Identifying Sexual Predators by SVM Classification with Lexical and Behavioral Features

Notebook for PAN at CLEF 2012

Colin Morris and Graeme Hirst

Department of Computer Science, University of Toronto
{colin, gh}@cs.toronto.edu

Abstract We identify sexual predators in a large corpus of web chats using SVM classification with a bag-of-words model over unigrams and bigrams. We find this simple lexical approach to be quite effective with an F_1 score of 0.77 over a 0.003 baseline. By also encoding the language used by an author's partners and some small heuristics, we boost performance to an F_1 score of 0.83. We identify the most "predatory" messages by calculating a score for each message equal to the average of the weights of the *n*-grams therein, as determined by a linear SVM model. We boost performance with a manually constructed "blacklist".

1 Introduction and motivation

Our tasks were to distinguish sexual predators from non-predators in a corpus of online chats which was highly biased toward the negative class, and then to identify lines written by these alleged predators which were most indicative of their bad behaviour. Of the approximately 98,000 chat participants (whom we will generically refer to as "authors") in the PAN training corpus, 142 are identified as being sexual predators. We subdivide the complement class of non-predators into "victims" (anyone who ever talks to a predator — we have 142 in our training corpus), and "bystanders" (those who have no interactions with predators).

Although we read existing literature on the linguistic characteristics of sexual predators, such as [4], [5], [6], and [7], unlike some of the other teams we make no *a priori* assumptions about the language of sexual predators, and only the barest assumptions about predator behaviour (we merely assume that predators and victims chat in pairs, rather than in larger groups). Rather, we use a naive machine learning approach, wherein predators are defined solely and completely by the behaviour and language of the 142 predators identified by the ground truth of the training set.

We use a set of standard lexical features and features that generically describe the behaviour of chat participants. It's our hope that, given the success of our approach, a *post hoc* analysis of feature weights will suggest an empirically defensible model of "predatory language", and perhaps add or remove evidentiary weight to existing theories of predator language and behaviour.

We hypothesize that our classifier will be more effective if it can be attuned to both the language of *predatoriness* and the language of *victimhood*. For example, we imagine

that an adult engaging in a sexually explicit chat with another consenting adult might use language not unlike that of a sexual predator. However, we would expect the other participant to be an eager participant in the former case, and reticent or evasive in the latter case.

Thus, an important aspect of our approach is that a given author's feature vector reflects not just that author's language and behaviour, but also the language and behaviour of his or her interlocutor(s). This gives our machine learning algorithm roughly twice as much information to base its model on; we expect at least some of this additional information to be useful for discrimination, since we don't expect the language of one author to wholly determine the language of his or her interlocutor, notwithstanding the effect of lexical entrainment [1].

For the second task of predatory message identification, we return to our set of lexical features from the classification task. We train a linear SVM model for distinguishing predators from non-predators using just lexical features, and use the resulting weights over unigrams and bigrams to induce a weighting of "predatoriness" over all terms. We flag all predator messages where the sum of the weights of the terms in the message is above a certain hand-tuned threshold, along with all messages which contain any terms in a hand-assembled "blacklist".

2 Features

Our feature set can broadly be divided into lexical features and what we'll term "behavioural features", which capture patterns in the ebb and flow of conversation. Feature vectors are calculated on a per-author basis.

2.1 Lexical features

We use a standard bag-of-words model, since this has been shown to be robust in the face of a wide variety of text classification problems. Having also experimented with term presence, tf-idf, and log of term frequency, we ultimately settled on simple term frequency as our metric. We used both unigrams and bigrams.

As noted above, a key aspect of our approach to lexical features was our consideration of the language of the focal author's interlocutors as well as that of the focal author themselves. Thus every token t that appears more often than our threshold (empirically set to 10) yields two features: the number of times the focal author utters t, and the number of times any of the focal author's interlocutors utters t. We will henceforth refer to features of the latter type as "mirror" features. If we take the following short, imagined exchange as an example:

Author1: hi alice
Author2: hi hi

then Author1 would be associated with the following vector:

$$\{hi : 1, alice : 1, hi\ alice : 1, OTHER\_hi : 2, OTHER\_hi\ hi : 1\}$$

and Author2 would be associated with a mirror vector:

$$\{hi : 2, hi\ hi : 1, OTHER\_hi : 1, OTHER\_alice : 1, OTHER\_hi\ alice : 1\}.$$

We experimented with a number of standard text preprocessing routines including lowercasing, stripping punctuation, and stemming. None of these routines improved performance, thus our final results use simple space-separated tokens as features.

We also tried to add "smarts" to our lexical features with some transformation rules. We introduced the following special tokens:

\SMILEY For smiley faces matching a collection of emoticons assembled from Wikipedia (http://en.wikipedia.org/wiki/List_of_emoticons). We also introduce the following refinements:
\SMILEY_happy
\SMILEY_sad
\SMILEY_silly
\SMILEY_other

\MALE_name For tokens matching a list of the 1,000 most common male given names in the United States.[1] We manually removed around 10 names which are more likely to appear as common nouns (e.g. "Guy").

\FEMALE_name As above, for female names. In cases where a name can be both male and female, we choose the sex for which the name is more popular.

\NUM For any sequence of digits. We also introduce the following refinements on this category:
\NUM_small For $n < 13$.
\NUM_teen For $13 \leq n < 18$.
\NUM_adult For $18 \leq n < 80$.
\NUM_large For $n \geq 80$.

\PHONE_num For tokens matching any number of patterns for a phone number, with or without area code, with a variety of possible delimiters.

To our disappointment, these transformations seemed to add little discriminative power to our model; we will elaborate on and discuss this later in our results section. Unless otherwise specified, all results given below use only the simplest lexical features, without preprocessing or transformation rules.

2.2 Behavioural features

In addition to using the language of our authors, we explored high-level conversational patterns in order to exploit the small amount of metadata associated with conversations (mostly in the form of timestamps). In addition to looking at what words authors use, we're interested to see *how* they use them.

[1] We sourced our name lists from http://www.galbithink.org/names/us200.htm, using births from 1990 to 1999. The figures ultimately come from United States Social Security Administration.

Because we became interested in the secondary problem of distinguishing predators from victims (see section 3.1), many of these features are concerned with the problem of "symmetry-breaking". That is, given two authors who speak to one another using very similar language (which we found is often the case with predators and victims), what non-lexical aspects of the conversation can be used to distinguish them?

We used two "author-level" features which were straightforward to calculate on a per-author basis:

NMessages The total number of messages sent by this author in the corpus.

NConversations The total number of conversations in the corpus which this author participates in.

These two features were quite strongly correlated with predatorhood. This is probably an unintended side effect of the corpus construction, and we shouldn't use this fact to draw any conclusions about predator behaviour, such as "predators talk a lot".

Because of the large imbalance between the positive and negative class in the corpus and because there were anomalies on both sides (that is, predators with very few messages or conversations, and non-predators with many messages and conversations), these features alone are not enough to attain a reasonable F-score.

Initiative We employ a number of features which can be thought of as approximating an author's tendency to "initiate" with their partner:

Initiations The number of times this author initiates a conversation by sending the first message (this is usually something like "hey" or "what's up?").

Initiation rate The above variable normalized by number of conversations.

Questions The number of times this author asks a question, where we roughly define a question as any message ending in a question mark or interrobang.

Question rate As above, but normalized by number of messages.

Attentiveness Another set of features correspond to an author's attempts to keep a conversation going, and perhaps their level of commitment to the conversation.

Response time Messages in our corpus come with timestamps which are not guaranteed to be correct in an absolute sense, but which we assume are at least correct with respect to some time offset; thus, we expect the time deltas between messages to be accurate. Unfortunately, we have only minute-level precision. In a conversation between authors A and B we measure A's response times as follows: when we first see a message from B, we record the timestamp $t0$. We pass by any subsequent messages from B until we encounter a message from A and record its timestamp $t1$. The response time is $t1 - t0$. We seek ahead to the next message from B and repeat this process until the end of the conversation. We measure the mean, median, and max response times for each author, aggregated over all response times (rather than over all conversations).

This measure falls apart somewhat with conversations involving more than two authors. However, one of the few assumptions we make about predators and victims is that they always speak in pairs — and this is certainly true in the training data.

Repeated messages We measure the lengths of "streaks" of messages from the focal author which are uninterrupted by an interlocutor. The shortest allowable streak length is 1. Again, we record the mean, max, and median repeated messages.

Conversation dominance Our last set of features can be thought of as reflecting the degree to which the focal author "dominates" his conversations.

Message ratio The ratio of messages from the focal author to the number of messages sent by the other authors in the conversation, aggregated over all conversations in which the focal author participates.

Wordcount ratio As above, but using the number of "words" (space-separated tokens) written by each author.

3 Machine learning techniques and tools

Our machine learning algorithm of choice was support vector machines, using the LIBSVM library [3]. We used a radial kernel, having also experimented with a linear kernel. We return to the linear kernel in the predator message task (below), since unlike the radial kernel, it allows us to inspect feature weights to get a rough idea of the discriminative power of various features.

In testing our models, we used cross-validation with $n = 5$.

3.1 Results postprocessing

After classifying unknown authors using our model, we experimented with two later filters for boosting performance. Both steps were motivated by our observation that a large proportion of false positives (usually more than 75%) were in fact victims; thus predators and victims were quite similar in our dataset with respect to our lexical and behavioural features.

The first and most obviously effective step hinged on the assumption that the likelihood of two predators talking to one another was negligbly small. Thus, with our set of predicted predators, we returned to our corpus of conversations and found any pairs that ever talked to one another. For every such pair, we flipped the label of the author in whom the SVM had the least confidence (in addition to predicted labels, LIBSVM yields the confidence of each prediction). This increased precision at a small cost to recall.

The second filter used a second SVM model with the specialized task of distinguishing predators from victims (rather than predators from non-predators). After the first classification, we would run our predator-victim classifier on the alleged predators, and keep only the authors that were again labelled as predators. The rationale behind

this step was that the differences between predators and bystanders are quite coarse. This is due to the nature of the training set, where the non-predatory conversations tend to be very different from predatory conversations in terms of topic (e.g. IRC chatrooms on web programming), or in the relationship between interlocutors (e.g. short chats between anonymous strangers on Omegle, which contrast with predators and victims who tend to have repeated, sustained conversations).

Because predators and victims are discussing the same topics and are virtually identical in terms of number and length of conversations, we need to look to more fine-grained differences. This is what motivated our "symmetry-breaking" behavioural features such as message ratio, number of repeated messages, and number of initiations.

3.2 Predatory messages task

We trained a linear model for discriminating predators from non-predators using only our lexical features. We then treated the weight assigned to each term as an approximation of the "predatoriness" of that term. We assigned a predator score to each message equal to the sum of the weights of all unigrams and bigrams in the message, and flagged as predatory all messages with a predator score above a certain threshold. We hand-tuned this threshold so that what we deemed was a reasonable proportion of messages were flagged (approximately 2% to 5%).

We also build by hand a "blacklist" of 122 *n*-grams (including morphological variations and spelling variants) which automatically flag a message as predatory. Because we begin from the assumption that the messages we're classifying are all from predators to victims, we can choose words which have no conceivable place in an appropriate conversation between an adult and a child. Thus, these words don't automatically signal a message as predatory (since they may be employed in conversations between consenting adults), but they do signal a message as predatory when the message is from a predator to a victim.

Our blacklist focuses on terms which are sexually explicit, pertain to the exchange of photos, or pertain to arranging meetings. In an analysis of 51 chats between sexual predators and victims, Briggs et al [2] found that 100% of predators initiated sexually explicit conversations, 68.6% sent nude photos, and 60.8% scheduled a face-to-face meeting. We expect this blacklist to strictly increase recall, at a trivial cost to precision, if any.

Finally, we heavily penalize very short messages (those consisting of four or fewer space-separated tokens). This is based on the assumption that such short messages are unlikely to convey enough propositional content to be "predatory" (except, perhaps, with respect to the surrounding context), and on the volatility of taking averages over a small set of values.

4 Results

4.1 Predator classification

Using the default parameter settings for LIBSVM ($\gamma = 1/nfeatures$, $C = 1$), gave precision of 0.91, recall of 0.28, and F_1 score of 0.43 on the PAN training data. The large

Table 1. Cross-validated results ($n = 5$) on the predator classification task. The first row uses our optimized settings of c and γ with all features described in section 2 but without lexical transformation rules and without any postprocessing of results. Subsequent rows add or subtract features or steps for comparison. Note that the second row corresponds to the configuration used for our main submission to the competition. The last row is our baseline, resulting from labelling every author as a predator.

Variation	Recall	Precision	F1-Score
-	0.73	0.88	0.80
Partner flip	0.73	0.92	0.81
Predator-victim classification	0.65	0.89	0.76
Predator-victim classification and partner flip	0.65	0.91	0.76
Transformation rules	0.71	0.90	0.80
Transformation rules and partner flip	0.70	0.93	0.80
Only lexical features	0.74	0.93	0.82
Only lexical features with partner flip	**0.74**	**0.95**	**0.83**
Only focal lexical features	0.69	0.87	0.77
Only behavioural features	0.70	0.47	0.56
Baseline	1.0	0.001	0.003

disparity between recall and precision suggested that we needed to penalize errors in one class above those in the other. Setting the parameter $w1$ to 15, thus penalizing false negatives 15 times more than false positives, gave precision 0.63, recall 0.65 and F_1 score 0.64, thus optimizing F_1 score.

We performed a grid search to optimize the setting of parameters C and γ, varying them on a logarithmic scale. We settled on $C = 100$ and $\gamma = 10^{-4}$.

Table 1 gives our basic cross-validated results on the training data, along with the results associated with certain variations. Section 3.1 describes the "partner flip" and "predator-victim classification" filters. Our set of transformation rules are described in section 2.1. "Only focal lexical features" means that we only count the words used by the author under consideration (the "focal author") and not their interlocutors – see section 2.1.

Precision and recall alone don't give a full picture of the nature of our errors, since there is a hidden "third class" beyond predators and non-predators. There is a relatively high degree of confusion between predators and "victims" (those who chat with predators). Table 2 gives the confusion matrix for these classes in a basic run, and table 3 gives the confusion matrix for the same run following our "partner flip" filter. Note that these confusion matrices aren't square because in our classification scheme the "victim" and "bystander" classes are conflated into the class of "non-predators".

4.2 Message classification

Our results for the message classification subtask are given in table 4, evaluated on the ground truth given for the test data. Because our training data contains no ground truth for the message classification task, we're unable to give cross-validated results.

Table 2. Class confusion in a basic run.

Class	Predator	Non-predator	Total
Predator	104	38	142
Victim	8	134	142
Bystander	6	97532	97538
Total	118	97704	

Table 3. Class confusion in a basic run followed by our partner flip filter (see section 3.1).

Class	Predator	Non-predator	Total
Predator	103	39	142
Victim	3	139	142
Bystander	6	97532	97538
Total	112	97710	

Table 4. Results of the message classification task on the evaluation data.

Run	Precision	Recall	F_1 score	F_3 score
Standard run (submission)[a]	0.445	0.187	0.263	0.198
Standard run	0.544	0.192	0.**284**	0.205
Low predatoriness threshold	0.192	0.403	0.260	**0.403**
Low threshold, only weights	0.176	0.345	0.232	0.345
Only blacklist	**0.565**	0.181	0.274	0.194
Baseline	0.094	0.530	0.160	0.363

[a]For the sake of clarity and completeness, we include here our results as reported on the competition website, which are hindered by a bug which caused messages by alleged predators *and* victims to be considered. All other results reported here were obtained after this bug was fixed.

In preparing our submission, we didn't know that F_3 score would be the evaluation metric, nor what proportion of predator messages would be flagged. Thus our particular "standard" threshold, which resulted in high precision and low recall, put us in a relatively poor position. The "Low predatoriness threshold" run uses the same methods but a much lower minimum predatoriness score for messages (-0.03 rather than 0.01[2]), with the aim of improving recall and therefore F_3 score.

Note that our baseline involves selecting every message as predatory, even though it does not have 1.0 recall. This is because the pool of "predators" whose messages we classified was based on our classification in the previous step, rather than the ground truth (and thus, because we didn't achieve perfect recall in the first subtask, some predators don't even have their messages considered in this subtask). The interdependence of the subtasks also means that our baseline applies uniquely to our results, and not to those of other teams, who may have higher or lower baselines.

5 Discussion

5.1 Predator classification

Perhaps the most interesting feature of table 1 is the robustness of simple lexical features. Of our innovations – the mirror lexical features for conversational partners (see section 2.1), the partner flip and predator-victim classification filters, transformation rules, and behavioural features – only the mirror lexical features have an unambiguously positive effect on results, and some seem to diminish F-score when compared to the lexical baseline.

Table 5. The top and bottom 10 lexical features associated with predatorhood according to a linear SVM model. As in section 2.1 we use the convention that *OTHER_* preceding an *n*-gram denotes the use of that *n*-gram by the focal author's partner(s), rather than the focal author themselves. Note that in constructing this list, we set the minimum appearance threshold for *n*-grams to 30 rather than the typical 10, in an attempt to filter out spurious features.

Rank	*n*-gram
1	*OTHER_wtf*
2	*??? ???*
3	*hiiiii*
4	*asl*
5	*OTHER_no.*
6	*OTHER_hi*
7	*??*
8	*?*
9	*hello?*
10	*there*

Rank	*n*-gram
1	*???*
2	*now*
3	*now u?*
4	*so wat*
5	*hi*
6	*wat*
7	*OTHER_:(*
8	*so*
9	*around*
10	*what*

[2] Feature weights are not necessarily distributed symmetrically about 0; thus it would be facile to say that positive weights are predatory and negative weights are "anti-predatory".

Table 6. Statistics reflecting the distribution of our behavioural features across predators, victims, and bystanders.

Feature	Predator avg	Victim avg	Bystander avg
NMessages	288.58	296.28	8.44
NConversations	14.20	12.80	1.53
MessageRatio	0.523	0.486	0.471
WordcountRatio	0.560	0.455	0.472
NQuestions	35.70	42.49	1.39
MessageLength	2.658	2.060	1.705
Initiations	11.30	7.73	0.66
AvgResponseTime (minutes)	0.798	1.610	0.630

The partner flip step was generally effective, especially in maximizing $F_{0.5}$ score which was the evaluation measure for the competition. The improvement shown in table 1 is small because the partner flip is a step that's most effective in high-recall, low-precision runs, whereas ours tended to be the opposite. While our predator-victim classifier was quite accurate (having a cross-validated accuracy of 0.93 when applied to the predators and victims in our training data), it wasn't ultimately able to increase our F-score in the classification of predators and non-predators. Again, we suspect that the picture might have been different if our results had been skewed toward high recall and low precision rather than the opposite.

Omitting behavioural features seems to give a slight (0.03) increase in cross-validated F-score. A naive interpretation of this might be that behavioural features are actually harmful to accuracy. In fact, they do convey useful information about predatoriness, since our 12 behavioural features alone attain an F-score of 0.56, which is well above baseline (and which would place in the middle of the competition results). We suspect that the score increase when omitting these features is due to random noise. Applied to the evaluation data, it was the purely lexical model that gave a slightly lesser F-score.

We suspect that the negligible effects of our innovations are because the Pareto principle is at play in the data, wherein 20% of our features capture 80% of the instances in our corpus (in fact, the ratio may be more like 1% to 99%). This is supported by the fact, as noted above, that our mere 12 behavioural features can attain a stunningly high F-score of 0.56 on our highly imbalanced dataset (where the random baseline is 0.03). We claim that our transformation rules and behavioural features carry useful information about predatorhood, but that they unfortunately don't provide enough *new* information on top of our simple lexical features to increase performance.

Table 5 gives the 10 top and bottom lexical features associated with predatorhood. While we know that our simple lexical features are very effective at identifying predators, the feature weightings are surprisingly opaque. While the top 100 features contains a handful of obviously sexual *n*-grams (e.g. 18:*sexy*, 23:*wanna fuck*), the vast majority are common function words (e.g. 10:*there*, 24:*you*, 28:*my*, 40:*and*). Thus, it's not obvious how to draw a meaningful picture of predator language based on these weights.

Table 6 gives the average of some of our behavioural features across our three classes of authors. Although our behavioural features ultimately offered no improve-

ment on top of our lexical features, they were able to form a reasonably accurate classification model alone, and their distribution may offer some insights into predator and victim behaviour (in a way that our lexical features have not). As noted earlier, the trends in number of messages and conversations are artefactual and not much should be read into them. However, it's interesting that predators consistently send more and longer messages than their victim counterparts. Predators also initiate conversations almost twice as often as victims, and take, on average, less than half as long to respond to messages. The standard deviation for average response time among victims is 6.733, quite large compared to 1.053 for predators and 2.267 for bystanders. This suggests that the distribution for victims has a long tail, with victims often waiting long periods of time to respond.

These numbers paint a behavioural picture of the predator as someone who dominates conversations, and who is the more "eager" participant, tending to initiate conversations, and keep them going by responding quickly and voluminously.

5.2 Message classification

Despite our ranking of *n*-grams based on linear SVM weights being difficult to interpret, they were fairly effective at classifying messages. Our approach gives the highest precision of all submissions, our original submission achieving 0.445 precision, and 0.544 following a bugfix, above the next highest precision submission of 0.350, and well above the baseline of 0.092.

Our initial parameter setting gives an F_1 score of 0.284, which is well above the baseline of 0.169. Our best F_3 score is achieved by setting a low threshold for predatorscore, giving $F_3 = 0.403$. To our surprise, our baseline of labelling every message as predatory achieves an F_3 score of 0.363, which bests all but the aforementioned run, and which handily exceeds all submissions to the competition.

The "Low threshold, only weights" row of table 4 shows that our SVM weights alone achieve a respectable F_1 score (0.232, exceeding the baseline of 0.160).

As we would expect, the blacklist alone achieves the highest precision, at a cost to recall. We were surprised to see that precision was only 0.565, since we had constructed our blacklist in such a way that we thought all terms would be unambiguously "predatory". Examining the false positives from this run reveals that most could be argued to belong to the class of predatory messages, for example:

```
<conversation id=027600c74917a8d2438070be950fc2b6>
    <message line=40>i wanna kiss, etc</message>
    <message line=42>lick</message>
</conversation>

<conversation id=0730400af8a1b5a8aa88146baf417191>
    <message line=15>so you wont be sleeping naked tonight
        I take it</message>
    <message line=71>so what are you wearing?</message>
    <message line=84>so does she have a cam?</message>
    <message line=90>what would you show me on cam?</message>
</conversation>
```

References

1. Brennan, S.E., Clark, H.H.: Conceptual pacts and lexical choice in conversation. Journal of Experimental Psychology: Learning, Memory, and Cognition 22 (1996)
2. Briggs, P., Simon, W.T., Simonsen, S.: An exploratory study of internet-initiated sexual offenses and the chat room sex offender: Has the internet enabled a new typology of sex offender? Sexual Abuse: A Journal of Research and Treatment 23 (2011)
3. Chang, C.C., Lin, C.J.: LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology 2, article 27 (2011), software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm
4. Malesky, L.A.: Predatory online behavior: Modus operandi of convicted sex offenders in identifying potential victims and contacting minors over the Internet. Journal of Child Sexual Abuse 16(2), 23–32 (2007)
5. Marcum, C.D.: Interpreting the intentions of Internet predators: An examination of online predatory behavior. Journal of Child Sexual Abuse 16(4), 99–114 (2007)
6. 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 (2011)
7. Pendar, N.: Toward spotting the pedophile: Telling victim from predator in text chats. In: First IEEE International Conference on Semantic Computing. pp. 235–241. Irvine, CA (2007)