Mitigating Cybercrime and Online Sexual Grooming Of Minors On Social Media Using Machine Learning: A Desktop Survey

Abstract— Cyber threats such as identity deception, cyber bullying, identity theft and online sexual grooming have been witnessed on social media. These threats are disturbing to the society at large. Even more so to minors who are exposed to the Internet and might not even be aware of these threats. This paper describes a brief overview of different developments on cybersecurity methodologies that have been implemented to ensure safety of minors on social media, particularly; online sexual grooming. A desktop survey on machine learning technologies that have used to detect online grooming is presented in this paper. The aim is to consolidate most of the work done in the past by scholars in this area of research, in order to give insights on various algorithms that have been proposed and the reported performance results.

Keywords—Sexual predatory; online sexual grooming; pedophile; cyberpodophilia

I. INTRODUCTION

Social media has become hugely popular these days, with different platforms at our fingertips, including Facebook, MySpace, Whatsapp and online gaming. Frequent use of these platforms by children, such as online gaming has been increasing, Kontostathis et al. [1]. Unfortunately, social media like any other e-communication medium contains cyber threats, Hugo et al., [2], such as cyber bullying, information theft, identity deception and online sexual grooming. Consequently, these threats can endanger children who are exposed to Internet. Consider, for an example, the problem of online sexual grooming, where an adult in this case; pedophile (a person with sexual interest in children) uses social media and randomly selects a minor with an intention to deceive or groom for sexual gain.

Child grooming or sexual grooming: is defined by Harms [3], as "*a communication process by which a perpetrator applies affinity seeking strategies, while simultaneously engaging in sexual desensitization and information acquisition about targeted victims in order to develop relationships that result in need fulfillment"* such as physical sexual solicitation. As such, the term pedophile or sexual predator is used to describe such people and these terms are often used interchangeably.

Girouard [4] found that attempts to solicit children in the cyberspace have been common. He further reports that one in seven kids (ages 9-17 years old) have been sexually approached online. A recent report by a South African newspaper, Beeld [5], points out that a similar problem has been found in the Singaporean study. Where they indicate that the most affected children range from the age of 12 years. This is certainly a critical issue to the society especially that most of sexually abused children have voluntarily agreed to meet their abuser [6]. Therefore, having automated ways to identify or detect a person attempting to approach a child with malicious intents can proactively protect children from being physically or sexually abused.

The aim of this paper therefore is; to give a brief overview of machine leaning technologies and algorithms that have been employed towards solving this problem. Also, to give a summary of findings from researchers in terms of how these techniques perform in detecting or identifying online sexual grooming.

The sections presented are as follows: section II gives a detailed literature review. Section III is a summary that analyses the work presented in section II. Section IV presents the conclusion and recommended future work.

II. LITERATURE REVIEW

This section is divided into subsections which are organized according to different text categorization features used by different scholars. The datasets used, algorithms applied and their performance results are also presented.

A. Lexical Features

The work done by Pendar [10], aims to identify online pedophile by flagging suspicious chat interactions from a collection of online chat logs. The main objective of his work is to assess the feasibility of machine learning classifiers towards automated recognition of online sexual predators.

1) Dataset: Pervert Justice (PJ) dataset was used. Perverted Justice[1] is a non-profit organization that initiated a sting operation to capture online predators. They trained police officers and volunteers to pose as minors online in order to attract and capture online sexual predators. This organization collected chat logs of previously convicted sexual predators and made them publicly available for use. Pendar splits this data into two documents: pedophiles only chats and victims only chats. Thus, distinguishing predators

[1] www.pervertedjustice.org

from victims. He argues that given a chat produced by a pedophile; a classifier should be able to accurately categorize that chat accordingly, and vice versa when a chat is produced by a victim.

2) Algorithms: He proposes to use two different classifiers namely, Support Vector Machines (SVM) and k-Nearest Neighbours (k-NN).

3) Experimental results: His experimental results indicates that k-NN models based on trigrams have a high accuracy rate of 94% in detecting pedophiles compared to 90% accuracy of SVM models. Thereby concluding that with these classification models, it is attainable to develop tools that can auto detect online pedophiles. However, he notes that the performance of these models is very low when unigrams and bigrams are used with an average score of 60%.

B. Lexical Features and Luring Communication Theory

Kontostathis and colleagues [11], [12], also embarked on the same domain of study dating back from 2009 to 2012. They have developed a system called ChatCoder which has been evolving from version to version.

The first task in their study, ChartCoder [11], used phrase-matching to generate features and group these phrases into categories.

1) Dataset: for both chatCoder1 and chatCoder2 Kontostathis also utelised PJ dataset and ChatTrack

2) Algorithm: they used key phrase matching and rulebased techniques for chatCoder1 and and decicion trees to validate accuracy of chatCoder2.

Their software was able to distinguish pedophile from no-pedophile communications, and, k-Means to cluster only pedophile conversations and they realized that there were four categories of communication. April used PJ and ChatTracker datasets.

3) Experimental results: Kontostathis reports on 93% accuracy results to distiguish predator vs victim,.

The second task's objective (ChatCoder2) [12], was to analyze, index and categorize communicative strategies used by sexual predators to lure their victims. To advance from previous work, they used machine learning techniques such as J8 tool with different machine learning libraries. On both tasks they were guided by Luring Communication Theory (LCT) model, one of the psychological theories established by Olson [13].

The LCT model states and defines five categories that predators use to attract their victims: gaining access, deceptive trust development, grooming and physical approach. Therefore, Kontostathis and colleagues, apply this model to label sentences in a conversation according to their related LCT phases. Their results indicate that, the system is able to determine predatory sentences.

They report that the system is able to accurately classify non-predatory sentences with a score of approximately 75%. However their work did not focus on how accurate can their system correctly classify sentences based on a specific LCT phases.

Similarly, Gupta et al. [14], also uses LCT model to do an empirical analysis on various pedophile messages to gain insight of how predators behave. They propose to use Linguistic model LICW (Linguistic Inquiry and Word Count), a word count and language analysis software. They used this toll to perform psycho linguist profiles for each of the six phases of LCT as proposed by [12]. Their objective is to unpack useful textual patterns towards understanding of online predatory behavior.

4) Dataset: They used PJ dataset, where they randomly selected 75 conversations. They manually labeled these chats with the help of their psychology specialist.

In their analysis, they report that relationship forming stage is the most prominent stage than any other stage even grooming stage as anyone would expect. While their study is on auto detection of online grooming, they do not provide any automated classification system with regards to the identified stages.

Leveraging on Gupta's and Kontostathis's work, Cano et al. [15], propose to automatically categorize such stages using a binary classification method known as Support Vector Machines. For features, [13] propose to employ various feature representation methods ranging from lexical, semantic, syntactical, discourse patterns and LICW.

5) Dataset: Using a dataset that Kontostathis (chatCoder2) has already labeled. Her prepossessing includes stemming, tagging and changing misspelled and emoticons into English words using a dictionary which she created.

6) Algorithm: She proposes to use SVM for classification. For experimental setting, Cano merges all the stages that are labeled as (trust development, grooming, and physical approach) as positive set and data that is labeled "other" as negative set. Their experimental results show remarkable performance on unigrams and discourse patterns on each stage. They later combine all the features to detect each stage and the results are as follows:

7) Experimental results: Their experimental resuts based on three grooming categories indicates that: Trust Development: Precision is 79%, recall is 82% ; Grooming: Precision is 88% and recall is 88% Approach: precision is 87 and recall is 89%. Their results indicate the feasibility of a system that can indicate how advanced the grooming is using these stages or phases. This suggests that, systems like this can be used as a preventative measure by addressing grooming through intervention before physical meeting.

In their work Hugo et al. [16], propose a different technique, instead of applying traditional classification methods such as SVM or Naïve Bayes, they propose a chained model: a simple sequence of classifiers.

Their objective is to effectively perform predator identification by using a set of local classifiers specialized at classifying certain segments of a document. They divide a document into equal segments which they assume, relates to known different grooming stages. Local classifiers are trained per segmented document, and their resulting outputs are merged using a chain strategy to be an input on the next segment. An example, of a local classifier training given by [17], says, segments s1, s2, s3 will have their corresponding local classifiers c1, c2, c3. However, the classifier c2 is dependent on c1, in that the output of c1 is used as an input for c2. Same applies to c3, where the output c2 becomes an input in this classifier. Eventually, c3 is a classifier that can be used to predict a pedophile at any stage.

8) Algorithm: The chained model was used. They also incorporate a ring-based mechanism, whereby the chained classifier is iterated as many times as possible to further improve performance of the model. Their aim is cover a broader scope of a document, as chained classifiers is used to handle multi-labeled data and to incorporate dependences among different labels.

9) Dataset: PAN12. PAN12 dataset is a combination of PJ (a small proportion), krnjin[2] , irclog[3] , omegle[4] data collection. It was collected by Inches et al., [8], researchers for a data forensic organisation known as CLEF [8].The data is setup in such a way that it articulates a realistic scenario in which predator data is small compared to regular conversations.

10) Experimental results: Their experimental results indicate that the suggested classifier model performs better than global classifier and comparable to the models that were evaluated from PAN12, with 96% precision, 59% recall. Note, by global classifier they refer to a classifier that is built with the entire document labeled as predator instead of segmenting it, such as that of [11].

C. Behavioral Features and Lexical Features

Moris [18], also used behavioral features and lexical features. For behavioral features, he extracted information from the document such as a number of messages an author sent and a total number of conversations an author was involved in.

1) Dataset: Moris used PAN12 dataset.

2) Algorithm: SVM algorithm was used.

3) Experimental Results: He reports on a good on accuracy F1 of approximately 83% on lexical features and a reasonable F0.5 of 56% on behavioral features

Similar to Moris's work, Yun-Gyuna Cheang et al. [19], are also detecting suspicious predatory chat, however using real data, that is, with real victims instead of pseudo-victims. Their goal is to compare the performance of different machine learning algorithms on a real setting environment.

4) Dataset: They used game data from an entity called MovieStarPlannet [5]and PAN12.They created sub-datasets of MovieStarPlannet to perform and test various prepressing techniques, including stemming.

5) Algorithm: they used multi layer perceptron. For feature construction, they use bag of words, sentiment and rule breaking features. In this case, rule braking, is when a user tries to use forbidden words in the game environment. If used, they trigger an alert either to block or warn the user. To avoid these triggers the user would then either misspell the word or shorten the word. Their aim for incorporating rule braking features is to understand the behavior of the predator.

6) Experimental results: Their findings show that rule breaking features are relatively useful. They suggest that these features can be useful and more robust for big data. Their experimental results have an accuracy score of 92% using multi-layer perceptron.

D. Sentiment, Fixated Discourse and Lexical Features

In their work to detect pedophile conversations in chat logs, [20], proposed a fixated discourse, a sequence of related words, bag of words and sentiment and content based features. A Fixated discourse is a pattern in which a predator is unwilling to change the topic and will not allow the victim to divert from or interrupt the subject of the matter [20]. In such cases, the predator will always ignore any interruptions and go back to the topic of interest mainly, sexual-related conversations.

1) Dataset: They used three different datasets, child sexual grooming dataset from Perverted Justice (PJ), cybersex adult- adult[6] logs available online and NPS[7] child-child chat logs. With child PJ dataset as positive dataset and the other two datasets as benign.

2) Algorithm: They used SVM binary classification on this task.

3) Experimental results: Their experimental results indicate that high level features accurately classify pedophiles versus cybersex conversations with a score of 97%, compared to low level features classifying pedophiles from NPS with a score of 94%.

[2] http://krijnhoetmer.nl/irc-logs
[3] http://irc.netsplit.de
[4] http://inportb.com

[5] http://www.moviestarplanet.dk
[6] http://faculty.nps.edu/cmartell/NPSChat.htm
[7] http://oocities.org/urgrl2lf

Recently, Ebrahimi et al., [7, 21], also leveraged on lexical features. The main objective in their work [7] is to address the problem of negative dataset by applying semi supervised anomaly detection technique. They argue that, by using a binary classification to solve this problem might not generalize in a real setting [7]. They note that it not realistic to use negative dataset when training the model because of its diversity and computationally difficult to acquire all categories of negative dataset for a model to learn. Rather, training a model on a positive set only and test it on unlabeled data is more realistic.

4) Dataset: PAN12 dataset was used

5) Algorithim: semi-supervised Support Vector Machine(One-class classification) was used. The model is trained on predatory chats only and tested on unlabbeled data.

6) Experimental results: This model has never been experimented in this domain, and their contribution show good results. Their experimental results, show an accuracy of 98%.

In [21], Ebrahimi and colleagues further exploit deep learning techniques, Convolutional Neural Network (CNN) to experiment auto detection of sexual predator identification using the same the dataset. Their experimental results show good performance of f1 measure of approximately 80%.

After conducting a literature survey and documenting all findings per work; following are the two sections which describe the methodology used and a summary of the results.

III. Experimental Settings and Results

After realizing that Inches [8], has already published a detailed overview of all articles that were featured in the CLEF 2012 conference, we decided to consider only papers that used PAN12 data but not featured in his report, to avoid duplicating his work. In the end, we were left ten articles including various works which are based on Perverted Justice Dataset.

In table 1, a summary report of our findings is described.

TABLE I. SUMMARY OF RESULTS

TABLE II. MACHINE LEARNING CATEGORIES			
Author	Year	Algorithm	Performance Measure (Accuracy or f1)
Pendar [8]	2007	k-NN, SVM	Accuracy: 94%, 90%
Kontostathis et al. [9]	2010	Decision Tree	Accuracy = 93%
McGhee et al. [10]	2011	Rule-based	Accuracy= 75%
Cano et al. [14]	2014	SVM	f1=85%
Escalante et al. [15]	2013	Chain classifier	f1=73%
Moris[16]	2013	SVM	f1=83%
Yun-Gyuna et al. [17]	2015	Multilayer Perceptron	Accuracy =92%
Bogdanova[18]	2014	SVM	Accuracy =97%
Ebrahimi et al. [7]	2016	SVM(One-class)	Accuracy= 98%
Ebrahimi et al. [21]	2017	CNN	f1=80%

IV. CONCLUSION

From the different algorithms used by the various researchers, we found that most of the methods are based on supervised learning, with little to no attention on other methods such as unsupervised or reinforcement learning. We also note that, the used models show good results. However the main drawback is that of data labeling as it differs from one scholar to another. Meaning that the performance of these models is constrained by the kind of labeling that was chosen by the respective researcher. We anticipate that the use of other models can contribute in this area, especially those that are capable of labeling data such as unsupervised learning to semi-supervised techniques.

Based on the on what has been learned from this paper and taking the work that is already done by other scholars into consideration, we plan to use semi-supervised deep learning models as future work to improve accuracy on CNN models. Also to leverage on new ML models that have emerged such as LSTM (Long-Short-Term Memory) and have been shown to be efficient in Natural Language Processing and text categorization.

References

[1] A. Kontostathis, L. Edwards, and A. Leatherman, "Text mining and cybercrime," Text Mining: Applications and Theory. John Wiley & Sons, Ltd, Chichester, UK, pp. 149-164, 2010

[2] H.J. Escalante, E.Villatoro-Tello, S.E. Garza, A.P. López-Monroy, M. Montes-y-Gómez, L.Villaseñor-Pineda, "Early detection of deception and aggressiveness using profile-based representations," Expert Systems with Applications, 89, pp. 99-111, 2017

[3] C. Harms, "Grooming: An operational definition and coding scheme," Sex Offender Law Report, 8(1), pp. 1-6, 2007.

[4] C. Girouard, “The National Center for Missing and Exploited Children,” US Department of Justice, Office of Justice Programs, Office of Juvenile Justice and Delinquency Prevention, 2008.

[5] http://www.ntewerk24.com/za/beeld , 2018 February 09.

[6] J. Wolak, K.J. Mitchell, and D. Finkelhor, “Online Victimization of Youth: Five Years Later,” Bulletin 07-06-025, National Center for Missing and Exploited Children, Alexandia, Alexandria, VA, 2006.

[7] M. Ebrahimi, C.Y. Suen, O. Ormandjieva and A. Krzyzak, “Recognizing predatory chat documents using semi-supervised anomaly detection,” Electronic Imaging, pp. 1-9, February 2016.

[8] G. Inches and F. Crestani, “Overview of the International Sexual Predator Identification Competition at PAN-2012,” In CLEF (Online working notes/labs/workshop), vol. 30, September 2012.

[9] M. Meyer, “Machine learning to detect online grooming,” 2015.

[10] N. Pendar, “Toward spotting the pedophile telling victim from predator in text chats. In Semantic Computing,” 2007, ICSC 2007. International Conference , IEEE, pp. 235-241, September 2007.

[11] A. Kontostathis, “Chatcoder: Toward the tracking and categorization of internet predators”. Proc. Text Mining Workshop 2009 Held In Conjunction With The Ninth Siam International Conference On Data Mining (SDM 09), Sparks, Nv, May 2009.

[12] 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), pp. 103-122, 2011.

[13] 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. Communication Theory,” 17(3), pp. 231-251, 2007.

[14] A. Gupta, P. Kumaraguru, and A. Sureka, ,”Characterizing pedophile conversations on the internet using online grooming. arXiv preprint arXiv:1208.4324, 2012.

[15] R. O’Connell, “A typology of child cybersexploitation and online grooming practices,” tech. reprt, Cyberspace Research Uniti, University of Central Lancashier, 2003.

[16] A.E. Cano, M. Fernandez, and H. Alani, “Detecting child grooming behaviour patterns on social media,” In International Conference on Social Informatics, Springer, Cham, pp. 412-427, November 2014.

[17] H.J. Escalante, E. Villatoro-Tello, A. Juárez, M. Montes-y-Gómez, and L. Villaseñor, “Sexual predator detection in chats with chained classifiers” Proc. 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, pp. 46-54, 2013

[18] C . Morris, “Identifying online sexual predators by svm classification with lexical and behavioral features,”, Master of Science Thesis, University Of Toronto, Canada, 2013.

[19] Y.G. Cheong, A.K. Jensen, E.R. Guðnadóttir, B.C. Bae, and J. Togelius, “Detecting predatory behavior in game chats,” IEEE Transactions on Computational Intelligence and AI in Games, 7(3), pp. 220-232, 2015.

[20] D. Bogdanova, P. Rosso, and T. Solorio, “Exploring high-level features for detecting cyberpedophilia” Computer speech & language, 28(1), pp. 108-120, 2014.

[21] M. Ebrahimi, C.Y. Suen, and O. Ormandjieva, “Detecting predatory conversations in social media by deep convolutional neural networks,”, Digital Investigation, 18, pp. 33-49, 2016.