Detecting Predatory Behavior in Game Chats

Yun-Gyung Cheong, Alaina K. Jensen, Elín Rut Guðnadóttir, Byung-Chull Bae, and Julian Togelius

***Abstract*—While games are a popular social media for children, there is a real risk that these children are exposed to potential sexual assault. A number of studies have already addressed this issue, however, the data used in previous research did not properly represent the real chats found in multiplayer online games. To address this issue, we obtained real chat data from *MovieStarPlanet*, a massively multiplayer online game for children. The research described in this paper aimed to detect predatory behaviors in the chats using machine learning methods. In order to achieve a high accuracy on this task, extensive preprocessing was necessary. We describe three different strategies for data selection and preprocessing, and extensively compare the performance of different learning algorithms on the different data sets and features.**

***Index Terms*—Chat, data mining, game data, natural language processing (NLP), preprocessing, sexual predator, text classification.**

I. Introduction

SOCIAL media is becoming increasingly prevalent today and children are frequent users of many different types of social media, such as online games [1]. Unfortunately, this can put them in a fragile position and enable others to take advantage of them.

In 2011, the EU Kids Online Project conducted a survey of 25 142 children aged 9–16 in 25 different European countries, regarding their activity online, which shows that 93% of 9–16-year-olds go online at least once a week. The purpose of the study was to assess risk factors such as bullying, being approached by strangers and receiving sexual messages. The study showed that one in eight children have encountered sexual images and received sexual messages online. Few of the children exposed to sexual content were actually upset by it, which suggests that in many cases, children do not understand the risk associated with such content or messages [2].

Definition

Morris and Hirst [3] define sexual predation as having two characteristics: "*age disparity*: a predator is an adult who chats with an underage individual" and "*inappropriate intimacy*: the adult must introduce or encourage intimate conversation." In our experiments, this definition is modified by the omission of the *age disparity* element, because of the context and circumstances of the game *MovieStarPlanet*. Among the rules stated on *MovieStarPlanet*, is the rule "Don't write things that are sexually suggestive…."[1] The rules also forbid the exchange of personal information such as addresses, phone numbers, or social network profiles. Because the rules specifically prohibit these behaviors, we define a *sexual predator* as follows.

1) Anyone who initiates sexually suggestive language. This can be either obvious as in "Let's have sex" or subtle as in "What does your underwear look like?"
2) Anyone who welcomes this type of language, and responds with similar language.
3) Anyone who tries to gain physical access to other users of the game (i.e., "Let's meet in real life").

In the context of this project, a user receives a predator (P) label based on this definition regardless of age, because the rules of the game strictly prohibit this type of language without respect to which person is using it. According to our definition, the term sexual predator is synonymous with the term rule breaker in the context of *MovieStarPlanet*. The labeling was done by the moderators of *MovieStarPlanet*, often based on reports from users.

A. Challenges

We consider the task of labeling predators in a chatlog as a machine learning task, specifically a text classification task using supervised learning. The uniqueness of our work has much to do with our collaboration with *MovieStarPlanet* and the use of their data. This entails a unique set of challenges as the text below, which affect both the methodology and outcome of our experiments.

The style of chatting in *MovieStarPlanet* is very different from other forms of written text, due to a very high level of misspellings, slang, grammatical errors, and seemingly meaningless symbols. Some of these characteristics are common in chat data as opposed to other online text data [4], but possibly even more so in the *MovieStarPlanet* data set due to the young age (8–15 years) of the chat participants.

Second, the data set distribution between predators (P) and nonpredators (NP) was not uniform. While the predator chatlogs we were given sometimes spanned a long period of time,

Manuscript received November 30, 2013; revised June 01, 2014; accepted March 02, 2015. Date of publication April 21, 2015; date of current version September 11, 2015. This work was supported in part by the Hongik University under a new faculty research support fund.

Y.-G. Cheong is with the Department of Computer Engineering, Sungkyunkwan University, Suwon 440-746, Korea (e-mail: ygcheong@gmail.com).

A. K. Jensen is currently working independently (e-mail: alainapaina@gmail.com).

E. R. Guðnadóttir is with the IT Department, Financial Supervisory Authority, Reykjavik 105, Iceland (e-mail: elinrut@gmail.com).

B.-C. Bae is with the School of Games, Hongik University, Sejong 339-701, Korea (e-mail: byungchull@gmail.com).

J. Togelius is with the Department of Computer Science Engineering, New York University, New York, NY 10003 USA (e-mail: julian.togelius@gmail.com).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCIAIG.2015.2424932

[1] http://info.moviestarplanet.co.uk/terms-conditions.aspx

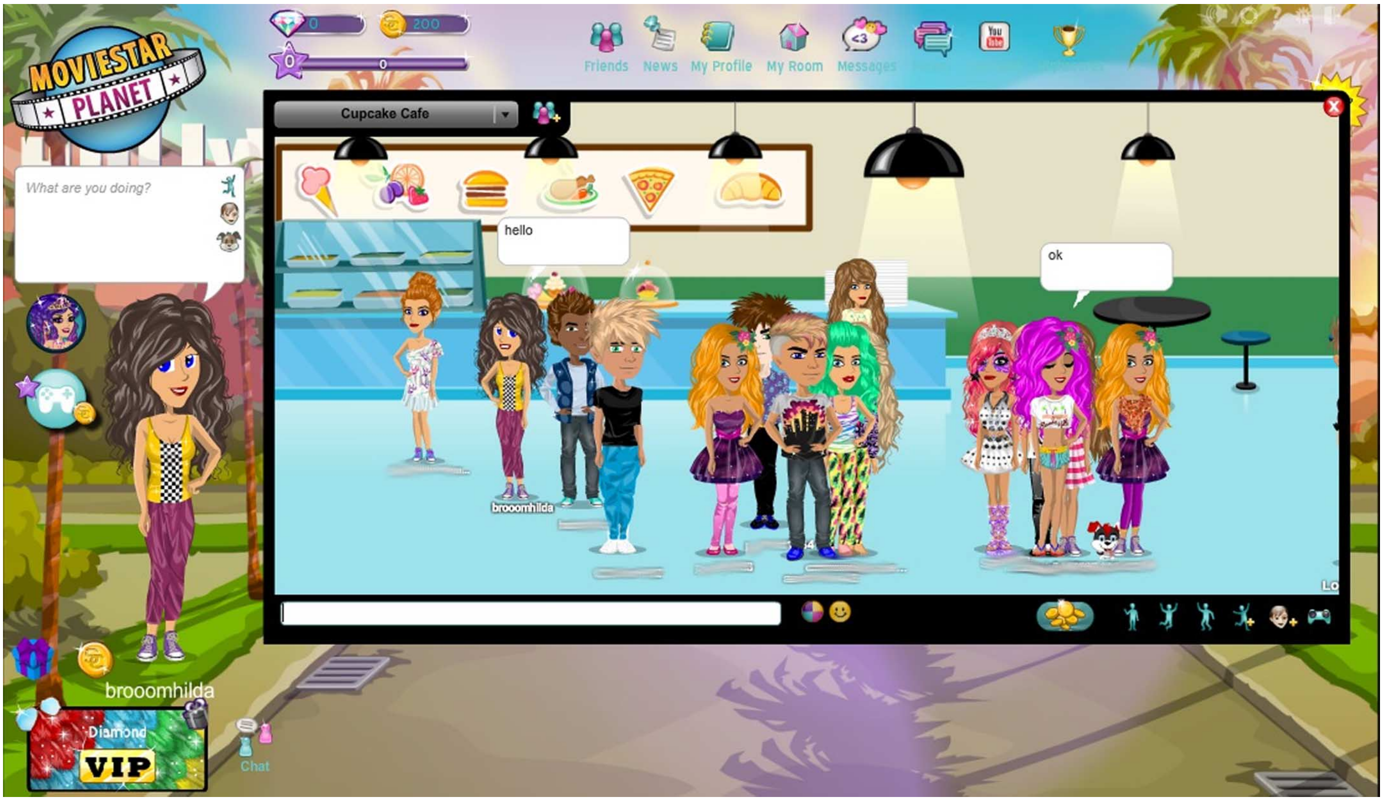

Fig. 1. A screenshot of a chatroom in the *MovieStarPlanet* game.

the normal chatlogs we were provided with took place during a shorter time frame.

Third, the nature of the game often calls for a language that may be similar to a predatory language, e.g., the children can be in a relationship, they very often talk about dating, being single, looking for or having a boyfriend/girlfriend and loving their boyfriend/girlfriend. They also appear to frequently engage in family role play, e.g., "Pretend I am your dad/mom/sister/brother." This sort of conversation within the normal chats easily causes a large number of false positives when trying to detect predatory behavior.

Finally, one of the greatest challenges is that the users of the game frequently try to circumvent the automatic safety nets in the game by using creative spelling and adding extra spaces, symbols, or line breaks when using words from *MovieStarPlanet* 's blacklist of words. To address this challenge, we have constructed a feature set which is designed to detect this type of behavior.

B. *Our Approach*

In order to address the problem, by detecting predators and other rule-breakers in game-based chats, our work applied standard natural language preprocessing methods and machine learning algorithms on recent data from the successful online game and community for children, *MovieStarPlanet* (Fig. 1) [5]. These data were collected from the actual game as it was played, meaning that the data and our experimental results have maximum ecological validity [6]. Text classification algorithms have successfully been used for finding sexual predators in the past; our problem is different both because of the way we define sexual predators, and especially because we wish to do so for the context of an online game for children.

In this paper, we examine the *MovieStarPlanet* game chat data to address the following three hypotheses.

- Lexical information of a chat [i.e., Bag of Words (BoW) representation in our work] in combination with supervised classifiers can discriminate between a predator and a non-predator in a real game chat.
- Behavioral information (i.e., rule-breaking features in our work) can be used to predict whether the chatter is a sexual predator or not in real game data.
- The last part of a chat just before being caught can signal if the speaker is a sexual predator.

C. *Contributions*

This paper is mainly an application paper presenting a unique case study. However, we also introduce a method, based on existing algorithms, which can be used on variations of the same problem and on different data sets. The contributions of this paper include the following.

- We present the first published use of machine learning algorithms to identify sexual predators in a real setting, based on real data where no participant is a pseudovictim, and validate it in a live system where it was actually used to catch predators.
- We address the problem of rule-breaking behavior in addition to sexual predation.
- We address the problem of detecting predators and rule breakers in game chats, which are considerably different

in nature to ordinary chats as they feature an element of role playing.
- We propose a number of new features extracted from chat logs (e.g., one letter lines, nonletter words, consecutive identical letters) which we have not seen used in the literature.
- We compare several different ways of posing the classification problem, in addition to comparing classification and preprocessing algorithms.
- We additionally validate the method using a standard data set.

D. Overview

The paper is organized as follows. Section II reviews the previous efforts and approaches to automatic detection of online predators including a description of available data in the field. Section III contains detailed description of the data set our data preprocessing method. This is followed by sections describing the features and how they were created (Section IV). We then discuss the methodology used in the experiment and the results (Section V, Section VI, and Section VII). Finally, Section VIII summarizes the major findings from the experiments, and the future work.

II. Related Work

A. Publicly Available Data

Pendar [7] defines the types of data for predators as follows:

1) predator/other:
   a) predator/victim (victim is underage);
   b) predator/pseudovictim (volunteer posing as child);
   c) predator/pseudovictim (law enforcement officer posing as child);
2) adult/adult (consensual relationship).

There is a general lack of publicly available chatlogs of type 1a because of privacy issues [1], [8], whereas chats of type 1b have been made available through nonprofit, volunteer based organizations, specifically perverted-justice.com.

1) Perverted Justice: The Perverted Justice Foundation (PJ) [9] is a nonprofit organization based in the United States, which is dedicated to finding and convicting pedophiles and other sexual predators who use chatrooms to find their victims. Their goals also include creating an atmosphere in regional chatrooms which is not conducive to predatory behavior.

The method of this organization is to hire and train adult volunteers to pose as adolescents in chatrooms [1], especially regional chatrooms and social networking sites. When encountering a sexual predator online, these volunteers are able to chat with them until the predator incriminates him or herself with predatory language, which is then reported to the police and often results in criminal convictions. The number of predators which have been convicted due to perverted-justice.com since 2004 is currently 550. Files containing full chatlogs of convicted sexual offenders are available for download on perverted-justice.com, along with analysis of key chat segments, and information about each offender.

These data have been used in most of the classification experiments which have been done so far in sexual predator detection [1]. This is the most credible data source for predators which is publicly available, however, its use is controversial [1], because the victims are not real, therefore this is data of type 1b above.

2) PAN 2012: PAN 2012 was a competition held in conjunction with CLEF 2012 (Conference and Labs of the Evaluation Forum), in September 2012 in Rome, Italy. The PAN2012 data set [8] consists of chatlogs drawn from various sources including true positives (predators) from PJ (see Section II-A1), negative data from IRC logs [10] (chats about generic topics), and false positives from the Omegle repository [11] (consensual chats about sex). The data set consists of a training set and a testing set. Overall, the data set has an extreme class imbalance problem, meaning that the nonpredators far outweigh the predators. The training data set consists of 66 914 conversations produced by 97 671 users of which 142 were labeled as a predator (0.15%). The conversations that involve the predators account for 4.52% of the total posts [8].

The PAN 2012 Identifying Sexual Predators data set is undoubtedly well researched and thoughtfully constructed, and served its purpose well in the PAN competition. However, a few drawbacks in the data set have been noted, including the following.

- The positive (P) data included in the data set is drawn from PJ, which has been described as controversial, as the predators are real but the victims are not. The question remains how much this data reflects the actual problem [1], [3], [12].
- The negative data and the false positives are both drawn from sources which are entirely different than PJ [8], which could degrade the quality overall, as the slang, acronyms, etc., used in chat dialogue can be specific to a particular online context [13].
- As Internet slang is always evolving [14], selecting chatlogs from several years ago and combining them with newer chatlogs might have misleading effects.

We also note that research on sexual predator identification has not yet been done in the context of an online game for children, but only in the wider context of chatlogs in general. Chatting in the context of game-playing can be considerably different from other types of chatting. That data offered by *MovieStarPlanet* is of the coveted type 1a (i.e., real predator–real victim under age) which is very difficult to obtain [1], [8], and does not include the aforementioned drawbacks.

B. Previous Approaches

Due to lack of access to reliable data [1], [8], the problem of detecting online sexual predation has not been studied much until recently [4]. The strategies used so far have been based on text classification focusing on detecting lines as predatory and discriminating between a predator or victim. Some of the work has been based on communication theories [15]–[17], while others base their research on the assumption that predatory language is different from other types of language [4], [18]. Preprocessing and filtering strategies are often used during the training stage [12], [16], [18], [19].

One of the earliest experiments was conducted in 2007 by Pendar [7]. To discriminate between a predator and victim in a text containing a predatory conversation, he collected 701 text logs from PJ. Each log was divided into two files; a file containing chat lines written by the predator and a file containing chat lines written by the pseudo-victim, thus, the corpus ended up consisting of 1402 files. Pendar experimented with unigrams, bigrams and trigrams to construct the feature set. The preprocessing step removed stopword and fixed repeated letters in words (e.g., “nooo” to “no”), but did not apply stemming or spelling correction. The best results were achieved when using trigrams and the k-NN classifier (F measure of 0.943 with $k = 30$).

McGhee *et al.* [17] developed a program called *ChatCoder* which can determine the lines containing predatory language from a given text. The program has undergone several versions. Its rule-based version has shown to outperform its simple phrase-matching version, however, its machine learning version has not shown significant improvement over its rule-based version. *ChatCoder* is built upon communication theories proposed by Olson *et al.* [20], in which the process of predation consists of five phases: gain access, deceptive trust development, grooming, isolation, and approach. This model was later elaborated by Leatherman [21] to develop a coding system for the context of online sexual predation. The deceptive trust development phase can contain four subcategories: personal information, information about relationships, information about favorite activities, and compliments. Kontostathis *et al.* [1] developed rules to identify predatory text, and noted that both the Leatherman and Olson models turned out to be too complicated for conversations that take place in an online environment. Based on these findings, McGhee *et al.* [17] simplified the Olson's model to contain only three classes: exchange of personal information, grooming, and approach. They also coded lines containing none of the classes, to filter out nonpredatory lines for the training phase.

Bogdanova *et al.* [4] used sentiment analysis to test whether a chat was predatory or not. They used PJ data to form the positive data set, and collected cybersex logs available online and NPS chat corpus to form the negative data set. Their approach was drawn from research suggesting that pedophiles often behave in a distinct manner; they are emotionally unstable and suffer from psychological problems. Therefore, Bogdanova *et al.* [4] attempted to detect predatory text using the features of emotional markers—words that express joy, sadness, anger, surprise, disgust, and fear. Positive and negative words, emoticons, and imperative sentences were also considered. They also used communication-model features borrowed from McGhee *et al.* [17], the words associated with approach, relationship, family, communicative desensitization, and sharing information. Other features included the word usages helpful to detect neuroticism level (e.g., percentages of personal and reflexive pronouns and modal obligation verbs) and fixated discourse features (the unwillingness to the change a topic within a conversation).

TABLE I
RESULTS OF THE PAN2012 COMPETITION FOR THE TASK OF IDENTIFYING PREDATORS. ADAPTED FROM [8]

Participant	**Precision**	**Recall**	**F1**	**F0.5**
Villatoro-Tello et al. [19]	0.98	0.77	0.87	0.93
Parapar and el. [18]	0.94	0.67	0.78	0.87
Morris and Hirst [23]	0.97	0.60	0.75	0.87
Eriksson and Karlgren [25]	0.86	0.89	0.87	0.86
Peersman et al. [16]	0.89	0.60	0.71	0.81
Kontostathis et al. [15]	0.36	0.67	0.47	0.39
Bogdanova et al. [4]	0.03	0.22	0.05	0.03

The rest of this section details PAN 2012 competition (see Table I for the results). Villatoro-Tello *et al.* [19]—their system took the first place in the PAN2012 competition for the predator detection task (see Table I)—present a two-step framework to detect predators in a chat. The first step of their system detect suspicious conversations, and the second step identifies the predator from victim in a suspicious conversation. For the first step they trained a classifier with a corpus containing conversations they had labeled suspicious, and for the second step a classifier was trained with predatory conversations, which were divided into a victim and predator interventions. No preprocessing (e.g., removal of punctuation marks, stopwords, or stemming) was used. The data however went through prefiltering, where conversations with only one participant were removed, as well as conversations with less than 6 interventions per user, and finally conversations containing long sequences of characters were also removed. This filtering resulted in 90% reduction ratio of conversations/users. Villatoro-Tello *et al.* experimented both with SVM and NN (two layer, single hidden layer of ten units) with BoW feature representation. The best performance of F measure ($\beta = 0.5$) of 0.9346 was reached using NN classifier with binary weighting in both steps.

Drawn from psycholinguistics research, Parapar *et al.* [18] incorporated linguistic inquiry and word count (LIWC) [22], which measures to what degree different categories are used by people, in the features. The feature set also contains terms (TF–IDF) and chat-based features, and they experimented with unigrams, bigrams, and trigrams and the combination of them. No stemming was used and no feature dimensionality reduction was applied. Instead, they used SVM due to the large number of features. Chat-based features represent the activity of the chatters, such as number of subjects contacted by a chatter, the percentage of conversations initiated by a chatter, etc. For the PAN 2012 competition a run consisting of TF–IDF based on unigrams and chat-based features was nominated, which resulted in the third place for the first subtask. An interesting aspect of their research is that the use of the LWIC features did not improve the results.

Morris and Hirst [3], [23] used lexical and behavioral features. The lexical features used BoW representation of unigrams and bigrams, and the behavioral features consisted of information that can be extracted from the conversations, such as the number of messages sent by an author and the total number of conversations which this author participated in. To identify predators they used an SVM classifier with a radial kernel and two filters to distinguish predators from victims, since a large portion of the false positives were victims. Using only lexical features they managed to get an F score of 0.77. The behavioral features did not enhance the results when used on top of the lexical features, but used alone they gave a reasonable classification with F score of 0.56.

TABLE II
GENERAL STATISTICS FOR P (PREDATOR) AND NP (NONPREDATOR) CLASSES IN THE RAW DATA

	Number of Users	*Number of lines*	*Unique words*	*Misspelled words*
P	59	40,413	1,921	648
NP	8,707	62,704	16,135	9,799

TABLE III
COMPARISON IN THE NUMBER OF LINES PER USER BETWEEN P AND NP CLASSES IN THE RAW DATA. FOR INSTANCE, THERE WERE 5063 (58.15%) NPS THAT EACH ENTERED FIVE LINES OR LESS IN THE DATA SET

	1-5	6-10	11-15	16-20	21-25	26+
P	2	1	1	0	3	50
NP	5,063	1,712	882	451	274	325

Peersman *et al.* [16], [24] presented a three-stage approach which combined predictions of the three levels: the level of individual post, the level of the user, and of the entire conversation. They used two SVM classifiers, one to detect a predatory post, and another one to classify a chatter as a predator or nonpredator. To identify predatory posts, token unigram features representing a post was used. To identify users, a single instance vector containing all posts from the same user was used. The results from these two classifiers were then combined to level out the high recall results from the post classifier and the high precision results from the user classifier.

As discussed above, rule-based systems and machine learning approaches have been actively used to detect online sexual predators. Although only a few of the contributions have been discussed in this section, there is a common denominator amongst them despite different approaches. It is feasible to discriminate between a predator and nonpredator with relative high accuracy using supervised classifiers in combination with simple BoW representation [7], [16], [19], [24]. In addition, high-level, behavioral features have been being incorporated into the feature set [3], [4], [17], [18].

III. DATA PREPROCESSING

We employed two types of data set (P: predator data; NP: nonpredator data) provided by *MovieStarPlanet*, which consisted of all of the verbal communication from different users of the game—including statuses, comments on videos and forum postings, as well as public and private chats from chatrooms and games. All user IDs and IP addresses were anonymized in these data, to protect the *MovieStarPlanet* users.

A. Raw Data

The raw data that we received were classified as either unlabeled (that is, normal and presumably nonpredator) or labeled (as predator).

1) Unlabeled NP Data: Two normal chat data were given by *MovieStarPlanet*, where each data contains approximately 65 000 lines of 15 min of gameplay across the entire U.K. site on a particular date. One normal chat data was used for NP data in our training set (after extensive preprocessing). The other was used for unlabeled testing (as will be described in Section VI). As these normal chat data were all essentially unlabeled, *MovieStarPlanet* and we presume that these files did not contain users which fulfilled our definition of a sexual predator.

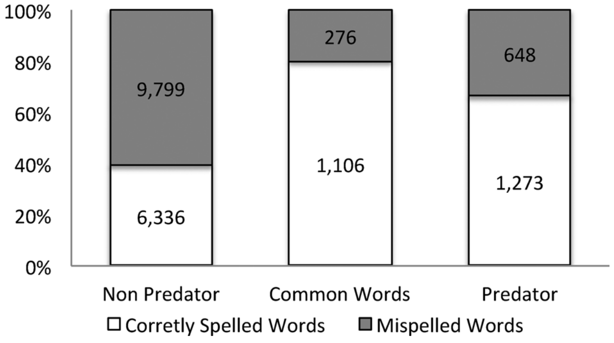

Fig. 2. Distribution of misspelled words for the categories of nonpredator, predator, and common words between the two.

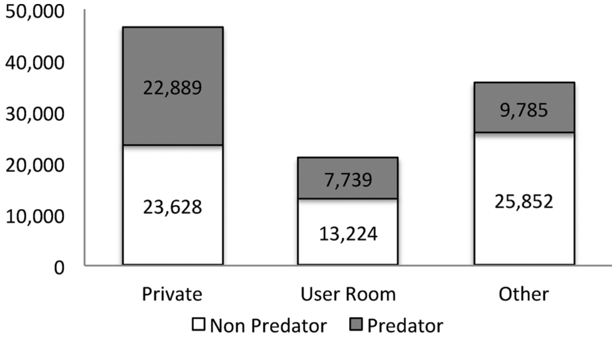

Fig. 3. Distribution of chat types of recipients between predator and nonpredator in the raw data. The x-axis denotes the chat types and the y-axis denotes the numbers of lines written in the conversation that take place in the chat type.

2) Labeled Predator (P) Data: Unlike the NP data described above, predator data were already labeled as predator (manually by human moderators from *MovieStarPlanet*), containing the full chatlogs (sometimes spanning up to three months) typed by 59 predators.

B. Descriptive Statistics of Raw Data

This section describes the statistical description of the raw data sets provided by *MovieStarPlanet*. Table II characterizes the general statistics for P and NP in terms of the number of users, lines, unique words, and misspelled words.

1) Number of Users: It is noted that there is an imbalance between the number of users in both classes. We emphasize that these figures are valid only for the data set that we worked with. In fact, predators comprise a much smaller percentage than 1% of all *MovieStarPlanet* users.

2) Number of Lines: As shown in Table II, the number of total lines do not show significant class imbalance (40 413 lines for P and 62 704 lines for NP). However, the length of the chats per user differs greatly between the P and the NP sets (see Table III), which results from the long timespans of the chats in the P class, sometimes encompassing several months, as opposed to the 15-min chatlogs in the NP class.

3) Word Usage: The NP data set contained 16 135 unique words and the P data set contained 1921 unique words. There

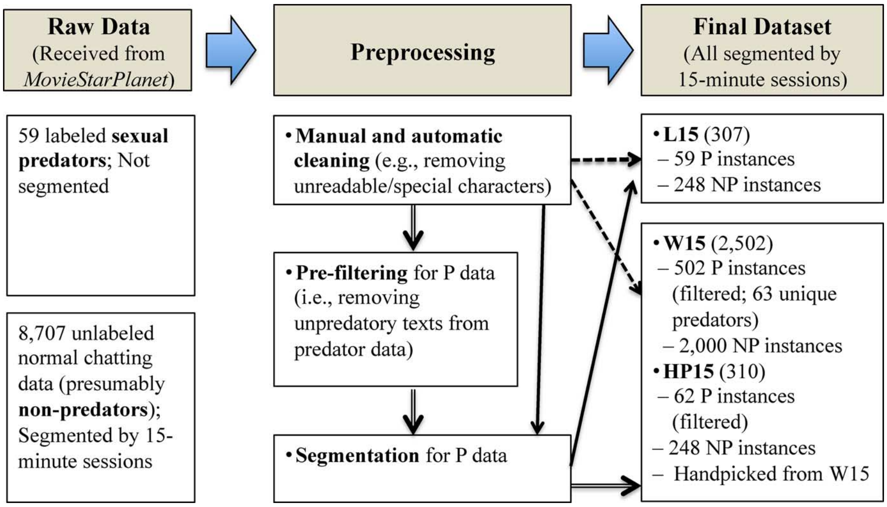
Raw Data
(Received from *MovieStarPlanet*)
Preprocessing
Final Dataset
(All segmented by 15-minute sessions)
59 labeled **sexual predators**; Not segmented
8,707 unlabeled normal chatting data (presumably **non-predators**); Segmented by 15-minute sessions
• **Manual and automatic cleaning** (e.g., removing unreadable/special characters)
• **Pre-filtering** for P data (i.e., removing unpredatory texts from predator data)
• **Segmentation** for P data
• **L15** (307)
– 59 P instances
– 248 NP instances
• **W15** (2,502)
– 502 P instances (filtered; 63 unique predators)
– 2,000 NP instances
• **HP15** (310)
– 62 P instances (filtered)
– 248 NP instances
– Handpicked from W15

Fig. 4. Overview of the data preparation process and final data subsets after preprocessing. Solid lines denote the flow of unfiltered P (predator) data preprocessing; double lines denote the flow of filtered P data processing; dotted lines denote the flow of NP (nonpredator) data preprocessing. Note that HP15 is a hand-picked subset of W15. For HP15, only one instance is selected per user. Note also that the P data in W15 and HP15 are filtered and the P data in L15 are not filtered.

were 1382 words in common between the two classes, which encompasses 71.94% of the P words, but only 8.56% the NP words. The misspelled words in common between P and NP were only 276, accounting for 14% of P and 1.71% of NP. This indicates that P and NP, at least in the data set we received, misspell in different ways. Fig. 2 shows the distribution of the misspelled words for each set.

It is interesting to note the difference in the percentage of misspellings in both classes—60.73% for NP and 33.73% for P. Nevertheless, this difference could be simply due to the large number of users present in NP, each with their own unique style of writing and misspelling. It is also important to point out that most of the misspellings recorded in these statistics were probably one time misspellings, each carried out by one user. It was our view that misspellings could be a useful feature when describing the behavior of a user, but could also be a stumbling block when creating a BoW feature. Therefore, misspellings became a focus on which several different techniques were used.

4) Chat Types: We divided the chatting types into three categories where chats can take place in the game: private, user room, and other. In Fig. 3, the private category denotes a 1–1 conversation between two confirmed friends, which cannot be viewed by other users. In a user room, it is possible to have either private or public conversation. The other category describes a public chat room or any of the other publicly available forums in the game. The figure shows an interesting trend. While the distribution between P and NP is equal in the private category, the lines written by P are much less than those written by NP in the user room and the other categories. This suggests that predators in this data set tend to have private chats.

C. Data Set Preparation

This section details the procedure to prepare data sets that consists of three steps: cleaning, prefiltering, and segmentation, as illustrated in Fig. 4.

1) Preprocessing: We constructed the P data set from the initially given predator data containing 59 sexual offenders' full chat logs for the maximum duration of three months. We employed one of the two regular chat data sets for our training purposes: NP, and the other one for the final testing (Section VI). Fig. 5 includes examples of predatory language, and Fig. 6 includes examples of nonpredatory language (which possibly include some other type of offense than sexual predation). The resulting P and NP data sets were then given as input to the Weka software package [26] with standard natural language processing (NLP) operations (e.g., tokenizing, filtering stop words, stemming, etc.).

who wants to have *** with me
can i fiddle with ur body?
what is your email
it with me i can feel it now uhh uhh uhh ohh harder HARDER
if u want to be my gf tell me where u live then

Fig. 5. Examples of lines of predatory language included in the offender data. Segments including these lines were labeled as ”sexual predator.”

Someone give me an auto and if u do i will give u one back! ;p
want to swap accounts
f* off other wise i will brske ur a*
ITS SPRING SUMMER HOLIDAYS BEACH ICE-CREAM AND I AM FREE NO SCHOOL FOR 2 WEEKS !!!!!!!
Nothing whats wrong with UR HAIR ITS WEARD

Fig. 6. Examples of lines of nonpredatory language, including some other type of offence than sexual predation, in the offender data. These lines were not included as decisions for labeling as sexual predator.

We ran the initial testing on this data, using the full (up to three months) texts of the unsegmented P data and the 15-min

NP data sets, and obtained disappointing results. We hypothesized that this was due to the noise within the predator chats, caused by a high percentage of nonpredatory text (see Fig. 6), with only short segments of chatting which were predatory in nature). As random undersampling of predator data can potentially eliminate important examples[27], we manually filtered out lines that do not show sexual predatory behavior. This prefiltering process has been used in previous approaches [12], [16], [18], [19] and can be automated using software packages [17] or a rule-based system [12].

As we manually eliminated the lines that are clearly not predatory in nature, we noted that many of the would-be "victims" of those predators were actively participating in the sexual conversation along with the original predator, and thus four more predators were found based on our definition of predatorhood as described in Section I. The total number of predators after preprocessing was 63.

2) Resulting Data Subsets: The P and NP data did not match in time span due to the way it was collected, which caused a class imbalance. To alleviate this problem, we created data subsets. All data subsets maintained a distribution of 20% P data and 80% NP data. The NP data were randomly undersampled to achieve this distribution, and the P data were prepared differently for each data set as described later. To solve the mismatched time span problem, predator data were always concatenated into 15–min segments in each data subset in order to match the duration of NP chat.

When we manually went through the predator data to determine the predatory lines, it was noted that many of the violations (which were recorded by the human moderators) were located in the end of the chat log. Thus, we hypothesize that the last 15 min of a predator chat might contain the predatory behavior. In order to test this hypothesis, a data subset was created which contained the last 15-min segment of each predator chat, which constituted one tuple (or instance) per user. This data subset was named L15 (last 15 min of unfiltered predator chats). For this data set, the unfiltered data were used in order to avoid the cases where the last 15 min of the file might have been filtered out.

We used the filtered predator data (of which nonpredatory texts were eliminated) to create two other data subsets: W15 (whole filtered predator chats divided into 15-min segments) and HP15 (single hand-picked 15-min segments of filtered predator chats). For W15, the entire filtered predator chats containing predatory texts were simply divided up into 15-min segments. For HP15, we manually picked only a single 15-min segment per user that contains the most predatory language (see Fig. 5) and also gave precedence to the longest chat segments.

The number of users in W15 (see Table IV) still shows a class imbalance problem. However, this data subset compensates for the user class imbalance by counting predators more than once, i.e., dividing their long chats into 15-min segments, each representing one instance in W15. Therefore, the number of instances for W15-P is actually 502, which corresponds to 2000 instances for W15-NP without a significant class imbalance problem. The number of predators differ among the data sets. L15 set uses the unfiltered NP data set, which resulted in 59 predators. W15 uses all the predators found during the manual filtering process, 63 predators. One predator was ignored due to an error in the data preparation stage for the HP15 set, which resulted in 62 predators.

TABLE IV
GENERAL STATISTICS FOR ALL THE DATA SUBSETS, DIVIDED INTO STATISTICS FOR P AND NP IN EACH DATA SET

	Number of Users	*Number of lines*	*Unique words*	*Misspelled words*
W15-P	63	9,138	3,193	1,350
W15-NP	2,000	20,699	7,989	4,090
L15-P	59	1,063	926	249
L15-NP	248	3,388	2,532	882
HP15-P	62	1,220	1,086	339
HP15-NP	248	3,388	2,532	882

TABLE V
FEATURE VECTOR

Feature type	Sub-category	Size of the feature vector
BoW	No sub-category	dynamically generated
Sentiment	Emoticons	3
	Sentiment scores	2
Rule Breaking	Behavioural Features	7
	Blacklist/Alert list features	2

IV. FEATURE EXTRACTION

To model the predatory behavior, we constructed a feature vector that consists of the three types of features (see Table V): 1) BoW features of the chat texts which vary depending on the vocabulary diversities of the data set (i.e., based on bigram model, 442 word features for W15, 865 words features for HP15, and 804 word features for L15); 2) five features relating to sentiment analysis; and 3) nine rule-breaking features that are specifically designed to represent rule-breaking behaviors in our data set.

A. Bag of Words

BoW has been established in numerous previous paper as an effective feature for text classification [3], [7], [18], [25]. In all our BoW representations, we employed TF–IDF weighting and bigram language model, which are widely accepted and utilized in PAN2012. We pruned the BoW features that occurred higher than top 1% of frequency and that occurred lower than the bottom 1% of the total words. Stopwords (e.g., and, or) were also eliminated.

As our initial attempts were not successful, we hypothesized that the cause of this could be the abundant misspellings in the data set. This is because, with a BoW, each word or consecutive words in the data set essentially becomes one feature, and the frequency of that word within the entire message becomes the value of that feature. This is only effective if the words are spelled in a uniform way. For instance if the word "address" is spelled as "adress," "addres," "adres" in addition to the correct spelling within the data set, which is actually only one word, it will create four different features and thus the BoW results will be skewed.

Although misspellings can have an important meaning in the context of the chat as argued in [7], [19], we conclude that, at least in our data set, any potentially meaningful misspellings are dwarfed by the number of superfluous misspellings. As stated before, this is undoubtedly due to the young age of the chat participants in this case. Therefore, we corrected the misspelling using the Jazzy automatic spell checking API [28], before creating a BoW from it.

B. *Sentiment Features*

Our sentiment features consisted of emoticon scores (three features) and sentiment scores (two features). Three types of emoticons—positive [e.g., :-), :*], neutral (e.g., :-P), and negative [e.g., :-(, 8(, etc.]—were extracted from the data using regular expressions. We computed positive and negative sentiment scores, using AFINN-111 wordlist [29] which includes labeled and scored sentiment words. The scores were collected and added together separately as a negative sentiment score and a positive sentiment score, to avoid the negative and positive numbers negating each others effect. We used AFINN-111 list for our chat data because AFINN-111 sentiment wordlist was specifically created for microblogging (i.e., twitter) and was more effective on this type of the Internet content than the standard but older ANEW list [29].

C. *Rule-Breaking Features*

In addition to the BoW and sentiment features, nine other features were created, which we call rule-breaking features. The premises for creating these features were the observations made on the data set about the ways in which users try to avoid being caught when breaking the rules. Most of them involve avoiding the blacklist words in various ways. These rule-breaking features are further divided into blacklist/alertlist features, and behavioral features.

1) Blacklist/Alert List Features: MovieStarPlanet uses an extensive wordlist including alert words and blacklist words, with many spelling variations on each word. As *MovieStarPlanet* provided their full alert/blacklist for our project, we counted the number of alert or blacklist words within a text.

Alert words do not necessarily indicate bad behavior. These words can be said in the context of the game, but frequent usage of such words will alert the moderator. On the other hand, blacklist words are not allowed in the context of the game, and are blocked. In some cases, however, it's possible to type a word which contains a blacklist word, surrounded by other letters or symbols, without that word being blocked. We therefore derived two features from the alert/blacklist: a count of alert words found in the text and a count of how many times a blacklist word was enclosed in another word or phrase.

Table VI shows the average count of alert words and blacklist words appearing in each data set. As seen from the table, clear distinction is made between P and NP in terms of the number of occurrences of alert and blacklist words across all the three data sets.

TABLE VI
AVERAGE OCCURRENCE OF ALERT WORDS AND BLACKLIST WORDS PER USER

	Alert Words	*Blacklist Words*
W15-P	0.30	0.34
W15-NP	0.02	0.00
L15-P	0.49	0.56
L15-NP	0.05	0.00
HP15-P	0.76	0.35
HP15-NP	0.05	0.00

2) Behavioral Features: These features are defined based on the observation on rule-breaking behavior from the data set. Each feature is the count of occurrences of the following.

- One letter lines (LL)—A user types a blacklist word, by typing one letter, hitting enter, typing the next letter, hitting enter, etc., until the full or partial blacklist word has been spelled out.
- One word lines (WL)—A user types a forbidden phrase by typing one word at a time.
- Lines (NL)—The number of lines itself could possibly indicate suspicious behavior.
- Spaces (SP)—A user types blacklist words with spaces in between the letters in a word (e.g., “s e x”).
- Nonletter words (NLW)—A user types blacklist words by typing symbols or numbers in place of letters, for instance “s*x.” This feature records a count of the words which contain symbols and/or numbers in addition to letters.
- Consecutive identical letters (CL)—A user types many consecutive identical letters inside a word, for example “seeeeex.” Therefore, any word with more than two consecutive identical letters is counted in this feature.
- Misspellings (MS)—A user types misspelled words to avoid being caught.

A simple statistical analysis of these features on the data revealed there is little distinct patterns between P and NP classes.

V. EVALUATION

To test our hypotheses mentioned in the Section I, we conducted a series of machine learning prediction on the three data sets we constructed using the Weka 3.7.11 version [26]. The machine learning algorithms include naïve Bayes (NB), J48 decision tree (DT), multilayer perceptron (MLP), logical regression (LR), IBk as a k-nearest neighbor algorithm (k-NN), and support vector machine (SVM). For J48, we experimented with different parameters and set the number of minimum instances under each node as 20, as this setting produced the highest accuracy rate. In a similar fashion, the number of neighbors participating in the voting was set 5 ($k = 5$) for the k-NN algorithm and the polynomial kernel was chosen for SVM. Unless mentioned otherwise, the Weka default setting was employed.

Each of these experiments was run on every data subset. The methodology behind the classification of the *MovieStarPlanet* data consists of the following three steps.

1) Extract all the 14 features and BoW features per instance.
2) Prepare a particular feature vector per instance for a particular data set.
3) Train and test on six classification algorithms under the fivefold cross validation.

Extracting BoW was done individually for each data set and produced different number of BoW features (865 for the HP15 set, 442 for the W15 set, and 804 for the L15 set).

TABLE VII
Results of Using Blacklist Features Only in Accuracy in P (Acc.), F1 Measure in P (F1), F0.5 Measure in P (F0.5), Precision in P (P), and Recall in P (R), for Each Data Set, for the Algorithm Mentioned

Dataset	Acc.	F1	F0.5	P	R
HP15	.82	.15	.3	1.00	.08
W15	.83	.26	.47	.99	.15
L15	.82	.09	.21	1.00	.05

We ran a stratified cross-validation test provided by Weka for HP15 and L15. For the W15 set contains multiple chatting instances of a single predator, we manually created five train and test sets. We inspected the W15 predator set, and found out that there are three predators of whose instances occupy the majority of the whole W15 P set (99 instances, 71 instances, and 36 instances, respectively). Eliminating these outliers leave us 296 instances in P. Therefore, we designed each test set to contain approximately 60 instances in a stratified manner, making sure that a particular predator's instances appear in the train set only or in the test set only.

For each test set, its corresponding train set includes the P instances that are not present in the test set. The NP instances for the test set and the train set were randomly chosen from the 2000 instances in the original W15 to keep the 20:80 ratio between P and NP.

A. *Baseline Prediction Using Blacklist Only*

Before we experimented with machine learning methods, we classified the predators based on the blacklist feature only as our baseline. Any chats that violated the words in the blacklist were classified as predators, and the ones without blacklist violation were classified as nonpredators. Table VII exhibits the basic classification results that will be used as a baseline. This classification on the W15 data set resulted in accuracy $= 0.83$, $F1 = 0.26$, $F0.5 = 0.47$. The majority of the predator instances did not violate the blacklist condition. The classification results obtained from HP15 and L15 were worse than that of W15: accuracy $= 0.82$, $F1 = 0.15$, $F0.5 = 0.3$ for HP15 and accuracy $= 0.82$, $F1 = 0.09$, $F0.5 = 0.21$ for L15. Although the accuracies were reasonably high even with this simple method, it is noted that F values are low due to many false negatives. It is obvious that blacklist features alone are not sufficient to detect sexual predators successfully.

B. *All Features*

We ran experiments using a combination of BoW, rule-breaking, and the sentiment features. Table VIII summarizes the results. Most of the six algorithms (except NB) applied on the W15 data set produced good performances overall, with the highest accuracy of 0.93, the $F1$ value of 0.78, and the $F0.5$ value of 0.86. Among the six machine learning algorithms, SVM and MLP produced the best result for HP15 and W15, respectively, and showed good results across the three data sets.

The prediction accuracies were enhanced from the baseline for HP15 (increase of 0.008 using SVM) and W15 (increase of 0.10 in using MLP), and improved marginally for L15 (0.01 in accuracy using DT). The $F1$ values were greatly improved, 0.56 for HP15, 0.52 for W15, and 0.43 for L15. This indicates that our method (learning the frequent word usage, rule-breaking behavior, and sentiment) are more effective for detecting predators than relying on the blacklist only.

TABLE VIII
Results of Using All Features in Accuracy in P (Acc.), *F1* Measure in P (F1), F0.5 Measure in P (F0.5), Precision in P (P), and Recall in P (R), for Each Data Set, for the Algorithm Mentioned. Standard Deviation Values (SD) are Also Provided for W15 in Parentheses. The Best Performance in Each Data Set Is Shown in Bold

Data	Alg.	Acc.	$F1$	$F0.5$	P	R
HP15	NB	.86	.67	.65	.64	.69
	DT	.86	.57	.68	.78	.45
	LR	.76	.60	.50	.45	**.89**
	kNN	.84	.34	.53	**.87**	.21
	SVM	**.89**	**.71**	**.73**	.75	.68
	MLP	.89	.70	**.73**	.66	.75
W15	NB	.79(2.20)	.57(.04)	.51(.04)	.48(.04)	.72(.07)
	DT	.91(3.99)	.74(.16)	.78(.06)	.80(.05)	**.73(.23)**
	LR	.89(3.11)	.70(.08)	.74(.09)	.76(.09)	.66(.10)
	kNN	.85(1.72)	.49(.14)	.62(.07)	.71(.06)	.40(.17)
	SVM	.92(2.51)	.77(.09)	.85(.06)	**.91(.05)**	.68(.12)
	MLP	**.93**(1.74)	**.78(.06)**	**.86(.04)**	**.91(.04)**	.70(.11)
L15	NB	.80	.46	.47	.48	.44
	DT	**.83**	**.52**	**.55**	**.57**	.48
	LR	.72	.49	.41	.38	.70
	kNN	.26	.34	.24	.20	**.97**
	SVM	.81	.47	.49	.50	.44
	MLP	.81	.40	.46	.50	.34

It is noted that all the algorithms performed worse on L15 (accuracy $= 0.83$, $F1 = 0.52$, $F0.5 = 0.55$) than on W15 and HP15. L15 was created to test the hypothesis that predators used their most predatory language during the last 15 minutes of their chat history, and this result indicates that our hypothesis regarding L15 was false. Therefore, we focus on HP15 and W15 in the rest of this section, and exclude the results obtained when using L15.

C. *Bag of Words Features*

To test the effectiveness of BoW for detecting rule-breaking users, we applied the six machine learning algorithms using BoW features only. Table IX shows the results that the best result was obtained when using MLP on the HP15 set (accuracy $= 0.90$, $F1 = 0.73$, $F0.5 = 0.73$). While MLP produced the best performance for HP15, its result for W15 was disappointing. On the other hand, SVM showed reliably good results for both data sets.

It is sensible that HP15 excelled W15 when using BoW (extracted from words used in the chat) alone, for HP15 contains the most predatory language. Particularly, for HP15 the use of BoW outperformed the result obtained when using all the features (see Table VIII), however, the difference was marginal (difference in accuracy $= 0.01$, $F1 = 0.02$, $F0.5 = 0.0$).

From this result, we can draw a conclusion that collecting a large-scale text corpus that correctly represents the predatory

TABLE IX
Results of Using Bag of Words Features in Accuracy in P (Acc.), F1 Measure in P (F1), F0.5 Measure in P (F0.5),Precision in P (P), and Recall in P (R), for Each Data Set, for the Algorithm Mentioned. Standard Deviation Values (sd) Are Also Provided for W15 in Parentheses. The Best Performance in Each Data Set Is Shown in Bold

Data	Alg.	Acc.	$F1$	$F0.5$	P	R
HP15	NB	.86	.67	.65	.64	.69
	DT	.83	.42	.55	.68	.31
	LR	.75	.57	.49	.44	**.82**
	kNN	.83	.29	.48	**.85**	.18
	SVM	.89	.72	.72	.72	.71
	MLP	**.90**	**.73**	**.73**	.77	.69
W15	NB	.77(2.21)	.55(.05)	.49(.04)	.46(.04)	**.68(.08)**
	DT	.87(3.84)	.50(.22)	.69(.06)	**.88(.05)**	.37(.2)
	LR	.82(3.2)	.50(.11)	.56(.09)	.54(.09)	.46(.13)
	kNN	.83(1.81)	.43(.08)	.55(.09)	.66(.09)	.33(.09)
	SVM	**.88(2.84)**	**.63(.12)**	**.72(.09)**	.80(.09)	.52(.13)
	MLP	.73(29.95)	.45(.27)	.68(.33)	.74(.32)	.51(.36)

TABLE X
Results of Using Rule-Breaking Features in Accuracy in P (Acc.), F1 Measure in P (F1), *F0.5* Measure in P (F0.5), Precision in P (P), and Recall in P (R), for Each Data Set, for the Algorithm Mentioned. Standard Deviation Values (sd) Are Provided in Parentheses. The Best Performance in Each Data Set Is Shown in Bold

Data	Alg.	Acc.	$F1$	$F0.5$	P	R
HP15	NB	.86	.56	.65	.74	.45
	DT	.86	.57	.68	.78	.45
	LR	.86	.57	.67	.76	.45
	kNN	.82	.40	.50	.59	.31
	SVM	.84	.38	.56	**.83**	.24
	MLP	**.87**	**.63**	**.70**	.77	**.53**
W15	NB	.86(3.28)	.62(.13)	.65(.06)	.65(.05)	.65(.19)
	DT	**.91(2.46)**	.73(.1)	**.80(.04)**	**.84(.04)**	.66(.16)
	LR	.90(3.04)	.72(.13)	.77(.03)	.80(.02)	.67(.20)
	kNN	**.91(2.71)**	**.75(.09)**	**.80(.06)**	**.84(.05)**	.68(.14)
	SVM	.90(2.80)	.73(.09)	.79(.04)	.77(.05)	**.69(.15)**
	MLP	**.91(2.15)**	.73(.09)	.79(.06)	.82(.05)	**.69(.18)**

TABLE XI
Results of Using Sentiment Features Combined With Rule-Breaking Features in Accuracy in P (Acc.), F1 Measure in P (F1), F0.5 Measure in P (F0.5), Precision in P (P), and Recall in P (R), for Each Data Set, for the Algorithm Mentioned. Standard Deviation Values (sd) are Provided in Parentheses. The Best Performance in Each Data Set Is Shown in Bold

Data	Alg.	Acc.	$F1$	$F0.5$	P	R
HP15	NB	.86	.56	.59	.74	.45
	DT	.86	.57	.60	.78	.45
	LR	**.87**	.59	.62	.78	.47
	kNN	.83	.40	.44	.67	.29
	SVM	.85	.40	.45	**.89**	.26
	MLP	**.87**	**.60**	**.63**	.79	**.49**
W15	NB	.85(3.87)	.59(.16)	.62(.09)	.62(.07)	.63(.20)
	DT	.89(4.29)	.74(.11)	**79(.05)**	**.82(.04)**	.70(.18)
	LR	**.91(3.04)**	.73(.13)	.78(.04)	.81(.04)	.69(.20)
	kNN	.90(.38)	.74(.02)	**.79(.04)**	**.82(.04)**	.68(.06)
	SVM	**.91(2.77)**	.75(.10)	**.79(.05)**	.81(.04)	.71(.15)
	MLP	**.91(1.95)**	**.77(.07)**	**.79(.06)**	.81(.05)	**.74(.12)**

chats will enable us to detect sexual predation, without further needs of behavioral features.

D. Rule-Breaking Features

Next, we tested our hypothesis that rule-breaking features (that include particular behaviors to avoid being caught and the prohibited word list) are as effective as BoW features. As shown in Table X, the algorithms performed on W15 produced the best result using the k-NN model (accuracy $= 0.91$, $F1 = 0.75$, $F0.5 = 0.80$). A comparison of this result and the best result using all the features reveals that rule-breaking features were useful as much as all the features were. We observed only slight decreases of 0.01 in accuracy, 0.03 in $F1$, and 0.06 in $F0.5$ from the best results when using all the features. The most striking result emerged from the comparison between the results using rule-breaking features and those using BoW features. By using only nine rule-breaking features, the algorithms on W15 showed performances higher than those by using 442 BoW features with increases of 0.03 of accuracy, 0.12 of $F1$, and 0.08 of $F0.5$. This finding strongly supports our hypothesis regarding rule-breaking features that learning the predators' behavior would be useful for predator detection. We also found that for HP15, using rule-breaking features produced worse results than those using BoW features. This implies that lexical features play a more important role than behavioral features do, when the chat corpus contains high level of predatory wordings.

E. Sentiment Features

We examined sentiment features to test if they can help other features detect sexual predation. Adding sentiment features to rule-breaking features lead to little gain in performance for W15 (difference of 0.001 in accuracy, 0.02 in $F1$). On the contrary, this addition resulted in a trivial decrease for HP15 (differences of 0.003 in accuracy, 0.03 in $F1$, 0.07 in $F0.5$). Therefore, we concluded that the type of sentiment features that we created in this work did not contribute to performance improvement.

F. Blacklist and Alert List Features

Additionally, we tested the effectiveness of using the blacklist combined with the alert list, as these lists are practically used by *MovieStarPlanet*. As shown in Table XII, the best performance for W15 was accuracy $= 0.85$, $.F1 = 0.46$, and $F0.5 = 0.63$. The best result for HP15 was 0.86 with $F1 = 0.57$ and $F0.5 = 0.68$.

It is noted that algorithms performed worse on W15 than HP15 within a small margin. This suggests that blacklist and alert list work well when the corpus contains high level of predatory wordings.

TABLE XII
RESULTS OF USING THE COMBINATION OF BLACKLIST AND ALERT LIST FEATURES IN ACCURACY IN P (ACC.), F1 MEASURE IN P (F1), F0.5 MEASURE IN P (F0.5), PRECISION IN P (P), AND RECALL IN P (R), FOR EACH DATA SET, FOR THE ALGORITHM MENTIONED. STANDARD DEVIATION VALUES (SD) ARE PROVIDED IN PARENTHESES. THE BEST PERFORMANCE IN EACH DATA SET IS SHOWN IN BOLD

Data	Alg.	Acc.	$F1$	$F0.5$	P	R
HP15	NB	**.86**	**.57**	**.68**	.78	**.45**
	DT	**.86**	**.57**	**.68**	.78	**.45**
	LR	.85	.43	.60	**.82**	.29
	kNN	**.86**	**.57**	**.68**	.78	**.45**
	SVM	.84	**.57**	**.68**	.78	**.45**
	MLP	**.86**	**.57**	**.68**	.78	**.45**
W15	NB	**.85(1.69)**	.43(.11)	.61(.07)	.82(.06)	.30(.11)
	DT	**.85(1.64)**	.43(.11)	.62(.06)	.84(.06)	.30(.11)
	LR	**.85(2.27)**	**.46(.14)**	**.63(.08)**	.83(.07)	**.33(.13)**
	kNN	**.85(1.83)**	.43(.12)	.61(.07)	.82(.06)	.30(.12)
	SVM	.84(.85)	.32(.05)	.52(.07)	**.91(.09)**	.20(.04)
	MLP	**.85(1.79)**	.42(.12)	.61(.06)	.82(.06)	.30(.12)

G. Discussions

This section summarizes our results, stressing the significant findings. All the results we obtained outperformed the simple method relying on blacklist/alert list. There is an urgent need to employ a data-driven approach in the practical game or chatting system as the basic method is not sufficient to detect sexual predators.

Inspecting Tables VIII–XII reveals that algorithms applied on W15—the data that contains all the predatory chats including subtle expressions—performed exceptionally well in many feature set/algorithm combinations. The best results for W15 was obtained using MLP on all the features ($F1 = 0.78$, $F0.5 = 0.86$, accuracy $= 0.93$).

Considering the characteristics of our corpus which was created by young children and contains lots of slangs and grammatical errors, we believe that these results were promising.

We also discovered that BoW features were useful for the HP15 data set; the best result for HP15 was obtained using MLP on BoW features ($F1 = 0.73$, $F0.5 = 0.73$, accuracy $= 0.90$).

However, considering the poor results on W15 when using BoW features only (see Table IX), these results suggest that BoW features are very effective when the data set contains clear predatory wordings.

The rule-breaking features were found to be as useful as BoW features, especially for W15 ($F1 = 0.75$, $F0.5 = 0.80$, accuracy $= 0.91$). This finding needs to be highlighted as using rule-breaking features can save tremendous space and computation time, which are essential for big data processing. On the other hand, the Sentiment features were found to be the least useful.

Finally, our hypothesis regarding L15 was false; this means that the portion of the last 15 segment of chat before being caught was unable to signal whether the speaker was a predator or not.

TABLE XIII
NUMBER OF USERS LABELED AS PREDATORS IN THE UNLABELED DATA SET, BY THE MODELS BASED ON NAIVE BAYES AND BOW + BLACKLIST FEATURES, AND NAIVE BAYES AND BLACKLIST FEATURES ONLY, FOR ALL THREE DATA SUBSETS

	W15	L15	HP15
Naive Bayes and BoW+Blacklist features	325	647	112
Naive Bayes and Blacklist features	68	29	29

VI. TESTING ON UNLABELED DATA

A testing was performed on our unlabeled data set using the combination of the NB algorithm and the BoW and Blacklist features (what we considered the best model for the HP15 data set in our previous informal experimentations which are not included in this paper).

Table XIII reports the number of predators found using each of the best models. It is interesting to note the number of labeled predators in L15 compared with W15 and especially HP15, which seems to indicate a high number of false positives, as 647 predators in a 15-min chatlog seems highly unlikely.

Additional testing on the unlabeled data was also performed using a model which built upon blacklist/alertlist features only, because these results had far fewer predator labels compared to the results on the model above. Then, we asked a *MovieStarPlanet* moderator to investigate these predator cases manually.

This analysis alone, when reported to *MovieStarPlanet*, resulted in 11 users being permanently locked out of the system, as well as two users being added to a watchlist for further investigation. The moderator who collaborated with us held the view that these users should have been detected immediately at the time of the incidents. It is interesting that the results were very helpful even when using blacklist/alertlist features only, which leads us to the need for further collaboration to build more sophisticated models.

VII. TESTING ON THE PAN2012 DATA

Finally, to test the generalization of our approach, we conducted evaluations on the PAN2012 data. We obtained the training and the test corpus from the PAN2012 site [30] and prepared the three data sets by segmenting the chat into 15-min intervals: W15, HP15, and L15. For HP15, we handpicked the 15-min portion that contains the worst sexual discourse from each predator.

We ran the fivefold cross validation on the training set and obtained the best accuracy of nearly 93% ($F1 = 0.78$) for the W15 set when all the features and a Multilayerperception algorithm was used. Considering the fact that the PAN2012 data set includes consensual sexual conversations as the major false positive examples, the result is promising. This suggests that our approach is general enough to detect sexual predators in the regular setting domain and not only limited to MSP data. Interestingly, however, we obtained a similar best accuracy (nearly 97%) with a significantly lower $F1$value (0.12) when the learned model was used to predict the test set provided by PAN2012. We believe that the poor $F1$-measure on the test set was partly due to the difference in size between the training and the test sets. Overall, the test set size is more than two times of the training set size. If a fair amount of the test data set contain behaviors that are not present in the training set, this could be disadvantageous to data-driven models.

VIII. Conclusion

This paper presents a data-driven, text classification approach to detect sexual predators using real chat data, which was provided by the game company *MoviStarPlanet*. We created data subsets for testing different preprocessing strategies and features unique to *MovieStarPlanet*: bag of words, sentiment features, and rule-breaking features. Rule-breaking features are designed to capture behaviors intended to avoid typing forbidden words. To fully exploit the bag of words features we used automatic spell checking, which improved the classification accuracy significantly.

Given all of the above together with machine learning classification algorithms, our approach has achieved a classification result with 92.51% accuracy, $F1$measure of 0.78, and $F0.5$measure of 0.86 when a multilayer perceptron algorithm was applied on all the features. Similar approaches have been presented for PAN2012 [3], [23]. However, we believe that this study is the first work that demonstrated the feasibility of the text classification approach for detecting predators in a real game chat corpus.

We also tested several hypotheses that we set out in the initial stage of this paper. First, our experimental studies have shown that BoW representation is useful for predicting predators when the data contains predatory wordings. This result reinforces the previous work [7], [16], [19], [24], which successfully distinguished predators from nonpredators using supervised classifiers in combination with simple BoW representation

Second, we discovered that behavioral features were as useful as BoW representation, especially when the data contains less severe predatory language. The use of nine RB features reduces the number of features greatly, and, therefore, will be extremely useful for processing big data. These features are more robust than a simple blacklist function, and can cross over into other NLP areas and games where this type of behavior is common. Finally, our results revealed that the last part of a chat was insignificant for detecting predators in our data set.

To continue our future work, it would be essential to collect more sexual predator data. Particularly, we are interested in validating the finding regarding rule-breaking features with different, larger corpora. In addition, current features could also be improved, starting with a better spell checker that would improve the BoW features. We plan to add contextual features, taking into account information about the recipient of the message and whether it was sent in a private, public or other setting. It would also be interesting to integrate our approach into the *MovieStarPlanet* game system, where a history of some users is directly accessible.

Acknowledgment

The authors would like to thank H. Del Angel Bernal, K. B. Lange, and J. B. Jensen for the test on the PAN2012 data. They would especially like to thank P. Jakobsen, J. H. Michaelsen, and the moderators at *MovieStarPlanet* for providing the data and the support for this study.

References

[1] A. Kontostathis, L. Edwards, and A. Leatherman, *Text Mining and Cybercrime*. New York, NY, USA: Wiley, 2010, pp. 149–164.

[2] L. Sonia, H. Leslie, G. Anke, and Ó. Kjartan, "Risks and safety on the Internet: The perspective of European children: Full findings and policy implications from the EU Kids online survey of 9–16 year olds and their parents in 25 countries," *EU Kids Online, Deliverable D4. EU Kids Online Netw.*, 2011.

[3] C. Morris, "Identifying online sexual predators by SVM classification with lexical and behavioral features," M.S. thesis, Dept. Comput. Sci., Univ. Toronto,, Toronto,, ON, Canada, 2013.

[4] D. Bogdanova, P. Rosso, and T. Solorio, "On the impact of sentiment and emotion based features in detecting online sexual predators," in *Proc. 3rd Workshop Comput. Approaches Subjectiv. Sentiment Anal.*, Stroudsburg, PA, USA, 2012, pp. 110–118.

[5] May 2014 [Online]. Available: http://www.moviestarplanet.dk/

[6] E. R. Guðnadóttir *et al.*, "Detecting predatory behaviour in online game chats," in *Proc. 2nd Workshop Games NLP (GAMNLP-13) Int. Conf. Interact. Digit. Storytelling (ICIDS)*, Istanbul, Turkey, Nov. 2013.

[7] N. Pendar, "Toward spotting the pedophile telling victim from predator in text chats," in *Proc. Int. Conf. Semantic Comput.*, Washington, DC, USA, 2007, pp. 235–241.

[8] G. Inches and F. Crestani, "Overview of the international sexual predator identification competition at PAN-2012," in *Proc. CLEF (Online Working Notes/Labs/Workshop)*, 2012.

[9] May 2014 [Online]. Available: http://www.perverted-justice.com/

[10] May 2014 [Online]. Available: http://www.irclog.org/ and http://krijn-hoetmer.nl/irc-logs/

[11] May 2014 [Online]. Available: http://www.omegle.com/

[12] J. M. G. Hidalgo and A. A. C. Díaz, "Combining predation heuristics and chat-like features in sexual predator identification," in *Proc. CLEF (Online Working Notes/Labs/Workshop)*, 2012.

[13] P. Hohenhaus, "Elements of traditional and reverse purism in relation to computer-mediated communication," *Linguistic Purism in the Germanic Lang.*, vol. 75, p. 204, 2005.

[14] Y. Yan, "World wide web and the formation of the Chinese and English Internet slang union," *Comput.-Assisted Foreign Lang. Educ.*, vol. 1, p. 005, 2006.

[15] A. Kontostathis, A. Garron, K. Reynolds, W. West, and L. Edwards, "Identifying predators using ChatCoder 2.0," in *Proc. CLEF (Online Working Notes/Labs/Workshop)*, 2012.

[16] C. Peersman, F. Vaassen, V. Van Asch, and W. Daelemans, "Conversation level constraints on pedophile detection in chat rooms," in *Proc. CLEF (Online Working Notes/Labs/Workshop)*, 2012.

[17] I. McGhee *et al.*, "Learning to identify Internet sexual predation," *Int. J. Electron. Commerce*, vol. 15, no. 3, pp. 103–122, 2011.

[18] J. Parapar, D. E. Losada, and A. Barreiro, "A learning-based approach for the identification of sexual predators in chat logs," in *Proc. CLEF (Online Working Notes/Labs/Workshop)*, 2012.

[19] E. Villatoro-Tello, A. Juárez-González, H. J. Escalante, M. Montes-y Gómez, and L. V. Pineda, "A two-step approach for effective detection of misbehaving users in chats," in *Proc. CLEF (Online Working Notes/Labs/Workshop)*, 2012.

[20] L. N. Olson, J. L. Daggs, B. L. Ellevold, and T. K. Rogers, "Entrapping the innocent: Toward a theory of child sexual predators' luring communication," *Commun. Theory*, vol. 17, no. 3, pp. 231–251, 2007.

[21] A. Leatherman, "Luring language and virtual victims: Coding cyberpredators' on-line communicative behavior," Senior honor's thesis, Media Commun. Studies, Ursinus College, Collegeville, PA, USA, 2009.

[22] J. W. Pennebaker, C. K. Chung, M. Ireland, A. Gonzales, and R. J. Booth, "The development and psychometric properties of LIWC2007," LIWC Inc., Austin, TX, USA [Online]. Available: http://www.liwc.net/LIWC2007LanguageManual.pdf

[23] C. Morris and G. Hirst, "Identifying sexual predators by SVM classification with lexical and behavioral features," in *Proc. CLEF (Online Working Notes/Labs/Workshop)*, 2012.

[24] C. Peersman, W. Daelemans, and L. Van Vaerenbergh, "Predicting age and gender in online social networks," in *Proc. 3rd Int. Workshop Search Mining User-Generated Contents*, 2011, pp. 37–44.

[25] G. Eriksson and J. Karlgren, "Features for modelling characteristics of conversations," in *Proc. CLEF (Online Working Notes/Labs/Workshop)*, 2012.

[26] M. Hall *et al.*, “The Weka data mining software: An update,” *SIGKDD Explor. Newsl.*, vol. 11, no. 1, pp. 10–18, Nov. 2009.

[27] N. V. Chawla, “Data mining for imbalanced datasets: An overview,” in *Data Mining and Knowledge Discovery Handbook*. New York, NY, USA: Springer-Verlag, 2005, pp. 853–867.

[28] “Jazzy automatic spell checking API,” May 2014 [Online]. Available: http://moderntone.blogspot.dk/2013/02/tutorial-on-jazzy-spell-checker.html

[29] F. Å. Nielsen, “A new ANEW: Evaluation of a word list for sentiment analysis in microblogs,” ArXiv, 2011 [Online]. Available: arXiv:1103.2903

[30] May 2014 [Online]. Available: http://www.uni-weimar.de/medien/webis/research/events/pan-12/pan12-web/authorship.html

Yun-Gyung Cheong received the B.S. degree in 1996 and the M.S. degree in 1998 in information engineering from Sungkyunkwan University (SKKU). In 2007, she received the Ph.D. degree in computer science from North Carolina State University, Raleigh, NC, USA.

She is an Assistant Professor at Sungkyunkwan University, Korea. Her research interests lie in artificial intelligence with emphasis on its use in discourse planning for narrative, games, and user interfaces. Before joining SKKU, she was a Postdoctoral Fellow at the Center for Computer Games Research at the IT University of Copenhagen and a researcher at Samsung Advanced Institute of Technology..

Alaina K. Jensen received the B.A. degree in English from Brigham Young University, Provo, UT, USA, in 2007 and the M.Sc. degree in software development and technology from the IT University of Copenhagen, Denmark, in 2013.

Her previous work experience includes Maersk Line Finance IT, where she was involved in the development of software for financial reporting. She is currently on sabbatical for family reasons. Her research interests include NLP and interactive web design.

Elín Rut Guðnadóttir received the B.A. degree in psychology from the University of Iceland in 2004 and the M.Sc. degree in software development and technology from the IT University of Copenhagen, Denmark, in 2013.

She is currently working in the IT Department, Financial Supervisory Authority Iceland, where she is implementing technical standards on supervisory reporting. Her main interest lies in data mining and revealing hidden facts about interesting data sets.

Byung-Chull Bae received the B.S. degree in 1993 and the M.S. degree in 1998 degree in electronics engineering from Korea University, Korea, and the Ph.D. degree in computer science from North Carolina State University, Raleigh, NC, USA, in 2009.

He is an Assistant Professor at School of Games, Hongik University, Korea. He has worked at LG Electronics and Samsung Electronics as a research engineer, and worked for IT University of Copenhagen, Denmark, as a visiting scholar and a lecturer. His research interests include interactive storytelling and affective computing.

Julian Togelius received the B.A. degree from Lund University, the M.Sc. degree from the University of Sussex, U.K., and the Ph.D. degree from the University of Essex, U.K.

He is an Associate Professor in the Department of Computer Science Engineering, New York University, New York, USA. He works on all aspects of computational intelligence and games and on selected topics in evolutionary computation and evolutionary reinforcement learning. His current main research directions involve search-based procedural content generation in games, game adaptation through player modeling, automatic game design, and fair and relevant benchmarking of game AI through competitions. He has previously worked at IDSIA in Lugano and at the IT University of Copenhagen.

Dr. Togelius is a past chair of the IEEE CIS Technical Committee on Games, and an Associate Editor of the IEEE TRANSACTIONS ON COMPUTATIONAL INTELLIGENCE AND AI IN GAMES.