Privacy Protection Protocol in Social Networks Based on Sexual Predators Detection

Zeineb Dhouioui
BESTMOD Laboratry
ISG Tunis
University of Tunis
dhouioui.zeineb@hotmail.fr

Jalel Akaichi
BESTMOD Laboratry
King Khaled University
Guraiger, Abha, KSA
jalel.akaichi@kku.edu.sa

ABSTRACT

Social networks offer great opportunities for communication and no one can deny the benefits of these sites; however, we cannot ignore the privacy defects such as fraudsters, spam, hackers and sexual predators. This paper presents a privacy protection protocol in order to protect users from risks that may be encountered in their social networks activities. In particular, we focus on the sexual predators detection. For this purpose , we use text mining tools to classify doubtful conversations based on lexical and behavioral features extraction. Moreover, we are able to flag potential predators, by computing the predatorhood score according to the sum of features weights. Different experiments have been carried out based on comparative study between two machine learning algorithms: support vector machines (SVM) and NaÃŕve Bayes (NB). The results are very promising.

Keywords

Community Detection; Privacy protection; security; social networks; predators; text mining; machine learning

1. INTRODUCTION

Recently and with the increasing popularity of social networks, those sites witness many risks that threat users privacy. Although the widespread use of social networks, these later have become an attractive ground for false and malicious users. Online social networks witness some irregular and illegal anomalies which can be suspicious individuals such as spammers, sexual predators and online fraudsters. Indeed, cybercrime is a major problem that requires a better understanding to be predicted and consequently deeply proposing preventive solutions. Thus, we propose a protocol to control users privacy.

Among illegal anomalies, we commonly find that sexual predators and their victims are generally children or teenagers. Usually, we use the term grooming attack defined by [11] as a communication process by which a perpetrator applies affinity seeking strategies, while simultaneously acquiring information about sexually desensitizing targeted victims in order to develop relationships that match what perpetrator want (e.g. physical sexual molestation). Due to the huge number of possible messages, traditional approaches proposed to detect online sexual predators are ineffective and powerless.

With the privacy issues, conversations content is always inaccessible. Another challenge in detecting online social networks predators is that their behaviors change too fast and regularly. Hiding the true identity (name, age, gender and location) since it is possible to provide false personal information (false profiles), makes the identification of predators difficult. Moreover, this task is challenging because chat data are specific.

In this paper, we introduce a method allowing sexual predators identification based on behavioral and lexical features. This method can handle different languages of conversations texts. Additionally, it guarantees to flag potential predators according to the predatorhood score.

This paper is organized as follows: in section 2, we depict the state of the art online predators detection in social networks; in section 3, we describe the proposed architecture; in section 4, we illustrate experiments and discuss and evaluate the results. Finally, in section 5, we conclude and propose possible directions for future works.

2. STATE OF THE ART

Recently [12] and with the increasing popularity of social networks, those sites witness sex crimes since offenders seek to develop relationships or to seduce underage teenagers: cyber pedophilia. Authors list some risk behavior:

- Posting personal online information
- Interacting with unknown people
- Using the Internet to make rude and nasty comments to others
- Sending personal information to unknown people met online
- Downloading images from file-sharing programs
- Visiting X-rated sites on purpose
- Using the Internet to embarrass or harass youths people
- Talking online to unknown people about sex

ICC '16, March 22-23, 2016, Cambridge, United Kingdom

DOI: http://dx.doi.org/10.1145/2896387.2896448

In [1], a two steps approach has been proposed to identify sexual predators based on social network dialogues. The first step is devoted to detect suspicious conversations i.e. in which predators participate; the second consists in identifying the sexual predators conversations from over all users dialogues. According to experiments, Random forests give the best results for the first step, while neural networks are effective in the second step. The proposed methodology for identifying sexual predator conversations runs based on two classification processes as follows:

- The first classification process ensures detecting conversations of sexual predators
- The second classification process guarantees selecting predators dialogues

For the pre-processing step, authors create and employ 3 lexical resources as dictionaries:

- Emoticons: : -) is considered as happy
- Contractions: isnt́ is normalized as is not
- SMS vocabulary: 4u = for you

After the pre-processing step features are extracted with the POS-tagger. This paper [4] treats sexual predator detection in chat conversation using sequences of classifiers. In fact, the idea is to divide documents into three parts referring to different stages of predation. Rahman Miah and his colleague propose a method to distinguish between child-exploitation, adult-adult and general-chatting dialogues based on text categorization approach and psychometric information [10]. Furthermore, authors in [2] have affirmed that standard text-mining features are suitable to discriminating general-chatting from child-exploitation conversations, but are inadequate to differentiate between child exploitation and adult-adult conversations . Recently, Michalopoulos and Mavridis proposed in [7] a probabilistic method which offers three classes of the chat interventions:

- Gaining Access: refers to the intention of predators to acquire the victim approach.
- Deceptive Relationship: designates the deceptive relationship that the predator attempts to create with the child, and is considered as preliminary to a sexual attack.
- Sexual Affair: obviously, refers to a sexual affair intention of the predator with the victim.

Undoubtedly, the previous categories represent stages that a sexual offender follows to approach minors. Each user (victim or sexual predator) is modeled according to their set of interventions; hence only one conversation is generated for each user saving the order of interventions. Then, those conversations are classified into sexual predator or the rest type of user. Conversations are classified into 3 segments based on local classifiers which are inspired from chain based classifiers. Sentiment in texts can be useful to detect online sexual predators. In this paper [3], a list of high level features has been proposed, containing sentiment features. Authors employed a corpus including predators chats obtained from http://www.perverted-justice.com as well as two negative datasets.They treat predators detection task using the natural language processing (NLP) techniques. This task is hard due to specificity of the chat data since online chatting involves very fast typing. Moreover, chat data are characterized by huge amount of mistakes, misspellings, specific slang.Some sexual features proposed in [6] are used in this work such as:

- Percentage of approach words: this set of words contain verbs like meet, and nouns such as car and hotel
- Percentage of relationship words: including dating words
- Percentage of communicative desensitization words: these words refer to family members names
- Percentage of words expressing sharing information: giving basic information such as age, location and sending photos
- The imperative sentences imply tendencies to be dominant.

Based on the analysis of the conversations published at www.perverted-justice.com, several characteristics of predators language have been exposed [3]:

- Fixated discourse: predators force the minor to discuss on a sex-related topic
- Implicit/explicit content: starting with ordinary compliments, predators reach sexual conversation
- Implicit/explicit content: starting with ordinary compliments, predators reach sexual conversation
- Offenders know the immorality of their acts; however, they hold responsibility for the victim.
- Predators habitually act as children
- To minimize the risk of being discovered and prosecuted, they require deleting chat logs, mainly advising victims not to tell anyone, and having offline meeting.

This paper [6] aims at evaluating and classifying sexual predators strategies to establish relationships with kids via internet. The Chat Coder determines lines in the chat log that contains predatory language. Labels used in this work arise from the communicative model which is described in this work. Authors represent a description of the rule-based approach which identifies predatory communication in chat log. Generally, the predator starts grooming the child victim when this latter trusts him. Deceptive trust is manifested in the whole communication that congregates the predator and the victim. Certainly, winning the victim trust is fundamental to achieve next level of "the entrapment cycle" such as isolation and approach. Grooming is defined as a strategy used by sexual abusers to push their victims to accept the sexual conduct. Grooming is categorized into communicative desensitization and reframing.

- Communicative desensitization: denotes the use of vulgar sexual language by the offender with the intention to desensitize the victim. Indeed, to achieve this in online predation, pornographic images are sent and sexual slang terms or netspeak are employed instead of daily words for example welcum instead of welcome.

- Reframing: refers to the attempt of sex offenders to reassure the victim by experiencing online sexual advances. It is defined as contact or sex play between victim and predator that may be communicated in ways that would make it beneficial to the victim later in life.

In addition to grooming, the victim must be isolated physically and emotionally. Physical isolation refers to the time that the predator spends alone with the victim, and mental isolation is defined as a growing emotional dependency on the predator (i.e friendship and guidance). However, complete physical isolation is not achieved virtually i.e. over the Internet, the predator reach isolation if the victim chats without supervision. Moreover, predation is more effective with minors who are isolated that is means if they have weak paternal or maternal relationships or if they have a limited number of friends. This information is deduced by asking questions about social life of the minor, or by providing excessive sympathy and support. In order to facilitate abuse and win victim control for more exploitation, the predator tries to isolate the victim. Once the victim trust is reached and grooming begins, the predator seeks to approach the victim by proposing meetings for sexual purposes which is the final step.
Online conversations or instant chat present a daily activity of the majority of people. Despite the benefits of these tools, they rather attract cybercriminal acts [5]. Commonly, features extraction is used in most research. We find lexical and behavioral features. Indeed, lexical features such as unigram and bigram can be obtained from the preliminary conversation text, and also weighting using TF-IDF or the cosine similarity and emoticons counting. These features can be derived using the LIWC tool. Commonly, stemming or stop-word elimination is not applied with lexical features in order to conserve author style with misspelling and grammatical errors. Behavioral features ensure detecting user actions within a conversation; for instance, it includes the number of times a user initiates a conversation, the number of questions asked, the frequency of turn-taking, and eventually capture intention of grooming or hooking. Usually, only one set of features: one of the most common for each author is created in order to describe him and consequently his predator potential will be exploited. Other researchers propose the Language Model (LM) of a single author, in addition to the LM of the two participants in the chat. Classification approaches include SVM, neural networks, maximum entropy, decision trees, KNN and NaÃŕve Bayes. Social networks pedophiles groom underage victims with the intention of attracting and engaging them in sexually explicit texts or video conversations, if not arranging real meetings [9]. Consequently, it is crucial to protect adolescents by detecting doubtful conversations automatically. Online text chat that contains sexually explicit content is categorized into two types:interaction between predator and victim and the consensual interaction between two adults The first type comprises sub-types according to Pendar [9]:

- predator/ other:
  - predator/victim: in this case the victim is underage
  - predator/pseudo-victim: child but volunteer
  - predator/pseudo-victim: law enforcement officer pretends to be a child
- adult/adult called also consensual relationship

In this work [9], Pendar used a set of SVM and distance-weighted k-NN classifiers. The author did not apply the words stemming the words. Moreover, he extracted the n-grams after deleting all the stop words. Subsequently, feature extraction is performed. Two main characteristics are attributed in [8] to define predatorhood in this paper:

- Age disparity: Typically, a predator is an adult communicating with an underage social network user. Victims are usually adolescents.
- Inappropriate intimacy: the predator introduces an intimate conversation generally a sexual one

3. 3P (PRIVACY PROTECTION PROTOCOL) FOR SOCIAL NETWORKS

Given a social network $G(E, V)$ and a member U_i that looks for a high safety level. U_i verifies the degree of the privacy protection by running the following protocol: U_i executes the selection function which guarantees to find trusted users U_i executes the control function which serves to manage privacy and personal information such as: Photos, status, Location sharing. An automatic process is executed called anomalies detection including the identification of illegal behavior and malicious individuals like: spammers and online fraudsters; in our work, we will focus on sexual predators identification. In this paper, we explore the use of text mining techniques to detect suspect conversation and to detect predators in social networks. Through the lexical and behavioral features, the nature of conversations is determined. A weight is manually assigned to each feature. The value of these weights will be summed in the last step in order to determine the potential predators. In this section, we propose a general architecture 3P for social networks: Privacy Protection Protocol, which will be illustrated in figure 1.
Step 1: raw data collection: this step consists of gathering conversations from Facebook. Firstly, we annotated manually into positive and negative conversation. This is will be used in order to train a classifier to distinguish malicious users.
Step 2: Features classification: two kinds of features are proposed in this paper; behavioral and lexical.
Step 3: Lexicon development: this step focuses on the informal language of online social networks. For this reason, 3 types of lexicon were created:

- Exchange of personal information
- Grooming
- Approach

Step 4: Conversation classification: based on features we can classify conversation into two classes positive and negative.
Step 5: Flagging users according to predator degree by summing the features weights.

The main challenging task is to obtain real conversations and how to handle these ones with 3 languages: Arabic, French, and English. Grooming stages are as follows:

- Relationship (friendship forming)

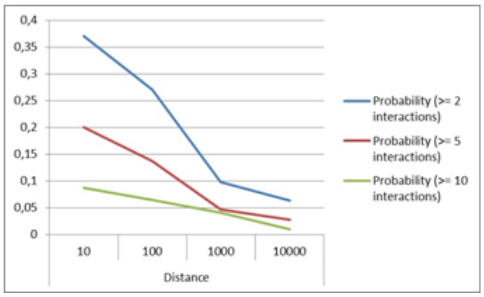

Figure 1: Overall framework of sexual predators identification approach

Table 1: Behavioral and lexical features

Behavioral features	Lexical feature
The number of times a user initiates a conversation	Percentage of approach words
The number of questions asked	Percentage of relationship words
The number of intimate conversations	Percentage of communicative desensitization words: these words refer to family members names, for instance, (mom, dadâĂę.)
The frequency of turn-taking	Percentage of words expressing sharing information
Intention of grooming or hooking	Percentage of isolation
	Number of emoticons

- Risk assessment and exclusivity
- Sexual affair

We detailed the proposed method as follows, indeed, in this work, there is no preprocessing stage of conversations texts due to the special characteristics of these conversation texts such as neglecting grammar rules, using abbreviations and emotions. For instances hug (< :d>) and kiss (: - *) can reflect an introduction of sexual stage. But we will generate n-grams. However, pre-filtering is crucial to reduce the computational task by eliminating conversations that contains only one participant or very short ones (less than 10 interventions for both users), or those containing many and several unrecognized characters. Then, we proceed to features classification. We distinguish two types of features in the following table 1 [2]:
Conversation can be finally classified according to the presence of features. Indeed, if the conversation contains the intention of sexual harassment. The last step involves on the identification of potential predators based on the sum of features weights.

4. EXPERIMENTS

4.1 Metrics

In order to compare the performance of the proposed classification method, evaluation has been carried out on the following metrics used in classification studies:the accuracy, the precision, the recall and the F-measure.

$$Accuracy = \frac{a+d}{a+b+c+d}$$

$$Precision = \frac{a}{a+b}$$

$$Recall = \frac{a}{a+c}$$

$$F-measure = \frac{2\times Precision\times Recall}{Precision+Recall}$$

Where:

- a: the number of conversations correctly assigned to this class
- b: the number of conversations incorrectly assigned to this class
- c: the number of conversations incorrectly rejected to this class
- d: the number of conversations correctly rejected to this class

We compared the performances of different feature sets using the most famous and used learning algorithms: NaÃŕve Bayes and SVM classifiers.

4.2 Data

The data was gathered by collecting some real conversations since online harassment is a taboo issue. This phase was the most difficult because internet users refuse to give conversations. Fortunately, we could recover some conversations of our friends and we are hopeful to collect more real conversations in Arabic language for our next works. 20 conversations are collected, 15 among them will be used for the training model and 5 for the cross validation. As we said previously, there is no text pre-processing task. However, N-Grams are generated to analyze its influence on the generated data.

4.3 Results and discussions

The features extraction is the process of extracting the main characteristics of the text. In our context, we need to be able to identify the words that express misbehavior. The training model contains 9 conversations of sexual predators and 6 of non-sexual predators classified manually. In tables 2 and 3, weights are assigned manually to the behavioral features, the frequency of these features in the collected conversations is then computed. Since we handle a classification task with two classes and to highlight the effectiveness of our method, we can use two simple machine learning techniques which are the SVM and the naive bayes NB for the cross validation. We can conclude clearly that the SVM algorithm outperformed the Naive Bayes in all cases 2. For SVM, a conversation has positive classification is 82.2% likely to be correct. However, it gives negative classification is only 62.2 % likely to be correct.
This is clearly visible in the F-measure rates: 75.2% for Positive and 71.57% for Negative. This is due the fact that the model in this case study was built with many more malicious conversations. This would be surprising due to the rough privacy settings proposed by social networks developers. We collected unlabeled conversations. Results in figure 3show that 30% of conversations are misclassified. This misclassification is due to ambiguity in the Arabic language.

Table 2: The assigned weights and frequencies of the behavioral features

Lexical features	Assigned weights	Frequency
The number of times a user initiates a conversation	4	85%
The number of questions asked	3	90%
The number of intimate conversations	5	65%
The frequency of turn-taking	3	40%
Intention of grooming or hooking	5	62%

Table 3: The assigned weights and frequencies of the lexical features

Lexical features	Assigned weights	Frequency
Percentage of approach words	5	60%
Percentage of relationship words	4	75%
Percentage of communicative desensitization words: these words refer to family membersâĂŹ names	1	30%
Percentage of words expressing sharing information	2	81%
Percentage of isolation	3	76%
Number of emoticons	3	80%

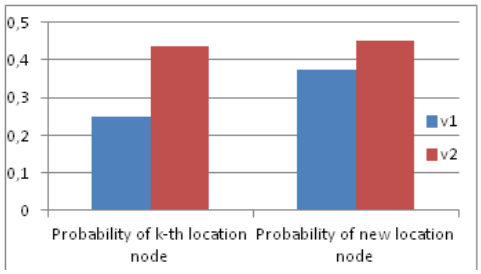

Figure 2: Comparison between NB and SVM

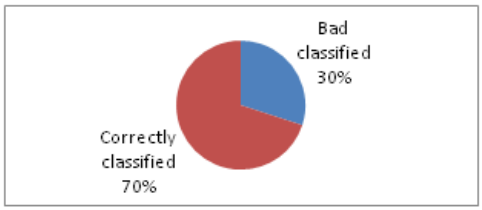

Figure 3: Classification of unlabeled data

5. CONCLUSION

This paper offers a protocol protecting users privacy in social networks based especially on sexual predator detection in instant conversation using features classification. Our protocol is useful for several networks. The main advantage of the proposed method is that we can handle different languages of conversations texts. We build a supervised classifier for detecting potential sexual predators from online social networks conversations. As features we used behavioral and lexical terms extracted from texts. The results shows that SVM classifier has a better performance than naive bayes in terms of various metrics such as precision, recall and the F-measure. It would be very interesting to include additional features and use the largest number of conversations. As ongoing work, we will generate a new corpus for determining suspicious conversation by adding linguistic features.

6. REFERENCES

[1] Y. Alemán, D. Vilarino, and D. Pinto. Searching sexual predators in social networks.

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

[3] D. Bogdanova, P. Rosso, and T. Solorio. 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*, pages 110–118. Association for Computational Linguistics, 2012.

[4] H. J. Escalante, I. LabTL, L. E. E. No, E. ú Villatoro-Tello, A. Juárez, and M. Montes-y Gómez. Sexual predator detection in chats with chained classifiers. *WASSA 2013*, page 46, 2013.

[5] G. Inches and F. Crestani. Overview of the international sexual predator identification competition at pan-2012. In *CLEF (Online Working Notes/Labs/Workshop)*, volume 30, 2012.

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

[7] D. Michalopoulos and I. Mavridis. Utilizing document classification for grooming attack recognition. In *Computers and Communications (ISCC), 2011 IEEE Symposium on*, pages 864–869. IEEE, 2011.

[8] C. Morris. *Identifying online sexual predators by svm classification with lexical and behavioral features*. PhD thesis, MasterâĂŹs thesis. Department of Computer Science, University of Toronto, Toronto, Ontario, Canada, 2013.

[9] N. Pendar. Toward spotting the pedophile telling victim from predator in text chats. In *null*, pages 235–241. IEEE, 2007.

[10] M. W. RahmanMiah, J. Yearwood, and S. Kulkarni. Detection of child exploiting chats from a mixed chat dataset as text classification task. In *Proceedings of the Australian Language Technology Association Workshop*, pages 157–165, 2011.

[11] 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 *CLEF (Online Working Notes/Labs/Workshop)*, 2012.

[12] J. Wolak, D. Finkelhor, K. J. Mitchell, and M. L. Ybarra. Online" predators" and their victims: myths, realities, and implications for prevention and treatment. *American Psychologist*, 63(2):111, 2008.