Recognizing Predatory Chat Documents using Semi-supervised Anomaly Detection

Mohammadreza Ebrahimi, Ching Y. Suen, Olga Ormandjieva, Adam Krzyzak; Department of Computer Science, Concordia University; Montreal, Quebec/Canada

Abstract

Chat-logs are informative documents available to nowadays social network providers. Providers and law enforcement tend to use these huge logs anonymously for automatic online Sexual Predator Identification (SPI) which is a relatively new area of application. The task plays an important role in protecting children and juveniles against being exploited by online predators. Pattern recognition techniques facilitate automatic identification of harmful conversations in cyber space by law enforcements. These techniques usually require a large volume of high-quality training instances of both predatory and non-predatory documents. However, collecting non-predatory documents is not practical in real-world applications, since this category contains a large variety of documents with many topics including politics, sports, science, technology and etc. We utilized a new semi-supervised approach to mitigate this problem by adapting an anomaly detection technique called One-class Support Vector Machine which does not require non-predatory samples for training. We compared the performance of this approach against other state-of-the-art methods which use both positive and negative instances. We observed that although anomaly detection approach utilizes only one class label for training (which is a very desirable property in practice); its performance is comparable to that of binary SVM classification. In addition, this approach outperforms the classic two-class Naïve Bayes algorithm, which we used as our baseline, in terms of both classification accuracy and precision.

Introduction

During the past decade, automated online Sexual Predator Identification from chat documents has boomed by means of pattern recognition techniques capable of flagging likely predators for the attention of law enforcement. The most common approach has been presented in PAN-2012 international competition [1] which was specifically engineered to accomplish the following two tasks [2]:

- Finding the predators vs. victims
- Finding the predatory messages in a predatory document

The first task seems to be more important for law enforcement since it can help them to limit their search space drastically. It is worth mentioning that the second task has not been as successful as the first one due to the fact that it requires deeper natural language analysis.

The first task can be performed in two steps [3]:

- Identifying the predatory documents in the entire conversation corpus
- Searching in participants of predatory documents in order to distinguish the sexual predator and victim

In this paper we focus on the first step mentioned above (i.e. identifying the predatory conversations), since it will be the most proper area for helping the investigators in real-world applications. Accordingly, the main motivation behind using One-class SVM on this kind of data and treating the problem as an anomaly detection problems is making a classifier which is able to learn from only one class label instead of what we have in the traditional binary classification. Figure 1 depicts the different granularity levels for designing classifiers in online sexual predator identification.

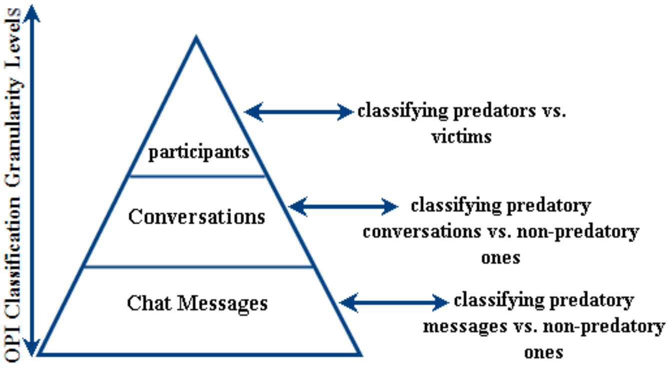

OPI Classification Granularity Levels
participants
Conversations
Chat Messages
classifying predators vs. victims
classifying predatory conversations vs. non-predatory ones
classifying predatory messages vs. non-predatory ones

Figure 1. Classification Granularity Levels and their corresponding classification problem in SPI

Section 2 describes the current status of SPI, section 3 explains the proposed approach which is based on semi-supervised anomaly detection, and section 4 dissects the document recognition process we conducted on SPI problem including pre-processing, feature extraction and pattern classification. Also, the result of comparing different methods is described in this section.

Motivation

According to researchers who participated in PAN-2012, There has been a major weakness in the data set: The non-predatory and non-sexual samples were exclusively gathered from publicly available IRC logs which mainly contain the chats about computer and web technologies; therefore cannot represent "general conversations" [4]. The samples in general conversation category (which are also non-predatory) must include countless topics such as sport, music, games, computer, etc. In practice, it is not an easy task to assemble such a training data set. As a result, the current top-ranked algorithms in PAN2012 may have learned how to distinguish computer-related chats vs. sexual-related chats instead of identifying actual predatory chats in online cyber space. Accordingly, one can expect that their performance will decrease in real-world applications. In other words, we believe that although the top-ranked algorithms in PAN-2012 had significant F_1-score on test data set (87% for the winner), since they require general samples that are able to represent the non-predatory data properly, their performance will decrease significantly in practical environments such as law enforcement. In this work, we propose a novel way to handle this problem by eliminating the need for having both class labels in the train data set. Due to the absence of one of the class labels in the training process, our applied method will be more practical at the expense of having a lower, but still acceptable, F_1-score. Using only one class label in training process categorizes this approach as a semi-supervised classification method. Furthermore, in order to guarantee the efficiency of our approach we aim to beat the baseline (naïve Bayes algorithm) in terms of F_1-Score.

Note that each chat conversation represents a document in our recognition process; hence, in the remaining parts of this paper we use document and conversations interchangeably.

Related Work

Perhaps the first successful attempt for using machine learning in SPI problem was done by Pendar by means of weighted K-NN classifier to distinguish predators from underage victims [5]. To the best of our knowledge, the first empirical system with capability of determining predatory messages in chat logs is ChatCoder1 (and Chatcoder2) implemented and evolved by Kontostathis and her colleagues [6] [7]. The system uses a rule based approach in conjunction with decision trees and instance-based learning methods (K-NN). It is worth mentioning that in order to deal with the issue of learning imbalance data, [8] has already introduced a general approach using a weighted version of KNN algorithm to mitigate the problem of imbalanced data in text categorization which is not specifically related to the SPI.

Recently, the PAN-2012 conference has acted as a boost for applying machine learning techniques to this area. The main strength of this conference is providing the first publicly available official data set which was specifically engineered for sexual predator identification task. Researchers tuned their proposed methods against the same training data and reported their performance on the test data. Several machine learning algorithms have been used to solve SPI problem in this competition. These algorithms cover a wide range of classification algorithms such as maximum entropy-based classification [9], K-NN [10], Support Vector Machine [4] and Neural Networks [3]. Eventually, one team has been announced as the winner based on their classification accuracy and an augmented F-measure. The winner team [3] has used a two-step binary classification approach called SCI (Suspicious Conversation Identification) and VFP (Victim From Predator Disclosure) using SVM and Neural Networks. Accordingly we have used SVM as the state-of-the-art method to compare the performance of our anomaly detection approach with. Escalante and his colleagues [11] proposed a new method based on learning a chain of three local classifiers corresponding to three segments of each document (i.e. conversation) but the approach could not outperform that of the winner in PAN-2012.

A related research has been done on cyber bullying by Kontostathis which is very close to predator identification [12]. They utilize a different supervised learning algorithm based on latent Semantic Indexing which is called Essential Dimensions of LSI for identifying cyber bullying. They built their own data set using Form spring.me, a questin-and-answer popular website.

As the most recent work, [13] have proposed enriching the traditional bag-of-word language model by adding other feature types including sentiment features, psycho-linguistic features and discourse patterns. Eventually, they have used binary classification for the actual predator identification task.

Generally, the algorithms used in PAN-2012 can be considered as the state of the art in sexual predator identification. While in regard to anomaly detection, there is a wide variety of unsupervised, supervised, and semi-supervised models. A comprehensive survey of anomaly detection has been done in [14]. The authors have categorized the anomaly detection methods into six major categories: clustering based, classification based, nearest neighbor based (also includes density based methods), statistical, Information theoretic and spectral methods. We use a slightly different taxonomy to show the place of the method we use based on the learning method that is used for anomaly detection. We avoid describing different methods and foundations of anomaly detection since it is beyond the scope of this article. Instead, we focus on the specific anomaly detection method (i.e. one-class SVM) that yielded the desirable results in this application domain. Figure 2 illustrates the taxonomy of most common anomaly detection techniques as well as the position of semi-supervised techniques.

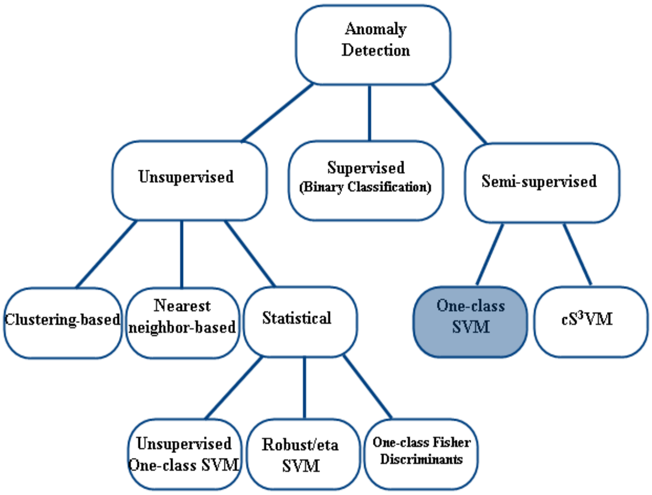


Figure 2. Position of Semi-supervised and SVM-based techniques in the taxonomy of anomaly detection techniques

One-class SVM has been highlighted in the figure. For the sake of completeness, the unsupervised SVM-based algorithms are shown as well. The corresponding leaf nodes of the taxonomy will be introduced in the next section.

Recently, several works have addressed the problem of anomaly detection in micro-blogs or short messages especially in Twitter [15], [16]. In [17], Kumaraswamy et al. have used domain-specific features encoded as first order logic for textual anomaly detection. Anomaly detection methods have not been applied to the SPI problem. As a new area of application, we examined this approach on the PAN-2012 data and we discuss the results of this approach and compare it with other widely used two-class classification methods in the following sections. Next section discusses the adaptation of the notion of anomaly detection to sexual predator identification.

Anomaly Detection

Let data set $\boldsymbol{D}$ be defined as $\boldsymbol{D} \subset \boldsymbol{X} \times \boldsymbol{Y}$ where $\boldsymbol{X} = \{\chi_1, \chi_2, ..., \chi_{\boldsymbol{n}}\}$ is the set of $\boldsymbol{n}$ documents. Each document is denoted by an m-dimensional feature vector

$\chi_i = (\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_m)^T$ where $\boldsymbol{i} \in \{1,2,..., \boldsymbol{n}\}$ and $\boldsymbol{x}_j$ is the j^{th} feature value of vector χ_i where $\boldsymbol{j} \in \{1,2,..., \boldsymbol{m}\}$. Also let $\boldsymbol{Y} = \{p, np\}$ represent the set of two class labels corresponding to predatory and non-predatory instances respectively. In a probabilistic setting, it is assumed that each document χ_i is roughly drawn from probability distribution $\mathrm{P}(\chi)$. The anomaly detection task is defined as finding a probability distribution P such that $\mathrm{P}(\chi)$ is near one for the majority of samples considered as normal and contrarily close to zero for the majority of anomalous samples. One can choose $\boldsymbol{l} \in \Re$ as the threshold for recognizing a document as a predatory one when $\mathrm{P}(\chi) \le \boldsymbol{l}$. The notion of anomaly is a domain-specific concept which is related to the properties of the problem domain. This means that an anomalous sample in a specific domain might be considered as normal in another area of application. Figure 3 shows the probabilistic view of anomaly detection in SPI problem for a conversation with only two features $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$.

Anomaly detection which is also known as novelty or outlier detection is often referred to finding instances which do not conform to the underlying pattern of normal data [14]. The following two research questions arise in regard to rationalization of applying semi-supervised methods to sexual predator identification:

Why not use unsupervised anomaly detection?

This can be shown that supervised and semi-supervised anomaly detection methods outperform unsupervised methods in terms of performance [18]. We focused on semi-supervised techniques due to their superior predictive power compared to that of unsupervised methods. Although according to [18], the predictive power of semi-supervised methods comes at the expense of having weakness in identifying actual novel samples, in the domain of sexual predator identification, this weakness does not have a drastic impact due to the lack of such novelties that we may deal with in another domain such as network intrusion detection.

Why not use supervised anomaly detection?

As already mentioned, in this application domain, providing non-predatory samples is not practical at all. So we utilize a semi-supervised anomaly detection method that is capable of learning from only one class label in contrast to the binary (i.e. two-class) classification methods. In this setting, positive and negative instances are mapped to predatory and non-predatory conversations.

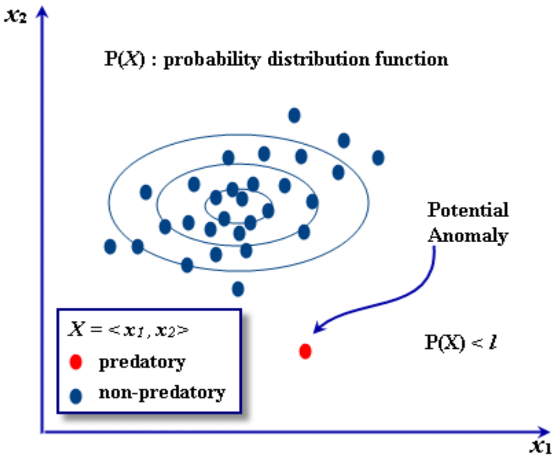


Figure 3. Probabilistic view of anomaly detection in SPI setting (while predatory samples are considered anomalous)

Moreover, one of the circumstances that justifies using an anomaly detection approach is when the data is naturally imbalanced. Due to the fact that predatory samples are rare compared to non-predatory ones, we usually deal with data sets containing several hundred predatory conversations among several hundred thousands of non-predatory conversations.

It is worth mentioning that one can apply a reverse notion of anomaly in a manner that considers predatory conversations as normal ones and non-predatory conversations as anomalous.

One-class SVM

One-class SVM has been introduced by Scholkopf as a novelty detection technique and has been widely used in the area of anomaly detection [19]. The algorithm is a variation of ν-SVM [20] which uses parameter $\nu \in [0,1]$ to control the fraction of support vectors as well as fraction of outliers (anomalies). It is worth mentioning that in original SVM choosing the best regularization parameter $\boldsymbol{C} \in [0, \infty)$ is a real challenge. ν-SVM tries to ameliorate this problem by introducing parameter $\boldsymbol{\nu}$ that indirectly affects the regularization. The main idea of One-class SVM is providing an algorithm which returns a function f with output +1 in a small region capturing most of the data points, and -1 elsewhere. The constrained optimization problem is defined as follows [21]:

$$\min_{\boldsymbol{w}, \boldsymbol{\xi} \in \Re, \boldsymbol{\rho} \in \Re} \left(\frac{1}{2} \|\boldsymbol{w}\|^2 + \frac{1}{\boldsymbol{\nu l}} \sum_i (\xi_i - \boldsymbol{\rho}) \right) \tag{1}$$

$$\mathit{with\,regardto}: \boldsymbol{w}$$

$$\mathit{subjectto}: (\boldsymbol{w}.\Phi(\boldsymbol{\chi}_i)) \ge \boldsymbol{\rho} - \xi_i, \quad \xi_i \ge 0$$

In which n is the number of conversations in data set, ρ parameterizes a hyperplane in the feature space, w is the weight vector, ξ is the slack variable which penalizes the objective function and Φ is the internal mapping function used in the kernel. Note that in this notation (.) denotes the inner dot product.

The optimization problem can be solved by using the following Lagrangian in which $\alpha_i, \beta_i \geq 0$ [21].

$$L(w, \xi, \rho, \alpha, \beta) =$$

$$\frac{1}{2}\|w\|^2 + \frac{1}{\nu l}\sum_i (\xi_i - \rho) - \sum_i \alpha_i \left((w.\Phi(\chi_i)) - \rho + \xi_i \right) - \sum_i \beta_i \xi_i \quad (2)$$

Finally, the decision function for one-class SVM will be obtained as follows:

$$f(\chi) = \mathrm{sgn}(w.\Phi(\chi)) \quad (3)$$

Besides the original method described above, there is another variant of semi-supervised SVM-based technique for anomaly detection called cS^{3}VM [22]. This method is based on the cluster assumption (i.e. there is a one-to-one mapping between clusters and classes.) Since the optimization problem in this setting is non-convex, the authors leverage a method to convert the non-convex optimization problem to a convex one by using a method called smoothing in an iterative manner.

Based on the given taxonomy, there are also several unsupervised methods for anomaly detection. One-class SVM can naturally be used in an unsupervised setting as well [23]. Moreover, there are two unsupervised variations of SVM which have been recently introduced by Amer et al. [23] called robust-svm and eta-svm. Since these versions are completely unsupervised, they sacrifice the performance (i.e. accuracy, precision and recall) too much, so we chose to use the original method in this study. Using one-class SVM has led to acceptable results in the area of anomaly detection, but it has not been utilized in such a problem yet. In the following section, we describe the data set as well as the results of applying this method on SPI problem.

Document Recognition Process

This part describes the pattern recognition process that we have conducted on SPI problem including the data set, pre-processing, feature extraction, and pattern classification. Figure 4 shows the process that we have conducted on the chat log data set.

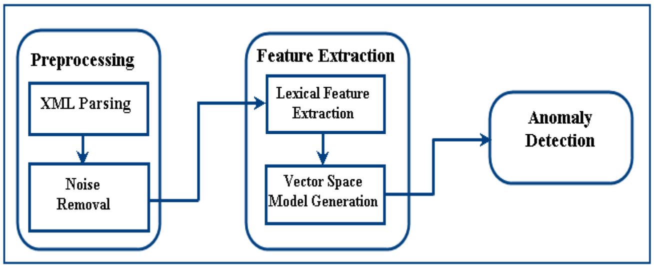
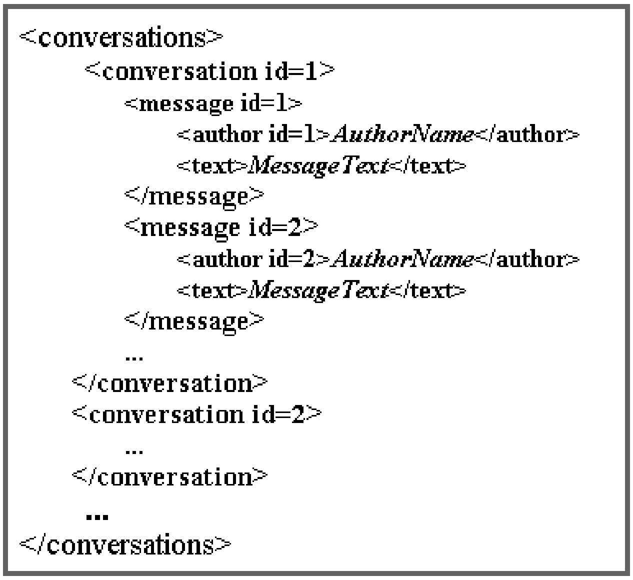


Figure 4. Proposed Recognition process performed on chat log documents

Data set

We used the training and testing data set in PAN-2012 which is the largest and most practical publicly available data set so far according to our knowledge [1]. Both the training set and test set are in XML format. The data schema has been shown in Figure 4.



```
<conversations>
     <conversation id=1>
          <message id=1>
               <author id=1>AuthorName</author>
               <text>MessageText</text>
          </message>
          <message id=2>
               <author id=2>AuthorName</author>
               <text>MessageText</text>
          </message>
          ...
     </conversation>
     <conversation id=2>
          ...
     </conversation>
     ...
</conversations>
```

Figure 5. Document schema of conversations in PAN-2012's data set

The data set is highly imbalanced. It contains 66927 conversations in the training set and 155128 conversations in the test set. There are 2016 and 3737 predatory conversations in training and testing set respectively. These predatory conversations are related to 142 (out of 97695 unique users) and 254 (out of 218,716 unique users) predators respectively. The total number of exchanged messages in the training corpus is 903,607. Another challenging aspect of the data set is the large number of consensual sexual conversation between adults which should be ideally recognized as non-predatory documents because of the fact that no minor is involved in this consensual chats.

Problem Setting

In our experimental setting we chose Naïve Bayes as a common binary text classifier as our baseline. Also we tried to simulate the results of the winner team in PAN-2012 for identifying predatory conversations based on Support Vector Machines. We performed two main categories of experiments: 1) training the model with only non-predatory samples, and 2) training the model with only predatory samples. Table 1 shows the experiments we have conducted. We refer to each experiment by its shortened name and describe the corresponding results in the next section.

Table 1. Different Experiments Conducted in this setting

Experiment No.	Experiment Short Name	Experiment Description
1	*Train-NP-B*	Train one-class SVM on non-predatory conversations and bigram features
2	*Train-P-B*	Train one-class SVM on predatory conversations and bigram features
3	*Test-P-B*	Test one-class SVM on predatory conversations and bigram features
4	*Train-P-B-NR*	Train one-class SVM on predatory conversations with bigram features after noise removal

5	*Test-P-B-NR*	Test one-class SVM on predatory conversations with bigram features after noise removal

It is worth mentioning that instead of using binary classification; one can formulate the problem as a multi-class classification problem with *m* classes in which *m-1* classes are devoted to topical classes in negative category such as (commercial, political, etc.) and the m^{th} class represents the predatory conversations. However, this design might not be as efficient as binary classification although it does some extra work beyond the scope of our problem definition.

The training set have been evaluated via k-fold cross validation with *k=10* and micro-averaging the results for each fold. In order to evaluate the performance of algorithms, four common performance criteria have been used: accuracy, precision, recall, and F-Measure. Normally, the last measure is calculated as the harmonic mean of precision and recall and called F_1-score, unless one wants to weigh either precision or recall more than the other one. The general formula for F-score is as follows [24]:

$$F_\beta = \frac{(\beta^2+1)PR}{\beta^2 P + R}; \quad (0 \le \beta \le \infty) \quad (4)$$

In PAN-2012 international competition, both β=1 and β=0.5 were used as the main performance measures. The latter was used to put more emphasis on precision and could raise controversies. Accordingly, in order to consider precision as important as recall, we use β=1 and calculated the widely-acceptable F1-score as the performance measure. We used RapidMiner™ [25] as an open-source powerful tool for our preprocessing and also LibSVM [26] for C-SVM and One-class SVM. The designed pre-processing steps are available on github at https://github.com/mohammadrezaebrahimi/pre-process-PAN.git as an XML file which can be imported in Rapidminer. The process includes parsing, feature extraction, noise removal and feature selection tasks which are described in the remaining of this section.

Preprocessing and Feature Extraction

In the preprocessing phase, we parsed the XML data and extracted the raw textual document for each conversation. As most of the successful approaches in this domain, we leveraged the bag-of-words model for feature extraction in our experiments and generated both unigram and bigram representation of the data to examine the performance of training on these two different features. Typically, there are three options for data representation in text classification: 1) binary representation in which the occurrence of the specified term is encoded as 1 or 0 otherwise, 2) Term Frequency (number of occurrences), and 3) TF/IDF which has several different variants. We used the normalized weighting schema which is calculated as follows:

$$\mathbf{w}_{tf/idf}(t,c) = \mathbf{log}(1+\mathbf{tf}(t,c))\mathbf{log}\left(\frac{N}{1+\mathbf{df}(t)}\right) \quad (5)$$

Whereas ***t*** and ***c*** denote *term* and the *conversation* in which term has appeared respectively, ***N*** is the total number of conversations and *tf(t,c)* is the term frequency of term ***t*** in conversation ***c***. Finally, *df(t)* is the number of documents in which the term ***t*** has occurred. We used TF/IDF weighting schema based on the approach of the winner team in PAN-2012. Afterwards the unigram or bigram features were obtained by regular *tokenization* and *stop-word removal* in RapidMiner™. The resultant unigram and bigram vector space models for training data set contain 45450 and 280378 features respectively.

As a side note, among bag-of-words approaches, unigram and bigram features are the most common representation techniques among bag-of-words approaches used in this domain. While Pendar has used trigram features some other researchers [27] have used Kernel-based features in character level instead of word level. But their method's performance is not as successful as bag-of-words methods.

Also it is wise not to use *stemming* while we are dealing with conversational documents which usually have informal writing styles. Because performing noise removal (in term level) as well as stemming will distort the stylistic patterns the authors use in their conversations. According to [3] the predator may try to maintain the connection by writing "soryyyyyyyyy" in case the child feels bad about the inappropriate intimacy. As a result we did not use any stemming for dimensionality reduction in our preprocessings.

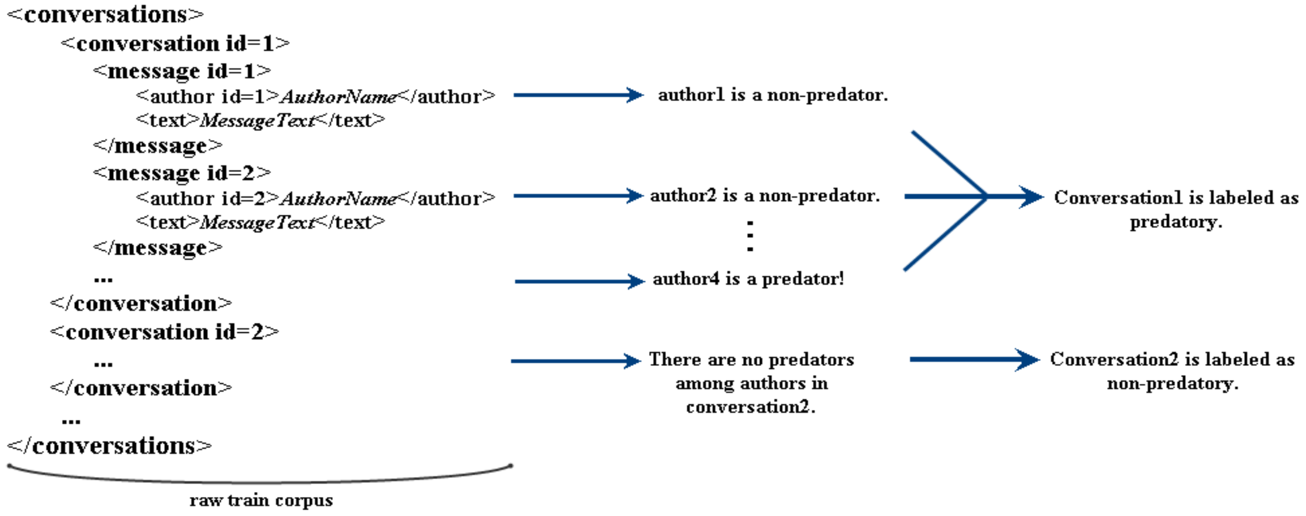

<conversations>
<conversation id=1>
<message id=1>
<author id=1>AuthorName</author>
<text>MessageText</text>
</message>
<message id=2>
<author id=2>AuthorName</author>
<text>MessageText</text>
</message>
...
</conversation>
<conversation id=2>
...
</conversation>
...
</conversations>
raw train corpus
author1 is a non-predator.
author2 is a non-predator.
author4 is a predator!
There are no predators among authors in conversation2.
Conversation1 is labeled as predatory.
Conversation2 is labeled as non-predatory.

Figure 6. Labeling Conversations in Training Data

Figure 6 depicts the preprocessing procedure used for labeling conversations as predatory or non-predatory. Note that although we have labeled both predatory and non-predatory conversations in training data set, we use only one of these two classes in model training unlike binary classifiers which leverage both of the class labels.

Feature Selection

In order to select the most salient features we fed the primary features obtained from the previous phase into a supervised feature selection algorithm called Information Gain. That is the amount of reduction in the entropy that might be obtained by leveraging feature t. Information gain for dataset D and candidate feature t is calculated based on the following formula:

$$IG(D \mid T) = H(D) - H(D \mid T) \qquad (6)$$

In which H() represents information entropy. We conducted several feature selection experiments to conduct the best bigram feature set. The feasibility of each feature set was based on the performance of classification using that feature set. We calculated the information gain for each of the features in the data set and then sorted them in increasing order of their corresponding information gain. Then the top k-percent of the ordered set was selected each time to make five feature sets. Table 2 shows the feature sets in this experiment:

Table 2. Different feature sets and their corresponding top-k selected features

No.	Top K-Percentage	Number of features
1	60%	168227
2	70%	196265
3	80%	224302
4	90%	252340
5	100%	280378

Then we performed one-class SVM classification algorithm on each of the above five data sets and measured the ultimate performance by four criteria: Accuracy, Precision, Recall, and F_1-measure. The following diagram shows the performance for mentioned feature sets. As it can be seen, the feature set containing 224302 features has the best performance. We used this feature set for building the classification model.

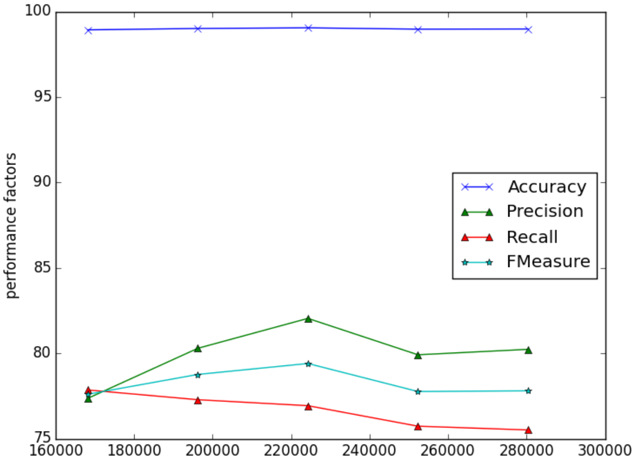


Figure 7. Changes of performance criteria versus number of features

Pattern Classification

In this part, we describe the achieved results and compare them with the base line and SVM as highly-standard binary classification method which was used by winner of PAN-2012. The training has been done via 10-fold cross validation and then the resultant model has been applied on the standard test set described in section 4-1. First we assess whether the one-class SVM should be trained on non-predatory or predatory conversations. In the first case, we trained the model on negative samples by filtering out the predatory samples. In this case, one-class SVM learns the distribution of none-predatory conversations. Tables 3 and 4 show the results for training the model on non-predatory conversations and predatory ones respectively. For a discussion on parameter optimization, please refer to the last section of this part. From comparing these two tables it can be inferred that training the model on predatory conversations yields better results.

Table 3. Results of training on Non-predatory samples (Experiment *Train-NP-B*)

Learning Algorithm	Performance Measures[a] (%)			
	Acc.	Pre.	Rec	F_1
Naïve Bayes (with Laplace smoothing)	84.9	16.6	100	28.5
SVM (regularization parameter C=10)	99.5	99.2	84.8	91.5
One-Class SVM (lower bound parameter nu=0.1)	24.2	2.3	59.4	4.5
One-Class SVM (lower bound parameter nu=0.13)	76.7	4.4	32.1	7.7

Table 4. Results of training on predatory samples (Experiment *Train-P-B*)

Learning Algorithm	Performance Measures[b] (%)			
	Acc.	Pre.	Rec	F_1
Naïve Bayes (with Laplace smoothing)	84.9	16.6	100	28.5
SVM (regularization parameter C =10)	99.5	99.2	84.8	91.5
One-Class SVM (lower bound parameter nu=0.2)	**98.0**	**65.1**	**70.7**	**67.8**

[a.] Indicates percentage of accuracy, precision, recall and F_1-Score.

But when we apply the model on the test set, the results are not so promising (Table 5). Particularly the precision rate is too low.

Table 5. Results of testing on predatory samples before noise removal (Experiment *Test-P-B*)

Learning Algorithm	Performance Measures[c] (%)			
	Acc.	Pre.	Rec	F_1
Naïve Bayes *(with Laplace smoothing)*	81.4	10.8	91.8	19.3
SVM *(regularization parameter C =10)*	98.4	75.5	50.4	60.5
One-Class SVM *(lower bound parameter nu =0.2)*	68.7	**5.8**	79.0	11.3

[b.] Indicates percentage of accuracy, precision, recall and F_1-Score.

We believe that this behavior is due to the fact that the anomaly detection algorithms are more sensitive to noise than binary classification algorithms. Accordingly, we conducted a new series of experiments after doing a naïve noise removal procedure to examine the effect of noise removal on performance improvement. For our noise removal procedure, we simply omitted the conversation with just one participant. Tables 6 and 7 show the results after performing noise removal on the train and test data respectively. As we expected, even though the performance of all of the algorithms has increased after removing useless data, the noise removal procedure affects the performance of one-class SVM more significantly compared to that of other methods. Accordingly, the F-measure rises from 11% to 75%. This confirms our hypothesis about the sensitivity of one-class SVM to the noise.

Table 6. Results of training on predatory samples after noise removal (Experiment Train-P-B-NR)

Learning Algorithm	Performance Measures[d] (%)			
	Acc.	Pre.	Rec	F_1
Naïve Bayes *(with Laplace smoothing)*	84.3	13.1	100	23.2
SVM *(regularization parameter C =10)*	99.9	99.9	95.7	97.8
One-Class SVM *(lower bound parameter nu =0.2)*	**99.0**	**80.2**	**75.5**	**77.8**

[c.] Indicates percentage of accuracy, precision, recall and F_1-Score.

Table 7. Results of testing on predatory samples after noise removal (Experiment Test-P-B-NR)

Learning Algorithm	Performance Measures[e] (%)			
	Acc.	Pre.	Rec	F_1
Naïve Bayes *(with Laplace smoothing)*	80.3	10.7	91.9	19.2
SVM *(regularization parameter C =10)*	98.5	78.1	50.1	61.0
One-Class SVM *(lower bound parameter nu =0.2)*	**98.2**	**70.7**	**44.5**	**54.6**

Indicates percentage of accuracy, precision, recall and F_1-Score.

As it can be observed, one-class SVM outperforms the baseline and its performance is comparable to binary SVM. Figure 7 summarizes the above results at a glance.

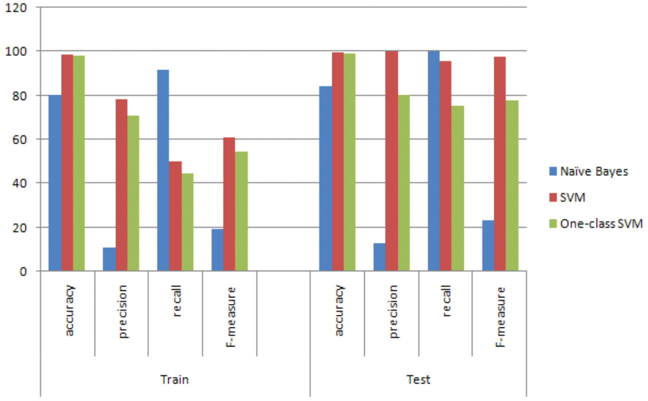


Figure 7. Comparison of the anomaly detection approach with naïve Bayes and SVM

To summarize, we observed that when we added a noise removal module into the process, One-Class SVM out-performs the base line (Naïve Bayes) and its performance is comparable with two-class SVM in this application domain.

We can also draw the following two subsidiary conclusions: Firstly, NB is superior with a high percentage of recall (100% on train set and 91% on test set) which implies that in terms of lower leakage rate (i.e. false negative), the base line defeats other approaches. Secondly, SVM outperforms other methods with the highest percentage of precision (78.13%). In other words SVM has the lowest false alarm rate (i.e. false positive) among the applied methods.

Parameter Optimization Remarks

As we discussed earlier, one-class SVM needs the parameter $\boldsymbol{\nu} \in [\mathbf{0}, \mathbf{1}]$ to be tuned. Although the parameter is bounded, it turns that this parameter optimization is a challenging task for which there is no exact formal solution. In order to estimate a good value for this parameter we used the exhaustive grid search which simply tries the entire set of combinations of parameters in a classification problem and chooses the best parameter setting based on the performance criterion (i.e. F$_1$-score). In this case, we considered $\boldsymbol{\nu}$ as the main parameters for tuning. Using a linear discretization, we chose 15 discrete points out of the interval of parameter $\boldsymbol{\nu}$ in a linear manner into 15 points: [0.66, 0.13,0.2, 0.26, …, 1]. Based on the performance evaluation, $\nu = 0.13$ in experiment setting *Train-NP-B* and $\nu = 0.2$ in experiment setting *Train-P-B-NR* revealed the best performance results. We used the same approach for estimating the value of regularization parameter in SVM binary classification. Although this approach does not necessarily lead us to the global optimum, it is a typical parameter setting approach which is used excessively in practical pattern recognition tasks.

Conclusion and Future Works

We carried out a novel successful application of anomaly detection for online predator identification which is of more use in practice compared with the current binary classification approaches that require non-predatory samples to be learned. Although as a semi-supervised technique we only used the predatory samples to train our model, as the results show, not only our approach outperforms the baseline learning algorithm (Naïve Bayes), also it is even comparable to the state-of-the-art binary classification algorithms.

In order to increase the performance of our model, we plan to combine the Naïve Bayes algorithm with the current model through designing an ensemble of heterogeneous classifiers. This way, we aim to also obtain the benefit of high recall rate of Naïve Bayes algorithm. Also we plan to apply other mentioned semi-supervised anomaly detection approaches on the data set in order to compare the performance of the method with them. As a side note, in practice we need to conduct feature selection technique or dimensionality reduction to improve the performance. Accordingly, we would also increase the performance of our method via supervised feature selection techniques such as information gain or gain ratio.

Acknowledgment

This research has been sponsored by Natural Sciences & Engineering Research Council of Canada (NSERC) and also Office of Vice President, Research & Graduate Studies at Concordia University (OVPRGS). The authors would like to thank Colin Morris for illuminating discussions and sharing his experiences about his previous work on sexual predator identification.

References

[1] "PAN-2012." [Online]. Available: http://pan.webis.de.

[2] G. Inches and F. Crestani, "Overview of the International Sexual Predator Identification Competition at PAN-2012," CLEF (working notes), 2012.

[3] E. Villatoro-Tello, A. Juárez-González, H. J. Escalante, M. Montes-y-Gómez, and L. Villaseñor-Pineda, "A Two-step Approach for Effective detection of Misbehaving Users in Chats," Rome, Italy, 2012.

[4] C. Morris, "Identifying Online Sexual Predators by SVM Classification with Lexical and Behavioral Features," Master of Science Thesis, University Of Toronto, Canada, 2013.

[5] N. Pendar, "Toward spotting the pedophile telling victim from predator in text chats," Washington, USA, 2007, pp. 235–241.

[6] A. Kontostathis, "Toward the tracking and categorization of internet predators," in *In Proceeding of Text Mining Workshop 2009 held in conjunction with Ninth Siam International Conference Data Mining*, 2009.

[7] I. Mcghee, J. Bayzick, A. Kontostathis, L. Edwards, A. Mcbride, and E. Jakubowski, "Learning to Identify Internet Sexual Predation," *Int. J. Electron. Commer.*, vol. 15, no. 3, pp. 103–122, 2011.

[8] S. Tan, "Neighbor-weighted K-nearest Neighbor for Unbalanced Text Corpus," *Expert Syst Appl*, vol. 28, no. 4, pp. 667–671, May 2005.

[9] G. Eriksson and J. Karlgren, "Features for modelling characteristics of conversations," Rome, Italy, 2012.

[10] I.-S. Kang, C.-K. Kim, S.-J. Kang, and S.-H. Na, "IR-based k-Nearest Neighbor Approach for Identifying Abnormal Chat Users," Rome, Italy, 2012.

[11] H. J. Escalante, E. Villatoro-Tello, A. Juárez, M. Montes-y-Gómez, and L. Villaseñor, "Sexual predator detection in chats with chained classifiers," in *Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, Atlanta, Georgia, 2013, pp. 46–54.

[12] A. Kontostathis, K. Reynolds, A. Garron, and L. Edwards, "Detecting Cyberbullying: Query Terms and Techniques," in *Proceedings of the 5th Annual ACM Web Science Conference*, New York, NY, USA, 2013, pp. 195–204.

[13] A. Cano, M. Fernandez, and H. Alani, "Detecting Child Grooming Behaviour Patterns on Social Media," in *Social Informatics*, vol. 8851, L. Aiello and D. McFarland, Eds. Springer International Publishing, 2014, pp. 412–427.

[14] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM Comput. Surv.*, vol. 41, no. 3, p. Article No. 15, 2009.

[15] P. Anantharam, K. Thirunarayan, and A. Sheth, "Topical anomaly detection from Twitter stream," in *4th Annual ACM Web Science Conference (WebSci '12)*, New York, NY, USA, 2012, pp. 11–14.

[16] J. Guzman and B. Poblete, "On-line relevant anomaly detection in the Twitter stream: an efficient bursty keyword detection model," in *ACM SIGKDD Workshop on Outlier Detection and Description (ODD '13)*, New York, NY, USA, 2013, pp. 31–39.

[17] R. Kumaraswamy, A. Wazalwar, T. Khot, J. Shavlik, and S. Natarajan, "Anomaly Detection in Text: The Value of Domain Knowledge," in *Florida Artificial Intelligence Research Society Conference*, 2015.

[18] N. Görnitz, M. Kloft, K. Rieck, and U. Brefeld, "Toward Supervised Anomaly Detection," *J Artif Int Res*, vol. 46, no. 1, pp. 235–262, Jan. 2013.

[19] B. Schölkopf, R. Williamson, A. Smola, J. Shawe-Taylor, and J. Platt, "Support Vector Method for Novelty Detection," *Adv. Neural Inf. Process. Syst.*, vol. 12, pp. 582–588, 2000.

[20] B. Schölkopf, A. J. Smola, R. C. Williamson, and P. L. Bartlett, "New Support Vector Algorithms.," *Neural Comput.*, vol. 12, no. 5, pp. 1207–1245, 2000.

[21] B. Schölkopf, J. C. Platt, J. C. Shawe-Taylor, A. J. Smola, and R. C. Williamson, "Estimating the support of a high-dimensional distribution," vol. 13, no. 7, 2001.

[22] O. Chapelle, M. Chi, and A. Zien, "A Continuation Method for Semi-supervised SVMs," in *Proceedings of the 23rd International Conference on Machine Learning*, New York, NY, USA, 2006, pp. 185–192.

[23] M. Amer, M. Goldstein, A. Slim, and S. Abdennadher, "Enhancing One-class Support Vector Machines for Unsupervised Anomaly Detection, ODD'13," Chicago, IL, USA, 2013, pp. 8–15.

[24] N. Chinchor, "MUC-4 Evaluation Metrics," in *Proceedings of the 4th Conference on Message Understanding*, Stroudsburg, PA, USA, 1992, pp. 22–29.

[25] I. Mierswa, M. Wurst, R. Klinkenberg, M. Scholz, and T. Euler, "YALE: Rapid Prototyping for Complex Data Mining Tasks," in *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY, USA, 2006, pp. 935–940.

[26] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, pp. 27:1–27:27, 2011.

[27] M. Popescu and C. Grozea, "Kernel Methods and String Kernels for Authorship Analysis," in *Notebook for PAN at CLEF 2012*, 2012.

Authors Biography

Mohammadreza Ebrahimi is a research assistant at CEnter of PAttern Recognition and Machine Intelligence (CENPARMI) at concrdia University in Canada. He has done several practical projects in the area of Document Categorization, Text Mining, and Textual Pattern Recognition. He holds the Power Corporation of Canada Graduate Award in 2015 and Concordia University 25th Anniversary Award in 2015 which both are granted based on academic excellence.

Dr. Ching Y. Suen is the Director of CENPARMI and the Concordia Chair on AI & Pattern Recognition. He received his Ph.D. degree from UBC (Vancouver) and his Master's degree from the University of Hong Kong. He has served as the Chairman of the Department of Computer Science and as the Associate Dean (Research) of the Faculty of Engineering and Computer Science of Concordia University.

Dr. Olga Ormandjieva is an Associate Professor in the Computer Science and Software Engineering Department at Concordia University, Montreal, Quebec (Canada) and a member of the Ordre des Ingénieurs du Québec OIQ). She holds a Ph.D. in Computer Science (2002) and Master's Degree in Computer Science and Mathematics (1987).

Dr. Adam Krzyzak is a Professor at Department of Computer Science at Concordia University. He has served as associate editor of Pattern Recognition Journal and also Research Fellow in Department of Computer Science at McGill University.He holds his Ph.D in Computer Science from University of Technology in Poland.