Identifying Sexual Predators by SVM Classification with Lexical and Behavioral Features

Colin Morris and Graeme Hirst

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

September 20, 2012

Colin Morris and Graeme Hirst

Predator classification

Lexical features

- Term frequency over unigrams and bigrams.
- Experimented with lowercasing, stemming, stripping punctuation etc. Best results came from using space-separated tokens.
- *n*-gram appearance threshold of 10 to lower dimensionality and reduce noise.

"Mirror" lexical features

- Goal: model attuned simultaneously to the language of "predatoriness" and "victimhood".
- Thus, each n-gram g yields two features: the number of times the focal author utters g, and the number of times any of the focal author's interlocutors utters g.

Example

Author1: hi alice
Author2: hi hi

Yields the following feature vector for Author1:

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

with Author2 associated with the following mirror vector:

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

Behavioural features

"Behavioural features" reflect the structure of conversations, rather than their content. Examples:

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.

Initiative

Initiations The number of times this author initiates a conversation by sending the first message.

Initiation rate As above, but normalized by number of conversations.

Questions The number of times this author asks a question.

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

Attentiveness

Response time The time between between the focal author's messages and the most recent preceding non-focal message.

Repeated messages The average length of "streaks" of messages from the focal author which are uninterrupted by an interlocutor. The shortest allowable streak length is 1.

Conversation dominance

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.

SVM classification

- We use support vector machines (SVMs).
  - Successful and robust in the face of a variety of text classification tasks.
- LIBSVM implementation.
- Radial kernel.
- Set $C = 100$, $\gamma = 10^{-4}$ based on a grid search of parameter values.
- Featuresets and parameterisations tested by cross-validation on training set, with $n = 5$.

Partner flip

- We found that many of our false positives were in fact *victims*.
- Thus, if two alleged predators (per the SVM classification step) talk to each other, we flip the label of the one in whom we have the least confidence.

Effects on results

- Lexical features alone: $F_1 = 0.82$
- Behavioural features alone: $F_1 = 0.56$
- Lexical + Behavioural features: $F_1 = 0.80$
- "Partner flip" postprocessing step increases precision at a small cost to recall.
- Behavioural features fail to improve results when combined with lexical features, but perform respectably by themselves. They also show interesting and dramatic variations across our three classes.

Results (cont'd)

Feature	predator	victim	bystander
NMessages	288.58	296.28	8.44

Results (cont'd)

Feature	predator	victim	bystander
NMessages	288.58	296.28	8.44
NConversations	14.20	12.80	1.53

Results (cont'd)

Feature	predator	victim	bystander
NMessages	288.58	296.28	8.44
NConversations	14.20	12.80	1.53
MessageRatio	0.523	0.486	0.471

Results (cont'd)

Feature	predator	victim	bystander
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

Results (cont'd)

Feature	predator	victim	bystander
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
QuestionRate	0.078	0.069	0.096

Results (cont'd)

Feature	predator	victim	bystander
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
QuestionRate	0.078	0.069	0.096
MessageLength	2.658	2.060	1.705

Results (cont'd)

Feature	predator	victim	bystander
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
QuestionRate	0.078	0.069	0.096
MessageLength	2.658	2.060	1.705
InitiationRate	0.796	0.604	0.431

Results (cont'd)

Feature	predator	victim	bystander
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
QuestionRate	0.078	0.069	0.096
MessageLength	2.658	2.060	1.705
InitiationRate	0.796	0.604	0.431
AvgResponseTime (minutes)	0.798	1.610	0.630

Colin Morris and Graeme Hirst

Message classification

SVM weights

- We train a linear SVM model using the same lexical features described earlier.
- We treat the weight the SVM model assigns to an *n*-gram as a proxy for its "predatoriness".
- Calculate the predatoriness of a message as a sum of the weights of all unigrams and bigrams. Flag the message if its score is above an arbitrary threshold.
- Add a considerable penalty to the score of short messages (≤ 4 tokens).

SVM weights

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*

Blacklist

- We manually construct a "blacklist" of 122 unigrams and bigrams which have no conceivable appropriate use in a conversation between an adult and a child.
- We expect this to strictly increase recall at no cost to precision.

Effects on results

- Blacklist alone and SVM weights alone both exceed random baseline. Best results come from combining both approaches.
- But blacklist approach doesn't give 1.0 precision. We found many messages containing blacklisted terms which weren't flagged as predatory. For example:

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

- We consider their exclusion debatable.

Conclusions

- Simple bag of words + SVM approach gives excellent results with low effort.
- Behavioural features add little new information, but provide interesting insights into predator behaviour.
- The weights assigned to terms by a linear SVM, while opaque, are useful for selecting predatory messages.
- A manually constructed blacklist is a simple and effective augmentation.