Digital Forensics supported by Machine Learning for the detection of Online Sexual Predatory Chats

CH Ngejane[a,b,*], JHP Eloff[a], TJ Sefara[b], VN Marivate[c]

[a] *Cyber-security & Big Data Science Research Group*
Department of Computer Science, University of Pretoria, South Africa
[b] *Council for Scientific and Industrial Research, Pretoria, South Africa*
[c] *Data Science for Social Impact Research Group*
Department of Computer Science, University of Pretoria, South Africa

Abstract

Chat-logs are informative digital footprints available on Social Media Platforms (SMPs). With the rise of cybercrimes targeting children, chat-logs can be used to discover and flag harmful behaviour for the attention of law enforcement units. This can make an important contribution to the safety of minors on SMPs from being exploited by online predators. The problem is that digital forensic investigation is mostly manual. Thus, a daunting task for forensic investigators because of the sheer volume and variety of data. The solution that is proposed in this paper employs a Digital Forensic Process Model that is supported by Machine Learning (ML) methods to facilitate the automatic discovery of harmful conversations in chat-logs. ML has already been successfully applied in the domain of text analysis for the discovery of online sexual predatory chats. However, there is an absence of approaches that show how ML can contribute to a digital forensic investigation. Thus, the contribution of this paper is to indicate how the tasks in a digital forensic investigation process can be organised so to obtain usable ML results when investigating online predators.

Keywords:
digital forensic investigation, cybersecurity, machine learning, cyber safety, online sexual predatory conversation

*Corresponding author
Email address: hombakazi.ngejane@gmail.com (CH Ngejane)

1. Introduction

Sexual predators who target children in cyberspace are a global problem these days. In March 2019, BBC News reported that the British police received 1600 crimes related to online child predatory within 11 months [1]. Examining online sexual predatory conversations is a daunting task for digital forensic investigators, especially so when taking into consideration the volume, variety and rate of these types of incidents. Online sexual grooming is a crime perpetrated by pedophiles who use social media platforms to target children online and result in a digital forensic investigation.

Mainly, digital forensic investigation processes involve- identifying, collecting and acquiring evidence to examine, analyse and present conclusive findings to relevant stakeholders [2]. Also, these processes aim to build a portfolio of evidence for the court of law [3]. Nowadays, the volume and the variety of data presented for digital forensic investigations has increased significantly. Internet and social media platforms (SMPs) are the leading cause of this increase. Thus, rendering the manual investigation done by forensic experts a daunting task. Therefore has resulted in an increased burden of digital forensic investigation backlogs [4].

Intelligent technologies, such as Machine Learning (ML), have the potential to support the digital forensic investigation process. These technologies can automate the said manual digital forensic investigation processes when analysing significant volumes and a large variety of data such as to be found in chat-logs. They can also fast-tract action and assist law enforcement units to investigate and deal with cyber-incidents proactively. As a result, the evidence can be used in the court of law against the predators, thus minimising the spread of online sexual grooming.

Recently, ML models have been used in digital forensics to address social cyber-related threats such as intrusion detection and digital text forensics.

In the case of online sexual predatory, ML can be used to classify text as

either predatory or non-predatory. The ML-based online sexual predatory behaviour identification systems for digital forensics have to take into consideration the sensitive nature of their outcomes. For an ML model to be successful, it requires predatory data, which is scarce, and, contains highly imbalanced and unstructured data [5, 6]. These challenges can undermine the accuracy of ML models resulting in the inability to generalise in a real-world setting. Digital forensic investigators need to understand how and why the classification takes place [7]. Also, understand when does the model fail in order to ensure transparency and fairness.

The work at hand proposes the use of ML in digital forensic investigation. The objective is threefold; 1. To improve efficiency in terms of time and reducing human effort by automating digital forensic investigation manual tasks; 2.To explain to digital forensic investigators how the model discovers an online sexual predatory conversation through ML interpretability; and 3. Ensure ML transparency so that digital forensic experts can understand the possible limitations of the model. Thus, developing a trust relationship between forensic experts and ML tools so that forensic experts can make informed decisions about the admissibility of the evidence. This approach is guided by a Digital Forensic Process Model (DFPM) developed by the same research group of the paper in hand, see [3].

The main contributions of this work are as follows:

- To integrate the use of ML in the DFPM by overlaying the ML processes in the Digital Forensic Investigation (DFI) tasks
- To classify online sexual predatory conversations from non-sexual predatory conversations as accurately as possible using ML conventional and modern methods
- To apply ML interpretability to decompose the model's decision making using sexual predatory conversations important features including words and phrases

- To probe the ML model's performance by using the what-if tool to discover the model's limitations

The rest of the paper is as follows: Section 2 provides background information on online predatory conversations (commonly known as online sexual grooming), an overview of related works and the proposed DFPM background. Section 3 outlines the integrated ML in the context DFI tasks; followed by the analysis of results in Section 4. Section 5 concludes our work and outlines directions for future work.

2. Background

The increased use of SMPs has brought about several challenges regarding cyber safety [8, 9]. Cyberbullying [10] and deception (e.g. online sexual predatory) are just two of the issues that arise with children [11]. As such, it is essential to identify specific behavioural patterns in an online conversation that best describe a sexual predatory conversation.

A term used to describe online child sexual exploitation is "Online grooming." [12]. Harms [13] defines this term as a:

> *communication process by which a perpetrator applies affinity seeking strategies, while simultaneously engaging in sexual desensitisation and information acquisition about targeted victims in order to develop relationships that result in need fulfilment (e.g. sexual molestation).*

The investigation of online sexual predatory conversations as a prevalent task for digital forensic is common [14]. As part of the grooming process, studies indicate that predators build a trust relationship with their victims to increase the victim's dependence on them. Therefore it is crucial to understand the predator's grooming behaviour for forensic investigations. This depends on manual analysis of chat logs, which usually requires a considerable amount of time. Thus, posing a need for automated and intelligent ways of discovering online

sexual predatory behaviour. Various scholars have proposed a number of ML models as detailed in [15] and [16].

2.1. Related Work

The related work for the research at hand mainly focuses on the following two fields:

1. Machine learning models for the discovery of online sexual predatory.
2. Digital Forensic Process Model.

2.1.1. Machine learning models for the discovery of online sexual predatory

The subsequent studies focused on discovering behavioural patterns in an online conversation based on the theory of luring communication (TLC) presented by Olson et al. [17], O'Connell [18], Whittle et al. [19]. The TLC framework states that online sexual exploitation of minors occurs in various grooming stages. Therefore, these stages can be used to trace whether a conversation is predatory or not. Olson et al. [17], also noted that online sexual grooming behavioural patterns display similar characteristics as those in physical interactions. Therefore, they can be useful and relevant digital evidence to investigate further or to prosecute the perpetrator.

Accordingly, Miah et al. [20] proposed a method to categorise online conversations into three sexual grooming stages: (i) Child exploitation: Elements of sexual exploitation of a child in a conversation. (ii) Sexual fantasies: Consensual conversations between peers with a high degree of sexual content. (iii) General chatting: General conversations without sexual content.

They applied text classification techniques such as Naive Bayes, Classification via Regression (CvR) and J48-Decision Trees. To train these models, they extracted term-based psychometric features using a tool called Linguistic Inquiry and Word Count (LIWC) [21].

Similarly, Kontostathis et al. [22] and McGhee et al. [23] developed a tool called ChatCoder to mimic human annotators in labelling predatory conversations according to their associated grooming stages. To achieve this, they

constructed a dictionary of unique terms and phrases, used to label each line in a conversation. They used phrase matching and rule-based algorithms to classify a conversation as either associated (or not) to grooming stages.

Leveraging on the annotated data from McGhee et al. [23], Cano et al. [24] aimed to enhance accuracy when classifying online grooming stages by using a Support Vector Machine (SVM) classifier. They trained the classifier using various features, namely bag-of-words (BOW), term frequency-inverse document frequency (TFIDF), syntactical, sentiment polarity, content, LIWC [21], and discourse patterns.

Black et al. [25] proposed a linguistic and content analysis approach to examine online predatory conversations utilising LIWC. They aimed to determine the congruence between offline versus online predator behaviour patterns by using five sexual grooming stages as proposed by O'Connell [18]. They randomly sampled 44 conversations from the Perverted Justice dataset and segmented predator lines into five equal parts using word count in order to model the grooming stages.

Meyer [26] proposed to detect an adult pretending to be a child in an online chat by employing a two-fold approach. They first classified adults and children in chats, and later examined each child to identify real children from fake children using text analysis. The author's results showed that it is possible to distinguish genuine children from adults pretending to be children. Therefore, they concluded that their model could be used to alert children about the actual age of the person in the online conversation.

Recently, Liu et al. [27] and Ebrahimi et al. [6] proposed to address this problem through the use of deep learning approaches. Unlike the traditional ML methods previously proposed, where BOW and TFIDF have been used to train the models, the authors proposed to use sentence and word embeddings as features to compress feature dimension. However, the authors did not consider a classification of grooming stages, which is vital in understanding the behaviour of a sexual predator [18].

Amato et al. [28] proposed to identify online offensive behaviour using a

two-step detection approach. They first used Markov chains to discover normal behaviour from a sampled dataset. They used an activity detection framework to identify unexplained behaviour based on healthy behaviour. They validated their work by conducting a series of experiments using a dataset extracted from Facebook.

Souri et al. [29] employed a five-factor model approach to identify personality profiles using ML methods. They aimed to discover online behavioural patterns that can be traced by investigating user's online interactions.

On the other hand, Kloess et al. [30], proposed a qualitative approach of thematic analysis to discover incriminating processes of online sexual predatory. Using a case study, they found that predators used indirect or direct grooming processes to chat or initiate contact with their victims. Whereas, Lorenzo-Dus and Kinzel [31] proposed the corpus assisted discourse (CAD) approach and an LIWC tool to understand a predator's grooming language. Their findings show that the LIWC tool lacked transparency compared to CAD, which could reveal sophisticated features associated with grooming stages.

Lastly, Anderson et al. [14] proposed a digital forensic investigation framework to automate the analysis of online grooming detection using the ML system. They employed BoW to extract words that can be associated with online predatory patterns and used fuzzy-rough to select important words as input for the twin fuzzy support vector machine classifier.

Based on the works reported above, it is evident that the use of ML effective in identifying online sexual grooming activities. Online sexual identification has become increasingly important as more and more chat systems are now moving off large chat rooms onto platforms such as WhatsApp, Telegram, to name a few. For the organisations that run these systems, it is crucial to be able to assist in identifying the misuse of their platforms. However, prior works focused mainly on optimising the classification performance of the ML models and on different feature extraction processes. Experiments were conducted with various ML models to improve performance in terms of prediction accuracy.

The limitation in respect of previous studies is that none considered inter-

preting the decision making of their proposed ML models. ML interpretability, as defined by Molnar [32], is the degree to which humans can comprehend why the model has made a particular prediction decision. For the forensic investigation task, prediction accuracy is not enough. In addition to prediction, the model should explain how it came to its prediction [32]. This is essential for forensic investigators as they are required to interpret and present their findings in the court of law or relevant stakeholders [33, 34]. Interpreting the model may also help forensic expert to understand when the model might be limited. In this regard, we have used a what-if tool, which allows humans to analyse, probe and visualise model's performance quickly [35].

2.1.2. Digital Forensic Process Model

The DFPM, as described by Kohn et al. [3], is a set of phases that are used to aid forensic investigation by explaining how the evidence should be revealed. Within this process model, the Digital Forensic Investigation (DFI) phase, depicted in Fig. 1, is of particular interest. DFI is used to determine and analyse extracted digital evidence to identify relevant and admissible information for judicial review [3, 36].

For this study, DFPM was chosen over the ISO/IEC 27043:2015. ISO/IEC 27043:2015 furnishes guidelines based on idealised models for common incident investigation processes for various cases involved in digital forensics [37]. Unlike the DFPM, the ISO/IEC 27043:2015 is not a detailed, low-level guide, but rather a guide that provides an overview of all incident investigation principles and processes without prescribing particular details within each of the investigation tasks [38].

In the case of online predatory conversations, DFI can be used to guide the tasks of extracting incriminating behaviour in a conversation. However, the limitation with DFI, is that it is relatively manual, time-consuming and dependent on human expertise. Thus the need to automate these tasks using ML models. As such, the following section describes the selected DFI tasks, wherein, each task is followed by how the ML process can fit in to complement

the task.

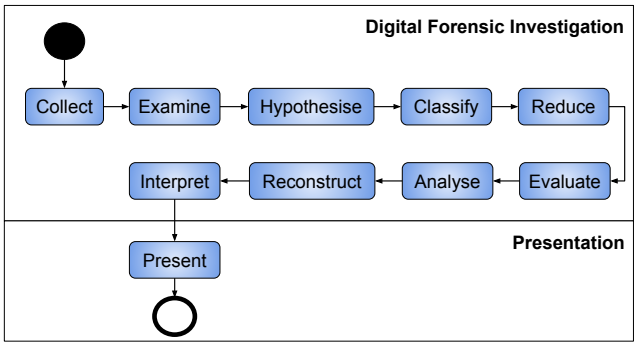


Figure 1: The DFI phase of an Integrated Digital Forensic Process Model [3]

The tasks of the DFI, see Fig. 1 are as follows:

Collect: In the collection task, the investigator collects evidence for examination and investigation. The investigator works on a copied version of the evidence to ensure that the original evidence is changed [3]. As such, in this work data is sourced from relevant social media platforms and stored in a local database for processing, extraction of features and analysis.

Examine: The examine task ensures the formatting of data into a visible or human-readable form for processing [3]. In the current work, HTML character codes from the dataset are converted to the standard ASCII letters.

Hypothesise: The forensic investigator formulates a hypothesis based on inferred assumptions to determine the root cause of the incident [3]. In this paper, it is hypothesised that an online conversation can either be predatory or non-predatory in nature. Therefore, an ML model can be used as an assisting tool to classify these conversations accurately.

Classify: In this task, data is categorised into groups of similar features [3]. For the current work, each conversation is classified as either predatory or non-predatory using ML models.

Reduce: Is a task similar to feature extraction in ML, which is a type of dimensionality reduction that efficiently represents essential parts of the data. In this paper, the strategy used for feature extraction only selects features that

are useful for facilitating online sexual predator discovery [39].

Evaluate: The investigator's findings are evaluated to test whether the formulated hypothesis holds true or not [3]. In the current work, a model is validated on unseen data to test its performance and accuracy by using known ML metrics.

Analyse: The organised data is analysed and examined against the hypothesis to obtain potential evidence [3]. In this paper, the ML models are analysed and tested against the hypothesis. Furthermore, the performance of these models is analysed using ML statistical measures.

Reconstruct: The inferred sequence of events from the investigation is used to reconstruct how and why the incident occurred [3]. In this paper, several input conversations are analysed by assigning a prediction probability to each line of a conversation so to trace the classification decision.

Interpret: If the evaluation was a success, then interpreting the evidence to produce meaningful and contextually legal statements [3] is the next step. In this paper, the reconstructed ML results are conveyed in the context of digital forensic investigations.

The DFPM, particularly the DFI phase, is used to guide the methodology reported in this paper. The DFI tasks were taken into consideration to integrate the use of ML in the digital forensic investigation process. Therefore, the DFI tasks are used as a basis to structure the remainder of the subsequent sections. Starting in Section 3, the dataset used is described, followed by the pre-processing methods for data cleaning and experiments. Results and analysis are presented in Section 4.

3. The Integration of ML models with DFPM

This section is structured according to the DFI phase of the DFPM Fig. 1. It starts by discussing the data acquisition from the PAN, which stands for Plagiarism analysis, Author identification and Near-Duplicate detection [15]. PAN is a series of related scientific and shared tasks for a digital text forensic

investigation.

3.1. Collect

In 2012, PAN initiated the sexual predator identification task, where, they collected online conversations data with the following properties: a small percentage of true positives (i.e. sexual predatory chats and non-consensual); a large percentage of false positives (consensual conversations, with similar topics as "sexual predatory chats") and a large proportion of false negatives (general conversations between people talking about any other topic); as a result, the data is highly imbalanced. They aimed to supply researchers with a standard structure to assess techniques for identification of online sexual predatory.

For false positive conversations, Inches and Crestani [15] used *omegle* repository[1]. This repository consists of anonymous conversations between two peers who happen to be online at the same time, and who engage in various topics including aggressive, consensual sexual conversations or normal conversations. Moreover, for true negative conversations, the team collected data from the Internet Relay Chat (IRC) service[2], where the nature of the dataset included interactions between two or more users per conversation.

The predatory conversations, in other words, the true positives in PAN-12 data, were collected by a non-profit organisation called Perverted Justice (PJ)[3]. PJ actively collects evidence against sexual predators using social media platforms (e.g. online conversations) for the attention of law enforcement officers. This organisation employs trained volunteers who pose as children (approximately 12 years old) and engage in online chat-rooms with potential sexual predators. These decoys play along with their predators online, until their predators request for a physical meeting. PJ then works hand in hand with police officers to arrest the predators, should they arrive for the meeting.

[1] `http://omegle.iportb.com/`

[2] `http://irc.netsplit.de/`

[3] `http://www.perverted-justice.com/`

Notice that, in order to circumvent criminal entrapment, PJ's decoys wait for their predator to initiate physical contact with them and not the other way round. Entrapment is defined by Subramanien and Whitear-Nel [40], as the act of coercing a criminal act in an innocent person's mind to convict that person of the committed crime later. Therefore, the conviction of online predators who initiate to meet with their pseudo-victims may be permissible in court in this case.

The PAN-12 data comprises of both training and testing corpora in XML format. A training corpus contains about 66 000 documents where each document is a conversation. Each conversation is labelled by a unique identifier containing a series of messages. Each message is labelled by a unique line number in the conversation as produced by an author. Lastly, a conversation line of text is aligned with the author and timestamped. Fig. 2 represents an XML structure of a conversation.

```
<conversations>
    <conversation id=1>
        <message id=1>
            <author id=1>AuthorName</author>
            <text>MessageText</text>
        </message>
        <message id=2>
            <author id=2>AuthorName</author>
            <text>MessageText</text>
        </message>
        ...
    </conversation>
    <conversation id=2>
        ...
    </conversation>
    ...
</conversations>
```

Figure 2: XML structure of a PAN-12 conversation [41]

As part of data exploration, when training XGBoost, boosted trees are built, extracting importance scores for each word or phrase is relatively simple. In gen-

eral, importance provides a score that shows how important each word or phrase was during the creation of the boosted trees within the model [42]. Table 1 shows an excerpt of a chat transcript from PAN-12 with highlighted essential features. The dataset contains more than ten thousand conversations with ten words or more in a message, as shown in Fig. 3.

Table 1: Example of a chat based on the first 200 important features

Author	**Text**
1	:-*
2	Hey gf am going to Smiss u
1	im gonna miss u 2
1	i might b on some 2morrow
1	cause we gotta wait for gpa 2 finish his work
2	What time u going to be on?
1	donno yet
2	Ok loves u
2	Take lots of pics
1	i dont got a camera
1	but gmas got one
2	See u can take some silly
1	ok
2	:-*
1	:-*
1	brb
2	Cause ur going be my wife someday :-x
1	back
1	:-*

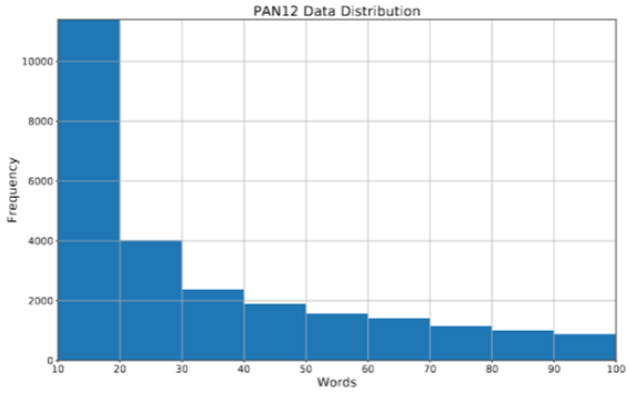


Figure 3: PAN12 data distribution showing word frequency for each conversation or document that has at least 10 words

3.2. Examine

The dataset contains several short text abbreviations, emoticons, slang, digits, symbols, character repetitions, HTML character codes, URLs, and conversations of different lengths, all of which increase the perplexity of the models. However, some words carry significant information, as shown in Table 2. Hence, we converted HTML character codes to the standard ASCII letters. Moreover, we applied lemmatisation to convert words that have different inflectional endings to their stem.

Table 2: Sample of words from the dataset

Symbol	**Meaning**
asl	age & sex & location
brb	be right back
u	you
:-*	kiss
gf	girl friend
m/f	male or female
n	and

3.3. Hypothesis

In our pursuit to understand the behaviour of sexual predators in online chat-logs, we ran several experiments with the PAN-12 dataset to extract features that better explain how machine learning models would identify sexual predatory conversations. This enabled us to provide insightful information that can be useful for forensic investigations. To that end, we trained the selected models using ML supervised techniques. The hypothesis is that there are important features such as phrases or words that are connected to an online predatory conversation.

3.4. Classify

We utilised four types of models for classification, namely: Logistic Regression (LR), XGBoost, Multilayer Perceptrons (MLP), and Long Short-Term Memory (LSTM).

LR. is a mathematical modelling algorithm for predicting binary or multi-class classification problems [43]. We implemented LR by setting multi-class to 'multinomial'. Such that, the minimised loss is the multinomial loss fit across the whole probability distribution. To solve the optimisation problem, we used the limited memory Broyden–Fletcher–Goldfarb–Shanno (L-BFGS) algorithm [44].

XGBoost. short for eXtreme gradient boosting, is an efficient implementation of the gradient boosted trees [45, 42]. One of the major advantages of using XGBoost is its scalability to solve real-world problems using a lesser amount of resources [42].

MLP. is an artificial neural network framework and a non-parametric estimator used for both classification and regression [46]. It is a class of a feed-forward artificial neural network. We implemented MLP using three fully connected layers activated by *rectified linear unit* and *sigmoid* on the output. We used dropout layers between fully connected layers with a probability of 0.5.

LSTM. is an artificial recurrent neural network architecture (RNN) built to deal with long time-dependencies [47]. The advantage of using LSTM models is their ability to overcome vanishing gradient problems mostly encountered when training traditional RNN [48]. We implemented a bidirectional LSTM (BiLSTM) which uses two independent LSTMs for end-to-end sequence processing. It is coupled with an embedding input layer and two fully connected layers activated by a *rectified linear unit* and *sigmoid* on the output. We use dropout layers between fully connected layers with a probability of 0.5.

3.5. Reduce

Reduce is a task similar to feature extraction in ML, which is a type of dimensionality reduction that efficiently represents essential parts of the data.

Deep learning models like LSTM can be used to extract features automatically. We used TFIDF to extract maximum features of 100k to train LR, XGBoost and MLPs. TFIDF is a statistical measure used to assess how important a word is to a document in a corpus. In TFIDF, we extracted unigrams, bigrams, and trigrams to be used as features. The next step is to do feature selection strategy that selects features that are most useful for the current problem. Feature selection is an ML technique of selecting a subset of appropriate features for use in model building [39]. Feature selection can be used to identify and eliminate unnecessary, noisy, less informative. It is also used to identify redundant features from data that do not contribute to the accuracy of a model or may reduce the accuracy of the model. For feature selection strategy to get an optimal solution, we used unigrams to select individual words as features to help in the identification of words in a conversation are connected with sexual predatory behaviour.

Moreover, to identify phrases connected with sexual predatory behaviour, we used a combination of bigrams and trigrams to select a range of two or three words in a conversation enabling the selection of phrases as features. For BiLSTM, we encoded data up to a maximum threshold of 150 words in a conversation. The conversations that were shorter than 150 words; were padded with

zeros. Conversely, conversations that have more than 150 words were truncated.

As shown in Table 3, LR and XGBoost generated essential features based on the provided n-grams. Some of the features do not make sense; hence, the model provides ranked corresponding weights for each feature. We observed that features with low weights do not make sense, but features with high weights are reasonable features.

It is crucial to note that the list of features as presented in Table 3 contain stop words and noisy features which can also signify a valuable meaning in a conversation. Several scholars, including [49], recommend minimal or no data cleaning in order to preserve the meaning or the tone of the author in the conversation. For example, to preserve author's tone, Villatoro-Tello et al. [49] indicates that *in the grooming phase, the predator may amend the relationship by an emphasised "soooooorrrrryyyy" when the minor felt threatened by any obtrusive language.* Thus, we also did not consider data cleaning in this work.

Table 3: Top 12 important features

Unigram		**Bigram & Trigram**	
LR	**XGBoost**	**LR**	**XGBoost**
asl	u	hey asl	hi asl
http	asl	u from	u there
m	http	nice to meet	i hope
f	hi	a girl	a girl
the	m	hi m/f	can call me
's	f	how old	love you
male	ok	18 m	hey sexy
from	you	your name	how old
haha	the	horny girl	sorry i
name	back	what is	sweet dream
a	miss	looking for a	my mom
horny	hope	f*ck you	miss me

3.6. Evaluate

This section discusses the performance metrics used to assess the quality of the models.

Accuracy. is the total number of correctly predicted examples. It is calculated as follows:

$$Accuracy = \frac{TrueP + TrueN}{TrueP + TrueN + FalseP + FalseN} \tag{1}$$

Since the PAN-12 dataset is imbalanced, accuracy is not enough to evaluate the quality of the models. We consequently used the following metrics that do not contain false positive because we are concerned with true positives.

Precision. calculates the total number of positively predicted examples that are relevant. The formulation is calculated as follows:

$$Precision = \frac{TrueP}{TrueP + FalseP} \tag{2}$$

Sensitivity or recall. measures how good a model is at predicting the positives. It is also called true positive rate. The formulation is calculated as follows:

$$Sensitivity = \frac{TrueP}{TrueP + FalseN} \tag{3}$$

$F_1 score.$ is the harmonic mean of the recall and precision. We use F_1score since there is a high variance between precision and sensitivity for skewed datasets. The formulation is calculated as:

$$F_1 score = 2 \times \frac{precision \times sensitivity}{precision + sensitivity} \tag{4}$$

The Confusion matrix. shown in Table 4 evaluates the quality of the classifier. The diagonal elements (TrueN & TrueP) are the number of examples where the predicted label is the same as the true label, while off-diagonal elements represent examples that are mislabelled by the classifier.
where:

Table 4: Structure of a binary confusion matrix

	Predicted label	
True label	**Non-predator**	**Predator**
Non-predator	TrueN	FalseP
Predator	FalseN	TrueP

- TrueP (True positive) is the number of predator examples that are predicted as predatory.
- TrueN (True negative) is the number of non-predator examples that are predicted as non-predatory.
- FalseP (False positive) is the number of non-predator examples that are predicted as predatory.
- FalseN (False negative) is the number of predator examples that are predicted as non-predatory.

4. Results

The results section comprises of the three remaining DFI tasks, namely: analyse, reconstruct and interpret.

4.1. Analyse

We divided the PAN-12 dataset into 67% for training and 33% for testing. Table 5 illustrates the experimental results of the models. We used ten epochs to train MLP and BiLSTM on a batch size of 512 and observed good performance on the accuracy of above 98% for all the models when using both TFIDF and embeddings as input. Since the data is skewed, we further tested the models for precision, sensitivity and F_1score. LR, MLP, and BiLSTM obtained an F_1score of approximately 70%, while XGBoost obtained the lowest F_1score of 57%.

Confusion metrics tables, as shown from Table 6-9, indicate that there were a total number of 16223 non-predatory conversations, of which, 0.2% (26) were predicted as sexual predatory conversations by XGBoost, and 0.2% (25) were predicted as sexual predatory conversations by LR. Whereas 0.6% (101) were predicted as sexual predatory conversations by MLP, and 1.4% (220) were predicted as sexual predatory conversations by BiLSTM. We later investigate why and when the models made incorrect predictions (see Section 4.3.1).

Since the data is imbalanced, we use $F_1 score$ to compare performance with other studies. In Table 10, we show some of the work previously done on online sexual predatory conversations using PAN-12 dataset. Vilariño et al. [50] used PAN-12 to identify online sexual predatory chats using multinomial naive Bayes to obtain an $F_1 score$ of 0.0354, which is outperformed with an $F_1 score$ of 0.0498 when using a dictionary of sexual terms. Kontostathis et al. [51] used decision trees to identify online sexual predatory chats with an $F_1 score$ of 0.47. Vartapetiance and Gillam [52] obtained an $F_1 score$ of 0.48 by manually constructing rules. Better performance is obtained by Kang et al. [53] with an $F_1 score$ of 0.7 using K-nearest Neighbours (KNN). On the other hand, we obtained better $F_1 score$, which is above 0.7 on LR, MLP, and BiLSTM. KNN used by Kontostathis et al. [51] does not provide feature importance after training. However, LR and XGBoost provide important features (shown in Table 1) when trained, and these features were later used to answer our questions as defined in Section 3.3. Although some of the above studies obtained their top performance, however, they do not answer the research questions asked in this paper.

4.2. Reconstruct and Interpret

The objective of this section is to trace the ML prediction decision. This is done by reconstructing the model's overall accuracy per predicted predatory conversation. For example, we trace the accuracy by checking the most relevant parts (i.e. phrase or words) with the highest prediction probability in a predatory conversation. At the same time, interpret the results in the context

Table 5: Performance measures after training the models. BiLSTM is weighted by encoding the data set before creating a word-embedding layer, while other models weighted with TFIDF

Algorithm	**Weighting**	**Accuracy**	**Precision**	**Sensitivity**	F_1 **Score**
LR	TFIDF	0.985	0.921	0.572	0.706
XGBoost	TFIDF	0.98	0.893	0.426	0.577
MLP	TFIDF	0.985	0.779	0.704	0.740
BiLSTM	Embedding	0.98	0.643	0.783	0.706

of digital forensic investigation.

A TFIDF with unigrams, bigrams and trigrams was performed in the whole corpus to obtain a feature vector representation. We also trained XGBoost and LR classifiers to map each conversation to their predicted category and to extract essential features.

4.2.1. Which phrases are connected with Sexual Predatory behaviour?

In an attempt to find key phrases that are connected with sexual predatory activities, we used TFIDF with bigrams and trigrams. We combined groups of five lines from a predatory conversation and applied a greedy search algorithm which states the following:

Select a predatory conversation, split it into lines of text and use the model to predict each line. Starting with the lines with the highest prediction probability; Sequentially merge lines by adding their results until two conversations are left, or the prediction probability is high enough.

The aim of the greedy search algorithm is to highlight words or phrases that are important only to the predatory label. Important words shown in Table 3 are important to both labels (non-predatory and predatory).

For example, in Fig. 4, we selected one predatory conversation and segmented it into 12 lines of smaller conversations. We used the models in Table 5 to predict the probability of each line coming from a predatory conversation. We recorded the probability result, as shown in the "proba" attribute. For the next steps,

Table 6: Confusion matrix of the LR

	Predicted label	
True label	**Non-predatory**	**Predatory**
Non-predatory	16198	25
Predatory	217	290

Table 7: Confusion matrix of the XGBoost

	Predicted label	
True label	**Non-predatory**	**Predatory**
Non-predatory	16197	26
Predatory	291	216

Table 8: Confusion matrix of the MLP

	Predicted label	
True label	**Non-predatory**	**Predatory**
Non-predatory	16122	101
Predatory	150	357

Table 9: Confusion matrix of the BiLSTM

	Predicted label	
True label	**Non-predatory**	**Predatory**
Non-predatory	16003	220
Predatory	110	397

Table 10: Summary of some of the work being done on sexual predator using PAN-12 data

Authors	**Features**	**Method**	**F1**
[50]	cosine similarity metric	Dictionary of terms	0.0498
[50]	cosine similarity metric	multinomial Naive Bayes	0.0354
[52]	pattern matching	similarity	0.48
[51]	rules	JRIP	0.47
[53]	n-grams	k-NN	0.70

we sequentially merged these segments based on the combinations resulting in the highest prediction probability. We do this until the prediction probability is the highest it can be or there are only two segments left.

The phrases and words shown in Fig. 4 and 5 respectively are part of the 100k important features (from Table 3) the model uses to make predictions. Using the greedy search algorithm helps to know which words are used to predict a conversation as predatory.

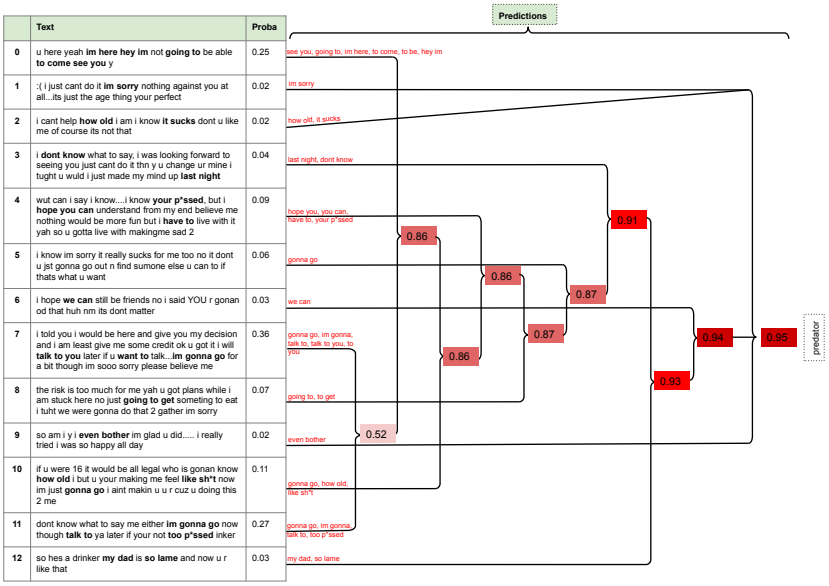
Therefore, results show that the parts of the conversation with the highest prediction probability contribute more to the overall prediction. Ideally, the segmented predictions indicate how the model came to its final decision, for a digital forensic investigation to examine and interpret to make an informed conclusion about the model's overall prediction.

4.2.2. Which words best are connected with Sexual Predatory behaviour?

In an attempt to find relevant words that are connected with sexual predatory activities, we used TFIDF with unigrams. Similarly, we segmented a predatory conversation into groups of five lines and made predictions using a greedy algorithm, as stated in 4.2.1.

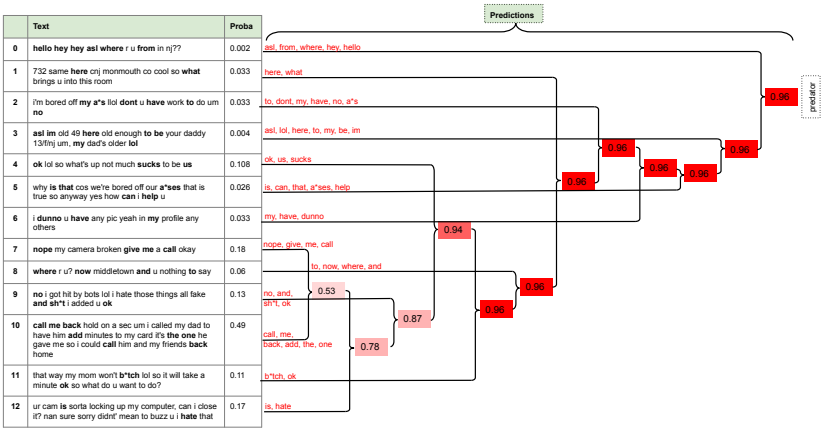
For example, Fig. 5 shows the words that are associated with sexual predatory behaviour in a conversation. We followed the same process, as described in section 4.2.1.

The difference between Fig. 5 and Fig. 4 is mainly features, wherein, Fig. 5



	Text	Proba
0	u here yeah **im here hey im** not **going to** be able **to come see you** y	0.25
1	:(i just cant do it **im sorry** nothing against you at all...its just the age thing your perfect	0.02
2	i cant help **how old** i am i know **it sucks** dont u like me of course its not that	0.02
3	i **dont know** what to say, i was looking forward to seeing you just cant do it thn y u change ur mine i tught u wuld i just made my mind up **last night**	0.04
4	wut can i say i know....i know **your p*ssed**, but i **hope you can** understand from my end believe me nothing would be more fun but i **have to** live with it yah so u gotta live with makingme sad 2	0.09
5	i know im sorry it really sucks for me too no it dont u jst gonna go out n find sumone else u can to if thats what u want	0.06
6	i hope **we can** still be friends no i said YOU r gonan od that huh nm its dont matter	0.03
7	i told you i would be here and give you my decision and i am least give me some credit ok u got it i will **talk to you** later if u **want to** talk...**im gonna go** for a bit though im sooo sorry please believe me	0.36
8	the risk is too much for me yah u got plans while i am stuck here no just **going to get** someting to eat i tuht we were gonna do that 2 gather im sorry	0.07
9	so am i y i **even bother** im glad u did..... i really tried i was so happy all day	0.02
10	if u were 16 it would be all legal who is gonan know **how old** i but u your making me feel **like sh*t** now im just **gonna go** i aint makin u u r cuz u doing this 2 me	0.11
11	dont know what to say me either **im gonna go** now though **talk to** ya later if your not **too p*ssed** inker	0.27
12	so hes a drinker **my dad** is **so lame** and now u r like that	0.03

Figure 4: Greedy algorithm showing how important **phrases** (extracted with XGBoost) are associated with sexual predatory. Important phrases shown in red appeared more than once and the cumulative use of these important **phrases** increased the probability of the conversation being predatory. The long texts are truncated at the end.



	Text	Proba
0	**hello hey hey asl where** r u **from** in nj??	0.002
1	732 same **here** cnj monmouth co cool so **what** brings u into this room	0.033
2	i'm bored off **my a*s** lol **dont** u **have** work **to** do um **no**	0.033
3	**asl im** old 49 **here** old enough **to be** your daddy 13/f/nj um, **my** dad's older **lol**	0.004
4	**ok** lol so what's up not much **sucks** to be **us**	0.108
5	why **is that** cos we're bored off our **a*ses** that is true so anyway yes how **can** i **help** u	0.026
6	i **dunno** u **have** any pic yeah in **my** profile any others	0.033
7	**nope** my camera broken **give me** a **call** okay	0.18
8	**where** r u? **now** middletown **and** u nothing **to** say	0.06
9	**no** i got hit by bots lol i hate those things all fake **and sh*t** i added u **ok**	0.13
10	**call me back** hold on a sec um i called my dad to have him **add** minutes to my card it's **the one** he gave me so i could **call** him and my friends **back** home	0.49
11	that way my mom won't **b*tch** lol so it will take a minute **ok** so what do u want to do?	0.11
12	ur cam **is** sorta locking up my computer, can i close it? nan sure sorry didnt' mean to buzz u i **hate** that	0.17

Figure 5: Greedy algorithm showing how important **words** (extracted with XGBoost) are associated with sexual predatory behaviour in a conversation. Important **words** shown in red some appeared more than once and the cumulative use of these important **words** increased the prediction probability of the conversation being predatory

uses words as features and Fig. 4 uses phrases. Overall, we observed that the presence of features, either words or phrases, from each conversation, greatly impacts the model's decision making as it can be seen from the greedy search algorithm (Fig. 5 and Fig. 4). The revealed features can be linked back to Table 3 where LG and XGBoost were used to extract important features in the whole corpus. We observed that when there are more important features in the conversation, the prediction value increases. The prediction value tells us that the conversation is predatory when it is close to one.

It is also interesting to note that our findings are similar to ChatCoder [23, 54], a software that uses rule-based ML models to identify online predatory conversations. As they have also reported and as shown in Fig. 5 and Fig. 4, it evident that ML models turn out to do better on general content rather than sexual aggressive content as one would expect. Again, this might be a crucial piece of evidence for digital forensic investigators, that, predators usually do not use massive sexual content when grooming their victims. The difference between their work and the work at hand is that they also used human experts to annotate their dataset manually. We, on the other hand, we let the model intelligently extract relevant features.

Our work does not try to identify what is the best way to present the model's chosen features to the users. We believe this is best left for User Interaction/User Experience/Human Computer Interaction research which intersects with Machine Learning systems. An example is work by [55] which looks at how to best have a NLP model that is used for Clinical text analysis to get feedback from users and update the view from user interactions. Another example is that of Information Visualisation that assists a user in general exploration of conversational text [56], different to our goal directed work in identifying parts of conversations that may be connected to a sexual predator. We foresee that a system such as that discussed in our paper, would still need optimisations that take into account the user experience. These would likely involve working directly with investigators of sexual predator crimes and understanding how they would react to the models outputs and optimising how this information

is presented to them for better efficacy in their work. Those optimisations are beyond the scope of this work.

4.3. Model Limitations

It is crucial when using ML models to be able to understand the limitations of models. This is exemplified in recent work to provide insight into ML models by including Model Cards [57] to model users (in this case, the investigators). In this part of the work, we do not plan to exhaustively identify limitations of models but provide a few examples to show that this is an essential part of the process of using ML in the DFPM process. Each ML model trained with data from this process would need to have its limitations explored. In this section, we show and discuss some of these limitations for illustration.

4.3.1. When and why do the models make mistakes?

This section explores the performance of the BiLSTM model. We used the What-if tool[4] to inspect the model where it makes prediction mistakes, as the confusion matrices in Table 6-9, show that the model has misclassified some of the conversations.

According to Wexler et al. [35], the What-if tool can be used to investigate model performances:

1. on a dataset using up to two models;
2. for a range or the entire features of the dataset;
3. using optimisation strategies;
4. by organising inference results into histograms, scatter plots or confusion matrices;
5. on manipulations to individual data point values; and

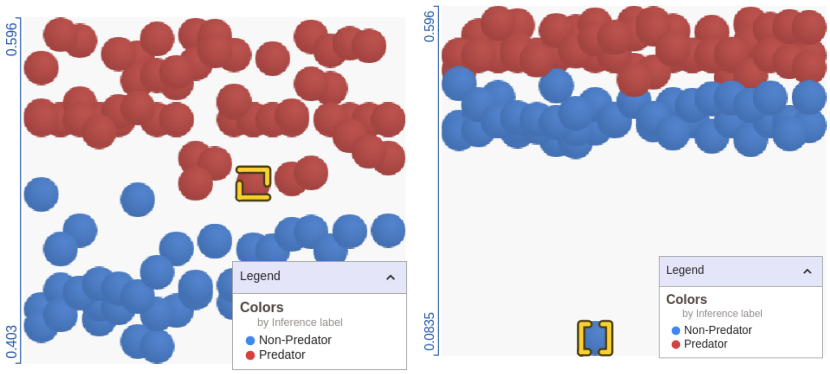
[4] `https://pair-code.github.io/what-if-tool/`

6. by arranging data in similarity to a certain data point using cosine similarity.

Therefore options 4 and 5 were considered for this paper with interest on false positives(i.e. non-predatory conversation wrongly classified as predatory) and false negatives (i.e. predatory conversation wrongly classified as no-predatory). To achieve this, we first investigated false-positive data points using the What-if tool, as shown in Fig 6, with a highlighted data point that contains the text:

"May I ghelp you? id love to be ghelped what is ghelping?"

We observed that this data point was incorrectly classified by the model as predatory with a probability of 0.594 while its accurate label is non-predatory, hence a false positive. We ran the predictions for each word in a text and found that the first word *"May"* contained a high probability of 0.65, thus causing the model to classify the whole text incorrectly. After removing this word from the text, we again ran the inferences on the What-if tool. The data point subsequently reverted to its accurate label (non-predatory) with a probability of 0.8.



(a) Before (b) After

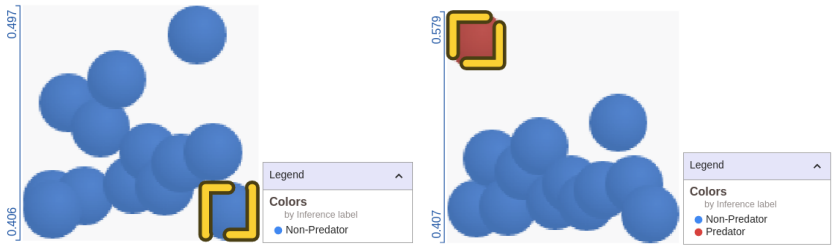
Figure 6: What-if tool running inferences to show how the model reacts to the modification of a false positive data point

Lastly, we investigated the false negative data points shown in Fig 7 with a

data point containing the following text:

"hello beautiful! u're here?"

We observed that the model wrongly classified this data point as non-predatory with a probability of 0.41, hence a false negative. We further investigated each of the words by running predictions, which resulted in the word "u're" having a lower probability of 0.17 which potentially caused the whole text to be misclassified. After removing this word from the text, we ran the inferences on the What-if tool, and the data point reverted to its accurate label (predatory) in Fig. 7 with a probability of 0.58. In essence, digital forensic investigators can use tools such as the what-if tool to understand the model's limitations and manipulate data points to examine the reasons why.



(a) Before (b) After

Figure 7: What-if tool running inferences to show how the model reacts to the modification of a false negative data point

To answer the question *When does a model make mistakes?*, we observed that certain words and phrases are essential features to identify a conversation as either predatory or non-predatory. A model can, therefore, make mistakes when a conversation contains features that overlap as essential features for both predatory and non-predatory classes. Similarly, when a word or phrase in a conversation carries more weight as an essential feature for one class yet appearing in a conversation of another class can also cause the model to incorrectly classify the whole conversation because of that particular phrase or word. Thus, resulting in a model predicting a conversation as predatory or non-predatory

with a probability score of close to 0.5.

5. Conclusion and Future Work

Overall, this work aimed to show the use of ML in digital forensic investigation processes. Mainly, to support digital forensic investigators to discover incriminating concepts or predatory behaviour in online chats systematically.

The Interpretability of ML was examined by applying the greedy search algorithm using the trained models in Table 3 on the predatory conversation to determine which terms allow to mark a conversation as predatory. We captured the false-negative cases in F_1 score since our aim was to learn what happens when a positive case is identified, that is, what makes it positive.

Through considering the decision-making cues of ML models, we were able to identify words and phrases that are more likely to be linked to sexual predatory behaviour.

Moreover, we used the What-if Tool to evaluate the ML model's performance so that digital forensic experts can understand the limitations of the models. We discovered that the model makes a mistake when: (i) the conversations consist of features which are important to both the classes (predatory and non-predatory), (ii) consists of features which are less important to both classes, (iii) and features that overlap in both classes.

Due to the scarcity of publicly available data and the sensitive nature of predatory related datasets, only the PAN-12 dataset was used in this work. Therefore, at the moment, being able to perform a quantitative test for generalisation is not yet possible. Thus, our work only considered qualitative analysis of different ML models on top of the cross-validated performance measures on the predictive power of ML on the task at hand. It is also desirable to include analysis that may quantify how well this approach would work across different predatory datasets. As future work, we are looking into getting local datasets possibly labelled by forensic experts, that will assist in an improved understanding of the predatory behaviour in general.

Availability of data and material

The datasets analysed during the current study are available in the PAN-12 repository, https://pan.webis.de/clef12/pan12-web/author-identification.html

Competing Interests

The authors declare that they have no competing interests.

Funding

Not Applicable.

Author's Contributions

CH contributed to writing the paper, Cybersecurity concepts, sourced the dataset, Natural Language Processing and Machine Learning modelling and analysis. JHP contributed to Cybersecurity concepts, Integrated Digital Forensic Process Model and reviewed Manuscript. TJ contributed in Natural Language Processing and Machine Learning modelling and analysis and Reviewed Manuscript. VN contributed in Natural Language Processing and Machine Learning modelling and analysis and Reviewed Manuscript. All authors read and approved the final Manuscript.

Acknowledgements

Not Applicable.

References

References

[1] BBC, New campaign challenges online child sex predators, 2019. URL: https://www.bbc.com/news/uk-scotland-47584148, Accessed: 25.04.2019.

[2] S. Grzonkowski, N. A. Lekhac, et al., Enabling trust in deep learning models: A digital forensics case study, arXiv preprint arXiv:1808.01196 (2018).

[3] M. D. Kohn, M. M. Eloff, J. H. Eloff, Integrated digital forensic process model, Computers & Security 38 (2013) 103–115.

[4] B. Hitchcock, N.-A. Le-Khac, M. Scanlon, Tiered forensic methodology model for digital field triage by non-digital evidence specialists, Digital investigation 16 (2016) S75–S85.

[5] N. Pendar, Toward spotting the pedophile telling victim from predator in text chats, in: International Conference on Semantic Computing (ICSC 2007), IEEE, 2007, pp. 235–241.

[6] M. Ebrahimi, C. Y. Suen, O. Ormandjieva, Detecting predatory conversations in social media by deep convolutional neural networks, Digital Investigation 18 (2016) 33–49.

[7] L. Arras, F. Horn, G. Montavon, K.-R. Müller, W. Samek, What is relevant in a text document?: An interpretable machine learning approach, PloS one 12 (2017) e0181142.

[8] P. K. Smith, F. Thompson, J. Davidson, Cyber safety for adolescent girls: bullying, harassment, sexting, pornography, and solicitation, Current Opinion in Obstetrics and Gynecology 26 (2014) 360–365.

[9] S. Edwards, A. Nolan, M. Henderson, A. Mantilla, L. Plowman, H. Skouteris, Young children's everyday concepts of the internet: A platform for cyber-safety education in the early years, British Journal of Educational Technology 49 (2018) 45–55.

[10] Q. Li, Bullying in the new playground: Research into cyberbullying and cyber victimisation, Australasian Journal of Educational Technology 23 (2007).

[11] E. Van der Walt, J. H. Eloff, J. Grobler, Cyber-security: Identity deception detection on social media platforms, Computers & Security 78 (2018) 76–89.

[12] T. Kucukyilmaz, B. B. Cambazoglu, C. Aykanat, F. Can, Chat mining: Predicting user and message attributes in computer-mediated communication, Information Processing & Management 44 (2008) 1448–1466.

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

[14] P. Anderson, Z. Zuo, L. Yang, Y. Qu, An intelligent online grooming detection system using ai technologies, in: 2019 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), IEEE, 2019, pp. 1–6.

[15] G. Inches, F. Crestani, Overview of the international sexual predator identification competition at pan-2012, in: CLEF 2012 Evaluation Labs and Workshop, volume 30, 2012.

[16] C. Ngejane, G. Mabuza-Hocquet, J. Eloff, S. Lefophane, Mitigating online sexual grooming cybercrime on social media using machine learning: A desktop survey, in: 2018 International Conference on Advances in Big Data, Computing and Data Communication Systems (icABCD), IEEE, 2018, pp. 1–6.

[17] L. N. Olson, J. L. Daggs, B. L. Ellevold, T. K. Rogers, Entrapping the innocent: Toward a theory of child sexual predators' luring communication, Communication Theory 17 (2007) 231–251.

[18] R. O'Connell, A typology of child cyber exploitation and online grooming practices, Preston, UK: University of Central Lancashire (2003).

[19] H. Whittle, C. Hamilton-Giachritsis, A. Beech, G. Collings, A review of online grooming: Characteristics and concerns, Aggression and violent behaviour 18 (2013) 62–70.

[20] M. W. R. Miah, J. Yearwood, S. Kulkarni, Detection of child exploiting chats from a mixed chat dataset as a text classification task, in: Proceedings of the Australasian Language Technology Association Workshop 2011, 2011, pp. 157–165.

[21] Y. R. Tausczik, J. W. Pennebaker, The psychological meaning of words: LIWC and computerized text analysis methods, Journal of Language and Social Psychology 29 (2010) 24–54.

[22] A. Kontostathis, L. Edwards, J. Bayzick, A. Leatherman, K. Moore, Comparison of rule-based to human analysis of chat logs, Communication Theory 8 (2009).

[23] I. McGhee, J. Bayzick, A. Kontostathis, L. Edwards, A. McBride, E. Jakubowski, Learning to identify internet sexual predation, International Journal of Electronic Commerce 15 (2011) 103–122.

[24] A. E. Cano, M. Fernandez, H. Alani, Detecting child grooming behaviour patterns on social media, in: International conference on social informatics, Springer, 2014, pp. 412–427.

[25] P. J. Black, M. Wollis, M. Woodworth, J. T. Hancock, A linguistic analysis of grooming strategies of online child sex offenders: Implications for our understanding of predatory sexual behavior in an increasingly computer-mediated world, Child Abuse & Neglect 44 (2015) 140–149.

[26] M. Meyer, Machine learning to detect online grooming, 2015.

[27] D. Liu, C. Y. Suen, O. Ormandjieva, A novel way of identifying cyber predators, arXiv preprint arXiv:1712.03903 (2017).

[28] F. Amato, A. Castiglione, A. De Santo, V. Moscato, A. Picariello, F. Persia, G. Sperlí, Recognizing human behaviours in online social networks, Computers & Security 74 (2018) 355–370.

[29] A. Souri, S. Hosseinpour, A. M. Rahmani, Personality classification based on profiles of social networks' users and the five-factor model of personality, Human-centric Computing and Information Sciences 8 (2018) 24.

[30] J. A. Kloess, C. E. Hamilton-Giachritsis, A. R. Beech, Offense processes of online sexual grooming and abuse of children via internet communication platforms, Sexual Abuse 31 (2019) 73–96.

[31] N. Lorenzo-Dus, A. Kinzel, So is your mom as cute as you?': Examining patterns of language use by online sexual groomers, Journal of Corpora and Discourse Studies 2 (2019) 1–30.

[32] C. Molnar, Interpretable Machine Learning, Lulu. com, 2019.

[33] M. K. Rogers, Psychological profiling as an investigative tool for digital forensics, in: Digital Forensics, Elsevier, 2016, pp. 45–58.

[34] S. Satpathy, C. Mallick, S. K. Pradhan, Big data computing application in digital forensics investigation and cyber security, International Journal of Computer Science and Mobile Applications, National Conference on "The Things Services and Applications of Internet of Things" (2018) 129–136.

[35] J. Wexler, M. Pushkarna, T. Bolukbasi, M. Wattenberg, F. Viégas, J. Wilson, The what-if tool: Interactive probing of machine learning models, IEEE transactions on visualization and computer graphics 26 (2019) 56–65.

[36] D.-O. Jaquet-Chiffelle, E. Casey, M. Pollitt, P. Gladyshev, A framework for harmonizing forensic science practices and digital/multimedia evidence, Technical Report, OSAC/NIST, 2018.

[37] I. Standard, Information Technology — Security Techniques — Investigation principles and processes, 2015.

[38] A. Valjarević, H. Venter, R. Petrović, Iso/iec 27043: 2015—role and application, in: 2016 24th Telecommunications Forum (TELFOR), IEEE, 2016, pp. 1–4.

[39] X. Deng, Y. Li, J. Weng, J. Zhang, Feature selection for text classification: A review, Multimedia Tools and Applications 78 (2019) 3797–3816.

[40] D. Subramanien, N. Whitear-Nel, The exclusion of evidence obtained by entrapment: An update, Obiter 32 (2011) 634–650.

[41] M. Ebrahimi, C. Y. Suen, O. Ormandjieva, A. Krzyzak, Recognizing predatory chat documents using semi-supervised anomaly detection, Electronic Imaging 2016 (2016) 1–9.

[42] T. Chen, C. Guestrin, Xgboost: A scalable tree boosting system, in: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, ACM, 2016, pp. 785–794.

[43] D. G. Kleinbaum, K. Dietz, M. Gail, M. Klein, M. Klein, Logistic regression, Springer, 2002.

[44] F. E. Curtis, X. Que, A quasi-Newton algorithm for nonconvex, nonsmooth optimization with global convergence guarantees, Mathematical Programming Computation 7 (2015) 399–428.

[45] J. H. Friedman, Greedy function approximation: a gradient boosting machine, Annals of statistics (2001) 1189–1232.

[46] E. Alpaydin, Introduction to machine learning, MIT press, 2009.

[47] M. Schuster, K. K. Paliwal, Bidirectional recurrent neural networks, IEEE Transactions on Signal Processing 45 (1997) 2673–2681.

[48] B. Plank, A. Søgaard, Y. Goldberg, Multilingual part-of-speech tagging with bidirectional long short-term memory models and auxiliary loss, in: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), 2016, pp. 412–418.

[49] E. Villatoro-Tello, A. Juárez-González, H. J. Escalante, M. Montes-y Gómez, L. V. Pineda, A two-step approach for effective detection of mis-

behaving users in chats, in: CLEF 2012 Evaluation Labs and Workshop, volume 1178, 2012.

[50] D. Vilariño, E. Castillo, D. Pinto, I. Olmos, S. León, Information retrieval and classification based approaches for the sexual predator identification, in: CLEF 2012 Evaluation Labs and Workshop, 2012.

[51] A. Kontostathis, A. Garron, K. Reynolds, W. West, L. Edwards, Identifying predators using chatcoder 2.0., in: CLEF 2012 Evaluation Labs and Workshop, 2012.

[52] A. Vartapetiance, L. Gillam, Quite simple approaches for authorship attribution, intrinsic plagiarism detection and sexual predator identification, in: CLEF 2012 Evaluation Labs and Workshop, 2012.

[53] I.-S. Kang, C.-K. Kim, S.-J. Kang, S.-H. Na, Ir-based k-nearest neighbor approach for identifying abnormal chat users., in: CLEF 2012 Evaluation Labs and Workshop, 2012.

[54] A. Kontostathis, L. Edwards, A. Leatherman, Text mining and cybercrime, Text Mining: Applications and Theory. John Wiley & Sons, Ltd, Chichester, UK (2010) 149–164.

[55] G. Trivedi, Clinical text analysis using interactive natural language processing, in: Proceedings of the 20th International Conference on Intelligent User Interfaces Companion, 2015, pp. 113–116.

[56] E. Hoque, Visual text analytics for online conversations: Design, evaluation, and applications, in: Companion Publication of the 21st International Conference on Intelligent User Interfaces, 2016, pp. 122–125.

[57] M. Mitchell, S. Wu, A. Zaldivar, P. Barnes, L. Vasserman, B. Hutchinson, E. Spitzer, I. D. Raji, T. Gebru, Model cards for model reporting, in: Proceedings of the conference on fairness, accountability, and transparency, 2019, pp. 220–229.