Knowledge-Based Systems 259 (2023) 110039

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

Online grooming detection: A comprehensive survey of child exploitation in chat logs

Parisa Rezaee Borj *, Kiran Raja, Patrick Bours

NTNU, Teknologivegen 22, Gjøvik, 2815, Norway

ARTICLE INFO

Article history:
Received 5 July 2022
Received in revised form 23 August 2022
Accepted 14 October 2022
Available online 21 October 2022

Keywords:
Cyber grooming
Child exploitation
Online predators
Chat analysis
Stylometry
Text analysis
Keystroke dynamics

ABSTRACT

Social media platforms present significant threats against underage users targeted for predatory intents. Many early research works have applied the footprints left by online predators to investigate online grooming. While digital forensics tools provide security to online users, it also encounters some critical challenges, such as privacy issues and the lack of data for research in this field. Our literature review investigates all research papers on grooming detection in online conversations by looking at the psychological definitions and aspects of grooming. We study the psychological theories behind the grooming characteristics used by machine learning models that have led to predatory stage detection. Our survey broadly considers the authorship profiling research works used for grooming detection in online conversations, along with predatory conversation detection and predatory identification approaches. Various approaches for online grooming detection have been evaluated based on the metrics used in the grooming detection problem. We have also categorized the available datasets and used feature vectors to give readers a deep knowledge of the problem considering their constraints and open research gaps. Finally, this survey details the constraints that challenge grooming detection, unaddressed problems, and possible future solutions to improve the state-of-the-art and make the algorithms more reliable.

1. Introduction

Different online messaging platforms such as chat functions and instant messaging applications in social networking sites have evolved as an alternative to a standard communication medium. Such platforms allow individuals to exchange messages peer-to-peer without explicit content moderation, which can be exploited for various malicious intents. The internet-facilitated malicious activities vary from normalizing certain destructive behaviors by online extremism organizations [1,2], disseminating fake news [3,4] and spammers [5] to drastically impacting users' mental health [6]. Targeted chats can be used to spread hatred [7], manipulate the victim for propaganda, coordinate criminal or terrorist activities [1,2,8], radicalization, and in the worst of scenarios to target under-aged online users (minors and children) for sexual favors and abuse [9,10]. Unlike in public chats or discussions, targeted messages, in most cases, exploit an already existing online relationship with the other group members in the network [1,2,11]. In cases where such prior relation does not exist, the malicious actor spends time building the relation, often referred to as online grooming, and eventually targets the victim [12].

Research on the retrospective view of online grooming experienced by minors showed that 25% of the minor participants talked with adult strangers [13,14]. More importantly, 65% of those who spoke with a stranger experienced sexual solicitation from an adult stranger. 23% of participants revealed that they had conversations with stranger adults that followed a grooming pattern, and around 38% of them established a confidential relationship with the groomer [13]. Another report by the National Center for Missing and Exploited Children (NCMEC)[1] shows that more than a million child abuse cases were reported in 2019 [14]. Also, the technology companies reported to the US National Center for Missing and Exploited Children (NCMEC) over 45 million photographs and videos of sexually abused children, and New York Times claimed that this number increased twice in only one year [14]. The increasing number of these reports leads to a concern that requires attention.

The internet offenses against adolescents vary from exchanging child pornography to finding potential victims, engaging in a dangerous relationship, and normalizing certain destructive behaviors to lower the child's inhibition. Much research in digital

* Corresponding author.
E-mail addresses: parisa.rezaee@ntnu.no (P.R. Borj), kiran.raja@ntnu.no (K. Raja), patrick.bours@ntnu.no (P. Bours).

[1] https://www.missingkids.org/home.

https://doi.org/10.1016/j.knosys.2022.110039

forensics has been produced in detecting online sexual predators. However, the majority of them have focused on children's images and videos [9,15–17]. For instance, Lee et al. [17] provided a comprehensive survey on child sexual abuse material detection. The main focus was distribution methods, policy and legal framework dimensions, and detection applications and implementations. They mostly surveyed information about image hash database, keywords,web-crawler, detection based on filenames and metadata, and visual detection [17].

Detecting such offenses in public communications on social media and public chats is relatively easy, as they can be monitored by employing content moderators who assess the content manually or through an automated mechanism such as using profanity filters [18,19]. Automated public chat/discussion moderation algorithms can be devised using large-scale training data. However, several challenges can be foreseen in devising and using the algorithms effectively. For instance, large-scale data may not be available for training moderating algorithms, or the privacy regulations impose restrictions on using such data even when available [20–23]. Such challenges have a hindering impact on the advancements for preventing misuse of online messaging platforms.

A robust and automated surveillance system that increases children's security on online platforms requires an in-depth knowledge of a predator's behavior. Understanding the online predators' patterns facilitates better detection mechanisms, thereby educating children to react appropriately in dangerous situations. At the same time, digital forensics cases require operational evidence that can be used in court, which leads to the analysis of massive amounts of data and increases the forensics investigation load [24]. Since monitoring private messages in different applications is more challenging, in this research, we mainly focus on cases where child predators use different applications such as chat rooms and social network applications (Twitter, Facebook, and Instagram) to engage in a relationship with minors.

Despite the importance of the grooming problem, there is a lack of algorithmic surveys for grooming detection on online chat logs. Few research surveys focused on online harassment and sexual predation on online platforms [25–27]. For instance, Razi et al. [27] reviewed various approaches for sexual risk detection from a human-centered view considering sex trafficking, sexual harassment, and sexual grooming, and Miljana et al. [25] investigated the diversity of cyber-aggression, cyberbullying, and cyber-grooming and identified their target categories. We mainly focus on analyzing all the works related to peer-to-peer chat communication for sexual abuse of minors on online platforms, considering the tremendous threats to children and minor victims. Especially our primary goal is to provide an extensive survey on a scenario where a predator (i.e., a malicious actor with an intent to get sexual favors from minors) targets a victim.

1.1. Contributions

To the best of our knowledge, this is the first survey that reviews all research work on grooming detection focusing on chat logs. We detail and compare all research works based on an algorithmic performance perspective, considering the psychological theories behind the grooming characteristics used by machine learning methods for grooming detection in chat conversations. The contributions of this work are listed below:

- Using an algorithmic performance evaluation perspective, we propose a conceptual framework for systematically reviewing online grooming detection literature, mainly on chat conversations.
- We survey the psychological investigation of the grooming procedure by various research works. Also, our survey shows how machine learning models have applied grooming attributes for grooming stage detection based on psychological theories.
- This research gives a profound explanation of feature sets along with their constraints and potential solutions. Also, the available datasets are listed considering the limitations to supplement the readers with the state-of-the-art.
- This survey discusses the role of authorship profiling in grooming detection in the early stages by studying the existing works on text mining and keystroke dynamics that cope with detecting the age and gender of authors on social media.
- We also categorize author profiling research papers based on feature sets and data for age and gender detection works.
- This survey details open research problems and the potential gaps in online grooming detection literature to benefit the reader with a piece of more profound knowledge.
- Present potential future works to improve the algorithms in real-time scenarios.

Fig. 1 represents the overall contributions of this research work.

It should be noted that our primary focus in this survey is chatlogs and short text analysis for grooming detection. This research paper does not include the dark web investigation for child exploitation material such as image processing, hash databases, and distribution methods [9,15–17]. The taxonomy of this survey paper is illustrated in Fig. 2.

The remainder of this paper is organized as follows. Section 2 first gives an overview of online grooming definitions and analyzes the psychological perspectives of sexual predators and different grooming characteristics. Section 3 breaks down the problem into different categories where Section 3.1 presents the available datasets for the research, and Section 3.2 discusses the various feature vectors that have been used in research works. We will also discuss how the performance of each method for grooming detection is evaluated in Section 3.3. The paper continues by surveying various grooming detection techniques in Section 3.4. Section 3.4.1 describes how the grooming characteristics were used for detecting online grooming stages. Section 3.4.2 details previous research works for predatory conversation detection, and Section 3.4.3 summarizes the techniques for predatory identification. Section 3.4.4 summarizes the authorship profiling techniques for cyber-grooming purposes in online platforms. Section 4 discusses the challenges and open gaps of grooming detection and possible solutions along with its constraints, and finally, we conclude the paper in Section 5.

2. Online grooming

2.1. Definition of online grooming

Online grooming is defined as a process performed by malicious actors such as pedophiles to entrap their victims [23,28]. However, it is recommended not to use the term 'pedophile' to define grooming as it is used only after a precise clinical diagnosis and cannot be used for all offenders [29]. One of the first comprehensive definitions of grooming was given by Craven et al. [29] as below:

> "*process by which a person prepares a child, significant others, and the environment for the abuse of this child. Specific goals include gaining access to the child, gaining compliance, and maintaining the child's secrecy to avoid disclosure. This process strengthens the offender's abusive pattern, as it may be used to justify or deny their actions.*"

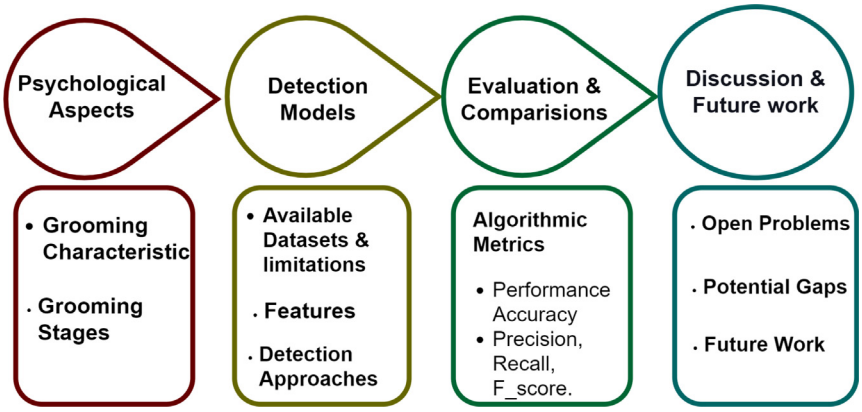

Fig. 1. Overall contributions of this research work.

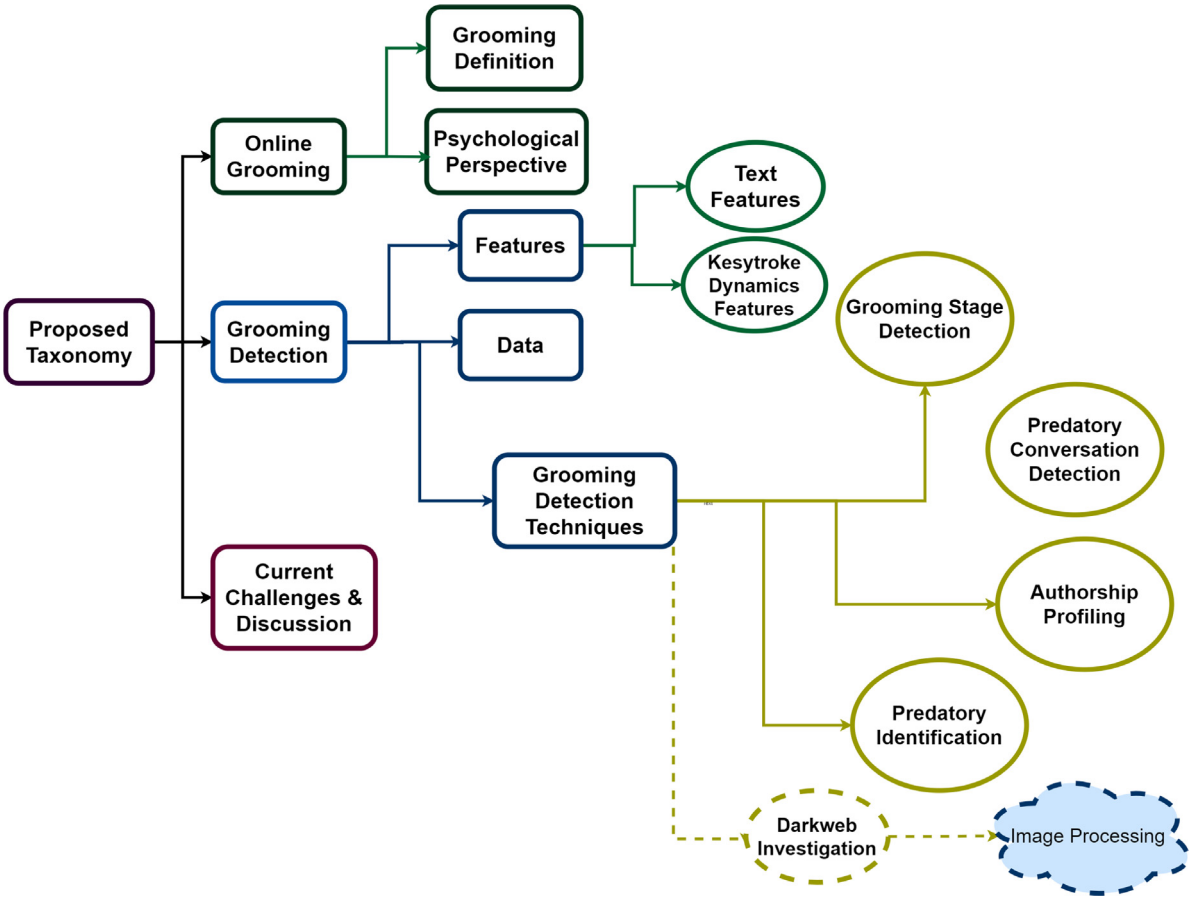

Fig. 2. The proposed taxonomy for online grooming detection problem.

Online grooming is a process to gain, persuade, and engage a child in sexual activity where the internet is used as a medium for access. The offender tries to avoid disclosure by keeping the victim's secrecy [30–33]. One should consider that the grooming's psychological effect might be as intense as the physical effects since it can change the victim psycho-socially. The following section will discuss some previous research works that have investigated the psychological aspects of online grooming for the reader's convenience.

2.2. Psychological perspectives of online grooming

Child grooming has been researched in many works, including social and psychological areas [12,34–36]. Predators have different interests, such as curiosity viewing, cyber sex, or distribution of child abuse material. The individual differences between predators make online grooming detection complicated. Psychological research has shown that grooming detection is multifaceted and complex due to its variation in the period, type, and intensity. Therefore, it is difficult to predict where the grooming starts and when it is done [37]. In some cases of child abuse, it is challenging to detect the incident before the predator gains access to the victim physically [37]. In addition, some predators like to meet the victim in person (called hands-on child predators or Contact Child Sex Offenders (CCSO)). At the same time, some are just fantasy-driven, and they are not willing or interested in meeting the child in the real world (Fantasy

Child Sex Offender (FCSO)) [35]. Predators who intend to meet their victims in person have different motives than those who merely have sexual fantasies about their victims [35]. Predators who target their victims online without meeting them in the real world face a minor punishment, making them less reticent to perform harmful online actions. The threat of the fantasy-driven predator is as critical as the threat of the contact-drive predator because they are more likely to repeat their online acts. At the same time, they can harm the victims both physically and psychologically [35]. Investigating the demographic features of online predators has also indicated that online predators are mostly younger than offline child abusers, and they are single and unemployed [34].

Chiu et al. [35] have performed exploratory research to investigate the difference in the content of an online predatory conversation of a contact-driven predator versus that of a fantasy-driven predator who does not have any intention to meet the victim in person. The data included 4353 messages where there were 12 victims and nine predators. They coded each conversation in various manners with a computer annotation program, defined different hypotheses for messages, and explored their theory by applying Statistical Discourse Analysis (SDA) [36]. It was found that generally, predators talk about their prior experience with the new victim to build a confidential relationship. They also use particular words such as first-person pronouns and negative and positive emotional expressions. However, a difference is that some predators use grooming tactics to convince the victim to meet online while the other predators do not [35].

Online grooming's manner and timing can differ from face-to-face cases [38]. Although they might have some similarities in conversing about traveling, parents, analyzing parents' work time, and sexual conversations about past relationships [39]. Whittle et al. [12] studied the impact of online grooming on victims by interviewing eight young victims who were abused through online grooming. It was shown that as the child's level of vulnerability increases, online grooming would affect the victim more adversely.

Further, it can be challenging for police and family or community members to detect the grooming before the abuse occurs as predators vary in their strategies regarding their fear of getting caught [31]. When the parents and people around the victim do not know grooming tactics, it is difficult to distinguish a child grooming conversation from an adult's typical interaction with the child. Winters and Jeglic [31] investigated the possibility of grooming recognition by people. They performed an experiment where participants were asked to read some vignettes and rate the likelihood of a person being a child molester. It was discovered that people might not detect the potential writing pattern of a child predator, and giving hindsight could bias the result of rating by overestimating the likelihood that if the person is a predator [31].

3. Online grooming detection

The goal of grooming detection is to build operational evidence to apply in a court of law while challenging considering the tremendous number of online cases. Pattern recognition and machine learning methods have facilitated extensive data analysis, including investigating chat logs in an automated manner. They have been well explored for finding the potential threats in online platforms. The approaches typically consist of collecting the relevant data, extracting the most relevant features, and devising a classifier for arriving at a decision [28,33,40–48]. These approaches, if suitably engineered, can also provide a faster processing time to detect predators at an early stage and decrease online threats to young victims.

To give a complete overview of online grooming detection, we present relevant data and feature vectors used in different research works in the following sections. We continue discussing how previous works [33,40,41] tried to detect online grooming with different perspectives. Some works [28,33,40,42,43] have explored various phases of grooming to investigate the different themes in online conversations, while others focused on predatory conversation detection [49–51], predatory identification [41,52,53], or author profiling [44–48].

3.1. Datasets

Online grooming mostly happens on private chat logs on social media or open chat platforms. Therefore, the relevant data for online grooming detection should have the same characteristics as the chat logs on mentioned platforms. Due to the non-availability of such data, many research works have identified datasets with similar characteristics for devising algorithms for online grooming and predator conversation detection [28,33,33,40–51]. They have used different datasets for grooming detection, and Table 1 represents an overview of various datasets used in various research for grooming detection. We detail the same in the following sections and discuss the relevance of such datasets for the problem.

3.1.1. Datasets for online grooming detection

A popular dataset source for predatory detection is the perverted justice website.[2] Perverted Justice Foundation, more commonly known as Perverted-Justice (often shortened to PeeJ or PJ), is an American organization based in California and Oregon where police officers pretend to be children to attract and trap predators. Ashcroft et al. [58] used the Perverted-Justice website to include texts written by predators and also used book reviews, blogs, and chat logs for non-predatory texts. Along with the PJ dataset, Sulaiman et al. [55] used Literotica (www.literotika.com) data that contains conversations between adults that express their passion legally about sexual topics.

Recently, multiple works [51–53,63,64,69] have used the PAN2012 [41] dataset in which the primary goal was to identify sexual predators. The dataset contains chat conversations from 4 different sources. Two of these are regular IRC chats that contain non-sexual chats. These two sources are from two websites, i.e., http://www.irclog.org/ and http://krijnhoetmer.nl/irc-logs/. The third source used for the PAN2012 dataset was chatlogs from the Omegle chat service. The Omegle chat service intends to connect two adults for a chat randomly. A large part of the conversations on Omegle has a sexual character. These are not classified as predatory conversations, as the participants in these chats are all adults. The data on sexual topics between adults cover the false-positive cases in the dataset for grooming detection. Finally, the fourth source used was (parts of) conversations published on PJ. Complete conversations on PJ have often been split into multiple conversations in the PAN2012 dataset, depending on the time between the messages (for details, see [41]). Approximately 4% of the conversations in the PAN2012 dataset are from PJ and therefore classified as predatory, while the remaining conversations are considered non-predatory.

Online game platforms also provide a possibility for private and public communication. In games where children play, there will also be a potential risk that predators form a threat to minor users. Following the same motivation, Cheong et al. [74] used MovieStarPlanet as a source for data to detect grooming behavior in online game platforms.

[2] http://www.perverted-justice.com/.

Table 1
A summary of datasets used for grooming detection in previous works.

Data	Sources	Ref
Predatory data	www.perverted-justice.com	[23,28,33,40–43,54–60]
	PAN2012	[51–53,61–73]
	MovieStarPlanet	[74]
Non-predatory data	www.literotika.com	[40,42,55]
	http://www.irclog.org	[41]
	http://krijnhoetmer.nl/irc-logs	[41]
	Omegle	[41]
	Twitter	[46,75–80]
	Blogs, Book Reviews	[44,58,59]
	British National Corpus(BNC)	[81,82]

3.1.2. Datasets for attributes detection

Previous works [81,82] have used the British National Corpus[3] (BNC) to address online grooming detection by looking at the author's style and author information. For instance, Koppel et al. [81] were able to identify the gender of the author with reasonable accuracy, analyzing a large corpus of formal written texts (both fiction and non-fiction) from the British National Corpus. Tam and Martell [83], and Lin [84] used the data collected from chat rooms for age detection and author profiling. Also, some previous research works [46,75–78] have used tweets to determine the online authors' demographic attributes.

3.1.3. Dataset constraints

Each of the used datasets has limitations that challenge online grooming detection. We detail these limitations of the datasets in the cyber-grooming problem below:

- **Imbalanced Data**: Grooming conversations are a small portion of the massive number of conversations taking place on online platforms. A dataset for online grooming detection should follow a similar distribution. In [85], it was shown that the number of predatory queries on a particular peer-to-peer network was approximately only 0.25% of all queries. The percentage of predatory conversations might be higher for online chat applications, but it will probably still be low compared to the percentages of non-predatory conversations. Also, the percentage depends heavily on the particular chat application. For example, an application-specific for children will attract more sexual predators than a chat to discuss football or car models. It can be assumed that a dataset resembling a real-life situation will be biased, making it challenging to find the predatory behavior patterns for devising efficient machine learning methods. The highly imbalanced data has made detecting sexual predators on online platforms complicated. Therefore, it is critical to have balanced data using machine learning methods to solve a problem while it is highly imbalanced. Some papers attempted to consider this setting and have tried to address this as an imbalanced data problem [62,72,73,86].
- **Non-Standard Structure**: Chat logs, blogs, and tweets are short texts, and it is more challenging to analyze them than the standard text, in which context information and large sentence constructs give enough information. Short chat texts have different structures, contain spurious information, and are full of grammar errors, abbreviations, slang words and phrases, and spelling mistakes. So, it is challenging to gain information about the conversation partners to detect grooming.
- **Security and Privacy Issues**: A crucial challenge in finding online predators is gathering the data. Access to archived data of chats between victims and predators is challenging due to significant privacy and legal issues. Online service providers for chat platforms do (generally) not record chat data; even if it is collected, this data is not publicly available for research. In addition, collecting this data requires the informed agreement of the participants.

3.2. Features for online grooming detection

Predatory conversation data has different characteristics when compared to non-predatory conversation data. The characteristic differences stem from the writing style between two specific persons, between classes of persons (e.g., adults versus children), or between the themes of the conversation or text. Previous works have applied different feature vectors to capture different characteristics of the predators for grooming detection [48,51,62–64,71–73]. Different methods to extract information from chat logs, including stylometry, Keystroke Dynamics (KD), and features that capture the psychological characteristics of the authors, such as Linguistic Inquiry and Word Count (LIWC) and authors' activities like chat-based feature vectors, have been explored. This section first gives an overview of different stylometry features such as the one-hot representation of a word/text and distributed representations of word vectors that are the statistical analysis of variations in the writing style. Then, it introduces LIWC, chat-based characteristics, and KD feature vectors.

- **Bag of Words** (BoW) is the conventional method for creating a one-hot representation of a text. It can be regarded as a dictionary of all possible words or tokens that do not consider their relationship. BoW representation provides a high-dimensional vector (for instance, 10 000 or more), where a text is represented in a sparse manner where most of the values in BoW vectors are zero except the ones that represent the dictionary words in the text. A disadvantage is that it can provide the same feature vectors for different texts with different meanings. Various techniques have been further proposed to code the non-zero entries in BoW feature representations to improve the feature representation, such as Binary, Term Frequency (TF), Term Count (TC), and Term Frequency–Inverse Document Frequency (TF-IDF) [87–90]. BoW features for sexual predatory detection has been well used by a number of previous works [23,49,51–53,59,60,62–67,69–74].
  Suppose that the data set D is defined as $D = X * Y$ where X represents the set of n documents, i.e. $X = \{d_1, d_2, \ldots, d_n\}$. Each document can be represented by an m-dimensional feature vector, where m is the size of the dictionary. Document i is represented as $d_i = (f_1^i, f_2^i, \ldots, f_m^i)$, where f_j^i (for $j \in \{1, \ldots, m\}$) represents the feature value for the jth word in the dictionary in document d_i. Y contains the class labels for the sexual predatory detection problem. The values of Y can be represented by $\{predatory, non-predatory\}$ or simply by 0 and 1, representing predatory and non-predatory,

[3] http://www.natcorp.ox.ac.uk/.

respectively. The BoW models (Binary, TC, TF, and TF-IDF) determine the feature values based on the word occurrence in each document without concerning where each word has occurred in the document and the relationships between the words.

We provide a simple illustration of each BoW technique for the reader's convenience. Suppose we have three documents, where document 1 is *"I live with my parents. I work at a hospital"*, document 2 is *"My parents are at work"* and document 3 is *"I work at a hospital"*. BoW considers all the words found in the three documents as *live, with, parents, are, at, work, I, hospital, my, a*, and the size m of the dictionary is ten in this example. So, each document is represented by a feature vector of length ten. Each BoW model calculates the feature vector based on the explanation given below.

- **Binary**: If a word is present at least once in the document, the entry value will be set as 1. So, the feature vectors for each document in the above example will be: $d_1 = [1, 1, 1, 0, 1, 1, 1, 1, 1, 1]$, $d_2 = [0, 0, 1, 1, 1, 1, 1, 0, 0, 1, 0]$, and $d_3 = [0, 0, 0, 0, 1, 1, 1, 1, 0, 1]$. Multiple researchers have applied binary representations for extracting lexical information to detect predators [51–53,62].
- **Term Count (TC)**: The entry value shows the number of appearances of the words in the document. The weight of the vector displays the number of words in the text. The feature vectors for the example will be: $d_1 = [1, 1, 1, 0, 1, 1, 2, 1, 1, 1]$, $d_2 = [0, 0, 1, 1, 1, 1, 1, 0, 0, 1, 0]$, and $d_3 = [0, 0, 0, 0, 1, 1, 1, 1, 0, 1]$. Borj et al. [51] used TC as one of the feature vectors for analyzing the sexual predatory conversations, and the results showed that it could detect online grooming conversations with good performance.
- **Term Frequency (TF)**: The fraction of the words' appearances is the entry value for each word in the text representation. In other words, it is a normalized version of the TC values.

$$TF(t) = \frac{n_t}{n_d}, \tag{1}$$

where n_t is the number of times word t appears in a document d, and n_d is the total number of terms in the document d. The feature vector of d_1 of the example will be:
$d_1 = [\frac{1}{10}, \frac{1}{10}, \frac{1}{10}, 0, \frac{1}{10}, \frac{1}{10}, \frac{2}{10}, \frac{1}{10}, \frac{1}{10}, \frac{1}{10}]$,
$d_2 = [0, 0, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, 0, 0, \frac{1}{6}, 0]$,
and $d_3 = [0, 0, 0, 0, \frac{1}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5}, 0, \frac{1}{5}]$.
Many documents or chat conversations contain many common words, while some are repeated more in some discussions. For instance, a higher frequency of words that express compliments may indicate online grooming occurring in a predatory conversation. Accounting for the frequency of a word in online chat logs can therefore help detect grooming conversations [51,52,65].
- **Term Frequency–Inverse Document Frequency (TF-IDF)**: If a word appears in many documents, it might not provide enough information for discriminating between different types of documents [91]. TF-IDF can provide feature vectors with vital information as it primarily considers the critical terms in each document. Therefore, it gives a lower value to the words in many documents and a higher value to the discriminative words seen in particular chat conversations. In other words, discriminatory terms have more power to distinguish the documents from each other, and TF-IDF applies this to enhance the feature vector [91].
TF-IDF computes the feature values based on the term frequency and the inverse of the document frequency [91]. Eq. (2) shows how TF-IDF is computed:

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right) \tag{2}$$

Here, the frequency of word i in document j is represented by $tf_{i,j}$. Furthermore, N is the total number of documents, and df_i represents the number of documents containing word i, and finally, $w_{i,j}$ is the feature value of word i when representing document j in the TF-IDF vector representation. For instance, the TF-IDF value for the term *live* in document d_1, will be:

$$w_{live,d_1} = 0.1 \times \log\left(\frac{3}{1}\right) = 0.158 \tag{3}$$

The term *live* only appears in document 1, while for example the term *parents*, that appears in 2 documents, would get a value of $w_{parents,d_1} = 0.058$. The term *at*, appearing in all 3 documents, would get a value of $w_{at,d_1} = 0$.
It should be noted that TF-IDF is the most common representation technique used for sexual predatory detection problems among the BoW techniques as it can distinguish the most discriminative words for representing a predatory conversation [49,51–53,63–65,69–73].
The description so far assumed that the dictionary on which the BoW features are based contained single words, while also pairs of consecutive words (called bigrams) are used in many papers. In some cases, using bigrams improved performance in detecting sexual predators [66,67,92]. Although Pendar [23] used trigram features (a combination of 3 consecutive words in a text), the use of unigram and TF-IDF feature vectors has shown a better performance.

- **Word Embedding** is a distributed representation of a text. While the statistical analysis of word occurrence is one of the primary methods for text analysis, the approaches mentioned above do not capture the meanings of the words. In addition, chat texts are short, and the bag of words features have sparsity problems. Distributed representations of word vectors create the word vector structures based on word analogies, concerning their several dimensions of difference. For example, word2vec [93] and GloVe [94] feature vectors provide the most used distributed representations for text analysis with dimensions of the feature vectors between 100 and 300. Word2Vec and Glove are described as follows:
  - **Word2vec**: is a distributed representation of word vectors. There are two techniques to compute the Word2vec features: the continuous Bag of word model and skip-gram model [93]. Continuous Bag of word (CBoW) is the simple extension of a bigram model. Fig. 3 is an example of the CBoW method where given the four surrounding words: 'you', 'have', 'close', and 'friends', it is desired to predict 'any' as the middle word for this context.
  The skip-gram model gets one word as an input and predicts the context words [93]. For instance, suppose we have only one word w_t that we desire to predict given context words w_{t-1}, w_{t-2}, w_{t+1} and w_{t+2}. It can be said that skip-gram is the opposite of CBoW, where the target word is the input while the context words are the outputs [93].

– **GloVe:** To give a short description of the aspects of the GloVe method, assume that matrix X contains the word-word co-occurrence counts where X_{ij} represents the number of times word j occurs in the context of the word i. Now

$$X_i = \sum_k X_{ik} \tag{4}$$

is the number of times any word appears in the context of word i, and hence,

$$P_{ij} = P(j|i) = \frac{X_{ij}}{Xi} \tag{5}$$

is the probability that the word j occurs in the text of the word i. For word k that is related to word i, but not to word j, the ratio P_{ik}/P_{jk} is significant, and similarly, for word k related to word j but not related to word i the ratio P_{ik}/P_{jk} is small. Therefore, it can be noted that the ratio can distinguish relevant words from irrelevant words better than the raw probability [94]. So, the model for GloVe is extracted from the general model below (more details can be found in [94]):

$$F(\omega_i, \omega_j, \omega_k) = \frac{P_{ik}}{P_{jk}} \tag{6}$$

The distributed representation of word vectors such as GloVe and Word2vec can distinguish the meaning of the words used in different contexts. However, the BoW feature vectors do not consider the analogies and changes because of a word's different meanings and locations in the text. For instance, combining the words 'dog' and 'toy' can result in different word vectors applying distributed representation of word vectors. At the same time, the BoW models give the same feature vector regardless of different meanings. Some works have used Word2vec, and GloVe feature vectors for detecting sexual predatory conversations [51,64,68]. There exist other word embedding systems like BeRT [95], ELMo [96] and fastText [97], but these have been less used in cyber grooming detection so far.

- **Affective features & LIWC:** : Some child offenders display feigned emotion and affection to make the impression that they are in love with the minor victim [98]. Tightening the trust link by showing false emotion is a technique that some predators perform to get the minor victim under control for further harmful actions [98]. Capturing the psychological characteristics by the words can reveal the affective features of a conversation. Linguistic Inquiry and Word Count (LIWC) [99] provides psycholinguistic profiles for the conversations revealing the emotional and psychological aspects of the data where it considers the level to which groups use different categories of words. LIWC features capture the psycholinguistic characteristics of the documents, including the affective characteristics displayed by a child groomer. It analyzes textual documents and provides various personal interest categories (e.g., love, emotion, work, home, leisure), psychological categories, and punctuation groups. Parpar et al. [70] have categorized 80 types of LIWC features that were used for predatory detection in chatrooms.
- **Chat-based features** capture the authors' activity in online conversations, such as the ratio of initiating the topics of conversation by the user, the percentage of written lines by a user, and the time spent online. Online predators mostly initiate the conversation topics to gain enough information about the victim and assess the risk. Their primary way to gather information is by asking many questions [49]. Chat-based features capture all these predators' actions, such as the percentage of a conversation started by a user. It has also been shown that online predators are emotionally unstable and prone to lose their temper and be anxious [92]. The chat-based characteristics determine the type of conversation, for example, if it is negative or anxious. Parapar et al. [70] pointed out that the activity of the author and the time of the chats (e.g., if the chats happened late at night) could also be used as a feature for detecting predatory chats.
- **Keystroke Dynamics (KD) features**: Keystroke Dynamics is a behavioral biometric that can authenticate or identify users based on how they type on a keyboard. In KD, one can only measure when a key is pressed down and released, giving (together with the key code of the key that is pressed) two raw timing features per key used from which several features can be extracted. First, per typed key, one can calculate the time the key was held down by looking at the difference between the time the key was pressed and rereleased. This feature is called duration and is sometimes referred to as the hold time in literature. Second, one can calculate the latency between pairs of keys, i.e., the time elapsed between releasing the first key and pressing the following key. Latency is sometimes also referred to as flight time. There are four variations of latency, and the one described above is also referred to as the RP-latency [100,101]. Alternatively, one can look at the time between pressing the first and next key (PP-latency) or the time between the release of these two keys (called RR-latency). Finally, PR-latency is the elapsed time between pressing the first key and releasing the following key (Please refer to [100,101] for an in-depth introduction to KD). Besides its natural use for authentication, KD can also be used to determine the authors' emotional state [102–105], and emotional states can again help to detect the predators as they are not emotionally stable [92]. Borj and Bours [50] used KD feature vectors to detect authors that lied about their demographic information, such as age and gender, in chat conversations.

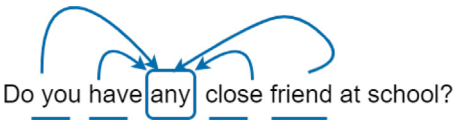

Fig. 3. CBoW example.

Fig. 4 displays a summary of the proposed data and features.

3.2.1. Constraints of the feature vectors

Some constraints challenge discriminative and stable feature extractions for chat messages. For example, the chat messages do not follow the standard pattern for writing, and they contain many slang words. The BoW techniques cover all the information, including non-sense words and slang, by creating a large sparse matrix. The sparse feature matrix can impact the performance of many machine learning methods for detecting sexual predators. Cardei and Rebedea [62] extracted different types of features such as question ratio, underage expression ratio, above age existence, and slang words ratio and combined them with the BoW feature sets. As mentioned above, the sparse feature vector can negatively impact the grooming detection performance. Therefore, applying a feature selection method helped improve the performance, considering the most discriminative feature space for sexual predatory detection from the chat logs. There are several methods for feature selection, including Mutual Information (MI),

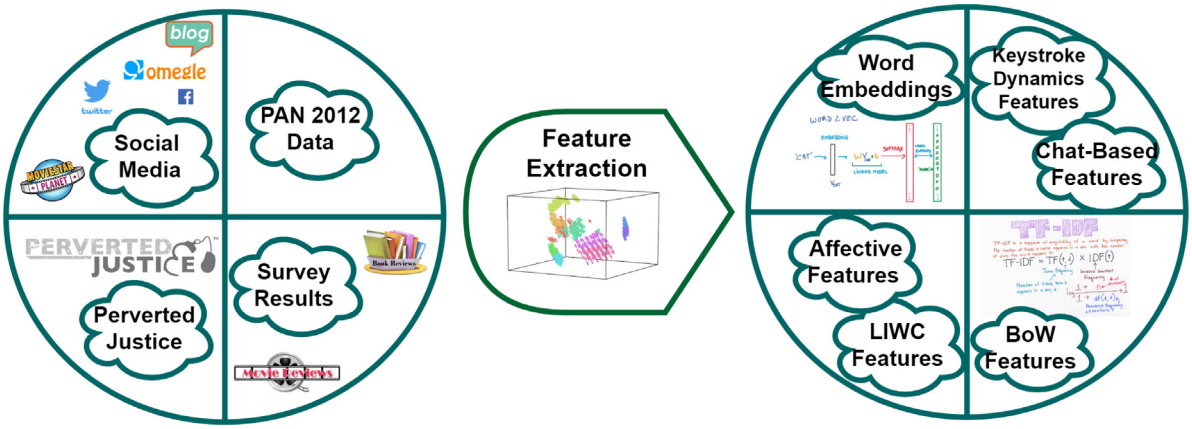

Fig. 4. Data and features in sexual predatory detection.

chi-squared, or frequency-based feature selection. For instance, Cardei and Rebedea [62] used MI to consider the power of presence/absence of a term in making the right classification decision for each class by using SVM [62].

Zuo et al. [72] proposed a fuzzy-rough feature selection approach that captures the uncertainty resulting from the lack of rigid boundaries in the dataset's various classes by demonstrating the concept's lower and upper edges. They first performed feature extraction based on BoW and TF-IDF approaches. Zuo et al. [72] then reduced the feature space by applying the fuzzy-rough method to select the most discriminative features and speed up the process. Finally, using the reduced feature sets, the authors [72] experimented with the online grooming detection on the PAN2013 dataset by using four classifiers, including Gaussian Naive Bayes (GNB), Random Forest (RF), AdaBoost (AB), and Logistic Regression (LR) [72]. The BoW methods do not cover the relationship between the words and make the same feature space for words with different contexts. Thus, the same feature sets of the various conversations decrease the performance for detecting the predators.

The word embedding methods such as GloVe and Word2vec cover the relationship between words and the semantic information in chat conversations. However, the pre-trained word embeddings trained using general documents such as Google News data are unsuitable for sexual predatory detection problems because chat logs contain many out-of-vocabulary and slang words [64]. For instance, the chat sentence 'r ur parents der' contains some words such as 'der' and 'ur' that are not defined in the dictionary of words [64].

3.3. Performance metrics

Cyber grooming detection is posed as a two-class classification problem. The predatory conversations class is usually considered a positive class and the non-predatory conversations as negative. The True Positive (TP) samples are considered all the samples in the positive class classified as positive samples by the classification algorithm. Similarly, True Negative (TN) are negative samples classified as negatives by the classification algorithm. A positive sample classified as a negative sample is considered a False Negative (FN) classification, and a False Positive (FP) is defined as a negative sample classified as positive.

Many works have applied the standard evaluation metrics for analyzing the performances of their methods based on TP, TN, FN, and FP, such as Accuracy, Precision, Recall, and F-score. We give a brief definition for each metric below:

- **Accuracy** is the fraction of correct predicted labels for all samples, i.e.

$$Acc = \frac{TP + TN}{TP + FP + TN + FN} \tag{7}$$

- **Precision** is the ratio of the detected relevant samples (TP, i.e., correctly identified sexual predators or predatory conversations) and all detected samples (contains both TP and FP samples), i.e.,

$$P = \frac{TP}{TP + FP} \tag{8}$$

  Precision indicates the probability that a sample classified as predatory is, in fact, predatory.
- **Recall** is the fraction between detected relevant samples (TP) and all the actual relevant samples, i.e.

$$R = \frac{TP}{TP + FN} \tag{9}$$

  Recall indicates the probability that a predatory sample will be detected as such by the classification algorithm,
- **F-score:** is the weighted harmonic mean between precision and recall and is defined as

$$F_\beta = (1 + \beta^2) \cdot \frac{P \cdot R}{\beta^2 \cdot P + R} \tag{10}$$

  where β is a positive real factor and can be varied to put more weight on either precision or recall.

The ten best results of the PAN2012 competition for identifying predators are given in Table 2. The ranking for the PAN2012 competition was based on the $F_{0.5}$ score, so we kept that ranking here too. Further, Table 3 presents various works and their best performance for the problem based on the metrics mentioned above.

3.3.1. Constraints of performance metrics

The predictive scores resulting from different classification models for finding sexual predators have an essential role in many areas, particularly the cases related to law enforcement decisions (for example, in courtrooms). Therefore, the fairness of the machine learning methods should be analyzed and considered carefully to avoid any mistake that can harm people's lives. For instance, the data for grooming detection is highly imbalanced, negatively impacting the accuracy metric's relevance [86]. The amount of positive samples (predatory conversation) is much lower than the number of negative samples. In the PAN2012 dataset, the positive samples amounted to just 4%, and in Latapy

Table 2
The best results reported on PAN2012 dataset, ranked based on $F_{0.5}$ scores.

Participants	Precision	Recall	$F_{0.5}$	F_1	F_2
Villatoro et al. [71]	0.98	0.79	0.93	0.87	0.82
Snider[a]	0.98	0.72	0.92	0.83	0.76
Parapar et al. [106]	0.94	0.67	0.87	0.78	0.71
Morris & Hirst [67]	0.97	0.61	0.87	0.75	0.66
Eriksson & Karlgren [107]	0.86	0.89	0.86	0.87	0.89
Peersman et al. [44]	0.89	0.59	0.81	0.72	0.64
Grozea & Popescu[a]	0.76	0.64	0.73	0.70	0.66
Sitarz[a]	0.73	0.63	0.71	0.67	0.64
Vartapetiance & Gillam[a]	0.62	0.39	0.55	0.48	0.42
Kontostathis et al. [a]	0.36	0.67	0.39	0.47	0.57

Note
[a]Indicates no corresponding article available.

Table 3
The accuracy obtained by various state-of-art works in detecting online grooming.

Ref	Year	Accuracy	$F_{0.1}$	$F_{0.5}$
Pendar et al. [23]	2007	–	0.94	–
Parapar et al. [106]	2012	–	0.84	–
Villatoro et al. [71]	2012	0.92	0.87	0.93
Bogdanova et al. [92]	2012	0.97	–	–
Cheong et al. [74]	2015	0.93	0.78	0.86
Ashcroft et al. [58]	2015	0.99	–	–
Ebrahimi et al. [63]	2016	0.99	0.77	–
Ebrahimi et al. [64]	2016	–	0.80	–
Cardei et al. [62]	2017	–	–	0.95
Escalante et al. [65]	2017	–	0.94	–
Zuo et al. [72]	2018	0.73	–	–
Zuo et al. [73]	2019	0.76	–	–
Misra et al. [66]	2019	–	0.58	–
Bours et al. [53]	2019	–	0.94	0.97
Borj & Bours [49]	2019	0.98	0.86	–
Muñoz et al. [68]	2020	0.88	0.42	–
Fauzi & Bours [52]	2020	0.95	0.90	0.93
Borj et al. [51]	2020	0.99	0.96	0.98
Ngejane et al. [69]	2021	0.98	0.70	–

et al. [85], for peer-to-peer networks, it amounted to only 0.25%. If a method provides high accuracy, it does not always mean that it will be efficient as the size of the negative class is enormous and can cover the incompetence of the method for detecting positive cases. For example, if a dataset would contain 99% negative samples, then simply classifying every test sample as negative would already result in a 99% accuracy. Therefore, the approaches should be cautious and not select a technique based on accuracy alone. It is desirable to integrate more human-centered models for developing and evaluating grooming detection techniques to avoid any lifetime negative impact on people's lives [27].

It should also be considered that the F_1-score gives the same weight to the recall and precision, while it can be problematic for cyber-grooming detection. Inches and Crestani [41] observed that it was important not to overload law enforcement with investigating many false-positive cases. A false-positive sample would mean that the law enforcement agency would investigate a falsely accused person, taking time but not leading to any actionable results. Many false-positive cases would mean that less time could be spent on actual positive cases. Considering this, Inches and Crestani [41] used F_β with $\beta = 0.5$ for ranking the performance results of the PAN2012 competition. Other researchers have followed this suggestion [71]. Some papers [52,53] used a β value higher than one that would emphasize recall and aim for a lower number of false negative classifications. A lower false negative value would lead to fewer undetected positive samples [49,51–53].

3.4. Online grooming detection techniques

Researchers have conducted online grooming detection in different ways. While some considered the stage direction of the chat logs, others focused on identifying the predators and detecting suspicious messages. It also has been shown that looking into the demographic attributes of the users facilitates the detection task for finding adults soliciting minors. The remainder of this section will detail the different techniques of online grooming detection, including grooming stage detection, predatory conversation detection, predatory identification, and authorship profiling.

3.4.1. Online grooming stage detection

Grooming can consist of different stages, whether online or in real life. Researchers have considered different stages in the grooming process ranging from 3 to 6 stages. This section will describe various research works and discuss the stages identified in the grooming process. We also describe how the various stages are detected within a conversation.

In many cases, victim selection is considered the first stage in the grooming process [30,108]. Researchers believe that selecting a victim depends on many factors such as interest/attractiveness, ease of access, or perceived weak points and vulnerabilities of the child. Some research works [30,108] showed that the victim's physical characteristics play a prominent role in being targeted (42%). Predators mainly target children with vulnerable family conditions, such as living with single parents, custodial cases, and drug or mental problems [30]. The predators can also threaten children with psychological vulnerabilities. Psychological issues increase the chances of isolating the victim from others while the victim suffers from some problems such as low self-esteem, low confidence, insecurity, neediness, or naivety [30,108].

After finding the victim, the offender attempts to develop a trusted friendship. Olson et al. [30] have described this phase of grooming as:

> Deceptive trust development is the ability of a child molester to cultivate relationships with potential victims and possibly their families, intended to benefit the perpetrator's sexual interests.

The deceptive trust development has the grooming process's primary role where predators obtain much information about the victim by being helpful, showing attention, and sharing secrets from previous relationships [31]. Thus, the child/victim gets the impression of having a confidential and exciting relationship that should be kept secret. The main goal of the predator in this phase is to control and manipulate the victim for further actions [30,31].

Predators try to minimize the risk of danger by asking many questions, such as about other users of the victim's computer and if the parents have the passwords to access the conversations [32]. Predators also make the victims aware that their relationship is not appropriate to avoid jail in legal cases [32].

Generally, it can be said that sexual predators use different language themes in online conversations [109]. Each theme displays distinctive cognition used by online sexual offenders, such as discourse content, online solicitation, and fixated discourse [109,110]. Notably, the discourse content demonstrates a pattern that persuades the grooming without being obvious to a child as there is no sexual topic or explicit harassment. Instead, the predators display their emotions and behavior in different patterns while minimizing the risk of being detected and preparing the victim mentally for further abuse [109,110]. To understand trust in online predatory conversations, researchers found several patterns of compliments behavior that show how by giving compliments, the predators frame the grooming process and gain the victims' trust [110].

Table 4
Grooming stages and their characteristics.

Stage	Grooming characteristics
Friendship forming	Questions about profile exchange information: (1) Exchanging email address; (2) Asking the age/gender/location/name; (3) personal information/details about family.
Trust development	Conversations about favorite hobby and activity; Giving compliment; Pictures; Building mutual trust; Showing feelings like anger, love, etc.
Risk assessment	Conversations about the relationship with parents and friends; Acknowledging wrong doing; Questions to determine if the child is alone; Assessing the risk of conversations.
Exclusivity	Expressing feeling of love and exclusiveness; Other way of communication.
Sexual	Conversations about body and intimate parts; Sexual content; Sexually oriented compliments; Giving body description; Exchanging sexual pictures; Fantasy control and aggression.
Conclusion	Arrange further contact and meeting

Since online grooming has different stages, many researchers [28,33,40,42,43] tried to detect each stage in suspicious chat conversations using machine learning methods and grooming characteristics. Previous works have used various types of grooming characteristics that display the mentioned purpose of the predators. The overall overview of the grooming characteristics that different papers [28,33,40,42,43] have used is presented in Table 4.

One of the first works for identifying the grooming stages was done by Kontostathis [33]. His tool annotated each message of the conversations based on the theory from Olsen et al. [30]. According to Olsen's theory, predatory conversation has three subsets: grooming, isolation, and approach. The main focus of a predator is developing a deceptive trust to catch the victim. According to the theory, the chat conversation starts with gaining access to the victim, followed by building a deceptive relationship with the minor. The last stage is initiating and keeping sexually abusive contact [30]. To test the three stages hypothesis, Kontostathis et al. [56] downloaded 288 chat conversations from the Perverted Justice (PJ) website (http://www.perverted-justice.com/). The PJ website published only chat conversations of convicted predators. Kontostathis et al. [56] also developed a dictionary that contains the words and phrases annotated based on the predatory phase to cluster each stage in a chat using k-means clustering by [33]. Following the same idea, they developed a tool to annotate conversations into specific grooming stages. They also used phrase matching and rule-based techniques for classifying sentences into the various grooming stages [56].

Michalopoulos and Mavridis [57] have also considered online grooming as a three-phase procedure: Gaining access, Deceptive relationship, and Sexual affair. Term Frequency–Inverse Document Frequency (TF-IDF) features (for details see Section 3.2) were extracted for each phase and various machine learning models, such as Naive Bayes (NB), Support Vector Machine (SVM), Maximum Entropy (ME), Expectation–Maximization (EM), and EMSIMPLE [111] were trained. Their method computes the probabilities that a chat conversation belongs to each grooming class pattern. Using a linear combination of the probabilities, the method decides whether the chat conversation is a grooming conversation or not [57].

Escalante et al. [61] covered the different grooming stages using chain-based classifiers to divide the conversation into three distinct segments. They used a local classifier for each segment and combined their results with various strategies. Each local classifier is related to each stage in online grooming, i.e., gaining access, deceptive relationship, and sexual stage. It was supposed that the vocabulary usage differs in each conversation segment, and combining each segment's result could provide a good clue for making the correct decision if a conversation is predatory or not [61].

Online grooming was considered a four-stage process in [43]. The stages were trust level, grooming, seeking a physical approach, and others. However, one should realize that the online grooming stages are not necessarily sequential. Predators might return to earlier stages to decrease the chance of losing the victims' trust and assess their risks. They might even skip stages in the grooming process. Each stage of grooming can also be identified using feature sets such as Bag of Words (BoW), syntactical, i.e., Part of Speech (POS), sentiment, content (Complexity, readability, length), psycho-linguistic (LIWC dimensions) [112], and discourse patterns.

Online grooming was also considered as a six-stage process where the stages are: (1) friendship forming, (2) relationship forming, (3) risk assessment, (4) exclusivity, (5) sexual, and (6) conclusion [28,40]. Each stage contains some grooming characteristics presented in Table 4. For instance, in the friendship-forming stage, the features such as exchanging personal information like name, age, and location are considered [40].

A predator measures the danger and threat level by asking if the child is alone or if nobody else reads the conversation, and this type of message belongs to the risk assessment stage. Black et al. [39] investigated the similarities and differences in various grooming procedures by applying Linguistic Inquiry and Word Count (LIWC) and content analysis for analyzing different phases of online grooming. They discovered that assessing risk and potential for victimization can be detected by analyzing 40% of an online conversation. However, the first 20% of that conversation can already indicate signs of grooming [39].

During the exclusive stage, predators build a confidential relationship with the victim and try to get his/her trust. One should consider that the predator might not talk about sexual topics with the victim during this stage to avoid losing trust so that he can finally approach the victim in person and execute some harmful action [49]. The time a predator spends in each stage varies, depending on their personality and condition. The exclusive stage is where the predator has assessed the risk, and the theme of the conversation changes [40,42]. Once a predator believes that the victim might be emotionally and mentally ready, the conversation goes over to the sexual stage, where sexual topics can be brought up by asking questions such as *'Did you ever touch yourself?'* [40,42].

The combinations of the words that indicate the various grooming stages are used as feature vectors to classify grooming stages in some papers [40,42]. Gunawan [40] tested the six stages theory applying SVM and K-Nearest Neighbors (KNN) to identify grooming conversations. In addition to the characteristics mentioned above, some other features, such as asking about the victim's relationship with the parents, can indicate online grooming [42]. Using words in biology, body, or feeling categories, as well as arranging further contact and meeting, are very discriminative features in detecting online grooming stages [42]. Furthermore, the features that demonstrate the *'other way to contact'*, *'reframing'*, or *'asking hot pictures'* are also valuable clues for detecting online grooming. Pranoto et al. [42] detected online grooming by training a logistic mathematical model applying all features mentioned above.

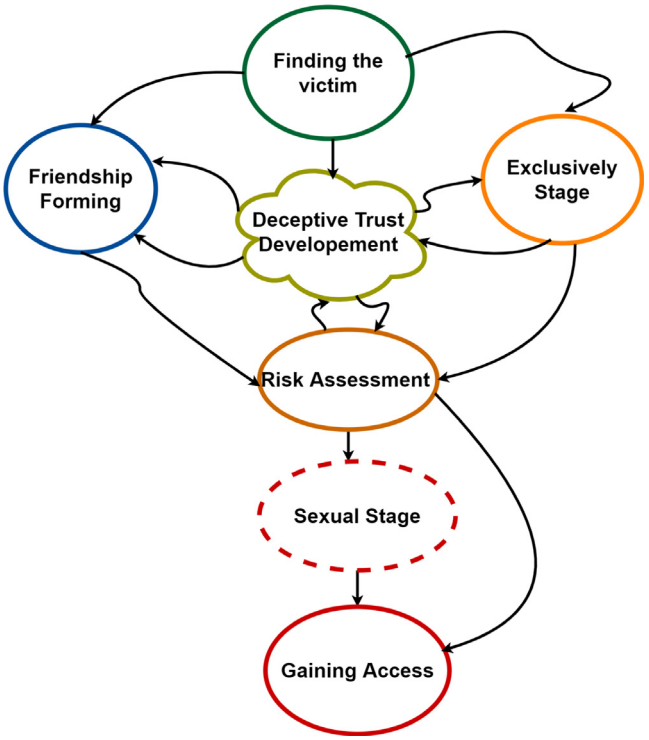

Fig. 5. Grooming stages.

Gupta et al. [28] tested if the sexual stage is the main stage in predatory conversations by annotating 75 predatory conversations according to the six grooming stages and extracting features using LIWC. They found that contrary to the hypothesis, the friendship stage played a central role in online grooming, contributing to 40% of the messages in a predatory conversation [28]. This could be explained by the fact that a predator tries to build a deceptive trust friendship, which is a tedious task, and consequently, most of the messages belong to the friendship-building stage [28].

Fig. 5 summarizes and shows the different stages of online grooming, and Table 5 gives a short overview of the most relevant papers.

3.4.2. Predatory conversation detection

One of the main areas for grooming detection is predatory conversation detection. This section discusses how various earlier works have addressed the problem of predatory conversation detection in online platforms. One of the first attempts to detect grooming conversations was made by Villatoro et al. [71] at the PAN2012 competition. Their method applied a two-stage classification scheme where the first stage distinguished the predatory conversations from the non-predatory ones. The conversations classified as predatory in the first stage were further used in the second stage to identify the predator and victim in suspicious conversations [71]. Fauzi and Bours [52] performed the same experiment with a performance improvement, applying a new method called soft voting. The proposed soft voting technique calculates the probability that a conversation is predatory based on the average probability from 3 selected classifiers. If the average probability is over 0.5, then the conversation is classified as predatory, otherwise, as non-predatory [52].

Pre-processing has a vital role in changing the performance of text mining. The chat logs have many slang words and non-sense terms that do not provide valuable information for the model and can negatively influence the result. Borj et al. [51] investigated the impact of preprocessing on data in the performance of predatory conversation detection in the two-stages model. The preprocessing can be done by applying tokenization, stop-word removal, removing the words that are longer than 20 letters, as well as sentences with less than seven words as they do not provide any information to be classified in any class [51,62]. The authors showed how preprocessing could increase the performance of the models for sexual predatory conversation detection [51].

As mentioned earlier, the highly imbalanced dataset is a big challenge in grooming detection. There are several techniques for handling the imbalanced data for learning algorithms, such as Sample Method, Cost-Sensitive Method, Kernel-Based Method, and Active Learning Method [113]. Cardei and Rebedea [62] tested several techniques to cope with the imbalanced data in grooming detection problem: cost-sensitive technique, sampling techniques such as BalanceCascade [114], and clustering-based method using k-means [115]. The cost-sensitive technique performed best when testing on the PAN2012 dataset for predator detection, where a cost matrix defines the penalty for misclassifying a sample. Zuo et al. [73] proposed a new method for handling the imbalanced data issue in sexual predator detection. They combined an adaptive fuzzy inference-based activation function with the artificial neural networks (ANNs) and extracted BoW and TF-IDF as feature sets to classify the data sets.

Some early research works are on the early identification of gestures/actions in texts with as little information as possible [53,65,116]. Most research uses complete conversations to detect online grooming. However, it is better to detect a predatory conversation as early as possible to reduce the risk of actual harm to the victims. For instance, Dulac-Arnold et al. [116] used a Markov Decision Process to classify documents into topics while processing separate sentences. The method has two main components: it can include either the following sentence or make a final decision about the topic of the text. Escalante et al. [65] provided a model to detect sexual predator threats and aggressive acts in the early stages. Their proposed method uses profile and sub-profile representations and the document vector space representation for the investigation of threats. To detect online grooming in its early stages, Bours and Kulsrud [53] proposed different methods, including message-based, author-based, and conversation-based. For the message-based model, they used LR and Ridge classifiers on the PAN2012 data and found that some words such as “sweetie”, “hun”, “mwah”, and “lil slut” indicated predatory conversations strongly. In the author-based technique, they considered each author’s messages in an entire conversation at once. For the conversation-based model, they first classify conversations as predatory or non-predatory based on the conversation and then determine the predator based on only the messages from each chatter. Various models such as LR, Ridge, NB, SVM, and NN were trained by BoW and TF-IDF features, while the best performance was obtained using TF-IDF features and the NB classifier [53].

Predators often show their emotions, such as fear and anger, that reflect their frustration of being in danger of getting caught or not receiving what they want [59]. A set of high-level features indicating emotional states such as emotional instability, inferiority, loneliness, low self-esteem, and emotional immaturity are affective attributes that highlight a predatory conversation [59]. For instance, the percentage of positive words, the percentage of anger or sadness words, and the percentage of relationship words can capture the emotional characteristics of a writer. Data can be labeled into different emotions such as anger, disgust, fear, joy, sadness, and surprise [117]. Java WordNet Similarity library (a Java implementation of Perl Wordnet) [118] and Resnik’s similarity measure [119] are used for extracting affective features [59]. Applying the affective feature sets showed that it is challenging

Table 5
Online grooming stage detection by different papers.

Author	Number of stages	Data	Features	ML technique
Kontostathis [33]	3	www.perverted-justice.com	Grooming characteristics	K-mean Decision Tree
Kontostathis et al. [56]	3	www.perverted-justice.com	Grooming characteristics	Phrase matching Rule-based techniques
Escalante et al. [61]	3	PAN2012	BoW	Ring-based Classifier Decision Tree
Michalopoulos et al. [57]	3	www.perverted-justice.com	TF-IDF	Naive Bayes SVM Maximum Entropy EM and EMSIMPLE Decision Tree
Cano et al. [43]	4	www.perverted-justice.com	BoW, POS, sentiment, complexity, readability, length, psycho-linguistic (LIWC dimensions)	Decision Tree SVM
Gunawan et al. [40]	6	www.perverted-justice.com www.literotika.com	Grooming characteristics	SVM, K-NN Decision Tree
Pranoto et al. [42]	6	www.perverted-justice.com www.literotika.com	Grooming characteristics Decision Tree	Binary Logistic Regression
Gupta et al. [28]	6	www.perverted-justice.com	LIWC to create psycho-linguistic profile based on grooming characteristics	Logistic Regression Decision Tree

to distinguish a child-exploiting conversation from an adult–adult conversation with a sexual topic rather than detecting child grooming conversation from the general chat messages [59].

Some chat conversations may implicitly show signs of predatory behavior without containing any direct terms that explain if the aim of the conversation is grooming. Semantic analysis can cover this issue by investigating pseudo-intelligent information in chat data without the need for human intervention to characterize implicit anomalous conversations. In this case, a distributed representation of the context captures the semantic representation of the data. Following this idea, Munoz et al. [68] extracted the Word2Vecfeature space and used Convolutional Neural Networks (CNN) for grooming detection, analyzing the chat conversations with an accuracy of 0.88. To capture the Word2Vecfeature set, they pre-processed the PAN2012 dataset by Tweet Tokenizer [68]. Then, they extracted the features using the Skip-gram model implemented in TensorFlow, where the feature set has a dimension of 128. They also used Noise Contrastive Estimation (NCEloss) for optimization.[4]

All the mentioned techniques for predatory conversation detection need large amounts of data that has both predatory and non-predatory conversations. Ebrahimi et al. [63] proposed an anomaly method that avoids gathering non-predatory conversations to have a practical model without analyzing the non-related conversations. The model is based on a semi-supervised one-class SVM and does not require non-predatory samples for training. Later, Ebrahimi et al. [64] used CNN for predatory chat detection. The CNN model gained a better performance compared to the semi-supervised anomaly model.

3.4.3. Predatory identification

Few earlier works mentioned above have also tried to identify the predators [51–53,71]. From a forensic's perspective, it is crucial to identify the predators for further actions. This section will discuss various methods [23,58,67,70,74,120] that have been used for predatory identification.

Pendar [23] proposed one of the first models for distinguishing online predators from victims, using SVM and k-Nearest Neighbors (KNN) machine learning methods. He [23] used the Perverted Justice website to collect the data and extracted n-gram features, including document frequency and character ratios. Similarly, Cheong et al. [74] tried to detect the predators in online game chat platforms, applying a combination of inherent features with the BoW representations. Morris [67] provided a method for predatory identification where the predator's language is learned simultaneously with the language used by the victims. An SVM model was trained by lexical features, such as n-grams, and behavioral features that capture the author's conversation flow pattern, such as turn-taking and message length [67].

From a psycholinguistic perspective, there is a relationship between word usage and personality, social conditions, and consequently emotional states [120]. Thus, Part of Speech (POS) tags (pronouns, auxiliary verbs, etc.) can reveal helpful information about the author of a text and his/her emotional state to show how honest or deceptive an author is. Parapar et al. [70] extracted various features exploring this concept to distinguish predator behavior from a victim's behavior. Their study reported that predators engage primarily in 1-on-1 conversations and less in conversations involving multiple persons. Other features such as time of chat were observed closer to midnight, and predators' linguistic profiles showed that they mostly use first-person pronouns [70]. The authors also showed that emotional expression could be a good indicator of the deceptive language of a predator. Their results indicated that the deceptive trust phase and language pattern include effective use of loving, time, and location words. They trained an SVM model for predatory identification based on three different feature spaces such as LIWC features (Psycholinguistic features), TF-IDF features, and chat-based features. Parapar et al. [70] also noted that the content-based features indicated characteristics such as how active, anxious, or intense an author was.

Another way to distinguish a predator from a victim is to determine a child from an adult based on the writing style. Ashcroft et al. [58] showed that it is possible to distinguish a child from an adult, although it could be more challenging in short text messages compared to books and reviews. The authors extracted the data from various resources for grooming

[4] The details of the implementation by Munoz et al. [68] can be found on the GitHub repository https://github.com/gisazae/Tensorflow-Examples/blob/master/IntegracionCorpus_checkpoint.ipynb.

Table 6
Summary of research works on sexual predatory conversation detection and sexual predatory identification.

Author	Objective of the research	Data	Features	ML technique
Ashcroft et al. [58]	Distinguish child from predator	Reviews on Amazon, Reviews on Spagetti Book Club, Blog-kid and blog-adults, http://perverted-justice.com	POS-tags	AdaBoost
Bogdanova et al. [59]	Predatory conversation detection	NPS chat corpus, Cybersex Logs, http://perverted-justice.com	Emotional characteristics, n-grams	SVM
Borj et al. [51]	Predatory conversation detection & predatory identification	PAN2012	BoW, TF, TF-IDF, GloVe	SVM, NB, RF
Bours et al. [53]	Predator detection	PAN2012	BoW, TF-IDF	LR, Ridge, NB, SVM, NN
Cardei et al. [62]	Predatory conversation detection & predator identification	PAN2012	BoW & behavioral characteristics	SVM
Cheong et al. [74]	Detecting predatory behavior in game chats	MovieStarPlanet	BoW, sentiment features, rule-breaking features, blacklist/alert list terms, behavioral characteristics	DT, MLP, KNN, SVM, LR, NB
Ebrahimi et al. [63]	Predatory conversation detection as an anomaly	PAN2012	TF-IDF	1-class SVM
Ebrahimi et al. [64]	Predatory conversation detection	PAN2012	GloVe, Word2Vec, BoW	SVM, CNN, NN
Escalante et al. [65]	Early detection of threats in social media	PAN2012, Kaggle, UANL	TF, TF-IDF, Profile Specific Representation (PSR), Subprofile Specific Representation (SSR)	NN, KNN, SVM,NBM
Fauzi et al. [52]	Predatory conversation detection & predatory identification	PAN2012	BoW, TF, TF-IDF	NB, SVM, NN, KNN, RF, Ensemble Model
Miah et al. [60]	Predatory conversation detection	http://www.fugly.com, http://chatdump.com, http://perverted-justice.com	Term-based features, Psychometric characteristics	NB, Regression
Misra et al. [66]	Authorship attribution of online predatory conversations	PAN2012	Character unigram and bigrams	CNN
Morris et al. [67]	Predatory identification	PAN2012	Lexical features, Behavioral features	SVM
Muñoz et al. [68]	Grooming detection	PAN2012	Word2Vec	CNN
Ngejane et al. [69]	Predatory conversation detection	PAN2012	TF-IDF, Embedding	LR, XGBoost, MLP, BiLSTM
Parapar et al. [70]	Predatory identification	PAN2012	TF-IDF, LIWC, chat-based & Content-based features	SVM
Pendar et al. [23]	Predatory identification	http://perverted-justice.com	n-grams	SVM, KNN
Villatoro-Tello et al. [71]	Predatory identification	PAN2012	BoW, TF-IDF	NN,SVM
Zuo et al. [72]	Grooming detection	PAN2013	BoW, TF-IDF	GNB, LR, AdaBoost, Fuzzy Interpolation
Zuo et al. [73]	Predatory conversation detection	PAN2013	BoW, TF-IDF	ANN

detection. Their data included the reviews of children's books by children between 7 and 15 years old, reviews on Amazon, blog posts from blogger.com, the chat between adults and children, and predatory conversations from the PJ website. In addition, linguistic features such as stop and function words, letters of the alphabet, punctuation, and numbers were extracted. They also considered grooming, sexual features, and the POS tags to gain more information about each document in conjunction with an Adaboost classifier [58].

Table 6 presents all the related works for grooming detection with a focus on predatory conversation detection and predator identification. Finally, Table 7 presents a taxonomy classification of the proposed approaches for grooming stage detection, predatory conversation detection, and predatory identification.

3.4.4. Author profiling

The enormous amount of data that law enforcement agencies have to investigate to find predators requires an automated system. Automated systems facilitate the detection task for finding adults soliciting minors. Instead of detecting predators or predatory conversations, one might also have a closer look at chatters directly. It is known that predators often use a fake identity while online searching for potential victims. For example, they might conceal the gender or age while making initial contact [48,50]. Thus, profiling the authors based on their writing style has been explored for detecting predators, forensics, security, and marketing. Author profiling classifies the authors based on various aspects, including age, gender, native language, or personality type (see Fig. 6).

It is logical to consider stylometry for author profiling, given that predator detection in online conversation is conducted using

Table 7
Categorization of works according to objectives in predator detection.

Objective		
Grooming stage detection	Predatory conversation detection	Sexual predatory identification
Cano et al. [43]	Bogdanova et al. [59]	Ashcroft et al. [58]
Egan et al. [109]	Borj et al. [51]	Borj et al. [51]
Escalante et al. [61]	Bours & Kulsrud [53]	Cardei & Rebedea [62]
Gunawan et al. [40]	Cardei & Rebedea [62]	Cheong et al. [74]
Gupta et al. [28]	Ebrahimi et al. [64]	Fauzi & Bours [52]
Kontostathis [33]	Ebrahimi et al. [63]	Misra et al. [66]
Kontostathis et al. [56]	Escalante et al. [65]	Morris [67]
Michalopoulos & Mavridis [57]	Fauzi& Bours [52]	Pendar [23]
Pranoto et al. [42]	Miah et al. [60]	Villatoro et al. [71]
	Misra et al. [66]	
	Munoz et al. [68]	
	Zuo et al. [72]	
	Zuo et al. [73]	

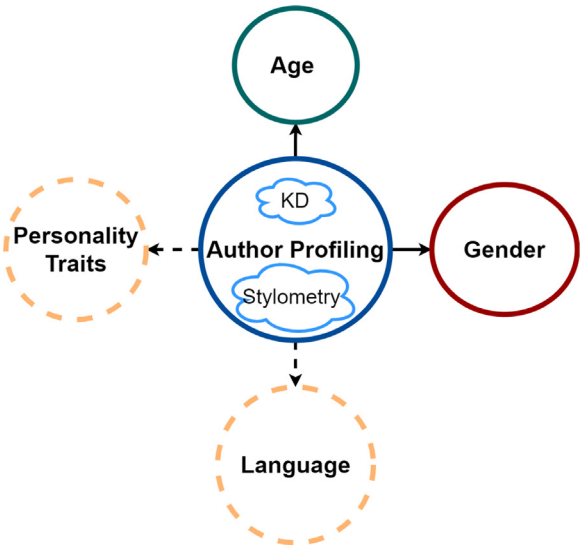

Fig. 6. Author profiling.

chat logs [81,82]. Alternatively, one can also consider the typing rhythm of an author for author profiling. Typing rhythm, or Keystroke Dynamics (KD), has also been used in previous research to detect the age and gender of an author [121–126]. Section 3.4.4 displays the overall view of authorship profiling. The primary focus in grooming detection is age and gender, as predators might conceal their actual age and gender to trap victims.

- **Author Profiling based on Keystroke Dynamic:** Few earlier works have investigated the feasibility of identifying the gender of an author by measuring KD and applying various machine learning models [121–123]. Giot and Rosenberger [121] presented a method for gender recognition using KD with more than 91% accuracy on GREYC keystroke database. The authors extracted five different features from each sample, including duration, RP-, PP-, and RR-latencies, and a combination of these four timing values in conjunction with an SVM classifier to identify gender. Fairhurst and Da Costa-Abreu [122] also identified the gender of users in a social network environment using the GREY dataset performing KNN, Decision Tree (DT), NB, and fusion techniques. The fusion techniques were Dynamic Classifier Selection based on Local Accuracy class (DCS-LA), Majority Voting, and Sum rule-based [122]. Idrus et al. [123] used a set of 5 short texts (17–24 characters long) with Majority Voting and detected the author's gender with an accuracy of 92.1%. Besides gender detection, a few earlier works have also investigated the possibility of detecting the age of an author by applying keystroke dynamics [50,124–126]. Pentel [125] proposed an age and gender detection model using KD information on various machine learning methods such as SVM, KNN, and Random Forest (RF). Tsimperidis et al. [124] created a database with a free text keylogger called 'IRecU' and used it to predict the age of the users. It was shown that the accuracy performance improved by decreasing the number of age groups. In another work by Tsimperidis et al. [126], the average values of the keystroke durations and PP-latency features were considered for age detection.
- **Stylometry based Authorship Profiling:**
As the first attempt to determine the authors' age and gender by stylometry, one can consider the works as early as 2002 and 2003 by Koppel et al. [81] and Argamon et al. [82] respectively. Koppel et al. [81] detected the gender of the authors on the BNC[5] corpus by extracting simple lexical and syntactic features with an accuracy of 80%. Argamon [82] showed a difference between various genders in writing style based on the BNC corpus. Notably, the usage of pronouns and some noun modifiers vary between genders, where women use many pronouns, and men use more noun specifiers. In the remainder of this section, we summarize author profiling by early research works focusing on the applied data and used features.

  – **Stylometry Data for Author Profiling**: Earlier works have investigated the feasibility of automatically predicting age and gender on short texts such as chat logs and social network platforms [44–47]. Peersman et al. [44] provided a method to distinguish between adults and adolescents. They extracted the data from a Belgian online social networking platform called Netlog. The size of the data has a significant impact on the performance of detecting age and gender. The accuracy of author profiling in short blog segments is lower than on the lengthy messages [45]. Nguyen et al. [46] tried to distinguish Twitter users' age where the sentences were short (on average less than ten words). Their main goal was to investigate how age can influence language usage in the dataset extracted from 3000 Dutch Twitter users. Different age categories were defined, and the age was predicted as a continuous variable applying content features, and stylistic features [46]. They also considered the impact of gender on the performance of age detection, where age and gender are assumed as inter-dependent

5 http://www.natcorp.ox.ac.uk/.

variables, and the obtained correlations were around 0.74 [46]. PAN[6] held a series of competitions for author profiling over recent years. In PAN2013 [47], they covered multilingual platforms where they collected English and Spanish data. The data contains various themes to provide a realistic setting. The data was collected from Netlog[7] and blog posts,[8] and it was labeled by users' demographic information. To get more reliable data, they selected the authors with blogs with at least 1000 words; their data is gender-balanced with various age groups. In PAN2015 [127], the main goal was to investigate the authors' various demographic information, such as age, gender, language variety, and personality traits. The data was collected from Twitter in different languages such as English, Arabic, Portuguese, and Spanish. Data for PAN2016 [79] and 2017 [80] was also extracted from Twitter. The purpose of PAN2018 [128] was to detect the gender by texts and images. The data was based on the PAN2017 [80] corpus extended by images shared on the Twitter timelines. This dataset was gender-balanced, and each author had at least 100 tweets and ten images in the dataset.

- **Stylometry Features for Author Profiling**: Statistical analysis of word usage can give enough information to detect the author's age, gender, and native language [47,75]. A combination of content-based and style-based features provides a good clue for gender and age detection [75]. Style-based features cover function words and POS such as articles, auxiliary verbs, conjunctions, prepositions, and pronouns. Content-based features can contain the most frequent words in the data where the number of these frequently used words can be chosen based on the data and research goal [75,77]. In addition to the content-based and style-based features, slang words and message length can also provide good information about the author of a text [129]. Peersman et al. [44] used n-grams of words or characters and morphological, lexical, and semantic features for recognizing adults versus adolescents.
  The participants of the PAN2013 competition [47] have mainly used stylistic and content features. Stylistic features contain frequencies of punctuation marks, capital letters, quotations, and POS tags. They also cover emoticons and URL links. Content features were captured using different approaches such as Latent Semantic Analysis, BoW, TF-IDF, dictionary-based words, topic-based words, entropy-based words, sentiment words, emotion words, and slang [47]. The PAN2015 participants [127] applied style-based and content-based features and their combination in n-gram models for feature extraction. They also extracted psycholinguistic features such as polarity words and emotions using NRC (a polarity dictionary that evaluates the polarity value of a word [130]), or LIWC [127,131,132]. The goal of the PAN2016 [79] was age and gender detection from a cross-genre perspective. Many competition participants used stylistic features such as the frequency of using specific words such as function words and slang. For gender detection, they considered sentences that discriminated the female from male, such as 'my wife ...' or 'my man ...' to distinguish men from women. Many participants also combined stylistic features with models such as POS, n-gram models, readability index, and vocabulary richness [79].
  Basile et al. [133] proposed a model that could detect gender and language for all language varieties. They trained their model with data extracted from Twitter from PAN2016 and PAN2017. POS and emojis were extracted as feature sets to distinguish the gender. Using unigram models, they also excluded some specific words to increase the discriminative power of the features [133]. For instance, they considered only words that start with an uppercase letter, or only words that start with a lowercase letter, and so on. In addition, the frequency of geographical names was considered a feature for language detection. Their model was built based on a linear SVM with 3-, 4-, and 5-grams features combined with the mentioned features [133]. PAN2018 [134], participants used deep learning in addition to the traditional feature space, such as content-based or style-based. Some authors combined the traditional feature space, such as stylistic features, with word embedding [134], while some represented the documents with word embedding feature space [135,136]. In [137], the authors combined the POS tag n-grams with syntactic dependencies for gender detection to capture the verbal constructions. Daneshvar and Inkpen [138] extracted various types of word and character n-grams.
  It is critical to discover the users' age and gender as soon as possible to avoid harming children. The main idea is to focus on more vulnerable people in online threats, such as children or teenagers with particular personality traits. Early author profiling (EAP) was proposed by López-Monroy et al. [139] to increase the relevant groups' security by highlighting the target group in the early stages [139]. The feature extraction of EAP was based on meta-word, where word vectors represent the most discriminative features. A clustering method captures the meta-words, and the centroid of each cluster represents a profile meta-word set. The word occurrence in each document can provide the most similar meta-word based on Euclidean distance. Meta-word feature space also captures the semantic relationship between the words [139].
- **Fusion in Author Profiling**: Chat logs can have both text features and keystroke dynamics characteristics. It is shown that a combination of both feature sets can improve the authorship profiling performance where the texts are short and might not provide enough information for predatory detection. As an example, one can consider the work by Li et al. [48], where they collected the chat conversation from the Skype platform and extracted keystroke dynamics and stylometry features to detect the gender of online authors. In addition to the timing feature sets, they considered the ratio of applying the delete key, the number of letters in the word, and the number of characters in each message to train an RF-based gender prediction model [48].

Table 8 summarizes author profiling research works mainly for age and gender detection.

- **Challenges for Reliable Author Profiling** Author profiling confronts some challenges that make it challenging to have a reliable profiling technique. We detail them below:

6 https://pan.webis.de/.
7 http://www.netlog.com.
8 http://blogspot.com.

Table 8
A summary of authorship profiling papers.

Authorship profiling method	Author	Object	ML technique
Keystroke dynamics	Giot et al. [121]	Gender detection	SVM
	Fairhurst et al. [122]	Gender detection	K-NN, DT, NB, Fusion Models
	Pentel et al. [125]	Age & Gender detection	SVM, K-NN, RF
	Tsimperidis et al. [124]	Age detection	ANN
	Tsimperidis et al. [126]	Age detection	RF, SVM, NB, Multi-Layer Perceptron, RBF
Text analysis	Argamon et al. [75]	Age & Gender detection	Multinomial Regression (BMR)
	Koppel et al. [81]	Gender detection	Exponential Gradient Algorithm
	Schler et al. [77]	Age & Gender detection	Multi-Class Real Winnow (MCRW)
	Goswami et al. [129]	Age & Gender detection	Naive Bayes
	Nguyen et al. [46]	Age detection	Logistic & Linear Regression
	López-Monroy et al. [139]	Early author profiling	SVM, Naive Bayes
	Basile et al. [133]	Gender & Language detection	SVM
	Peersman et al. [44]	Age & Gender detection	SVM
	Daneshvar et al. [138]	Gender detection	SVM
Both	Li et al. [48]	Gender detection	Random Forest

- **Data Constraints**: One of the common issues in author profiling is the difficulty of labeling the data. The researchers have used the information provided by online users to label the data with some risks of incorrect labels where users have lied about their age and gender. Author profiling can also be challenging when no suitable training data is available for the model. The problem arises when the training corpus does not have the same pattern as the testing data, and training in such a situation challenges the author profiling. It is not straightforward to detect the age or gender of the authors without accessing the practical training corpus. A cross-domain gender detection was introduced as a solution to cope with this problem [140]. Note that the data size and domain similarities substantially impact the performance of gender detection in cross-domain gender detection [140].
- **Privacy Issues**: Privacy issues play a vital role in using the user's data for any research on author profiling. The regulations like GDPR and national privacy guidelines often prevent social media platforms from disclosing data for research. Further, to build a reliable author profile, a history of the author is often required, and obtaining history needs retrospective data posing a significant challenge for advancing author profiling for predator detection.
- **Low Accuracy**: Author profiling by stylometry or keystroke dynamics has a lower accuracy than physical methods such as fingerprint and face recognition. Even though it is theoretically possible to apply physical techniques for author profiling in social media and chat rooms, it is expensive to implement and is challenged by stricter privacy regulations.

4. Discussion on open problems & potential gaps

This work presented a detailed survey of the latest advancements and challenges of grooming detection in chat logs and social media. This section details the constraints that limit online grooming detection in real-life scenarios.

4.1. Challenges in dataset

A dataset for cyber grooming detection can be challenged with different constraints such as availability, privacy issues, imbalanced essence, non-standard structure, and the unreliability of online data. Accessing private chat conversations is illegal or highly challenging in most countries, making it difficult to collect relevant data for grooming analysis. Testing on actual data is critical for providing the techniques that work on actual grooming datasets and makes applications reliable further. One should also consider that many applications do not collect typing rhythm information. At the same time, it could easily be implemented in future systems, making the cyber grooming detection by KD techniques feasible [23,41].

The necessity of pertinent data also leads to the challenge of the highly imbalanced dataset in grooming detection, where the amount of predatory conversations data is much lower than everyday conversations data [62,72,85]. The imbalanced nature of grooming data will lead to a sub-optimal classifier that gives more weight to one class over the other and results in underfitting or oversampling [86]. It is challenging to train a reliable machine learning model on an imbalanced dataset. The mentioned problem arises where the dataset has a skewed distribution with features such as class overlapping, small sample size, and small disjuncts. The grooming dataset overlaps and disjuncts with non-predatory chatlogs where chatters talk about the same topics in both cases [141]. So, it is critical to design an application that considers the imbalanced essence of the data in this problem and, nevertheless, provides good performance for cyber grooming detection [86].

Internet websites and applications are the only sources that can provide actual data, while the unreliability of the metadata information can challenge it. Users might provide fictitious information on online platforms for various reasons, so metadata information given to the online data from unknown users is not reliable [142,143]. Training machine learning techniques with incorrect labels will lead to inoperative applications. Therefore, researchers should collect data where the metadata is correct and confirmed for having a reliable grooming detection module.

4.2. Topic and context modeling

Grooming conversations do not follow the same pattern as natural language. They have various language themes depending on the characteristics of the predator, and the condition of the chatlog [109]. The predator performs the grooming dialogues so that his aim is unclear to the victim or the family. Predators do not show their motivation explicitly, and the grooming conversation is not of a sexual topic nature in many cases [31]. To minimize risk, predators express their emotions so that the victims trust them while their primary incentive is hidden [110]. A deep understanding of the chatlogs that reveal these dangerous motives can be gained by semantic analysis. Most early research

works have used Context-free models such as word2vec and GloVe models for feature extractions, providing a single word embedding for each term. Contextual models such as BERT[9] can represent each term based on the chat context and lead to a more profound knowledge for semantic analysis [144]. The chat logs have multiple new slang words that may not be available in the learned vocabulary of these language models. A robust semantic analysis can provide better feature vectors for new terms and slang words for training better algorithms.

4.3. Transferability of detection approaches in cross-domain settings

Applying a different domain data such as Google News for training a model and using the model for another domain such as chat logs can challenge the performance of the semantic analysis. Different domains have different context-specific expressions and terms that have different meanings for the new domain's context. The words between different domains often are not discriminative enough to result in high-performance semantic analysis and classification. Many words are domain-specific, and it is challenging to convey well across another domain [145]. For instance, a sentence in a book review with a positive tone such as 'it can take all my time' might reflect a negative connotation in an electronic service of a website [146].

Most deep learning techniques for semantic analysis require a large amount of data for training. In contrast, data collecting in grooming detection is limited, and there is not much available grooming data for training deep learning models. The conventional pre-trained models such as Word2Vec and GloVe can infer the low-level information while gaining more profound information requires extensive training. It is advised to apply Bidirectional Encoder Representations from Transformers (BERT) [95,144] to cope with the transferability in cross-domain for cyber grooming detection as it has profoundly bidirectional contextualization and allows the model to gain information from various representations with different positions [146].

4.4. Cross-language challenges

Machine learning models perform well where the training and testing datasets follow the same feature space and distribution. Distribution variation in feature space where the data language changes can lead to a performance drop [147–150]. Language syntax and semantics can vary based on the language family, and the approaches learned using one language may fail to scale up for another. Using datasets with different languages for the grooming detection problem can be challenged where language variations, vernaculars, dialects, and country status can represent the conversation data differently. Applying labeled data for short text analysis to train the classification model is challenging if the language used in test data differs from the training dataset [151]. Despite the potential problem, no works are attempting to address this problem, and this is mainly due to challenges in accessing similar datasets from a different family of languages.

In earlier research works, transfer learning has been used to cope with cross-language text classification [147–150]. Transfer learning does not require training and testing datasets to be identically distributed. At the same time, the model in the target domain might not need to be trained from scratch, which will reduce the training time and data in many cases [152].

[9] https://github.com/google-research/bert.

4.5. Limited understanding of psychological aspects

Getting a deeper insight into the predators' modus operandi and their motivation has the potential to improve detecting offenders before the crime happens [153]. From the psychological perspective, one of the critical limitations of anonymous studies of pedophiles and the predatory problem is their reliance on self-report for explaining their sexual interest [154]. They might not show their actual behavior due to lack of trust, social desirability, and fear of losing anonymity [154]. One technique to cope with this constraint is to apply online community data where predators have shown their genuine interest without being asked directly. However, it is impossible to understand their actual motives through police records or pure online datasets. Therapists and suitable interviewers must talk with the offenders and analyze their correspondence and written fantasies when predators trust the interviewer completely [153].

An interdisciplinary approach between various areas such as psychology, linguistics, computer scientists, and law enforcement agencies such as police is needed to develop profound knowledge about predators. The interdisciplinary findings would lead to better and more reliable algorithms by exploiting complementary knowledge from different domains.

4.6. Real-time analysis

Most research papers in our review proposed algorithms where entire conversations were analyzed. It means that the detection happens after the harm has occurred. For instance, Gupta et al. [28] created a machine learning model based on affective feature sets that used the whole conversations for detecting six different stages of the predatory conversations. Similarly, in another work by Ringenberg et al. [24], they distinguished contact-driven and fantasy-driven sexual solicitors based on the entire conversations extracted from the PJ website. They did not integrate their models in a real-time situation where harm might have already been inflicted before law enforcement had a chance to prevent it. Only a few research papers tried to detect grooming before and during the incident [155–157]. For instance, Michalopoulos et al. [157] designed a model that monitored the messages during the conversations and sent a warning signal to parents in case of high-risk exploitation. MacFarlane et al. [156] proposed a model considering three main concepts in a message: intentions, locations, and times. In case of detecting all these three concepts, the moderator of the online game can terminate the conversation by blocking the suspicious users avoiding any harmful action. So, the lack of a proper system that prevents grooming leads to a vital need to create a system that integrates into a real-time situation and prevents any harm by detecting the risky content effectively and meaningfully beforehand.

4.7. Deceptive features

Online child groomers might access publicly available data on how children write and learn the writing style by analyzing this data. It leads to a risk of a predator imitating children's writing styles, which can avoid possible detection. The extracted features from the imitated behavior are deceptive and challenge the performance and reliability of automatic grooming detection. For instance, despite syntactic complexity correlating with the deceiver's age, some research studies demonstrated that deceivers can create less complex sentences for grooming purposes [50,158,159]. Molesters can follow this strategy to challenge online grooming detection techniques based on deceptive features

by making less complex sentences. The risk of imitating the victim's behavior by the groomer is inevitable, and implementing an author profile model that detects deception remains challenging.

5. Conclusions

Online platforms allow predators to fake their real identities, decreasing the threat of getting caught. The enormous number of suspicious grooming cases challenges grooming detection, and an automatic surveillance system requires a deep understanding of predatory behavior. We propose an algorithmic survey that systematically details online grooming detection literature focusing on chat conversations. We start our investigation of the grooming problem by looking into child grooming psychological theories and how researchers have applied these theories to define grooming characteristics for automated detection by machine learning models. This research details feature sets, their constraints, and potential solutions to the grooming detection problem. Also, the available datasets are categorized by discussing the restrictions to supplement the readers with the grooming literature. Further, we broadly review various research papers in chat logs for predatory conversation detection and predatory identification. Since molesters might conceal their real identities to trap the victims, this research also investigates various works that applied authorship profiling for age and gender detection to find child groomers by categorizing the authorship profiling literature based on features and datasets. We finalized our survey by discussing constraints that challenge grooming detection, open problems, and possible future solutions.

CRediT authorship contribution statement

Parisa Rezaee Borj: Revision and modification, Investigation, Writing – original draft, Editing. **Kiran Raja:** Revision and modification, Supervision, Reviewing and editing. **Patrick Bours:** Revision, Supervision, Reviewing and editing.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Parisa Rezaee Borj reports financial support, article publishing charges, and equipment, drugs, or supplies were provided by Norwegian University of Science and Technology. Parisa Rezaee Borj reports a relationship with Norwegian University of Science and Technology that includes: employment.

Data availability

No data was used for the research described in the article.

References

[1] M. Gaikwad, S. Ahirrao, S. Phansalkar, K. Kotecha, Online extremism detection: A systematic literature review with emphasis on datasets, classification techniques, validation methods, and tools, in: IEEE Access, Vol. 9, IEEE, 2021, pp. 48364–48404.

[2] M. Gaikwad, S. Ahirrao, S. Phansalkar, K. Kotecha, Multi-ideology ISIS/Jihadist white supremacist (MIWS) dataset for multi-class extremism text classification, in: Data, Vol. 6, MDPI, 2021, p. 117.

[3] S.R. Sahoo, B.B. Gupta, Multiple features based approach for automatic fake news detection on social networks using deep learning, in: Applied Soft Computing, Vol. 100, Elsevier, 2021, 106983.

[4] J.V. Tembhurne, M.M. Almin, T. Diwan, Mc-DNN: Fake news detection using multi-channel deep neural networks, in: International Journal on Semantic Web and Information Systems, Vol. 18, IJSWIS, IGI Global, 2022, pp. 1–20.

[5] S.R. Sahoo, B.B. Gupta, Classification of spammer and nonspammer content in online social network using genetic algorithm-based feature selection, in: Enterprise Information Systems, Vol. 14, Taylor & Francis, 2020, pp. 710–736.

[6] H.A. Bouarara, Recurrent neural network (RNN) to analyse mental behaviour in social media, in: International Journal of Software Science and Computational Intelligence, Vol. 13, IJSSCI, IGI Global, 2021, pp. 1–11.

[7] G. Weimann, Using the internet for terrorist, in: Hypermedia Seduction for Terrorist Recruiting, Vol. 25, IOS Press, 2007, p. 47.

[8] J. Bartlett, L. Reynolds, The State of the Art 2015: A Literature Review of Social Media Intelligence Capabilities for Counter-Terrorism, Demos London, 2015.

[9] C.M. Steel, E. Newman, S. O'Rourke, E. Quayle, An integrative review of historical technology and countermeasure usage trends in online child sexual exploitation material offenders, in: Forensic Science International: Digital Investigation, Vol. 33, Elsevier, 2020, 300971.

[10] C. Peersman, C. Schulze, A. Rashid, M. Brennan, C. Fischer, iCOP: Live forensics to reveal previously unknown criminal media on P2P networks, in: Digital Investigation, Vol. 18, Elsevier, 2016, pp. 50–64.

[11] D. Bright, C. Whelan, S. Harris-Hogan, Exploring the hidden social networks of 'lone actor'terrorists, in: Crime, Law and Social Change, Vol. 74, Springer, 2020, pp. 491–508.

[12] H.C. Whittle, C. Hamilton-Giachritsis, A.R. Beech, Victims' voices: The impact of online grooming and sexual abuse, in: Universal Journal of Psychology, Vol. 1, Citeseer, 2013, pp. 59–71.

[13] E.A. Greene-Colozzi, G.M. Winters, B. Blasko, E.L. Jeglic, Experiences and perceptions of online sexual solicitation and grooming of minors: a retrospective report, in: Journal of Child Sexual Abuse, Vol. 29, Taylor & Francis, 2020, pp. 836–854.

[14] E. Quayle, Prevention, disruption and deterrence of online child sexual exploitation and abuse, in: Era Forum, Vol. 21, Springer, 2020 pp. 429–447.

[15] L. Sanchez, C. Grajeda, I. Baggili, C. Hall, A practitioner survey exploring the value of forensic tools, ai, filtering, & safer presentation for investigating child sexual abuse material (csam), in: Digital Investigation, Vol. 29, Elsevier, 2019, pp. S124–S142.

[16] F. Anda, N.-A. Le-Khac, M. Scanlon, DeepUAge: improving underage age estimation accuracy to aid CSEM investigation, in: Forensic Science International: Digital Investigation, Vol. 32, Elsevier, 2020, 300921.

[17] H.-E. Lee, T. Ermakova, V. Ververis, B. Fabian, Detecting child sexual abuse material: A comprehensive survey, in: Forensic Science International: Digital Investigation, Vol. 34, Elsevier, 2020, 301022.

[18] A. Malm, R. Nash, R. Moghadam, Social network analysis and terrorism, in: The Handbook of the Criminology of Terrorism, Wiley Online Library, 2017, pp. 221–231.

[19] P. Chitrakar, C. Zhang, G. Warner, X. Liao, Social media image retrieval using distilled convolutional neural network for suspicious e-crime and terrorist account detection, in: 2016 IEEE International Symposium on Multimedia, ISM, IEEE, 2016, pp. 493–498.

[20] P. Wisniewski, The privacy paradox of adolescent online safety: A matter of risk prevention or risk resilience? in: IEEE Security & Privacy, Vol. 16, IEEE, 2018, pp. 86–90.

[21] K. Sarikakis, L. Winter, Social media users' legal consciousness about privacy, in: Social Media+ Society, Vol. 3, SAGE Publications, Sage UK: London, England, 2017, 2056305117695325.

[22] B.E. Duffy, N.K. Chan, "You never really know who's looking": Imagined surveillance across social media platforms, in: New Media & Society, Vol. 21, SAGE Publications, Sage UK: London, England, 2019, pp. 119–138.

[23] 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.

[24] T. Ringenberg, K. Misra, K.C. Seigfried-Spellar, J.T. Rayz, Exploring automatic identification of fantasy-driven and contact-driven sexual solicitors, in: 2019 Third IEEE International Conference on Robotic Computing, IRC, IEEE, 2019, pp. 532–537.

[25] M. Mladenović, V. Ošmjanski, S.V. Stanković, Cyber-aggression, cyberbullying, and cyber-grooming: a survey and research challenges, in: ACM Computing Surveys, Vol. 54, CSUR, ACM, New York, NY, USA, 2021, pp. 1–42.

[26] C.H. Ngejane, G. Mabuza-Hocquet, J.H. 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.

[27] A. Razi, S. Kim, A. Alsoubai, G. Stringhini, T. Solorio, M. De Choudhury, P.J. Wisniewski, A human-centered systematic literature review of the computational approaches for online sexual risk detection, in: Proceedings of the ACM on Human-Computer Interaction, Vol. 5, ACM, New York, NY, USA, 2021, pp. 1–38.

[28] A. Gupta, P. Kumaraguru, A. Sureka, Characterizing pedophile conversations on the internet using online grooming, 2012, arXiv preprint arXiv:1208.4324.

[29] S. Craven, S. Brown, E. Gilchrist, Sexual grooming of children: Review of literature and theoretical considerations, in: Journal of Sexual Aggression, Vol. 12, Taylor & Francis, 2006, pp. 287–299.

[30] L.N. Olson, J.L. Daggs, B.L. Ellevold, T.K. Rogers, Entrapping the innocent: Toward a theory of child sexual predators' luring communication, in: Communication Theory, Vol. 17, Oxford University Press, 2007 pp. 231–251.
[31] G.M. Winters, E.L. Jeglic, I knew it all along: The sexual grooming behaviors of child molesters and the hindsight bias, in: Journal of Child Sexual Abuse, Vol. 25, Taylor & Francis, 2016, pp. 20–36.
[32] E. Carmody, T.D. Grant, Online grooming: moves and strategies, in: Language and Law/Linguagem e Direito, Vol. 4, 2017, pp. 103–141, URL https://publications.aston.ac.uk/id/eprint/40097/.
[33] A. Kontostathis, Chatcoder: Toward the tracking and categorization of internet predators, in: Proc. Text Mining Workshop 2009 Held in Conjunction with the Ninth Siam International Conference on Data Mining (SDM 2009). SPARKS, NV. MAY 2009, Citeseer, 2009.
[34] K.M. Babchishin, R. Karl Hanson, C.A. Hermann, The characteristics of online sex offenders: A meta-analysis, in: Sexual Abuse, Vol. 23, Sage Publications, Sage CA: Los Angeles, CA, 2011, pp. 92–123.
[35] M.M. Chiu, K.C. Seigfried-Spellar, T.R. Ringenberg, Exploring detection of contact vs. fantasy online sexual offenders in chats with minors: Statistical discourse analysis of self-disclosure and emotion words, in: Child Abuse & Neglect, Vol. 81, Elsevier, 2018, pp. 128–138.
[36] M.M. Chiu, N. Lehmann-Willenbrock, Statistical discourse analysis: Modeling sequences of individual actions during group interactions across time, in: Group Dynamics: Theory, Research, and Practice, Vol. 20, Educational Publishing Foundation, 2016, p. 242.
[37] L. Miller, Sexual offenses against children: Patterns and motives, in: Aggression and Violent Behavior, Vol. 18, Elsevier, 2013, pp. 506–519.
[38] R. O'Connell, A Typology of Cybersexploitation and Online Grooming Practices, Cyberspace Research Unit, University of Central Lancashire, 2003, pp. 22–41.
[39] 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, in: Child Abuse & Neglect, Vol. 44, Elsevier, 2015 pp. 140–149.
[40] F.E. Gunawan, L. Ashianti, S. Candra, B. Soewito, Detecting online child grooming conversation, in: 2016 11th International Conference on Knowledge, Information and Creativity Support Systems, KICSS, IEEE, 2016 pp. 1–6.
[41] G. Inches, F. Crestani, Overview of the international sexual predator identification competition at PAN-2012, in: CLEF (Online Working Notes/Labs/Workshop), Vol. 30, 2012.
[42] H. Pranoto, F.E. Gunawan, B. Soewito, Logistic models for classifying online grooming conversation, in: Procedia Computer Science, Vol. 59, Elsevier, 2015, pp. 357–365.
[43] 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.
[44] C. Peersman, W. Daelemans, L. Van Vaerenbergh, Predicting age and gender in online social networks, in: Proceedings of the 3rd International Workshop on Search and Mining User-Generated Contents, 2011 pp. 37–44.
[45] C. Zhang, P. Zhang, Predicting Gender from Blog Posts, University of Massachussetts Amherst, USA, 2010.
[46] D. Nguyen, R. Gravel, D. Trieschnigg, T. Meder, " How old do you think I am?" A study of language and age in Twitter, in: Proceedings of the International AAAI Conference on Web and Social Media, Vol. 7, 2013.
[47] F. Rangel, P. Rosso, M. Koppel, E. Stamatatos, G. Inches, Overview of the author profiling task at pan 2013, in: CLEF Conference on Multilingual and Multimodal Information Access Evaluation, CELCT, 2013, pp. 352–365.
[48] G. Li, P.R. Borj, L. Bergeron, P. Bours, Exploring keystroke dynamics and stylometry features for gender prediction on chat data, in: 2019 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics, MIPRO, IEEE, 2019, pp. 1049–1054.
[49] P.R. Borj, P. Bours, Predatory conversation detection, in: 2019 International Conference on Cyber Security for Emerging Technologies, CSET, IEEE, 2019, pp. 1–6.
[50] P.R. Borj, P. Bours, Detecting liars in chats using keystroke dynamics, in: Proceedings of the 2019 3rd International Conference on Biometric Engineering and Applications, 2019, pp. 1–6.
[51] P.R. Borj, K. Raja, P. Bours, On preprocessing the data for improving sexual predator detection: Anonymous for review, in: 2020 15th International Workshop on Semantic and Social Media Adaptation and Personalization (SMA, IEEE, 2020, pp. 1–6.
[52] M.A. Fauzi, P. Bours, Ensemble method for sexual predators identification in online chats, in: 2020 8th International Workshop on Biometrics and Forensics, IWBF, IEEE, 2020, pp. 1–6.
[53] P. Bours, H. Kulsrud, Detection of cyber grooming in online conversation, in: 2019 IEEE International Workshop on Information Forensics and Security, WIFS, IEEE, 2019, pp. 1–6.
[54] S. Ali, H. Abou Haykal, E.Y.M. Youssef, Child sexual abuse and the internet—A systematic review, in: Human Arenas, Springer, 2021 pp. 1–18.
[55] N.R. Sulaiman, M.M. Siraj, Classification of online grooming on chat logs using two term weighting schemes, in: International Journal of Innovative Computing, Vol. 9, 2019.
[56] A. Kontostathis, L. Edwards, J. Bayzick, A. Leatherman, K. Moore, Comparison of rule-based to human analysis of chat logs, in: Communication Theory, Vol. 8, 2009.
[57] D. Michalopoulos, I. Mavridis, Utilizing document classification for grooming attack recognition, in: 2011 IEEE Symposium on Computers and Communications, ISCC, IEEE, 2011, pp. 864–869.
[58] M. Ashcroft, L. Kaati, M. Meyer, A step towards detecting online grooming–Identifying adults pretending to be children, in: 2015 European Intelligence and Security Informatics Conference, IEEE, 2015, pp. 98–104.
[59] D. Bogdanova, P. Rosso, T. Solorio, Exploring high-level features for detecting cyberpedophilia, in: Computer Speech & Language, Vol. 28, Elsevier, 2014, pp. 108–120.
[60] 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.
[61] H.J. Escalante, E. Villatoro-Tello, A. Juárez, M. Montes, L. Villaseñor-Pineda, Sexual predator detection in chats with chained classifiers, in: Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, 2013, pp. 46–54.
[62] C. Cardei, T. Rebedea, Detecting sexual predators in chats using behavioral features and imbalanced learning, in: Nat. Lang. Eng. Vol. 23, 2017 pp. 589–616.
[63] M. Ebrahimi, C.Y. Suen, O. Ormandjieva, A. Krzyzak, Recognizing predatory chat documents using semi-supervised anomaly detection, in: Electronic Imaging, Vol. 2016, Society for Imaging Science and Technology, 2016, pp. 1–9.
[64] M. Ebrahimi, C.Y. Suen, O. Ormandjieva, Detecting predatory conversations in social media by deep convolutional neural networks, in: Digital Investigation, Vol. 18, Elsevier, 2016, pp. 33–49.
[65] H.J. Escalante, E. Villatoro-Tello, S.E. Garza, A.P. López-Monroy, M. Montes-y Gómez, L. Villaseñor-Pineda, Early detection of deception and aggressiveness using profile-based representations, in: Expert Systems with Applications, Vol. 89, Elsevier, 2017, pp. 99–111.
[66] K. Misra, H. Devarapalli, T.R. Ringenberg, J.T. Rayz, Authorship analysis of online predatory conversations using character level convolution neural networks, in: 2019 IEEE International Conference on Systems, Man and Cybernetics, SMC, IEEE, 2019, pp. 623–628.
[67] C. Morris, Identifying Online Sexual Predators by SVM Classification with Lexical and Behavioral Features (Master of Science Thesis), University of Toronto, Canada, 2013.
[68] F. Muñoz, G. Isaza, L. Castillo, SMARTSEC4cop: Smart cyber-grooming detection using natural language processing and convolutional neural networks, in: International Symposium on Distributed Computing and Artificial Intelligence, Springer, 2020, pp. 11–20.
[69] C. Ngejane, J. Eloff, T. Sefara, V. Marivate, Digital forensics supported by machine learning for the detection of online sexual predatory chats, in: Forensic Science International: Digital Investigation, Vol. 36, Elsevier, 2021, 301109.
[70] J. Parapar, D.E. Losada, A. Barreiro, Combining psycho-linguistic, content-based and chat-based features to detect predation in chatrooms, in: J. UCS, Vol. 20, 2014, pp. 213–239.
[71] E. Villatoro-Tello, A. Juárez-González, H.J. Escalante, M. Montes-y Gómez, L.V. Pineda, A two-step approach for effective detection of misbehaving users in chats, in: CLEF (Online Working Notes/Labs/Workshop), Vol. 1178, 2012.
[72] Z. Zuo, J. Li, P. Anderson, L. Yang, N. Naik, Grooming detection using fuzzy-rough feature selection and text classification, in: 2018 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), IEEE, 2018, pp. 1–8.
[73] Z. Zuo, J. Li, B. Wei, L. Yang, F. Chao, N. Naik, Adaptive activation function generation for artificial neural networks through fuzzy inference with application in grooming text categorisation, in: 2019 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), IEEE, 2019, pp. 1–6.
[74] Y.-G. Cheong, A.K. Jensen, E.R. Guðnadóttir, B.-C. Bae, J. Togelius, Detecting predatory behavior in game chats, in: IEEE Transactions on Computational Intelligence and AI in Games, Vol. 7, IEEE, 2015, pp. 220–232.
[75] S. Argamon, M. Koppel, J.W. Pennebaker, J. Schler, Automatically profiling the author of an anonymous text, in: Communications of the ACM, Vol. 52, ACM, New York, NY, USA, 2009, pp. 119–123.
[76] A. Mukherjee, B. Liu, Improving gender classification of blog authors, in: Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, 2010, pp. 207–217.
[77] J. Schler, M. Koppel, S. Argamon, J.W. Pennebaker, Effects of age and gender on blogging, in: AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs, Vol. 6, 2006, pp. 199–205.

[78] J.D. Burger, J. Henderson, G. Kim, G. Zarrella, Discriminating Gender on Twitter, Tech. rep., Mitre Corp Bedford MA Bedford United States, 2011.
[79] F. Rangel, P. Rosso, B. Verhoeven, W. Daelemans, M. Potthast, B. Stein, Overview of the 4th author profiling task at PAN 2016: cross-genre evaluations, in: Working Notes Papers of the CLEF, Vol. 2016, 2016 pp. 750–784.
[80] F. Rangel, P. Rosso, M. Potthast, B. Stein, Overview of the 5th author profiling task at pan 2017: Gender and language variety identification in twitter, in: Working Notes Papers of the CLEF, 2017, 1613–0073.
[81] M. Koppel, S. Argamon, A.R. Shimoni, Automatically categorizing written texts by author gender, in: Literary and Linguistic Computing, Vol. 17, Oxford University Press, 2002 pp. 401–412.
[82] S. Argamon, M. Koppel, J. Fine, A.R. Shimoni, Gender, genre, and writing style in formal written texts, in: Text-the Hague then Amsterdam then Berlin-, Vol. 23, Walter De Gryuter & Co, 2003, pp. 321–346.
[83] J. Tam, C.H. Martell, Age detection in chat, in: 2009 IEEE International Conference on Semantic Computing, IEEE, 2009, pp. 33–39.
[84] J. Lin, Automatic Author Profiling of Online Chat Logs, Tech. rep., Naval Postgraduate School, Monterey CA, 2007.
[85] M. Latapy, C. Magnien, R. Fournier, Quantifying paedophile queries in a large p2p system, in: 2011 Proceedings IEEE INFOCOM, IEEE, 2011, pp. 401–405.
[86] P.R. Borj, K. Raja, P. Bours, Detecting sexual predatory chats by perturbed data and balanced ensembles, in: 2021 International Conference of the Biometrics Special Interest Group, BIOSIG, IEEE, 2021, pp. 1–5.
[87] K.S. Jones, A statistical interpretation of term specificity and its application in retrieval, in: Journal of Documentation, MCB UP Ltd, 1972.
[88] M. Lan, C.L. Tan, J. Su, Y. Lu, Supervised and traditional term weighting methods for automatic text categorization, in: IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 31, IEEE, 2008 pp. 721–735.
[89] H.C. Wu, R.W.P. Luk, K.F. Wong, K.L. Kwok, Interpreting tf-idf term weights as making relevance decisions, in: ACM Transactions on Information Systems (TOIS), Vol. 26, ACM, New York, NY, USA, 2008 pp. 1–37.
[90] G. Salton, C. Buckley, Term-weighting approaches in automatic text retrieval, in: Information Processing & Management, Vol. 24, Elsevier, 1988, pp. 513–523.
[91] S. Robertson, Understanding inverse document frequency: on theoretical arguments for IDF, in: Journal of Documentation, Vol. 60, Emerald Group Publishing Limited, 2004, pp. 503–520.
[92] D. Bogdanova, P. Rosso, T. Solorio, On the impact of sentiment and emotion based features in detecting online sexual predators, in: Proceedings of the 3rd Workshop in Computational Approaches to Subjectivity and Sentiment Analysis, 2012, pp. 110–118.
[93] X. Rong, Word2vec parameter learning explained, 2014, arXiv preprint arXiv:1411.2738.
[94] J. Pennington, R. Socher, C.D. Manning, Glove: Global vectors for word representation, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP, 2014, pp. 1532–1543.
[95] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, 2018, arXiv preprint arXiv:1810.04805.
[96] M.E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, L. Zettlemoyer, Deep contextualized word representations, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), Association for Computational Linguistics, New Orleans, Louisiana, 2018, pp. 2227–2237, http://dx.doi.org/10.18653/v1/N18-1202, URL https://aclanthology.org/N18-1202.
[97] A. Joulin, E. Grave, P. Bojanowski, T. Mikolov, Bag of tricks for efficient text classification, 2016, arXiv preprint arXiv:1607.01759.
[98] C. Barber, S. Bettez, Deconstructing the online grooming of youth: Toward improved information systems for detection of online sexual predators, in: International Conference on Information Systems, ICIS, AIS eLibrary, 2014.
[99] J.W. Pennebaker, R.L. Boyd, K. Jordan, K. Blackburn, The Development and Psychometric Properties of LIWC2015, Tech. rep., 2015.
[100] P.S. Teh, A.B.J. Teoh, S. Yue, A survey of keystroke dynamics biometrics, in: The Scientific World Journal, Vol. 2013, Hindawi, 2013.
[101] S.P. Banerjee, D.L. Woodard, Biometric authentication and identification using keystroke dynamics: A survey, in: Journal of Pattern Recognition Research, Vol. 7, Citeseer, 2012, pp. 116–139.
[102] C. Epp, M. Lippold, R.L. Mandryk, Identifying emotional states using keystroke dynamics, in: Proceedings of the Sigchi Conference on Human Factors in Computing Systems, 2011, pp. 715–724.
[103] A. Kołakowska, A review of emotion recognition methods based on keystroke dynamics and mouse movements, in: 2013 6th International Conference on Human System Interactions, HSI, IEEE, 2013, pp. 548–555.
[104] A. Kołakowska, Recognizing emotions on the basis of keystroke dynamics, in: 2015 8th International Conference on Human System Interaction, HSI, IEEE, 2015, pp. 291–297.
[105] R. Bixler, S. D'Mello, Detecting boredom and engagement during writing with keystroke analysis, task appraisals, and stable traits, in: Proceedings of the 2013 International Conference on Intelligent User Interfaces, 2013, pp. 225–234.
[106] J. Parapar, D.E. Losada, A. Barreiro, A learning-based approach for the identification of sexual predators in chat logs, in: CLEF (Online Working Notes/Labs/Workshop), Vol. 1178, 2012.
[107] G. Eriksson, J. Karlgren, Features for modelling characteristics of conversations: Notebook for PAN at CLEF 2012, in: CLEF 2012 Evaluation Labs and Workshop, Rome, Italy, 17-20 September 2012, 2012.
[108] D. Finkelhor, Current information on the scope and nature of child sexual abuse, in: The Future of Children, JSTOR, 1994, pp. 31–53.
[109] V. Egan, J. Hoskinson, D. Shewan, Perverted justice: A content analysis of the language used by offenders detected attempting to solicit children for sex, Antisocial Behav.: Causes Correl. Treat. 20 (3) (2011) 273297.
[110] N. Lorenzo-Dus, C. Izura, "Cause ur special": Understanding trust and complimenting behaviour in online grooming discourse, in: Journal of Pragmatics, Vol. 112, Elsevier, 2017, pp. 68–82.
[111] G. Celeux, G. Govaert, A classification EM algorithm for clustering and two stochastic versions, in: Computational Statistics & Data Analysis, Vol. 14, Elsevier, 1992, pp. 315–332.
[112] Y.R. Tausczik, J.W. Pennebaker, The psychological meaning of words: LIWC and computerized text analysis methods, in: Journal of Language and Social Psychology, Vol. 29, Sage Publications, Sage CA: Los Angeles, CA, 2010, pp. 24–54.
[113] H. He, E.A. Garcia, Learning from imbalanced data, in: IEEE Transactions on Knowledge and Data Engineering, Vol. 21, Ieee, 2009, pp. 1263–1284.
[114] X.-Y. Liu, J. Wu, Z.-H. Zhou, Exploratory undersampling for class-imbalance learning, in: IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), Vol. 39, IEEE, 2008, pp. 539–550.
[115] N.V. Chawla, N. Japkowicz, A. Kotcz, Special issue on learning from imbalanced data sets, in: ACM SIGKDD Explorations Newsletter, Vol. 6, ACM, New York, NY, USA, 2004, pp. 1–6.
[116] G. Dulac-Arnold, L. Denoyer, P. Gallinari, Text classification: A sequential reading approach, in: European Conference on Information Retrieval, Springer, 2011, pp. 411–423.
[117] C. Strapparava, R. Mihalcea, Semeval-2007 task 14: Affective text, in: Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007), 2007, pp. 70–74.
[118] D. Hope, Java wordnet similarity library, 2008.
[119] P. Resnik, Using information content to evaluate semantic similarity in a taxonomy, 1995, arXiv preprint cmp-lg/9511007.
[120] J.W. Pennebaker, M.R. Mehl, K.G. Niederhoffer, Psychological aspects of natural language use: Our words, our selves, in: Annual Review of Psychology, Vol. 54, Annual Reviews, 4139 El Camino Way, PO Box 10139, Palo Alto, CA 94303-0139, USA, 2003, pp. 547–577.
[121] R. Giot, C. Rosenberger, A new soft biometric approach for keystroke dynamics based on gender recognition, in: International Journal of Information Technology and Management, Vol. 11, Inderscience Publishers Ltd, 2012, pp. 35–49.
[122] M. Fairhurst, M. Da Costa-Abreu, Using keystroke dynamics for gender identification in social network environment, in: 4th International Conference on Imaging for Crime Detection and Prevention 2011 (ICDP 2011), IET, 2011, pp. 1–6.
[123] S.Z.S. Idrus, E. Cherrier, C. Rosenberger, P. Bours, Soft biometrics for keystroke dynamics: Profiling individuals while typing passwords, in: Computers & Security, Vol. 45, Elsevier, 2014, pp. 147–155.
[124] I. Tsimperidis, S. Rostami, V. Katos, Age detection through keystroke dynamics from user authentication failures, in: International Journal of Digital Crime and Forensics, Vol. 9, IJDCF, IGI Global, 2017, pp. 1–16.
[125] A. Pentel, Predicting age and gender by keystroke dynamics and mouse patterns, in: Adjunct Publication of the 25th Conference on User Modeling, Adaptation and Personalization, 2017, pp. 381–385.
[126] I. Tsimperidis, A. Arampatzis, A. Karakos, Keystroke dynamics features for gender recognition, in: Digital Investigation, Vol. 24, Elsevier, 2018 pp. 4–10.
[127] F. Rangel, P. Rosso, M. Potthast, B. Stein, W. Daelemans, Overview of the 3rd author profiling task at PAN 2015, in: CLEF, sn, 2015, p. 2015.
[128] F. Rangel, P. Rosso, M. Montes-y Gómez, M. Potthast, B. Stein, Overview of the 6th author profiling task at pan 2018: multimodal gender identification in twitter, in: Working Notes Papers of the CLEF, 2018.
[129] S. Goswami, S. Sarkar, M. Rustagi, Stylometric analysis of bloggers' age and gender, in: Proceedings of the International AAAI Conference on Web and Social Media, Vol. 3, 2009, pp. 214–217.
[130] S.M. Mohammad, S. Kiritchenko, X. Zhu, NRC-Canada: Building the state-of-the-art in sentiment analysis of tweets, 2013, arxiv preprint arXiv:1308.6242.

[131] M. Arroju, A. Hassan, G. Farnadi, Age, gender and personality recognition using tweets in a multilingual setting, in: 6th Conference and Labs of the Evaluation Forum (CLEF 2015), Springer, 2015, pp. 22–31.
[132] M. Giménez, D.I. Hernández, F. Pla, Segmenting target audiences: Automatic author profiling using tweets, in: CEUR Workshop Proceedings, 2015.
[133] A. Basile, G. Dwyer, M. Medvedeva, J. Rawee, H. Haagsma, M. Nissim, N-GrAM: New groningen author-profiling model—Notebook for PAN at clef 2017, in: CEUR Workshop Proceedings, Vol. 1866, 2017.
[134] B.G. Patra, K.G. Das, D. Das, Multimodal author profiling for Twitter, in: Notebook for PAN at CLEF, 2018.
[135] R. Veenhoven, S. Snijders, D. van der Hall, R. van Noord, Using translated data to improve deep learning author profiling models, in: Proceedings of the Ninth International Conference of the CLEF Association (CLEF 2018), Vol. 2125, 2018.
[136] R.K. Bayot, T. Gonçalves, Multilingual author profiling using lstms, in: Proceedings of the Ninth International Conference of the CLEF Association (CLEF 2018), 2018.
[137] J. Karlgren, L. Esposito, C. Gratton, P. Kanerva, Authorship profiling without using topical information: Notebook for PAN at CLEF 2018, in: CLEF (Working Notes), 2018.
[138] S. Daneshvar, D. Inkpen, Gender identification in twitter using N-grams and LSA, in: Proceedings of the Ninth International Conference of the CLEF Association (CLEF 2018), CEUR-WS, 2018.
[139] A.P. López-Monroy, F.A. González, T. Solorio, Early author profiling on Twitter using profile features with multi-resolution, in: Expert Systems with Applications, Vol. 140, Elsevier, 2020, 112909.
[140] R. Dias, I. Paraboni, Cross-domain author gender classification in Brazilian portuguese, in: Proceedings of the 12th Language Resources and Evaluation Conference, 2020, pp. 1227–1234.
[141] W.-C. Lin, C.-F. Tsai, Y.-H. Hu, J.-S. Jhang, Clustering-based undersampling in class-imbalanced data, in: Information Sciences, Vol. 409, Elsevier, 2017, pp. 17–26.
[142] L. Reichart Smith, K.D. Smith, M. Blazka, Follow me, what's the harm: Considerations of catfishing and utilizing fake online personas on social media, in: J. Legal Aspects Sport, Vol. 27, HeinOnline, 2017, p. 32.
[143] A. Romanov, A. Semenov, O. Mazhelis, J. Veijalainen, Detection of fake profiles in social media-literature review, in: International Conference on Web Information Systems and Technologies, Vol. 2, SCITEPRESS, 2017 pp. 363–369.
[144] A. Rogers, O. Kovaleva, A. Rumshisky, A primer in bertology: What we know about how bert works, in: Transactions of the Association for Computational Linguistics, Vol. 8, MIT Press, 2020, pp. 842–866.
[145] L. Li, W. Ye, M. Long, Y. Tang, J. Xu, J. Wang, Simultaneous learning of pivots and representations for cross-domain sentiment classification, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 34, 2020, pp. 8220–8227.
[146] B. Myagmar, J. Li, S. Kimura, Transferable high-level representations of bert for cross-domain sentiment classification, in: Proceedings on the International Conference on Artificial Intelligence, ICAI, The Steering Committee of The World Congress in Computer Science, Computer, 2019, pp. 135–141.
[147] C. Wan, R. Pan, J. Li, Bi-weighting domain adaptation for cross-language text classification, in: Twenty-Second International Joint Conference on Artificial Intelligence, 2011.
[148] J. Zhou, S. Pan, I. Tsang, S.-S. Ho, Transfer learning for cross-language text categorization through active correspondences construction, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 30, 2016.
[149] J. Huang, O. Kuchaiev, P. O'Neill, V. Lavrukhin, J. Li, A. Flores, G. Kucsko, B. Ginsburg, Cross-language transfer learning, continuous learning, and domain adaptation for end-to-end automatic speech recognition, 2020, arxiv preprint arXiv:2005.04290.
[150] C. Yu, Y. Chen, Y. Li, M. Kang, S. Xu, X. Liu, Cross-language end-to-end speech recognition research based on transfer learning for the low-resource Tujia language, in: Symmetry, Vol. 11, Multidisciplinary Digital Publishing Institute, 2019, p. 179.
[151] M. Imran, P. Mitra, J. Srivastava, Cross-language domain adaptation for classifying crisis-related short messages, 2016, arXiv preprint arXiv:1602.05388.
[152] C. Tan, F. Sun, T. Kong, W. Zhang, C. Yang, C. Liu, A survey on deep transfer learning, in: International Conference on Artificial Neural Networks, Springer, 2018, pp. 270–279.
[153] J.R. Blalock, M.L. Bourke, A content analysis of pedophile manuals, in: Aggression and Violent Behavior, Elsevier, 2020, 101482.
[154] S.J. Jones, C. Ó Ciardha, I.A. Elliott, Identifying the coping strategies of nonoffending pedophilic and hebephilic individuals from their online forum posts, in: Sexual Abuse, SAGE Publications, Sage CA: Los Angeles, CA, 2020, 1079063220965953.
[155] L. Penna, A. Clark, G. Mohay, A framework for improved adolescent and child safety in MMOs, in: 2010 International Conference on Advances in Social Networks Analysis and Mining, IEEE, 2010, pp. 33–40.
[156] K. MacFarlane, V. Holmes, Agent-mediated information exchange: Child safety online, in: 2009 International Conference on Management and Service Science, IEEE, 2009, pp. 1–5.
[157] D. Michalopoulos, E. Papadopoulos, I. Mavridis, Artemis: protection from sexual exploitation attacks via SMS, in: 2012 16th Panhellenic Conference on Informatics, IEEE, 2012, pp. 19–24.
[158] B.M. DePaulo, J.J. Lindsay, B.E. Malone, L. Muhlenbruck, K. Charlton, H. Cooper, Cues to deception, in: Psychological Bulletin, Vol. 129, American Psychological Association, 2003, p. 74.
[159] V. Pérez-Rosas, R. Mihalcea, Experiments in open domain deception detection, in: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, 2015, pp. 1120–1125.