Natural Language Engineering **23** (4): 589–616. © Cambridge University Press 2017
doi:10.1017/S1351324916000395

*Detecting sexual predators in chats using behavioral features and imbalanced learning**

CLAUDIA CARDEI and TRAIAN REBEDEA

Department of Computer Science, University Politehnica of Bucharest, 060042 Bucharest, Romania
e-mails: claudia.cardei@gmail.com, traian.rebedea@cs.pub.ro

(*Received 23 November 2015; revised 19 December 2016; accepted 20 December 2016; first published online 31 January 2017*)

Abstract

This paper presents a system developed for detecting sexual predators in online chat conversations using a two-stage classification and behavioral features. A sexual predator is defined as a person who tries to obtain sexual favors in a predatory manner, usually with underage people. The proposed approach uses several text categorization methods and empirical behavioral features developed especially for the task at hand. After investigating various approaches for solving the sexual predator identification problem, we have found that a two-stage classifier achieves the best results. In the first stage, we employ a Support Vector Machine classifier to distinguish conversations having suspicious content from safe online discussions. This is useful as most chat conversations in real life do not contain a sexual predator, therefore it can be viewed as a filtering phase that enables the actual detection of predators to be done only for suspicious chats that contain a sexual predator with a very high degree. In the second stage, we detect which of the users in a suspicious discussion is an actual predator using a Random Forest classifier. The system was tested on the corpus provided by the PAN 2012 workshop organizers and the results are encouraging because, as far as we know, our solution outperforms all previous approaches developed for solving this task.

1 Introduction

In a world where almost everything one can think of can be found online and where digital communication has become the most used form of person-to-person communication, more and more people use chat (instant messaging) programs in order to discuss and share information with other people worldwide. Many chat systems allow a higher level of anonymity as compared to real-life communication and interactions. This has been shown to be useful in some situations, as people tend to disclose more information in chat rooms, where they perceive themselves as

*This work has been partially funded by the Sectorial Operational Programme Human Resources Development 2007–2013 of the Romanian Ministry of European Funds through the Financial Agreement POSDRU/159/1.5/S/132397. Moreover, Claudia Cardei would like to thank Google for the Anita Borg scholarship granted in 2014 which partly funded this work on sexual predator identification in online conversations.

https://doi.org/10.1017/S1351324916000395 Published online by Cambridge University Press

more anonymous, than they would do in a normal face-to-face relationship (Whitty 2002). However, not all people have the best intentions when using chat systems for interacting with peers and the increased level of anonymity can favor their inappropriate behavior. This can lead to situations where one participant in a chat may try to exploit their chat partner in order to obtain sexual favors. As children and teens are among the most active online users, especially on social networks, instant messaging, and discussion forums, they can easily become victims of such behavior. We define a sexual predator as a person who tries to obtain sexual favors in a predatory manner (by committing sexual crimes or by trying to abuse people that are underage, usually under 18 years old). Online sexual predators have very much to do with online child sex exploitation.

Predatory acts and other offences related to incorrect sexual behavior directed toward underage people are significant problems in day-to-day life. Recent studies inform that one in twelve children experienced a sexual victimization, including thirty-two per 1,000 who experienced a sexual assault, and twenty-two per 1,000 who experienced a completed or attempted rape (Finkelhor *et al.* 2005). Although the situation is different for most Internet-initiated sex acts – where the cases of violence are rarer, and the teens usually agree to meet the offender more than once (or are even in love with the offender) – the problem is by all matters serious (Wolak, Finkelhor and Mitchell 2004). For example, Wolak *et al.* (2010) conclude that the total number of online sex crimes committed in the United States in 2000 represent approximately seven per cent of the total number of arrests for statutory rape that were reported during the same period. However, we consider that this number is even larger now, as the usage of online chat is more widespread nowadays, both on computers and on mobile devices that support chat technologies. Moreover, other studies show that over sixty per cent of the first online encounters between predators and victims happened in chat rooms (Wolak *et al.* 2004; Mitchell, Finkelhor and Wolak 2007).

1.1 Problem overview

Since most of sex crimes initiated online start from a chat conversation, it is important to develop a system that automatically detects sexual predator behavior in online chats. This system can be easily used to monitor all the different chat programs available on a specific device and to alert an adult every time it recognizes such abnormal behavior in a conversation of an underage person.

The problem of identifying a predator in an online chat may be divided even further into two separate tasks, each focusing on a different level of analysis:

- Identifying if a chat conversation contains sexual misbehavior by indicating the participant in the chat who is a potential predator;
- Identifying the exact messages of the predator; these are the most distinctive utterances of the predator's misbehavior in each specific conversation.

The system presented in this paper aims to solve only the first task, thus the output will be a specific participant in a chat conversation that has shown a potential

https://doi.org/10.1017/S1351324916000395 Published online by Cambridge University Press

predatory behavior. We consider that such a system would suffice for any user, either a parent or a local authority such as a police officer, especially if the provided accuracy is high, accompanied by a very low false positive rate. This would leave up to the user to solve the second task by manually examining the conversation in order to spot the utterances that are more likely to denote a predatory conduct, as this second task is much more difficult. Thus, we consider that the current technologies in natural language processing are not yet mature enough to solve this second task with a good accuracy.

As the sexual predator identification problem can be clearly treated as a supervised machine learning task, we need labeled data to train one or more text classifiers. To this extent, the corpus provided by the PAN 2012 competition (Inches and Crestani 2012) is the largest dataset used for detecting sexual predators in online conversations. It contains a collection of conversations crawled from different chat repositories and Internet Relay Chat (IRC) channels which has been split into training and test files, which have also been used for the competition. A list with the sexual predator users is provided for both the training and the test datasets.

As the chats in the dataset aim to approximate reality as close as possible, less than four per cent of the conversations in the corpus contain a potential sexual predator, as the percentage of predators among normal chat users is very low. This may be a huge problem, since the dataset is imbalanced and most of the classifiers require a relatively balanced training (and test) corpus for achieving a good classification accuracy. Nevertheless, there are several methods that have been successfully applied when dealing with imbalanced datasets in machine learning (He and Garcia 2009). Out of the three widely used methods for classifying imbalanced data listed below, we have used the first two techniques as we will show in the following parts of the paper:

- Sampling Methods;
- Cost-Sensitive Methods;
- Kernel-Based Methods and Active Learning Methods.

After running several experiments to identify the best solution to solve this problem, we are proposing a two-stage classifier: the former detects if a conversation contains a potential offender, while the latter identifies the actual sexual predator from the list of participants to a given chat. We use standard text categorization methods based on the Bag of Words (BoW) model, together with feature selection and with some empirical features that were designed to quantify the behavior of a sexual predator in a chat as opposed to a normal user. Among these empirical features, we can highlight the following: the number of questions asked by a participant to a chat, the percentage of slang words and of words with sexual connotations within the vocabulary of a user, or the age of the user if it can be detected in the conversation. Alongside the behavioral features (BF), we have also used a BoW model and selected the most relevant words within the training chat corpus by employing feature selection. The most relevant words are used as additional features to improve the detection of whether a conversation contains inappropriate behavior or whether a user is a potential predator.

https://doi.org/10.1017/S1351324916000395 Published online by Cambridge University Press

The paper shows that both feature selection employed on the BoW and the BF are essential steps in building a powerful method for the identification of sexual predators in chat conversations. By combining these two types of features, we have obtained a classifier that has an overall performance of 0.957 on the test corpus, which is, to our knowledge, the best result obtained on the PAN 2012 predator dataset.

In should be taken into consideration that this result may also be attributed to the way the sexual predator dataset from PAN 2012 has been created. As explained in the next section, the corpus contains conversations of predators and their victims mixed with chats of adults engaged in sexual discussions and even cybersex and IRC discussions on various other topics. Although the dataset has a distinct training and test set, the chats are drawn from the same data sources and have a similar distribution of topics.

The outline of this paper is as follows. In Section 2, we provide a general description for sexual predator identification in online chats, focusing on the PAN 2012 competition and dataset, along with the most successful solutions and results. However, we also highlight other important works related to predator detection in chat conversations, as well as several theoretical elements that are used in the proposed solution. Section 3 presents the system that has been designed for solving the predator identification problem, together with alternatives that were considered. Then, in Section 4, are provided details about the various experiments conducted in order to identify the best solution, together with the results of these experiments and a comparison between the proposed alternatives. Section 5 concludes the paper by highlighting the most important contributions of the presented solution and by discussing the future work.

2 Related work

The following subsections present the most important results and solutions developed for predator detection in the recent years. In the first part, the focus is on the PAN 2012 competition which proposed the sexual predator identification task and offered a corpus for training, as well as on the best solutions developed by the participants to PAN. In the second part of this section, we present other methods and results from researchers that did not take part into the PAN competition, together with more details on the chat corpus used in developing and testing the proposed solution.

2.1 Sexual predator identification at PAN 2012

PAN (Plagiarism Analysis, Authorship Identification, and Near-Duplicate Detection – http://pan.webis.de/) is a lab on uncovering plagiarism, authorship, and social software misuse, organized as part of CLEF (Conference and Labs of the Evaluation Forum – http://clef2012.clef-initiative.eu/). In 2012, PAN proposed a distinct sub-task under the Author Identification topic, focused on Sexual Predator Identification in chat conversations (Inches and Crestani 2012).

https://doi.org/10.1017/S1351324916000395 Published online by Cambridge University Press

For this task, the organizers collected and provided two collections of chats: one for training (which also contained the list of expected sexual predators) and another one used for evaluating the submitted solutions. The conversations dataset consists of an XML file with all the chat discussions, each one having a set of messages representing the utterances of the various participants to each chat. All messages contain the following information: the ID of the user, the timestamp, and the actual text.

In order to build the whole collection (both for training and evaluation), the organizers collated chats from different online sources:

- Perverted Justice website[1] where one can find conversations between convicted sexual predators and volunteers posing as underage teenagers;
- Omegle repository[2] which contains anonymous online conversation between consenting adults who use abusive language or even engage in cybersex;
- IRC channels[3,4] which includes logs of IRC discussions on various topics.

The workshop organizers made available about thirty per cent of this collection as a training set, consisting of 66,927 conversations which contained over 90,000 unique users. They also provided the list of the 142 sexual predators which could be found within these conversations. The other seventy per cent of the corpus was used to assess the submitted solutions. For the test dataset, only the conversations were provided and each submitted solution provided a list of the sexual predators identified in these discussions.

In order to measure the performance of the submitted solutions, standard metrics for text classification (Manning, Raghavan and Schütze 2008) were used: precision (P), recall (R), and F-score. However, the F-score with precision and recall being equally weighted – called the F1 measure – is not always the best performance indicator. The same is true for sexual predator detection, as it is more important for the system to have a very high confidence that the identified participants really are sexual offenders than to aim to detect all the sexual predators. Therefore, for assessing the performance of the submitted solutions, the organizers proposed to favor precision over recall, thus using an F-score with the weight set to 0.5:

$$F = \frac{(\beta^2 + 1)PR}{\beta^2 P + R}, \text{where } \beta = 0.5 \tag{1}$$

Although it is important to retrieve as many suspects as possible (recall), the motivation behind the emphasis on precision is to optimize the time of the police officers who will receive the output of the program and will have to review the list of possible suspects.

Sixteen submissions were made within this competition at PAN 2012 and the best solution achieved a $F_{0.5}$-score of 0.9346, developed by Villatoro-Tello *et al.* (2012).

[1] http://www. perverted-justice.com/
[2] http://omegle.inportb.com/
[3] http://www.irclog.org/
[4] http://krijnhoetmer.nl/irc-logs/

https://doi.org/10.1017/S1351324916000395 Published online by Cambridge University Press

Table 1. *Top ten results for the sexual predator competition at PAN 2012*

$F_{0.5}$-score	Participant
0.9346	Villatoro-Tello *et al.* (2012)
0.9168	Tim Snider
0.8691	Parapar, Losada and Barreiro (2012)
0.8652	Morris and Hirst (2012)
0.8638	Eriksson and Karlgren (2012)
0.8137	Peersman *et al.* (2012)
0.7316	Popescu and Grozea (2012)
0.7060	Rachel Sitarz
0.5537	Vartapetiance and Gillam (2012)
0.3946	Kontostathis *et al.* (2012)

Table 1 contains the results of the top ten solutions which were submitted for this competition.

We should take into consideration that the rather high scores obtained on the PAN sexual predator dataset might be explained by the fact that all predator conversations have been extracted from the same source, which contains the chat logs of predators which have been identified and prosecuted as sexual offenders. There might be other chat conversations between predators and their victims which are more complex or where the predator has a more complex behavior which have not been added to the Perverted Justice website, mainly because the offenders have not been convicted. As the work presented in this paper uses the PAN predator dataset, the same considerations apply to our results as well.

2.2 Previous results in sexual predators identification

The winning solution at PAN 2012 was developed by Villatoro-Tello *et al.* (2012), who proposed a two-step approach for identifying the sexual predators. Thus, they employed two different classifiers, one used to detect suspicious conversations (Suspicious Conversations Identification – SCI classification) and the second one for predator disclosure (Victim From Predator – VFP classification). Before training the classifiers, they filtered the chats by removing conversations with only one participant, conversations with less than six messages per user and conversations with long sequences of unrecognized characters. For representation, they used a BoW model to determine the features for both classifiers. The actual features were computed using term frequency-inverse document frequency weighting. For the first classifier, the feature vector represented all the text within a conversation, while for the second classifier it consisted of all the messages for each user. They used Multi-Layer Perceptrons with ten neurons on the hidden layer for both classifiers.

Later, the same authors proposed a different approach which makes use of a three-stage chain classifier (Escalante *et al.* 2013). Each model of this chain classifier is trained for a different stage related to how a sexual predator approaches the victim: gaining access, deceptive relationship, and sexual affair/proposal. Each stage

https://doi.org/10.1017/S1351324916000395 Published online by Cambridge University Press

was represented by an equally length fragment extracted from the conversation. However, this solution obtained a lower performance than the approach proposed during the competition.

Peersman *et al.* (2012) used a similar approach with two classification models: one for separating posts by predators from posts by other participants, and another one for excluding the users without any suspicious behavior, both trained using Support Vector Machines (SVMs). They also incorporated a filtering module to increase the precision of the first classifier. They relied on the probability outputs of the SVM classifier and averaged the outputs of the ten most suspicious posts of each user in order to determine whether an user is suspicious or not. Then the outputs of both classifiers were combined using a linear combination.

Parapar *et al.* (2012) developed a classifier that distinguishes predators from non-predators by concatenating all the messages that a specific person wrote in all conversations from the corpus, resulting in a document per user. They removed words with more than twenty characters and the ones that appear in less than ten documents. After that, the remaining words were weighted using a frequency-inverse document frequency scheme. They also considered word bigrams and trigrams during their experiments to find their role in the overall performance. Moreover, they incorporated Linguistic Inquiry and Word Count (Pennebaker, Francis and Booth 2001) features that estimate the degree to which people use different categories of words, and also chat-based features (the number of subjects contacted by a given user, the percentage of conversations initiated by a given user, etc.). Then they trained a SVM classifier using all these features.

Morris and Hirst (2012) used both lexical features and some BF in their solution. For lexical features, they computed a vector of messages for each user along with the response of the interlocutors concatenated by the special prefix/tag OTHER_. They also transformed some of the other lexical features into special tags such as SMILEY_, MALE_name, FEMALE_name, NUM for numbers, and PHONE_number for telephone numbers. They considered several BF metrics such as initiative of a user, attentiveness, and conversation dominance. Using these features, they trained a SVM classifier for this task. However, after that, they applied a filter consisting of another SVM with the task of distinguishing predators from victims.

Other approaches include the ones of Eriksson and Karlgren (2012) which have used a Maximum-Entropy classifier with the profile of each user and Popescu and Grozea (2012) who combined k-Nearest Neighbors and Random Forest based on string kernels with character 5-gram presence bits. Kontostathis *et al.* (2012) developed ChatCoder 2.0 which uses a document per user approach. They used features such as the number of pronouns in a line, number of personal nouns, number of relationship nouns, or number of communicative desensitization words (such as 'kiss', 'bra', 'horny', 'sex') to train a C4.5 decision tree classifier.

Besides the PAN 2012 sexual predator identification competition, there have also been several other approaches worth mentioning. Bogdanova, Rosso and Solorio (2012b) propose a classifier that uses features based primarily on sentiments, such as emotional markers (joy, sadness, anger, surprise, disgust, or fear words) and on the offender's psychological profile. They observed that predators are not willing to step

https://doi.org/10.1017/S1351324916000395 Published online by Cambridge University Press

aside from the sex-related conversation, concluding that predators have a somehow fixated discourse or predominant topic. They suggest to use lexical chains in order to model and identify this feature in the conversations (Bogdanova, Rosso and Solorio 2012a). Considering these features, they employed a Naive Bayes classifier trained on chat logs from the Perverted Justice site, cybersex chat logs and the NPS chat corpus[5] and report an accuracy of ninety-four per cent for distinguishing Perverted Justice conversations from cybersex chat logs and ninety-two per cent in separating Perverted Justice from NPS chats.

Vartapetiance and Gillam (2012) used an approach based on deception detection. They developed a set of features which best characterize the predators based on different cues/topics discovered by manually inspecting a part of the corpus: address (asking the children about the address of the house or of some place to meet), parents (asking the children about the presence or absence of their parents while chatting, making sure the chat logs will be later deleted, and the children will not discuss with their parents), age of the predator, and the intentions of the predator. Their main algorithm analyzes each chat and detects lexical patters for each of the aforementioned topics. In a subsequent research, the authors improved their categories by including a wider list of word patterns and by filtering conversations that were discussing computer topics (Vartapetiance and Gillam 2014). Moreover, they investigated the use of machine learning algorithms (decision trees, Naive Bayes, and SVM) for selecting the best attributes. The main conclusion is that SVMs proved the best solution for these features and the results obtained using a SVM would place them on the fourth place in the PAN competition.

Our solution has been inspired by the various approaches developed for sexual predator detection. Similar to Villatoro-Tello *et al.* (2012), we employ a two-stage classification strategy in order to solve the two main problems of detecting a sexual predator: identifying suspicious discussions and separating predator from victim in these conversations. However, we go beyond their solution by integrating behavioral features that prove very powerful in differentiating predators from victims, in a similar way to Vartapetiance and Gillam (2012) or Morris and Hirst (2012). Our BF aim to detect more complex issues about the phrase structure of predators (such as textual complexity differences) or about the intentional elements expressed in the dialogue (e.g. invitations, guiding the discussion, questions). Moreover, we have treated the problem of the imbalanced dataset more thorough than any previous work. All these improvements allowed us to obtain the best result known so far in online sexual predator identification.

2.3 Learning from imbalanced datasets

He and Garcia (2009) present in a comprehensive study the state-of-the-art methods for dealing with imbalanced datasets for learning algorithms. They motivate their review by the increasing volume of datasets with under-represented classes, which

[5] http://faculty.nps.edu/cmartell/NPSChat.htm

https://doi.org/10.1017/S1351324916000395 Published online by Cambridge University Press

prove to be a challenge for the existing methods used for supervised learning. The study provides three main methods to deal with imbalanced datasets:

- Sampling Methods;
- Cost-Sensitive Methods;
- Kernel-Based Methods and Active Learning Methods.

According to He and Garcia (2009), sampling methods for handling imbalanced datasets can be divided into several categories. In the proposed solution, a cost-sensitive solution has been employed to handle the imbalance in the sexual predator dataset. However, several sampling methods have also been tested, including random undersampling and oversampling, informed undersampling using BalanceCascade (Liu, Wu and Zhou 2009), and CBO – a clustering-based method using k-means (Jo and Japkowicz 2004).

Cost-sensitive methods are based on a cost matrix which defines the penalty of misclassifying an instance. The purpose of the cost matrix, $C(i, j)$, is to help finding a hypothesis that minimizes the overall cost on the training data set (Bayes conditional risk):

$$R(i|x) = \sum_j P(j|x)C(i, j) \tag{2}$$

$P(j|x)$ is the probability of each class j for an instance x, while $C(i, j)$ is the cost or penalty for predicting that an instance belongs to class i when it actually belongs to class j. Several specific cost-sensitive methods have been developed for the most important classifiers: for AdaBoost (Sun *et al.* 2007), for decision trees (Maloof 2003), for neural networks (Kukar *et al.* 1998), etc. Moreover, Domingos (1999) proposed MetaCost, a general method that allows any classifier to be cost-sensitive by wrapping a cost-minimizing procedure around the classifier.

2.4 Corpus description and brief analysis

As already mentioned in the general description of the PAN 2012 sexual predator identification competition, the training collection has approximately 67,000 chats with more than 90,000 users. Among these conversations, only around 2,000 have a sexual predator within the list of participants. Moreover, about sixty-eight per cent of the chats are between two users and the rest are multi-party discussions. After a brief analysis of the subset of the corpus containing only the chats between two users, an initial observation was that many conversations had very few lines as can be seen in Figure 1. These very brief conversations can mislead the training of the classifiers, since no meaningful information can be extracted from them even by employing manual human inspection. As a result of this preliminary investigation, we decided to use for training a subset of the corpus consisting of conversations with exactly two users and more than twenty messages.

Table 2 contains additional details about the outcome of this filtering phase: The resulted corpus has a little over 8,000 conversations out of which about ten per cent of the chats contain a predator. We considered that this is a simple method for removing some of the outliers in the train dataset without losing useful information when training the classifiers. It is also important to notice that no sexual predator

https://doi.org/10.1017/S1351324916000395 Published online by Cambridge University Press

Table 2. *Size of the corpus after filtering*

Filter	# conversations	# conversations with potential predator
No filter	66,927	2,016
+ Exactly two users	45,741	1,088
+ More than twenty messages	8,069	817

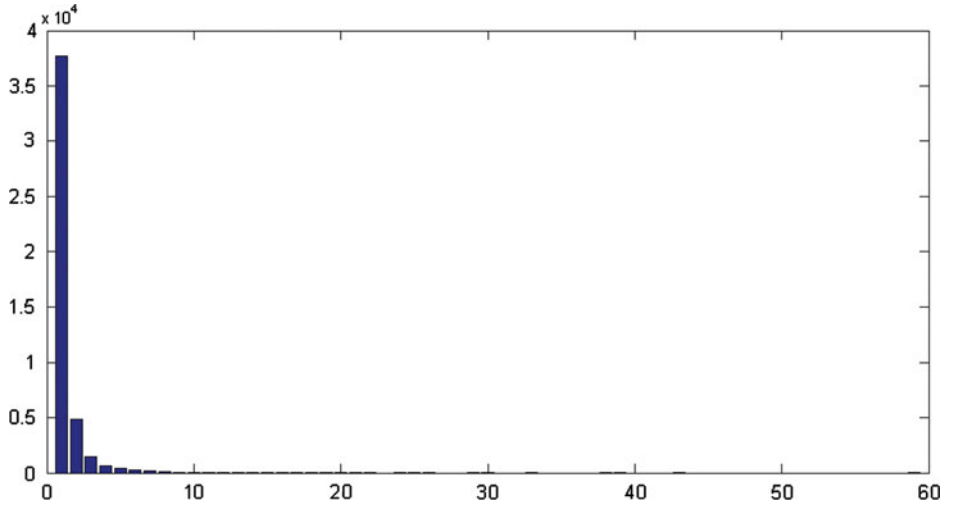


Fig. 1. (Colour online) Histogram for the number of messages in conversations between two users.

has been lost in this filtering phase as all predators have also been involved in at least one longer conversation that has not been removed. An interesting side-effect of this phase is that the filtered corpus is still imbalanced; however, the ratio of chats with predators versus chats without predatory content has increased several times.

Since the corpus was created by merging chats from different websites and IRC channels, the conversations cover a diversity of topics. For the interested reader, Appendix A contains selected representative logs for each of the four types of conversations present in the analyzed dataset: one chat from the Perverted Justice website, one from the Omeagle repository containing a cybersex sex between consenting adults, and two conversations from an IRC channel – one covering general topics and one on a more technical topic.

Two distinct representations of the analyzed corpus were used. The first representation helps to classify at the conversation level and is helpful for determining whether a conversation contains a potential predator. In this case, each feature vector represents an entire chat conversation. The second one, which is used to determine whether a user is a predator, classifies at the participant level and each feature vector represents an unique participant. The training corpus used for the conversation-level representation is the one containing only chats with only two users and more than twenty messages. However, the corpus used for participant-

https://doi.org/10.1017/S1351324916000395 Published online by Cambridge University Press

Table 3. *Distribution of classes for each type of chat representation*

Type of representation	Corpus size	Predatory class	Non-predatory class
Conversation level	8,069	817	7,252
Participant level	97,689	142	97,547

level representation contains, for each user, the list of all utterances that were authored by him in all conversations, regardless of the number of participants and messages. Table 3 presents the size of the training corpus and the distribution of the classes for each type of representation. From the data in this table, one can easily see that solving the problem of predator identification at the participant level is more difficult due to the highly imbalanced dataset. This is the main reason for employing a two-stage classification strategy.

3 System description

This section is divided into two main subsections. The first subsection presents the most important methods and algorithms used for developing the system: feature selection methods that help reducing the dimensionality of the feature space and the classifiers used to distinguish predators from regular users. The second subsection describes in detail the proposed solution, including the features and the classification pipeline.

3.1 Methods and algorithms

3.1.1 Feature selection for text classification

Feature selection is seldom used for high-dimensionality machine learning tasks, especially in text classification and information retrieval. In text classification, when using the BoW model, feature selection extracts a subset of the vocabulary and further uses only these terms as features to train the classifier. It has two main advantages: it reduces the training time of the classifier by considering a substantially smaller feature vector, and it also improves the accuracy of the classifier by removing noisy features.

There are several methods used for feature selection in text classification: mutual information (MI), chi-squared, or frequency-based feature selection (Manning *et al.* 2008) are among the most widely used. For the proposed solution, we have selected MI which measures how much information the presence/absence of a term contributes to making the correct classification decision on each class c. Moreover, MI expresses the relative entropy between the joint distribution and the product distribution of two random variables (Cover and Thomas 2012).

$$I(U;C) = \sum_{e_t \in \{1,0\}} \sum_{e_c \in \{1,0\}} P(U = e_t, C = e_c) \log \frac{P(U = e_t, C = e_c)}{P(U = e_t)P(C = e_c)} \tag{3}$$

where U is a random variable quantifying the presence of a term in a document and it takes values $e_t = 1$ (the document contains term t) and $e_t = 0$ (the document does

https://doi.org/10.1017/S1351324916000395 Published online by Cambridge University Press

not contain t), and C is a random variable that takes values $e_c = 1$ (the document is in class c) and $e_c = 0$ (the document is not in class c).

The results section contains additional details about the usage of MI for feature selection and provides a discussion on the optimum number of features for detecting sexual predators on the analyzed corpus.

3.1.2 Description of employed classifiers

SVMs have been very popular for many text classification tasks, mainly due to the fact that they perform well for high-dimensional feature spaces and also allow the usage of different kernels, such as string kernels. Moreover, Sequential Minimal Optimization (SMO) (Platt 1998) is a method for training SVMs in significant less time which has permitted training SVMs for larger corpora.

Another widely used category of text classifiers are based on decision trees. Among them, Random Forest (Breiman 2001) grow a collection of unpruned decision trees by selecting a random subset of the features at each node. Afterwards, an instance is classified by traversing each tree and collecting their classification, the final decision being the majority vote of all the classifiers. The algorithm uses bagging in order to sample with replacement a training subset for each tree in the forest. When growing the tree, at each node, the variable on which the split is done is the best from m randomly selected variables ($m \ll M$, where M is the total number of features). The random selection of the features at each step is preferred to the classical decision tree training algorithm in order to break the correlation between the trees in the forest. Moreover, no pruning is used. Most often, m is recommended to be equal or close to $\sqrt{M}$.

An important advantage of Random Forest over SVMs is the training speed, which makes them more popular for larger datasets. For the sexual predator detection task, both methods have offered comparable results therefore we decided to test both of them in the proposed solution.

As sexual predator chats are quite infrequent, MetaCost (Domingos 1999) has been employed to handle the imbalanced dataset. Its main advantage is that it allows any classifier to be cost-sensitive by wrapping a cost-minimizing procedure around the classifier. MetaCost uses a cost matrix to define different weights for misclassifying an item.

3.2 Proposed solution

3.2.1 Features for sexual predator detection

Since we considered two distinct representations of the chats (conversation level and participant level), we developed a different set of features for each representation. The features used for both representations can be divided into two categories: 'lexical' features that are extracted from the raw text using the BoW model, and 'behavioral' features that are engineered to reflect how a user acts in the chat by taking into consideration the task at hand.

https://doi.org/10.1017/S1351324916000395 Published online by Cambridge University Press

Table 4. *Top fifteen most relevant lexical features for detecting sexual predator ordered descending using MI*

Conversation-level representation	Participant-level representation
k	call
back	home
gonna	leave
kewl	house
2	back
mom	phone
home	talk
brb	tomorrow
gotta	mom
call	time
doin	ready
nite	shower
lol	long
tomorrow	tonight
ya	meet
dont	night

Table 5. *Vocabulary size before and after feature selection*

Type of representation	Before MI	After MI
Conversation level	4,246	1,000
Participant level	12,868	50

To build the set of lexical features, we first applied a pre-processing step for the whole corpus. This included tokenization, stop-words removal, and also discarding long words with over twenty letters. For the remaining words, a BoW model was built and, after preliminary experiments, we have opted for binary features instead of frequency-inverse document frequency. Then the most discriminative features were selected by employing MI feature selection. Table 4 shows the top fifteen most relevant lexical features for detecting sexual predators within the training corpus, ordered descending by MI score. Using this procedure, the vocabulary size has been reduced considerably both for the conversation-level representation and for the participant-level one, as can be seen in Table 5. The experiments for determining the optimum number of features selected using MI are described in the next section.

As for the behavioral/interactional attributes, we started from the premise that there is a need for higher level linguistic features that are able to capture discourse, semantic, and pragmatic elements which can easily separate predatory discussions from non-predatory ones, as well as predators from victims. Part of the BF described next were introduced after manual inspection of the conversations, being also inspired from research in textual complexity and the analysis of chat logs in educational settings.

https://doi.org/10.1017/S1351324916000395 Published online by Cambridge University Press

Although some of these features are empirical or defined according to the task at hand (e.g. underage or above age existence, usage of slang), many are in accordance with research in the behavior of online predators when attracting their victims using chat conversations (Malesky 2007; Whittle *et al.* 2013) such as usage of sexual expressions, initiating conversations, trying to be a mentor which may manifest in linguistic differences which may be measured using a readability index or other textual complexity measures. Our choice of behavioral/interactional features is based on current research in the behavior of sexual predators in online conversations and from an empirical analysis of the PAN chat corpus. However, other interactional features might have been used such as dialog acts (Core and Allen 1997) detected in the chat utterances or even sequences of relevant dialog acts.

Next, we present the full list of behavior and interactional features used for both levels of representation.

- *Question ratio:* Percentage of questions asked from the total number of messages. This number was computed heuristically by counting question marks, patters such as 'is it' where the verbs appear before nouns/pronouns, and lines containing one of the *Five Ws and H* followed by an auxiliary verb and the 'you' pronoun.
- *Underage expressions ratio:* Percentage of expressions that may reflect that the user is underage from the total number of messages. We counted expressions such as 'don't know', 'didn't do', 'never did', etc.
- *Above age existence:* Existence of a message containing the answer to an ASL[6] request and the answer should be a number above 18.
- *Underage existence:* As above, but the age should be between 10 and 17.
- *WordNet*[7] *words ratio:* Percentage of the words that are present in WordNet from the total number of words. In this case, WordNet is used as a dictionary to separate the formal vocabulary from slang and other informal vocabulary.
- *Slang words ratio:* Percentage of slang words from the total number of words. For this, a list of the most common English slang words has been crawled from the Online Slang Dictionary[8].
- *Sexual words ratio:* Percentage of words with sexual connotation from the total number of words. The list of words with sexual connotation was also obtained from the Online Slang Dictionary.
- *Non-words ratio:* Percentage of words that do not exist in either WordNet or the Online Slang Dictionary. The motivation behind this feature is to count the ratio of misspelled words.
- *Negations ratio:* Percentage of messages that contain a negation from the total number of messages. This feature was added because, in many cases, the victim usually refuses the predator's requests.

[6] http://en.wikipedia.org/wiki/Age/sex/location
[7] http://wordnet.princeton.edu
[8] http://onlineslangdictionary.com

https://doi.org/10.1017/S1351324916000395 Published online by Cambridge University Press

- *Flesch Reading Ease Score (FRES)*: Score computed by the formula for the FRES test used to assess the readability of a text (Flesch 1948): $FRES = 206.835 - 1.015\left(\frac{total_words}{total_sentences}\right) - 84.6\left(\frac{total_syllables}{total_words}\right)$

These BF can be used for both types of chat representations. However, for the participant-level representation, the features reflecting the existence of a message containing the age of the participant will be modified to consider all conversations in which an user takes part, so they will become the following:

- *Above age ratio:* Percentage of conversations that contain the answer at an ASL request from the total number of conversations in which the user takes part. The age in this answer is a number above 18.
- *Underage ratio:* As above, but the age is between 10 and 17.

Moreover, for the participant-level representation, two new features have been introduced that have no added value for the conversation-level representation:

- *Initiation ratio:* Percentage of conversations that the user had initiated from the total number of conversations that he/she took part in.
- *Average response time:* The average time that the user responds to the messages sent by the other participants.

This set of behavioral attributes is lacking features that could describe the actual interaction between that user and the other participants in the chat. To tackle this issue, for both representations, we enhanced the feature set by adding the following new features. For the conversation-level representation, we divided each conversation into two documents (since each conversation had only two users after the intial filtering of the corpus), each one containing the messages of a single user. For these two new documents, we computed the BF vector and the new features were calculated as the difference between the feature values of the user who initiated the conversation and the feature values of the other participant. This allowed us to capture information whether the initiator of the chat conversation had a different behavior within the discussion than his/her interlocutor.

For the participant-level representation, this mechanism needed to be slightly changed as each user might have participated in more than a single conversation. Thus, for each participant we extracted all partners that interacted with him/her in the entire corpus and we computed, for each BF, the average value of all his/her partners. These mean scores were the new features quantifying the difference in interaction between a user and all his/her interlocutors.

Table 6 shows the final feature vector size for each representation. We chose 1,000 lexical features for the conversation-level representation and fifty lexical features for the participant-level corpus. As described in the next section, these values were selected as they offered the best results after performing cross-validation on the training set. More details about the performed experiments, along with their results, are presented in the following section.

Table 7 shows the BF sorted by their relevance, whereas Table 8 presents the top fifteen most relevant features for detecting a sexual predator. As can be observed

https://doi.org/10.1017/S1351324916000395 Published online by Cambridge University Press

Table 6. *Final feature vector size*

Set of features	Conversation level	Participant level
Lexical features (BoW)	1,000	50
Behavioral features	10	12
Interaction behavioral features	10	12
TOTAL	1,020	74

Table 7. *Behavioral features sorted by their relevance using MI. Features prefixed with DIF represent the difference between the feature values of the participants, while the ones starting with PRTN are computed for the chat interlocutor/partner*

Conversation-level representation	Participant-level representation
flesch_reading	PRTN_negation
num_expr_underage	PRTN_initiation
slang_words	PRTN_slang
num_questions	initiation
negation	negation
sex_words	num_expr_underage
aboveage	PRTN_sex_words
underage	PRTN_underage
DIF_num_expr_underage	PRTN_wordnet
non_words	num_questions
DIF_negation	PRTN_num_questions
wordnet	non_words
DIF_flesch_reading	avg_rsp_time
DIF_wordnet	PRTN_flesch_reading
DIF_sex_words	flesch_reading
DIF_aboveage	slang_words
DIF_underage	PRTN_avg_rsp_time
DIF_non_words	PRTN_num_expr_underage
DIF_slang_words	wordnet
DIF_num_questions	PRTN_non_words
–	aboveage
–	sex_words
–	underage
–	PRTN_aboveage

from these tables, for the conversation-level representation, the BF are not as powerful as the lexical ones, and the interaction behavioral features (IBF) are actually poor. However, for the participant-level representation, the situation is completely different: Among the most relevant attributes are the behavioral ones, especially the features which reflect the interaction between the user and his/her partners.

A possible explanation for this result is that separating conversations that might be predatory, have some sexual content, or contain other messages that resemble predatory conversations is a simpler task than separating the actual predator users

https://doi.org/10.1017/S1351324916000395 Published online by Cambridge University Press

Table 8. *Top fifteen most relevant features for detecting a sexual predator using MI*

Conversation-level representation	Participant-level representation
k	call
back	PRTN_NEGATION
kewl	PRTN_INITIATION
2	home
mom	PRTN_SLANG_WORDS
home	leave
brb	INITIATION
gotta	house
call	NEGATION
nite	back
FLESCH_READING	NUM_EXPR_UNDERAGE
lol	phone
doin	talk
asl	time
m	tomorrow

Table 9. *Cost matrix for the suspicious conversation classifier*

	Actual non-predators	Actual predators
Expected non-predators	0	1
Expected predators	7	0

from other participants in chats between two consenting adults. Thus, lexical features are very discriminative for this task, while behavioral and interactional features are more relevant for separating predator users from suspicious conversations from non-predators.

3.2.2 Chained classification

For an improved detection of sexual predators, the problem was split into two different tasks and a chain of two classifiers was employed to solve them. The first task consisted of identifying suspicious conversations and for this task the classifier was trained on the conversation-level representation and corpus. As mentioned at the beginning of this section, in order to handle the imbalanced dataset, we used a MetaCost classifier with the cost matrix represented in Table 9. From this matrix, one can easily notice that we penalized false negatives seven times more than false positives. This way we are increasing the recall for this task at the cost of a slightly lower precision, thus ensuring that most of the suspicious conversations containing a predator are handled to the second classification step. As will be discussed in the next section, the classifier wrapped inside the MetaCost function is a SVM trained using SMO.

https://doi.org/10.1017/S1351324916000395 Published online by Cambridge University Press

Table 10. *Distribution of classes for the second classifier with/without SCI*

	Corpus size	Predatory class	Non-predatory class
Without SCI	97,689	142	97,547
With SCI	381	136	245

Table 11. *Cost matrix for the second classifier used for predator detection*

	Actual non-predators	Actual predators
Expected non-predators	0	1
Expected predators	2	0

In fact, the main role of the first classifier is to filter out as many non-predator users as possible by removing all conversations with non-predatory content. This filtering performed by the suspicious conversation identifier (SCI) is thus useful to reduce the data imbalance for the second classifier trained for predator detection. In Table 10 is highlighted the effect of this filtering stage on the dataset used for training the predator identification classifier. Without SCI, the participant-level corpus has over 90,000 users out of which only 142 are predators, while with SCI the dataset will contain only 381 users and a balanced distribution of the two classes. As a conclusion, the second classifier is trained on the participant-level corpus from which the users that did not participate in any suspicious conversations (as identified by the SCI) were removed, thus leading to a much smaller dataset.

The chain classifier acts on two different types of data. In the first stage, SCI works on conversations to filter out as many chats which are considered safe. These filtered out chats do not resemble discussions containing predators in the training set. The second stage acts on all the users from the chats which are considered suspicious to discriminate between actual predators and safe users. Although not tested in the current work, it is highly possible that the SCI phase actually acts as a filter for chats which contain a specific vocabulary, probably including sexual references and invitations to meetings. This is also demonstrated by the fact that lexical features work very well for this task as shown in this section and the subsequent one.

The best solution for the second classifier proved to be a Random Forest classifier wrapped again in a MetaCost procedure with the cost matrix from Table 11. The forest grows ten decision trees by selecting at each node the best values among seven randomly chosen attributes. More details are provided in the next section.

4 Experiments and results

The final solution was implemented using the Weka data mining toolkit (Hall *et al.* 2009) which offers implementations of the most popular machine learning algorithms and provides an easily accessible Java library. The chats were tokenized

https://doi.org/10.1017/S1351324916000395 Published online by Cambridge University Press

using Apache Lucene's[9] linguistic pipeline; although Lucene is a full-featured text search engine, we only needed access to the tokenizers. For interfacing with WordNet, we have opted for the RiTa library[10] that was designed for various natural language processing applications tasks.

The next three subsections provide details on the experiments that have been undertaken in order to reach the final solution, providing comparisons of different classifiers, the role of feature selection on the lexical features, and the influence of the various type of features for the overall performance. In the first subsection, we present the results of an initial experiment which did not include the SCI stage described in the previous section. The next subsection discusses the experiments undertaken to choose the most suitable classifier for the SCI filtering phase. In the final subsection, it is shown how the two classifiers (SCI and predator detection) can be integrated, resulting a much smaller and more balanced dataset for the participant-level classifier trained for predator detection.

4.1 One-stage participant-level classification

The first version of our solution included only the participant-level classifier with the BF described in the previous section. As the training (and test) dataset was extremely imbalanced, we also decided to employ MetaCost for the full participant-level corpus. Using only the BF, the best result was obtained using a Random Forest (ten trees, four features) classifier, and the performance was $F_{0.5} = 0.904$. This result was obtained using a cost matrix for MetaCost of $[0, 1; 5, 0]$. The main problem encountered for these tests was that although the precision was very good, the recall was too low.

To overcome this, we then developed and added the IBF to test if they improve the previous results. As several participants in the full training dataset only participated in one single user conversation, they did not have a partner and the IBF could not be computed. Thus, we decided not to include them in the training, leading to a corpus of 94,784 non-predators and 138 predators. IBF slighlty improved the overall $F_{0.5}$ score to 0.929, but only by increasing the precision and not the recall. A third set of experiments also included lexical features (LEX), besides BF and IBF. As mentioned in the previous section, lexical features are extracted using a BoW model and applying MI for feature selection. However, in this context the lexical features did not improve the results. Table 12 presents all the experiments made on the full-participant corpus using four-fold cross-validation.

4.2 Conversation-level classifier

Using a single classifier did not allow us to achieve both good precision and good recall mainly due to the highly imbalanced corpus. In order to improve the latter metric, we have used the solution proposed by Villatoro-Tello *et al.* (2012). Thus, a

[9] http://lucene.apache.org/core/
[10] http://www.rednoise.org/rita/

https://doi.org/10.1017/S1351324916000395 Published online by Cambridge University Press

Table 12. *Experiments for the full participant-level corpus*

Classifier	Features	Cost matrix	*P*	*R*	$F_{0.5}$
RF (10 t, 4 f)	BF	[0 1; 10 0]	0.929	0.739	0.883
RF (10 t, 4 f)	BF	[0 1; 7 0]	0.954	0.732	0.899
RF (10 t, 4 f)	BF	[0 1; 5 0]	0.971	0.711	0.904
RF (15 t, 4 f)	BF	[0 1; 5 0]	0.947	0.754	0.900
RF (10 t, 5 f)	BF+IBF	[0 1; 7 0]	1	0.659	0.906
RF (10 t, 5 f)	BF+IBF	[0 1; 5 0]	1	0.688	0.916
RF (15 t, 5 f)	**BF+IBF**	**[0 1; 5 0]**	**1**	**0.725**	**0.929**
RF (20 t, 5 f)	BF+IBF	[0 1; 5 0]	1	0.696	0.919
RF (20 t, 7 f)	BF+IBF+50 LEX	[0 1; 5 0]	0.979	0.688	0.902
SMO	BF+IBF+50 LEX	[0 1; 5 0]	0.638	0.703	0.650
RF (20 t, 7 f)	BF+IBF+100 LEX	[0 1; 5 0]	0.900	0.522	0.786
RF (20 t, 10 f)	BF+IBF+100 LEX	[0 1; 5 0]	0.954	0.601	0.853
SMO	BF+IBF+100 LEX	[0 1; 5 0]	0.573	0.681	0.591

two-stage classification approach was employed by adding a classifier that filters out the non-predator users which are considered in the second step – the actual predator detection. This new classifier was trained on the conversation-level corpus and its main role was to remove all conversations with no suspicious behavior/vocabulary, hence removing only non-predatory users. This stage is called suspicious conversation identification and, since the conversation-level corpus is highly imbalanced as well, we used MetaCost for this task as well.

Several experiments were made for determining the best solution for the conversation-level classifier which is used for SCI. We started only with the BF, and then added the IBF. However, the results were very poor using only these set of features (see Table 13). Thus, we decided to also add lexical features (LEX) and this solution significantly improved the performance for both classifiers employed for this task (SMO and Random Forest). Table 13 presents the results of several experiments performed for the SCI stage, including a simple baseline based on the majority. These results were obtained using four-fold cross-validation. It can be easily observed that SVMs trained using SMO outperform the Random Forest classifiers using the same features. Moreover, the BF and IBF always improve the results obtained with the lexical features alone. At last, using MI feature selection improves the results and the maximum performance is obtained for 1,500 lexical features; with more lexical features the performance starts to degrade.

4.3 Final solution: two-chain classifier

Analyzing the results obtained by the two approaches, we decided to merge them and to train a two-chain classifier. The first one is the conversation-level classifier used for SCI, which filters out as many non-predator users as possible without removing predator users. The second one is the participant-level classifier which performs the actual predator detection. In order to build the chain, the second classifier is trained

https://doi.org/10.1017/S1351324916000395 Published online by Cambridge University Press

Table 13. *Experiments for the conversation-level corpus used for SCI*

Classifier	Features	Cost mat.	*P*	*R*	$F_{0.5}$
Baseline/Maj.	–	[0 1; 10 0]	0.101	1	0.123
RF (10 t, 4 f)	BF	[0 1; 7 0]	0.531	0.743	0.563
RF (10 t, 5 f)	BF+IBF	[0 1; 7 0]	0.562	0.709	0.586
RF (15 t, 10 f)	BF+IBF+10 LEX	[0 1; 7 0]	0.913	0.819	0.892
SMO	BF+IBF+100LEX	[0 1; 7 0]	0.877	0.909	0.883
RF (20 t, 25 f)	BF+IBF+600LEX	[0 1; 7 0]	0.953	0.771	0.910
SMO	*BF+IBF+600LEX*	*[0 1; 7 0]*	*0.913*	*0.933*	*0.916*
SMO	600 LEX	[0 1; 7 0]	0.904	0.924	0.907
RF (20 t, 30 f)	BF+IBF+1000LEX	[0 1; 7 0]	0.956	0.776	0.913
SMO	*BF+IBF+1000LEX*	*[0 1; 7 0]*	*0.938*	*0.931*	*0.936*
SMO	1000 LEX	[0 1; 7 0]	0.930	0.938	0.931
SMO	**BF+IBF+1500LEX**	**[0 1; 7 0]**	**0.938**	**0.938**	**0.938**
SMO	1500 LEX	[0 1; 7 0]	0.936	0.938	0.936

Table 14. *Number of predators eliminated by the SCI stage*

ID	Details of the classifier	Number of predators eliminated
[1]	SMO - BF+IBF+600 LEX	6
[2]	SMO - BF+IBF+1000 LEX	6
[3]	SMO - BF+IBF+1500 LEX	7

on a subset of the participant level corpus which contains only the users who took part in at least one of the conversations marked as suspicious by the SCI.

It should be noted that the SCI filtering stage may also eliminate some of the predators, not only non-predatory users. Table 14 shows the number of predators removed by each of the best three classifiers used for the SCI stage and it can be easily seen that the results are very good.

For building the reduced dataset used for training the participant-level classifier, we kept the three best performing conversation-level classifiers trained using four-fold cross validation as described in the previous subsection. Then, we evaluated the whole chat corpus using each of them and retained only the conversations classified as suspicious. Next, only the users taking part in those conversations were selected to train the second classifier. Using this substantially smaller dataset (whose characteristics are presented in the previous section), we trained several combinations of for SCI and actual predator detection. Their results obtained using four-fold cross-validation are presented in Table 15 together with a simple baseline classifier which predicts the majority class.

As BF are more relevant for this task than for the SCI filtering phase which is based mostly on lexical features, we have ran a series of ablation tests to measure the performance of each BF. Table 16 contains the results for the top five BF taken into pairs (current user and interlocutor scores), which highlights that sex words are the most relevant, while slang/dictionary words, the number of questions and

https://doi.org/10.1017/S1351324916000395 Published online by Cambridge University Press

Table 15. *Experiments for the filtered participant-level corpus*

Classifier	Features	Cost mat.	P	R	$F_{0.5}$
Baseline/Maj.	–	[0, 1; 2, 0]	0.403	1	0.458
[1]+RF(10t,5f)	BF+IBF	[0, 1; 2, 0]	0.971	1	0.977
[2]+RF(10t,5f)	**BF+IBF**	**[0, 1; 2, 0]**	**0.993**	**1**	**0.994**
[3]+RF(10t,5f)	BF+IBF	[0, 1; 2, 0]	0.985	1	0.987
[1]+RF(10t,7f)	BF+IBF+50LEX	[0, 1; 2, 0]	0.955	0.941	0.952
[2]+RF(10t,7f)	**BF+IBF+50LEX**	**[0, 1; 2, 0]**	**1**	**0.971**	**0.994**
[3]+RF(10t,7f)	BF+IBF+50LEX	[0, 1; 2, 0]	0.978	0.978	0.978
[1]+RF(10t,7f)	BF+IBF+100LEX	[0, 1; 2, 0]	0.961	0.897	0.947
[2]+RF(10t,7f)	BF+IBF+100LEX	[0, 1; 2, 0]	0.964	0.993	0.969
[3]+RF(10t,7f)	BF+IBF+100LEX	[0, 1; 2, 0]	0.969	0.919	0.958

Table 16. *Ablation tests for behavioral features in filtered participant-level corpus*

Feature set	$F_{0.5}$
BF+IBF+50LEX	0.994
−slang	0.981
−num_expr_underage	0.976
−wordnet	0.976
−num_questions	0.972
−sexwords	0.932

underage expressions also play an important role. However, one other important advantage of the BF is that using a very small number of features (20–22), we can achieve the same or better performance than using hundreds of lexical features in less time.

Although for the conversation-level corpus, the classifier taking into consideration the first most relevant 1,500 words achieved the best results, when tested along with the second classifier for actual predator detection the results were poorer than for the other classifiers. Moreover, the filtered participant-level classifier also seems to benefit when adding a small number of lexical features selected using MI. However, the performs degrades quickly if the number of lexical features is increased above 50.

For the final evaluation on the test dataset, we took the best two chain-classifiers with regard to cross-validation performance (see bolded rows in Table 15). The results of their evaluation on the test corpus is presented in Table 17. As it can be seen, the best two classifiers selected using cross-validation also offer very good results on the test dataset and their performance is very similar. Moreover, both results ($F_{0.5} = 0.957$ and $F_{0.5} = 0.955$) outperform all solutions submitted in the PAN 2012 competition, where the best solution achieved $F_{0.5} = 0.934$ (see Table 1). As far as we know, this is the best solution achieved until now for this task, improving the performance of a sexual predator detection system from 0.935 to 0.957. These new

https://doi.org/10.1017/S1351324916000395 Published online by Cambridge University Press

Table 17. *Final evaluation on the test corpus*

First classifier	Second classifier	*P*	*R*	$F_{0.5}$
SMO - BF+IBF+1000 LEX	RF - BF+IBF	0.986	0.850	**0.955**
SMO - BF+IBF+1000 LEX	RF - BF+IBF+50 LEX	1	0.818	**0.957**

results show that an automated solution for predator detection can be deployed in practice and it would assist police officers with very high confidence.

5 Conclusions

This paper presents a solution for identifying sexual predators in chat conversations which makes use of behavioral and interactional features and employs techniques for handling the imbalanced distribution inherent to this task. The developed system is based on a two-chained classification strategy. The former detects if a conversation contains suspicious behavior/vocabulary thus filtering out as many non-predatory participants as possible, while the latter handles the actual separation of predator users from non-predators given a suspicious conversation. For both classifiers, we used three sets of attributes: binary BoW lexical features selected using MI, BF which quantify the actions of a user in a chat, and interactional features which aim at determining a difference in the behavior of the two interlocutors in a conversation.

The first classifier acts as a filtering stage and it detects whether a conversation contains suspicious predatory behavior. The second classifier was then trained only on a subset of the participants – the ones that took part in the conversations marked as suspicious by the first classifier. In our experiments, we compared two classifiers which perform very well for text classification tasks: SVMs trained using the SMO algorithm and Random Forests. After several experiments, a SMO classifier was chosen for the first step, while for the actual predator detection step a Random Forest performed better due to the smaller number of lexical features. Moreover, the results highlighted that the first classifier relies more on lexical features than on behavioral ones, while the second classifier behaves the opposite way: among its most relevant features were several behavior attributes: the percentage of negations for a user, the percentage of initiated conversations, or the percentage of slang words. Moreover, we have showed that the second classifier performs almost the same with or without lexical features, thus the presented behavioral and interactional features have the best discriminative power. Using this approach, we obtained an $F_{0.5}$ score of **0.957** on the PAN 2012 test corpus, which is the best result for sexual predator identification as far as we know.

With regard to future work, there are several ideas that might be explored. One of them involves using features which reflect the sentiment of each chat participant. For this task, we may use SenticNet (Cambria and Hussain 2012) which is an affective resource that assigns semantic and sentic values to concepts. Another interesting experiment would be using topic modeling techniques, such as Latent Dirichlet Allocation (Blei, Ng and Jordan 2003), to detect the prevalent topics from each

https://doi.org/10.1017/S1351324916000395 Published online by Cambridge University Press

conversation. It would be interesting to see if the lexical features used by the first classifier can be replaced by a smaller set of attributes reflecting the topics discussed in each chat.

The proposed approach could be extended to other types of deceptive online conversations, like detecting unsolicited financial scams or messages aimed at obtaining personal details from unsuspecting victims. However, the BF have been designed to detect sexual predators and are not useful for any other type of deceptive behavior. Moreover, in order to detect frauds, scams, or any other type of predatory behavior using a similar approach, it is required both to have a consistent corpus of chat conversations to train the classifiers and also for fraudulent users to engage in complex conversations with their victims.

References

Blei, D. M., Ng, A. Y., and Jordan, M. I. 2003. Latent dirichlet allocation. *The Journal of Machine Learning Research* **3**(Jan): 993–1022.

Bogdanova, D., Rosso, P., and Solorio, T. 2012a. Modelling fixated discourse in chats with cyberpedophiles. In *Proceedings of the Workshop on Computational Approaches to Deception Detection*, Association for Computational Linguistics, Avignon, France, pp. 86–90.

Bogdanova, D., Rosso, P., and Solorio, T. 2012b. On the impact of sentiment and emotion based features in detecting online sexual predators. In *Proceedings of the 3rd Workshop in Computational Approaches to Subjectivity and Sentiment Analysis*, Association for Computational Linguistics, Jeju, Republic of Korea, pp. 110–8.

Breiman, L. 2001. Random forests. *Machine Learning* **45**(1): 5–32.

Cambria, E., and Hussain, A. 2012. *Sentic Computing: Techniques, Tools, and Applications*. Dordrecht: Springer Netherlands.

Core, M. G., and Allen, J. 1997. Coding dialogs with the damsl annotation scheme. In *AAAI Fall Symposium on Communicative Action in Humans and Machines*, Boston, MA, pp. 28–35.

Cover, T. M., and Thomas, J. A. 2012. *Elements of Information Theory*. New York, NY: John Wiley & Sons.

Domingos, P. 1999. Metacost: a general method for making classifiers cost-sensitive. In *Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, pp. 155–64.

Eriksson, G., and Karlgren, J. 2012. Features for modelling characteristics of conversations. In *Proceedings of CLEF 2012 (Online Working Notes/Labs/Workshop)*, CEUR-WS, Rome, Italy.

Escalante, H. J., Erro, L., Villesanor, E. Villatoro-Tello, A. Juá rez, and Montes-y Gómez, M. 2013. Sexual predator detection in chats with chained classifiers. In *Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, Association for Computational Linguistics, pp. 46–54, Atlanta, Georgia.

Finkelhor, D., Ormrod, R., Turner, H., and Hamby, S. L. 2005. The victimization of children and youth: a comprehensive, national survey. *Child Maltreatment* **10**(1): 5–25.

Flesch, R. 1948. A new readability yardstick. *Journal of Applied Psychology* **32**(3): 221.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. 2009. The weka data mining software: an update. *ACM SIGKDD Explorations Newsletter* **11**(1): 10–8.

He, H., and Garcia, E. A. 2009. Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering* **21**(9): 1263–84.

https://doi.org/10.1017/S1351324916000395 Published online by Cambridge University Press

Inches, G., and Crestani, F. 2012. Overview of the international sexual predator identification competition at PAN-2012. In *Proceedings of CLEF 2012 (Online Working Notes/Labs/Workshop)*, CEUR-WS, Rome, Italy.

Jo, T., and Japkowicz, N. 2004. Class imbalances versus small disjuncts. *ACM SIGKDD Explorations Newsletter* **6**(1): 40–9.

Kontostathis, A., Garron, A., Reynolds, K., West, W., and Edwards, L. 2012. Identifying predators using ChatCoder 2.0. In *Proceedings of CLEF 2012 (Online Working Notes/Labs/Workshop)*, CEUR-WS, Rome, Italy.

Kukar, M., and Kononenko, I. 1998. Cost-sensitive learning with neural networks. In *Proceedings of 13th European Conference on Artificial Intelligence*, Brighton, UK, pp. 445–9.

Liu, X.-Y., Wu, J., and Zhou, Z.-H. 2009. Exploratory undersampling for class-imbalance learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics* **39**(2): 539–50.

Malesky, L. A. 2007. Predatory online behavior: modus operandi of convicted sex offenders in identifying potential victims and contacting minors over the internet. *Journal of Child Sexual Abuse* **16**(2): 23–32. PMID: 17895230.

Maloof, M. A. 2003. Learning when data sets are imbalanced and when costs are unequal and unknown. In *Proceedings of ICML-2003 Workshop on Learning from Imbalanced Data Sets II*, vol. 2, Washington, DC.

Manning, C. D., Raghavan, P., and Schütze, H. 2008. *Introduction to Information Retrieval*. Cambridge: Cambridge University Press.

Mitchell, K. J., Finkelhor, D., and Wolak, J. 2007. Youth internet users at risk for the most serious online sexual solicitations. *American Journal of Preventive Medicine* **32**(6): 532–7.

Morris, C., and Hirst, G. 2012. Identifying sexual predators by svm classification with lexical and behavioral features. In *Proceedings of CLEF 2012 (Online Working Notes/Labs/Workshop)*, CEUR-WS, Rome, Italy.

Parapar, J., Losada, D. E., and Barreiro, A. 2012. A learning-based approach for the identification of sexual predators in chat logs. In *Proceedings of CLEF 2012 (Online Working Notes/Labs/Workshop)*, CEUR-WS, Rome, Italy.

Peersman, C., Vaassen, F., Van Asch, V., and Daelemans, W. 2012. Conversation level constraints on pedophile detection in chat rooms. In *Proceedings of CLEF 2012 (Online Working Notes/Labs/Workshop)*, CEUR-WS, Rome, Italy.

Pennebaker, J. W., Francis, M. E., and Booth, R. J. 2001. *Linguistic inquiry and word count: LIWC 2001*. Mahway, NJ: Lawrence Erlbaum Associates.

Platt, J. 1998. Sequential minimal optimization: a fast algorithm for training support vector machines. Technical Report msr-tr-98-14, Microsoft Research.

Popescu, M., and Grozea, C. 2012. Kernel methods and string kernels for authorship analysis. In *Proceedings of CLEF 2012 (Online Working Notes/Labs/Workshop)*, CEUR-WS, Rome, Italy.

Sun, Y., Kamel, M. S., Wong, A. K., and Wang, Y. 2007. Cost-sensitive boosting for classification of imbalanced data. *Pattern Recognition* **40**(12): 3358–78.

Vartapetiance, A., and Gillam, L. 2012. Quite simple approaches for authorship attribution, intrinsic plagiarism detection and sexual predator identification. In *Proceedings of the 6th PAN workshop at CLEF2012 on Uncovering Plagiarism, Authorship, and Social Software Misuse (PAN2012)*, Rome.

Vartapetiance, A., and Gillam, L. 2014. Our little secret: pinpointing potential predators. *Security Informatics* **3**(1): 1–19.

Villatoro-Tello, E., Juá rez-Gonzá lez, A., Escalante, H. J., Montes-y Gómez, M., and Pineda, L. V. 2012. A two-step approach for effective detection of misbehaving users in chats. In *Proceedings of CLEF 2012 (Online Working Notes/Labs/Workshop)*, CEUR-WS, Rome, Italy.

https://doi.org/10.1017/S1351324916000395 Published online by Cambridge University Press

Whittle, H., Hamilton-Giachritsis, C., Beech, A., and Collings, G. 2013. A review of online grooming: characteristics and concerns. *Aggression and Violent Behavior* **18**(1): 62–70.
Whitty, M. T. 2002. Liar, liar! an examination of how open, supportive and honest people are in chat rooms. *Computers in Human Behavior* **18**(4): 343–352.
Wolak, J., Finkelhor, D., and Mitchell, K. 2004. Internet-initiated sex crimes against minors: implications for prevention based on findings from a national study. *Journal of Adolescent Health* **35**(5): 11–20.
Wolak, J., Finkelhor, D., Mitchell, K. J., and Ybarra, M. L. 2008. Online "predators" and their victims: myths, realities, and implications for prevention and treatment. *American Psychologist* **63**(2), 111–128.

A Example of chat conversations from the corpus

- Chat between a pseudo-victim (agent impersonating a teenager) and a sexual predator:

User1: hi sweety
User2: o hi
User2: sorry
User1: hey baby what u doing?
User2: dang im actully doing hw
User2: lol
User2: report due tomorow
User2: lol
User1: on what?
User2: ann frank
User1: sex with a older man?
User2: huh
User2: u dont no who she was?
User1: oh thats better that my idea lol
User2: lol ur bad
User1: i know lol
User2: thats wut i like abt u
User1: ¡—smartass
User2: lol
User1: still wanna get naked with me?
User2: hell ya
User2: i canrt wait
User2: u aint backing out r u?
User1: noway
User2: cool
User2: can i tty tomorow??
User2: i just started this report lol
User1: what u looking 4ward 2 most?
User2: just havin fun
User1: sure sorry
User2: naw its ok
User2: but u b on tomorrow rite?
User1: i dont want in trouble
User2: o dang i dont ether
User1: cool c ya then cutie
User2: cya :-*

User1: :-P

- Chat between two consenting adults engaged in cybersex:

User1: Asl
User2: asl
User1: 17 f usa
User2: 18 m usa
User2: horny?
User1: Hell ya
User1: Wanna sex chat
User2: sure
User1: I wanna fuck you and suck your dick
User2: cn u fit 9in down ur throat

https://doi.org/10.1017/S1351324916000395 Published online by Cambridge University Press

User2: ????/
User1: Ill try
User2: k go ahead
User1: Ohh yaah I can feel it now
User2: so do u hve nice tits n a ass
User2: ?????/
User1: Yaa I do
User1: Big bad tits
User2: how big?
User1: Really big
User2: dds
User2: ?

User1: What's that

- Chat between two adults on general topics:

User1: hi
User1: asl?
User2: hey
User2: tell me something interesting about yourself
User2: 23/m/ire
User1: 17 f poland
User1: interesting, u say?
User2: cool
User1: hmm
User1: for example?
User2: just soemhting to start a conversation about
User2: something*
User2: i'm a film maker
User1: really?
User2: yeah i'm a script writer/director
User1: i'm student.. hmm.. i love play the guitar and i read a lot og polish books. i'm nobody interesting, sorry
User2: lol everyone's interesting somehow
User1: have u made any film?
User2: yeah i've made a good few
User2: i've done some work in tv as well
User1: can i know your name?
User2: it's steven
User2: you?
User1: Adelajda
User2: cool
User2: nice to meet you
User1: me too :)
User2: i'm looking at getting funding together for a script i wrote
User2: i hope to make the film this summer
User1: but can i know your last name? are u famous? ^^
User2: i'm steven murrya
User2: murray*
User2: and no i'm not famous... yet
User2: i'm only starting out in film making
User1: of*
User2: so what are you thinking of doing after school?
User1: i want to be a doctor
User1: i'm in class with medical profile
User2: cool my first cousin is a doctor
User2: it's a very good job you get to save people's lives

User1: not really, i want to be a dermatologist :P

- Chat between two adults on a technical topic:

User1: User2: I believe that's not possible
User1: You'd have to have lots of canvases, or one big canvas which you copy all the fragments onto
User2: User1: thanks for the answer...
User2: What's the best (= most performant) way to realise ~1000 scopes that get updated each second?
User2: ist currently using images where the src get's modified - and that's extremely slow
User1: What kind of image is being shown in each?
User1: and how large are they?
User2: a little icon, 4x4 pixel, png, no transparency
User2: each scope has 8 of them
User1: Hmm, I find it hard to visualise what you mean
User2: [inserts images]

https://doi.org/10.1017/S1351324916000395 Published online by Cambridge University Press

User2: 1.png and 0.png either show a peak or a plain line
User2: currently a JavaScript function is called every second to move the src attribute from each img to its right neighbour
User2: and depending on other data the left most img is then set to 0.png or 1.png
User2: => it looks like a scope that's flowing to the right
User2: (it looks a bit like the CPU graph in the MS Windows task manager, the direction of the flow is just opposite)
User1: Sounds like doing it with canvas and drawing lines (rather than using images) should have much lower overhead
User1: Not sure how well it'd work if you have a thousand to update per second, though
User1: (Drawing them all onto one big canvas would probably be faster than drawing lots of little ones, I'd expect)
User2: User1: yes, that's my reasoning. One big canvas, only one translate and 1000 little draws per second resulting in only one redraw should be very lightweight
User2: well, I guess that I'll have to stay not optimal for now...

User1: can't think of a better solution

https://doi.org/10.1017/S1351324916000395 Published online by Cambridge University Press