Grooming Detection using Fuzzy-Rough Feature Selection and Text Classification

Zheming Zuo, Jie Li, Philip Anderson, Longzhi Yang
Department of Computer and Information Sciences,
Faculty of Engineering and Environment
Northumbria University, Newcastle upon Tyne, UK
Email: {zheming.zuo, jie2.li, philip.anderson, longzhi.yang}@northumbria.ac.uk

Nitin Naik
Defence School of
Communications of Information Systems
Ministry of Defense, UK
Email: nitin.naik100@mod.gov.uk

***Abstract*—Online child grooming detection has recently attracted intensive research interests from both the machine learning community and digital forensics community due to its great social impact. The existing data-driven approaches usually face the challenges of lack of training data and the uncertainty of classes in terms of the classification or decision boundary. This paper proposes a grooming detection approach in an effort to address such uncertainty based on a data set derived from a publicly available profiling data set. In particular, the approach firstly applies the conventional text feature extraction approach in identifying the most significant words in the data set. This is followed by the application of a fuzzy-rough feature selection approach in reducing the high dimensions of the selected words for fast processing, which at the same time addressing the uncertainty of class boundaries. The experimental results demonstrate the efficiency and efficacy of the proposed approach in detecting child grooming.**

I. Introduction

Along with the rapid adoption of the Internet, smart phones and social networking Apps, more than 90% of the population in the UK are able to communicate through the cyberspace, in which one-quarter of them are under 25 years old [1]. Currently, using the Internet has been a part of children/teenagers' lives, such as gaming and socializing, etc. However, people can easily hide their identity online, which poses a great risk to children and teenagers. In fact, children and teenagers may not fully understand the risks that they are facing on the Internet, which could place them into many dangerous situations, such as providing personal information when talking to strangers online. Consequently, children and teenagers can be easily groomed online, that someone builds an emotional connection with the child or teenager to gain their trust for the purposes of sexual abuse, sexual exploitation or trafficking.

According to the report from National Society for the Prevention of Cruelty to Children (NSPCC) in 2014, 12% of 11-16 years old young people in the UK have received unwanted sexual messages, and 8% of 11-16 years old young people in the UK have received requests to send or respond to a sexual message [1]. This is further supported by a survey conducted by Barnardo's of their sexual exploitation services across the UK which indicated that 42% of the children supported had been groomed online [2]. Such crimes can affect the victim's life psychologically, physically, emotionally, behaviorally and psycho-socially [3] not only at the time the crime occurred, but potentially for many years after. Grooming detection is therefore very important for protecting children and teenagers and indeed is one of the many important tasks performed by the law enforcement agencies. However, digital forensic investigations are often time-consuming, adding to an already increasing backlog of investigations [4]. In addition and potentially, they can have an adverse psychological effect on the investigators.

A grooming detection system is therefore of great importance to automate the detection process, by analyzing the conversation text, such as chat room logs, text messages, emails or mobile Apps, in detecting the possible child grooming conversations [5]. A number of machine learning algorithms have been applied to address such issues in the literature. For instance, a rule-based classification approach was applied to classify the chat logs, thus to label the predatory posts [6]. The Support Vector Machine (SVM) was then adopted as the classifier to solve the text classification task, thus to raise an alarm when a grooming type activity in a conversation is detected [7]. A grooming detection system was proposed by establishing a logistic mathematical model based on some frequently appeared key characteristics to classify online conversation logs [8]. However, the uncertainty naturally included in the conversions may affect the efficiency of these approaches.

This paper proposes an approach for automatic child grooming detection using fuzzy-rough feature selection in an effort to address the uncertainty that comes with the nature of natural language conversation. In particular, the work firstly uses the bag of words (BoW) or term frequency-inverse document frequency (TF-IDF) features to identify a list of words for text classification in the context of digital forensics, which is followed by the application of fuzzy-rough feature selection. From this, a classifier is applied based on the extracted features for text classification. The project also extracted a grooming data set from the publicly available PAN'13 data set to support this study. The experiments demonstrate the work of the approach with promising results generated.

The rest of the paper is structured as follows. Section II introduces the theoretical underpinnings of fuzzy-rough feature selection method. Section III presents the proposed grooming detection approach. Section IV details the experiments for comparison and validation. The paper is concluded in Section V with future developments suggested.

978-1-5090-6020-7/18/$31.00 ©2018 IEEE

II. Background

The existing grooming detection methods and the fuzzy-rough feature selection approaches are briefly reviewed in this section.

A. Grooming Detection

The procedure of a conventional grooming detection system can usually be summarized in four steps: special characteristics identification, feature extraction, features selection and classification [5].

Special Characteristics Identification: Online child grooming conversation texts are usually complex, which are highly depending on the perpetrator characteristics and behaviors. As a consequence, the common pattern of the grooming characteristics may not be easily identified. However, based on the previous research, an online child grooming process would usually go through typical stages or progresses, and thanks to these stages, the related important grooming characteristics can be reasonably identified [5].

Feature Extraction: In general, different text data sources contain different number of characteristics and usually very noisy, which cannot be readily forwarded to the classifier. The aim of feature extraction is to extract the most important features, thus to build a uniform document representation for a given data set, with the Term Frequency-Inverse Document Frequency (TF-IDF) [9] being most commonly used in this setting. In particular, the features of the data are words or combinations of words from a list or a dictionary ($\mathbb{T} = T_1, \cdots, T_n$) that have been identified from the previous step, where $\mathbb{T}$ denotes the created dictionary and n is the number of the words in the dictionary. For a given data set $\mathbb{S}$ that contains m data instances (i.e., $\mathbb{S} = \{s_1, s_2, ...s_m\}$), the progress of the feature extraction extracts the important information of each data instance s_i and represents it into a fixed length feature set associated with their importance, based on the term listed in the created dictionary appeared frequency in data instance. This feature set is then the identify of data instance s_i, which can be expressed as: $s_i = \{w^i_{T_1}, w^i_{T_2}, \cdots, w^i_{T_n}\}$, where $w^i_{T_n}$ represents the importance of feature n in the given data instance s_i.

Features Selection: Based on the number of identified important grooming characteristics, a large number of features may be identified and extracted, which greatly increases the computational cost and also deteriorates the system performance. In order to address such issue, a features selection method is usually employed here to rank the extracted features, which then selects the most important and discriminative features for use in the next step of the process.

Classification: After the process of feature extraction and feature selection, a ready-to-use training data set is utilized and can be applied to the classification algorithm, such as Gaussian Naïve Bayes (GNB) classifier, AdaBoost, the logistic regression (LR) [10] and fuzzy interpolation [11], [12], [13] for system modeling. Note that the performance of classifier can be enhanced by involving an extra step of further normalizing the selected features prior to the classification phase.

B. Fuzzy Rough Feature Selection

Fuzzy-rough sets encapsulate the related but distinct concepts of vagueness (from fuzzy sets) and indiscernibility (from rough sets), both of which occur as a result of uncertainty in knowledge. Vagueness appearance is due to the lack of distinction or hard boundaries in the data itself. This is typical in human communication and reasoning. Rough sets can be said to model ambiguity resulting from a lack of information through set approximations.

A fuzzy-rough set is usually represented as a pair of fuzzy sets which expresses the lower and upper boundaries of the concept. The definitions of the fuzzy lower and upper approximations to approximate a fuzzy concept X are given as follows [14]:

$$\begin{aligned} \mu\underline{R_P}X(x) &= \inf_{y \in \mathbb{U}} I\left(\mu_{R_P}(x,y), \mu_X(y)\right), \\ \mu\overline{R_P}X(x) &= \sup_{y \in \mathbb{U}} T(\mu_{R_P}(x,y), \mu_X(y)), \end{aligned} \tag{1}$$

where I is a fuzzy implicator and T is a t-norm, and R_P is the fuzzy similarity relation induced by the subset of features P:

$$\mu R_P(x,y) = T_{a \in P}\{\mu_{R_a}(x,y)\}, \tag{2}$$

where $\mu_{R_a}(x,y)$ is the degree to which objects x and y are similar for feature a. $\mu_{R_a}(x,y)$ may be defined in many ways, and one example definition is as follows:

$$\mu_{R_a}(x,y) = \begin{cases} 1 - \frac{|a(x) - a(y)|}{|a_{\max} - a_{\min}|}, \\ \exp(-\frac{(a(x)-a(y))^2}{2\sigma_a^2}), \\ \max\left(\min\left(\frac{(a(y)-(a(x)-\sigma_a))}{\sigma_a}, \right.\right. \\ \left.\left. \frac{((a(x)+\sigma_a)-a(y))}{\sigma_a} \right), 0 \right), \end{cases} \tag{3}$$

where σ_a^2 is the variance of feature a. The choice of relation operation is highly depending on the intended application. Generally speaking, the last relation as expressed in Equation (3) may be appropriate for general feature selection purpose [15]. The fuzzy positive region of the decision feature $\mathbb{D}$ on an attribute subset P can be defined in a similar way to the original crisp rough set approach [16], such as:

$$\mu_{POS_{R_P}(\mathbb{D})}(x) = \sup_{X \in \mathbb{U}/\mathbb{D}} \mu_{\underline{R_P}X}(x). \tag{4}$$

An important application of fuzzy-rough sets is the discovery of dependencies between attributes in data analysis. This is of particular significance for feature selection and pattern classification. The fuzzy-rough dependency degree can be defined as:

$$\gamma'_P(\mathbb{D}) = \frac{\sum_{x \in \mathbb{U}} \mu_{POS_{R_P}(\mathbb{D})}(x)}{|\mathbb{U}|} \, . \tag{5}$$

A fuzzy-rough reduct R is defined as a subset of features which preserves the dependency degree of the entire dataset, i.e. $\gamma'_R(\mathbb{D}) = \gamma'_{\mathbb{C}}(\mathbb{D})$. Based on this, a fuzzy-rough feature selection algorithm can be constructed by using Eq. (5) to gauge subset quality. In [16], it has been shown that the dependency function is monotonic and those fuzzy discernibility matrices may also be used to discover the reducts. The fuzzy-rough feature selection algorithm is shown in Algorithm 1. Fuzzy-rough feature selection method can be extended to adapt to a wider range of real-world data mining and knowledge discovery problems such as mammographic risk analysis [17], [18]. However, it has not been applied for solving online child grooming detection problem.

Algorithm 1 The Fuzzy-Rough Feature Selection (FRFS) [16]

```
Require:
    $\mathbb{C}$: the set of all conditional attributes;
    $\mathbb{D}$: the set of decision attributes;
$R \leftarrow \{ \}$; $\gamma'_{best} = 0$; $\gamma'_{prev} = 0$
do
    $T \leftarrow R$
    $\gamma'_{prev} = \gamma'_{best}$
    foreach $x \in (\mathbb{C} - R)$
        if $\gamma'_{R \cup \{x\}}(\mathbb{D}) > \gamma'_{T}(\mathbb{D})$
            $T \leftarrow R \cup \{x\}$
            $\gamma'_{best} \leftarrow \gamma'_{T}(\mathbb{D})$
    $R \leftarrow T$
until $\gamma'_{best} == \gamma'_{prev}$
return $R$
```

III. The Proposed Approach

The proposed grooming detection approach is illustrated in Figure 1, which utilizes the fuzzy-rough feature selection method and the bag of words (BoW) or the term frequencyinverse document frequency (TF-IDF) model for text classification in the field of digital forensics. The general process of the fuzzy-rough feature selection-based text classifier include five phases, which are pre-processing, feature extraction, feature selection, feature normalization, and classification, which are detailed in the rest of this section.

A. Pre-processing

This step is usually required for text classification for grooming detection, as the datasets are often not well-structured and noisy, in order to boost up the speed of conducting the entire process. The dataset is commonly represented in the format of extensible markup language (XML), and a parser is then employed for extracting the only meaningful data by removing all the markups, via traversing each XML file from the root to all leaf nodes. Taking the PAN'13 Author Profiling data set [19] as an example, all the meta-data will be removed except the words included in the text conversions and the conversation IDs.

B. Text Feature Extraction

Text feature extraction is performed after the pre-processing stage, for generating uniformed document representation for each data instance with an unified length, because different data instances (or conversation chat log) in the dataset usually are of different lengths. In this work, two well-acknowledged text feature extraction approaches are employed, including the bag of words (BoW) [20] and term frequencyinverse document frequency (TF-IDF) [21]:

1) Bag of Words: Given a document, the text features extracted by the BoW model is essentially a set of words (terms) without tag, syntax, semantics etc [20].

2) Term FrequencyInverse Document Frequency: TF-IDF is performed in two steps, TF and IDF. Suppose that there are totally n XML documents and each of which is notated as d_i, i.e., $d_i \in \{d_1, d_2, ..., d_n\}$, if the word w appeared in p data instances ($p \leqslant n$), the TF-IDF of a word w can be calculated by:

$$\begin{aligned} TFIDF(w, d_i) &= TF(w, d_i) * IDF(w) \\ &= \frac{|\{d_i | w \in d_i\}\,|}{|\,d_i|} * \log_{\mathrm{e}}^{(\frac{n}{p})} \end{aligned} \tag{6}$$

where $|\cdot|$ represents the size of the set. Note that d_i is a set of important words, or selected features.

C. Fuzzy-Rough Feature Selection

Fuzzy-rough feature selection (FRFS) is a feature selection method for selecting discriminative and important features (or attributes) to describe the entire data set as introduced in Section II-B.. It can be used to cope with uncertainties associated with the data [14]. Essentially, the appearance of vagueness is caused by the lack of clear boundaries of concepts in the data. Particularly, in this work, the similarity function within the fuzzy-rough feature selection method is defined as:

$$\begin{aligned} \mu_{R_w}(d_1, d_2) = \max \Bigg(&\min \Bigg(\frac{\big(w(d_2) - (w(d_1) - \sigma_w)\big)}{\sigma_w}, \\ &\frac{\big((w(d_1) + \sigma_w) - w(d_2)\big)}{\sigma_w} \Bigg), 0 \Bigg), \end{aligned} \tag{7}$$

where $\mu_{R_a}(d_1, d_2)$ denotes the degree to which data instances (i.e., documents) d_1 and d_2 are similar for word w, σ_w is the covariance of word (or feature) w.

D. Text Feature Normalisation

There are some feature normalization strategies available in the literature such as min-max (MM) normalization which linearly transforming the selected text features X to the interval of [0,1], ℓ_1-normalization [22], and ℓ_2-normalization [23]. The recently proposed power normalization (PN) [23] is a data normalization technique which guarantees that the selected features are invariant to the number of extracted or selected words, and its corresponding performance is closely related to the power coefficient α. In this work, all the implementations of ℓ_1, ℓ_2 and PN normalization approaches are investigated in the context of text classification in the experimental section, which include ℓ_1PN, ℓ_2PN, PNℓ_1, and PNℓ_2 [24], [25]. The normalization approaches are summarized as follows:

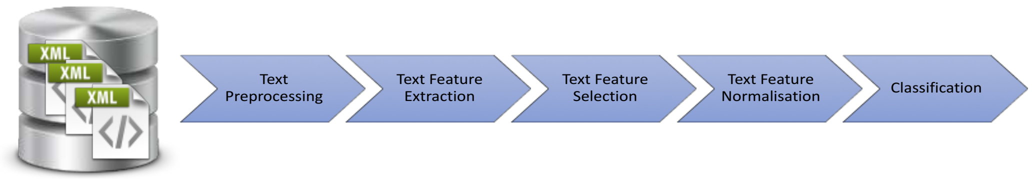


Fig. 1. The general framework of online text classification

$$X^{norm} = \begin{cases} \frac{X - \min(X)}{\max(X) - \min(X)}, & \text{if norm = 'MM';} \\ \frac{X}{||X||_1}, & \text{if norm = '}\ell_1\text{';} \\ \frac{X}{||X||_2}, & \text{if norm = '}\ell_2\text{';} \\ \text{sign}(X)||X||^{\alpha}, & \text{if norm = 'PN'.} \end{cases} \quad (8)$$

where $|| \cdot ||_1$ and $|| \cdot ||_2$ denote the taxicab and Euclidean norm respectively, $\alpha \in [0, 1]$ is the power coefficient in the PN method.

E. *Classification*

Without lose generality, for a given dataset, suppose that the normalized text features are $X = \{\{w_{11}, w_{12}, ..., w_{1n}\}, \cdots, \{w_{m1}, w_{m2}, ..., w_{mn}\}\}$, and $x_i = \{w_{i1}, w_{i2}, ..., w_{in}\}, 1 \leqslant i \leqslant m$ represents the ith data instance. the value of n refers to the number of selected features or words led by the feature selection approach. From this, a classier can be employed to perform the binary classification task, and four widely applied classifiers are used in this work.

1) Gaussian Naïve Bayes: In the Gaussian Naïve Bayes (GNB) classifier [10], the likelihood of text features appearance are assumed to be subjected to Gaussian distribution:

$$\mathcal{P}(x_i|c) = \frac{1}{\sqrt{2\pi\sigma_c^2}} \exp\left(-\frac{(x_i - \mu_c)^2}{2\sigma_c^2}\right), \quad (9)$$

where c denotes the class variable, and x_i, $1 \leqslant i \leqslant n$, represents the text feature vector of the ith data instance. The Gaussian distribution parameters σ_c^2 and μ_c can be learned or estimated based on the maximal likelihood.

2) Random Forest: A random forest is a meta estimator that fits a number of decision tree classifiers on various sub-samples of the data set and use averaging to improve the predictive accuracy and control over-fitting [10]. The sub-sample size is always the same as the original input sample size but the samples are drawn with replacement if bootstrap samples were used.

3) AdaBoost: AdaBoost [10] is a meta-estimator that starts by fitting a classifier on the original data set and then fits additional copies of the classifier on the same data set but where the weights of incorrectly classified instances are adjusted such that subsequent classifiers focus more on difficult cases.

4) Logistic Regression: The logistic regression (LR) [10] is a linear classifier in which the probabilities were employed for describing possible outcomes of a single trial that are modeled using the logistic function.

IV. Experiments

The proposed approach was validated and evaluated as reported in this section, using a reconstructed data set which was derived from the PAN'13 Author Profiling data set.

A. *The Data Set*

The PAN'13 author profiling data set [19] consists of a total number of 262,254 XML files in which, originally, 236,814 and 25,440 XML files were contained in the training and testing folders of the English corpus. This data set was originally proposed for the purpose of age and gender prediction. The details of this data set are listed in the upper part of Table I. A new data set was constructed by selecting all the grooming data items in the PAN'13 author profiling data set in addition to 731 normal coversations[†], to support the experiments in this work. Note that the generated online grooming data set is highly imbalanced and not specifically collected for digital forensics, as indicated by the low event rate for classes 'Pedophile' and 'Sex' in the upper part of Table I.

TABLE I. PAN'13 AUTHOR PROFILING DATA SET AND THE EXTRACTED ONLINE GROOMING DATA SET

Corpus	File sizes (262,254 V.S. 1,000)						
	Normal	*Pedophile*	*Sex*	*Total*	*ERP*	*ERS*	*ERT*
Training	236,626	164	24	236,814	0.07%	0.01%	0.08%
Test	25,359	72	9	25,440	0.28%	0.04%	0.32%
Online grooming dataset	731	236	33	1,000	23.60%	3.30%	26.9%

*ERP: Event Rate for Pedophile; ERS: Event Rate for Sex; ERT: Event Rate for Total.

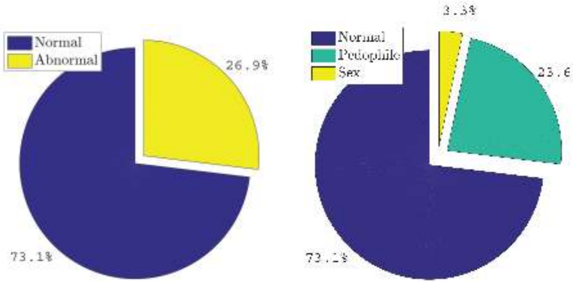
B. *Experimental Setups*

All the experiments were implemented in Python™ 2.7.14 and conducted using a HP® workstation with Intel® Xeon™ E5-1630 v4 CPU @ 3.70 GHz. The performances of four classifiers are evaluated using 10-Fold cross-validation for two classification tasks:

1) Binary classification: The two class labels are 'Normal' and 'Abnormal'.

2) Multi-label classification: The three assigned classes are 'Normal', 'Pedophile', and 'Sex', respectively.

[†] The IDs of the list of the selected XML files in forming the online grooming data set is available at: http://www.lyang.uk/PAN-DigitalForensics



(a) Binary classification (b) Multi-label classification

Fig. 2. Data instances percentage for both learning tasks

C. Result Analysis

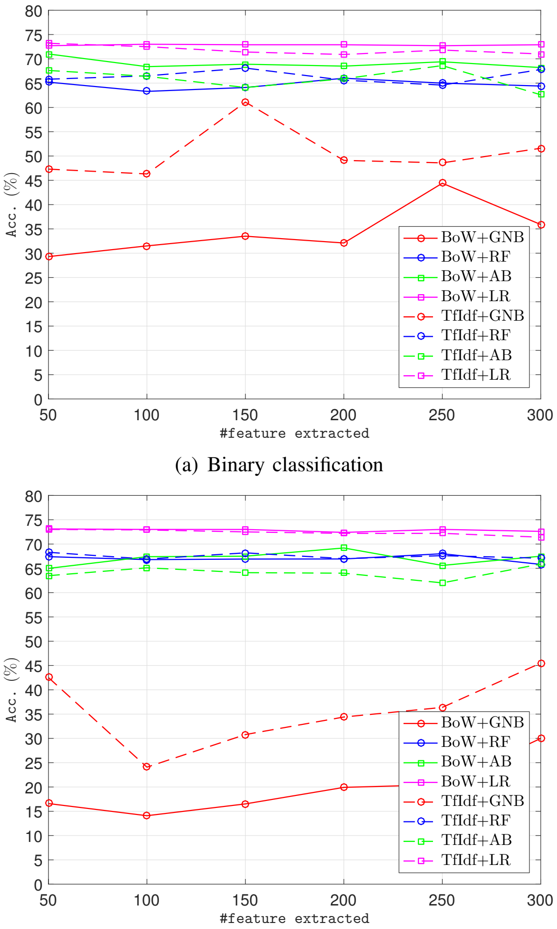
1) The basic experiment: To investigate the performance of the proposed grooming detection approach, this experiment was conducted using 4 classifiers, including Gaussian Naive Bayes (GNB), Random Forest (RF), AdaBoost (AB) and Logistic Regression (LR). The experiment applied the feature extraction approach with BoW model where the number of extracted features are ranging from 50 to 300 with interval of 50 from totally 46,703 features/words in 1,000 XML files, without the use of feature selection, and also used the feature normalization approach with the MM method. The performance, for binary and multi-label classifications, are reported in Table II.

TABLE II. PERFORMANCE THE BASIC EXPERIMENT IN ACCURACY.

Classification Type	*Classifier*	*Extracted feature set size*					
		50	*100*	*150*	*200*	*250*	*300*
Binary	GNB	29.30	31.50	33.50	32.10	**44.40**	35.90
	RF	65.20	63.30	64.10	**66.01**	65.00	64.40
	AB	**71.00**	68.39	68.89	68.50	69.39	68.20
	LR	72.70	**73.00**	72.90	72.90	72.70	72.90
Multi-label	GNB	16.59	14.09	16.52	19.92	20.48	**30.00**
	RF	67.40	66.80	66.91	66.93	**68.01**	65.81
	AB	64.99	67.40	67.50	**69.19**	65.60	67.50
	LR	**73.11**	73.01	73.00	72.40	73.01	72.61

2) Impact of different feature extraction methods: This experiment evaluates the performance of text classification using the BoW and TF-IDF models for extracting text features, without taking the feature selection phase. The performance is shown in Figure 3 for both binary and multi-label classifications.

3) Impact of various feature normalization strategies: This experiment investigates the impact of eight feature normalization techniques, including MM, ℓ_1, ℓ_2, PN, ℓ_1PN, ℓ_2PN, PNℓ_1, and PNℓ_2. The number of features (i.e., words) in both BoW and TF-IDF models start from 50.



(a) Binary classification

(b) Multi-label classification

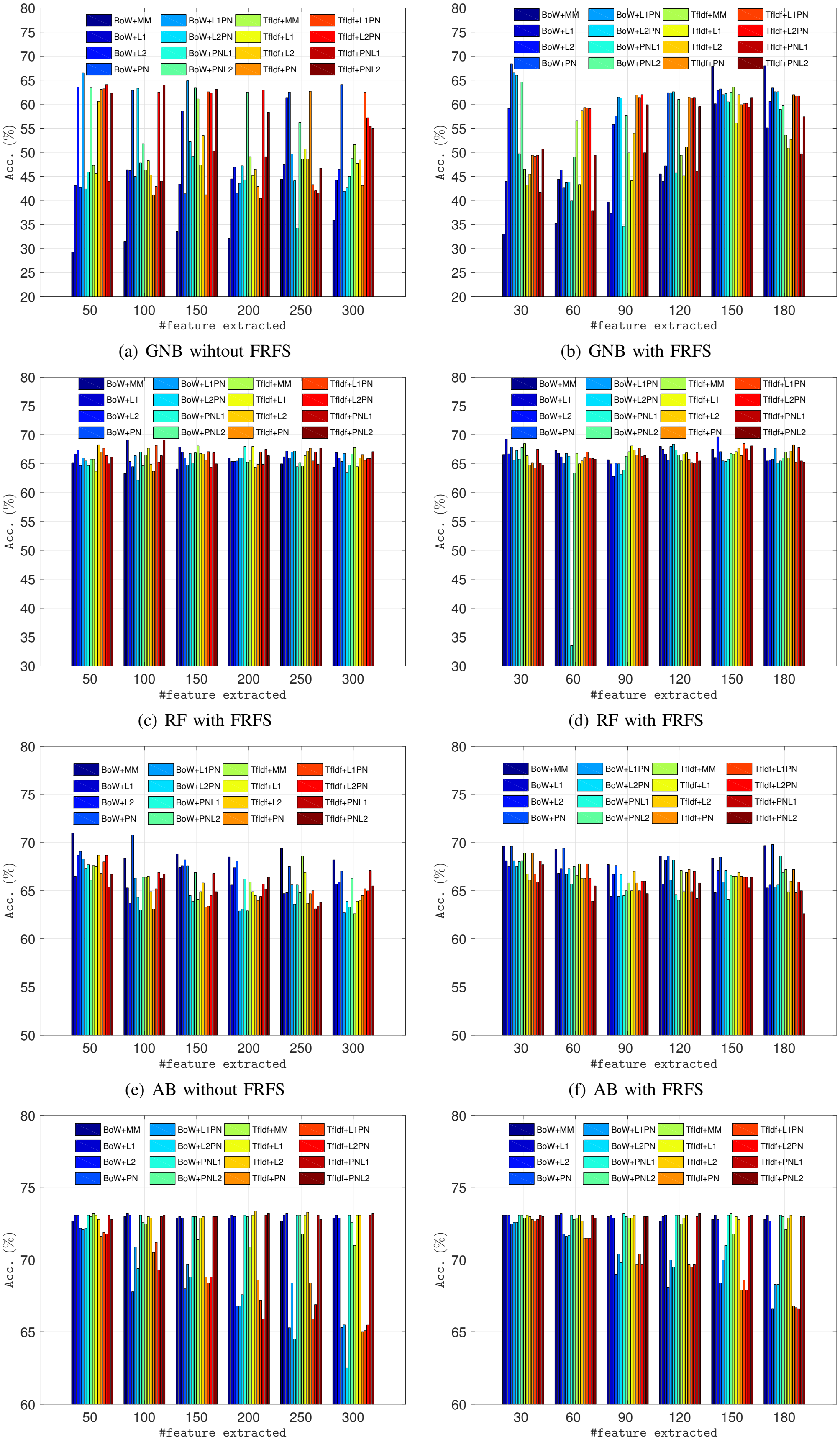
Fig. 3. Performance comparison in accuracy percentage between the BoW and the TF-IDF model with the varying number of features extracted with the same (MM) normalization technique.

Based on the results shown in the left columns of Figures 4 and 5, the best performance are further summarized in the Table III. It is noticeable that, the GNB classifier is suitable for PN-based normalization strategies (i.e., PN, ℓ_1PN, ℓ_2PN, and PNℓ_2) in both binary and multi-label classification tasks using BoW or TFIDF features; the RF classifier works better with the basic normalization (i.e., ℓ_1, ℓ_2, and MM), and the PN-based normalization techniques(i.e., PN, PNℓ_1, PNℓ_2, and ℓ_2PN), when a small number of (from 50 to 150) and a large number of (from 200 to 300) features are used; AB classifier is able to produce better performance in binary and multi-label classification using the basic normalization and the PN based strategies; the LR classifier works better with the basic normalization techniques (i.e., ℓ_1, ℓ_2, and MM) in both classification tasks.

TABLE III. BEST PERFORMANCE OBTAINED WITHOUT EMPLOYING THE FUZZY-ROUGH FEATURE SELECTION. COLOR CODES: BOW+MM, BOW+ℓ_1, BOW+ℓ_2, BOW+PN, BOW+ℓ_1PN, BOW+ℓ_2PN, BOW+PNℓ_1, AND BOW+PNℓ_2. RESULTS LED BY TF-IDF IN *italic* AND BOW-BASED IN NORMAL.

Classification Type	*Classifier*	*#feature extracted*					
		50	*100*	*150*	*200*	*250*	*300*
Binary	GNB	66.50	*63.99*	64.89	*62.99*	*62.69*	64.10
	RF	*68.30*	69.11	*68.10*	68.01	*67.70*	*67.80*
	AB	71.00	70.80	68.89	68.50	69.39	68.20
	LR	*73.20*	73.20	73.00	*73.40*	*73.30*	*73.20*
Multi-label	GNB	56.93	53.10	*41.40*	52.99	52.95	*45.80*
	RF	*68.90*	68.60	68.49	68.80	*69.71*	69.51
	AB	*67.32*	67.40	67.50	69.19	*67.92*	68.21
	LR	73.11	73.11	*73.20*	73.11	73.21	73.11

4) Impact of feature dimensions: In this experiment, the fuzzy-rough feature selection method was employed. Specifically, the selected number of features/attributes was set to



(a) GNB wihtout FRFS

(b) GNB with FRFS

(c) RF with FRFS

(d) RF with FRFS

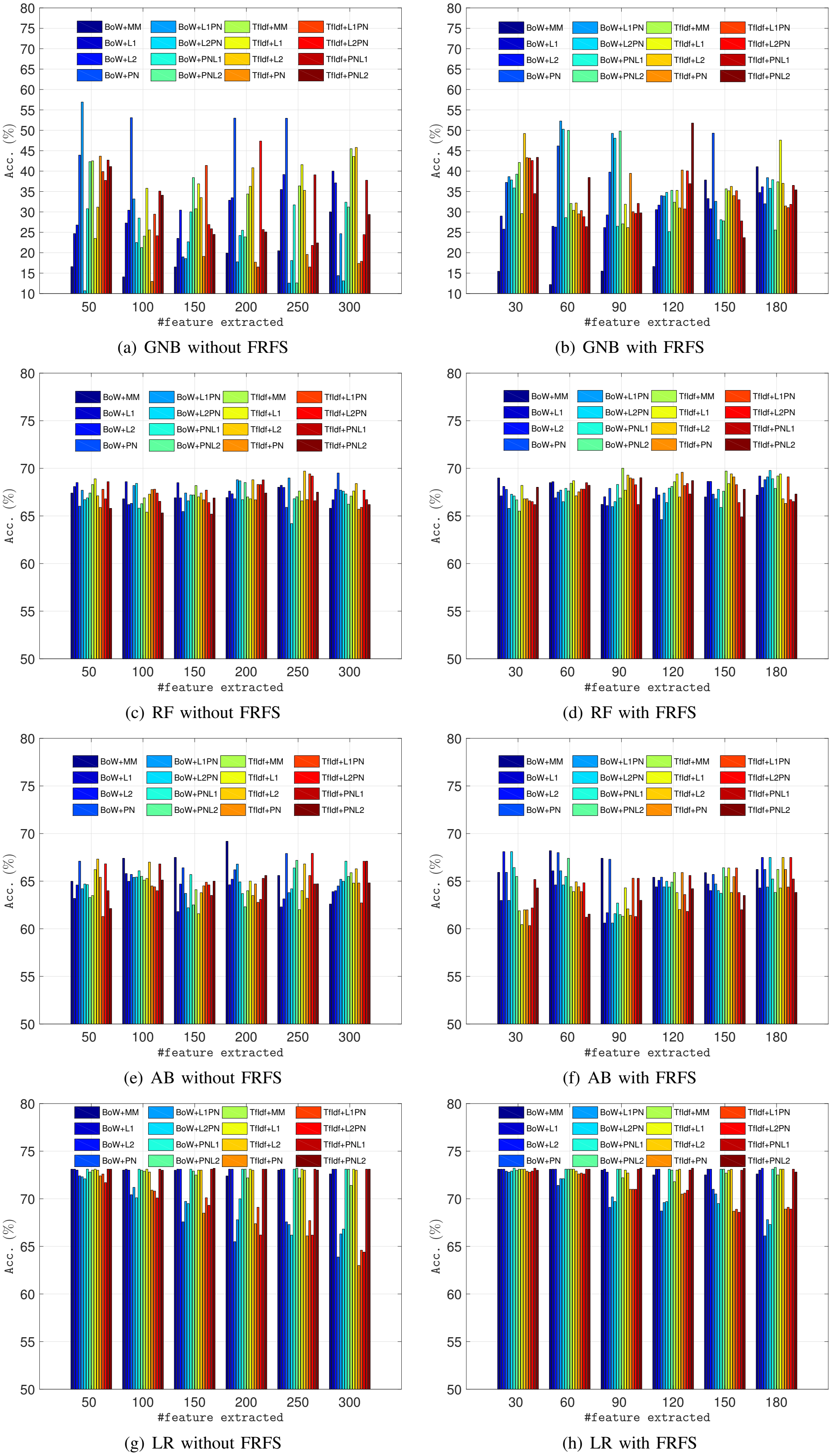
(e) AB without FRFS

(f) AB with FRFS

(g) LR without FRFS

(h) LR with FRFS

Fig. 4. Performance in accuracy with and without fuzzy-rough feature selection applied under different feature normalization strategies and feature set sizes for binary classification.



(a) GNB without FRFS

(b) GNB with FRFS

(c) RF without FRFS

(d) RF with FRFS

(e) AB without FRFS

(f) AB with FRFS

(g) LR without FRFS

(h) LR with FRFS

Fig. 5. Performance in accuracy with and without fuzzy-rough feature selection applied under different feature normalization strategies and feature set sizes multi-label classification.

60% to the originally extracted number of features. That is, accordingly, 30, 60, 90, 120, 150, and 180 features were selected. The results for this experiment are reported in the right columns of Figures 4 and 5.

Generally speaking, the performance achieved using FRFS with 150 features as shown in Table IV) is better than that without the use of the FRFS as shown in Table III. In particular, the accuracies generated by using 30 features resulted from the use of FRFS are better than those when 150 or 300 features were used. Given the high degree of imbalance of the online grooming data set, the overall efficiency and effectiveness of the proposed grooming detection approach is promising, and the performance can be further improved if better data set and more carefully fine-tuned parameters are used.

TABLE IV. BEST PERFORMANCE OBTAINED BY INVOLVING THE FUZZY-ROUGH FEATURE SELECTION. COLOR CODES: BOW+MM, BOW+ℓ_1, BOW+ℓ_2, BOW+PN, BOW+ℓ_1PN, BOW+ℓ_2PN, BOW+PNℓ_1, AND BOW+PNℓ_2. RESULTS LED BY TF-IDF IN *italic* AND BOW-BASED IN NORMAL.

Classification Type	*Classifier*	*#feature extracted*					
		30	*60*	*90*	*120*	*150*	*180*
Binary	GNB	68.42	*59.31*	*62.00*	62.60	67.89	67.99
	RF	69.30	67.30	*68.10*	68.40	69.69	*68.30*
	AB	69.61	69.40	67.71	68.60	68.49	69.80
	LR	73.10	73.20	73.20	*73.21*	73.20	73.10
Multi-label	GNB	*49.22*	52.28	49.82	*51.78*	49.30	*47.60*
	RF	68.99	*68.70*	*70.10*	*69.59*	*69.71*	69.78
	AB	68.10	68.20	67.40	*65.91*	66.41	67.49
	LR	73.21	73.11	*73.21*	*73.21*	*73.20*	73.30

V. Conclusion

A child grooming detection system was proposed in this work by employing the fuzzy-rough feature selection method in addressing the uncertainty coming with the natural language conversations. An extracted data set has been used for system validation and evaluation. The experimental results revealed the power of the proposed approach in support online grooming detection. Although promising, there is room for improvement. Firstly, it is interesting to investigate how the proposed approach works on other forensics tasks. Also, it requires further investigation to test the proposed approach for dealing with online text streaming in real time.

References

[1] A. E. Cano, M. Fernandez, and H. Alani, "Detecting child grooming behaviour patterns on social media," in *Social Informatics*, L. M. Aiello and D. McFarland, Eds. Cham: Springer International Publishing, 2014, pp. 412–427.

[2] C. Foc and G. Kalkan, "Barnardo's online grooming survey 2016," 2016. [Online]. Available: http://www.barnardos.org.uk/publication-view.jsp?pid=PUB-2920

[3] D. Michalopoulos and I. Mavridis, "Utilizing document classification for grooming attack recognition," in *2011 IEEE Symposium on Computers and Communications (ISCC)*, June 2011, pp. 864–869.

[4] D. Lillis, B. A. Becker, T. O'Sullivan, and M. Scanlon, "Current challenges and future research areas for digital forensic investigation," in *Annual ADFSL Conference on Digital Forensics, Security and Law*, 2016.

[5] F. E. Gunawan, L. Ashianti, S. Candra, and B. Soewito, "Detecting online child grooming conversation," in *2016 11th International Conference on Knowledge, Information and Creativity Support Systems (KICSS)*, Nov 2016, pp. 1–6.

[6] I. McGhee, J. Bayzick, A. Kontostathis, L. Edwards, A. McBride, and E. Jakubowski, "Learning to identify internet sexual predation," *International Journal of Electronic Commerce*, vol. 15, no. 3, pp. 103–122, 2011.

[7] S. J. Pandey, I. Klapaftis, and S. Manandhar, "Detecting predatory behaviour from online textual chats," in *Multimedia Communications, Services and Security*, A. Dziech and A. Czyżewski, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 270–281.

[8] H. Pranoto, F. E. Gunawan, and B. Soewito, "Logistic models for classifying online grooming conversation," *Procedia Computer Science*, vol. 59, pp. 357 – 365, 2015, international Conference on Computer Science and Computational Intelligence (ICCSCI 2015). [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1877050915020657

[9] G. Salton, "Developments in automatic text retrieval," *science*, vol. 253, no. 5023, pp. 974–980, 1991.

[10] R. S. Michalski, J. G. Carbonell, and T. M. Mitchell, *Machine learning: An artificial intelligence approach*. Springer Science & Business Media, 2013.

[11] L. Yang and Q. Shen, "Closed form fuzzy interpolation," *Fuzzy Sets and Systems*, vol. 225, pp. 1–22, 2013, theme: Fuzzy Systems.

[12] J. Li, L. Yang, Y. Qu, and G. Sexton, "An extended takagi–sugeno–kang inference system (tsk+) with fuzzy interpolation and its rule base generation," *Soft Computing*, vol. 22, no. 10, pp. 3155–3170, May 2018.

[13] L. Yang, F. Chao, and Q. Shen, "Generalized adaptive fuzzy rule interpolation," *IEEE Transactions on Fuzzy Systems*, vol. 25, no. 4, pp. 839–853, Aug 2017.

[14] Y. Qu, Y. Rong, A. Deng, and L. Yang, "Associated multi-label fuzzy-rough feature selection," in *2017 Joint 17th World Congress of International Fuzzy Systems Association and 9th International Conference on Soft Computing and Intelligent Systems (IFSA-SCIS)*, June 2017, pp. 1–6.

[15] N. Mac Parthaláin and R. Jensen, "Fuzzy-rough feature selection using flock of starlings optimisation," in *Fuzzy Systems (FUZZ-IEEE), 2015 IEEE International Conference on*. IEEE, 2015, pp. 1–8.

[16] R. Jensen and Q. Shen, "New approaches to fuzzy-rough feature selection," *IEEE Transactions on Fuzzy Systems*, vol. 17, no. 4, pp. 824–838, Aug 2009.

[17] N. Mac Parthaláin, R. Jensen, Q. Shen, and R. Zwiggelaar, "Fuzzy-rough approaches for mammographic risk analysis," *Intell. Data Anal.*, vol. 14, no. 2, pp. 225–244, Apr. 2010. [Online]. Available: http://dl.acm.org/citation.cfm?id=1804307.1804313

[18] Q. Guo, Y. Qu, A. Deng, and L. Yang, "A new fuzzy-rough feature selection algorithm for mammographic risk analysis," in *2016 12th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD)*, Aug 2016, pp. 934–939.

[19] F. Rangel, P. Rosso, M. Koppel, E. Stamatatos, and G. Inches, "Overview of the Author Profiling Task at PAN 2013," in *CLEF 2013 Evaluation Labs and Workshop – Working Notes Papers, 23-26 September, Valencia, Spain*, P. Forner, R. Navigli, and D. Tufis, Eds., Sep. 2013.

[20] D. Jurafsky and J. H. Martin, *Speech and language processing*. Pearson London:, 2014, vol. 3.

[21] H. C. Wu, R. W. P. Luk, K. F. Wong, and K. L. Kwok, "Interpreting tf-idf term weights as making relevance decisions," *ACM Transactions on Information Systems (TOIS)*, vol. 26, no. 3, p. 13, 2008.

[22] J. Yang, K. Yu, Y. Gong, and T. Huang, "Linear spatial pyramid matching using sparse coding for image classification," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009, pp. 1794–1801.

[23] F. Perronnin, J. Sánchez, and T. Mensink, "Improving the fisher kernel for large-scale image classification," in *European conference on computer vision*. Springer, 2010, pp. 143–156.

[24] R. Cameron, Z. Zuo, G. Sexton, and L. Yang, "A fall detection/recognition system and an empirical study of gradient-based feature extraction approaches," in *UK Workshop on Computational Intelligence*. Springer, 2017, pp. 276–289.

[25] Z. Zuo, L. Yang, Y. Peng, F. Chao, and Y. Qu, "Gaze-informed egocentric action recognition for memory aid systems," *IEEE Access*, vol. 6, pp. 12 894–12 904, 2018.