ISSN:2229-6107

INTERNATIONAL JOURNAL OF
PURE AND APPLIED SCIENCE & TECHNOLOGY

E-mail :
editor.ijpast@gmail.com
editor@ijpast.in

www.ijpast.in

INTERNATIONAL JOURNAL OF
PURE AND APPLIED SCIENCE & TECHNOLOGY
IJPAST

ISSN :2229-6107
www.ijpast.in
Vol 10,Issuse 2.April 2020

Logical Models for Categorizing Talk About Personal Hygiene Online

THANGELLA PRIYANKA[1] , RAJAMPALLI PRIYA BHAVANA[2] , VALLAMSHETLA POOJITHA[3]
DANDA REVATHI[4]
SUPERVISOR , J VIJAYA SREE
Assistant Professor[1,2,3,4]

Abstract

The purpose of this research is to create a logistic mathematical model that can identify whether or not a given online interaction constitutes grooming. There are a variety of reasons why this work is crucial: the rise in the number of people using the Internet, the proliferation of social media, the proliferation of online crime of all kinds, and the epidemic of online sexual abuse. In youngsters has a wide range of physiological and physical effects. According to the Child Exploitation and Online Protection unit of the UK's Home Office Serious Organized Crime division, online grooming was the most often reported suspected Internet activity in 2009–2010. The elements of a grooming chat are identified by analyzing over 160 online script interactions.

1. Introduction

The goal of this study is to develop a mathematical model for determining whether or not a given conversation script in an online setting is a grooming chat. Cambridge Online Dictionary1 defines "grooming conversation" as "the criminal practice of befriending a child, typically online, with the aim of convincing the child to engage in sexually exploitative behavior." Several factors inspired this work. The first is that there has been a dramatic increase in the number of Internet users all over the world in recent years. In the United States, for instance, 67% of all homes with kids are online, 84% of kids ages 12 to 17, and 97% of kid's ages 18 to 24 have access to the web2. The proliferation of social media platforms like Face book and Twitter, and the subsequent rise in media coverage, is a further factor. The third issue is the proliferation of cybercrime, particularly online grooming. Fourth, sexual abuse of minors is a crime that has physical, emotional, behavioral, and psychological repercussions.

According to reports received by Child Exploitation and Online Protection Service in 2009 and 2010, online grooming was the most often reported suspected Internet behavior. (CEOP). Some examples of online grooming include making inappropriate sexual advances or encouraging a youngster to engage in sexual activity. The Child Exploitation and Online Protection Centre (CEOP) in the United Kingdom was set up as part of the UK's Home Office Serious Organized Crime agency (SoCA) to study the

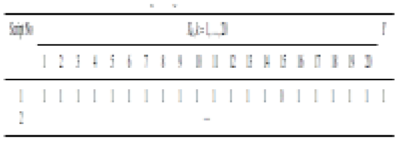
ISSN :2229-6107
www.ijpast.in
Vol 10,Issuse 2.April 2020

Danger of sexual crimes against children both online and in person and to aid in the development of strategies for safeguarding them. CEOP has observed that the perpetrators are likely to have a high level of IT literacy, as well as an awareness of the law enforcement system and how to circumvent it. The Internet has made it easier for individuals to do both good and harmful things. It's far simpler for sexual predators to create new identities and hide their tracks this way. Sexual predators may quickly compile a list of prospective victims with the help of the Internet.4. Sexual offenders may seize unforeseen openings, although they are often Part, specific amounts of preparation are necessary to accomplish an act5. They should plan, choose, follow, and catch their victims in order to carry out an assault. About 71% of sexual offenders are so dangerous to society that they must be locked up for the rest of their lives.6. However, Internet conversations often leave digital traces. It is possible to use this electronic evidence in forensic investigations of sexual offenders. (Al-Zaidy, Fung, 2012). Additionally, their digital footprints might be used by law enforcement to investigate and prosecute the crime. The data collected digitally will help the mathematical model being constructed for law enforcement.

2. Research Methods

The following is the method that was used to conduct this study. The first thing we did was going to http://www.perverted-justice.com/ and randomly downloads 111 chat scripts and 48 scripts from www.literotika.com. More than five hundred talks between juvenile victims or law enforcement and predators that specialize in grooming youngsters may be found on the defunct site. Police masquerading as underage people. It was previously established that all of the interactions were instances of internet grooming. On the latter, adults may openly discuss their sexual desires without fear of legal repercussions. Elzinga et al.7 and Wollis8 have taken a similar tack. One hundred of these dialogues will be utilized as training data, and another fifty-nine will be used as testing data. We shall arbitrarily sample from both data sets to quantify the presence of grooming traits, we secondly constructed the term frequency-inverse inverse document frequency (tf-idf) matrix. There are 20 distinct categories for various aspects of personal presentation, which will be discussed in further depth in Section 3. Table 1 is a sample conversation script illustrating the grooming persona. Here, everything except trait 15 of grooming is represented in dialogue script no. 1. This is definitely a grooming discussion, hence Y = 1. If there are no grooming details in the script, those details will be initialized to 0. A script has Y = 0 if and only if it is a grooming dialogue.

Table 1: An example of the grooming conversation's characterization.



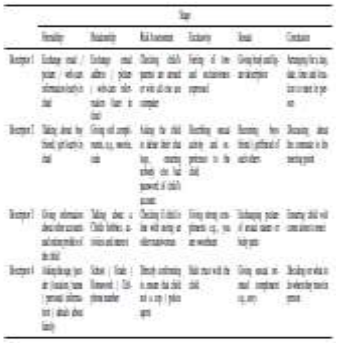
Script No	X_j (j = 1,…,20)																				Y
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1
2										–											

3. Supporting Theory

Grooming Process

Many researchers4, 9-11 have broken down the grooming process into distinct phases because of its complexity. The following are the phases that were identified by O'Connel10. First, the pedophile forms a friendship with the youngster by getting to know them and learning more about them via extensive questioning. How long a pedophile spends doing this depends on the individual. Second-stage relationship development includes developing the friendship further, At this point; the adult may start talking to the kid about things like school and family life. While not every adult goes through this faze, those who do usually keep in touch with the kid to keep up the charade that they're best buddies. Third-stage risk assessment: referring to the section of the chat in which pedophiles inquire as to the computer's location and user count. The pedophile is attempting to gauge the likelihood that someone would notice his inappropriate behavior toward the child, such as the child's parents, guardians, or older siblings. Fourth, after the risk assessment phase, the speed of dialogues often shifts into the exclusive phase. Sometimes the pedophile may direct what he wanted to speak about, such as issues such as said by an adult, and the victim will believe that the pedophile understood all they said. At this point, you're probably attempting to create trust by telling each other a little secret. Phase 5: The Sexual Phase. This is when the pedophile might ask inquiries such, "Have you ever been kissed?" or "Do you have ever touched yourself?" At this point in the discourse, an adult often positions the introduction such that there is a strong feeling of shared push. Since the youngster had never entered the subject before, it was impossible for them to detect the shift in intensity. Welner11 also suggested the following six steps. First, the abuser identifies a vulnerable youngster, then they earn the child's confidence, then they provide a need, then they isolate the child, then they sexualize the connection, and last, they maintain control. The Lanning4 grooming method includes finding children who are easy prey, getting their peers involved, getting them used to being touched, isolating them, and making them feel responsible. Table 2 displays Gupta9's recommended steps along with a brief description of each.

Table 2. Descriptor for Stage of Online Grooming9.



	Stage					
	Friendship	Relationship	Risk Assessment	Exclusivity	Sexual	Conclusion
Descriptor 1	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Descriptor 2	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Descriptor 3	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Descriptor 4	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

Grooming Characteristics

Inquiring into the potential dangers of a discussion (X1): Predators always consider the possibility that their chat with a prospective victim may be overheard by the victim's parents. The parents may bring up the incident in discussion if they are aware of who the victim is. Face legal consequences because their parents gave police a tip. When conversing with a possible victim, predators often inquire as to who uses the computer, where it is stored, and whether the victim's parents know the password to the chat program. The offender then warns the victim that their acts may have legal repercussions. (X2). Criminals tell victims this so they won't be caught themselves. Predators use this strategy to make sure that prospective victims are aware of the risks yet continue the discussion anyhow. Predators may avoid being reported to the authorities if they initiate conversation with their victims about their desires. Child predators often inquire about the victim's parents as one of the first points of conversation. (X3). Potential victims whose relationships with their parents are hostile or unpleasant are at a higher risk. Youth who spend a lot of time online, particularly those whose parent-child relationships are tense or nonexistent, are vulnerable to sexual exploitation. Another feature that's similar to the second is the presence or absence of a companion or adult. (X4). If no trusted adult or sibling tells their parents about the conversation, the predator may feel secure from legal repercussions.12, 13. Contact via non-Internet means (X5): At this point, predators are looking for alternatives to Internet-based communication; they want to feel fulfilled by moving from text-based to media-based to voice-based communication because it's more exciting for them. At this stage, predators are attempting to establish mutual trust with their prospective victims. If they are successful in doing so, the subsequent stages of their relationships with their victims will be much more straightforward. Ten (X7): Feel-related word use in conversation; this includes both predators and prospective victims discussing their emotions. Terms pertaining to biology, the body, and sex (X8): Any conversation that intends to become sexual will inevitably utilize terms from this category.8. Sexually predatory language (X9) includes the usage of terms for sexual organs that are common among minors.13.

Slang or commonly-used terminology for intimate body parts (X10): terminology for the sexual category that is also used in everyday speech.7. Reframing (X11): At this point, sexual themes related to predators14 and the reframing of sexual conduct into non sexual word, such as linking sexual act to fooling about, practicing, and teaching13, will be introduced gradually into the dialogue.

Predators may ask prospective victims for images with a sexual theme if they seem attractive (X12). Photos of a sexual nature may be a source of imagination or used to blackmail prospective victims into submission.7, 10, 12, and 13.

In phase X13, the predator will desensitize the victim via communication in an effort to make the victim feel more at ease talking about sexual topics. Some predators would make up typos by replacing non-sexual words with sexual ones; for example, they could change the letter "p" in "pick" to a "d" and claim it was an accident. Sharing sexual experiences and wants (X14): At this point, predators will try to find out whether their prospective victims have any sexual urges or are open to being touched sexually in the future. At this point, the predator may also inquire as to the sexual history of the possible victim. How often, if at all, do you engage in sexual activity? Predators believe that prey who have never had sex before would find it simpler to engage in sexual activity with them since it is no longer taboo to discuss sexuality.12. Sexual context was discussed for the first time (X15), before predators' fantasies had developed to the kern10 level. At this point in the dialogue, we will begin to broach the topic of fantasy, although at a very preliminary level (X16).Activity based on acting out fantasies (X17): the dialogue has progressed to this point, but there is no obvious sign of closeness.

Term Frequency and Inverse Document Frequency

Next, we'll use a standard method for evaluating documents called term frequency-inverse document frequency (TF-IDF) to determine how often each term appears in the discourse. (Russell, 2014). The primary objective of this procedure is to identify the frequency with which certain terms appear in a given document or corpus of text. To count how many documents include that phrase and discover its inverse-document-frequency (IDF).

Binary Logistic Regression

Generalized form the equation for the binary logistic model is

$$\ln\left(\frac{P}{1-P}\right) = b_0 + b_1X_1 + b_2X_2 + \cdots + b_kX_k \qquad (1)$$

Where X1, X2, X3, X4, X5, Xk, X, and Y are all binary variables taking on the values 0 or 1. If Xk = 1, then features of conversations about grooming of type k exist (k = 20). If Y = 1, then the model thinks it's a grooming talk.

Performance Measures

The constructed logistic model is assessed based on the confusion matrix shown in Table 3. The correctness of the model, as specified by Eq. (1), is also considered. (2).

The meaning of "contingency table" is listed in Table 3.

		Actual	
		Yes	No
Prediction	Yes	True Positive (TP)	False Positive (FP)
	No	False Negative (FN)	True Negative (TN)

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FN + FP} \qquad (2)$$

4. Results

Feature Extractions

The grooming qualities of each dialogue script were first identified and then classified into 20 categories (see below). Table 4 provides a summary of the grooming features found in a sample of 100 discussion transcripts. The following are some intriguing inferences drawn from the tabulation. Some of the most typical grooming behaviors include "using word in biology, body, and sexual category," "introduced sexual stage," "using word in feeling category," "arranging further contact and meeting," "telling the sexual preference or desire and sexual experience," and "calling intimate part using popular name or using slang word," in that order. The language used in discussions about personal hygiene seldom includes terms relating to raising children. The 33 conversations that do not include grooming lack access to seven of the twenty grooming qualities. However, numerous grooming cues are picked up in otherwise normal discourse. Because of this, discussing proper grooming techniques might be difficult to categorize. Using words from the "feeling" category, "using words from the "biology, body, and sexual" category, "calling intimate part using popular name or using slang," "introduced sexual stage," "fantasy enactment initial stage," and "fantasy enactment-based activity" are all common in grooming conversations that otherwise wouldn't be considered grooming.

Model Development

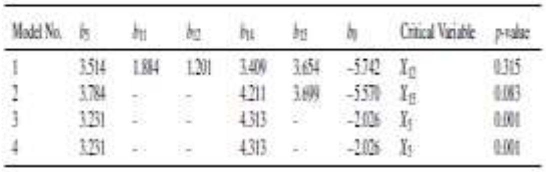
Each independent variable is pre-tested for its possible association with the dependent variable Y before the logistic model is established. Statistical tests such as the paired t-test, Spearman's rank correlation coefficient, and Pearson's correlation coefficient may be used to assess a relationship's strength. But the first since all the information is binary, option is chosen. Table 5 displays the p-values determined by applying the paired t-test to each independent variable. From 0.000 to 0.9999, the p-values cover a large range. As the size of the p-value indicates, the independent variables we're interested in are those that are highly important to or have a major impact on the dependent variable. We define statistical significance as a p-value for an independent variable being less than a predetermined threshold; in this example, we choose a p-value of 0.25. Based on this criterion, the following factors need not be included in the logistic model creation process: The X-series numbers 5-11-12-14-15

Based on this information, we may reduce Eq. (1) to its simplest form.

$$\ln\left(\frac{P}{1-P}\right) = b_0 + b_5X_5 + b_{11}X_{11} + b_{12}X_{12} + b_{14}X_{14} + b_{15}X_{15}. \qquad (3)$$

Step-wise regression employs three techniques—the Enter Method, the Forward Stepwise Method, and the Backward Stepwise Method—to calculate the model coefficients b0, b5, b11, b12, b14, and b15. Table 6 displays the outcomes of the sequential regression analysis. The table lists the estimated model coefficients (b0, b5, b11, b12, b14, and b15), the procedures that were used, and the results. With emphasis on the most important variable in each model and its degree of significance. The most important factor is unmistakably the one with the greatest p-value and the coefficient value that is the lowest relative to the other variables in the model. The paired t-test outcome of Eq requires intuitive evaluation. (3). this leads to the conclusion that a reliable independent variable must be capable of accurately distinguishing between grooming and non-grooming discussions.

Table 6. Logistic models have been developed for categorizing online grooming discourse.



Model No.	b_5	b_{11}	b_{12}	b_{14}	b_{15}	b_0	Critical Variable	p-value
1	3.514	1.884	1.201	3.409	3.654	–5.742	X_{12}	0.315
2	3.784	–	–	4.211	3.699	–5.570	X_{15}	0.083
3	3.231	–	–	4.313	-	–2.026	X_5	0.001
4	3.231	–	–	4.313	-	–2.026	X_5	0.001

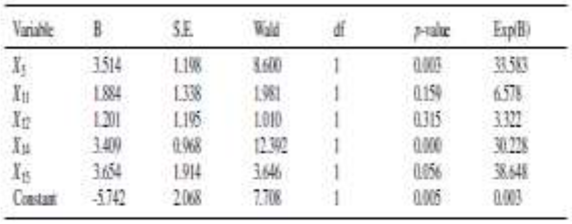
The frequency with which it appears in a talk about grooming should be higher than in other types of conversations. Table 4 shows that X14 is present in 85% of scripts for conversations about grooming, but in just 4% of scripts for conversations about anything else. In grooming discussions, X5 is mentioned 67% of the time, yet it does not have zero frequency outside of discussions about grooming (94%). Sixty percent of grooming-related conversations include the variable X11, whereas ninety-seven percent do not. X12 and X15 are interesting factors to consider. Only 34% of conversations about grooming mention variable X12, but 94% of conversations involving anything else do not. Conversely, X15 is mentioned 97% of the time while talking about grooming but just 18% of the time when talking about anything else. This finding suggests that X15 is a reliable predictor of grooming-related speech, whereas X12 is reliable for non-grooming-related conversation.

Table 6 displays five models, each with its own set of coefficients for the independent variables: b5 = 3.514, b11 = 1.884, b12 = 1.201, b14 = 3.409 and b15 = 3.654. The numerical number gives a precise description of the related variable's impact on the model's forecast. B11 and b12, the two smallest coefficients, have values that are drastically different from those of the other coefficients. The coefficient for X12 is 1.202, and its corresponding p-value is 0.315, making it the least significant variable. Variable X11 is the second, less important one. The second model does not have X11 or X12. The coefficients for this model are as follows: b5 = 3.784, b14 = 4.211 and b15 = 3.699, with a p-value of 0.083 for X15, the least significant variable in the model. We see that when X11 and X12 are removed, the model uncertainty drops from 31.5 percent to 8.3 percent. A model with an exceptionally low p-value is produced when the number of independent variables is reduced further. Model using only two variables (X5 and X14) and a little (0.1%) margin of error.

Testing's of Significance for Model 1, Model 2, and Model 3

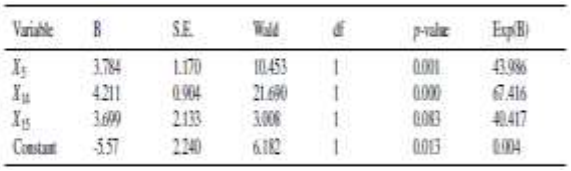
The outcomes of the significance tests for Models 1, 2, and 3 are discussed below. Tables 7, 8, and 9 detail the outcomes of the testing, in that order. According to the data, X5 and X14 seem to be the two most important factors in every analysis. The significance of X15, the second most important variable, is followed by using X11 and X12 as the means. The calculated p-value and Wald statistic allowed for these inferences to be made.

Table 7. A statistical overview of X5, X11, X12, X14, and X15, five independent variables.



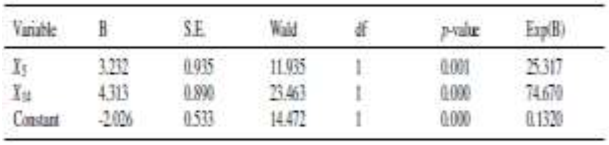
Variable	B	S.E.	Wald	df	p-value	Exp(B)
X_5	3.514	1.198	8.600	1	0.003	33.583
X_{11}	1.884	1.338	1.981	1	0.159	6.578
X_{12}	1.201	1.195	1.010	1	0.315	3.322
X_{14}	3.409	0.968	12.392	1	0.000	30.228
X_{15}	3.654	1.914	3.646	1	0.056	38.648
Constant	-5.742	2.068	7.708	1	0.005	0.003

The results of the statistical tests for X5, X14, and X15 are summarized in Table 8.



Variable	B	S.E.	Wald	df	p-value	Exp(B)
X_5	3.784	1.170	10.453	1	0.001	43.986
X_{14}	4.211	0.904	21.690	1	0.000	67.416
X_{15}	3.699	2.133	3.006	1	0.083	40.417
Constant	-5.57	2.240	6.182	1	0.013	0.004

Table 9 presents a summary of the statistical tests conducted on X5 and X14, two independent variables.

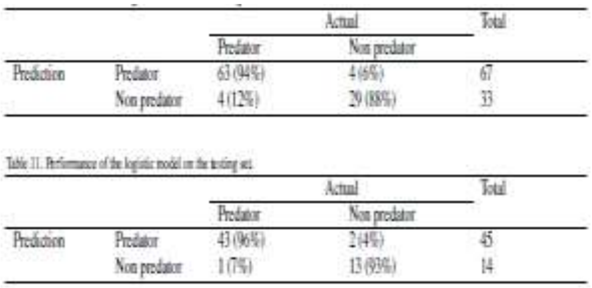


Variable	B	S.E.	Wald	df	p-value	Exp(B)
X_5	3.232	0.935	11.935	1	0.001	25.317
X_{14}	4.313	0.890	23.463	1	0.000	74.670
Constant	-2.026	0.533	14.472	1	0.000	0.1320

Model Evaluation

Here, we assess the Table 6 model's efficacy in terms of its accuracy and recall (Table 3). Table 10 displays results from 100 training set scripted conversations, whereas Table 11 displays findings from 59 testing set scripted conversations. It should be noted that only the third model is used for performance evaluation. However, Increases in recall with the first and second models are likely to come at the expense of accuracy. Training set accuracy is 92%, while testing set accuracy is 95%. This means the third model is rather accurate.

Training set results for the logistic model are shown in Table 10.



		Actual		Total
		Predator	Non predator	
Prediction	Predator	63 (94%)	4 (6%)	67
	Non predator	4 (12%)	29 (88%)	33

Table 11. Performance of the logistic model on the testing set.

		Actual		Total
		Predator	Non predator	
Prediction	Predator	43 (96%)	2 (4%)	45
	Non predator	1 (7%)	13 (93%)	14

5. Conclusions

The goal of this study is to develop a logistic mathematical model that can identify whether or not an online interaction that follows a predetermined script constitutes grooming. Given the proliferation of social media and the rise in crime that has followed, this work is crucial. The research has pinpointed five key aspects of the grooming discussion. These traits include the following questions: "Other way to contact? ", "Reframing? ", "Asking hot pictures? ", "Telling the sexual preference or desire and sexual experience? ", and "Introducing sexual stage?" In addition, a mathematical model of logistic processes has been developed using this data. The model can detect grooming conversations with 95% accuracy, including 96% true positive and 93% true negative and only 4% false positive and 7% false negative, based on an evaluation of the model's performance using the training data set of 100 scripts and the testing data set of 59 scripts.

References

1. Cambridge Online Dictionary, http://dictionary.cambridge.org/

2. Institute of Health Economics. Sexual Exploitation of Children and Youth over the Internet: A Rapid Review of the Scientific Literature. Alberta, Canada. 2010.

3. Cruise, TK. Sexual Abuse of Children and Adolescents. In Psychologists: Helping Children at Home and School III: Handouts for Families and Educators. 2010.

4. Lanning, KV. Child Molesters: A behavioral analysis for professionals investigating the sexual exploitation of children. Office of Juvenile Justice and Delinquency Prevention, Office of Justice Programs, U.S. Department of Justice. 2010.

5. Andrews, DA., and Bonta, J. The Psychology of Criminal Conduct. New Providence, NJ: Matthew Bender & Company, Inc. 2010.

6. Herkov, M. What is Sexual Addiction? http://psychcentral.com/lib/what-is-sexual-addiction/000748. Retrieved on August 6, 2014.

7. Elzinga, P., Wolff, K. E., & Poelmans, J. Analyzing Chat Conversations of Pedophiles with Temporal Relational Semantic Systems. European Intelligence and Security Informatics Conference, European Intelligence and Security Informatics Conference (EISIC), 2012. p. 242–249.

8. Wollis, M. A. Online Predation: A Linguistic Analysis of Online Predator Grooming. College of Agriculture and Life Sciences, Social Sciences of Cornell University, Research Report, 2011 http://dspace.library.cornell.edu/bitstream/1813/23125/2/Wollis%2c% 20Melissa%20-%20Research%20Honors%20Thesis.pdf

9. Gupta, A., Kumaraguru, P., and Ashish, S. Characterizing Pedohile Conversation on the Internet using Online Grooming. Computing Research Repository. arXiv preprint arXiv: 1208.4324 (2012).

10. O'Connel, R. A typology of cybersex exploitation and online grooming process. Technical Report. Cyberspace Research Unit, University of Central Lancashire, the United Kingdom. 2014. http://netsafe.org.nz/Doc_Library/racheloconnell1.pdf.