Identifying Online Sexual Predators by SVM Classification with Lexical and Behavioral Features [1]

Master of Science paper,

Department of Computer Science, University of Toronto

Colin Morris

January 30, 2013

[1]This paper incorporates portions of [Morris and Hirst, 2012], a paper I wrote with Graeme Hirst and presented at the PAN 2012 lab in Rome.

Abstract

We present a method for picking out sexual predators from a collection of online chats, and for identifying messages which are especially suggestive of predatory behaviour. We use support vector machines with unigram and bigram counts — lexical features which have proven robust in the face of a variety of text classification problems. Because each text is the product of two (or, occasionally, more) authors, we use separate counts for each *n*-gram, one being the number of times the *n*-gram is uttered by the author under consideration, and the other the number of times it is uttered by that author's partner(s). In this way, we train our model simultaneously on the characteristics of "predator-like language" and of "victim-like language".

We augment these lexical features with what we term "behavioural features", which capture patterns in the ebb and flow of an author's conversations (e.g., turn-taking behaviour, message length), as well as in the larger constellation of an author's conversations (e.g., number of conversations, number of distinct conversational partners).

Finally, we experiment with some post-processing steps following our round of SVM classification which increase precision by filtering out false positives — in particular, "victims" who are labelled as predators, a phenomenon that we found greatly confounded our classifier.

We deployed this method in the sexual predator task at PAN 2012 lab on "Uncovering Plagiarism, Authorship, and Social Software Misuse". There, on an unseen corpus of 219,000 authors, 254 of them predators, our method retrieved 159 authors, 154 of them predators, giving a precision of 0.969 and a recall of 0.606, thus ranking fourth out of 16 submissions to the tasks. On the subtask of retrieving "predatory" messages, we achieved the highest precision of all submissions, and ranked third in F-score.

Ultimately, we found that, while our (dual) lexical features and postprocessing were mutually helpful, our behavioural features failed to add significant discriminative power. However, a classifier trained on these features alone performed well above baseline, and we observed strong variations in their distribution across classes, which may prove useful *per se*.

Contents

1 Introduction

1.1 Motivation

The traditional approach to identifying online sexual predators wherein officers or volunteers pose as minors in chat rooms and follow up on predatory overtures is inadequate in the face of the immense number of communication channels online and the number of messages passing through them. This suggests the need for automated tools capable of flagging likely online predators for the attention of law enforcement. Indeed we know that some such systems already exist. For example, Facebook is known to automatically flag some messages suspected of being predatory and pass them on to law enforcement[1].

A secondary concern is, having identified a possible predator, highlighting inculpatory evidence (presumably to be reviewed by a human). This is the motivation for classifying messages based on "predatoriness". The scenario where this message classification might be useful — and the one we kept in mind in developing our algorithm — would be something like the following:

1. With an automated algorithm, an investigator identifies a relatively small number of suspicious authors from a large collection of online chats.

2. For each suspicious author, the investigator runs the message classification algorithm to identify a relatively small set of the "worst" messages from all the author's chats. If these messages demonstrate predatory behaviour, the investigator confirms this and opens a full investigation.

3. Otherwise the investigator discards the author from consideration and continues to the next author (an efficient approach), or reads the author's chat logs in full to find any missed evidence (a conservative approach).

Although our goal is to devise a classification system for predators and predatory messages, our methods give us empirical indicators of predator language and behaviour "for free". This knowledge is likely to be generally useful for any system seeking to identify predators, or even outside of a computational context.

1.2 Tasks

This work began as a submission to one of the tasks in the 2012 PAN lab, *Uncovering Plagiarism, Authorship, and Social Software Misuse*[2]. Given a large corpus of web chats, 0.1% of them being predatory in nature, we had two related subtasks. The first, which we call the **predator identification task**, was to identify the predatory authors. The second, the **predatory messages task**, was to flag the messages of predatory chats which were "most indicative of misbehaviour". Note that performance on the second task is dependent on performance on the first task. An algorithm that performs poorly on the first task will be evaluating the degree

[1] See for instance `http://www.reuters.com/article/2012/07/12/us-usa-internet-predators-idUSBRE86B05G20120712`

[2] See `http://www.webis.de/research/events/pan-12`

of purported predatoriness of more non-predator messages (all of which are guaranteed to be false positives) and missing out on classifying the messages of true predators (which will inevitably include what will be false negatives).

In both cases, we have access to a large, representative training corpus, provided for the PAN lab, with which to train a model (and for developing an algorithm through cross-validation). We describe these corpora in more detail in section 3.2.

1.3 Defining predatorhood

Semantically, we think of "predatorhood" as having two essential ingredients. The first is an *age disparity*: a predator is an adult who chats with an underage individual. The precise divide between "underage" and "adult" varies with jurisdiction, but the age of 18 is a reasonable point. Wolak et al. [2010] points out that the vast majority of victims of online sexual predators are adolescents rather than young children, a fact which is frequently missing from the public perception of online predators.

The second ingredient is an element of *inappropriate intimacy*. That is, the adult must introduce or encourage intimate conversation. This is frequently sexual in nature, but may be less overt (e.g., *you're very pretty*, *I love you*). This may also manifest itself in invitations to meet the victim in person.

Though much has been written on the subject, we will defer any discussion of predator psychology or patterns of predator behaviour until later, when analysing our results. This is because we approached the task from a completely naive point of view, with no preconceptions about what predators are like or how they behave. Rather, we cast a wide net of features and left it to our machine learning tools to identify which are predictive of predatorhood.

1.4 Defining predatory messages

Defining a predatory message is more difficult than defining predatorhood, and there are a number of questions without obvious answers that we can ask:

Do we wish to consider each message in isolation, or should we consider the enclosing context as well in deciding whether a message is predatory? PAN organizers made no comment on this. In our imagined use case (a human reviewing a small selection of messages to confirm or deny a suspected author of being a predator), a sequence of three messages that together give strong evidence of predatorhood would certainly be useful.

Do we consider the victim's messages? For example, the other party in a conversation saying "I'm 13" is evidence of an age disparity, a necessary element of a predatory situation. Again, it's easy to see how this would be useful in our imagined use case, but in the domain of the PAN task, only messages sent by predators were considered in message classification.

Do we consider message predatoriness as binary? Or should we assign messages a numerical score? Or do we just wish to rank messages? Semantically, it seems most reasonable to think of messages as varying on a scale, especially if we take a statistical view of "message predatoriness" of a message *m* as being another way of talking about the probability of an author being a predator given that they uttered *m*. The PAN task required binary labels, but our algorithm outputs a floating-point value.

Should we aim to label a particular number of messages per author? Either a constant value, or as a proportion of their total messages? PAN organizers gave no indication of this, leading to the number of flagged messages varying drastically between participants, from a low of 51 to a high of over 77,000.

In preparing our methods for PAN, we deferred to organizers' answers to these questions (and, where none were available, we guessed). But it's plain to see that there are a number of other plausible choices that could have been made which would have resulted in a different task.

2 Related work

Prior to PAN 2012, there was little existing literature on identifying online sexual predators using computational techniques.

One notable early example is Pendar [2007], who used lexical features (unigrams, bigrams, and trigrams) with SVMs and k-NN machine learning. A notable difference is that he used no negative dataset — that is, no conversations not including a predator. He set the task of distinguishing predator from victim. This is a much more limited task, compared to our experimental setup, which is intended to include innocuous conversations and sexual conversations between consenting adults in the data. However, in developing our own algorithms, we found "victims" to be the greatest source of false positives. Thus, just distinguishing predator from victim is far from trivial. Pendar reports strong results using trigrams (F-scores of up to 0.908 with SVMS and 0.943 with k-NN), but surprisingly low results (F-scores between 0.415 and 0.575, where baseline is 0.5) with unigrams and bigrams.

Kontostathis et al have been working and reporting on their "ChatCoder" system since 2009 [Kontostathis, 2012]. Like Pendar, they evaluate their system on the task of distinguishing predator from victim. In contrast with Pendar, their methods are very domain-specific. ChatCoder 1 [Kontostathis, 2009] uses a simple dictionary of words and phrases grouped into broad categories (e.g. compliments, approach, isolation) and uses counts of these phrases as features to build a C4.5 decision tree. ChatCoder 2 [McGhee et al., 2011] does much the same thing but with richer, rule-based features. For instance a chat line is labelled as "grooming" if it "contains a communicative desensitization adjective (*horny*, *naked*) and either a first or second person pronoun or an action verb". These rules arise out of an *ad hoc* 'communicative model' of predator behaviour.

Between the release of the PAN sexual predator task and the publication of results, Bogdanova et al. [2012a] reported on "considerable variation in the length of sex-related lexical chains" between predators and non-predators, and suggested that "this could be a valuable feature in an automated pedophile detection system".

A few months later, Bogdanova et al. [2012b] presented their results on a dataset much smaller than the PAN corpus (they attempt to distinguish approximately 60 predators from 60 non-predators), claiming an accuracy of 0.94 using Naive Bayes with their lexical chain-related features combined with "emotional markers" as features. However, their PAN submission ultimately ranked 13th of 16, with an $F_{0.5}$ score of 0.03 (whereas the top five submissions all ranked above 0.85), suggesting that their featureset wasn't as valuable as they'd hoped, or that they weren't able to harness it with the appropriate learning techniques (they declined to submit a notebook of their methods to PAN).

3 Materials

3.1 Terminology

It will help to introduce a consistent terminology when referring to the materials of our experiments:

Author A participant in a conversation. We prefer this to the (more accurate, but sesquipedalian) "interlocutor".

Predator We contemplate the empirical properties of sexual predators in 1.3, but for the purposes of developing our algorithm and training our SVM models, a predator was defined only with reference to the ground truth provided at the outset of the PAN competition. When it comes to our corpus, an author is a predator if and only if they are identified as such in our ground truth file.

Victim A victim is an author who participates in a conversation with a predator.

Bystander A bystander is an author who is neither a victim nor a predator.

Message A block of text sent by a particular author, with an associated timestamp. The timestamps were not guaranteed to be correct in an absolute sense, but were presumably guaranteed to at least be correct with respect to some common time offset.

Conversation A sequence of messages from one, two, or many authors. It transpired that, by construction, a conversation in the corpus never contains a gap of more than 25 minutes between consecutive messages.

3.2 Chat corpus

Our training corpus was assembled by the organizers of the PAN 2012 lab, and contains web chats distributed over 97,689 authors, represented by anonymized numerical IDs. 142 of these authors were identified as being sexual predators. During the development of our algorithms, we were unaware of the provenance of these chats, and made no attempt to discover it. During the PAN lab, the corpus's creator, Giacomo Inches, disclosed the methods by which he created it.

Inches followed the division of web chats established by Pendar [2007] into:

I Predator/Other

(a) Predator/Victim (victim is underage)

(b) Predator/Pseudo-Victim (volunteer posing as child)

(c) Predator/Pseudo-Victim (law enforcement officer posing as child)

II Adult/Adult (consensual relationship)

3.2.1 Predator chats

This portion of the corpus corresponds to type (I) above. Inches notes that conversations of types (Ia) and (Ic) are difficult to come by, especially in the quantities required for an adequate corpus. Thus, he resorts to type (Ib). Conversations of this type are publicly available at the website `http://www.perverted-justice.com` (PJ). I'll call this subset of the corpus `C1`.

Figure 1 gives an example of a conversation in this subcorpus.

3.2.2 Nonpredator chats

Inches divides type (II) conversations (perhaps confusingly) into "false positives" (which he defines as *people talking about sex or shared topic with the "sexual predator"*) and "false negatives" (*general conversations between users on different topics*).

The "false negatives" are sourced from IRC chat archives available at `http://www.irclog.org/` and `http://krijnhoetmer.nl/irc-logs/`. These are said to be representative of "general conversations" containing "a variety of messages in length and duration". I'll call this subset of the corpus `C2`. An example of a chat from this subcorpus is given in figure 2.

The "false positives" are sourced from archived conversations taking place on the website `www.omegle.com`. These archives are available at `http://omegle.inportb.com/`. Inches notes that these Omegle conversations contain "abuse language and general silliness online" and that sometimes users "engage in cybersex". I'll call this subset of the corpus `C3`. An example of a sexual chat from this subcorpus is given in figure 3. Many conversations are not sexual in nature, and seem unlikely to induce false positives, such as the conversation in figure 4

3.2.3 Weaknesses

Because the components of the corpus are sourced from different contexts, it's important to try to minimize incidental or artefactual differences that allow us to discriminate between them merely on that basis. Our goal is to discover features and algorithms that accurately discriminate predatory chats from non-predatory ones. But if, for instance, our predator chats are all sourced from instant messaging conversations and our non-predator chats come from IRC chatrooms, then a "IM vs. IRC" classifier will serve our purpose just as well (and we may accidentally create one).

Figure 1: A chat from subcorpus `C1`, between a predator and a "pseudo-victim". Author2 is the predator.

Author1: u there?
Author2: Heey Sweety
Author2: Get some sleep?
Author1: hi
Author1: i just woke up lol
Author1: im so srry n please dont ever say i hadda enuff of u
Author2: Ididn't know what happened
Author1: im still tired
Author2: I was hoping you were getting some sleep
Author1: lol i did
Author1: i am gonna go back to bed but wanted to get on n tell u im srry
Author2: Not gonna call?
Author1: im just tired
Author1: i will tomrorow?
Author2: Ok Sweety, get some sleep
Author1: ty for bein nice
Author1: not bein mad
Author1: im just tired
Author2: Not mad
Author2: I understand
Author1: ty i will be on tomorrow
Author1: bfn :-*
Author2: After school?
Author1: yah
Author2: LOl MySweet
Author2: Try to get your friend to sott a pic or two
Author1: ok will do
Author2: shoot
Author1: nite greg
Author2: Nite Sweety
Author2: :-*gfn
Author2: See you tomorrow

Figure 2: A chat from subcorpus C2, an innocuous conversation over IRC involving several chatters.

Author1: telecon today?
Author2: apparently not
Author2: there is no agenda at least
Author2: is there something we should talk about
Author2: it is possible that I or masayuki will propose a change to wheel events
Author2: perhaps just a minor change
Author3: Author1, Author2, we can have a call if you like
Author2: delta values should probably be doubles
Author3: no!!!!!!
Author3: well, ok
Author2: but I'm not sure
Author3: if it's okay with MS
Author1: yep
Author2: I need to discuss with masayuki first
Author1: sorry, thought you meant is it ok to have a conference :-)
Author2: Author1: anything particular to discuss about?
Author1: the real last call! :-) .... and any node list updates?
Author2: I didn't meet jresig when I was in US...
Author2: so far only sicking's proposal can handle perhaps the most needed mutation event, DOMCharacterDataModified
Author2: it is possible that we'll implement it right after FF4, with moz prefix
Author2: so that people can try it out

To that end, we point out the following artefactual features that differ systematically between the three principal components of our corpus:

Number of conversation participants All conversations in C1 and C3 contain one or two active participants (the case where there's one participant corresponds to one author sending messages to another and receiving no responses). A large proportion of the conversations in C2 have more than two participants. This is just the nature of IRC, which is inherently a system for many simultaneous chatters. Thus, we can significantly boost the precision of any baseline system by discounting any conversation with more than two participants. We suggest that this is not an interesting criterion for discrimination, and says nothing about the nature of predator or victim.

Number of conversations No author in C3 engages in more than one conversation in the corpus. This is because authors on Omegle are anonymized. If an individual were to engage in a new conversation, he or she would do so under a new (anonymous) identity. Authors in C1 tend to engage in many conversations.

Number of distinct conversational partners C2 and C3 are snapshots of a particular chat channel, taken from an omniscient point of view. Given a particular author in one of these corpora, we expect the corpus to contain all their conversations in a particular time period (or at least a representative sample thereof). Conversations in C1 are all captured from the perspective of the pseudo-victim. In particular, this means that we will only see a predator's interactions with the corresponding pseudo-victim. We have

Figure 3: A sexual chat from subcorpus C3, taking place over Omegle. This is presumably the sort of conversation that Inches has in mind when he refers to this as the "false positive" corpus. There is an element of "intimacy" but no indication of an age disparity, thus we don't think of either party as a predator.

Author1: hi
Author2: hihi
Author2: m or f??
Author1: akjfkjasdkjfajkdfjkakdjflkajdlkfjaldkjflkjdflkjadljfkalkjdfaljkdfaljkd;lfjakldjf
Author1: alkjdflkjadlfjalkjdfalkjdf;lajddflkajdflkjad;lfjkalkjfajkfajkdfakljdflajkdflkajdflkajflkjadf-lkjadfljadflkjadlfjkadflkjalkjfajdf;lajdflkjalfkjakjdfajflajf;jkafd;jalfkdja;lkjdfaljkf;aljfaljf;ajdl
Author1: im a female a hot female
Author1: nothing to do
Author2: male
Author1: i need someone
Author1: someone who can take away my lonlyness
Author2: ok
Author1: i hope that can be you
Author2: I'm here
Author1: so how can you take away my problem
Author1: ?
Author1: wait i need to get off my panty to get masturbating
Author2: and you know that I just do what you have
Author1: oh really
Author1: i mean right now
Author2: yes
Author2: let's connect the webcam msn??
Author1: i hope we can do mutual masturbating
Author1: my webcam is broken
Author1: and thats why im bored
Author1: how about you lick my vagina
Author1: hold my tits
Author1: squeeze them
Author1: ah
Author2: yes
Author1: i like that feeling
Author2: send me a picture of you naked wheel
Author2: yes
Author1: wait can you be my pet
Author1: like you do everything i want example imagine me naked lick my vagina
Author2: yes
Author1: lick me everywhere
Author2: yes
Author2: your vagina and very Saburo
Author2: yes
Author2: give me your msn speak better for it
Author2: yes vagina masturbating to my face .... will go
Author2: hihi

Figure 4: An innocuous chat from subcorpus `C3` (Omegle). This shows that not every conversation in this subcorpus is what we would think of as a "false positive". Conversations frequently range over mundane subjects, and are often quite short (as in this case).

Author1: ask me 5 questions and i will answer them truthfully
Author2: you first.
Author1: you want me to ask you questions first?

no access to other conversations they engage in with authors other than the pseudo-victim. The upshot of this is that each author in `C1` has only one distinct conversational partner.

We discuss potential improvements to the corpus in section 8.1.

3.3 Predatory messages ground truth

We know that the ground truth for the predatory messages task (that is the list of IDs corresponding to messages in the test set sent by predators which were indicative of "misbehaviour") was constructed late in the competition, after teams had already submitted their results. We also know that the labelling was done by just one "expert", which Inches admits led "to exclusion of possibly relevant lines or the over-consideration of some others". We know nothing of the inclusion criteria the expert had in mind, nor his or her process in evaluating messages (e.g., whether he or she read them sequentially, or in isolation). We discuss some surprising discoveries we made in the ground truth when analyzing our results in section 7.2, and suggest improvements in section 8.2. The annotator labelled only the messages that were selected by at least one team, a fact which doesn't affect the results we present here, but which does limit the usefulness of this ground truth for any future experiments.

4 Features

The first step in training our model of predatory chats is to convert our data into a set of discrete numerical features. At the core of our algorithm, we classify at the author level (rather than classifying conversations or messages). We turn each author into a vector of features which describe the nature of their conversations in the corpus. 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.

4.1 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. In the bag-of-words model we treat a text as an unordered collection ("bag") of terms. 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 — that is, individual words, and

pairs of consecutive words. The use of bigrams increases performance while also greatly increasing the size of the feature-space, and thus the size of each vector; this increase is not unmanageably large though.

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 for young people in the United States.[3] 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.

[3]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.

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.

4.2 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.

Because we became interested in the secondary problem of distinguishing predators from victims (see section 5.3), 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 (pseudo-) 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 relate to the spurious features described in section 3.2.3, and although we found them to be strongly correlated with predator-status, 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.

4.2.1 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 As above, but 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.

4.2.2 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.

4.2.3 Conversation dominance

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

5 Classification method

5.1 Overview

5.1.1 Author classification

The core of our algorithm is as follows:

1. Create a vector corresponding to each author in the corpus, using the features described in the previous section.
2. Train an SVM model on the authors in our training set.
3. Label the authors in our test set using the model trained in 2.
4. Apply some postprocessing to these results.

These points all bear some elaboration. We discuss our reasons for using support vector machines and the parameters used in steps 2 and 3 in section 5.2. We discuss the postprocessing routines we use in step 4 in section 5.3.

As we iteratively refined our methods, we used cross-validation with $n = 5$. That is, we sliced our training corpus into five uniform, disjoint slices, used four of those as the training set in step 2, and used the remaining one as the test set in step 3. We repeated this five times, using each slice as the test set once, and report statistics aggregated over the five runs.

Our reported results, unless otherwise noted, use PAN's test corpus, to which we had no access during the development of our methods, as our test set, and the entirety of the training corpus as the training set.

5.1.2 Message classification

Our approach to choosing messages sent by predators which are "indicative of misbehaviour" emanates from our above approach to author classification. The set of putative predators that is the output of that algorithm becomes the input to our message classification algorithm.

This means that, so long as our author classification accuracy has imperfect precision, our algorithm will be operating on some messages not sent by true predators. This was a conceit of the PAN competition, and while it meant that the ranking of the message classification task was biased, it does reflect realistic considerations (cf. our imagined use case in section 1.1). Thus, having messages from a mix of predators and non-predators in this task is not an unrealistic choice.

5.2 Support vector machines

Our machine learning algorithm of choice was support vector machines, using the LIBSVM library [Chang and Lin, 2011]. SVMs have been shown to be effective at a variety of text classification tasks. SVMs also have

the advantage of dealing well with the high dimensionality of our data (we have on the order of 100,000 lexical features) — both in terms of speed, and the ability to disregard irrelevant features.

Unless otherwise noted, we used a radial kernel (`-t 2` in LIBSVM), having found it to give superior results compared to a linear kernel. The only time we return to a linear kernel is when we need a weighting over our features, such as when we want to determine the most and least "predatory" terms in our corpus.

We defer a detailed discussion of further SVM parameterization to section 6, so that we can refer to their effects on classification accuracy.

5.3 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.

5.3.1 Partner flip

The first and most obviously effective step hinged on the assumption that the likelihood of two predators talking to one another was negligbly small[4]. 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. We called this the "partner flip" filter.

5.3.2 Predator-victim classification

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

[4]The number of such cases in our training corpus was zero. We later learned that this was the case for the test set as well, and that it was by construction (see section 3.2)

"symmetry-breaking" behavioural features such as message ratio, number of repeated messages, and number of initiations.

5.4 Predatory messages task

In addition to the radial-kernel model, we also 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, at this point, 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 normally 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. [2011] found that 100% of predators initiated sexually explicit conversations, 69% sent nude photos, and 61% scheduled a face-to-face meeting. We expect this blacklist to strictly increase recall, at a trivial cost, if any, to precision.

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.

6 Results

6.1 Evaluation metric

We begin the presentation of our results with a discussion of an appropriate evaluation metric.

Because of the imbalance between classes, simple accuracy (#correctly classified/#total) is a poor choice of evaluation metric. Having 250 predators in the test set and on the order of 10,000 total authors, with a similar ratio in the training set, a system that labelled every author as a non-predator would achieve a baseline accuracy of about 0.98. Clearly, despite the impressively high number, this system is trivial and useless, whereas a system that detected every predator with a 0.1 false positive rate would be much more useful, despite having a lesser accuracy of around 0.9.

F-measure, the harmonic mean of precision and recall, is a standard evaluation metric in information retrieval and is robust in the face of an imbalance between the positive and negative class. However, the question of which version of F-measure to use arises. For any positive β, we can define F_β as:

$$F_\beta = (1+\beta^2) \cdot \frac{\text{precision} \cdot \text{recall}}{(\beta^2 \cdot \text{precision}) + \text{recall}}$$

The choice of β roughly corresponds to the degree of importance we place on recall over precision, or the cost we wish to associate with false negatives over false positives. F_1 measure (if no β is mentioned, "F measure" is assumed to refer to F_1 measure) gives equal weight to precision and recall. A larger choice of β corresponds to more weight on recall, and a smaller choice to more weight on precision.

For our task, the choice of β is not obvious. In developing our algorithm, we placed a higher emphases on recall. We imagined that the cost of a false positive would be small. A false positive wouldn't correspond to falsely accusing someone of being a sexual predator. Human (expert) judgement is still the gold standard for this task, and we imagined our algorithm producing a set of 'suspicious' chatters which would be reviewed by a law enforcement officer. Presumably, a false positive could be identified as such by a human and discarded in a short time span (say, 1-10 minutes). On the other hand, there's a high amount of utility associated with catching a predator (and therefore a high cost to false negatives).

To our surprise, the organizers of PAN 2012 used $\beta = 0.5$ in ranking submissions, weighting precision over recall. They apparently imagined our algorithms being used in quite a different scenario. As described in section 1.1, Facebook is known to use algorithms to automatically flag messages suspected of being sent by sexual predators. These suspcious messages are read by employees and then sent to police if appropriate. In this case, because of privacy concerns, there is a considerable cost to a false negative, far beyond the time it takes an employee to read the message. They wish to minimize the intrusion on users' privacy and read as few messages as possible.

This suggests that choice of evaluation metric is highly application-sensitive. In this section, we hedge by generally discussing F_1 measure. We can bias our results toward either precision or recall by simply adjusting the parameter to our SVM classifier ($-w1$ in LIBSVM) that controls the relative cost of false negatives and positives, thus maximizing F measure for any particular choice of β. In our case, F_1 measure is a good predictor of how well a particular configuration generalizes to other choices of β.

6.2 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 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.

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 4 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 PAN 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

Table 2: Class confusion in a basic run.

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

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 5.3 describes the "partner flip" and "predator-victim classification" filters. Our set of transformation rules is described in section 4.1. "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 4.1.

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".

Table 3: Class confusion in a basic run followed by our partner flip filter (see section 5.3).

	Labelled as		
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.

6.3 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.

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

Table 5: The top and bottom 10 lexical features associated with predatorhood according to a linear SVM model. As in section 4.1 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*

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

7 Discussion

7.1 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 4.1), 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. This is at least heartening in that it shows that our results aren't dependent on the artefactual features that we highlighted in section 3.2.3.

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 improvement 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.

7.2 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 to PAN, 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 predator-score, 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, even though the expert annotator thought otherwise. 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>
```

7.3 Comparison to other PAN submissions

We present a brief comparison of our methods and results with those of the other teams participating in the sexual predator task in PAN 2012[6].

[6]A complete table of results is available in Inches and Crestani [2012], or online at `http://www.uni-weimar.de/medien/webis/research/events/pan-12/pan12-web/authorship.html`.

7.3.1 Predator classification

Of 16 submissions to the predator classification task, our method placed fourth with an $F_{0.5}$-score of 0.865, trailing Parapar et al. [2012] with 0.869, Snider with 0.917 and Villatoro-Tello et al. [2012] with 0.935.

Inches and Crestani [2012] point out that SVMs were the most common machine learning tool used by competitors, with neural networks, maximum-entropy, decision trees, *k*-NN, and naive Bayes also making appearances. Features were all either lexical or behavioural, except in the case of one team that used character-level features with a string kernel.

Parapar et al. used tf-idf lexical features along with behavioural features which resembled our own quite closely. Villatoro-Tello et al., the top-scoring team, applied a "pre-filtering" step to remove conversations that were clearly irrelevant (based on number of participants, number of messages sent, and the presence of "noisy" text), thus reducing the number of considered conversations and authors by about 90%. From there, they took a two-step classification approach: the "Suspicious Conversation Identification" step identified conversatons thought to be between predator and victim, and then the "Victim From Predator" step attempted to choose the conversation participants who are the predators. This pre-filtering step was apparently very effective, though it's not clear that such a step could be done cleanly in a "real-world" scenario, rather than the artefactually heterogeneous corpus used in the PAN competition.

Eriksson and Karlgren [2012], who ranked fifth, slightly below us, were, to our knowledge, the only other team to overtly incorporate in the features of a given author the language of their conversational partner (they use the terms SLEX for *self*-lexical and OLEX for *other*-lexical features).

7.3.2 Message classification

Any fine-grained comparison of results with other competitors in the message classification subtask is bound to be confounded by the dependence on the previous task, but we give a brief overview of other approaches to the subtask. A striking fact is that the top-ranking team, Popescu and Grozea [2012], submitted *all* messages sent by their alleged predators — essentially a baseline approach, with no smarts. This is not even accounted for by the dependence on the predator classification task since Popescu and Grozea ranked only seventh in that task. What this does suggest is that F_3-score is a poor evaluation metric, since it leads to trivial algorithms dominating; if F_1-score had been used, Popescu and Grozea would have ranked only fifth.

Among other submissions, the use of a dictionary of "perverted" terms to score messages (similar to our "blacklist" approach) was common, as was the use of terms correlated with predatorhood in the training set using tf-idf.

Our submission ranked fifth of fourteen by F_3-score and third by F_1-score. Our precision was highest by a fairly wide margin (0.45, with the next-highest submission being 0.36), but our recall of 0.19 hurt our performance under the F_3-score metric.

8 Future work

8.1 Corpus

The PAN corpus has a number of virtues as a corpus for testing algorithms for sexual predator detection, and for comparing their performance with others. It is large, contains a realistic ratio of predatory to non-predatory conversations, and includes cases that are difficult to classify (e.g., adults engaging in consensual sexual conversation). However, as noted in section 3.2.3, it has certain obvious flaws that made our task easier than it would have been in a "real-world" situation. One future direction could be improving the realism of the PAN corpus.

Use of the PJ dataset for true positives introduces the problem of having a limited view of a predator's conversations: we only see the conversations they engage in with the pseudo-victim, making "number of distinct conversational partners" an excellent feature for increasing precision. We could remedy this by augmenting the PJ dataset with "fake conversations" between the predator and a new author. This could be accomplished by sourcing conversations from the Omegle dataset, and setting the predator as one of the authors. This means "number of distinct conversational partners" is no longer an attractive feature, and also simulates the realistic scenario wherein predators engage in some conversations which are non-predatory.

A more subtle problem with the use of the PJ dataset is that we're observing the behaviour of predators with "pseudo-victims" (adult volunteers acting as children). It's not clear how closely their behaviour mirrors that of true victims, or what effect this has on the behaviour of the predators with whom they're interacting. But as Inches and Crestani [2012] and others have noted, interactions between real predators and real victims are scarce and difficult to come by.

We also noted that "number of conversations" is also a strong feature for identifying Omegle conversations (which all belong to the negative class). We could remedy this by merging distinct Omegle authors in such a way that the distribution of number of conversations per author is statistically similar to the distributions in the IRC corpus and the PJ corpus.

8.2 Message ground truth

As we discussed in section 7.2, we found the ground truth for which messages were predatory to be incomplete. Also, only the messages that were put forth by at least one team were reviewed by a human for the ground truth, making it difficult to use in evaluating any future experiments. For future experiments of this nature, we would suggest the construction of a ground truth based on the majority decision of a panel of three or more judges, each given clear written instructions. (The PAN ground truth was the result of a single evaluator working from intuition, rather than a well-defined mode.) We feel this would be desirable, even if it came at the cost of having a smaller test set (say, 1,000 messages rather than 10,000).

References

D. Bogdanova, S. Petersburg, P. Rosso, and T. Solorio. Modelling fixated discourse in chats with cyberpedophiles. *EACL 2012*, page 86, 2012a.

D. Bogdanova, S. Petersburg, P. Rosso, and T. Solorio. On the impact of sentiment and emotion based features in detecting online sexual predators. *WASSA 2012*, page 110, 2012b.

Susan E. Brennan and Herbert H. Clark. Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 22, 1996.

Peter Briggs, Walter T. Simon, and Stacy Simonsen. 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.

Chih-Chung Chang and Chih-Jen Lin. 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`.

G. Eriksson and J. Karlgren. Features for modelling characteristics of conversations. *in Forner et al.*, 2012.

P. Forner, J. Karlgren, and C. Womser-Hacker, editors. *CLEF 2012 Evaluation Labs and Workshop — Working Notes Papers*, Rome, September 2012.

G. Inches and F. Crestani. Overview of the international sexual predator identification competition at PAN-2012. In *CLEF 2012 Evaluation Labs and Workshop — Working Notes Papers. Rome, Italy*, 2012.

A. Kontostathis. Chatcoder: Toward the tracking and categorization of internet predators. In *Proc. Text Mining Workshop 2009 held in conjunction with the Ninth Siam International Conference on Data Mining (SDM 2009). Sparks, NV. May 2009*. Citeseer, 2009.

April Kontostathis. Chatcoder. `http://www.chatcoder.com`, 2012.

I. McGhee, J. Bayzick, A. Kontostathis, L. Edwards, A. McBride, and E. Jakubowski. Learning to identify internet sexual predation. *International Journal of Electronic Commerce*, 15(3):103–122, 2011.

Colin Morris and Graeme Hirst. Identifying sexual predators by SVM classification with lexical and behavioral features. *in Forner et al.*, 2012. Available at `http://ftp.cs.toronto.edu/pub/gh/Morris+Hirst-PAN-2012.pdf`.

Javier Parapar, David E. Losada, and Alvaro Barreiro. A learning-based approach for the identification of sexual predators in chat logs. *in Forner et al.*, 2012.

N. Pendar. Toward spotting the pedophile telling victim from predator in text chats. In *Semantic Computing, 2007. ICSC 2007. International Conference on*, pages 235–241. IEEE, 2007.

Marius Popescu and Cristian Grozea. Kernel methods and string kernels for authorship analysis. *in Forner et al.*, 2012.

Esaú Villatoro-Tello, Antonio Juárez-González, Hugo Jair Escalante, Manuel Montes y Gómez, and Luis Villasenor-Pineda. A two-step approach for effective detection of misbehaving users in chats. *in Forner et al.*, 2012.

J. Wolak, D. Finkelhor, K.J. Mitchell, and M.L. Ybarra. Online predators and their victims. *Psychology of Violence*, 1:13–35, 2010.