**TELKOMNIKA**, Vol.16, No.3, June 2018, pp. 1239~1248
ISSN: 1693-6930, accredited **A** by DIKTI, Decree No: 58/DIKTI/Kep/2013
**DOI:** 10.12928/TELKOMNIKA.v16i3.6745


# A Simple Classifier for Detecting Online Child Grooming Conversation

**Fergyanto E. Gunawan*[1], Livia Ashianti[2], Nobumasa Sekishita[3]**
[1]Industrial Engineering Department, BINUS Graduate Program-Master of Industrial Engineering,
Bina Nusantara University, Indonesia
[2]Master Program of Computer Science, Bina Nusantara University, Indonesia
Jl Kebon Jeruk Raya No 27, Phone: +62-21-534-5830, Fax: +62-21-530-0244
[3]Department of Mechanical Engineering, Toyohashi University of Technology
Toyohashi, Aichi 441-8580, Japan
*Corresponding author, e-mail: fgunawan@binus.edu

***Abstract***

*The massive proliferation of social media has opened possibilities for the perpetrator conducting the crime of online child grooming. Because the pervasiveness of the problem scale, it may only be tamed effectively and efficiently by using an automatic grooming conversation detection system. The current study intends to address the issue by using Support Vector Machine and k-nearest neighbors' classifiers. Besides, the study also proposes a low-computational cost classification method, which classifies a conversation using the number of the existing grooming conversation characteristics. All proposed methods are evaluated using 150 textual conversations of which 105 are grooming, and 45 are non-grooming. We identify that grooming conversations possess 17 features of grooming characteristics. The results suggest that the SVM and k-NN can identify grooming conversations at 98.6% and 97.8% of the level of accuracy. Meanwhile, the proposed simple method has 96.8% accuracy. The empirical study also suggests that two among the seventeen characteristics are insignificant for the classification.*

***Keywords****: online child grooming; support vector machine; k-nearest neighbors; grooming classifier*



## 1. Introduction

Online child grooming is defined as a process to approach, persuade, and engage a child, the victim, in sexual activity by using the Internet as a medium. Perpetrators approach the victim to build not only sexual but also emotional relationship [1]. The massive proliferation of social media has opened possibilities for the perpetrators to conduct the crime of online child grooming in a larger sale [2]. According to the Child Exploitation and Online Protection Agency, online child grooming is the most reported crime in the UK in 2009–2010 [2]. It affects the victim life psychologically, physically, emotionally, behaviorally, and psycho-socially [3].

For revealing this type of crimes, investigator usually relies on the conversation texts where the grooming patterns are carefully analyzed [4]. With the vast amount of textual conversation data, the process becomes severe and requires a significant amount of time. The manual approach of investigating grooming pattern is also error Rome [4]; besides, the grooming process may take minutes, hours, days, or months [5-7].

For the reason described above, it is important to develop an automatic system to analyze a conversation text and to detect the possibility of the online child grooming conversation. During the last five years, a number of research works have been addressing the issue using various pattern detection schemes including using *k*-means clustering by Kontostathis, Edwards, and Leatherman [4], a ruled-based approach by McGhee et al. [8], Support Vector Machine (SVM) by Pandey, Khapaftis, and Manandhar [9]. Recently, Pranoto, Gunawan, and Soewito [10] developed a grooming detection system utilizing a logistic regression model. SVM method seems to work best for the text-based classification according to Ref. [11]. However, SVM has also been demonstrated for the image-based classification such as detection of corona artery disease [12] and breast cancer [13]. Reference [14] used SVM for developing an intrusion detection system. This study intends to propose a simple method to detect an online child grooming conversation. In doing so, the study firstly identifies the main

*Received July 8, 2017; Revised February 5, 2018; Accepted February 28, 2018*

characteristics of the type of conversations. The proposed method is developed on the basis of the number of existing characteristics.

## 2. Relevant Theories

### 2.1. The characteristics of online child grooming

Online child grooming conversation texts are complex as it varies in duration, type, and intensity depending on the perpetrator characteristics and behavior. However, in general, O'Connel [15] and Gupta [16] have identified the typical stages in an online child grooming process. The first is the friendship forming stage. The perpetrator tries to get introduced to the child and then to establish a possibility of exchanging name, location, age, etc. Furthermore, the perpetrator inquiries other online information related to the child, requesting photos to confirm that the child is indeed a child.

The second is the relationship forming stage. The perpetrator and the child talk about family, school, interest, and hobbies of the child so that he can exploit them by deceptively making the child believes that they are in a relationship. The third is the risk assessment stage. The perpetrator tries to gauge the level of threat and danger by talking to the child. He ensures that the child is alone, and nobody else is reading their conversations.

The fourth is the exclusivity stage. The perpetrator tries to gain the complete trust of the child. Often, the concept of love and care are introduced by the perpetrator in this juncture. The fifth is the sexual stage. The perpetrator and the child talk about sexual activities and developing sex fantasy. Finally, the sixth is the conclusion stage. In this stage, the perpetrator approaches the child for a meeting in person. These stages of online child grooming may or may not occur in a sequence. The frequency, order, and extent of the occurrence of these stages may vary from chat to chat. On the basis of the previous work [10], and Refs. [2-16], we have identified 17 grooming characteristics, see Table 1, and their relation to the grooming stages are presented in Table 2. These characteristics would be used to classify the online conversation texts on the current study.

Table 1. The identified grooming characteristics.

| No. | Characteristics, Description, and Source |
|---|---|
| 1 | Asking profile. Perpetrator and victim exchange information about personal info, such as, name, age, and location [16]. |
| 2 | Other way contact. Perpetrator and victim talk about another way to communicate, such as, phone, email, and social media [16]. |
| 3 | Asking picture. Perpetrator asks victim to send a his/her picture or vice Vera [16]. |
| 4 | Giving compliment. Perpetrator compliments the victim in order to make the victim happy and flattered [16]. |
| 5 | Talking about activity, favourite hobby, and school. Perpetrator and victim talk about daily activities, favourite hobbies and victim's school Activities [16]. |
| 6 | Talking about friend and relationship. Perpetrator and victim talk about friendships or relationships, such as, asking about relationship with another person [16]. If the victim is not in a relationship with another person, it's easier for perpetrator to get closer. |
| 7 | Asking questions to know the risk of conversation. Perpetrator tries figure out the risk of their conversation, whether their conversation is known by victim's parental [10]. Usually, perpetrator will ask about anyone who uses victim's computer, location of the computer, and whether victim's parents know the password of the chat application. |
| 8 | Acknowledging wrong-doing. Perpetrator will inform to potential victim what they are doing is wrong, and have legal risks for perpetrator [10]. By telling this to victim, perpetrator has a purpose, which is perpetrator will be free from legal cases that will make him/her jailed in the future. |
| 9 | Asking if the child is alone or under adult or friend supervision. Perpetrator wants to make sure the victim whether is alone or under Supervision [2]. |
| 10 | Trying to build mutual trust. Perpetrator trying to build the mutual trust from victim, the next level relationships will be easier for perpetrator if perpetrator gain the trust from the victim [10], [16]. |
| 11 | Using falling in love words. In conversation between perpetrator and the victim, they use words to express they are in love [2], [16] . |
| 12 | Using word to express feeling. In a conversation between the perpetrator and victim, they use words to express their feelings [10]. |
| 13 | Using word about biology, body, intimate parts, and sexual category. In a conversation between the perpetrator and the victim, they use words that contain sexual context [10]. |
| 14 | Asking hot picture. Perpetrator asks victim for sexual theme photos or vice versa [10], [16]. These pictures can be used as fantasy or a tool to threaten victim to obey the perpetrator. |
| 15 | Introducing sexual stage. Conversation started with talking about sexual context, such as ask about sex experiences [10], [16]. |

Table 1. The identified grooming characteristics.

| No. | Characteristics, Description, and Source |
|---|---|
| 16 | Sexual stage. Conversation has entered the stage of sexual fantasies with words that show the interaction of activities and involve intimacy [16]. |
| 17 | Arranging further contact and meetings. Perpetrator tries to get the victim address in order to have a meeting at the victim's house or to invite victim to meet somewhere [10, 16]. |

Table 2. The relation between the 17 grooming characteristics and grooming stages.

| No | Grooming Stage | Example Characteristics |
|---|---|---|
| 1 | Friendship forming | Asking profile |
| 2 | | Other way of contact |
| 3 | | Asking picture |
| 4 | | Giving compliment |
| 5 | Relationship forming | Talking about activity, favourite, hobby, school |
| 6 | | Talking about friend and relationship |
| 7 | Risk assessment | Asking questions to know the risk of conversation |
| 8 | | Acknowledging wrong-doing |
| 9 | | Asking if the child is alone or under adult or friend supervision |
| 10 | Exclusivity | Trying to build mutual trust |
| 11 | | Using falling-in-love words |
| 12 | | Using words to express feeling |
| 13 | Sexual | Using words about biology, body, intimate parts, and sexual category |
| 14 | | Asking hot picture |
| 15 | | Introducing sexual stage |
| 16 | | Sexual stage |
| 17 | Conclusion | Arranging further contact and meetings |

### 2.2. Support Vector Machine Classification Method

In the present study, we only use the Support Vector Machine (SVM) for linearly separable data. The SVM is a numerical method to compute a hyperplane for separating a two-class dataset. It can easily be extended to multiple-class problem. The SVM establishes the hyperplane, governed by $(\mathbf{w},b)$, by using the support vectors, which are the data points that are closest to the hyperplane. The following SVM formulation is derived from Refs. [17-18]; readers are advised to the two sources for detail exposition. We consider the point sets $\mathbf{x}_i \in \Re^d$, as the support vectors, with the categories $y_i \in [-1,+1]$. The hyperplane that separates $y_i = -1$ from those of $y_i = +1$ should satisfy

$$\langle \mathbf{w}, \mathbf{x} \rangle + b = 0, \tag{1}$$

where $\mathbf{w} \in \Re^d$, $\langle \mathbf{w}, \mathbf{x} \rangle$ denotes the inner dot product of the vectors $\mathbf{w}$ and $\mathbf{x}$, and $b$ is a scalar constant. The hyperplane is obtained by solving:

$$\min_{\mathbf{w},b} L_p = \tfrac{1}{2}\langle \mathbf{w}, \mathbf{w} \rangle - \sum_i \alpha_i \left[ y_i \left( \langle \mathbf{w}, \mathbf{x}_i \rangle + b \right) - 1 \right]. \tag{2}$$

where $\alpha_i \geq 0$. For the case where the data are linearly not separable, the feature vector $\mathbf{x}_i$ would be transformed with a kernel function. Two types of the kernel functions would be evaluated: polynomial type where $K(\mathbf{x},\mathbf{y}) = \left(1 + \langle \mathbf{x}, \mathbf{y} \rangle\right)^d$ and radial basis function (RBF) type where $K(\mathbf{x},\mathbf{y}) = \exp\left(-\langle \mathbf{x}-\mathbf{y}, \mathbf{x}-\mathbf{y} \rangle / \left(2\sigma^2\right)\right)$. The parameter $d$ is an integer, and would be evaluated for $d = 1,2$, and $3$, and $\sigma$ has a positive value.

### 2.3. k-nearest neighbor classification method

The *k*-Nearest Neighbor (*k*-NN) is an instant-based learning algorithm for the classification of the query data on the basis of the training dataset. For the classification

purpose, firstly, the method selects the most similar *k* data to the query data from the training dataset. The term similar is usually quantified by using the Euclidian distance [19]. Secondly, the method evaluates the classes of those *k*-selected data. Finally, the query data is assumed belong to the dominant class of the classes.

### 2.4. Accuracy Indicator

The classification accuracy is computed by:

$$\text{Accuracy}=\frac{\text{TP}+\text{TN}}{\text{TP}+\text{TN}+\text{FP}+\text{FN}} \quad (3)$$

where TP stands for True Positive, TN for True Negative, FP for False Positive, and FN for False Negative [20].

## 3. Research Method

The research procedure is schematically shown in Figure 1 and a few important steps are briefly explained in the following.

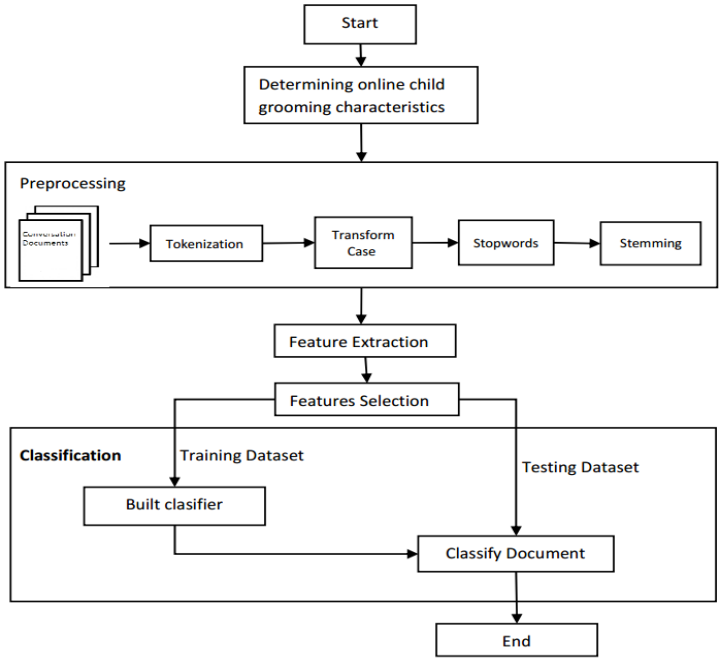


Figure 1. The research procedure

### 3.1. Dataset Preparation

Two types of conversation texts are required for this research: the first type is the conversations of actual online child grooming and the second type is the conversations of non-grooming conversations but has grooming characteristics. The first type conversations are randomly selected from Perverted Justice website [21], a website that contains more than 500 texts of grooming conversations involving perpetrators and children, juvenile victims, or undercover law enforcements. Only 105 texts are selected. The source has also been used by the previous researchers [4], [8-10], [22].

The second type conversations are selected from Literotic website [23]. The web contains conversation scripts of people expressing their sexual passion legally. Literotic defines that the website purpose is ``this chat room is here for adults interested in erotic subject, so be

aware of that before you enter!". Fourty five non-grooming conversation texts are randomly selected from the site.

### 3.2. Preprocessing

The text of online conversation contains many noises from the perspective of document classification. Those noises should be minimized or eliminated, if possible, prior the analysis to determine the grooming characteristics. The all texts in this research are subjected to the following processes. Tokenization: non-letter characters in the document would be removed and each document is partitioned into words. Transformation: words in the document would be transformed into lowercase. Stopword elimination: words which frequently exists across document but not significantly useful would be erased. Stemming: words in the document would be reduced into their root using porter algorithm. Generating 3-gram: words in the document would be formed into 3-grams of 1 continuous sequence formed of 3 words from the document.

### 3.3. Feature Extraction

Texts that have been preprocessed would be transformed into a vector space model (VSM). The features are words or combinations of words that form the word list. The word list is denoted with $T_1, T_2, \ldots, T_t$. The feature extraction result from a document $d_m$ is transformed into vector $d_m = \{w_{T_1,m}, w_{T_2,m}, \ldots, w_{T_t,m}\}$ where $m \in M$, $M$ is number of documents and $w_{T_i,m}$ is weight calculation results from feature $i$ using TF-IDF that represents how important features $i$ in document $m$ and all documents in the dataset.

### 3.4. Feature Selection

Feature extraction results from each document in VSM would be used to create a grooming characteristic vector. Grooming characteristics used are 17 characteristics that have been determined in Table 1 the vector is denoted $C_m = \{c_{m,1}, c_{m,2}, c_{m,3}, \ldots, c_{m,17}\}$ where $m \in M$ and $c_{m,j}$ is a value that indicates whether or not the characteristic $j$ in the document $m$. If document $m$ does not contain characteristic $j$ then $c_{m,j} = 0$. If document $m$ contains characteristic $j$ then $c_{m,j} = 1$. To determine grooming characteristic $j$ value in document $m$, $c_{m,j}$, features from extraction will be selected in accordance with database which stores words or combinations of words that describe each grooming characteristic. The value of features that have been selected will be summed. If the result is 0 then $c_{m,j} = 0$ and if the result is greater than 0 then the characteristic value $j$ in the document $c_{m,j} = 1$.

### 3.5. Classification

The classification would be performed using SVM (see Subsection 2.2), *k*-NN (see Subsection 2.3) and our proposed method, which is based on the number of grooming characteristics in the document.

## 4. Results and Discussion

We have analyzed 150 conversation texts consisting of 105 grooming conversations, randomly taken from www.perverted-justic.com, and 45 non-grooming conversations, randomly taken from www.literotika.com. We have identified seventeen grooming characteristics by learning those grooming conversation and by considering previous works. Those characteristics are then represented in a vector space. These characteristics and their frequencies of occurrence in grooming and non-grooming conversations are presented in Table 3.

Table 3. The frequency of occurrence of grooming characteristics on the 105 grooming and 45 non-grooming conversations on the current study.

| No | Grooming Characteristics | Frequency | |
|---|---|---|---|
| | | Grooming | Non-Grooming |
| 1 | Asking profile | 97 | 2 |
| 2 | Other way contact | 101 | 17 |
| 3 | Asking Picture | 102 | 10 |
| 4 | Talk About friend and relationship | 96 | 22 |
| 5 | Giving Compliment | 104 | 28 |
| 6 | Talk About Activity, Favourite, Hobby, school | 95 | 16 |
| 7 | Asking Question To Know Risk Of Conversation | 44 | 0 |
| 8 | Acknowledging wrong doing | 99 | 15 |
| 9 | Asking if child is alone or adult supervision or friend | 84 | 0 |
| 10 | Trying building mutual trust | 98 | 27 |
| 11 | Using word in falling-in-love | 70 | 7 |
| 12 | Using word in feel category | 105 | 42 |
| 13 | Using word in biology, body, intimate parts, and sexual category | 105 | 43 |
| 14 | Asking hot picture | 13 | 0 |
| 15 | Introduced sexual stage | 101 | 34 |
| 16 | Sexual Stage | 97 | 44 |
| 17 | Arrange further contact and meeting | 100 | 5 |

What makes automatic classification difficult is that the grooming characteristics also appear on non-grooming conversations as shown by Table 3. For example, the most prevalent characteristics, which is the 13th characteristics, “using word about biology, body, intimate parts, and sexual category” appears in 105 grooming text conversations and in 43 non-grooming text conversations.

Another insight shown by the table is that two characteristics, namely, the 14th characteristics, “asking hot picture”, and the 7th characteristics, “asking question to know risk of conversation,” apper rarely. We hypothesize: the two characteristics may not have significant contribution to the performance of the document automatic classification. This will be empirically evaluated.

In the following, we are going to discuss the results in term of the classification accuracy for various classification methods and with or without the 7th or 14th grooming characteristics. The training set consists of 70 grooming and 30 non-grooming conversations. The testing set consists of 35 grooming and 15 non-grooming conversations. For the first research results, we compare the level of accuracy of the results by several SVM kernel functions, namely, RBF, quadratic, polynomial, and linear functions. The results, on the average accuracy, are shown in Table 4.

Table 4. The classification accuracy of the SVM method with the four kernel functions using with or without the 7th and 14th grooming characteristics.

| Grooming Characteristics | The Type of SVM Kernel | | | |
|---|---|---|---|---|
| | RBF | Polynomial | Quadratic | Linear |
| All seventeen | 83.8 | 97.6 | 98.6 | 98.6 |
| Without the 14$^{tn}$ | 87.4 | 97.6 | 98.6 | 98.6 |
| Without the 7$^{tn}$ | 89.4 | 96.6 | 97.4 | 97.8 |

These results reveal some interesting phenomena, some are expected, some are unexpected. We expect that the highest level of accuracy would be achieved by using all grooming characteristics. This expectation is materialized for the three types of SVM kernels: polynomial, quadratic, and linear. Using the RBF kernel, the results are rather unexpected: the accuracies without the 7th and 14th characteristics are better than using the all characteristics. The expectation that the 7th and 14th grooming characteristics would only slightly affect the level of accuracy is only materialized for the three kernel: polynomial, quadratic, and linear. Using all grooming characteristics, the level of accuracy by means of SVM method is within the range of 83–98% depending on the selection of the kernel function. In comparison to the method utilizing the logistic regression model, see Ref. [10], the three kernels provide slightly better accuracies. The RBF kernel produces a lower accuracy than the logistic model. For the

second research results, we also compare the level of accuracy by using a different classifier, that is the *k*-NN method with the k values of 1, 3, and 5. The results, in average, are depicted in Table 5.

Table 5. The classification accuracy of the *k*-NN method method with the *k* values of 1, 3, and 5, and with or without the 7th and 14th grooming characteristics.

| Grooming Characteristics | *k* Value | | |
|---|---|---|---|
| | 1 | 3 | 5 |
| All seventeen | 97.0 | 97.8 | 97.2 |
| Without the 14$^{tn}$ | 97.0 | 97.8 | 97.2 |
| Without the 7$^{tn}$ | 96.8 | 97.2 | 96.8 |

These results completely agree with our expectation. The highest average level of accuracy is achieved by using all grooming characteristics. This result is materialized for all values of *k*. The expectation that the 7th and 14th grooming characteristics will only slightly affect the level of accuracy is materialized for all of *k* values. The average level of accuracy in classification without the 14$^{th}$ characteristics is the same with using all grooming characteristics. Using all grooming characteristics, the average level of accuracy by means of the *k*-NN method is within the range of 96.8–97.8% depending on the *k* value. However, it is not clear whether increasing the *k* value will increase or decrease the level of accuracy.

Finally, we propose a simple classification method, which requires very low computational cost and makes it suitable for implementation in the electronic mobile devices. The proposed method is to classify the conversation on the basis of the existing number of grooming characteristics. This method is proposed by observing the fact that the number of grooming characteristics are markedly different; see Table 6.

Table 6. The distribution of the number of grooming characteristics in the 150 grooming and non-grooming textual conversations in the current study.

| The Number of Grooming Characteristics Contained in the Document | The Number of Documents | |
|---|---|---|
| | Grooming | Non-Grooming |
| 1 | 0 | 0 |
| 2 | 0 | 1 |
| 3 | 0 | 1 |
| 4 | 0 | 4 |
| 5 | 0 | 5 |
| 6 | 0 | 10 |
| 7 | 0 | 4 |
| 8 | 1 | 8 |
| 9 | 1 | 7 |
| 10 | 2 | 4 |
| 11 | 5 | 1 |
| 12 | 8 | 0 |
| 13 | 10 | 0 |
| 14 | 19 | 0 |
| 15 | 23 | 0 |
| 16 | 30 | 0 |
| 17 | 6 | 0 |

The table suggests that a conversation tends to be a grooming conversation if it contains the number of grooming characteristics within the range 8-17. Meanwhile, a conversation tends to be a non-grooming conversation if it contains about 2-11 grooming characteristics. Thus, the number of grooming characteristics can simply be used as a classifier; despite the fact, there is an overlap in the number of grooming characteristics between the two categories. To evaluate a text conversation, we can set certain threshold, evaluate the number of grooming characteristics, and decide that the conversation is grooming if its number of grooming characteristics is equals or exceeds the threshold.

We empirically evaluate the method by varying the threshold value from 1 to 17. If the number of grooming characteristics in a document is less than the threshold, the conversation

will be classified as a non-grooming conversation and vice versa. The results in the average accuracy are depicted in Table 7.

Table 7. The classification accuracy of the proposed method as a function of the threshold value.

| Threshold | The Level of Accuracy (%) | | |
|---|---|---|---|
| | All | Without the 14th | Without the 7th |
| 1 | 70.0 | 70.0 | 70.0 |
| 2 | 70.0 | 70.0 | 70.0 |
| 3 | 70.2 | 70.2 | 70.2 |
| 4 | 71.6 | 71.6 | 71.6 |
| 5 | 74.2 | 74.2 | 74.2 |
| 6 | 77.4 | 77.4 | 77.4 |
| 7 | 82.8 | 82.8 | 82.8 |
| 8 | 85.8 | 85.8 | 85.8 |
| 9 | 90.8 | 90.8 | 90.8 |
| 10 | 95.8 | 95.8 | 95.8 |
| 11 | 96.8 | 96.8 | 96.0 |
| 12 | 94.0 | 94.0 | 94.0 |
| 13 | 88.2 | 88.2 | 86.8 |
| 14 | 80.6 | 80.6 | 77.0 |
| 15 | 68.6 | 67.8 | 67.6 |
| 16 | 53.2 | 50.0 | 37.4 |
| 17 | 35.0 | 30.0 | 30.0 |

These empirical data suggest that the highest average level of accuracy is achieved at the threshold value of 11. The best threshold provides an accuracy level of 96.8%. The expectation that the $7^{th}$ and 4th grooming characteristics would only slightly affect the level of accuracy is materialized for all of the threshold values.

Finally, we compare the level of accuracy of the three classification methods: SVM, $k$-NN, and our proposal. For the SVM method, we only include the results of using the linear kernel as they are the best among the method. For the same reason, for the $k$-NN method, we include only the case of $k$=3. The comparison is presented in Table 8. The three methods support the hypothesis that the accuracy would slightly drop when the 7th and 14th grooming characteristics are excluded. In addition, these results suggest that the SVM classifier is able to classify the best in term of the accuracy. The proposed method, despite of its simplicity, also performs rather well.

Table 8. A comparison of the classification accuracy for SVM, *k*-NN, and proposed methods with and without the 14th or 7th grooming characteristics.

| Grooming Characteristics | Classification Method | | |
|---|---|---|---|
| | SVM | *k*-NN | Proposed Method |
| All | 98.6 | 97.8 | 96.8 |
| Without $14^{tn}$ | 98.6 | 97.8 | 96.8 |
| Withot $7^{tn}$ | 97.8 | 97.2 | 96.0 |

With empirical findings presented in the current and previous [10] works, the proposed seventeen characteristics seem to be highly representative and unique to differentiate a grooming textual conversation from a non-grooming one. Even when the characteristics are used in conjunction with a simple classification method, the classified conversation is very likely to be correct. Despite of these findings, we also note that the likelihood of the success of child online grooming increases when perpretrators employ identity deception and suggesting secrecy [24]. In the current classification framework, these behaviors have been identified as one of the grooming characteristics. However, it is of a great interest to understand to which extent the victim age, child or adult, affects the success of the current classification method, and we leave this issue as a future work. For the final note, the present work has opened a possibility for developing an automatic grooming detecting system on the devices that have low computational power.

## 4. Conclusion

Automatic system to detect online child grooming has an important role in analyzing the vast amount of conversation texts. For the reason, many studies have been performed using various pattern detection schemes. In the current work, seventeen characteristics of grooming conversation are identified and utilized for classification. Two traditional classification methods are used: SVM and *k*-NN. Moreover, this work proposes a simple classification method on the basis of the number of existing grooming characteristics in the conversation. The numerical analysis using empirical data suggests that the SVM method with the linear kernel is the best among others with the average level of accuracy 98.6%. Our proposed method, despite its simplicity, also performs well with the average level of accuracy 96.8%. The empirical study also suggests that two among the seventeen characteristics are insignificant for the classification accuracy.

## References

[1] 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*. 2007; 17 (3): 231–251.
[2] H. Whittle, C. Hamilton-Giachritsis, A. Beech, and G. Collings. A review of online grooming: Characteristics and concerns. *Aggression and violent behavior*. 2013; 18(1): 62–70.
[3] D. Michalopoulos, and I. Mavridis. *Utilizing document classification for grooming attack recognition*. IEEE Symposium on Computers and Communications (ISCC). 2011: 864–869.
[4] A. Kontostathis, L. Edwards, and A. Leatherman. Text mining and cybercrime, Text Mining: Applications and Theory. John Wiley & Sons Ltd, Chichester, UK. 2010.
[5] K.J. Mitchell, D. Finkelhor, and J. Wolak. Police posing as juveniles online to catch sex offenders: Is it working? *Sexual Abuse: A Journal of Research and Treatment*. 2005; 17(3): 241–267.
[6] P. Briggs, W.T. Simon, and S. Simonsen. An exploratory study of internet-initiated sexual offenses and the chat room sex offender: Has the internet enabled a new typology of sex offender? *Sexual Abuse: A Journal of Research and Treatment*. 2010; 1079063210384275.
[7] S. Webster, J. Davidson, A. Bifulco, P. Gottschalk, C.V.T. Pham, J. Grove-Hills, C. Turley, C. Tompkins, S. Ciulla, V. Milazzo, A. Schimmenti, and G. Craparo, European online grooming project (2012). http://www.europeanonlinegroomingproject.com/media/2076/european-online-grooming-project-final-report.pdf
[8] I. McGhee, J. Bayzick, A. Kontostathis, L. Edwards, A. McBride, and E. Jakubowski. Learning to identify internet sexual predation. *International Journal of Electronic Commerce*. 2011; 15(3): 103–122.
[9] S.J. Pandey, I. Klapaftis, and S. Manandhar. Detecting predatory behaviour from online textual chats, in: Multimedia Communications, Services and Security. Springer. 2012: 270–281.
[10] H. Pranoto, F.E. Gunawan, and B. Soewito. Logistic models for classifying online grooming conversation. *Procedia Computer Science*. 2015; 59:357–365.
[11] B. Baharudin, L. H. Lee, and K. Khan. A review of machine learning algorithms for text-documents classification. *Journal of advances in information technology*. 2010; 1(1): 4–20.
[12] I. Babaoglu, O. Findik, and E. Ulker. A comparison of feature selection models utilizing binary particle swarm optimization and genetic algorithm in determining coronary artery disease using support vector machine. *Expert Systems with Applications*. 2010; 37(4): 3177–3183.
[13] U.R. Acharya, E. Ng, J.H. Tan, and S.V. Sree, Thermography based breast cancer detection using texture features and support vector machine, *Journal of medical systems.* 2012; 36(3): 1503–1510.
[14] S.J. Horng, M.Y. Su, Y.H. Chen, T.W. Kao, R.J. Chen, J.L. Lai, and C.D. Perkasa, A novel intrusion detection system based on hierarchical clustering and support vector machines. *Expert systems with Applications*. 2011; 38(1): 306–313.
[15] R.A. O'Connel, Typology of cybersex exploitation and online grooming process, Tech. rep., Cyberspace Research Unit, University of Central Lancashire, the United Kingdom (2014). http://netsafe.org.nz/Doc_Library/racheloconnell1.pdf
[16] V. Gupta, and G. Lehal. A survey of text summarization extractive techniques, *Journal of Emerging Technologies in Web Intelligence*. 2010; 2(3): 258–268.
[17] N. Christianni, and J. Shawe Taylor. An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods. Cambridge University Press, Cambridge, UK. 2000.
[18] T. Hastie, R. Tibshirani, and J. Friedman. The Elements of Statistical Learning. Springer, New York. 2008.
[19] B. Baharudin, L.H. Lee, and K. Khan, A review of machine learning algorithms for text-documents classification. *Journal of advances in information technology*. 2010; 1(1): 4–20.
[20] D.M. Powers, Evaluation: From precision, recall, and F-measure to roc, informedness, markedness & correlation. *Journal of Machine Learning Technologies*. 2011; 2(1): 37–63.
[21] P. Justice, www.perverted-justice.com, data are downloaded from the site on January 2016

[22] M. Wollis. A linguistic analysis of online predator grooming. Disertation. College of Agriculture and Life Sciences; 2011.
[23] Literotic, http://www.literotic.com/, data are downloaded from the site on January 2016 (July 2016).
[24] E. Bergen, J. Davidson, A. Schulz, P. Schuhmann, and A. Johansson. The effects of using identity deception and suggesting secrecy on the outcomes of adult-adult and adult-child or -adolescent online sexual interactions. *Victims and Offenders*. 2014; 9: 276–298.