	Official run rank
grozea12-run-2012-06-14-1706b	63290	5790	0.0915	0.8938	0.1660	0.4762	1
kontostathis-run-2012-06-16-0317e	19535	3249	0.1663	0.5015	0.2498	0.4174	2
peersman12-run-2012-06-15-1559	4717	1688	0.3579	0.2606	0.3016	0.2679	3
sitarz12-run-2012-0615-1515	4558	1486	0.3260	0.2294	0.2693	0.2364	4
morris12-run-2012-06-16-0752-main	2685	1211	0.4510	0.1869	0.2643	0.1986	5
kern12-run-2012-06-18-1827b	15533	1357	0.0874	0.2095	0.1233	0.1838	6
eriksson12-run-2012-06-15-1949	10416	1122	0.1077	0.1732	0.1328	0.1633	7
prasath12-run-2012-06-15-2122	77255	1044	0.0135	0.1612	0.0249	0.0770	8
parapar12-run-2012-06-15-0959j	2037	105	0.0515	0.0162	0.0247	0.0174	9
vartapetiance12-run-2012-06-15-1411	607	91	0.1499	0.0140	0.0257	0.0154	10
vilarino12-run-2012-06-14-2121b	6787	48	0.0071	0.0074	0.0072	0.0074	11
bogdanova12-run-2012-06-14-1117	49	4	0.0816	0.0006	0.0012	0.0007	12
villatorotello-run-2012-06-15-2157g	50	1	0.0200	0.0002	0.0003	0.0002	13
gomezhidalgo12-2012-06-15-1900	400	0	0.0000	0.0000	0.0000	0.0000	14

5 Conclusions

We presented in this document the results of the first International Sexual Predator Identification Competition at PAN-2012 within CLEF 2012. Given a realistic and challenging collection containing chat logs involving two (or more) persons, the 16 participants to the competition had to identify the predators among all the users in the different conversations and identify the part (the lines) of the predator conversations which were the most distinctive of the predator bad behaviour.

For the first problem we can conclude that lexical and behavioural features should be used when dealing with this kind of task. However, there is no unique method to identify predators but different approaches could be used, from SVM to Maximum-Entropy algorithm. Having a pre-filtering step to prune irrelevant conversations seems an important addition to the systems. For the second problem the most effective methods appeared to be those based on filtering on a dictionary or LM basis, partly due to the lack of ground truth for this specific problem (if we exclude the one based on 5-gram characters presence bit). The identification of common set of features and a group of effective strategies to identify predators is an achievement for this first part of the task.

During the competition some issues were raised about the measurement of performances for the two problems, whether we should emphasise Precision or Recall and about the degree of subjectivity in the creation of the ground truth for problem 2. This is an achievement, too: with this competition we wanted to give researchers a unique place for comparing their methods but also for discussing and debating about future directions on this research area.

6 Acknowledgments

We want to thank all the PAN 2012 and CLEF 2012 organisers for their hard work and support, as well as the participants of the competition for their patience and suggestions.

Development of this competition was funded in part by the Swiss National Science Foundation (SNF) project "Mining Conversational Content for Topic Modelling and Author Identification (ChatMiner)" under grant number 200021_130208.

References

1. Ayala, D.V., Castillo, E., Pinto, D., Olmos, I., León, S.: Information retrieval and classification based approaches for the sexual predator identification - notebook for pan at clef 2012. In: Forner et al. [5]
2. Christopher D. Manning, P.R., Schütze, H.: Introduction to Information Retrieval. Cambridge University Press (2008)
3. Clough, P., Ferro, N., Forner, P., Gonzalo, J., Huurnink, B., Kekäälå inen, J., Lalmas, M., Petras, V., de Rijke, M.: CLEF 2011. ACM SIGIR Forum 45(2), 32 (Jan 2012)
4. Eriksson, G., Karlgren, J.: Features for modelling characteristics of conversations- notebook for pan at clef 2012. In: Forner et al. [5]
5. Forner, P., Karlgren, J., Womser-Hacker, C. (eds.): CLEF 2012 Evaluation Labs and Workshop - Working Notes Papers, 17-20 September 2012, Rome, Italy (2012)
6. Hidalgo, J.M.G., Díaz, A.A.C.: Combining predation heuristics and chat-like features in sexual predator identification - notebook for pan at clef 2012. In: Forner et al. [5]
7. Kang, I.S., Kim, C.K., Kang, S.J., Na, S.H.: Ir-based k-nearest neighbor approach for identifying abnormal chat users - notebook for pan at clef 2012. In: Forner et al. [5]
8. Kern, R., Klampfl, S., Zechner, M.: Vote/veto classification, ensemble clustering and sequence classification for author identification - notebook for pan at clef 2012. In: Forner et al. [5]
9. Kontostathis, A., Edwards, L., Leatherman, A.: Text mining and cybercrime. In: Text Mining, pp. 149–164. Wiley Online Library (2010)
10. Kontostathis, A., West, W., Garron, A., Reynolds, K., Edwards, L.: Identify predators using chatcoder 2.0 - notebook for pan at clef 2012. In: Forner et al. [5]
11. Latapy, M.: Quantifying Paedophile Queries in a Large P2P System. System pp. 401–405 (2011)
12. McGhee, I., Bayzick, J., Kontostathis, A., Edwards, L., McBride, A., Jakubowski, E.: Learning to Identify Internet Sexual Predation. International Journal of Electronic Commerce 15(3), 103–122 (Apr 2011)
13. Morris, C., Hirst, G.: Identifying sexual predators by svm classification with lexical and behavioral features - notebook for pan at clef 2012. In: Forner et al. [5]
14. Parapar, J., Losada, D.E., Barreiro, A.: A learning-based approach for the identification of sexual predators in chat logs - notebook for pan at clef 2012. In: Forner et al. [5]
15. Peersman, C., Vaassen, F., Asch, V.V., Daelemans, W.: Conversation level constraints on pedophile detection in chat rooms - notebook for pan at clef 2012. In: Forner et al. [5]
16. Pendar, N.: Toward Spotting the Pedophile Telling victim from predator in text chats. In: International Conference on Semantic Computing (ICSC 2007). pp. 235–241. No. c, IEEE (Sep 2007)
17. Popescu, M., Grozea, C.: Kernel methods and string kernels for authorship analysis - notebook for pan at clef 2012. In: Forner et al. [5]

18. Vartapetiance, A., Gillam, L.: Quite simple approaches for authorship attribution, intrinsic plagiarism detection and sexual predator identification - notebook for pan at clef 2012. In: Forner et al. [5]
19. Villatoro-Tello, E., Juárez-González, A., Escalante, H.J., Montes-Y-Gómez, M., Villaseñor-Pineda, L.: A two-step approach for effective detection of misbehaving users in chats - notebook for pan at clef 2012. In: Forner et al. [5]
20. Voorhees, E.M., Harman, D.K.: TREC: Experiment and Evaluation in Information Retrieval. Digital Libraries and Electronic Publishing, MIT Press (2005)
21. Yin, D., Xue, Z., Hong, L., Davison, B.D., Kontostathis, A., Edwards, L.: Detection of Harassment on Web 2.0. In: CAW 2.0 '09. Madrid, Spain (2009)