UTM UNIVERSITI TEKNOLOGI MALAYSIA
INTERNATIONAL JOURNAL OF INNOVATIVE COMPUTING
ISSN 2180-4370
Journal Homepage : https://ijic.utm.my/

Classification of Online Grooming on Chat Logs Using Two Term Weighting Schemes

Nur Rafeeqkha Sulaiman, Maheyzah Md. Siraj
School of Computing
Universiti Teknologi Malaysia
81310 UTM Johor Bahru, Johor, Malaysia
Email: rafeeqkha@graduate.utm.my, maheyzah@utm.my

Submitted: 24/07/2019. Revised edition: 12/10/2019. Accepted: 13/10/2019. Published online: 28/11/2019
DOI: https://doi.org/10.11113/ijic.v9n2.239

***Abstract*—Due to the growth of Internet, it has not only become the medium for getting information, it has also become a platform for communicating. Social Network Service (SNS) is one of the main platform where Internet users can communicate by distributing, sharing of information and knowledge. Chatting has become a popular communication medium for Internet users whereby users can communicate directly and privately with each other. However, due to the privacy of chat rooms or chatting mediums, the content of chat logs is not monitored and not filtered. Thus, easing cyber predators preying on their preys. Cyber groomers are one of cyber predators who prey on children or minors to satisfy their sexual desire. Workforce expertise that involve in intelligence gathering always deals with difficulty as the complexity of crime increases, human errors and time constraints. Hence, it is difficult to prevent undesired content, such as grooming conversation, in chat logs. An investigation on two term weighting schemes on two datasets are used to improve the content-based classification techniques. This study aims to improve the content-based classification accuracy on chat logs by comparing two term weighting schemes in classifying grooming contents. Two term weighting schemes namely Term Frequency – Inverse Document Frequency – Inverse Class Space Density Frequency (TF.IDF.ICSdF) and Fuzzy Rough Feature Selection (FRFS) are used as feature selection process in filtering chat logs. The performance of these techniques were examined via datasets, and the accuracy of their result was measured by Support Vector Machine (SVM). TF.IDF.ICSdF and FRFS are judged based on accuracy, precision, recall and F score measurement.**

***Keywords*—TF.IDF.ICSdF, FRFS, SVM, online grooming, classification**

I. INTRODUCTION

Social Networking Service (also social networking site, SNS or social media) is an online platform where Internet users use it to create social networks or social connections with other people. SNS users can create virtual profiles with personal information, photos, interests and etc. to form connections with other people. Since their introduction to the public, it has engrossed many users, who have integrated SNS in their daily activities or daily lives. Most of SNS that public use supports the existing social networks, but other SNS help strangers to communicate or know each other based on shared interests, hobbies, and even locations. Some sites are catered to various users, while other sites cater only to groups who share the same spoken language or the same race, gender, and even nationality.

Due to the presence of Internet, more and more time is spent online, thus, exposing minors to both positive and negative side of the Internet and one of the negative side of the Internet is online sexual grooming. Online sexual grooming or cyber grooming is defined as a process that is used by a person to approach, persuade, and engage a minor, who is the victim, in sexual activity through the internet. In other words, it is the process of preparing a child to participate in illegal sexual encounters by adult. Cybergroomers use the Internet platforms that are widely used by young people, such as social networking sites or chatroom, to approach the victim since the platforms allow cybergroomers to communicate anonymously or pseudonymously [1].

One example of cyber grooming is when an adult exchange sexual context and media by using the social media networking sites or chatrooms by pretending to be someone of the same age as the minors. By pretending to be the same age as the minors, the minors will be more comfortable and less suspicious to befriend the cybergroomers. Once the cybergroomers have the minors' trust, the minors will easily reveal their personal information and easily initiate a relationship with the minors. This type of harassment need to be stooped as it can affect the child emotionally and physically. Due to the experience of being harassed, their academic performance, social and psychological well-being can also be affected and haunt them for the rest of their lives [2].

II. Survey on Online Grooming

This section starts on overview of Machine Learning is the sentiment analysis and followed by machine learning. In the overview of machine learning, the application of machine learning in cyber security and text classification are discussed. Lastly, online grooming classification and its processes is reviewed.

A. Machine Learning

Machine learning is defined as computational methods using experience to improve performance or to make accurate predictions. Experience in this context states the former information that is accessible by the learner in the form of collection of electronic data and used it for analysis. The data are in the form of digitalized training sets labelled by human or any information gathered by interacting with the environment. Process of machine learning consists of modelling sufficient and accurate prediction algorithm with additional sample complexity to assess the sample size needed for the algorithm to learn the concept intended for the research. Basically, the precision of an algorithm depends on the difficulty of the idea classes measured and the scope of the training model [3].

Machine learning algorithms have been used by data scientists to discover patterns in big data. These algorithms can be grouped in to two groups based on how it learns to make predictions: supervised and unsupervised. Supervised machine learning which is the commonly used machine learning algorithms where it includes algorithms such as linear and logistic regression, multi-class classification, and SVM. It is named as supervised learning because it is guided by data scientists to teach the algorithm. The possible output of the supervised learning algorithm is identified and the training data is labelled with the right answers. Unsupervised learning is a true artificial intelligence where it has the idea where the machine learns on identifying patterns and difficult processes on its own [4].

There are three main processes in machine learning. The first process is input, followed by processing and model. Input process is the data collected. In processing process, algorithms are used to pre-process the data, apply learning algorithms and error analysis. Lastly, a model is built based on the data processed. In terms of technique, the machine learning has various kind of technique including SVM, Naïve Bayes, Classical Neural Network, Maximum Entropy and Bayesian Network.

1) Machine Learning in Online Security

Online security is a set of machineries and procedures that is intended to secure the computers, systems, programs, and data from attackers, illegal access, alteration, or damage. Machine learning can be used to enhance cyber security by providing assistant in collecting data and learn on how to solve the problem relevant to the data collected. Machine learning application is much easier than solving problems by hand. For example, email is the most favorable medium for phishing. Attackers steal a victim's personal data by getting the domain name of a certain web page and reroute the website's traffic to a fake website meant for phishing by sending a number of e-mails to the victim. Machine learning can be used to learn examples of phishing emails and normal emails to generate a model to detect phishing emails [5].

B. Related Work on Online Grooming Classification

Online grooming is defined as a procedure of approaching, persuading, and minors in appropriate (sexual) activities through the internet. The online groomers introduce themselves a child in order to connect with them sexually and emotionally[6]. Investigators analyze conversation texts in order to detect grooming patterns as manual approach on grooming detection is error prone. As a result, an automated system is developed to analyze chat logs of their conversation and to detect the likelihood of a grooming chat [7].

Various pattern of detection schemes has been addressed which are k-mean clustering by Kontostathis, Edwards and Leatherman [7], a ruled-based approach by McGhee, Bayzick, Kontostathis, Edwards, McBride and Jakubowski [8], SVM by Pandey, Klapaftis and Manandhar [9], and a logistic regression model by Pranoto, Gunawan and Soewito [10]. These detection schemes detect online grooming detection by identifying the main features of the style of the conversations.

The process of identifying a child grooming conversation is a difficult process because it differs in period, nature and strength depending on the criminal behavior. Generally, the processes of an online grooming conversations have been identified by both O'Connell [11] and Gupta and Lehal [12]. However, these stages may differ from one chat to another and may not follow the order.

The first stage of online grooming is the friendship making phase. In this stage, the groomer introduces himself to the victim and then establishes a possibility of exchanging personal detail such as name, age, location, and

gender. Besides that, the criminal asks for other online information which is connected to the victim and requesting photos of the child to approve that the victim is really a minor. Relationship stage comes after friendship is formed between the child and the criminal. They start to discuss on normal topics like school, family, hobbies and the interest of the victim so that the child can be misled into thinking they are in a relationship. The third stage is the risk assessment stage. This stage refers to the conversation where the groomer asks the child on the location and the users of the computer. This information helps the groomer to evaluate the probability of being noticed by other people such as the child's guardian or parents.

The fourth stage is the exclusivity stage. Cyber groomer tries to gain the victim's belief by inserting the idea of love and care into their conversation. Next, comes the sexual stage where they start talking about sexual activities and start creating sexual fantasies. Lastly, the groomer approaches the child to meet in person. The next step after categorization of online grooming process includes pre-processing, features extraction, features selection, and lastly classification.

1) Feature Selection

Transformation of text document into a feature vector is a necessary process when performing text classification. By transforming text documents into feature vectors, the vector can be processed by the computer and be performed statistical analysis on it. In order to do so, each feature of the vector, or each component, is predefined to transform every text into a vector with similar structure and size. Throughout this transformation, it is essential to increase as many features as possible to the feature vector so that information will not lose from the original text. Thus, feature selection is used at a later time to lessen the amount of features to increase the accuracy.

Feature selection mechanism plays an important role in data mining. Data mining operation deals with redundant feature. Some of the features are important while some of them may not be irrelevant to our research. Hence, feature selection plays an important role in eliminating the redundant features and extract the information. Feature selection should be done accurately so as to prevent affecting the data accuracy and producing different output than the original information. According to Meyer [13], feature selection is a compulsory process due to several reasons: it helps us in reducing noise in dataset, it helps in reducing computational load, and it helps in reducing overfitting which occurs in a complex model.

a) Fuzzy Rough Feature Selection

Fuzzy-rough feature selection (FRFS) delivers a resources whereby discrete or real-valued noisy-data (both) are efficiently reduced without user supplying information to the algorithm. FRFS can be used on data that has continuous or nominal decision attributes, and can be applied to regression as well as classification datasets. Extra information that is required for FRFS is in the form of fuzzy partitions for each feature that can be spontaneously derived from the data [14].

FRFS compresses the connected but then different ideas of fuzziness (from fuzzy sets) and indiscernibility (from rough sets) whereby these sets are a result of doubt in information. Fuzzy sets come from due to the lack of difference in the data itself while rough sets are said to model uncertainty ensuing from the absence of approximation through set approximation [15].

A research done by Zuo, Li, Anderson, Yang and Naik [15] to detect online grooming implemented TF.IDF or Bag of Words (BoW) to identify terms for classification process and afterwards with Fuzzy-Rough Feature Selection in order to identify the uncertainty that includes in the nature of natural language conversation. In this research, the authors used two types of dataset which are training dataset and online grooming dataset categorized into three categories namely Normal, Pedophile, and Sex. The proposed grooming detections consists of 4 steps: test pre-processing, text feature extraction, text feature selection, text feature normalization, and classification. Before applying FRFS, the processed dataset went through TFIDF or Bow in order to generate uniformed document representation for each data instance with unified length.

b) Term Frequency-Inverse Document Frequency-Inverse Class Space Density Frequency (TF.IDF.ICS$_\delta$F)

Ren and Sohrab [16] suggested a new approach based on TFIDF called Term Frequency-Inverse Document Frequency-Inverse Class Space Density Frequency (TF.IDF.ICS$_\delta$F). this method provides a positive one-sidedness on both mostly used and less used terms. Various researchers have attempted to improve the performance of Text Classification by manipulating statistical classification methods and machine learning methods. However, good indexing technique for a novel term weighting scheme is very much wanted to truly enhance the classification task.

TF.IDF.ICS$_\delta$F is proposed in order to make the following contributions to text classification research area: to suggest an automatic indexing method using the combination of document-based and class (category)-based approaches, the proposed TF.IDF.ICSδF term weighting gives a positive discrimination both to rare and frequent terms, the proposed TF.IDF.ICSδF approach is effective in high-dimensional and comparatively low-dimensional vector spaces, and the proposed class-indexing method expands the existing document-indexing method in term indexing and generates more informative terms based on a certain category through use of inverse class frequency (ICF) and ICSδF functions.

2) Classification

Text classification technology has a very important part in the use of document data. This is because traditional text classification can no longer aids in the growth of agile information. Statistical-based automatic text categorization has been applied in many fields due to it becoming a popular research point in the academic and industrial world as an efficient and practical technique. Because of the ever-growing Internet and new features of the internet, such as diverse data and extensive sources, an in-depth research on related technologies of text classification still need to be carried out [17].

The main goal of text classification is to decide whether the labelled term is alike another labelled text. Assume that a set of different dataset has a label or a class attached to it. The dataset will be transformed into feature vectors in pre-processing process and a model is built by using machine learning in classification process. Fig. 1 shows a sample of text classification.

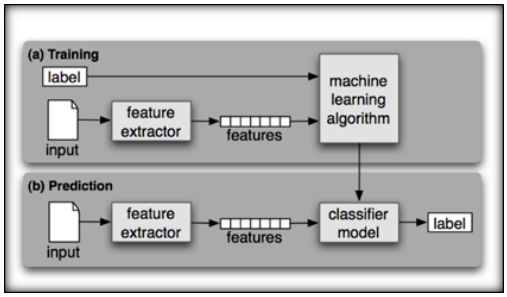

Fig. 1. Sample of text classification

a) Support Vector Machine (SVM)

Support Vector Machine (SVM) is one of the most theoretically well motivated and practically most effective classification algorithms in modern machine learning. SVM helps in working out the decision boundary that gives the optimal margin for classification task. In other words, SVM is a numerical method to calculate a hyperplane for separating two-class dataset. There are two types of case where SVM algorithm is used: linearly separable data and non-linear separable data [13].

For linearly separable data, SVM calculates on how to maximize the margin of hyperplane in order to increase the distance between decision boundary and the data of both classes. If the data is not possible to be separated (non-linear separable), the SVM algorithm automatically add an error term that approximates the number of misclassified samples.

Implementing SVM classifier techniques for classification is a good practice as SVM as the algorithm gives many advantages. SVM are known to have good generalization properties, to be insensitive to overtraining and to the curse-of-dimensionality due to its margin maximization and the regulation term [18].

b) Naïve Bayes

Bayessian classifiers assumes on how the data is created, and suggest a probabilistic model that symbolizes the assumptions made; then, labelled training example is used to evaluate the parameters of the model. Bayer's rule is used to classify new examples by selecting a specific class that is very likely created the example.

The simplest Bayes classifier is naïve Bayes model or idiot Bayes model. It is called the simplest because it assumes that all attributes of the examples are independent of each other based on the context of the class and this is called "naïve Bayes Assumptions". This assumption is however not true in real world tasks and naïve Bayes performs classification well. Due to the independent assumption, the parameters for each attribute can be learned separately which critically shortens learning, especially when the amount of the attributes are large [19].

3) Data Collection

Data collection is the initial step of this study where data collection of proper data or data set related to this study is performed. The collected data from various sources will undergo various processes in order to provide a suitable input for further processes such as feature selection and text classification. Datasets used for online grooming conversation classification consists of grooming conversations chat logs and non-grooming scripts.

a) Perverted-Justice Dataset

Perverted-justice (PJ) is a website that contains conversation scripts from various chatting websites. It contains almost 700 chat conversations between convicted online groomer and members of PJ posing as minors. The chats uploaded in this sites are all proven as online grooming conversation [10].

This website is formed by internet users who voluntarily participate in what Zingerle [20] described as "*vigilantes or cyber-vigilante communities, becoming self-appointed avengers of justice who wade through the Internet to hunt down unlawful netizens*." PJ members thought that the normal way of punishments by the law for criminal punishment is ineffective so they crowdsource with both the trial and the implementation of punishment by making use of social networks.

Galán-García, Puerta, Gómez, Santos and Bringas [21] states that conversations in PJ includes both real grooming conversations between online groomers and children, and conversations between PJ members who uses fake profiles to disguise as minors to make contact with predator. The recorded chat and phone conversations, and real life meetings become the evidences that they use to convict the

predators. PJ community also participate in the Dateline NBC investigative news program "To Catch a Predator"[20].

b) Literotica Dataset

www.literotika.com is an amateur porn stories website that contains conversation scripts of people expressing their passion legally [22]. According to Johnsdotter [23], Literotica is a well-attended website as stated by the company Compete. Compete is a company that provides site statistics; during the past years, Literotica had between 2.6 and 3.1 million unique visitors. It is one of the oldest and biggest erotic literature sources online comprising approximately 1.25 billion words [24].

c) Pre-Processing

Pre-processing is one of the process in text classification process. It is proven that time taken for pre-processing process takes from 50% up to 80% out of the whole text classification process. Thus, proving how important is pre-processing in text classification [25].

Pre-processing phase is the process of identifying and categorizing the most important features by converting textual data into data-mining ready data. The main goal of pre-processing is to convert documents into feature vectors to differentiate the text into exact words [26].

The conversations in this study will go through the following steps: tokenization: this process removes non-letter characters in the document and each document is partitioned into words. Tokenization also refer to as lexical analysis or text segmentation, transformation: words in the document is transformed into lowercase, stopping: 'Stop-words' are removed in this process, stop words are words that are frequently used in English and are not relevant in Information Retrieval (IR). Examples of stop words are 'the', 'of', 'and', 'to'[26], and stemming: this technique is the process of converting a word to its root or stem. In this process, a word is converted to its stem, which combines a language-dependent linguistic knowledge. For instance: the words wait, waits, waited, and waiting can be stemmed to the word 'WAIT'.

III. Our Proposed Framework

This section discusses on the details of the approach to reach the objectives of the research. The goal of this research as stated earlier is to detect online grooming towards minors through chat logs conversation by comparing two Term Weighting Schemes in classifying online grooming conversation logs. This research focus on using algorithms and datasets for testing and training. The algorithms used are TF.IDF.ICS$_\delta$F and Naïve Bayes algorithm, and SVM is used to identify and classify feature vectors. This chapter explains on the phases involved in the framework and their implementation. This chapter will be used as a guideline for overall process in this research.

A. Research Framework

Research framework shows the outlines of the research. Each phase matches to each objective. Since there are three objectives, three phases were identified in the research framework. The phases are labelled as Phase 1, Phase 2, and Phase 3. Each phase generates input for the following phase. The first step is the initial step which initiate the process of this research. Phase 1 begins with data collection and the collected data will then be pre-processed and represented into a proper form of structure. In Phase 2, feature selection and classification will be carried out. In this phase, the grooming conversation and non-online grooming conversation are classified. In Phase 3, the performance of term weighting schemes in Phase 2 is evaluated. Fig. 2 shows the research framework.

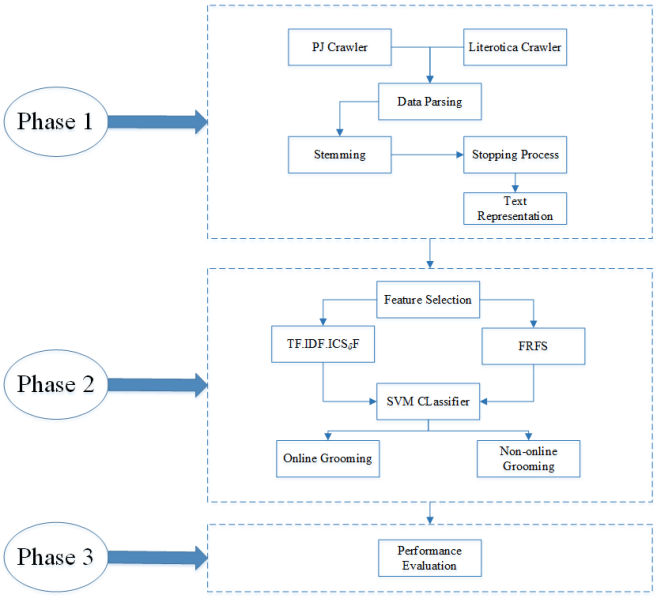

Fig. 2. Research framework

1) Phase 1

Phase 1 consists of data collection, pre-processing, and text representation. The data collection is the process where dataset from Perverted Justice (PJ) and Literotica is collected and built. The dataset collected will be used in Phase 2 for testing and training of data. Before pre-processing, the data will go through tokenization, transformation, stopping, and stemming to remove unwanted or repeated words to avoid any misrepresentation in the final result. Text representation will be used in Phase 2 for feature selection.

The data collection is the process where dataset from Perverted Justice (PJ) and Literotica is collected and built. The dataset collected will be used in Phase 2 for testing and training of data. In order to download texts from PJ and

Literotica, a crawler is used. The crawler is created in python by using python modules urlib and beautifulsoup4. The data is collected randomly in PJ and literotica.com. Urlib4 module fetches URLs using a variety of protocols in a very simple interface. Beautifulsoup4 is a python library for pulling data out of HTML and XML files. Beautifulsoup4 takes a little time to collect data and save hours or days of work. Unnecessary data object like usernames, images and other non-text subjects were excluded from crawling and only text was taken out for pre-processing. In Phase 2, this dataset is used for training and testing.

Pre-processing phase is the process of converting textual data into a data-mining ready structure, where the most important text-features that help to distinguish between text-categories are identified. The main goal of pre-processing is to represent each document as a feature vector to distinct the text into specific words. Stemming process transforms a word into its form while stopping is the process that eliminates stop-words such as 'and' and 'or'. Before classifying the dataset, the data need to be represented in a form that is understandable by classifiers. A correct representation of the dataset helps in increasing the accuracy and speed up the process. Text representation is used for classification phase.

a) Dataset

This study will use two datasets taken from Perverted-Justice and Literotica. Perverted-justice (PJ) is a website that contains conversation scripts from various chatting websites. It has more than 500 grooming conversations between grooming children predators with juvenile victims or law enforcement posing as teenagers. It contains more than 500 chat logs between online groomers and victims. Thus, there are 550 chat logs taken from this website to be used in this study. On the other hand, Literotica is an amateur porn stories website that contains conversation scripts of people expressing their passion legally. For research purposes, 200 data randomly selected from Literotica. Table I shows a description of the dataset labels.

TABLE I. Description of Dataset Labels

	Category	Label	Total
1	Perverted-Justice Dataset	PJ	400 chat logs
2	Literotica Dataset	LT	100 scripts

Two datasets will be used to do the proposed research. The first data set is a balanced dataset and contained 300 chat logs and 100 scripts. The second dataset is a balanced dataset containing 250 chat logs and 200 scripts. The reason for creating balanced and imbalanced dataset is to test the performance of aforementioned algorithms with different types of datasets. Imbalanced datasets closely resemble the real life data that algorithms might need to train on. Thus, it is imperative that the learning capacity of feature selection algorithm is independent of skew in data. Table II shows the usage of datasets.

TABLE II. Usage of Datasets to Perform Research

	Category	Training	Testing	Total
1	PJ	150	150	300
	LT	50	50	100
2	PJ	150	50	250
	LT	80	20	100

2) Phase 2

Objective of Phase 2 is to select and classify data based on feature selection: Term Frequency – Inverse Document Frequency – Inverse Class Space Density Frequency (TF.IDF.ICSδF) and Fuzzy-rough Feature Selection. Two main processes are executed in this phase which are feature selection and classification process. Phase 2 has the most critical work in the research framework.

Feature selection mechanism plays an important role in data mining. Data mining operation deals with redundant feature. Some of the features are important while some of them may not be irrelevant to our research. Hence, feature selection plays an important role in eliminating the redundant features and extract the information. Two term weighting schemes will be used in this research framework: TF.IDF.ICS$_\delta$F and Fuzzy-rough Feature Selection.

The main goal of text classification is to decide on a label for a text based on how alike it is to other text that is already labelled. The dataset will be transformed into feature vectors in pre-processing process and a model is built by using machine learning in classification process. Support Vector Machine will be used a classifier for the dataset. SVM is a numerical method to calculate a hyperplane for separating two-class dataset.

3) Phase 3

This is the last phase in this research. The results from classification process will be evaluated. The objective for this phase is to evaluate the accuracy of two term weighting schemes used during feature selection.

Performance evaluation is used to evaluate or measure the performance of the two weighting scheme. This phase will find the accuracy of detecting online grooming conversation.

Accuracy is the fraction of the classification result that are correct.

$$accuracy = \frac{TP + TN}{(TP + TN + FP + FN)} \quad (1)$$

Precision is the fraction of the predicted documents in a class that are correct.

$$precision = \frac{TP}{(TP + FN)} \quad (2)$$

Recall is the fraction of documents in a class that correctly predicted.

$$recall = \frac{TP}{(TP + FN)} \quad (3)$$

F-score is a weighted harmonic mean of precision and recall.

$$F\ score = \frac{2 \times precision \times recall}{precission + recall} \quad (4)$$

The results will be presented in the form pf graphs, tables and figures. The results will be evaluated for accuracy, precision, recall and F score. Table III shows the terminologies used in performance evaluation.

TABLE III. The Terminologies Used in Performance Evaluation

Terminology	Definition
TP	True Positive
TN	True Negative
FP	False Positive
FN	False Negative

IV. Conclusion

In this paper, the techniques on online grooming classification are reviewed and discussed along with feature selection techniques. Lastly, a framework on classification on online grooming based on chat logs is proposed. The datasets used for this projects are also stated and formulas for performance evaluations are previewed in tables.

References

[1] M. Ashcroft, L. Kaati, and M. Meyer. (2015). A Step Towards Detecting Online Grooming -- Identifying Adults Pretending to be Children. *2015 European Intelligence and Security Informatics Conference*, 98-104.

[2] H. C. Whittle, C. Hamilton-Giachritsis, and A. R. Beech. (2013). Victims' Voices: The Impact of Online Grooming and Sexual Abuse, *Universal Journal of Psychology*, 1(2), 59-71.

[3] M. Mohri, A. Rostamizadeh, and A. Talwalkar. (2012). *Foundations of Machine Learning*, 5, MIT Press.

[4] N. Castle. (2018). Supervised vs. Unsupervised Machine Learning, 2/12/2018, 2018; https://www.datascience.com/blog/supervised-and-unsupervised-machine-learning-algorithms.

[5] S. Abu-Nimeh, D. Nappa, X. Wang *et al.* (2007). A Comparison of Machine Learning Techniques for Phishing Detection. *Proceedings of the Anti-Phishing Working Groups 2nd Annual eCrime Researchers Summit 2007, Pittsburgh, Pennsylvania, USA, October 4-5, 2007*, 60-69.

[6] J. Wolak, K. J. Mitchell, and D. Finkelhor. (2006). *Online Victimization of Youth: Five Years Late*r, National Center For Missing & Exploited Children.

[7] A. Kontostathis, L. Edwards, and A. Leatherman. (2010). *Text Mining and Cybercrime, Text Mining: Applications and Theory.* John Wiley & Sons, Ltd, Chichester, UK.

[8] I. McGhee, J. Bayzick, A. Kontostathis *et al.* (2011). Learning to Identify Internet Sexual Predation. *International Journal of Electronic Commerce*, 15(3), 103-122.

[9] S. J. Pandey, I. Klapaftis, and S. Manandhar. (2012). Detecting Predatory Behaviour from Online Textual Chats. *International Conference on Multimedia Communications, Services and Security*, 270-281.

[10] H. Pranoto, F. E. Gunawan, and B. Soewito. (2015). Logistic Models for Classifying Online Grooming Conversation. *Procedia Computer Science*, 59, 357-365.

[11] R. O'Connell. (2003). *A Typology of Child Cybersexploitation and Online Grooming Practices*. Preston: University of Central.

[12] V. Gupta, and G. S. Lehal. (2010). A Survey of Text Summarization Extractive Techniques. *Journal of Emerging Technologies in Web Intelligence*, 2(3), 258-268.

[13] M. Meyer. (2015). *Machine Learning to Detect Online Grooming*.

[14] R. Jensen, and Q. Shen. (2009). New Approaches to Fuzzy-rough Feature Selection, *IEEE Transactions on Fuzzy Systems*, 17(4), 824.

[15] Z. Zuo, J. Li, P. Anderson *et al.* (2018). Grooming Detection using Fuzzy-Rough Feature Selection and Text Classification. *FUZZ-IEEE 2018-IEEE International Conference on Fuzzy Systems, 8th-13th July 2018, Rio de Janeiro, Brazil*, 1-8.

[16] F. Ren, and M. G. Sohrab. (2013). Class-indexing-based Term Weighting for Automatic Text Classification. *Information Sciences*, 236, 109-125.

[17] C. Liu, Y. Sheng, Z. Wei *et al.* (2018). Research of Text Classification Based on Improved TF-IDF Algorithm. *IEEE International Conference of Intelligent Robotic and Control Engineering (IRCE)*, 218-222.

[18] F. Lotte, M. Congedo, A. Lécuyer *et al.* (2007). A Review of Classification Algorithms for EEG-based Brain–computer Interfaces. *Journal of Neural Engineering*, 4(2), R1.

[19] A. McCallum, and K. Nigam. (1998). A Comparison of Event Models for Naive Bayes Text Classification. AAAI Technical Report WS-98-05, 41-48.

[20] A. Zingerle. (2015). Scambaiters, Human Flesh Search Engine, Perverted justice, and Internet Haganah: Villains, Avengers, or Saviors on the Internet? *ISEA Conference*.

[21] P. Galán-García, J. G. d. l. Puerta, C. L. Gómez *et al.* (2016). Supervised Machine Learning for the Detection of Troll Profiles in Twitter Social Network: Application to a Real Case of Cyberbullying. *Logic Journal of the IGPL*, 24(1), 42-53.

[22] F. E. Gunawan, L. Ashianti, S. Candra *et al.* (2018). Detecting Online Child Grooming Conversation. *TELKOMNIKA (Telecommunication, Computing, Electronics and Contro*l, 1-6.

[23] S. Johnsdotter. (2011). The Flow of Her Cum: On a Recent Semantic Shift in an Erotic Word. *Sexuality & Culture*, 15(2), 179-194.

[24] A. Lischinsky, and K. Gupta. (2017). Distant Reading Intimate Encounters: A Big Data Approach to Online Erotica.

[25] K. Morik, and M. Scholz. (2004). The Miningmart Approach to Knowledge Discovery in Databases. *Intelligent Technologies for Information Analysis*, 47-65, Springer.

[26] V. Srividhya, and R. Anitha. (2010). Evaluating Preprocessing Techniques in Text Categorization. *International Journal of Computer Science and Application,* 47(11), 49-51.